Available online at www.sciencedirect.com
ScienceDirect
Journal of Econometrics 137 (2007) 134–161
JOURNAL OF Econometrics
www.elsevier.com/locate/jeconom

Selection of estimation window in the presence of breaks

M. Hashem Pesaran[a,*], Allan Timmermann[b]

[a] University of Cambridge, UK
[b] University of California, San Diego, USA

Available online 24 April 2006

Abstract

In situations where a regression model is subject to one or more breaks it is shown that it can be optimal to use pre-break data to estimate the parameters of the model used to compute out-of-sample forecasts. The issue of how best to exploit the trade-off that might exist between bias and forecast error variance is explored and illustrated for the multivariate regression model under the assumption of strictly exogenous regressors. In practice when this assumption cannot be maintained and both the time and size of the breaks are unknown, the optimal choice of the observation window will be subject to further uncertainties that make exploiting the bias–variance trade-off difficult. To that end we propose a new set of cross-validation methods for selection of a single estimation window and weighting or pooling methods for combination of forecasts based on estimation windows of different lengths. Monte Carlo simulations are used to show when these procedures work well compared with methods that ignore the presence of breaks.

JEL classification: C22; C53

Keywords: Parameter instability; Forecasting with breaks; Choice of observation window; Forecast combination

1. Introduction

Consider the multivariate regression model subject to a single structural break at time $t = T_1$:

$$y_{t+1} = \mathbf{1}_{\{t \leqslant T_1\}}\boldsymbol{\beta}_1'\mathbf{x}_t + (1 - \mathbf{1}_{\{t \leqslant T_1\}})\boldsymbol{\beta}_2'\mathbf{x}_t + u_{t+1}, \quad t = 1, 2, \ldots, T. \tag{1}$$

[*] Corresponding author.
E-mail address: mhpl@econ.cam.ac.uk (M.H. Pesaran).

doi:10.1016/j.jeconom.2006.03.010

Here $\mathbf{1}_{\{t \leqslant T_1\}}$ is an indicator variable that takes the value of one when $t \leqslant T_1$, and is zero otherwise, while $\mathbf{x}_t$ is a $p \times 1$ vector of stochastic regressors, $\boldsymbol{\beta}_i$ $(i = 1, 2)$ are $p \times 1$ vectors of regression coefficients, and u_{t+1} is a serially uncorrelated error term with mean zero, assumed to be independently distributed of $\mathbf{x}_t$, possibly with a shift in its variance from σ_1^2 to σ_2^2 at the time of the break. Assuming that $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$ or $\sigma_1^2 \neq \sigma_2^2$, it follows that there is a structural break in the data generating process at time T_1 which we refer to as the break point.

Suppose that we know that $\boldsymbol{\beta}$ and/or σ have changed at T_1 and our interest lies in forecasting y_{T+1} given the observations $\{y_t, t = 1, \ldots, T\}$ and $\{\mathbf{x}_t, t = 1, \ldots, T\}$ which we collect in the information set $\Omega_T = \{\mathbf{x}_t, y_t\}_{t=1}^{T}$. How many observations should we use to estimate a model that, when used to generate forecasts, will minimize the mean squared forecast error (MSFE)? Here we are not concerned with the classical problem of identifying the exact point of the break, but rather the sample size that it is optimal to use to estimate the parameters of the model in order to forecast out-of-sample on the assumption that a structural break has in fact occurred. The standard solution is to use only observations over the post-break period $(t = T_1 + 1, \ldots, T)$ to estimate the model.

We show in this paper that this solution need not be optimal when the objective is to optimize forecasting performance. The intuition behind using pre-break data is very simple in the case where the regressors are strictly exogenous. If structural breaks characterize a particular time series, using the full historical data series to estimate a forecasting model will lead to forecast errors that are no longer unbiased, although they may have a lower variance. Provided that the break is not too large, pre-break data will be informative for forecasting outcomes even after the break. By trading off the bias and forecast error variance, one can potentially improve on the forecasting performance as measured by MSFE.

For the more general case where the regressors are not strictly exogenous, the potential gains in forecasting performance from using pre-break data can readily be demonstrated through Monte Carlo simulations. Not surprisingly, such gains can be quite large if the time and the size of the break are known. In the more common situation where these parameters are unknown, our simulations show that forecasting accuracy can be improved by pre-testing for a structural break. However, the overall outcome crucially depends on how well the location of the break point is estimated. In particular, information about the time of the break can be important to forecasting performance even in the absence of knowledge about the size of a break.

The main contributions of our paper are two-fold. First, we make a simple, yet largely overlooked theoretical point, namely that when the objective is to minimize out-of-sample MSFE it is often optimal to use pre-break data to estimate forecasting models on data samples subject to structural breaks. To establish this point we make a set of simplifying assumptions (such as strict exogeneity of the regressors) which are unlikely to hold empirically but allow us to get a set of theoretical results that are easy to understand. These assumptions also let us demonstrate analytically the factors determining the optimal estimation window both conditionally and unconditionally.

Our second main contribution is to propose a set of new methods that facilitate the practical implementation of the theoretical point that using pre-break data for parameter estimation can lead to improved forecasting performance. Although our paper focuses on the multivariate regression model, these methods are applicable to general dynamic models and are hence useful in a wide set of circumstances. It is commonplace to find that the timing of breaks as well as the size of any shift in parameters are poorly estimated. For this reason the proposed methods do not attempt to directly exploit the trade-off between the bias and forecast error variance but do so indirectly by searching across different starting points for the data window used to estimate the parameters of the forecasting model.

More specifically, we consider approaches based on cross-validation and forecast combination (pooling) procedures. Cross-validation based on pseudo-out-of-sample forecasting performance can be used as a criterion for selecting estimation windows of different length. One approach is to simply choose a *single* window size that leads to the lowest average out-of-sample loss over the cross-validation sample. This may work well if the break is well-defined and large. Another approach, which is likely to work better for relatively small breaks, is to *combine* forecasts based on different estimation windows, using equal weights or weights that are proportional to the inverse of the out-of-sample loss. Such methods may or may not condition on estimates of the break points and we consider both cases. Therefore, for this latter approach we shall consider forecast combinations with alternative weighting schemes and depending on whether knowledge of the breakpoints is used. This approach is inspired by the forecast combination literature where forecasts from alternative models are combined.[1] Here we pool forecasts obtained using the same model but estimated across different observation windows. Clearly, a meta forecast combination procedure that considers both sources of model uncertainty can be entertained. The forecast combination strategy can also be seen as a risk diversification strategy in the face of uncertainty regarding the breaks, and extends concerns over model uncertainty to a wider class of models where regression equations subject to breaks are viewed as different models.

The outline of the paper is as follows. Section 2 sets up the multivariate breakpoint model and considers the choice of observation window for the conditional and (for a simple data generating process) unconditional forecasting problem. Section 3 provides details of the implementation of the trade-off, cross-validation and combination methods. Section 4 reports Monte Carlo simulation results while Section 5 concludes.

2. Selection of the optimal window size

Many tests for structural breaks have been proposed. In the context of linear regression models Chow (1960) derived F-tests for structural breaks with a known break point, while Brown et al. (1975) derived Cusum and Cusum squared tests that are also applicable when the time of the break is unknown. More recently, contributions by Ploberger et al. (1989), Hansen (1992), Andrews (1993), Inclan and Tiao (1994), Andrews and Ploberger (1996) and Chu et al. (1996) have extended ests for the presence of breaks to account for heteroskedasticity and dynamics. Methods for estimating the size and timing of multiple break points have also been developed, see Bai and Perron (1998, 2003) and Altissimo and Corradi (2003).

Less attention has been paid to the problem of determining the size of the estimation window in the presence of breaks. To analyse this question, let m denote the starting point of the sample of the most recent observations to be used in estimation for the purpose of forecasting y_{T+1} based on the multivariate regression model (1) and information Ω_T. Also, let $\mathbf{X}_{m,T}$ be the $(T-m+1)\times p$ matrix of observations on the regressors such that

[1] For reviews of the forecast combination literature see Clemen (1989), Granger (1989), Diebold and Lopez (1996), and more recently Newbold and Harvey (2002) and Timmermann (2006).

$rank(\mathbf{X}_{m,T}) = p$, while $\mathbf{Y}_{m,T}$ is the $(T-m+1)\times 1$ vector of observations on the dependent variable whose value for period $T+1$ we are interested in forecasting. Defining the quadratic form $\mathbf{Q}_{\tau,T_i} = \mathbf{X}'_{\tau,T_i}\mathbf{X}_{\tau,T_i}$ so that $\mathbf{Q}_{\tau,T_i} = 0$ if $\tau > T_i$, the OLS estimator of $\boldsymbol{\beta}$ based on the sample from m to T $(m < T - p + 1)$ is given by

$$\widehat{\boldsymbol{\beta}}_T(m) = \mathbf{Q}^{-1}_{m,T}\mathbf{X}'_{m,T}\mathbf{Y}_{m,T}. \tag{2}$$

The error in forecasting y_{T+1} from (1) will be a function of the data sample used to estimate $\boldsymbol{\beta}$ and is given by

$$e_{T+1}(m) = y_{T+1} - \widehat{y}_{T+1} = (\boldsymbol{\beta}_2 - \widehat{\boldsymbol{\beta}}_T(m))'\mathbf{x}_T + u_{T+1}. \tag{3}$$

We implicitly assume that it is known that there is no break in the regression model in period $T+1$. Otherwise, the best forecast would need to consider the distribution from which new regression parameters are drawn after a break.[2]

To derive analytical results, we assume that $\mathbf{x}_t$ is strictly exogenous, in the sense that it is independently distributed of u_s for all s and t, such that

$$\mathrm{E}[\mathbf{x}_t | f(u_1, u_2, \ldots, u_{T+1})] = 0 \quad \text{for } t = 1, \ldots, T, \tag{4}$$

where $f(.)$ is a general function. In particular, it follows that $\mathrm{E}(\mathbf{x}_t|u_s) = 0$, for all t and s. While this assumption is clearly not empirically appropriate in many situations, it simplifies the analysis considerably and allows us to derive much stronger and clearer results than would otherwise be possible.

Autoregressive processes subject to structural breaks have been considered by Pesaran and Timmermann (2005). The advantage of that setup is that it is highly relevant to many empirical situations, but the theoretical results are complicated to interpret. Under strictly exogenous regressors, however, the factors determining the optimal window length under breaks become very transparent. This case therefore serves as a natural theoretical benchmark.

2.1. Conditional MSFE results

We shall consider the case where the prediction can be conditioned on the sequence of $\mathbf{x}_t$ values. Taking expectations of (3) conditional on $\mathbf{X}_T = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$, we get the conditional bias in the forecast error:

$$bias(m|\mathbf{X}_T) \equiv \mathrm{E}[e_{T+1}(m)|\mathbf{X}_T] = (\boldsymbol{\beta}_2 - \widehat{\boldsymbol{\beta}}_T(m))'\mathbf{x}_T. \tag{5}$$

Furthermore, using (2) it can be shown that

$$\begin{aligned} e_{T+1}(m) &= (\boldsymbol{\beta}'_2 - \mathbf{Y}'_{m,T}\mathbf{X}_{m,T}\mathbf{Q}^{-1}_{m,T})\mathbf{x}_T + u_{T+1} \\ &= (\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1)'\mathbf{Q}_{m,T_1}\mathbf{Q}^{-1}_{m,T}\mathbf{x}_T - \mathbf{u}'_{m,T}\mathbf{X}_{m,T}\mathbf{Q}^{-1}_{m,T}\mathbf{x}_T + u_{T+1}, \end{aligned} \tag{6}$$

where $\mathbf{u}_{m,T} = (u_m, u_{m+1}, \ldots, u_T)'$.

Squaring this expression and taking expectations, the conditional MSFE given $\mathbf{X}_T$ can be computed as follows:

$$\begin{aligned} MSFE(m|\mathbf{X}_T) &= \mathrm{E}[e^2_{T+1}(m)|\mathbf{X}_T] \\ &= \sigma^2_2 + \sigma^2_2(\boldsymbol{\mu}'\mathbf{Q}_{m,T_1}\mathbf{Q}^{-1}_{m,T}\mathbf{x}_T)^2 + \mathbf{x}'_T\mathbf{Q}^{-1}_{m,T}\mathbf{X}'_{m,T}\boldsymbol{\Sigma}_{m,T}\mathbf{X}_{m,T}\mathbf{Q}^{-1}_{m,T}\mathbf{x}_T, \end{aligned} \tag{7}$$

[2] A Bayesian procedure that allows for such a possibility is discussed in Pesaran et al. (2006).

where $\boldsymbol{\mu} = (\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1)/\sigma_2$, and $\boldsymbol{\Sigma}_{m,T} = \mathrm{E}(\mathbf{u}_{m,T}\mathbf{u}'_{m,T})$ is a $(T-m+1) \times (T-m+1)$ diagonal matrix with σ_1^2 on the first $T_1 - m + 1$ diagonal places and σ_2^2 on the remaining $T - T_1$ places. Using (7) note that

$$MSFE(m|\mathbf{X}_T) = \sigma_2^2(1 + \mathbf{B}_m + \mathbf{E}_m), \tag{8}$$

where $\mathbf{B}_m$ is the squared bias and $\mathbf{E}_m$ the efficiency term defined by

$$\mathbf{B}_m = \boldsymbol{\mu}'\mathbf{Q}_{m,T_1}\mathbf{Q}_{m,T}^{-1}\mathbf{x}_T\mathbf{x}'_T\mathbf{Q}_{m,T}^{-1}\mathbf{Q}_{m,T_1}\boldsymbol{\mu}, \tag{9}$$

and

$$\begin{aligned}\mathbf{E}_m &= \sigma_2^{-2}\mathbf{x}'_T\mathbf{Q}_{m,T}^{-1}(\sigma_1^2\mathbf{Q}_{m,T_1} + \sigma_2^2\mathbf{Q}_{T_1+1,T})\mathbf{Q}_{m,T}^{-1}\mathbf{x}_T \\ &= \mathbf{x}'_T\mathbf{Q}_{m,T}^{-1}\mathbf{x}_T + \psi(\mathbf{x}'_T\mathbf{Q}_{m,T}^{-1}\mathbf{Q}_{m,T_1}\mathbf{Q}_{m,T}^{-1}\mathbf{x}_T).\end{aligned} \tag{10}$$

Here the (proportional) break in the innovation variance parameter is given by $\psi = (\sigma_1^2 - \sigma_2^2)/\sigma_2^2$. For a given value of $\mathbf{x}_T$, and $1 < m \leqslant T_1 < T$, it is clear that since $\mathbf{Q}_{m,T} - \mathbf{Q}_{m+1,T} = \mathbf{x}_{m-1}\mathbf{x}'_{m-1} \geqslant 0$, then

$$\mathbf{x}'_T(\mathbf{Q}_{m,T}^{-1} - \mathbf{Q}_{m+1,T}^{-1})\mathbf{x}_T \leqslant 0, \tag{11}$$

and similarly[3]

$$\mathbf{x}'_T(\mathbf{Q}_{m,T}^{-1}\mathbf{Q}_{m,T_1}\mathbf{Q}_{m,T}^{-1} - \mathbf{Q}_{m+1,T}^{-1}\mathbf{Q}_{m+1,T_1}\mathbf{Q}_{m+1,T}^{-1})\mathbf{x}_T \geqslant 0. \tag{12}$$

The efficiency gain of adding one more observation to the pre-break estimation period (by reducing m) depends on whether $\sigma_1^2 \leqslant \sigma_2^2$ or $\sigma_1^2 > \sigma_2^2$. Under the former inequality $\psi \leqslant 0$, and using (11) and (12) in (10), it readily follows that

$$\mathbf{E}_m - \mathbf{E}_{m+1} \leqslant 0, \tag{13}$$

so that the forecast error variance is a decreasing function of the number of pre-break data points used in model estimation. Under $\sigma_1^2 > \sigma_2^2$ the outcome is ambiguous and we could have $\mathbf{E}_m - \mathbf{E}_{m+1} > 0$.

Letting $\mathbf{H}_m = \mathbf{Q}_{m,T_1}\mathbf{Q}_{m,T}^{-1}$, the change in the bias term from starting at observation m rather than observation $m+1$ is given by

$$\mathbf{B}_m - \mathbf{B}_{m+1} = \boldsymbol{\mu}'(\mathbf{H}_m\mathbf{x}_T\mathbf{x}'_T\mathbf{H}'_m - \mathbf{H}_{m+1}\mathbf{x}_T\mathbf{x}'_T\mathbf{H}_{m+1})\boldsymbol{\mu}.$$

Furthermore,

$$\begin{aligned}&\mathbf{H}_m\mathbf{x}_T\mathbf{x}'_T\mathbf{H}'_m - \mathbf{H}_{m+1}\mathbf{x}_T\mathbf{x}'_T\mathbf{H}_{m+1} \\ &\quad = (\mathbf{H}_m - \mathbf{H}_{m+1})\mathbf{x}_T\mathbf{x}'_T(\mathbf{H}_m - \mathbf{H}_{m+1}) + (\mathbf{H}_m - \mathbf{H}_{m+1})\mathbf{x}_T\mathbf{x}'_T\mathbf{H}_{m+1} \\ &\qquad + \mathbf{H}_{m+1}\mathbf{x}_T\mathbf{x}'_T(\mathbf{H}_m - \mathbf{H}_{m+1})'.\end{aligned}$$

By Lemma 1 in Appendix A, $\mathbf{H}_m - \mathbf{H}_{m+1} \geqslant 0$, and the matrices $\mathbf{x}_T\mathbf{x}'_T$, $\mathbf{H}_{m+1}$ are non-negative definite. Therefore,

$$\boldsymbol{\mu}'(\mathbf{H}_m\mathbf{x}_T\mathbf{x}'_T\mathbf{H}'_m - \mathbf{H}_{m+1}\mathbf{x}_T\mathbf{x}'_T\mathbf{H}_{m+1})\boldsymbol{\mu} \geqslant 0,$$

and so, irrespective of the sign of ψ,

$$\mathbf{B}_m - \mathbf{B}_{m+1} \geqslant 0. \tag{14}$$

[3] Note that $\mathbf{Q}_{m,T_1} - \mathbf{Q}_{m+1,T_1} = \mathbf{x}_{m-1}\mathbf{x}'_{m-1} \geqslant 0$, and by assumption $\mathbf{Q}_{m,T}$ is a positive definite matrix.

This means that the squared bias always increases, the earlier the pre-break estimation window is started.[4]

Consider now the overall change in $MSFE(m|\mathbf{X}_T)$ as the starting point of the window is changed from m to $m+1$:

$$\begin{aligned}\Delta_{m,m+1} &= MSFE(m|\mathbf{X}_T) - MSFE(m+1|\mathbf{X}_T)\\ &= \sigma_2^2[(\mathbf{B}_m - \mathbf{B}_{m+1}) + (\mathbf{E}_m - \mathbf{E}_{m+1})].\end{aligned} \tag{15}$$

When $\sigma_1^2 \leqslant \sigma_2^2$ (or $\psi \leqslant 0$), using (13) and (14), it is clear that as $m+1$ is decreased by one unit to m, $\Delta_{m,m+1}$ could rise or fall depending on whether the rise in $\mathbf{B}_m$ is higher or lower than the potential fall in $\mathbf{E}_m$. Similarly, as m is reduced to $m-1$, $\Delta_{m-1,m}$ could either rise or fall. Therefore, $MSFE(m|\mathbf{X}_T)$ need not be monotonic in m and there is a trade-off between an increase in the bias $\mathbf{B}_m$ as m is reduced compared to the potential increase in efficiency which results in a reduction in $\mathbf{E}_m$. In the case where $\sigma_1^2 > \sigma_2^2$ (and $\psi > 0$) the sign of $\mathbf{E}_m - \mathbf{E}_{m+1}$ is ambiguous and it might not be possible to reduce the MSFE by increasing the size of the pre-break sample. These results suggest that it may be optimal to use pre-break data points, particularly if $\sigma_1^2 \leqslant \sigma_2^2$. Appendix B presents a more formal analysis of when it is optimal to use pre-break data to estimate the parameters of the multivariate forecasting model.

For the univariate model ($p = 1$), some analytically tractable cases of special interest emerge. If there is no break in the mean ($\mu = 0$) and $\sigma_2^2 > \sigma_1^2$, it is optimal to use the full data set (and thus an expanding window) to estimate μ although the forecast error variance should be based on σ_2^2 and not on the weighted average of σ_1^2 and σ_2^2. In contrast, it is not optimal to include pre-break observations when either μ^2 is very large or σ_1^2 is much larger than σ_2^2 so that ψ is large. Hence if the break in the mean parameters is high or the pre-break error variance is much higher than the post-break error variance, then only post-break observations should be used in the estimation. However, even if a sizeable break in the mean has occurred, it may still be optimal to include pre-break data provided that the variance of the regression equation is smaller before the break occurred.

Using $v_1 = T_1 - m + 1$ and $v_2 = T - T_1$ to denote the number of pre-break and post-break observations, respectively, with $v = v_1 + v_2$ being the total length of the estimation window, we summarize our findings in the following proposition:

Proposition 1. *The optimal fraction of pre-break observations used to estimate the parameters of the multivariate regression model* (1) *with strictly exogenous regressors* (4) *is higher if*

(i) *the break in the mean parameters* (μ) *is small*;
(ii) *the variance parameter increases at the point of the break* ($\sigma_2^2 > \sigma_1^2$);
(iii) *the post-break window size* ($v_2 = T - T_1$) *is small.*

2.2. Unconditional MSFE results

The above results condition the choice of the optimal window size on the sequence of realizations of $\mathbf{x}_t$. This is clearly of greatest interest since most forecasts are conditioned on

[4] In the absense of assumption (4) this result need not hold as discussed by Pesaran and Timmermann (2005).

the available data. However, it is also of interest to investigate which factors determine the optimal window size on average, i.e. across the possible realizations of $\mathbf{x}_t$. Provided that a process is postulated for $\{\mathbf{x}_t\}$ one can integrate out the effect of $\mathbf{X}_T$ in the expression for the optimal window size and the resulting MSFE. In general, this can be done through Monte Carlo simulation. However, if the joint process generating $\{u_t, \mathbf{x}_{t-1}\}$ is sufficiently simple, analytical results can also be obtained. Considering the case with a single regressor, from (5) and (7) the first two unconditional moments of $e_{T+1}(m)$ are given by

$$\mathrm{E}[e_{T+1}(m)] = (\beta_1 - \beta_2)\mathrm{E}(\theta_m x_T),$$

$$\sigma_2^{-2}\mathrm{E}[e_{T+1}^2(m)] = 1 + \mu^2\mathrm{E}(\theta_m^2 x_T^2) + \mathrm{E}\left(\frac{x_T^2}{\sum_{t=m}^{T} x_{t-1}^2} + \frac{\psi x_T^2 \theta_m}{\sum_{t=m}^{T} x_{t-1}^2}\right), \tag{16}$$

where $\theta_m \equiv \theta_m(T_1, T) = \sum_{t=m}^{T_1} x_{t-1}^2 / \sum_{t=m}^{T} x_{t-1}^2$. The unconditional MSFE can be derived analytically in the special case where u_t and x_t are identically, independently and jointly normally distributed:

$$\begin{pmatrix} u_{t+1} \\ x_t \end{pmatrix} \sim iidN\left[\begin{pmatrix} 0 \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \omega^2 \end{pmatrix}\right]. \tag{17}$$

Using this model, Appendix C provides details and proves some interesting comparative static results summarized in the following proposition.

Proposition 2. *The unconditional MSFE of the forecasting model* (1) *subject to the restrictions in* (17) *with* $\mu_x = \psi = 0$ *is given by*

$$\mathrm{E}[e_{T+1}^2(m)] = \sigma^2 + \omega^2(\beta_1 - \beta_2)^2 \frac{v_1(v_1+2)}{v(v+2)} + \frac{\sigma^2}{v-2}.$$

Hence the optimal pre-break window that minimizes the MSFE is longer,

(i) *the smaller the signal-to-noise ratio* ω^2/σ^2;
(ii) *the smaller the size of the break* $(\beta_1 - \beta_2)^2$;
(iii) *the smaller the post-break window,* v_2.

The unconditional MSFE results are consistent with the conditional results established previously and confirm the intuition that the benefits from using more pre-break data increases when breaks are small, difficult to detect and occur late in the sample.

3. Determination of estimation window

In the context of the multivariate regression model with strictly exogenous regressors, the analytical results in Section 2 demonstrated the determinants of the trade-off involved in selecting the window used in estimating the parameters of a forecasting model whose performance is evaluated on the basis of its out-of-sample MSFE. While the multivariate regression model is in widespread use in forecasting experiments, in many cases the strict exogeneity (4) assumption cannot be maintained. Furthermore, in practical applications, neither the time of any break(s), nor their size is likely to be known, so one needs to consider how to treat the uncertainty surrounding these in order to construct empirically useful techniques.

To this end we consider in this section a broader class of strategies for dealing with breaks in forecasting situations. The first strategy selects a single best estimation window either by adopting cross-validation methods to a pseudo-out-of-sample forecasting experiment, by exclusively using post-break data or by using the (in-sample) trade-off embedded in Eq. (30). The second strategy borrows ideas from the literature on forecast combinations and combines forecasts from models estimated on different observation windows using a variety of robust weighting schemes. This strategy has the advantage that it bypasses the need for direct estimation of breakpoint parameters—a task that is often difficult in practice, particularly when it comes to determining the timing of small breaks, cf. Elliott (2005). Another advantage to this approach is that it is applicable to general dynamic models and for estimation methods other than least squares such as the maximum likelihood or the generalized method of moments.

Common to all methods is that the estimation window should exceed a minimum length, $\underline{\omega}$. This assumption is not necessarily restrictive since $\underline{\omega}$ can always be set equal to a very small number, i.e. the number of regressors plus one. In practice, however, to account for the very large effect of parameter estimation error in cases with few data points per estimated parameter, $\underline{\omega}$ should be reasonably large, say at least two to three times the number of unknown parameters. Furthermore, those forecasting methods that rely on cross-validation reserve the last $\tilde{\omega}$ observations to measure pseudo-out-of-sample forecasting performance used in ranking or weighting the various models.

We next provide specific details on the implementation of each approach. For simplicity we discuss the approaches under the assumption of a single break, but the generalization to multiple breaks is straightforward. For those approaches that pre-test for a break, we shall assume that an estimate of the break-point, $\hat{T}_1$, is available using methods such as those proposed by Bai and Perron (1998) or Altissimo and Corradi (2003).

3.1. Post-break window

Suppose that the forecaster has an estimate of the time of the break, $\hat{T}_1$. A simple estimation strategy is then to only use post-break data $[T_1+1:T]$ to estimate the parameters of the model, denoted by $\hat{\boldsymbol{\beta}}_{\hat{T}_1+1:T}$, where $\hat{\boldsymbol{\beta}}_{\hat{T}_1+1:T} = (\sum_{j=\hat{T}_1+1}^{T} \mathbf{x}_{j-1}\mathbf{x}_{j-1}')^{-1} \times \sum_{j=\hat{T}_1+1}^{T} \mathbf{x}_{j-1}' y_j$, and the resulting forecast is computed as $\hat{y}_{T+1}(\hat{T}_1) = \hat{\boldsymbol{\beta}}_{\hat{T}_1+1:T}' \mathbf{x}_T$. This strategy only involves estimating the time of the break and thus does not require the estimation of pre-break parameters or the post-break variance.

3.2. Trade-off method

The trade-off function is given by (30) in Appendix B, and is specified in terms of $\lambda = v_1/v$, the fraction of the pre-break observations. An estimate of the optimum value of λ, which we denote by $\hat{\lambda}^*$, is given by

$$\hat{\lambda}^* = \arg\min_{\lambda} \left\{ \lambda^2 (\hat{\boldsymbol{\mu}}' \hat{\boldsymbol{\Sigma}}_{v_1} \hat{\boldsymbol{\Sigma}}_v^{-1} \mathbf{x}_T)^2 + \frac{\lambda \hat{\psi}}{v} (\mathbf{x}_T' \hat{\boldsymbol{\Sigma}}_v^{-1} \hat{\boldsymbol{\Sigma}}_{v_1} \hat{\boldsymbol{\Sigma}}_v^{-1} \mathbf{x}_T) + \frac{1}{v} (\mathbf{x}_T' \hat{\boldsymbol{\Sigma}}_v^{-1} \mathbf{x}_T) \right\}, \tag{18}$$

subject to $0 \leqslant \hat{\lambda}^* < 1$, where[5]

$$\hat{\boldsymbol{\Sigma}}_v = \lambda \hat{\boldsymbol{\Sigma}}_{v_1} + (1-\lambda)\hat{\boldsymbol{\Sigma}}_{v_2}, \quad \hat{\psi} = (\hat{\sigma}_1^2 - \hat{\sigma}_2^2)/\hat{\sigma}_2^2, \quad \hat{\boldsymbol{\mu}} = (\hat{\boldsymbol{\beta}}_2 - \hat{\boldsymbol{\beta}}_1)/\hat{\sigma}_2,$$

$$\hat{\boldsymbol{\beta}}_1 = \left(\sum_{t=1}^{\hat{T}_1} \mathbf{x}_{t-1}\mathbf{x}'_{t-1}\right)^{-1} \sum_{t=1}^{\hat{T}_1} \mathbf{x}'_{t-1} y_t, \quad \hat{\boldsymbol{\beta}}_2 = \left(\sum_{t=\hat{T}_1+1}^{T} \mathbf{x}_{t-1}\mathbf{x}'_{t-1}\right)^{-1} \sum_{t=\hat{T}_1+1}^{T} \mathbf{x}'_{t-1} y_t,$$

$$\hat{\boldsymbol{\Sigma}}_{\hat{v}_1} = \hat{T}_1^{-1}\left(\sum_{t=1}^{\hat{T}_1} \mathbf{x}_{t-1}\mathbf{x}'_{t-1}\right), \quad \hat{\boldsymbol{\Sigma}}_{v_2} = (T - \hat{T}_1)^{-1}\left(\sum_{t=\hat{T}_1+1}^{T} \mathbf{x}_{t-1}\mathbf{x}'_{t-1}\right),$$

$\hat{T}_1$ is an estimate of the time of the break, and

$$\hat{\sigma}_1^2 = (\hat{T}_1 - p - 1)^{-1} \sum_{t=1}^{\hat{T}_1} (y_t - \mathbf{x}'_{t-1}\hat{\boldsymbol{\beta}}_2)^2, \quad \hat{\sigma}_2^2 = (T - \hat{T}_1 - p - 1)^{-1} \sum_{t=\hat{T}_1+1}^{T} (y_t - \mathbf{x}'_{t-1}\hat{\boldsymbol{\beta}}_2)^2.$$

We shall refer to estimation windows determined in this way as based on the trade-off method since it attempts to trade-off bias against reduction in parameter estimation error.

If both the pre- and post-break window sizes, v_1 and v_2, were to go to infinity, under broad conditions $\hat{\psi}$ and $\hat{\mu}$ could be consistently estimated. However, in most economic applications—and the ones that concern us here—even if the number of pre-break observations is very large, the number of post-break data points, v_2, is likely to be small and as a result it would not be possible to estimate λ^* consistently so other approaches that deal with the uncertainty that surrounds λ^* need to be considered.

3.3. Cross-validation

The cross-validation approach reserves the last $\tilde{\omega}$ observations of the data for an out-of-sample estimation exercise and chooses the estimation window that generates the smallest MSFE value on this sample. Since we further assume that a minimum of $\underline{\omega}$ observations is needed to estimate the parameters of the forecasting model, this means that $\underline{\omega}+\tilde{\omega}$ data points are required to adopt this method. For each potential starting point of the estimation window, m, the recursive pseudo-out-of-sample MSFE value is computed as

$$MSFE(m|T,\tilde{\omega}) = \tilde{\omega}^{-1} \sum_{\tau=T-\tilde{\omega}}^{T-1} (y_{\tau+1} - \mathbf{x}'_\tau \hat{\boldsymbol{\beta}}_{m:\tau})^2, \tag{19}$$

where $\hat{\boldsymbol{\beta}}_{m:\tau}$ is the OLS estimate based on the observation window $[m,\tau]$. Suppose again that the forecaster has an estimate of the time of the break, $\hat{T}_1$, and define $m^*(T,\hat{T}_1,\underline{\omega},\tilde{\omega})$ as that value of $m \in 1,\ldots,\hat{T}_1+1$, or $m \in 1,\ldots,T-\underline{\omega}-\tilde{\omega}$ (whichever is smallest since on efficiency grounds it would only be meaningful to search for windows that start prior to

[5] To simplify the derivation of $\hat{\lambda}^*$, we abstract from the dependence of $\boldsymbol{\Sigma}_{v_1}$ on v_1, and instead use the pre-break observations to estimate $\boldsymbol{\Sigma}_{v_1}$. Similarly, the structural parameters, $\boldsymbol{\mu}$ and ψ, are also based on pre- and post-break data.

$\hat{T}_1 + 1$) that minimizes the out-of-sample MSFE:

$$m^*(T, \hat{T}_1, \underline{\omega}, \tilde{\omega}) = \arg \min_{m=1,\ldots,\min(\hat{T}_1+1, T-\underline{\omega}-\tilde{\omega})} \left\{ \tilde{\omega}^{-1} \sum_{\tau=T-\tilde{\omega}}^{T-1} (y_{\tau+1} - \mathbf{x}'_\tau \hat{\boldsymbol{\beta}}_{m:\tau})^2 \right\}. \tag{20}$$

The corresponding forecast for period $T+1$ is computed as

$$\hat{y}_{T+1}(T, \hat{T}_1, m^*) = \mathbf{x}'_T \hat{\boldsymbol{\beta}}_{m^*:T}.$$

This approach can readily be implemented without an estimate of T_1, treating the break date as unknown. In this case the optimal break date is determined from

$$m^*(T, \underline{\omega}, \tilde{\omega}) = \arg \min_{m=1,\ldots,T-\underline{\omega}-\tilde{\omega}} \left\{ \tilde{\omega}^{-1} \sum_{\tau=T-\tilde{\omega}}^{T-1} (y_{\tau+1} - \mathbf{x}'_\tau \hat{\boldsymbol{\beta}}_{m:\tau})^2 \right\}, \tag{21}$$

so this method searches for m^* along the points $m = 1, \ldots, T - \underline{\omega} - \tilde{\omega}$ regardless of the break date.

3.4. Weighted average combination

In many empirical applications, as argued by Elliott (2005) and Paye and Timmermann (2004), it can be difficult to obtain a precise estimate of the time and size of a potential break. In particular, when the length of the evaluation sample ($\tilde{\omega}$) is short, the estimate of m^* is likely to be subject to considerable uncertainty. Rather than selecting a single (poorly determined) estimation window, it becomes attractive to combine forecasts based on different estimation windows. One approach that builds on ideas from the forecast combination literature is to let the forecast combination weights be proportional to the inverse of the associated out-of-sample MSFE-values. Suppose again that an estimate of the break date, $\hat{T}_1$, is available and assume that $\hat{T}_1 + 1 < T - \tilde{\omega} - \underline{\omega}$.[6] Then the combined (weighted) forecast is given by

$$\hat{y}_{T+1,W}(T, \hat{T}_1, \tilde{\omega}) = \frac{\sum_{m=1}^{\hat{T}_1+1} (\mathbf{x}'_T \hat{\boldsymbol{\beta}}_{m:T}) MSFE(m|T, \tilde{\omega})}{\sum_{m=1}^{\hat{T}_1+1} MSFE(m|T, \tilde{\omega})}. \tag{22}$$

The only point where knowledge of the time of the break ($\hat{T}_1$) is assumed in (22) is in determining the set from which m^* is selected. Values of m greater than $\hat{T}_1 + 1$ lead to inefficient estimators since they do not make use of all data points after the most recent break and can thus be disregarded.

Once again, this approach can be extended to treat the break date as unknown to avoid the need for having an estimate of T_1 and simply combine the optimal forecasts allowing for the minimal estimation and evaluation windows $\underline{\omega}, \tilde{\omega}$. In this case all values of $m \in 1, \ldots, T - \underline{\omega} - \tilde{\omega}$ are considered and the weighted average forecast is given by

$$\hat{y}_{T+1,W}(T, \underline{\omega}, \tilde{\omega}) = \frac{\sum_{m=1}^{T-\underline{\omega}-\tilde{\omega}} (\mathbf{x}'_T \hat{\boldsymbol{\beta}}_{m:T}) MSFE(m|T, \tilde{\omega})}{\sum_{m=1}^{T-\underline{\omega}-\tilde{\omega}} MSFE(m|T, \tilde{\omega})}. \tag{23}$$

[6] In general, the summation in (22) runs from $m = 1, \ldots, \min(\hat{T}_1 + 1, T - \tilde{\omega} - \underline{\omega})$.

3.5. Simple average combination (pooled forecast)

A particularly simple combination approach is to put equal weights on all forecasts generated subject to $m \leqslant \hat{T}_1 + 1$ (or, more specifically, $m \leqslant \min(\hat{T}_1 + 1, T - \underline{\omega})$):

$$\hat{y}_{T+1}(T, \hat{T}_1, \underline{\omega}) = (\hat{T}_1 + 1)^{-1} \sum_{m=1}^{\hat{T}_1+1} \mathbf{x}'_T \hat{\boldsymbol{\beta}}_{m:T}. \tag{24}$$

Alternatively, when an estimate, $\hat{T}_1$, is not available, subject to utilizing a minimal window length, $\underline{\omega}$:

$$\hat{y}_{T+1}(T, \underline{\omega}) = (T - \underline{\omega})^{-1} \sum_{m=1}^{T-\underline{\omega}} \mathbf{x}'_T \hat{\boldsymbol{\beta}}_{m:T}. \tag{25}$$

These forecasts build on the common finding in the literature on forecast combination that equal-weighted forecasts perform quite well and are difficult to beat, cf. Clemen (1989) and Stock and Watson (2001).

3.6. Multiple breaks

For simplicity, so far only the possibility of a single break is assumed, but in practice a time-series model may be subject to multiple breaks. The presence of multiple breaks complicates the relationship between the (squared) bias and forecast error variance since more breakpoint scenarios become possible. For example, under two breaks, the parameters may be trending upwards or downwards across break segments (making early data less useful for estimation) or alternatively be mean-reverting (making early data useful but more recent data less so). Clearly, the trade-off between using one pre-break data point versus none at all does not hinge on the absence of multiple breaks. However, things get more complicated once more than one pre-break data point is included. Consider, for example, the case where a break happened at time T_1 and another at time $T_1 - 1$. While inclusion of two data points may have been optimal in the absence of the additional break at time $T_1 - 1$, this could be overturned if this break (assuming that it could be detected) generated data from a model sufficiently different from that prevailing after time T_1.

Although the theoretical results get more complicated, the presence of multiple breaks need not complicate application of the above methods which extend in obvious ways. Formulas such as (21), (23) and (25) do not rely on an estimate of the timing (or size) of breaks which only indirectly show up in the patterns observed in the MSFE-values computed as a function of the window length. Furthermore, since the efficiency argument of using all post-break data only applies to the data after the most recent break, in the presence of multiple breaks one can use the cross-validation or trade-off methods as in (18) and (20) but (in the latter case) computing out-of-sample MSFE-values either for $m = 1, \ldots, \hat{T}_b + 1$ (in cases where earlier breaks are either believed to be difficult to detect or of a sufficiently small magnitude) or $m \in \hat{T}_{b-1} + 1, \ldots, \hat{T}_b + 1$, where $\hat{T}_{b-1}$ and $\hat{T}_b$ are the estimated dates of the penultimate and ultimate breaks.

3.7. Discussion

Although we focus on the multivariate regression model, the approaches proposed in this paper can readily be extended to handle different types of estimators required for general dynamic models of the form

$$y_{t+1} = g(\mathbf{x}_t; \boldsymbol{\theta}) + \varepsilon_{t+1}, \tag{26}$$

where $g(.)$ is a general nonlinear function. They can also be readily adapted to use with general loss functions $L(y_{t+1}, \hat{y}_{t+1})$ in addition to the quadratic loss that is commonly used and that we have adopted here.

The combination and cross-validation approaches do not depend on precise estimation of the size of the pre- and post-break parameters and, as we have seen, also do not necessarily require having an estimate of the time of the break(s), although it is an option to use such information when available. This is a disadvantage in the sense that they do not trade-off the effects described in Section 2. However, the trade-offs can be difficult to exploit in practice given the uncertainty surrounding the parameters characterizing any break(s), so in practice this may not be too much of concern.

These methods require choosing the two parameters $\underline{\omega}$ and $\tilde{\omega}$, the length of the minimal estimation window and the length of the evaluation window (although the pooling approach only requires the former). Choosing $\underline{\omega}$ is not much of an issue and is simply a feature that robustifies the combination methods against the influence of extreme forecasts that could result when the number of data points used to estimate a forecasting model is very small. Selection of $\tilde{\omega}$ is driven by the usual considerations faced by researchers attempting to partition a sample into in-sample and out-of-sample periods. If $\tilde{\omega}$ is set too large, then too much smoothing may result and forecasting methods that performed well earlier during the sample may be preferred over models that perform better closer to the end point, T. Conversely, if $\tilde{\omega}$ is set too short, then the ranking of forecasting models will be too noisy and affected too greatly by random variations. In our simulations we set $\underline{\omega}$ at 10% and $\tilde{\omega}$ at 25% of the sample, but other values could of course be used.

To obtain an estimate of the time of the break we shall use the method proposed by Bai and Perron (1998, 2003). This approach provides consistent estimates of the number of breaks under a broad set of conditions and (most importantly) allows for multiple breaks. The method can be implemented in several ways. We use the Schwarz Information Criterion to select the number of breaks, allow for up to three breaks and require a minimum of 10 observations between successive breaks (the simulation results are not sensitive to this assumption). Alternative approaches for detecting multiple structural breaks are also available. For example, Altissimo and Corradi (2003) propose a strongly consistent approach to estimating the number of breaks sequentially which asymptotically lets the probability of over- and underestimating the number of breaks converge to zero. They further propose a small sample correction to detecting the number of breaks which is demonstrated to perform well in Monte Carlo experiments.

In the case with a single break, the time of the break, T_1, can be consistently estimated under the conditions established by Bai (1997). Since it is not optimal for estimation purposes to only use part of the post-break sample $[T_1 + 1, T]$, it follows that expressions such as (20) and (21), (22) and (23), (24) and (25) will be pair-wise asymptotically equivalent. This, of course, requires that the number of pre- and post-break observations is very large, a condition that is unlikely to hold in practice in many situations.

4. Monte Carlo simulations

Our simulation setup assumes that data is generated according to a bivariate VAR(1) model

$$\begin{pmatrix} y_t \\ x_t \end{pmatrix} = \begin{pmatrix} \mu_{yt} \\ \mu_{xt} \end{pmatrix} + \mathbf{A}_t \begin{pmatrix} y_{t-1} \\ x_{t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{yt} \\ \varepsilon_{xt} \end{pmatrix}, \tag{27}$$

where $var(\varepsilon_{yt}) = \sigma^2_{\varepsilon_y t}$, $var(\varepsilon_{xt}) = \sigma^2_{\varepsilon_x t}$ and $cov(\varepsilon_{yt}, \varepsilon_{xt}) = 0$. The approach is general enough to relax these distributional assumptions and higher order dynamics can readily be accommodated by viewing (27) as being in companion form and letting $\mathbf{x}_t$ be a $p \times 1$ vector of predictor variables.

Breaks to the conditional mean are parameterized as follows:

$$\mathbf{A}_t = \begin{cases} \begin{pmatrix} a_{11} & a_{12} \\ 0 & a_{22} \end{pmatrix}, & t \leqslant T_1, \\ \begin{pmatrix} a_{11}+d_{11} & a_{12}+d_{12} \\ 0 & a_{22}+d_{22} \end{pmatrix}, & t > T_1. \end{cases} \tag{28}$$

This setup is similar to that adopted in Clark and McCracken (2001, 2005). In this parameterization, the values of d_{ij} indicate the size of a break. Breaks in $\mathbf{A}$ occur at time $T_1 + 1$ and affect the conditional distribution of y_t given y_{t-1} and x_{t-1}. a_{21} is always normalized at zero so x Granger-causes y, but not the reverse. We start the process from its pre-break stationary distribution and assume that while the mean parameters are affected by breaks, the unconditional or long-run mean is unaffected by breaks, i.e. $\boldsymbol{\mu}_t = (\mathbf{I} - \mathbf{A}_t)^{-1}\boldsymbol{\mu}_0$ for $t \geqslant T_1 + 1$, and $\boldsymbol{\mu}_0$ denotes the pre-break unconditional mean.

In the Monte Carlo simulations with no break to the variance we set $\sigma_{\varepsilon_y} = \sigma_{\varepsilon_x} = 1$ and $cov(\varepsilon_{yt}, \varepsilon_{xt}) = 0$. Breaks to the conditional variance are introduced by letting

$$\sigma_{\varepsilon_y t} = \begin{cases} 1, & t \leqslant T_1, \\ \zeta, & t > T_1, \end{cases}$$

where $\zeta > 0$ is a scaling factor that indicates a higher post-break variance if $\zeta > 1$ and a lower post-break variance if $\zeta < 1$.

To account for the importance of the position of the break in relation to the sample size, we consider a range of combinations of the break date and sample size by letting the break occur at 25%, 50% and 75% of the full sample size. The latter is varied from 100 to 200 observations.

As a natural benchmark we consider forecasts from a model that ignores breaks and uses all observations—this is an optimal estimation window in situations with no breaks. MSFE-values are reported relative to those produced by this benchmark, so a value of unity means the same MSFE performance as the benchmark, values above unity suggest worse forecasting performance and values below unity indicate better forecasting performance than the benchmark.

The parameter values assumed in the Monte Carlo experiments are shown in Table 1. Experiment 1 considers the case without a break. Experiment 2 introduces a relatively small break of -0.2 in the autoregressive parameter, a_{11}, that declines from 0.9 to 0.7 after

Table 1
Simulation setup

(I) *Single break*

Experiment no	d_{11}	d_{12}	d_{22}	σ_y	Comments
1	0	0	0	1	No break
2	−0.2	0	0	1	Small break in AR(1) dynamics
3	−0.4	0	0	1	Large break in AR(1) dynamics
4	0	0.5	0	1	Small break in marginal coefficient
5	0	1	0	1	Large break in marginal coefficient
6	−0.2	1	0	1	Break in dynamics, marginal coefficient
7	0	0	0	4	Increase in post-break variance
8	0	0	0	0.5	Decrease in post-break variance

(II) *Multiple breaks*

Experiment no	a_{11}	d_{11}	d_{12}	d^*_{11}	d^*_{12}	σ_y	Comments
9	0.9	−0.2	0	0	0	1	Mean reversion in AR dynamics
10	0.9	−0.2	0	−0.4	0	1	Decreasing trend in AR dynamics
11	0.3	0.2	0	0	0	1	Mean reversion in AR dynamics
12	0.3	0.2	0	0.4	0	1	Increasing trend in AR dynamics
13	0.9	0	1	0	0	1	Mean reverting break in marginal coefficient
14	0.9	0	1	0	2	1	Trended break in marginal coefficient
15	0.9	0	0	0	0	4	Increase in post-break variance
16	0.9	0	0	0	0	0.5	Decrease in post-break variance

Note: Under a single break it is assumed that $a_{11} = 0.9$. In both sets of experiments, $a_{12} = 1$, $a_{22} = 0.9$, $a_{21} = 0$ and $d_{21} = d_{22} = 0$. See Section 4 for details of the Monte Carlo experiments.

the break, while experiment 3 assumes that this break is somewhat larger at −0.4. Experiments 4 and 5 consider the effect of small and large breaks to the marginal coefficient of x_{t-1} on y_t by letting a_{12} change from 1 to 1.5 or 2 after the break. Experiment 6 studies the effect of a simultaneous break to a_{11} which declines from 0.9 to 0.7 and an increase in a_{12} from 1 to 2. Finally, experiments 7 and 8 change the volatility parameter by letting σ_{ε_y} increase from 1 to 4 (experiment 7) or decrease from 1 to 0.5 (experiment 8).

Table 2 reports the MSFE-values for one-step-ahead forecasts computed at the end of the sample using the methods introduced in Section 3. Results are based on 5000 Monte Carlo simulations. For comparison we also show MSFE results for the infeasible post-break, cross-validation, weighted average and trade-off methods under the assumption that the time of the break, T_1, though not its size, is known.

For experiment 1, that assumes there are no breaks, the forecasting method that generates the lowest out-of-sample MSFE-value is the expanding window. This is unsurprising in view of its efficiency properties in the absence of a break so that the longer the estimation window, the better. However, the various window determination methods also perform quite well—only the cross-validation approach and the pooling method that treat any break points as unknown generate efficiency losses of 2% or 3%.

Turning to the simulations under breaks (experiments 2–8), in experiment 2 where the autoregressive root of y goes from 0.9 to 0.7 at the time of the break, the ranking between the forecasting methods changes significantly. Now the full sample estimator performs worst among all the methods, followed by the weighted average method that conditions on an estimate of the time of the break, T_1. In contrast, the cross-validation and pooled

Table 2
MSFE-values under a single structural break

Panels	Known break date						Estimated break date					Unknown break date		
	Full sample	Post-break	Cross-validation	Weighted average	Pooled	Trade-off	Post-break	Cross-validation	Weighted average	Pooled	Trade-off	Cross-validation	Weighted average	Pooled
A: $T=100$, $T_1=25$, $\tilde{\omega}=25$, $\underline{\omega}=10$														
Exp. # 1	1	—	—	—	—	—	1.017	1.006	1.001	1.002	1.017	1.032	1.014	1.030
2	1	0.717	0.728	0.850	0.831	0.720	0.740	0.733	0.861	0.845	0.744	0.739	0.744	0.736
3	1	0.560	0.569	0.809	0.771	0.562	0.592	0.570	0.819	0.787	0.597	0.579	0.623	0.584
4	1	0.911	0.924	0.941	0.938	0.915	0.957	0.945	0.955	0.954	0.960	0.943	0.919	0.931
5	1	0.702	0.714	0.829	0.803	0.707	0.719	0.718	0.834	0.810	0.723	0.730	0.734	0.725
6	1	0.717	0.730	0.837	0.815	0.718	0.733	0.734	0.843	0.824	0.735	0.744	0.744	0.737
7	1	1.019	1.008	1.007	1.007	1.011	1.036	1.009	1.002	1.003	1.035	1.030	1.020	1.034
8	1	1.004	1.004	0.998	0.998	1.003	1.005	1.003	1.001	1.001	1.005	1.026	1.006	1.022
B: $T=100$, $T_1=50$, $\tilde{\omega}=25$, $\underline{\omega}=10$														
Exp.# 1	1	—	—	—	—	—	1.017	1.006	1.001	1.002	1.017	1.032	1.014	1.030
2	1	0.603	0.618	0.835	0.808	0.610	0.630	0.626	0.847	0.822	0.640	0.623	0.761	0.666
3	1	0.455	0.465	0.816	0.771	0.458	0.486	0.464	0.824	0.784	0.500	0.469	0.722	0.554
4	1	0.803	0.822	0.875	0.867	0.808	0.862	0.859	0.902	0.894	0.868	0.833	0.844	0.825
5	1	0.478	0.486	0.739	0.682	0.484	0.487	0.488	0.745	0.690	0.498	0.493	0.670	0.546
6	1	0.536	0.546	0.772	0.730	0.541	0.546	0.548	0.778	0.737	0.555	0.553	0.702	0.596
7	1	1.060	1.019	1.015	1.014	1.040	1.104	1.017	1.005	1.007	1.096	1.033	1.023	1.039
8	1	1.030	1.015	1.001	1.000	1.034	1.006	1.004	1.001	1.002	1.006	1.026	1.003	1.017
C: $T=100$, $T_1=75$, $\tilde{\omega}=25$, $\underline{\omega}=10$														
Exp. # 1	1	—	—	—	—	—	1.017	1.006	1.001	1.002	1.017	1.032	1.014	1.030
2	1	0.568	0.815	0.884	0.814	0.592	0.649	0.831	0.897	0.842	0.686	0.815	0.884	0.732
3	1	0.410	0.777	0.878	0.794	0.435	0.479	0.779	0.881	0.811	0.531	0.777	0.878	0.677
4	1	0.750	0.828	0.873	0.825	0.767	0.863	0.877	0.910	0.880	0.876	0.828	0.873	0.791
5	1	0.369	0.559	0.764	0.647	0.395	0.400	0.564	0.767	0.661	0.429	0.559	0.764	0.563
6	1	0.453	0.666	0.838	0.741	0.478	0.487	0.671	0.841	0.756	0.514	0.666	0.838	0.651
7	1	1.174	1.030	1.017	1.024	1.136	1.207	1.017	1.005	1.010	1.191	1.030	1.017	1.038
8	1	1.136	1.024	1.010	1.009	1.123	1.008	1.003	1.002	1.002	1.008	1.024	1.010	1.017

D: $T = 200$, $T_1 = 50$, $\tilde{\omega} = 50$, $\underline{\omega} = 20$														
Exp. # 1	1	—	—	—	—	—	1.002	1.000	1.000	1.000	1.001	1.020	1.009	1.016
2	1	0.671	0.676	0.837	0.817	0.672	0.674	0.676	0.842	0.822	0.675	0.684	0.708	0.688
3	1	0.518	0.522	0.812	0.773	0.518	0.530	0.522	0.818	0.782	0.531	0.527	0.596	0.542
4	1	0.916	0.923	0.943	0.941	0.918	0.924	0.925	0.946	0.944	0.925	0.933	0.922	0.926
5	1	0.710	0.717	0.823	0.802	0.711	0.711	0.717	0.826	0.806	0.713	0.724	0.736	0.723
6	1	0.684	0.689	0.814	0.793	0.685	0.686	0.689	0.817	0.796	0.687	0.696	0.712	0.698
7	1	1.009	1.004	1.004	1.004	1.003	1.013	1.001	1.000	1.000	1.012	1.021	1.011	1.018
8	1	1.004	1.004	1.000	1.000	1.004	1.000	1.000	1.000	1.000	1.000	1.018	1.006	1.013
E: $T = 200$, $T_1 = 100$, $\tilde{\omega} = 50$, $\underline{\omega} = 20$														
Exp.# 1	1	—	—	—	—	—	1.002	1.000	1.000	1.000	1.001	1.020	1.009	1.016
2	1	0.545	0.550	0.825	0.794	0.547	0.549	0.550	0.830	0.800	0.552	0.553	0.744	0.623
3	1	0.417	0.422	0.823	0.779	0.418	0.433	0.422	0.827	0.786	0.436	0.424	0.725	0.532
4	1	0.786	0.792	0.870	0.860	0.787	0.795	0.795	0.873	0.864	0.799	0.797	0.836	0.807
5	1	0.457	0.461	0.726	0.667	0.459	0.460	0.461	0.729	0.671	0.463	0.464	0.657	0.521
6	1	0.481	0.484	0.765	0.717	0.486	0.485	0.484	0.768	0.721	0.490	0.488	0.688	0.554
7	1	1.030	1.009	1.008	1.008	1.017	1.053	1.004	1.001	1.002	1.050	1.016	1.012	1.020
8	1	1.019	1.011	1.002	1.002	1.019	1.000	1.000	1.000	1.000	1.000	1.017	1.004	1.011
F: $T = 200$, $T_1 = 150$, $\tilde{\omega} = 50$, $\underline{\omega} = 20$														
Exp. # 1	1	—	—	—	—	—	1.002	1.000	1.000	1.000	1.001	1.020	1.009	1.016
2	1	0.484	0.708	0.878	0.802	0.492	0.499	0.709	0.879	0.810	0.513	0.708	0.878	0.702
3	1	0.357	0.683	0.867	0.783	0.364	0.380	0.683	0.867	0.790	0.398	0.683	0.867	0.654
4	1	0.668	0.750	0.871	0.815	0.679	0.704	0.761	0.878	0.827	0.713	0.750	0.871	0.764
5	1	0.314	0.476	0.775	0.652	0.322	0.319	0.476	0.775	0.656	0.331	0.476	0.775	0.548
6	1	0.393	0.594	0.847	0.749	0.399	0.399	0.595	0.847	0.755	0.413	0.594	0.847	0.639
7	1	1.089	1.017	1.010	1.014	1.064	1.192	1.011	1.005	1.009	1.171	1.017	1.010	1.020
8	1	1.066	1.018	1.008	1.009	1.056	0.999	1.000	1.000	1.000	0.999	1.018	1.008	1.011

Notes: The experiments are defined in Table 1. T is the total sample size, T_1 is the break point, $\tilde{\omega}$ is the size of the evaluation window and $\underline{\omega}$ is size of the minimum estimation window. All MSFEs are reported relative to the associated MSFE based on the full sample.

forecasting methods perform very well—nearly as well as the infeasible post-break or cross-validation methods that condition on the true value of T_1. The trade-off method performs in line with the post-break method.

When the size of the break to the AR(1) parameter of the predicted variable is increased (experiment 3), the cross-validation methods continue to perform best, independently of whether a pre-test for the break is undertaken. When measured against the full sample method, the performance of all methods that incorporate break point information improves as one would expect.

Under a small break to a_{12} (experiment 4), the best feasible methods are the two combination methods that do not pre-test for a break. The pooling method does particularly well across different values of the time of the break, T_1. As the break in a_{12} gets larger (experiment 5), the ranking changes and the approach that pre-tests for a break date followed by cross-validation across the dates $m = 1, \ldots, \hat{T}_1 + 1$ is best followed closely by the post-break and trade-off methods that also pre-test for a break. The post-break and trade-off methods perform well across different break dates, while the performance of the cross-validation approach deteriorates when T_1 gets close to T. These three methods are also best under a simultaneous break to a_{11} and a_{12} (experiment 6).

There is a simple explanation for these findings. When a break is large it becomes easier to detect and so using a pre-test to obtain an estimate of the time of the break leads to improved forecasting performance. Conversely, when a break is relatively small, it is better to treat the timing of the break as unknown and not attempt to estimate pre- and post-break parameters.

These observations can be confirmed from Table 3 which reports the average window sizes selected by a variety of methods. Panel A shows that when the size of the break is small (experiment 4), both the cross-validation and trade-off methods based on an estimate of T_1 use too long window sizes: When $T_1 = 25$, on average the window should be close to 75 observations, but instead it is 87 observations for these methods. This reflects a difficulty in detecting a relatively small break—the average estimate of $\hat{T}_1$ for this experiment is 16 observations, well below the true value of 25. As the break gets larger (experiment 6), the average window length increases to 78–79 observations, only a little higher than the post-break window of 75 observations. This is consistent with break date estimates that average 24 observations for experiments 5 and 6.

Another point that is worth noting is that as the time of the break approaches the end of the sample (i.e. $T_1 = 75$), the performance of the cross-validation and weighted average methods that pre-test for a break and consider estimation windows starting at $m = 1, \ldots, \hat{T}_1 + 1$ deteriorates. In both cases this is related to the imprecision in estimating the time of the break and (in the case of the weighted average forecast) to averaging over too many models that use too long a data sample for estimation purposes. In contrast, the post-break, trade-off and pooling methods (the latter assuming an unknown break date) perform quite well in these experiments. This happens for a subtle reason. When T_1 is close to the end of the sample, T, the true post-break window ($v_2 = T - T_1$) is short and so the results in Section 2 suggest that it is optimal to use relatively many pre-break observations. This effect is important for both the post-break and trade-off approaches that use close to 40 observations when the break is large (experiments 5 and 6) due in part to the fact that the Bai–Perron method tends to underestimate the time of the break (the average value of $\hat{T}_1$ in these experiments is 62, below the true value of 75).

Table 3
Average window sizes under a single structural break

Panels	Known break date		Estimated break date		Unknown break date
	Cross-validation	Trade-off	Cross-validation	Trade-off	Cross-validation
A: $T = 100$, $T_1 = 25$, $\tilde{\omega} = 25$, $\underline{\omega} = 10$					
Exp. # 1	—	—	98.0	98.1	82.0
2	78.4	78.0	79.1	79.6	70.9
3	76.3	77.0	76.3	78.0	68.4
4	83.0	81.0	87.5	86.7	75.9
5	79.2	78.0	79.2	78.8	72.4
6	78.6	77.9	78.5	78.7	70.8
7	91.8	95.0	96.9	96.0	81.8
8	88.8	83.6	97.8	98.0	81.7
B: $T = 100$, $T_1 = 50$, $\tilde{\omega} = 25$, $\underline{\omega} = 10$					
Exp. # 1	—	—	98.0	98.1	82.0
2	54.3	56.6	55.6	60.1	52.1
3	51.3	55.1	51.4	58.3	49.0
4	61.5	61.5	68.7	70.1	59.4
5	54.0	55.9	53.9	57.1	51.9
6	53.2	56.3	53.2	57.7	50.8
7	87.1	84.3	94.6	90.5	84.4
8	81.4	72.1	97.9	98.1	79.9
C: $T = 100$, $T_1 = 75$, $\tilde{\omega} = 25$, $\underline{\omega} = 10$					
Exp.# 1	—	—	98.0	98.1	82.0
2	54.8	40.7	59.6	52.4	54.8
3	54.5	37.6	55.1	45.4	54.5
4	55.9	46.6	70.0	67.3	55.9
5	44.7	37.2	45.5	39.5	44.7
6	47.1	38.0	48.0	41.9	47.1
7	83.8	65.2	92.0	82.2	83.8
8	80.8	61.6	98.1	98.4	80.8
D: $T = 200$, $T_1 = 50$, $\tilde{\omega} = 50$, $\underline{\omega} = 20$					
Exp. # 1	—	—	198.2	198.5	161.8
2	153.9	153.7	154.0	155.0	137.7
3	151.5	152.8	151.5	154.1	135.0
4	161.2	158.5	163.2	161.7	145.8
5	155.5	154.4	155.4	155.1	140.3
6	154.2	153.9	154.1	154.8	137.7
7	184.6	193.1	197.2	196.7	160.6
8	177.5	166.2	197.9	198.0	161.8
E: $T = 200$, $T_1 = 100$, $\tilde{\omega} = 50$, $\underline{\omega} = 20$					
Exp. # 1	—	—	198.2	198.5	161.8
2	103.3	110.2	103.4	112.0	98.5
3	101.2	109.1	101.4	111.7	96.0
4	111.0	115.9	111.6	117.8	106.3
5	104.0	109.0	104.0	110.3	99.2
6	103.2	108.8	103.3	110.3	98.0
7	174.2	173.9	193.0	185.7	167.8
8	160.2	143.7	198.1	198.5	156.4

Table 3 (*continued*)

Panels	Known break date		Estimated break date		Unknown break date
	Cross-validation	Trade-off	Cross-validation	Trade-off	Cross-validation
F: $T = 200$, $T_1 = 150$, $\tilde{\omega} = 50$, $\underline{\omega} = 20$					
Exp. # 1	—	—	198.2	198.5	161.8
2	87.4	70.8	88.6	77.5	87.4
3	86.3	68.1	86.3	75.0	86.3
4	86.7	78.7	95.0	91.0	86.7
5	75.7	66.1	75.8	68.2	75.7
6	77.8	67.9	77.8	70.1	77.8
7	165.6	136.9	183.6	156.0	165.6
8	158.9	129.8	198.3	198.7	158.9

The experiments are defined in Table 1. T is the total sample size; T_1 is the break point, $\tilde{\omega}$ is the size of the evaluation window and $\underline{\omega}$ is the size of the minimum estimation window. The reported window sizes are averages of the windows selected across 5000 replications.

While the post-break method based on an estimated value of T_1 generally performs quite well, it performs poorly when a break only affects the variance of the time series (experiments 7 and 8) and the Bai–Perron method fails to detect this break. Hence it is mainly in situations where a break is believed to affect the conditional mean that this method can be recommended.[7] In general, when the break is confined to the error variances, forecasts based on expanding ('full') estimation window seem to perform best.

The weighted average method combined with a pre-break test appears to be the superior method when the conditional mean is not subject to a break. Conversely, this method performs rather poorly when such a break occurs (in experiments 2–6). The explanation for this finding is again related to the estimates of the break dates. The average estimate of the break date is close to zero when a break either is not present or only affects the variance of the time series. In contrast, the weighted average that treats the time of the break as unknown performs better under breaks to the mean. Since this method averages over more models estimated on post-break data only (when the break occurs after 25 periods), it appears that the efficiency loss associated with using too few post-break observations (resulting from treating T_1 as unknown) is less important than the squared bias effect due to confining the averaging to models estimated on pre-break data. When the time of the break gets larger relative to the cross-validation window ($\tilde{\omega}$), the performance of the weighted average method that treats T_1 as unknown also deteriorates.

The timing of the break ($T_1 = 25, 50, 75$) has some impact on the performance of the various approaches. Among the methods that attempt to estimate the time of a possible break, the relative performance of the post-break and trade-off methods improves as the time of the break increases. Among the methods that treat the time of a break as unknown, the pooled method performs relatively well when the break occurs either early ($T_1 = 25$) or late ($T_1 = 75$) during the sample, while the cross-validation approach performs best when a break occurs in the middle of the sample ($T_1 = 50$). This finding of a great deal of sensitivity in out-of-sample forecasting performance to the timing of the break is mirrored by the analysis of Clark and McCracken (2005).

[7] Alternatively, a different method for detecting breaks to the variance should be adopted.

As the sample size gets larger (panels D–F in Table 2), the performance of the methods that test for a break systematically improves relative to the expanding window approach. Part of the reason for this lies in the improved precision of the estimates of the time of the break which for experiments 5 and 6 vary between 98 and 99 observations (the true value being 100). As a result, the post-break method generally performs rather well as does the trade-off method.

We conclude the following from these simulations. First, which method is best depends on the setup of the simulation experiment—in particular, the timing, size and nature of the break. Second, under a break to the coefficients determining the conditional mean of the predicted variable, an approach that pre-tests for a break and uses a set of inverse-MSE weights to compute a weighted average performs distinctly worse than the other approaches considered here. Third, an approach that pre-tests for a break and only uses post-break data generally performs well as does the cross-validation and pooling approaches that treat the size of the break as unknown. Fourth, the trade-off approach that determines the number of pre-break data points to include by trading off the squared bias against the reduction in forecast error variance generally leads to slightly higher out-of-sample MSFE-values compared to a simple post-break approach that ignores pre-break data. The reasons for this finding are clear. For a start, it is generally difficult to determine precisely the date of the break—something which is crucial when determining the optimal trade-off. Furthermore, and linked to the first point, estimates of the pre-break and post-break parameters—needed to determine the break size—are again plagued by errors and this will infect the trade-off. Even so, the trade-off method does improve on the post-break method in cases where the break only affects the variance. In general, however, an approach that uses estimates of a possible break date and then applies cross-validation to determine the estimation window appears to be a more robust way to proceed.

4.1. Results with multiple breaks

To account for the possibility of multiple breaks we also considered experiments with two breaks occurring at time T_1 and T_2, respectively. We assumed that the breaks occur at one- and two-thirds of the sample, respectively. Under this setup there are three break segments so the AR coefficients can now either decline, increase or mean revert over the sample:

$$\mathbf{A}_t = \begin{cases} \begin{pmatrix} a_{11} & a_{12} \\ 0 & a_{22} \end{pmatrix}, & t \leqslant T_1, \\ \begin{pmatrix} a_{11}+d_{11} & a_{12}+d_{12} \\ 0 & a_{22}+d_{22} \end{pmatrix}, & T_1+1 \leqslant t \leqslant T_2, \\ \begin{pmatrix} a_{11}+d_{11}^{*} & a_{12}+d_{12}^{*} \\ 0 & a_{22}+d_{22}^{*} \end{pmatrix}, & t \geqslant T_2+1. \end{cases} \tag{29}$$

To capture these possibilities, we again consider eight experiments. Experiment 9 assumes that a_{11} mean reverts from 0.9 to 0.7 and back to 0.9. Conversely, experiment 10 considers the effect of a decreasing trend in a_{11} which starts at 0.9, declines first to 0.7 and

Table 4
MSFE-values under multiple breaks

Panels	Known break date						Estimated break date					Unknown break date		
	Full sample	Post-break	Cross-validation	Weighted average	Pooled	Trade-off	Post-break	Cross-validation	Weighted average	Pooled	Trade-off	Cross-validation	Weighted average	Pooled
A: $T = 100$, $T_1 = 33$, $T_2 = 66$, $\tilde{\omega} = 25$, $\underline{\omega} = 10$														
Exp. # 9	1	0.906	0.934	1.015	0.991	0.915	0.949	0.946	1.011	0.986	0.959	0.934	1.015	0.939
10	1	0.512	0.533	0.695	0.639	0.517	0.576	0.559	0.783	0.741	0.580	0.533	0.695	0.575
11	1	0.989	1.001	1.005	1.000	0.998	1.015	1.004	1.004	1.003	1.013	1.001	1.005	0.980
12	1	0.740	0.759	0.813	0.796	0.746	0.795	0.790	0.853	0.836	0.799	0.759	0.813	0.761
13	1	0.702	0.766	1.049	0.982	0.736	0.744	0.768	1.051	0.989	0.779	0.766	1.049	0.837
14	1	0.213	0.225	0.571	0.420	0.220	0.223	0.227	0.580	0.432	0.233	0.225	0.571	0.324
15	1	1.120	1.030	1.020	1.021	1.092	1.176	1.016	1.007	1.011	1.161	1.030	1.020	1.039
16	1	1.079	1.026	1.006	1.005	1.070	1.009	1.005	1.002	1.002	1.009	1.026	1.006	1.016
B: $T = 200$, $T_1 = 66$, $T_2 = 132$, $\tilde{\omega} = 50$, $\underline{\omega} = 20$														
Exp.# 9	1	0.886	0.906	1.039	1.015	0.896	0.895	0.906	1.041	1.015	0.909	0.906	1.039	0.948
10	1	0.464	0.472	0.690	0.629	0.466	0.492	0.490	0.731	0.675	0.494	0.472	0.690	0.550
11	1	0.947	0.965	1.006	1.001	0.950	0.997	0.992	1.005	1.003	0.998	0.965	1.006	0.968
12	1	0.711	0.719	0.811	0.790	0.716	0.733	0.728	0.830	0.809	0.737	0.719	0.811	0.748
13	1	0.657	0.681	1.076	1.006	0.663	0.662	0.682	1.077	1.011	0.671	0.681	1.076	0.837
14	1	0.189	0.194	0.576	0.422	0.193	0.191	0.194	0.577	0.426	0.195	0.194	0.576	0.314
15	1	1.063	1.015	1.012	1.012	1.044	1.135	1.008	1.005	1.008	1.129	1.015	1.012	1.021
16	1	1.045	1.017	1.005	1.006	1.043	1.000	1.000	1.000	1.000	1.000	1.017	1.005	1.010

Notes: The experiments are defined in Table 1. T is the total sample size, T_1 and T_2 give the position of the first and the second break points, $\tilde{\omega}$ is the size of the evaluation window and $\underline{\omega}$ is the size of the minimum estimation window. All MSFEs are reported relative to the associated MSFE based on the full sample.

then to 0.5. Experiment 11 again assumes mean reversion in a_{11}, but now at a much lower level of persistence, namely from 0.3 to 0.5 and back to 0.3. Experiment 12 lets a_{11} follow an increasing trend, starting from 0.3, moving to 0.5 and then to 0.7 at the time of the break points. Experiments 13 and 14 consider breaks in the marginal coefficient a_{12} of x_{t-1} on y_t from zero to one and back to zero (experiment 13) and from zero to one to two (experiment 14). Finally Experiments 15 and 16 are similar to experiments 7 and 8 and assume an increase and a decrease in the post-break variance, respectively, occurring after the second break date, T_2. Again these parameter values are displayed in Table 1.

The results, presented in Tables 4 and 5, show considerable variation in performance across the various approaches to estimation window determination. In experiments 9–12 the best method under a sample size of 100 observations is either the cross-validation method or the pooled method with an unknown break date. In experiments 13 and 14 (break to the marginal coefficient, a_{21}), the post-break method performs best, followed closely by the trade-off and cross-validation methods. Again the weighted average method based on a pre-test for the time of any break(s) performs rather poorly under a break to the conditional mean of the predicted variable. In contrast, when the break only affects the variance (experiments 15 and 16), this method is in fact best, just as we found in the case with a single break.

Few of these conclusions are altered in the experiments with the larger sample size of $T = 200$, although the performance of the post-break method improves slightly as a result

Table 5
Average window sizes under multiple break

Panels	Known break date		Estimated break date		Unknown break date
	Cross-validation	Trade-off	Cross-validation	Trade-off	Cross-validation
A: $T = 100$, $T_1 = 33$, $T_2 = 66$, $\tilde{\omega} = 25$, $\underline{\omega} = 10$					
Exp. # 9	67.4	53.8	73.2	44.9	67.4
10	47.3	56.7	55.4	57.3	47.3
11	68.9	52.6	92.7	55.2	68.9
12	51.3	51.1	59.5	56.3	51.3
13	54.1	55.0	55.5	41.3	54.1
14	39.4	56.3	40.2	41.1	39.4
15	84.9	56.4	92.9	35.5	84.9
16	79.7	60.1	98.0	71.4	79.7
B: $T = 200$, $T_1 = 66$, $T_2 = 132$, $\tilde{\omega} = 50$, $\underline{\omega} = 20$					
Exp. # 9	113.2	100.3	115.5	80.9	113.2
10	79.6	108.5	89.1	92.3	79.6
11	115.8	95.6	170.0	98.3	115.8
12	84.9	91.4	92.4	100.0	84.9
13	83.2	109.2	83.3	76.9	83.2
14	73.2	111.8	73.3	75.9	73.2
15	168.8	115.1	188.1	56.0	168.8
16	156.5	119.8	198.3	152.2	156.5

Notes: The experiments are defined in Table 1. T is the total sample size, T_1 and T_2 give the position of the first and the second break points, $\tilde{\omega}$ is the size of the evaluation window and $\underline{\omega}$ is the size of the minimum estimation window.

of the more precise estimation of the time of the most recent break with a larger sample size. In fact, in experiments 13 and 14 the average estimates of the second break date is 131, close to the true value of 132, while in the smaller sample of 100 observations the corresponding estimates were 62, a bit below the true value of 66.

5. Conclusion

When interest lies in forecasting time-series with regression models that are subject to structural breaks, one might think that the parameters of the forecasting model should be estimated exclusively on data available after the most recent break. However, such an approach ignores two important facts. First, as we show in this paper, in choosing the estimation window there is in general an important trade-off between bias and forecast error variance which means that it is generally advantageous to include (some) pre-break information. Second, it can be difficult to precisely estimate the timing of one or multiple breaks, particularly when these are small and/or occur close to the boundaries of the data sample. It also follows from this point that, in practice, it is generally difficult to optimally exploit the bias–variance trade-off to determine the window size since the parameters characterizing the break points are imprecisely determined. Our results suggest that while little is lost by attempting to exploit the trade-off when the breaks only affect the error variances, the trade-off approach can improve on an approach that estimates the model parameters only on post-break data.

As a result of difficulties in estimating the time and size of the break(s), we proposed a range of alternative methods that either rely on pseudo-out-of-sample cross-validation or use forecast combination methods to combine forecasts from models whose parameters are estimated using different window sizes. These methods can be implemented without any knowledge of breaks to the underlying parameters. For many break processes we found that combination and pooling methods work well, particularly when the break is sufficiently small and hence difficult to detect.

Our findings are closely related to the work by Clark and McCracken (2005) on how the power of tests of predictive ability are affected by structural breaks. Clark and McCracken find that structural breaks can severely affect the out-of-sample predictive performance of econometric models. It follows from this work that researchers should be very careful in how they set up the out-of-sample forecasting experiment, paying close attention to any evidence of breaks. This is consistent with our results that estimation methods that display different degrees of sensitivity to structural breaks tend to produce very different out-of-sample forecasting performance.

Acknowledgements

We would like to thank two anonymous referees and an associate editor for many excellent suggestions on an earlier version of the paper. We also thank James Chu, David Hendry, Adrian Pagan and seminar participants at Erasmus University at Rotterdam, London School of Economics and University of Southern California for comments and discussion.

Appendix A

Lemma 1. *Let* $\mathbf{H}_m = \mathbf{Q}_{m,T_1}\mathbf{Q}_{m,T}^{-1}$, *where* $\mathbf{Q}_{m,T} = \mathbf{X}'_{m,T}\mathbf{X}_{m,T} = \sum_{t=m}^{T} \mathbf{x}_{t-1}\mathbf{x}'_{t-1}$ *and* $1 \leqslant m < T_1 < T$. *Then*

$$\mathbf{H}_m - \mathbf{H}_{m+1} \geqslant \mathbf{0},$$

where $\geqslant$ *denotes a positive semi-definite matrix.*

Proof of Lemma 1. Note that

$$\mathbf{H}_m - \mathbf{H}_{m+1} = (\mathbf{x}_{m-1}\mathbf{x}'_{m-1} + \mathbf{Q}_{m+1,T_1})\mathbf{Q}_{m,T}^{-1} - \mathbf{Q}_{m+1,T_1}\mathbf{Q}_{m+1,T}^{-1}.$$

Post multiplying by $\mathbf{Q}_{m,T} = (\mathbf{x}_{m-1}\mathbf{x}'_{m-1} + \mathbf{Q}_{m+1,T})$ and rearranging, we have that $\mathbf{H}_m > \mathbf{H}_{m+1}$ if

$$\mathbf{x}_{m-1}\mathbf{x}'_{m-1} + \mathbf{Q}_{m+1,T_1} > \mathbf{Q}_{m+1,T_1}\mathbf{Q}_{m+1,T}^{-1}\mathbf{x}_{m-1}\mathbf{x}'_{m-1} + \mathbf{Q}_{m+1,T_1}$$

or equivalently if

$$(\mathbf{Q}_{m+1,T}\mathbf{Q}_{m+1,T}^{-1} - \mathbf{Q}_{m+1,T_1}\mathbf{Q}_{m+1,T}^{-1})\mathbf{x}_{m-1}\mathbf{x}'_{m-1} > 0.$$

This holds provided that

$$(\mathbf{Q}_{m+1,T} - \mathbf{Q}_{m+1,T_1})\mathbf{Q}_{m+1,T}^{-1}\mathbf{x}_{m-1}\mathbf{x}'_{m-1} > 0$$

which is satisfied here since $\mathbf{Q}_{m+1,T} - \mathbf{Q}_{m+1,T_1} = \sum_{t=T_1+1}^{T}\mathbf{x}_{t-1}\mathbf{x}'_{t-1} > 0$ and $\mathbf{x}_{m-1}\mathbf{x}'_{m-1}$, $\mathbf{Q}_{m+1,T}$ are non-negative definite matrices.

Appendix B

This appendix considers determinants of whether it is optimal to use pre-break data to estimate the parameters of the multivariate regression model. Recall that $\mathbf{Q}_{m,T} = \sum_{t=m}^{T_1}\mathbf{x}_{t-1}\mathbf{x}'_{t-1} + \sum_{t=T_1+1}^{T}\mathbf{x}_{t-1}\mathbf{x}'_{t-1}$, and define

$$\boldsymbol{\Sigma}_v = v^{-1}\mathbf{Q}_{m,T} = \left(\frac{v_1}{v}\right)\boldsymbol{\Sigma}_{v_1} + \left(\frac{v_2}{v}\right)\boldsymbol{\Sigma}_{v_2},$$

where

$$\boldsymbol{\Sigma}_{v_1} = v_1^{-1}\sum_{t=m}^{T_1}\mathbf{x}_{t-1}\mathbf{x}'_{t-1}, \quad \boldsymbol{\Sigma}_{v_2} = v_2^{-1}\sum_{t=T_1+1}^{T}\mathbf{x}_{t-1}\mathbf{x}'_{t-1},$$

and, as noted earlier, $v_1 = T_1 - m + 1$, $v_2 = T - T_1$, $v = v_1 + v_2$, where v_1 measures the number of pre-break observations, while v_2 measures the number of post-break observations used in model estimation and v is the total length of the estimation window. Furthermore, define the fraction of pre-break observations as $\lambda = (v_1/v)$ so $1 - \lambda = (v_2/v)$ and $\boldsymbol{\Sigma}_v = \lambda\boldsymbol{\Sigma}_{v_1} + (1-\lambda)\boldsymbol{\Sigma}_{v_2}$. Then the bias term in (9) can be expressed as follows

$$\mathbf{B}_m = \lambda^2(\boldsymbol{\mu}'\boldsymbol{\Sigma}_{v_1}\boldsymbol{\Sigma}_v^{-1}\mathbf{x}_T)^2.$$

Similarly, using (10) we have

$$\mathbf{E}_m = \frac{1}{v}(\mathbf{x}'_T\boldsymbol{\Sigma}_v^{-1}\mathbf{x}_T) + \frac{\lambda\psi}{v}(\mathbf{x}'_T\boldsymbol{\Sigma}_v^{-1}\boldsymbol{\Sigma}_{v_1}\boldsymbol{\Sigma}_v^{-1}\mathbf{x}_T).$$

Hence, for a given value of the post-break window, v_2, the optimal pre-break window size is that value of v_1 that minimizes

$$f(v_1) = \lambda^2(\boldsymbol{\mu}'\boldsymbol{\Sigma}_{v_1}\boldsymbol{\Sigma}_v^{-1}\mathbf{x}_T)^2 + \frac{1}{v}(\mathbf{x}_T'\boldsymbol{\Sigma}_v^{-1}\mathbf{x}_T) + \frac{\lambda\psi}{v}(\mathbf{x}_T'\boldsymbol{\Sigma}_v^{-1}\boldsymbol{\Sigma}_{v_1}\boldsymbol{\Sigma}_v^{-1}\mathbf{x}_T). \tag{30}$$

This minimization problem simplifies considerably in the case of stationary regressors that are not subject to a break so that $\boldsymbol{\Sigma}_{v_1} = \boldsymbol{\Sigma}_{v_2} = \boldsymbol{\Sigma}_v$. Accordingly, we shall consider two cases separately, namely when $\boldsymbol{\Sigma}_{v_1} = \boldsymbol{\Sigma}_{v_2}$ and when $\boldsymbol{\Sigma}_{v_1} \neq \boldsymbol{\Sigma}_{v_2}$. In both cases $\boldsymbol{\Sigma}_{v_1}$ and $\boldsymbol{\Sigma}_{v_2}$ will be taken as given when choosing the optimal value of v_1. Since v_2 is given and the choice of v_1 does not affect v_2, the only assumption being made here is that $\boldsymbol{\Sigma}_{v_1}$ is invariant to the choice of v_1.

In the case where $\boldsymbol{\Sigma}_{v_1} = \boldsymbol{\Sigma}_{v_2}$, (30) simplifies to

$$f(v_1) = \lambda^2(\boldsymbol{\mu}'\mathbf{x}_T)^2 + \left(\frac{1+\lambda\psi}{v}\right)(\mathbf{x}_T'\boldsymbol{\Sigma}_{v_2}^{-1}\mathbf{x}_T).$$

Again the first term is the bias and the second term the efficiency. To obtain the optimal window size, v_1^*, note that

$$\frac{\partial f(v_1)}{\partial v_1} = \frac{2\lambda(1-\lambda)}{v}(\boldsymbol{\mu}'\mathbf{x}_T)^2 - \frac{1-\psi(1-2\lambda)}{v^2}\mathbf{x}_T'\boldsymbol{\Sigma}_{v_2}^{-1}\mathbf{x}_T.$$

Hence, assuming that $\boldsymbol{\mu}'\mathbf{x}_T \neq 0$ (i.e. there has been a break and the change in the regression coefficients, $\boldsymbol{\beta}$, are not fully compensated by elements of the conditioning vector, $\mathbf{x}_T$),

$$\lambda^* = \frac{v_1^*}{v_1^* + v_2} = \frac{(1-\psi)}{2(v_2 R_T - \psi)}, \tag{31}$$

where

$$R_T = \frac{(\boldsymbol{\mu}'\mathbf{x}_T)^2}{\mathbf{x}_T'\boldsymbol{\Sigma}_{v_2}^{-1}\mathbf{x}_T} > 0.$$

Note that since $\mathbf{x}_T'\boldsymbol{\Sigma}_{v_2}^{-1}\mathbf{x}_T > 0, f(v_1)$ can also be written as

$$f(v_1) = \mathbf{x}_T'\boldsymbol{\Sigma}_{v_2}^{-1}\mathbf{x}_T\left[\lambda^2 R_T + \left(\frac{1+\lambda\psi}{v}\right)\right],$$

and in the case where $\boldsymbol{\mu} = \mathbf{0}$ ($R_T = 0$), and $\psi = 0$, we have $f(v_1) = (v_1 + v_2)^{-1}(\mathbf{x}_T'\boldsymbol{\Sigma}_{v_2}^{-1}\mathbf{x}_T)$. Hence, in the absence of breaks, as to be expected, the minimum value of $f(v_1)$ is achieved for a maximum value of v_1.[8]

The solution in (31) is feasible if $0 \leqslant \lambda^* < 1$. To ensure that the inclusion of pre-break data is optimal it is further required that $\lambda^* > 0$, and $\partial^2 f(v_1)/\partial^2 v_1 > 0$ when evaluated at v_1^*. As noted above an optimal feasible solution is guaranteed only in the case where $\psi \leqslant 0$. In this case clearly $\lambda^* > 0$ and since

$$\frac{\partial^2 f(v_1)}{\partial^2 v_1} = 2(\mathbf{x}_T'\boldsymbol{\Sigma}_{v_2}^{-1}\mathbf{x}_T)\left[\frac{(1-\lambda)(1-3\lambda)}{v^2}R_T + \frac{1-\psi(2-3\lambda)}{v^3}\right],$$

[8] Once again recall that v_2 is given and does not vary with v_1.

when evaluated at λ^* we have[9]

$$(\mathbf{x}'_T\boldsymbol{\Sigma}_{v_2}^{-1}\mathbf{x}_T)^{-1}\frac{\partial^2 f(v_1^*)}{\partial^2 v_1} = \frac{(2v_2R_T - 1 - \psi)(2v_2R_T - 3 + \psi)v_2R_T}{2(v_2R_T - \psi)^2 v^2 v_2} + \frac{2}{v^3} - \frac{\psi(4v_2R_T - 3 - \psi)}{(v_2R_T - \psi)v^3}.$$

In the special case where $\sigma_1^2 = \sigma_2^2$ the above results simplify considerably. For λ^* we have

$$\lambda^* = \frac{1}{2v_2R_T} = \frac{\mathbf{x}'_T\boldsymbol{\Sigma}_{v_2}^{-1}\mathbf{x}_T}{2v_2(\boldsymbol{\mu}'\mathbf{x}_T)^2}.$$

This shows that the optimal fraction of the pre-break window depends inversely on the post-break window size and the 'effective' break size, $|\boldsymbol{\mu}'\mathbf{x}_T|$. It also varies directly with $\mathbf{x}'_T\boldsymbol{\Sigma}_{v_2}^{-1}\mathbf{x}_T$ which is likely to be of order p, the number of regressors.

This analysis can also be readily extended to the case where $\boldsymbol{\Sigma}_{v_1} \neq \boldsymbol{\Sigma}_{v_2}$, so long as we continue to assume that $\boldsymbol{\Sigma}_{v_1}$ does not vary with v_1. Under this assumption and using (30) we have

$$\frac{\partial f(v_1)}{\partial v_1} = \frac{2\lambda(1-\lambda)}{v}(\boldsymbol{\mu}'\boldsymbol{\Sigma}_{v_1}\boldsymbol{\Sigma}_v^{-1}\mathbf{x}_T)^2 + 2\lambda^2(\boldsymbol{\mu}'\boldsymbol{\Sigma}_{v_1}\boldsymbol{\Sigma}_v^{-1}\mathbf{x}_T)\left(\boldsymbol{\mu}'\boldsymbol{\Sigma}_{v_1}\frac{\partial\boldsymbol{\Sigma}_v^{-1}}{\partial v_1}\mathbf{x}_T\right) - \frac{1}{v^2}(\mathbf{x}'_T\boldsymbol{\Sigma}_v^{-1}\mathbf{x}_T) + \frac{1}{v}\left(\mathbf{x}'_T\frac{\partial\boldsymbol{\Sigma}_v^{-1}}{\partial v_1}\mathbf{x}_T\right) + \frac{\psi(1-2\lambda)}{v^2}(\mathbf{x}'_T\boldsymbol{\Sigma}_v^{-1}\boldsymbol{\Sigma}_{v_1}\boldsymbol{\Sigma}_v^{-1}\mathbf{x}_T) + \frac{2\lambda\psi}{v}\left(\mathbf{x}'_T\frac{\partial\boldsymbol{\Sigma}_v^{-1}}{\partial v_1}\boldsymbol{\Sigma}_{v_1}\boldsymbol{\Sigma}_v^{-1}\mathbf{x}_T\right),$$

where

$$\frac{\partial\boldsymbol{\Sigma}_v^{-1}}{\partial v_1} = -\frac{1-\lambda}{v}\boldsymbol{\Sigma}_v^{-1}(\boldsymbol{\Sigma}_{v_1} - \boldsymbol{\Sigma}_{v_2})\boldsymbol{\Sigma}_v^{-1}.$$

In this case the optimal choice of v_1 or λ is a complicated function of $\boldsymbol{\Sigma}_{v_1}$ and $\boldsymbol{\Sigma}_{v_2}$ but can be readily obtained using numerical techniques.

Appendix C

In this Appendix, we derive the unconditional MSFE expression used in Proposition 2. First, notice that under assumption (17),

$$\mathrm{E}(x_T^2\theta_m^2) = \mathrm{E}(x_T^2)\mathrm{E}(\theta_m^2).$$

Furthermore,

$$\theta_m \sim \frac{\chi_{v_1}^2(\lambda_1)}{\chi_{v_1}^2(\lambda_1) + \chi_{v_2}^2(\lambda_2)}, \tag{32}$$

where the total window size, $v = T - m + 1$, is split into a pre-break window size $v_1 = T_1 - m + 1$, and a post-break window size $v_2 = T - T_1$. $\chi_{v_1}^2(\lambda_1)$ is a non-central

[9] When $\psi \leqslant 0$, the second-order condition $\partial^2 f(v_1^*)/\partial^2 v_1 > 0$ will be satisfied if $v_2(\boldsymbol{\mu}'\mathbf{x}_T)^2 > \frac{3}{2}(\mathbf{x}'_T\boldsymbol{\Sigma}_{v_2}^{-1}\mathbf{x}_T)$, namely if the break size is sufficiently large.

chi-squared distribution with non-centrality parameter $\lambda_1 = v_1\mu_x^2$ and v_1 degrees of freedom. Likewise, $\chi^2_{v_2}(\lambda_2)$ is a non-central chi-squared distribution (independent of $\chi^2_{v_1}(\lambda_1)$) with non-centrality parameter $\lambda_2 = v_2\mu_x^2$ and v_2 degrees of freedom. Hence θ_m follows a doubly non-central beta distribution with parameters $v_1/2$ and $v_2/2$ and non-centrality parameters λ_1 and λ_2. Using a result due to Patnaik (1949), the first two moments of θ_m are approximately given by

$$\mathrm{E}(\theta_m) \approx \frac{v_1}{v_1+v_2} = \frac{v_1}{v} < 1,$$

$$\mathrm{E}(\theta_m^2) \approx \left(\frac{v_1}{v}\right)\frac{(1+kv_1)}{(1+kv)}, \tag{33}$$

where $k = (1+2\mu_x^2)^2/(2+8\mu_x^2)$. Assuming that $\psi = 0$ ($\sigma_1^2 = \sigma_2^2 = \sigma^2$), so that there is only a break in the conditional mean, the expected value of the last term in (16) can be shown to be given by

$$\mathrm{E}\left(\frac{x_T^2}{\sum_{t=m}^T x_{t-1}^2}\right) = \left(\frac{1}{2}\right)\exp\left(-\frac{1}{2}\lambda\right)(1+\delta^2) \times \sum_{j=0}^{\infty}\frac{(\frac{1}{2}\lambda)^j}{j!}\frac{\Gamma(\frac{1}{2}(v-2)+j)}{\Gamma(\frac{1}{2}v+j)}, \tag{34}$$

where $\lambda = v\mu_x^2$ and $\delta = \mu_x/\omega$. This expression can easily be evaluated numerically. Hence the unconditional MSFE is approximately given by

$$\mathrm{E}[e_{T+1}^2(m)] \approx \sigma^2 + (\beta_1-\beta_2)^2(\omega^2+\mu_x^2)\left(\frac{v_1}{v}\right)\frac{(1+kv_1)}{(1+kv)} + \left(\frac{\sigma^2}{2}\right)\exp\left(-\frac{1}{2}\lambda\right)(1+\delta^2) \times \sum_{j=0}^{\infty}\frac{(\frac{1}{2}\lambda)^j}{j!}\frac{\Gamma(\frac{1}{2}(v-2)+j)}{\Gamma(\frac{1}{2}v+j)}.$$

Tractable exact analytical results can be obtained when it is further assumed that $\mu_x = 0$. In this case the non-centrality parameters are zero,

$$\theta_m \sim Beta\left(\frac{v_1}{2}, \frac{v_2}{2}\right),$$

and the first two moments of θ_m are now given exactly by

$$\mathrm{E}(\theta_m) = \frac{v_1}{v},$$

$$\mathrm{E}(\theta_m^2) = \frac{v_1(v_1+2)}{v(v+2)}.$$

Also,

$$\mathrm{E}\left(\frac{x_T^2}{\sum_{t=m}^T x_{t-1}^2}\right) = \frac{1}{v-2},$$

and, unconditionally,

$$\mathrm{E}(\theta_m x_T) = \mathrm{E}(x_T)\mathrm{E}(\theta_m) = 0.$$

In total, we obtain the following expression for the unconditional MSFE:

$$\mathrm{E}[e_{T+1}^2(m)] = \sigma^2 + \omega^2(\beta_1 - \beta_2)^2 \frac{v_1(v_1+2)}{v(v+2)} + \frac{\sigma^2}{v-2}. \tag{35}$$

References

Altissimo, F., Corradi, V., 2003. Strong rules for detecting the number of breaks in a time series. Journal of Econometrics 117, 207–244.

Andrews, D.W.K., 1993. Tests for parameter instability and structural change with unknown change point. Econometrica 61, 821–856.

Andrews, D.W.K., Ploberger, W., 1996. Optimal changepoint tests for normal linear regression. Journal of Econometrics 70, 9–38.

Bai, J., 1997. Estimation of a change point in multiple regression models. Review of Economics and Statistics 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., 2003. Computation and analysis of multiple structural change models. Journal of Applied Econometrics 18, 1–22.

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

Chow, G., 1960. Tests of equality between sets of coefficients in two linear regressions. Econometrica 28, 591–605.

Chu, C.-S.J., Stinchcombe, M., White, H., 1996. Monitoring structural change. Econometrica 64, 1045–1065.

Clark, T.E., McCracken, M.W., 2001. Tests of equal forecast accuracy and encompassing for nested models. Journal of Econometrics 105, 85–110.

Clark, T.E., McCracken, M.W., 2005. The power of tests of predictive ability in the presence of structural breaks. Journal of Econometrics 124, 1–31.

Clemen, R.T., 1989. Combining forecasts: a review and annotated bibliography. International Journal of Forecasting 5, 559–583.

Diebold, F.X., Lopez, J.A., 1996. Forecast evaluation and combination. In: Maddala, G.S., Rao, C.R. (Eds.), Handbook of Statistics, vol. 14. North-Holland, Elsevier Amsterdam, pp. 241–268.

Elliott, G., 2005. Forecasting in the presence of a break. Mimeo, UCSD.

Granger, C., 1989. Combining forecasts—twenty years later. Journal of Forecasting 8, 167–173.

Hansen, B.E., 1992. Tests for parameter instability in regressions with I(1) processes. Journal of Business and Economic Statistics 10, 321–335.

Inclan, C., Tiao, G.C., 1994. Use of cumulative sums of squares for retrospective detection of changes of variance. Journal of the American Statistical Association 89, 913–923.

Newbold, P., Harvey, D.I., 2002. Forecasting combination and encompassing. In: Clements, M.P., Hendry, D.F. (Eds.), A Companion to Economic Forecasting. Blackwell, Oxford, pp. 268–283.

Patnaik, P.B., 1949. The non-central χ^2- and F-distributions and their applications. Biometrika 36, 202–232.

Paye, B., Timmermann, A., 2004. Instability of return prediction models. Journal of Empirical Finance, forthcomming.

Pesaran, M.H., Timmermann, A., 2005. Small sample properties of forecasts from autoregressive models under structural breaks. Journal of Econometrics 129, 183–217.

Pesaran, M.H., Pettenuzzo, D., Timmermann, A., 2006, Forecasting time series subject to multiple structural breaks. Review of Economic Studies, forthcoming.

Ploberger, W., Kramer, W., Kontrus, K., 1989. A new test for structural stability in the linear regression model. Journal of Econometrics 40, 307–318.

Stock, J.H., Watson, M., 2001. A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series. In: Engle, R.F., White H. (Eds). Festschrift in Honour of Clive Granger.

Timmermann, A., 2006. Forecast combinations. In: Elliott, G., Granger, C.W.J., Timmermann, A. (Eds.), Handbook of Economic Forecasting. North-Holland, Amsterdam, pp. 135–196.