BANK OF ENGLAND

Working Paper No. 490

Adaptive forecasting in the presence of recent and ongoing structural change

Liudas Giraitis, George Kapetanios and Simon Price

March 2014

BANK OF ENGLAND

Working Paper No. 490

Adaptive forecasting in the presence of recent and ongoing structural change

Liudas Giraitis,[1] George Kapetanios[2] and Simon Price[3]

Abstract

We consider time series forecasting in the presence of ongoing structural change where both the time-series dependence and the nature of the structural change are unknown. Methods that downweight older data, such as rolling regressions, forecast averaging over different windows and exponentially weighted moving averages, known to be robust to historical structural change, are found also to be useful in the presence of ongoing structural change in the forecast period. A crucial issue is how to select the degree of downweighting, usually defined by an arbitrary tuning parameter. We make this choice data-dependent by minimising forecast mean square error, and provide a detailed theoretical analysis of our proposal. Monte Carlo results illustrate the methods. We examine their performance on 97 US macro series. Forecasts using data-based tuning of the data discount rate are shown to perform well.

Key words: Recent and ongoing structural change, forecast combination, robust forecasts.

JEL classification: C100, C590.

(1) Queen Mary University of London. Email: l.giraitis@qmul.ac.uk
(2) Queen Mary University of London. Email: g.kapetanios@qmul.ac.uk
(3) Bank of England, City University London, CAMA and CFM. Email: simon.price@bankofengland.co.uk

The views expressed in this paper are those of the authors, and not necessarily those of the Bank of England. Giraitis's research was, in part, supported by ESRC grant RES062230790. This paper was finalised on 25 February 2014.

The Bank of England's working paper series is externally refereed.

Information on the Bank's working paper series can be found at
www.bankofengland.co.uk/research/Pages/workingpapers/default.aspx

Publications Group, Bank of England, Threadneedle Street, London, EC2R 8AH
Telephone +44 (0)20 7601 4030 Fax +44 (0)20 7601 3298 email publications@bankofengland.co.uk

ISSN 1749-9135 (on-line)

Summary

Forecasting is an important activity for central banks, not least because policy takes effect with a lag. Inevitably, policy is forward looking. Thus in many central banks, including the Bank of England, the published forecast is a key tool in communicating judgements about monetary policy and the economy. The Bank's forecast, published in the *Inflation Report*, represents the judgements of the Monetary Policy Committee and is not mechanically produced by a single model. However, many forecasting models - a "suite" of models - help the Committee determine its judgement, including simple largely atheoretical models of the type considered in this paper.

One common cause of forecast failure is that structural changes or "breaks" keep on occurring in the underlying relationships in the economy, and this paper addresses that problem, building on previous work undertaken in the Bank. The problem, almost by definition, is that we do not know what form the structural break took. If we did, we could model it: but then it would not be a structural break, but a known data-generation process. What we need are methods that are useful where there is the possibility of a wide range of types of structural change. The earlier work showed that a robust way of forecasting in such an environment is to discount past data so that more recent data is given more weight. This helps avoid forecast errors, as if there have been structural breaks in the past, the data pertaining to that period is given less weight compared to recent data where there may have been no or fewer breaks. This can be done in many ways. These include "rolling windows" where all data before a cut-off date is excluded, exponentially declining weights smoothly lowering the weight for distant data (often implemented as an exponentially weighted moving average), and other methods. But this raises the practical question of exactly how rapidly to downweight. The innovation of the paper is to choose this by using in-sample forecast performance.

The paper shows that in a wide variety of situations the method will have good statistical properties. What is more, it can handle any degree of persistence. Speaking somewhat loosely, "persistence" is the tendency for a series to be affected by its past behaviour. For example, a series that is simply a constant with some random white noise has no persistence. (In this case, the best forecast is to use all the data to calculate the mean as precisely as possible.) By contrast, in the classic random walk a series is equal to what it was last period plus a random white noise error, and so there is a high degree of persistence. (In that case, the best forecast ignores all except the last observation.) These examples show that the optimal rate of discounting past data is likely to depend upon persistence. We are also able to demonstrate that the method is very flexible. There are ways of including dynamics, similar to the widely used autoregressive (AR) method, known to produce good forecasts, where the series is solely related to a few of its own lags. We can also allow the weights to vary very flexibly using a non-parametric method which does not tie down the model to a specific form, and allow for other explanatory variables. The theory is for large samples, but we show using simulation ("Monte Carlo") methods that the methods work for short samples as well.

The proof of the forecast pudding is in the testing, so we apply the methods to a large number of economic variables from the United States using a sample from 1960 to 2008, comparing root mean square forecast errors (RMSFE), which is a standard criteria that penalises large forecast errors. Not all the series exhibit breaks, but in the typical (median) case the methods do better than an AR benchmark. The methods that work best are ones that allow for some dynamics. For some variables, such as financial spreads and some inflation series, they do spectacularly better. Moreover, in many cases the methods are significantly better (in the

statistical sense) than the benchmark, meaning that they do better much more often than would be expected by chance.

We conclude that the proposed technique of downweighting past data in a way determined by past forecast performance is likely to be a useful item in the forecaster's toolkit.

1 Introduction

It is widely accepted that structural change is a crucial issue in econometrics and forecasting. Clements and Hendry suggest forcefully (in e.g. 1998a,b) that such change is the main source of forecast error; Hendry (2000) argues that the dominant cause of forecast failures is the presence of deterministic shifts. Convincing evidence of structural change was offered by Stock and Watson (1996) who looked at many forecasting models of a large number of US time series, and found parameter instability in a substantial proportion. This issue remains relevant: in a survey of the literature on forecasting in the presence of instabilities for the *Handbook of Economic Forecasting*, Rossi (2012) writes 'the widespread presence of forecast breakdowns suggests the need for improving ways to select good forecasting models in-sample.' Our work on robust and data-driven forecasting is a contribution to precisely this end. As model parameters may change continuously, drift smoothly over time or change at discrete points in an unknown manner, and both within the sample and over the forecast period, we consider a general setting where the model structure and presence and type of structural change are all unknown.

There is a large literature on the identification of breaks, and forecasting methods robust to them (Rossi (2012)). However, the deeply practical need to forecast after a recent structural change, or during a period of such change, has received very little attention. As most forecast approaches are only effective in specific cases, the problem is compounded by the unknown and therefore unspecified nature of any structural change.

Detection of structural change has a long history, mainly in the context of structural breaks (although see Kapetanios (2007) for the case of smooth structural change). Seminal papers include Chow (1960), Andrews (1993) and Bai and Perron (1998). But the question of amendment of forecasting strategies then arises. While this has been tackled by many authors, a major contribution was made by Pesaran and Timmermann (2007). They concluded that, in the presence of breaks, forecast pooling using a variety of estimation windows provides a reasonably good and robust forecasting performance.

Nevertheless, most work on forecasting assumes that change has occurred when sufficient time has elapsed for post-break estimation.[1] In practice, the issue of change occurring in real time is a major consideration, which was partly addressed in Eklund, Kapetanios, and Price (2010). They considered a variety of forecasting strategies which can be divided into two distinct groups. In one case the forecaster monitors for change and adjusts methods once change has been detected. In the other the forecaster does not attempt to identify breaks, since that involves a substantial time lag. Instead break-robust forecasting strategies are used that essentially downweight data from older periods deemed to be irrelevant for the current conjuncture.

While moving in an interesting direction, Eklund, Kapetanios, and Price (2010) do not elaborate two issues: how much to downweight past data, and whether to do so monotonically. Clearly, any arbitrary discount factor is unlikely to be optimal. And neither may monotonicity: for example, if regimes (e.g., monetary policy) come and go then older data, from a period where the current regime previously held, might be more relevant than more recent data from other regimes.

In this paper, we suggest forecasting approaches that address these issues. Our main contribution is to introduce and analyse a cross-validation based method which selects a tuning parameter defining the downweighting rate of the older data. We show that the implied discount rate minimises the mean square error (MSE) of the forecast in the weighting schemes considered. Further,

[1] Exceptions include the interesting work of Clements and Hendry (2006) and Castle, Fawcett, and Hendry (2011).

we consider a nonparametric method for determining a flexible weighting scheme. The latter does not assume any particular shape for the weight function, nor monotonicity. We explore the properties of the new forecasting methods for a variety of models in terms of theory, with a Monte Carlo exercise and empirically. It turns out that the method is valid under a wide range of forms of structural breaks and persistence, and can be generalised in a number of practically important dimensions, most notably allowing varying dynamic structures.

A byproduct of our results is a new way to accommodate trends of a generic nature in forecasting. Unlike many forecasting approaches that require the removal of stochastic or other trends before forecasting, our methods can be directly applied to the level of the forecast series.

The rest of the paper is organised as follows. Section 2 presents our approach for forecasting in the presence of recent structural breaks. We provide its theoretical justification and asymptotic MSE, and describe some robust forecasting strategies. Section 3 includes an extensive Monte Carlo study in which these strategies are evaluated. In Section 4 the methods are used to forecast a large number of US macroeconomic time series, where we find results broadly consistent with the Monte Carlo study. Section 5 concludes. Proofs are reported in an Appendix.

2 Adaptive forecasting: econometric framework

2.1 Forecasting strategies

In this section we work with a simple location forecasting framework that is as general as possible while consistent with clear theoretical results. It may be summarised as

$$y_t = \beta_t + u_t, \qquad t = 1, \cdots, T, \tag{2.1}$$

where β_t is an unobserved persistent process, and u_t is a stationary dependent noise that is independent of β_t. Unlike most previous work we wish to place as little structure as possible on the process β_t. We do not specify whether β_t is stochastic or deterministic, or whether it is discontinuous or smooth. The noise process u_t is a stationary linear process with mean zero and finite variance σ_u^2. The persistent component $\beta_t \equiv \beta_{T,t}$ is allowed to be a triangular array, and can be a stochastic (unit root) or deterministic (bounded) trend. This set-up provides sufficient flexibility to our theoretical analysis of forecasting y_t, allowing for $\beta_{T,t}$ such as those used in locally stationary models (e.g. Dahlhaus (1996)), or in persistent stochastic unit root trend models. For simplicity of notation, we write $y_{T,t}$ as y_t and $\beta_{T,t}$ as β_t. It should be stressed that in robust forecasting, which is our focus, the structure of β_t is neither known nor estimated. Concerning our simple location conditional mean modelling, we note that our analysis can allow both the use of a generic model of the conditional mean of the process and robust forecasting around that model. We discuss details related to this extension in Section 2.8.

Eklund, Kapetanios, and Price (2010) find that simple forecasting of y_t, based on weighting schemes that discount past data, works well in practice. Examples include exponential weighting and forecast combinations based on different estimation windows. By varying a tuning parameter, such methods impose different shapes on the weight functions that downweight past data. Their weakness is that it is not clear how to select the tuning parameters. So data-dependent tuning methods for choosing these parameters are of great interest.

One way to calibrate parameters is by optimising on in-sample forecasting performance. This idea is not new. For example, Kapetanios, Labhard, and Price (2006) suggest forecasts where

different models are averaged with weights that depend on the forecasting performance of each model in the recent past. In what follows we formalise the above ideas, presenting a data-driven weighting strategy and developing its theoretical analysis.

We consider a linear forecast of y_t, based on (local) averaging of past values $y_{t-1}, \cdots, y_1$:

$$\hat{y}_{t|t-1,\,H} = \sum_{j=1}^{t-1} w_{tj,H} y_{t-j} = w_{t1,H} y_{t-1} + \cdots + w_{t,t-1,H} y_1, \tag{2.2}$$

with weights $w_{tj,H} \geq 0$ such that $w_{t1,H} + \cdots + w_{t,t-1,H} = 1$, parameterised by a single tuning parameter H. The latter defines the rate of downweighting the past observations (e.g., the width of the rolling window). The structure of weights $w_{tj,H}$ is described in Assumption 1. We assume that H takes values in the interval $I_T = [\alpha,\, H_{\max}]$, where $\alpha > 0$.

Assumption 1 *The function $K(x) \geq 0$, $x \geq 0$ is continuous and differentiable on its support, such that $\int_0^\infty K(u)du = 1$, $K(0) > 0$, and for some $C > 0$, $c > 0$*

$$K(x) \leq C \exp(-c|x|), \quad |\dot{K}(x)| \leq C/(1+x^2), \qquad x > 0, \tag{2.3}$$

where $\dot{K}$ is the first derivative of K. For $t \geq 1$, $H \in I_T$, set $k_{j,H} = K(j/H)$ and define

$$w_{tj,H} = \frac{k_{j,H}}{\sum_{s=1}^{t-1} k_{s,H}}, \qquad j = 1, \cdots, t-1. \tag{2.4}$$

Example 1 The main classes of commonly employed weights satisfy this assumption.
(i) *Rolling window* weights, with $K(u) = I(0 \leq u \leq 1)$.
(ii) *Exponential weighted moving average* (EWMA) weights, with $K(u) = e^{-u}$, $u \in [0, \infty)$. Then, with $\rho = \exp(-1/H)$, $k_{j,H} = \rho^j$ and $w_{tj,H} = \rho^j / \sum_{k=1}^{t-1} \rho^k$, $1 \leq j \leq t-1$.
(iii) *Triangular window* weights, with $K(u) = 2(1-u)I(0 \leq u \leq 1)$.

While the rolling window simply averages the H previous observations, the EWMA forecast uses all observations $y_1, \cdots, y_{t-1}$, increasingly downweighting the more distant past. In practice, forecasting of a unit root or trending process y_t is often conducted by averaging over the last few observations. When persistence in y_t falls, wider windows may be expected to yield smaller forecast MSE. It is also plausible that for a stationary process $\{y_t\}$ when dependence is sufficiently strong a forecast discounting past data will outperform the sample mean forecast $(y_t + \cdots + y_1)/t$. These observations, supported by the theory below, indicate that the 'optimal' selection of H depends on the unknown type of persistence in y_t. Thus, contrary to the usual practice of using a preselected value of H, a data based selection method for H is indicated.

2.2 Selection of the tuning parameter H

Given a sample $y_1, \cdots, y_T$, computation of the forecast $y_{T+1|T,\,H}$ requires selection of the parameter H. We use a cross-validation method, obtaining H by numerically minimising the mean squared forecast error of the in-sample forecasts, defined by the following objective function:

$$Q_{T,H} := \frac{1}{T_n} \sum_{t=T_0}^{T} (y_t - \hat{y}_{t|t-1,\,H})^2, \qquad \hat{H} := \operatorname{argmin}_{H \in I_T} Q_{T,H} \tag{2.5}$$

with starting point $T_0 = o(T)$, $T_n := T - T_0 + 1$. Define $H_{max} = T_0 T^{-\delta}$, $0 < \delta < 1$. We assume that T_0 and H_{max} are selected such that $T^{2/3} < H_{\max} < T_0 = o(T)$.

We will show that the forecast $\hat{y}_{T+1|T,H}$ of y_{T+1}, obtained with data-tuned weights (or $\hat{H}$), minimises the Mean Squared Error (MSE), $\omega_{T,H} := E(y_{T+1} - \hat{y}_{T+1|T,H})^2$ in H, hence making the forecast procedure (2.2) operational and optimal in the following sense. Let $H_{opt} = \text{argmin}_{H \in I_T} \omega_{T,H}$ be the optimal value of fixed parameter H minimising MSE $\omega_{T,H}$. Then

$$\omega_{T,\hat{H}} = \omega_{T,H_{opt}} + o(1), \qquad Q_{T,\hat{H}} = \omega_{T,\hat{H}} + o_p(1), \tag{2.6}$$

where the quantity $Q_{T,\hat{H}}$ is an estimate of the forecast error $\omega_{T,\hat{H}}$.

Below we verify that the minimisation procedure (2.5) provides optimal selection of H for basic forms of persistence in $y_t = \beta_t + u_t$. As an illustration, let $\hat{\sigma}^2_{T,u} := T_n^{-1} \sum_{j=T_0}^{T} u_j^2$ be the sample variance of u_t. We will show that, as $T \to \infty$,

$$\begin{aligned} Q_{T,H} &= \hat{\sigma}^2_{T,u} + E[Q_{T,H} - \sigma_u^2](1 + o_P(1)), \\ &= \hat{\sigma}^2_{T,u} + (\frac{\lambda_\beta H^m}{T^p} + \frac{\lambda_u}{H})(1 + o_P(1)), \quad H \to \infty, \end{aligned} \tag{2.7}$$

with some constants $\lambda_\beta \geq 0$, λ_u, and integers $m, p \geq 0$. The term $\lambda_\beta H^m / T^p$ is contributed by β_t while λ_u / H by u_t. These relations determine whether $\hat{H}$ takes finite values or is increasing with T.

For example, if the noise u_t in $y_t = \beta_t + u_t$ is sufficiently strongly dependent, then using exponential weights yields $\lambda_u < 0$. Then, no matter what β_t is, $Q_{T,H}$ reaches its minimum on a bounded interval, and the minimiser $\hat{H}$ of (2.7) remains bounded. Similarly, if y_t includes a linear or unit root trend β_t, then $H^m / T^p \geq H$, and the minimiser $\hat{H}$ of (2.7) again remains bounded. Under mildly persistent β_t, the minimiser $\hat{H}$ can also increase as a power of T. For example, if β_t is a bounded unit root trend, and u_t is an i.i.d. noise, then $Q_{T,H} = \hat{\sigma}^2_{T,u} + (\lambda_\beta H T^{-1} + \lambda_u H^{-1})(1 + o_P(1))$, $\lambda_u > 0$, which leads to $\hat{H} \sim c T^{1/2}$ (see Sections 2.5-2.6). Similar properties hold for the break in the mean model (see Section 2.6).

Notation. Beside $w_{tj,h}$, we will use the weights

$$w_{j,H} = k_{j,H} / \sum_{s=1}^{\infty} k_{s,H}, \quad j \geq 1. \tag{2.8}$$

Below, $a \wedge b = \min(a, b)$, $a \vee b = \max(a, b)$ and $I(A)$ is the indicator function; $a_T \sim b_T$ denotes that $a_T / b_T \to 1$, as T increases. We write $o_{p,H}(1)$ or $o_H(1)$ to indicate, that $\sup_{H \in I_T} |o_{p,H}(1)| \to_p 0$ or $\sup_{H \in I_T} |o_H(1)| \to 0$, as $T \to \infty$.

We will use the fact

$$\hat{\sigma}^2_{T,u} = \sigma_u^2 + O_p(T^{-1/2}), \tag{2.9}$$

which holds under Assumption 2 (see Proposition 4.5.2 in Giraitis, Koul, and Surgailis (2012)).

2.3 Properties of a forecast based on $\hat{H}$

Now we turn to the theoretical justification of the optimal properties of the selection procedure of H for $y_t = \beta_t + u_t$, where β_t is a persistent process (deterministic or stochastic trend) of

unknown type, and u_t is a stationary noise term. Our objective is to show that the forecast $y_{T+1|T,\hat{H}}$ of y_{T+1} with optimal tuning parameter $\hat{H}$ minimises the forecast MSE in the following sense: $\omega_{T,\hat{H}} = \omega_{T,H_{opt}} + o_P(1)$. Moreover, the property $Q_{T,\hat{H}} = \omega_{T,\hat{H}} + o_p(1)$ allows estimation of the forecast error.

The following assumption specifies the required properties of the stationary noise process u_t.

Assumption 2 *u_t is a stationary linear process*

$$u_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j}, \;\; t \in \mathbb{Z}, \qquad \varepsilon_j \sim \mathrm{IID}(0, \sigma_\varepsilon^2), \qquad E\varepsilon_1^4 < \infty, \tag{2.10}$$

such that $\sum_{k\in\mathbb{Z}} |\gamma_u(k)| < \infty$, $\sum_{k\geq n} |\gamma_u(k)| = o(\log^{-2} n)$ and $s_u^2 := \sum_{k\in\mathbb{Z}} \gamma_u(k) > 0$, where $\gamma_u(k) = \mathrm{Cov}(u_k, u_0)$.

Under Assumption 2, u_t has short memory, while its long-run variance s_u^2 is positive and finite. We will write $u_t \sim I(0)$ to denote that a stationary process u_t satisfies Assumption 2. Below $\beta_t \sim I(1)$ denotes a unit root process such that $\beta_t - \beta_{t-1}$ is an $I(0)$ process.

We shall consider the following types of persistent component β_t.

b1.	Constant	$\beta_t = \mu$.
b2.	Unit root	$\beta_t \sim I(1)$.
b3.	Bounded unit root	$T^{1/2}\beta_t \sim I(1)$.
b4.	Deterministic trend	$\beta_t = tg(t/T)$.
b5.	Bounded deterministic trend	$\beta_t = g(t/T)$.
b6.	Break in the mean	$\beta_t = \begin{cases} \mu_1, & t = 1, \cdots, t_0, \\ \mu_2, & t = t_0+1, \cdots, T. \end{cases}$

We suppose that, in (b4) and (b5), $g(x), x \in (0,1)$ is continuous and has a bounded second derivative, and in (b6), $\mu_1 \neq \mu_2$ and $\tau := T - t_0 = o(T)$.

We are now ready to analyse the properties of $Q_{T,H}$, $\hat{H}$ and the forecast error $\omega_{T,\hat{H}}$.

2.4 The case of a stationary process y_t

First we discuss the properties of the forecast in the case (b1) when $y_t = \mu + u_t$, $t \geq 1$ is a stationary process. We shall use the following notation: $\kappa_2 = \int_0^\infty K^2(x)dx$, $\kappa_0 = K(0)$, and

$$q_{u,H} := E\big(u_0 - \sum_{j=1}^{\infty} w_{j,H} u_{-j}\big)^2 - \sigma_u^2, \qquad \lambda_u := s_u^2\{\kappa_2 - \kappa_0\} + \sigma_u^2\kappa_0. \tag{2.11}$$

Theorem 1 *Suppose that $y_t = \mu + u_t, t \geq 1$, where u_t is a stationary $I(0)$ process.*
Then, as $T \to \infty$, for $H \in I_T$,

$$Q_{T,H} \;\; = \;\; \hat{\sigma}_{T,u}^2 + q_{u,H} + o_{p,H}(H^{-1}), \qquad \omega_{T,H} = \sigma_u^2 + q_{u,H} + o_H(H^{-1}), \tag{2.12}$$

where $q_{u,H} = \lambda_u H^{-1} + o(H^{-1})$, as $H \to \infty$.

Theorem 1 shows that $Q_{T,H}$ is a consistent estimate of $\omega_{T,H}$, and implies that the forecast $y_{T+1|T,\hat{H}}$ computed with the data-tuned $\hat{H}$ has the same MSE as $y_{T+1|T,H_{opt}}$. The latter can be estimated by $Q_{T,\hat{H}}$ as stated in the following corollary.

Corollary 1 *If $q_{u,H}$ reaches its minimum at some finite H_0, then*

$$\omega_{T,\hat{H}} = \omega_{T,H_{opt}} + o_p(1), \qquad Q_{T,\hat{H}} = \omega_{T,\hat{H}} + o(1) = \sigma_u^2 + q_{u,H_0} + o_p(1). \tag{2.13}$$

If $q_{u,H}$ reaches its minimum at infinity, then

$$\omega_{T,\hat{H}} = \omega_{T,H_{opt}} + o(T^{-1/2}), \qquad Q_{T,\hat{H}} = \omega_{T,\hat{H}} + O_p(T^{-1/2}) = \sigma_u^2 + O_p(T^{-1/2}). \tag{2.14}$$

Remark 1 Result (2.13) implies that the forecast with tuning parameter $\hat{H}$ has the same precision as the forecast based on H_{opt} that minimises the forecast error $\omega_{T,H}$. The sign of λ_u carries information about the location of the minimiser $\hat{H}$ of $Q_{T,H}$: for $\lambda_u < 0$, $Q_{T,H}$ reaches its minimum at some finite value H_0. In such a case, the error $\omega_{T,\hat{H}} = \sigma_u^2 + q_{u,H_0}$ of the optimal forecast is smaller than that of the sample mean, σ_u^2. The sign of λ_u is determined by two factors: the kernel K and the strength of dependence in u_t. For the rolling window kernel $K(u) = I(0 \leq u \leq 1)$, $\kappa_2 = \kappa_0 = 1$, and thus $\lambda_u = \sigma_u^2$ is always positive. However, for the exponential kernel $K(u) = e^{-u}$, $u \geq 0$, $\lambda_u = \sigma_u^2 - s_u^2/2$ which becomes negative when the long-run variance of u_t is sufficiently large: $s_u^2 > 2\sigma_u^2$, e.g., for an AR(1) model u_t with autoregressive parameter greater than 1/3. The fact that λ_u is smaller for exponential weights than for rolling windows suggests that EWMA weighting leads to a smaller forecast error and may outperform the latter.

If $q_{u,H}$ is a positive function, then $\hat{H}$ will take the largest possible value in I_T, and the forecast error $\omega_{T,\hat{H}} \to \sigma_u^2$ will be the same as for the sample mean. These observations are confirmed by simulation studies. Monte Carlo simulations in Table 2 show that for an AR(1) model u_t with parameter 0.7 the rolling window forecast does not outperform the sample mean, while the forecast based on EWMA weights reduces the relative MSE by 33%.

2.5 The case of a stochastic trend

In this section we analyse the properties of the forecast when $y_t = \beta_t + u_t$, $1 \leq t \leq T$ contains a stochastic trend β_t observed under a stationary noise u_t. We focus on two cases:

(b2), where β_t is a unit root $I(1)$ process, setting

$$q_{\beta,H}^{(2)} := E\big(\beta_0 - \sum_{j=1}^{\infty} w_{j,H}\beta_{-j}\big)^2, \qquad \lambda_\beta^{(2)} := s_{\nabla\beta}^2 \int_0^\infty \big(\int_x^\infty K(z)dz\big)^2 dx. \tag{2.15}$$

(b3), where $\beta_t = T^{-1/2}\tilde{\beta}_t$, $t = 1, \cdots, T$ and $\tilde{\beta}_t$ is a unit root $I(1)$ type process, setting

$$q_{\beta,H}^{(3)} := q_{\tilde{\beta},H}^{(2)}, \qquad \lambda_\beta^{(3)} := \lambda_{\tilde{\beta}}^{(2)}. \tag{2.16}$$

Here, $\nabla\beta_t = \beta_t - \beta_{t-1}$. In the following theorem, $q_{u,H}$ and λ_u are the same as in Theorem 1.

Theorem 2 *Let $y_t = \beta_t + u_t$, $t = 1, \cdots, T$ where u_t is a stationary $I(0)$ process.*

(i) If β_t is a unit root trend (b2), then, as $T \to \infty$, for $H \in I_T$,

$$Q_{T,H} \;=\; \hat{\sigma}_{T,u}^2 + q_{\beta,H}^{(2)} + q_{u,H} + o_{p,H}(H), \qquad \omega_{T,H} = \sigma_u^2 + q_{\beta,H}^{(2)} + q_{u,H} + o_H(H), \tag{2.17}$$

where $q_{\beta,H}^{(2)} + q_{u,H} = \lambda_\beta^{(2)} H + o(H)$, as $H \to \infty$.

(ii) If β_t is a bounded unit root trend (b3), then, as $T \to \infty$, for $H \in I_T$,

$$\begin{aligned} Q_{T,H} &= \hat{\sigma}^2_{T,u} + T^{-1}q^{(3)}_{\beta,H} + q_{u,H} + o_{p,H}\big(HT^{-1} + H^{-1}\big), \\ \omega_{T,H} &= \sigma^2_u + T^{-1}q^{(3)}_{\beta,H} + q_{u,H} + o_H\big(HT^{-1} + H^{-1}\big), \end{aligned} \tag{2.18}$$

where $T^{-1}q^{(3)}_{\beta,H} + q_{u,H} = \lambda^{(3)}_\beta HT^{-1} + \lambda_u H^{-1} + o\big(HT^{-1} + H^{-1}\big)$, as $H \to \infty$.

Theorem 2 implies that the forecast obtained using data tuned parameter $\hat{H}$ has the same MSE as using H_{opt}, which can be estimated by $Q_{T,\hat{H}}$. More precisely, the following holds.

Corollary 2 *For a unit root process β_t, as in (b2), $\hat{H}$ stays bounded, and*

$$\omega_{T,\hat{H}} = \omega_{T,H_{opt}} + o(1), \qquad Q_{T,\hat{H}} = \omega_{T,\hat{H}} + o(1) = \sigma^2_u + q^{(2)}_{\beta,H_0} + q_{u,H_0} + o_p(1), \tag{2.19}$$

where H_0 is the minimiser of $q^{(2)}_{\beta,H} + q_{u,H}$.

For a bounded unit root process β_t, as in (b3), $\hat{H}$ stays bounded, if $q_{u,H}$ reaches its minimum at some finite point H_0. Then

$$\omega_{T,\hat{H}} = \omega_{T,H_{opt}} + o_p(1), \qquad Q_{T,\hat{H}} = \omega_{T,\hat{H}} + o_p(1) = \sigma^2_u + q_{u,H_0} + o_p(1). \tag{2.20}$$

Otherwise, if $q_{u,H}$ reaches its minimum at infinity, then

$$\omega_{T,\hat{H}} = \omega_{T,H_{opt}} + O(T^{-1/2}), \quad Q_{T,\hat{H}} = \omega_{T,\hat{H}} + O_p(T^{-1/2}) = \sigma^2_u + O_p(T^{-1/2}).$$

Remark 2 When β_t is a unit root trend, the optimal $\hat{H}$ may require averaging over the last few observations, to minimise the effect of the noise u_t. As a rule $\hat{H}$ will not take large values, and the rolling window will be narrow, but not necessary consisting of a single last observation. To illustrate the selection of H for the rolling window, consider the example of a random walk plus noise $y_t = \beta_t + u_t$, where $\xi_t := \beta_t - \beta_{t-1} \sim \text{IID}(0, \sigma^2_b)$, $u_t \sim \text{IID}(0, \sigma^2_u)$ and $\sigma^2_u > \sigma^2_b/2$. Then,
$\omega_{t,1} = E(y_{t+1} - y_{t+1|t,1})^2 = E(y_{t+1} - y_t)^2 = E(u_{t+1} + \xi_{t+1} - u_t)^2 = 2\sigma^2_u + \sigma^2_b$;
$\omega_{t,2} = E(y_{t+1} - y_{t+1|t,1})^2 = E\big(y_{t+1} - (y_t + y_{t-1})/2\big)^2 = (3/2)\sigma^2_u + (5/4)\sigma^2_b$.
Hence, $\omega_{t,1} > \omega_{t,2}$, which implies $\hat{H} \geq 2$.

When β_t is a bounded unit root trend, $\hat{H}$ aims to minimise (2.17). If $q_{u,H}$ does not attain its minimum at a finite point, e.g. if $u_t \sim i.i.d.$, then $\hat{H}$ minimises $\lambda^{(3)}_\beta HT^{-1} + \lambda_u H^{-1}$ and increases as $\hat{H} \sim (\lambda_u/\lambda^{(3)}_\beta)^{1/2}T^{1/2}$. Then the forecast error satisfies $Q_{T,\hat{H}} = \hat{\sigma}^2_{T,u} + 2(\lambda_u\lambda^{(3)}_\beta)^{1/2}T^{-1/2} + o_p(T^{-1/2})$.

2.6 The case of a deterministic trend and a structural break

Next, we analyse the properties of the forecast of y_t, when $\beta_t = \beta_{T,t}$ is a deterministic trend and u_t is a stationary noise. We consider three cases:

(b4), where β_t is an unbounded trend: $\beta_t = tg(t/T)$, setting $\kappa_3 := (\int_0^\infty K(x)xdx)^2$,

$$q^{(4)}_{\beta,H} := c(g)(\textstyle\sum_{j=1}^\infty w_{j,H}j)^2, \quad \lambda^{(4)}_\beta := c(g)\kappa_3 \qquad c(g) := \int_0^1 \big(g(x) + x\dot{g}(x)\big)^2 dx. \tag{2.21}$$

(b5), where β_t is a bounded trend: $\beta_t = g(t/T)$, setting

$$q^{(5)}_{\beta,H} := c'(g)(\textstyle\sum_{j=1}^\infty w_{j,H}j)^2, \quad \lambda^{(5)}_\beta := c'(g)\kappa_3, \qquad c'(g) := \int_0^1 \dot{g}^2(x)dx. \tag{2.22}$$

(b6), where β_t models the break in the mean: $\beta_{t,T} = \mu_1; \quad t = 1, \cdots, t_0;$, $\beta_{t,T} = \mu_2, \quad t = t_0 + 1, \cdots, T$, where $\Delta = |\mu_1 - \mu_2| \neq 0$ and the post-break period $\tau = T - t_0 = o(T)$. We set

$$q^{(6)}_{\beta,H} := \Delta^2 T_n^{-1} \sum_{t=t_0}^{T} \big(\sum_{j=t-t_0}^{t-1} w_{j,H})^2, \quad G_{\tau,H} := \Delta^2 \int_0^{\tau/H} \big(\int_x^{\infty} K(v)dv \big)^2 dx. \tag{2.23}$$

Notations $q_{u,H}$ and λ_u are as in Theorem 1.

Theorem 3 *Let $y_t = \beta_t + u_t, t = 1, \cdots, T$ where u_t is an $I(0)$ process. Then, as $T \to \infty$, for $H \in I_T$ the following holds.*

(i) For unbounded trend β_t (b4),

$$Q_{T,H} \;=\; \hat{\sigma}^2_{T,u} + q^{(4)}_{\beta,H} + q_{u,H} + o_{p,H}(H^2), \qquad \omega_{T,H} = \sigma^2_u + \delta_g q^{(4)}_{\beta,H} + q_{u,H} + o_H(H^2), \tag{2.24}$$

where $q^{(4)}_{\beta,H} + q_{u,H} = \lambda^{(4)}_{\beta} H^2 + o(H^2)$, as $H \to \infty$, and $\delta_g := (g^2(1) + \dot{g}^2(1))/c(g)$.

(ii) For bounded trend β_t (b5),

$$\begin{aligned} Q_{T,H} &= \hat{\sigma}^2_{T,u} + T^{-2} q^{(5)}_{\beta,H} + q_{u,H} + o_{p,H}\big(H^2T^{-2} + H^{-1}\big), \\ \omega_{T,H} &= \sigma^2_u + T^{-2}\delta'_g q^{(5)}_{\beta,H} + q_{u,H} + o_H\big(H^2T^{-2} + H^{-1}\big), \quad \delta'(g) = \dot{g}^2(1)/c'(g), \end{aligned} \tag{2.25}$$

where $q^{(5)}_{\beta,H} \sim H^2 \lambda^{(5)}_{\beta}$, $q_{u,H} \sim \lambda_u H^{-1}$, as $H \to \infty$.

(iii) For break point in β_t (b6),

$$\begin{aligned} Q_{T,H} &= \hat{\sigma}^2_{T,u} + q^{(6)}_{\beta,TH} + q_{u,H} + o_{p,H}\big(HT^{-1} + H^{-1}\big), \\ \omega_{T,H} &= \sigma^2_u + \tilde{\lambda}^{(6)}_{T,H} + o_H\big(T^{-1}\big), \quad \tilde{\lambda}^{(6)}_{T,H} := \Delta^2 \big(\sum_{j=T+1-t_0}^{T} w_{j,H})^2 + q_{u,H}, \end{aligned} \tag{2.26}$$

where $q^{(6)}_{\beta,TH} + q_{u,H} = G_{\tau,H} HT^{-1} + \lambda_u H^{-1} + o\big(T^{-1}\big)$, as $H \to \infty$.

Theorem 3 obtains the asymptotic properties of $Q_{T,H}$ that allow the derivation of the following characteristics of the forecast $y_{T+1|T,\hat{H}}$.

Corollary 3 *For a deterministic trend, β_t, as in (b4), $\hat{H}$ stays bounded. For a linear trend $\beta_t = ct$,*

$$\omega_{T,\hat{H}} = \omega_{T,H_{opt}} + o(1), \qquad Q_{T,\hat{H}} = \omega_{T,\hat{H}} + o(1) = \sigma^2_u + q^{(4)}_{\beta,H_0} + q_{u,H_0} + o_p(1), \tag{2.27}$$

where H_0 is a minimiser of $q^{(4)}_{\beta,H} + q_{u,H}$.

For a bounded deterministic trend, β_t as in (b5), $\hat{H}$ stays bounded, if $q_{u,H}$ reaches its minimum at some finite point H_0. Then,

$$\omega_{T,\hat{H}} = \omega_{T,H_{opt}} + o(1), \qquad Q_{T,\hat{H}} = \omega_{T,\hat{H}} + o(1) = \sigma^2_u + q_{u,H_0} + o_p(1). \tag{2.28}$$

Otherwise, if $q_{u,H}$ reaches its minimum at infinity, then

$$\omega_{T,\hat{H}} = \omega_{T,H_{opt}} + O(T^{-2/3}), \qquad Q_{T,\hat{H}} = \omega_{T,\hat{H}} + O_p(T^{-1/2}) = \sigma^2_u + O_p(T^{-1/2}).$$

Remark 3 In the presence of a deterministic trend (b4), the optimal $\hat{H}$ will be small and the forecast will be based on averaging over the last few observations, but it may not consist of a single last observation, unless the noise u_t is negligible.

In the presence of a bounded smooth deterministic trend (b5) and $u_t \sim i.i.d.$, Theorem 3(ii) implies that the optimal $\hat{H}$ will tend to minimise $\lambda_\beta^{(5)}(H/T)^2 + \lambda_u H^{-1}$ and increase as $\hat{H} \sim (\lambda_u/2\lambda_\beta^{(5)})^{1/3}T^{2/3}$, yielding $Q_{T,\hat{H}} = \hat{\sigma}_{T,u}^2 + cT^{-2/3} + o_p(T^{-2/3})$, $c = (3/2)(2\lambda_u^2\lambda_\beta^{(5)})^{1/3}$.

The following corollary develops further the result of Theorem 3(iii). It shows that with a structural break in the mean, the time needed for the optimisation procedure to detect the break and switch the weighting to post-break data is proportional to $\sqrt{T}$.

Corollary 4 *Let y_t combine the break in the mean (b6) and an i.i.d. noise u_t.*
(i) If the post-break period $\tau = T - t_0$ satisfies $\tau/\sqrt{T} \to \infty$, then, with $\lambda_\beta := \Delta^2 \int_0^\infty \left(\int_x^\infty K(u)du \right)^2 dx$,

$$\hat{H} \sim (\lambda_u/\lambda_\beta)^{1/2}T^{1/2}, \qquad Q_{T,\hat{H}} = \hat{\sigma}_{T,u}^2 + 2(\lambda_u\lambda_\beta)^{1/2}T^{-1/2} + o_p(T^{-1/2}). \tag{2.29}$$

Moreover, $\omega_{T,\hat{H}} = \sigma_u^2 + O(r_T)$ and $\omega_{T,\hat{H}} = \omega_{T,H_{opt}} + O(r_T)$, where $r_T = T^{-1/2} + e^{-2c\tau/\sqrt{T}}$ and c is the same as in (2.3).
(ii) If $\tau = o(\sqrt{T})$, then $\hat{H}$ is not affected by the break, $\hat{H}/\sqrt{T} \to \infty$ and $Q_{T,\hat{H}} = \hat{\sigma}_{T,u}^2 + o_p(T^{-1/2})$, whereas $\omega_{T,\hat{H}} = \sigma_u^2 + \Delta^2 + o(1)$.

The proof of this corollary can be found in the appendix.

Example 2 If y_t contains the break in the mean (b6) and an i.i.d. noise u_t, then forecasting with the rolling window weights, will yield $c_\Delta = \Delta^2/3$ and $c_\Delta/\lambda_u = 3\sigma_u^2/\Delta^2$. Thus, in finite samples, if the time expired after the break $\tau > (\sigma_u/\Delta)\sqrt{3T}$, then the optimisation will tend to select the window width $\hat{H} \leq \tau$, and the forecast will be based on the data from the post-break period. However, for more recent breaks, such that $\tau < (\sigma_u/\Delta)\sqrt{3T}$, the forecast may not switch to the post-break data. The waiting time for such a switch is defined by the ratio σ_u/Δ and $\sqrt{3T}$. We briefly examine these matters with a Monte Carlo forecasting experiment based on 200 observations of the sequence $y_t = u_t + I(t \geq 160)$ where u_t are i.i.d.$(0, 1)$ normal random variables, and y_t has a break in the mean from 0 to 1 at time $t_0 = 160$. Over 1000 replications we get, as expected, that the full sample mean $200^{-1}\sum_{t=1}^{200} y_t$ produces a bad forecast for time period 201, compared to the sample mean $40^{-1}\sum_{t=160}^{200} y_t$ over the last 40 observations from the post break period. The relative MSE of the post break sample mean is 0.61, compared to the full sample mean. Both data-based exponential weighting and rolling window forecasts perform much better than the full sample mean, with relative MSEs of 0.65 and 0.69 respectively.

2.7 Examples

In order to get a better feel for the behaviour of the data-selected tuning parameters, we consider one single realisation of sequentially computed $\hat{H}_t$, $t = t_0, t_0 + 1, \cdots, T$ for two structural change experiments used in our Monte Carlo study below. We look at rolling window forecasts. Figures 1 and 2 report the starting point (solid line) of the data-selected rolling window for a structural break in the mean (Experiment 4 of our Monte Carlo study) and a unit root model (Experiment

11), respectively. The sample size T is 200 and the forecasting starts at $t_0 = 100$.[2] For comparison, we also report the first observation in the data-estimated rolling window when the model has no structural change (Experiment 1 in the Monte Carlo study), based on the same realisations of the noise u_t, as in the previous two cases (dotted line). The vertical distance between the diagonal (long dashes, the last observation in the window) and the starting point solid (dotted) line for a given $t = 100, \cdots, 200$ shows the time span of observations (a graphical realisation of the tuning parameter) used for forecasting, that is $t - \hat{H}_t$. It is clearly seen that, under structural change, the estimated tuning parameter selects a much smaller sample for forecasting than in the absence of structural change. Figure 1 shows that, for the structural break (at observation 110) the data-dependent method is attempting to get more information about the change immediately after the break by initially using a larger sample for forecasting. This then becomes smaller than that in the no-change case, as more data after the breakpoint accrue. Interestingly, after observation 125, the starting point of the rolling window is the first post break observation 111 (short-dashed line), as suggested by theory. Notice that 125 is close to the theoretical switching time $110+\sqrt{3(110)} = 128$ (see Example 2). Moreover, it remains at that point for much of the rest of the sample. In Figure 2, we can see that with a unit root, the window remains short throughout the sample. A final diagnostic for the method is the value of the estimated mean squared error obtained in real time. This is given in Figure 3, where the dotted line relates to the stationary case, the long-dashed line to the structural break case and the solid line to the unit root case. The smallest MSE is obtained in the stationary case followed by the structural break and finally the unit root, which is the ranking one would expect.

2.8 Extensions

Our proposed method extends in several practically relevant ways. In this section we briefly discuss some of these.

Nonparametric method

The above analysis presupposes a particular parametric form for the weight function. While that might be desirable from the usual motivation of parsimony, in some circumstances it will be restrictive. For example, monotonic downweighting might be counterproductive when data come from a process that follows a finite number of regimes. Data from the same regime as that holding during the latest forecast period may be more relevant than more recent data. To account for such possibilities, we construct a nonparametric weighting scheme.

Again we focus on the simple location model (2.1) assuming that β_t is some smooth deterministic function of t and u_t is a standardised IID$(0, 1)$ noise. We consider forecasts of y_t of the form

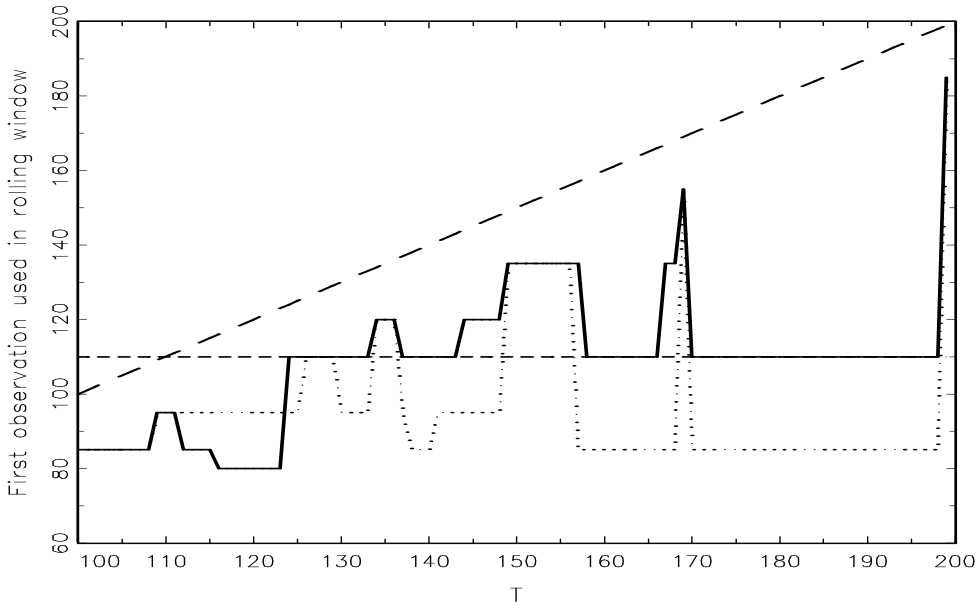
$$\hat{y}_{t|t-1} = \textstyle\sum_{j=1}^{t-1} w_{tj} y_{t-j}. \tag{2.30}$$

We wish to determine a nonparametric set of weights w_{Tj}, $j = 1, \cdots, T-1$, such that the forecast MSE of $\hat{y}_{T|T-1}$ is minimised subject to $\sum_{i=1}^{T-1} w_{Tj} = 1$. Letting $\tilde{\beta}_t = \beta_t - \beta_T$,

$$E(\hat{y}_{T|T-1} - y_T)^2 = \left(\textstyle\sum_{j=1}^{T-1} w_{Tj}\tilde{\beta}_{T-j}\right)^2 + \sigma_u^2 \textstyle\sum_{j=1}^{T-1} w_{Tj}^2.$$

[2]Details on how the parameter $\hat{H}_t$ is estimated are given in Section 3.

Figure 1: Realisation of the data-selected rolling window for a structural break. The solid line represents the starting point of the window for a structural break model with a break at observation 110 (Experiment 4 of the Monte Carlo study), and the dashed line (long dashes) shows the last observation in the window. The dashed line (short dashes) indicates the first post break observation, and the dotted line the beginning of the window when there is no structural change.



We construct the Lagrangean

$$L(\lambda, w_{T1}, \cdots, w_{T,T-1}) = \big(\sum_{j=1}^{T-1} w_{Tj}\tilde{\beta}_{T-j}\big)^2 + \sigma_u^2 \sum_{j=1}^{T-1} w_{Tj}^2 - \lambda\big(\sum_{j=1}^{T-1} w_{Tj} - 1\big).$$

Taking derivatives of L w.r.t. the w_{Tj}s and equating them to zero, gives equations

$$(\tilde{\beta}_{T-j}^2 + \sigma_u^2)w_{Tj} + \tilde{\beta}_{T-j} \sum_{i=1, i\neq j}^{T-1} \tilde{\beta}_{T-i} w_{Ti} = \lambda/2, \quad j = 1, \cdots, T-1.$$

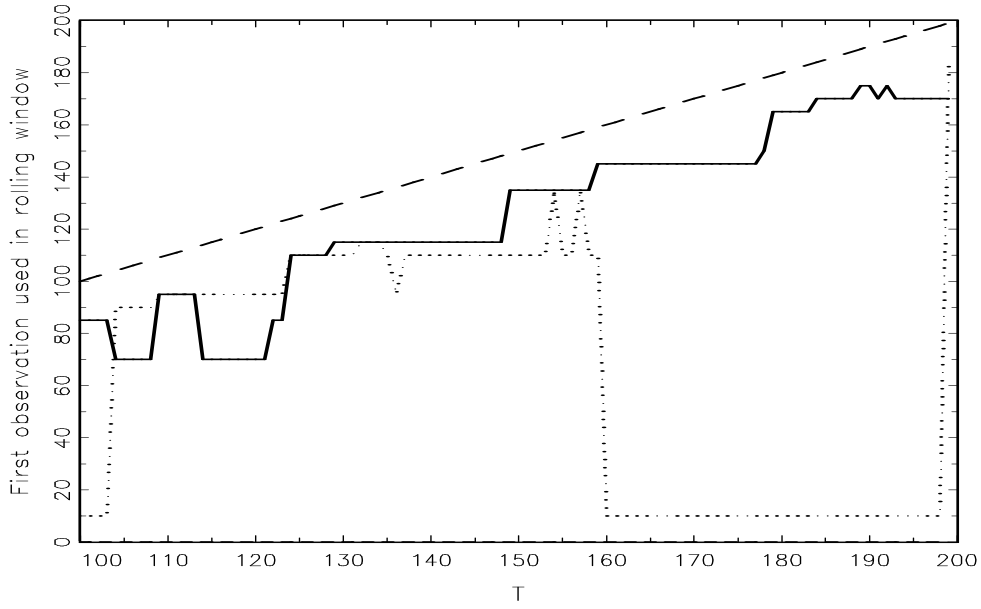
We need to solve this set of equations. As a system they are written as

$$(\tilde{B} + \sigma_u I)w_T = (\lambda/2)1, \qquad \text{or} \quad Bw_T = \Lambda, \tag{2.31}$$

where $\tilde{B} = (\tilde{\beta}_{T-j}\tilde{\beta}_{T-k})_{j,k=1,\ldots,T-1}$ is a $(T-1)\times(T-1)$ matrix, I is a $(T-1)\times(T-1)$ identity matrix, $w_T = (w_{Tj})_{j=1,\cdots,T-1}$ is a $(T-1)\times 1$ vector, 1 is $(T-1)\times 1$ unit vector, $B = \tilde{B} + \sigma_u I$ and $\Lambda = (\lambda/2)1$.

Then, $w_T = B^{-1}\Lambda$, and λ is determined such that the sum of the elements of $B^{-1}\Lambda$ is unity. This is not an operational procedure as β_T is unknown at time $T-1$. We suggest setting $\beta_t = \hat{\beta}_t$,

Figure 2: Realisation of the data-selected rolling window for a unit root. The solid line shows the starting point of the window for a unit root model (Experiment 11 of the Monte Carlo study), and the dashed line (long dashes) the last observation in the window. The dotted line indicates the beginning of the window when there is no structural change.



$t = 1, \cdots, T-1$ and $\beta_T = \hat{\beta}_T = \hat{\beta}_{T-1}$ where $\hat{\beta}_t$ denotes some estimator of β_t. This approach does not allow for a dependent u_t, but we discuss possible extensions of (2.1) below that make the assumption of a serially uncorrelated u_t more plausible.

The method can be extended to allow for time-varying variances $Eu_t^2 = \sigma_{u,t}^2$. Then, the forecast MSE takes the form

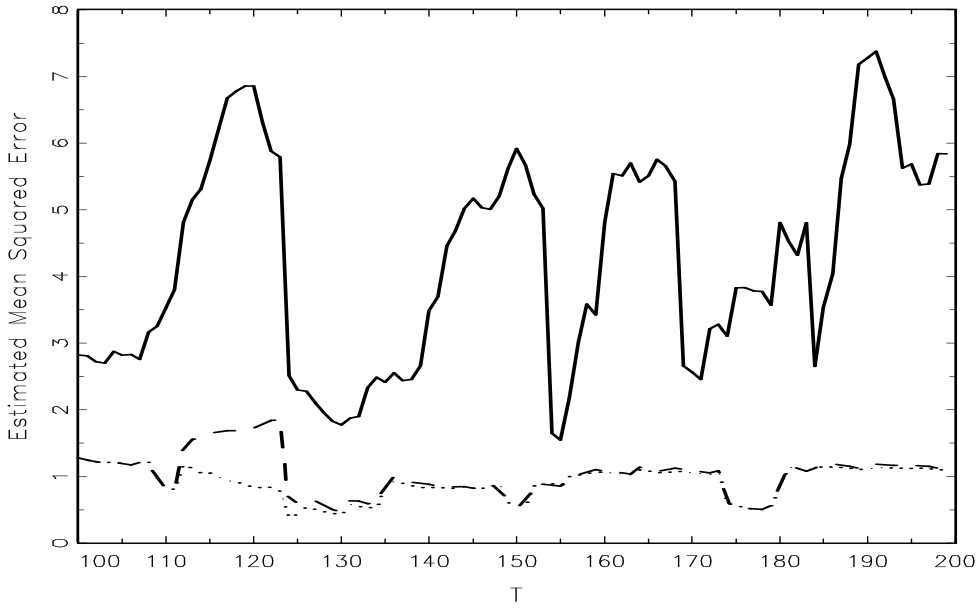
$$E(\hat{y}_{T|T-1} - y_T)^2 = \left(\sum_{j=1}^{T-1} w_{Tj}\tilde{\beta}_{T-j}\right)^2 + \sum_{j=1}^{T-1} w_{Tj}^2 \sigma_{u,T-j}^2.$$

Following the steps of the previous argument gives the following system of equations

$$(\tilde{B} + \tilde{I})w_T = (\lambda/2)1, \qquad \text{or} \quad Bw_T = \Lambda,$$

where $\tilde{I} = diagonal(\sigma_{u,T-1}^2, \cdots, \sigma_{u,1}^2)$ is $(T-1) \times (T-1)$ diagonal matrix. Once again this procedure becomes operational by replacing $\sigma_{u,t}^2$ with an estimate. We note that estimation of β_t and $\sigma_{u,t}^2$ is discussed widely in the literature when β_t and $\sigma_{u,t}^2$ are deterministic functions of time (see, e.g., Orbe, Ferreira, and Rodriguez-Poo (2005) and Kapetanios (2007)), and is examined in Giraitis, Kapetanios, and Yates (2011) for stochastic β_t.

Figure 3: Realisation of the estimated MSE. The dotted line shows the MSE for the stationary case, the long-dashed line for the structural break case and the solid line for the unit root case.



Subsamples

Another extension allows the forecast MSE to be evaluated and minimised over different sample periods, in order to select the optimal subsample and a specific tuning parameter. This is achieved by an extended two-parameter minimisation procedure given by

$$Q_{T,kH} := (T-k)^{-1} \sum_{t=k}^{T} (\hat{y}_{t|t-1,\,H} - y_t)^2, \qquad \{\hat{H}, \hat{k}\} := \text{argmin}_{H \in I_T, k \in \{k_{\min}, \cdots, k_{\max}\}} Q_{T,H,k}. \quad (2.32)$$

The selected values of $(\hat{H}, \hat{k})$ can then be used to construct forecasts based on the subsample $[\hat{k}, \cdots, T]$. This value of H may be different from that obtained by the optimisation in (2.5). Such a procedure, when building forecasts, seeks for an optimal subsample $y_{\hat{k}}, \cdots, y_T$ ('stability period') and an associated optimal tuning parameter $\hat{H} = \hat{H}(\hat{k})$. Observe that for the rolling window forecast, obviously $\hat{H} \leq T - \hat{k}$. However, using exponential downweighting, only data $y_{\hat{k}}, \cdots, y_T$ should be used.

The advantage of the two-parameter procedure becomes obvious in rolling window forecasts under the break in the mean, discussed in Example 2. If the rolling window is selected using all the data in a large sample $y_1, \cdots, y_T$, then it takes $\sqrt{T}$ time lags for the forecast to switch to the postbreak data. However, the switch may be faster when less observations are used (i.e., when $\hat{k} >> 1$ is selected, reducing the weight of irrelevant past information). Our theoretical findings show that the two-parameter minimisation will minimise the forecast MSE leading to the smallest possible MSE with optimal downweighting and the most relevant data subsample.

Dynamic weighting

Another simple way to allow for extra flexibility in the weight function is to allow the first p weights $w_1, \cdots, w_p$ ($p \geq 0$) to vary freely by specifying

$$\tilde{w}_{tj,H} = \begin{cases} w_j, & j = 1, \cdots, p, \\ K(j/H), & j = p+1, \cdots, t-1, \qquad H \in I_T, \end{cases} \tag{2.33}$$

and standardising the weights: $w_{tj,H} = \tilde{w}_{tj,H} / \left(\sum_{j=1}^{t-1} \tilde{w}_{tj,H} \right)$. This allows the first few lags of y_t to enter freely into the forecast rather than through a given parametric function, akin to an estimated AR process. Then, Q_T can be minimised jointly over $H, \tilde{w}_1, \cdots, \tilde{w}_p$, and, potentially, even p.

Conditional mean modelling

The location set-up in (2.1) does not allow for explicit conditional mean modelling. In this subsection we address this issue. It would be good if our analysis allows both the use of a generic model of the conditional mean of the process and robust forecasting around that model. Specifically, we would like to assume that the forecaster has a preferred model of the conditional mean which is known (at least up to a finite vector of unknown parameters), and then discuss how our robust adaptive forecasting methods can be applied to the residual from such a model. This allows considerable generality, and in practice allows application to realistic conditional mean models such as the widely used AR model.

In the conditional mean framework, one has a generic forecasting model of the form

$$x_t = g(z_t) + y_t \tag{2.34}$$

of the variable of interest x_t that produces forecasts $g(z_{t+1})$ based on a vector of predictor variables that may contain lags of x_t, or other generated variables such as, e.g., dummies to account for structural change. The process y_t in (2.34) is the part of x_t unexplained by $g(z_t)$. Assuming that the conditional mean function g has a known parametric structure up to an unknown finite dimensional parameter, fitting it to x_t gives rise to a parametric forecasting model. Clearly, such a model can be misspecified and may suffer problems associated with the presence of structural change in x_t, as discussed in the introduction. We will abstract from specification and estimation issues associated with g. This is because we wish to keep our discussion as general as possible and not related to the exact structure of g. Additionally, the presence of structural change in x_t is likely to complicate considerably any rigorous analysis of estimators of the unknown parameters. Moreover, our analysis of forecasting y_t will efficiently exploit any persistence remaining in y_t. Hence, it is sufficient to assume that fitting the model $g(z_t)$ to x_t produces y_t with an unspecified persistent structure that may combine dependence, trends and breaks. Once (2.34) is posited and the possibility allowed of suboptimal forecasts by $g(z_{t+1})$ due to structural change, it is important to consider ways in which additional forecasting of y_t may produce a superior forecast of x_t. In principle, a fixed conditional mean function g could be extended to a time-varying function, g_t, known or estimated by any of the currently available methods in the literature. However, under ongoing structural change, the properties of such an estimator may be difficult to determine.

In summary, for any given forecast $\hat{x}_t$ of x_t, based on information available up to time $t-1$, we shall write $x_t = \hat{x}_t + y_t$, $t = 1, \cdots, T$. Then we can use our robust methods to produce a forecast $\hat{y}_{T+1}$ of y_{T+1}, based on $y_t = x_t - \hat{x}_t$, $t \leq T$ and define the final forecast of x_{T+1} as

$x_{T+1}^{(forecast)} = \hat{x}_{T+1} + \hat{y}_{T+1}$. For example, we may set $\hat{x}_t \equiv 0$, and then $y_t = x_t,\ t \leq T$. Alternatively, we can fit to the data, x_t, some model of the form $g(x_{t-1}, x_{t-2} \cdots)$ and after obtaining its estimate, $\hat{g}$, we arrive at $\hat{x}_t = \hat{g}(x_{t-1}, x_{t-2} \cdots)$. It may be the case, as it sometimes is in policy institutions such as central banks, that g or $\hat{x}_t$ is obtained using informal judgements by policymakers. Note that any neglected dynamics or errors produced by such a fitting process will be accumulated in y_t and used subsequently to forecast y_{T+1}.

2.9 Theoretical conclusions

We conclude this section by noting some important implications of our analysis.

First, the dominant tendency in the forecasting literature of using models developed for non-forecast purposes, such as to generate impulse responses or policy analysis, may be counterproductive. Our arguments suggest that if good forecasting is the aim, then forecasting by averaging or appropriately downweighting past data, without engaging in further modelling, is a viable strategy.

Second, appropriately downweighting past can provide a general approach for handling trends of any nature. Our theoretical results show that this method applies for stochastic, linear or nonlinear deterministic trends and structural breaks without knowledge of the nature of the trend. It is therefore a tractable method for forecasting the levels of apparently nonstationary processes. As a result it bypasses difficult problems of combining appropriate detrending of level series with the subsequent forecasting of stationary processes. Importantly, the proposed forecasting approach continues to be valid if a series is actually stationary.

Finally, while theoretical results, such as, e.g., Remark 1, and small sample evidence indicate that an exponential kernel has theoretical advantages over a rolling window and is a very good choice in general, in a particular empirical application another kernel function may still be preferable. It is then worth noting that the MSE minimisation procedure determining the rate of downweighing past data can be used to select the kernel function, K, that produces the lowest MSE, among a set of admissible kernel functions.

3 Monte Carlo study

In this section we explore the finite sample performance of the forecasting strategies discussed in the previous section. We consider Monte Carlo experiments for the forecast of y_{T+1} based on the sample $y_1, \cdots, y_T$ for a number of specific designs for the simple location model (2.1) with β_t following a variety of processes analysed in the previous section. We also consider a variety of models with short memory dynamics. We analyse one-step ahead forecasts where the benchmark is the sample mean forecast $\hat{y}_{benchmark,T+1} = T^{-1} \sum_{t=1}^{T} y_t$ or an autoregressive $AR(1)$ forecast. The benchmarks disregard the possibility of structural change. We also consider a benchmark of the last available observation forecast, optimal when the process is a random walk. We compare the performance of the various forecasts in terms of relative MSE.

Design: data-generating processes. We consider the following location shift model (2.1) for generating the data:

$$y_t = \beta_t + u_t, \qquad t = 1, \cdots, T,$$

where u_t is either a standard normal IID$(0, 1)$ noise, or an $AR(1)$ process with parameter $\rho = 0.7$ or -0.7 and standard normal i.i.d. innovations. The process β_t is either a deterministic or stochastic

trend, or a process with a break in the mean. We consider the following data-generating processes, denoted in tables as $Ex1$–$Ex12$:

1. $y_t = u_t$.
2. $y_t = 0.05t + 5u_t$.
3. $y_t = 0.05t + 3u_t$.
4. $y_t = \begin{cases} u_t, & t \leq t_0 = T/2, \\ 1 + u_t, & t > t_0. \end{cases}$
5. $y_t = 2\sin(2\pi t/T) + 3u_t$.
6. $y_t = 2\sin(2\pi t/T) + u_t$.
7. $y_t = 2T^{-1/2}\sum_{i=1}^{t} v_i + 3u_t$.
8. $y_t = 2T^{-1/2}\sum_{i=1}^{t} v_i + u_t$.
9. $y_t = 0.5\sum_{i=1}^{t} v_i + 3u_t$.
10. $y_t = 0.5\sum_{i=1}^{t} v_i + u_t$.
11. $y_t = \sum_{i=1}^{t} v_i$.
12. $y_t = \sum_{i=1}^{t} u_i$,

where v_t is a standard normal IID$(0, 1)$ sequence. This selection of deterministic trends provides a variety of shaped functions driving the structural change in the unconditional mean of y_t.

$Ex1$ is the case of no structural change. Here, as long as the noise u_t is an i.i.d. or very weakly dependent process, the benchmark sample mean forecast should do best, and the robust methods at most should not lag far behind the benchmark. If u_t is a dependent process with persistent autoregressive dynamics then the AR benchmark should do best. Further, in this case, the robust forecast with EWMA weights should outperform the sample mean benchmark and rolling window (see Remark 1). Theory indicates that the exponential weights should outperform the rolling window, but it leaves open the possibility that the rolling window can outperform the benchmark when a stationary process y_t becomes persistent.

The functional form in $Ex2$ and $Ex3$ is a linear monotonic trend. While such trends may be unrealistic, at least for processes which have been detrended by applying filters or differencing, they provide a useful benchmark. Further, these trends are sufficiently subtle and minor to be swamped visually by the noise process. We consider different values for the variance of the noise process to explore such effects. The purpose of $Ex4$ is to introduce a break in the mean, to see if our robust methods can help under traditional structural change specifications. The break occurs at time $t_0 = T/2$, and the post-break time is greater than $\sqrt{T}$, as required by the theory. Hence the break is not 'too recent' and it will be taken into account by the robust forecasting method, leading to significant improvement of forecast quality comparing to full sample benchmarks. Moreover, the effect is amplified by the increase of dependence in the error process u_t.

Functions in $Ex5$ and $Ex6$ represent smooth cyclical bounded trends. These are more likely to remain after standard detrending and provide a realistic scenario. Moreover, wider oscillation of the trend in $Ex6$ relative to the variance of the noise process seems to lead to a stronger deterioration of the performance of the benchmark.

Next, $Ex7$ and $Ex8$ deal with a bounded stochastic trend β_t which is relevant for popular time-varying coefficient specifications in the macroeconometric and forecasting literature, while $Ex9$ and $Ex10$ deal with a random walk (unit root) process, observed under noise. Finally, $Ex11$ and $Ex12$ consider two versions of a standard random walk model, differing only in the persistence of the noise processes.

3.1 Forecast methods

We examine the robust forecasting methods using three classes of parametric weight functions.

Rolling window. This uses flat weights,

$w_{tj,H} = H^{-1}I(1 \leq j \leq H)$, $j = 1, \cdots, t-1$, for $H < t$, and
$w_{tj,H} = (t-1)^{-1}I(1 \leq j \leq t-1))$, for $H \geq t$,

giving equal weight to recent data and zero weight to older data. We denote it in the tables below by *Rolling* H where H is the window size.

Exponential (EWMA). This uses weights

$w_{tj,\rho} = \rho^{t-j} / \left(\sum_{k=1}^{t-1} \rho^k \right)$, $1 \leq j \leq t-1$, with $0 < \rho < 1$.

Here the main weight is placed on the last few data points, downweighting others to zero exponentially fast when ρ is small, and more equally when ρ is close to 1. We refer to this as *Exponential* ρ.

Polynomial method. This uses weights

$w_{tj,\alpha} = (t-j)^{-\alpha} / \left(\sum_{k=1}^{t-1} k^{-\alpha} \right)$, $1 \leq j \leq t-1$ with $\alpha > 0$.

The past is downweighted at a slower rate than with exponential weights. We refer to it as *Polynomial* α.

Methods with fixed tuning parameters. We consider forecasts with both fixed values of H and ρ, and data-selected values $\hat{H}$, $\hat{\rho}$ and $\hat{\alpha}$ for the tuning parameters. With polynomial weights we do not examine the fixed value cases. We set $H = 20, 30$ for rolling window and for exponential weights $\rho = 0.99, 0.95, 0.9, 0.8, 0.7$ and 0.5. Using fixed values allows us to compare the performance of the forecast with a data-tuned parameter with the best (smallest Monte Carlo forecast MSE) among the fixed cases. Our objective is to verify in simulations that these two MSEs, $\omega_{T,\hat{H}}$ and $\omega_{T,H_{opt}}$, are comparable, as indicated by Corollaries 1 to 3.

Nonparametric method. We also consider the nonparametric forecast method as in (2.30) and (2.31) based on the nonparametric weighting scheme. The corresponding results are referred to as *Nonparametric*.

Rolling $(\hat{k}, \hat{H})$ **method**. This is the rolling window forecast where $\hat{k}$ and $\hat{H}$ are selected minimising $Q_{T,kH}$ in H and k as in (2.32), referred to as *Rolling* $(\hat{k}, \hat{H})$.

Averaging method. The final robust method we examine is the averaging method of rolling window forecasts over different periods advocated by Pesaran and Timmermann (2007):

$$\bar{y}_{T+1|T} = \frac{1}{T} \sum_{H=1}^{T} \hat{y}_{T+1|T,\,H}, \qquad \hat{y}_{T+1|T,\,H} = \frac{1}{H} \sum_{t=T-H+1}^{T} y_t. \tag{3.1}$$

It combines rolling window forecasts of y_{T+1} using all possible windows that include the last available observation. A characteristic of this method is that it does not require selection of any tuning parameters apart from the mimimum sample size used for forecasting, which is usually of minor significance. We refer to this as *Averaging*.

Dynamic weighting method. This uses the weights defined in (2.33) with $p = 1$ and exponential K. We refer to it as *Dynamic weighting*.

Residual methods. We apply three methods to forecast $x_t = g(z_t) + y_t$, $t = 1, \cdots$, of (2.34). They fit to x_t the AR(1) dynamics $g(z_t) = \phi x_{t-1}$ and forecast residuals y_t by either a parametric or nonparametric method. The forecast of x_{t+1} based on $x_1, \cdots, x_t$ is $\hat{x}_{t+1} = \hat{\phi} x_t + \hat{y}_{t+1|t,\,\hat{H}}$.

Exponential AR method. It estimates the autoregressive parameter ϕ and the tuning parameter H jointly by minimising the forecast error $Q_{T,H} = Q_{T,H\phi}$ computed using $y_t = x_t - \phi x_{t-1}$ with exponential weights. We refer to it as *Exponential AR*.

The other two methods are two-stage methods, where the autoregressive parameter ϕ at x_{t-1} is estimated by OLS separately from the parameters associated with forecasting y_t.

Table 1: Monte Carlo Results. $T = 200$. One-Step Ahead Forecasts. $u_t \sim \text{IID}(0,1)$. Table reports relative mean squared error using a full sample mean benchmark

		Experiments										
Method		$Ex1$	$Ex2$	$Ex3$	$Ex4$	$Ex5$	$Ex6$	$Ex7$	$Ex8$	$Ex9$	$Ex10$	$Ex11$
Exponential	$\rho = \hat{\rho}$	1.085	0.699	0.436	0.791	0.802	0.253	1.029	0.691	0.622	0.212	0.042
Rolling	$H = \hat{H}$	1.066	0.694	0.448	0.807	0.804	0.276	1.005	0.696	0.627	0.272	0.153
Rolling	$H = 20$	1.039	0.658	0.413	0.762	0.759	0.264	0.977	0.668	0.618	0.323	0.268
	$H = 30$	1.027	0.654	0.420	0.768	0.764	0.312	0.965	0.685	0.659	0.403	0.373
Exponential	$\rho = 0.99$	1.003	0.736	0.570	0.833	0.836	0.556	0.964	0.766	0.750	0.592	0.562
	$\rho = 0.95$	1.040	0.656	0.412	0.751	0.759	0.258	0.972	0.652	0.595	0.271	0.192
	$\rho = 0.90$	1.102	0.690	0.426	0.778	0.795	0.242	1.023	0.667	0.592	0.211	0.104
	$\rho = 0.80$	1.234	0.772	0.472	0.861	0.892	0.263	1.142	0.733	0.645	0.196	0.062
	$\rho = 0.70$	1.414	0.888	0.538	0.983	1.028	0.299	1.304	0.833	0.731	0.208	0.048
	$\rho = 0.50$	1.947	1.231	0.737	1.352	1.431	0.411	1.783	1.137	0.998	0.268	0.041
Averaging		1.003	0.747	0.589	0.848	0.844	0.583	0.967	0.781	0.774	0.636	0.619
Nonparametric		1.102	0.671	0.415	0.770	0.778	0.239	1.015	0.667	0.595	0.242	0.155
Polynomial	$\alpha = \hat{\alpha}$	1.025	0.726	0.488	0.807	0.817	0.310	0.987	0.695	0.640	0.322	0.149
Rolling	$H = \hat{H}$, $k = \hat{k}$	1.061	0.720	0.473	0.812	0.825	0.283	1.011	0.699	0.636	0.243	0.106
Dynamic Weighting		1.162	0.719	0.452	0.807	0.822	0.260	1.082	0.711	0.630	0.214	0.045
Exponential AR		1.107	0.729	0.456	0.830	0.832	0.265	1.074	0.720	0.642	0.219	0.044
Exponential Residual		1.093	0.707	0.474	0.815	0.815	0.316	1.026	0.720	0.660	0.245	0.044
Nonparametric Residual		1.109	0.697	0.470	0.795	0.802	0.312	1.025	0.708	0.650	0.240	0.044
Last Observation		1.951	1.234	0.738	1.355	1.435	0.412	1.787	1.140	1.001	0.269	0.041
AR		1.000	0.805	0.599	0.826	0.870	0.381	0.978	0.844	0.790	0.310	0.052

Exponential residual method. It forecasts residuals $\hat{y}_t = x_t - \hat{\phi} x_{t-1}$ using exponential weights producing $\hat{H}$ and the forecast $\hat{y}_{t+1|t,\hat{H}}$. We refer to it as *Exponential Residual*.

Nonparametric residual method. It forecasts residuals $\hat{y}_t = x_t - \hat{\phi} x_{t-1}$ using the nonparametric forecast method. We refer to it as *Nonparametric Residual*.

3.2 Monte Carlo results

We choose a particular forecast starting point at time t_0 by any given method. Then one-step ahead forecasts $y_{t_0|t_0-1,H}, \cdots, y_{t|t-1,H}$, $t = t_0, ..., T$, are computed. The forecast evaluation period ends at T. Note that all forecasts for t are produced using only information up to $t-1$. To compare different forecast methods, as the performance criterion we use the forecast MSE relative to the benchmark of the sample mean of all data (MSE_{RR}). For method i, we compute $MSE_i = (T-t_0)^{-1} \sum_{t=t_0}^{T} (\hat{y}^{(i)}_{t|t-1} - y_t)^2$ and define the relative $MSE_{RR} = \frac{MSE_i}{MSE_0}$ where MSE_0 corresponds to the benchmark forecast by the sample mean. For all experiments, forecasting starts at $t_0 = 100$, and the sample size is $T = 200$. MSE_{RR} below unity shows that the forecast method outperforms the sample mean. We carry out 200 replications and report the average MSE_{RR} over these.

The relative MSE for models $Ex1$ to $Ex12$ obtained with our forecasting methods with data-selected and fixed tuning parameters are reported in Tables 1-3. In Table 1, the noise u_t is an i.i.d. standard normal process, whereas in Tables 2 and 3, the u_t are dependent variables, generated by stationary AR(1) processes with parameters $\rho = 0.7$ and $\rho = -0.7$ and i.i.d. standard normal innovations respectively.

The first column, labelled $Ex1$, corresponds to the stationary case $y_t = u_t$. In the i.i.d. case, as expected, the sample mean outperforms the forecasts for each method, especially those penalised by the loss of information from strong discounting. However, for sufficiently dependent u_t, discounting improves the forecast as indicated by Remark 1.

For the other experiments, in almost all cases, downweighting beats the sample mean in the sense that the MSE_{RR} is considerably below unity. Further, the full sample autoregressive model

Table 2: Monte Carlo Results. $T = 200$. One-Step Ahead Forecasts. $u_t \sim AR(0.7)$. Table reports relative mean squared error using a full sample mean benchmark

		Experiments										
Method		$Ex1$	$Ex2$	$Ex3$	$Ex4$	$Ex5$	$Ex6$	$Ex7$	$Ex8$	$Ex9$	$Ex10$	$Ex12$
Exponential	$\rho = \hat{\rho}$	0.632	0.418	0.285	0.591	0.465	0.141	0.572	0.358	0.351	0.110	0.008
Rolling	$H = \hat{H}$	0.964	0.605	0.425	0.956	0.707	0.238	0.878	0.559	0.580	0.248	0.100
Rolling	$H = 20$	1.027	0.678	0.433	1.027	0.764	0.272	0.936	0.618	0.635	0.341	0.213
	$H = 30$	1.045	0.684	0.450	1.014	0.775	0.318	0.941	0.637	0.689	0.431	0.320
Exponential	$\rho = 0.99$	0.978	0.732	0.572	0.966	0.813	0.547	0.911	0.713	0.744	0.610	0.555
	$\rho = 0.95$	0.865	0.561	0.374	0.852	0.642	0.232	0.790	0.518	0.527	0.262	0.149
	$\rho = 0.90$	0.758	0.484	0.327	0.739	0.558	0.181	0.696	0.442	0.437	0.176	0.065
	$\rho = 0.80$	0.656	0.419	0.287	0.623	0.480	0.150	0.597	0.374	0.366	0.130	0.028
	$\rho = 0.70$	0.606	0.393	0.269	0.568	0.444	0.137	0.551	0.344	0.335	0.112	0.016
	$\rho = 0.50$	0.617	0.411	0.278	0.569	0.451	0.138	0.562	0.351	0.340	0.106	0.008
Averaging		1.002	0.757	0.596	0.985	0.834	0.575	0.935	0.740	0.774	0.658	0.617
Nonparametric		0.970	0.593	0.386	0.968	0.694	0.218	0.893	0.556	0.551	0.235	0.106
Polynomial	$\alpha = \hat{\alpha}$	0.708	0.488	0.364	0.657	0.525	0.227	0.644	0.431	0.462	0.252	0.073
Rolling	$H = \hat{H}$, $k = \hat{k}$	0.824	0.532	0.359	0.788	0.591	0.192	0.740	0.470	0.475	0.197	0.068
Dynamic Weighting		0.610	0.393	0.268	0.577	0.441	0.137	0.560	0.348	0.341	0.112	0.011
Exponential AR		0.610	0.408	0.278	0.569	0.454	0.140	0.559	0.353	0.348	0.112	0.004
Exponential Residual		0.575	0.391	0.272	0.537	0.427	0.146	0.532	0.339	0.339	0.114	0.005
Nonparametric Residual		0.586	0.386	0.263	0.547	0.424	0.138	0.536	0.336	0.330	0.110	0.008
Last Observation		0.618	0.412	0.278	0.569	0.452	0.138	0.563	0.351	0.340	0.106	0.008
AR		0.556	0.380	0.248	0.556	0.406	0.140	0.515	0.370	0.374	0.144	0.011

although better than the sample mean forecast in the majority of cases is also beaten by down-weighting methods in several cases, particularly where there is a location shift or autoregressive dynamics. Generally, all these methods are useful, including the rolling window and averaging method. In the case of a fixed tuning parameter, for the model with a strong trend, the largest reduction of MSE_{RR} comes from the exponential weights with the highest discount rates. Although the tuned exponential weights are not the best, they are where they should be according to theory: comparable to the best fixed value methods and never among the poor performers. Note, e.g., that the exponential weights with a $\rho = 0.9$ fixed discount can perform both very well and considerably worse than the tuned exponential weights in a number of cases, illustrating the importance of data-dependent tuning.

Given that optimal fixed ρ for exponential weights cannot be observed in practice, our simulation study suggests the efficiency and usefulness of data based downweighting. The nonparametric method similarly offers a powerful alternative, for i.i.d. noise u_t slightly beating the tuned parameter methods in many cases. However, being designed for an i.i.d. noise u_t, in case of a dependent AR(1) noise this method is outperformed by the parametric tuning methods, unless coupled with an initial AR correction. It is also worth mentioning that while the benchmark full sample AR forecast is a good competitor in many cases, there are circumstances such as, for example, i.i.d. noise or autoregressive noise with negative autoregressive coefficients where it can perform considerably worse than robust downweighting methods.

Comparing exponential, rolling window and polynomial methods, the exponential method outperforms rolling windows while the latter beats polynomial windows when the noise u_t is dependent and is outperformed by it when the noise is i.i.d. The averaging method outperforms the benchmark but is beaten by the rolling windows with data-selected $\hat{H}$. The rolling window forecast using a data dependent window, $\hat{H}$, and an evaluation period $[\hat{k}, T]$, is equivalent to a rolling window with $\hat{H}$ and $k = 1$ under the i.i.d. noise, but outperforms it when the noise, u_t, is dependent.

It is worth noting that, in applications, one could select from a set of available forecasts with

Table 3: Monte Carlo Results. $T = 200$. One-Step Ahead Forecasts. $u_t \sim AR(-0.7)$. Table reports relative mean squared error using a full sample mean benchmark

		Experiments										
Method		$Ex1$	$Ex2$	$Ex3$	$Ex4$	$Ex5$	$Ex6$	$Ex7$	$Ex8$	$Ex9$	$Ex10$	$Ex12$
Exponential	$\rho = \hat{\rho}$	1.007	0.691	0.435	1.005	0.784	0.263	0.965	0.668	0.651	0.202	0.141
Rolling	$H = \hat{H}$	1.055	0.698	0.444	1.059	0.791	0.271	0.976	0.670	0.658	0.241	0.213
Rolling	$H = 20$	1.040	0.664	0.412	1.045	0.759	0.269	0.959	0.658	0.644	0.307	0.289
	$H = 30$	1.026	0.662	0.419	1.028	0.764	0.317	0.948	0.681	0.673	0.385	0.372
Exponential	$\rho = 0.99$	1.012	0.747	0.571	1.013	0.846	0.563	0.960	0.768	0.779	0.594	0.566
	$\rho = 0.95$	1.088	0.691	0.428	1.088	0.787	0.272	0.995	0.662	0.642	0.258	0.227
	$\rho = 0.90$	1.212	0.765	0.468	1.212	0.864	0.269	1.102	0.705	0.675	0.203	0.158
	$\rho = 0.80$	1.490	0.940	0.573	1.493	1.059	0.321	1.353	0.849	0.809	0.203	0.136
	$\rho = 0.70$	1.903	1.201	0.732	1.911	1.354	0.409	1.728	1.077	1.024	0.238	0.146
	$\rho = 0.50$	3.326	2.102	1.284	3.358	2.373	0.713	3.021	1.873	1.776	0.387	0.217
Averaging		1.007	0.756	0.587	1.008	0.852	0.589	0.961	0.783	0.807	0.645	0.615
Nonparametric		1.074	0.694	0.426	1.078	0.787	0.249	1.000	0.658	0.635	0.222	0.193
Polynomial	$\alpha = \hat{\alpha}$	1.001	0.801	0.547	1.001	0.873	0.353	0.980	0.730	0.718	0.293	0.251
Rolling	$H = \hat{H}$, $k = \hat{k}$	1.052	0.739	0.464	1.055	0.832	0.277	1.004	0.686	0.663	0.217	0.173
Dynamic Weighting		0.826	0.399	0.251	0.787	0.448	0.158	0.615	0.430	0.424	0.162	0.131
Exponential AR		0.581	0.399	0.252	0.563	0.449	0.161	0.575	0.438	0.433	0.162	0.122
Exponential Residual		0.573	0.434	0.393	0.557	0.482	0.368	0.559	0.541	0.559	0.298	0.194
Nonparametric Residual		0.585	0.419	0.381	0.565	0.465	0.360	0.543	0.534	0.549	0.292	0.192
Last Observation		3.340	2.111	1.289	3.372	2.383	0.716	3.033	1.881	1.784	0.389	0.218
AR		0.527	0.790	0.737	0.506	0.745	0.522	0.634	0.786	0.797	0.336	0.200

data-dependent and fixed discounting rates, the one minimising the criterion function $Q_{T,H}$ of (2.5), and respectively, the forecast MSE, $\omega_{T,\hat{H}} \sim Q_{T,\hat{H}}$. This possibility illustrates the wide relevance of our cross-validation approach.

In summary, the results suggest that robust forecasting methods with data-selected parametric downweighting are effective in the face of a variety of types of structural change, and in some cases prevent significant errors. For models with i.i.d. noise, nonparametric methods can be very effective. Further, exponential AR and residual methods seem to provide a very effective way to forecast under structural change in the presence of substantial short-run dynamics, and are likely to be preferable to the simpler methods that do not allow for short-run dynamics. It remains to be seen in the next section whether our proposed methods are effective in practice.

4 Empirical illustration

In this section we examine how our methods would have fared when applied to a wide range of US quarterly data series.[3] We are not trying to establish the best methods for particular data series, but instead to get an impression of whether the issues identified above are important in practice. Although not required with our methodology, so as not to disadvantage the simple location and an AR(1) benchmarks in all cases we transform series to stationarity. We use data on 97 US series for the US, taken from Eklund, Kapetanios, and Price (2010). The dataset includes real activity, prices and financial variables among others. Appendix C of Eklund, Kapetanios, and Price (2010) lists the series. The data span 1960Q1 to 2008Q3. We evaluate one step ahead forecasts over a long period starting in 1992Q2 and ending in 2008Q3. For each series, we compare MSEs to those from the full sample benchmark simple location and AR(1) models.[4]

[3]We take no account of real-time data revisions.

[4]The simple location model benchmark is the baseline model in our exposition and can perform well as a parsimonious forecasting strategy. The AR(1) is a standard forecast benchmark, and often the first lag is the critical one in AR forecasting. Elliott and Timmermann (2008) note that it is difficult to outperform simple approaches,

Table 4: Empirical relative mean squared error results for the US data with a full sample unconditional mean and AR(1) forecast as benchmarks.

		Full Sample Unconditional Mean Benchmark						Full Sample AR(1) Benchmark					
Method		Mean	Median	Min	Max	DM1	DM2	Mean	Median	Min	Max	DM1	DM2
Exponential	$\rho = \hat{\rho}$	0.614	0.652	0.024	1.288	1	48	0.883	0.979	0.073	1.402	4	22
Rolling	$H = \hat{H}$	0.779	0.923	0.110	1.163	2	34	1.947	1.052	0.298	11.811	32	16
Rolling	$H = 20$	0.876	0.979	0.185	1.764	11	32	2.725	1.087	0.393	24.211	37	10
	$H = 30$	0.862	0.987	0.222	1.610	4	29	2.812	1.050	0.373	25.043	36	8
Exponential	$\rho = 0.99$	0.908	0.957	0.462	1.211	3	38	4.218	1.062	0.718	35.741	33	6
	$\rho = 0.95$	0.744	0.840	0.164	1.150	2	45	2.000	1.015	0.353	13.232	32	19
	$\rho = 0.90$	0.680	0.728	0.086	1.197	4	50	1.473	1.034	0.291	7.102	31	22
	$\rho = 0.80$	0.653	0.621	0.057	1.435	4	52	1.126	1.049	0.194	3.244	24	23
	$\rho = 0.70$	0.686	0.598	0.039	1.772	8	49	1.029	1.078	0.125	2.068	22	21
	$\rho = 0.50$	0.876	0.638	0.024	2.774	21	47	1.148	1.026	0.073	3.237	27	14
Averaging		0.954	0.997	0.505	1.583	9	29	4.920	1.100	0.746	46.727	42	5
Nonparametric		0.808	0.904	0.111	1.738	7	31	2.084	1.059	0.337	13.355	36	12
Polynomial	$\alpha = \hat{\alpha}$	0.667	0.682	0.055	1.207	1	48	1.805	1.020	0.113	16.189	16	21
Rolling	$H = \hat{H}$, $k = \hat{k}$	0.706	0.792	0.089	1.135	2	50	1.545	1.045	0.264	8.071	30	24
Dynamic Weighting		0.624	0.616	0.023	1.401	3	53	0.912	1.018	0.074	2.213	3	21
Exponential AR		0.646	0.653	0.013	1.543	0	47	0.913	0.987	0.062	2.436	2	21
Exponential Residual		0.620	0.625	0.013	1.479	2	49	0.891	1.012	0.067	1.504	3	19
Nonparametric Residual		0.616	0.606	0.030	1.255	2	51	0.904	1.002	0.087	1.332	5	17
AR		0.709	0.899	0.024	1.164	3	52	-	-	-	-	-	-

The robust methods we report are those in the Monte Carlo study, and include rolling window forecasts, averaging across estimation periods, exponentially weighted moving average forecasts, polynomially weighted moving average forecasts and forecasts produced using nonparametric weights and residuals.

Table 4 contains the results. We report the median and mean MSE_{RR} relative to the full sample mean (equal weight) benchmarks. We also include the minimum and maximum MSE_{RR}. DM1 and DM2 report the number of significant Diebold-Mariano tests where the null is equality of the downweighting method and the benchmark. The alternative for DM1 is that the benchmark is the better forecast, and for DM2 that the downweighting method is superior. As in most cases for one of the two comparator models a form of rolling estimation is involved, the use of this test is valid (Giacomini and White (2005)).

In almost all cases, the data-dependent downweighting methods beat the sample mean benchmark. The median reduction in the optimised EWMA downweighting forecast is large, reaching over 30%. But this simple benchmark will not usually be applied in practice, as typically forecasts accounting for some dynamics are employed. Thus of much more interest is the more challenging AR benchmark.

The median statistics with respect to the AR model are typically greater than one, showing that the proposed methods fail to outperform a full sample AR. The natural interpretation of this is that only a minority of series suffer from structural change. Notwithstanding this, we note that the optimised exponential and the exponential AR beat the benchmark at the median, and elsewhere the forecast performance penalty at the median is small. This is particularly true for the dynamic methods (dynamic weighting and residual methods). The implication is that in this sample our methods are safe to use, in the sense that typically they will be, at worst, only slightly inferior to a full sample AR benchmark.

such as a parsimonious autoregressive model, that tend to generate relatively smooth and stable forecasts, without being subject to too much parameter estimation error. We also investigated an AR(p) benchmark, where p is chosen by the Bayesian information criterion, but found that both median and average forecast MSE, over all the series we consider, were higher compared to the AR(1) model.

In many cases, a relatively high fixed discount rate in the EWMA does well, although not on average beating the AR benchmark. But in some cases they do poorly. Note that even where the number of significant DM2 tests is high (favouring the downweighting method) the number of significant DM1 cases is invariably higher (see e.g. EWMA for $\rho = 0.90$ and 0.95). The point, of course, is that one fixed weight is unlikely to be right for all series. The data-dependent and fixed window rolling method also does poorly, as does the averaging method. Neither are the nonparametric and optimised polynomial methods particularly successful. But by contrast, in general the optimised EWMA downweighting and dynamic models (the dynamically weighted and residual based methods) do well, and we concentrate our discussion on these.

The mean MSE_{RR} of the optimised EWMA and dynamic methods is uniformly below the median, indicating that there is a predominance of well performing models and that sometimes, when structural change occurs, there are very large benefits to be had from the use of our proposed methods relative to an AR benchmark. The mean reduction in MSE is large enough to be practically important. In the best cases, for the dynamic models the improvement is sensational, with MSEs of less than 0.07. The exponential method is also an outstanding performer. In the worst cases, the optimised EWMA and the dynamic methods have MSEs between 1 and 1.5. While large, these are generally much lower than for the non-optimised (fixed tuning parameter) methods. These impressions are confirmed by formal tests. For the data-dependent exponential and the dynamic models in between 18 and 23% of cases the proposed method is significantly better than the AR benchmark, with less than 5% of cases where the benchmark is significantly better than the proposed robust method. This is strong evidence in support of the practical utility of data-dependent downweighting dynamic models.

Table 5 reports the series where the outperformance is most pronounced. It includes the 20 series with the smallest MSEs compared to the AR(1) for optimised *Exponential* and *Exponential AR* methods. There are very large improvements relative to the benchmark for all of these series, which are never unimportant practically and in some cases dramatic. The methods are particularly useful for forecasting spreads and inflation series. This is further strong evidence supporting the use of the optimised EWMA and dynamic methods.

5 Conclusions

Forecast methods that are known to be robust to historical structural change have been recently found to be useful forecasting tools under ongoing structural change. They include rolling regressions, forecast averaging over different windows and exponentially weighted moving averages. However, the typical practice of setting *a priori* the degree of downweighting older data is suboptimal by its nature. The alternative approach suggested here indicates that, although we do not know the structure of the model and the nature of structural change, we can make the choice of the tuning parameter data-dependent and select it by cross-validation using in-sample forecast performance. Such discounting has a number of attractive properties. It minimises asymptotic forecast MSE over the class of parametrically weighted moving average forecasts. Rather remarkably, it allows also the evaluation of the forecast error, and provides a framework for a number of new developments for forecasting under ongoing structural change. Both theory and small sample evidence suggest that exponential weighting may be most helpful and efficient, and that data selected tuning can provide a useful framework for avoiding large forecast errors. An especially useful finding is that our methods coupled with simple dynamic modelling, such as a low-order

Table 5: 20 series with the smallest relative MSE_{RR} for optimised *Exponential* and *Exponential AR* forecast and AR(1) forecast benchmark

Optimised *Exponential*		*Exponential AR*	
Spread AAA-FF	0.073	Spread AAA-FF	0.067
Spread BAA-FF	0.086	Spread BAA-FF	0.075
Spread 10Y-FF	0.119	Non-Borrowed Reserves of Depository Institutions	0.096
Non-Borrowed Reserves of Depository Institutions	0.155	Spread 10Y-FF	0.124
Spread 3M-FF	0.176	Spread 3M-FF	0.191
Spread 5Y-FF	0.198	Spread 5Y-FF	0.201
CPI-U: Medical care	0.212	Spread 6M-FF	0.253
Spread 6M-FF	0.243	CPI-U: Medical care	0.256
CPI-U: Durables	0.296	Commercial & Industrial Loans Outstanding	0.291
Commercial & Industrial Loans Outstanding	0.319	CPI-U: Durables	0.309
Manufacturing: average hourly earnings of production workers	0.399	Natural resources and mining employment	0.418
Natural resources and mining employment	0.422	Spread 1Y-FF	0.423
Spread 1Y-FF	0.429	Effective Federal Funds Rate	0.479
Construction: average hourly earnings of production workers	0.456	Construction: average hourly earnings of production workers	0.537
Consumer Credit Outstanding - Nonrevolving	0.540	Consumer Credit Outstanding - Nonrevolving	0.539
Consumer Price Index For All Urban Consumers: All Items	0.608	Manufacturing: average hourly earnings of production workers	0.560
CPI-U: Apparel	0.610	3-Month Treasury Bill: Secondary Market Rate	0.591
M2 Money Stock	0.614	M2 Money Stock	0.603
CPI-U: All Items Less Medical Care	0.637	Money Supply - M2	0.609
Money Supply - M2	0.647	M1 Money Stock	0.670
USA Prime Rate	0.672	PPI: Intermediate Mat. Supplies & Components	0.679
CPI-U: All Items Less Food	0.689	6-Month Treasury Bill: Secondary Market Rate	0.713
PPI: Intermediate Mat. Supplies & Components	0.719	CPI-U: Apparel	0.716
M1 Money Stock	0.726	USA Prime Rate	0.740
Nondurable goods manufacturing employment	0.746	Consumer Price Index For All Urban Consumers: All Items	0.784

autoregressive structure, can provide great improvements over standard forecasting methods in the presence of structural change, while having small costs in its absence.

The simulation study and the empirical exercise using a large number of US macroeconomic series show that fixed discount EWMA weighting, with a low discount rate, is often good, but is outperformed consistently by the data selected downweighting. Not all series exhibit breaks, but in many cases forecast performance is enhanced substantially and significantly relative to a full sample AR forecast, without a large penalty in other cases. Overall, we find strong support for our approach, motivated by the impossibility of knowing the optimal degree of discounting *ex ante*.

A Appendix: Proofs

A.1 Proof of Theorems 1-3 and Corollaries 1-4

In this section we establish the claims of Theorems 1-3 about $Q_{T,H}$ and $\omega_{T,H}$ for $y_t = \beta_t + u_t$, with β_t following models (b1) to (b6). The proof of Theorems 1-3 follows the main steps outlined below. Write

$$\begin{aligned} Q_{T,H} &= T_n^{-1}\sum_{t=T_0}^{T}(y_t - \hat{y}_{t|t-1,\,H})^2 = T_n^{-1}\sum_{t=T_0}^{T}\Big(\sum_{j=1}^{t-1} w_{tj,H}(y_t - y_{t-j})\Big)^2, \\ \omega_{T,H} &= E(y_{T+1} - \hat{y}_{T+1|T,H})^2 = E\big(\sum_{j=1}^{T} w_{T+1,j,H}(y_{T+1} - y_{T+1-j})\Big)^2. \end{aligned}$$

We will approximate $Q_{T,H}$ and $\omega_{T,H}$ by

$$Q_{T,H}^{(apr)} = T_n^{-1}\sum_{t=T_0}^{T}\Big(\sum_{j=1}^{T_1} w_{j,H}(y_t - y_{t-j})\Big)^2, \quad \omega_{T,H}^{(apr)} = E\big(\sum_{j=1}^{T} w_{j,H}(y_{T+1} - y_{T+1-j})\big)^2,$$

respectively, where $w_{j,H}$'s are as in (2.8), setting $T_1 = T_0 T^{-\delta/2}$. Recall that $H_{max} = T_0 T^{-\delta}$. Thus, $T_0/H_{max} = T^{\delta}$, $T_1/H_{max} = T^{\delta/2}$ and $T_1/T \leq T^{-\delta/2}$. Lemma A.1 below implies that uniformly in H:

$$Q_{T,H} = Q_{T,H}^{(apr)} + O_H(T^{-2}), \qquad \omega_{T,H} = \omega_{T,H}^{(apr)} + O_H(T^{-2}). \tag{A.1}$$

The proof of Theorems 1-3 is based on (A.1) and the following properties of $Q_{T,H}^{(apr)}$ and $\omega_{T,H}^{(apr)}$. Let $\tilde{Q}_{T,H}^{(apr)} := Q_{T,H}^{(apr)} - \hat{\sigma}_{T,u}^2$. For each of the cases (bi), $i = 1, \cdots, 6$, we will find deterministic approximating functions $\Gamma_{T,H}^{(i)}$, $\tilde{\Gamma}_{T,H}^{(i)}$ and a rate function $r_{T,H}^{(i)}$, such that as $T \to \infty$, uniformly in $H \in I_T$,

$$\text{(i)} \quad \big|E\tilde{Q}_{T,H}^{(apr)} - \Gamma_{T,H}^{(i)}\big| = o_H(r_{T,H}^{(i)}), \qquad \text{(ii)} \quad E\big|\tilde{Q}_{T,H}^{(apr)} - E\tilde{Q}_{T,H}^{(apr)}\big| = o_H(r_{T,H}^{(i)}), \tag{A.2}$$

$$\big|\omega_{T,H}^{(apr)} - \tilde{\Gamma}_{T,H}^{(i)} - \sigma_u^2\big| = o_H(r_{T,H}^{(i)}). \tag{A.3}$$

Then, (A.1)-(A.3) imply

$$\begin{aligned} Q_{T,H} &= \hat{\sigma}_{T,u}^2 + \Gamma_{T,H}^{(i)} + (E\tilde{Q}_{T,H}^{(apr)} - \Gamma_{T,H}^{(i)}) + (\tilde{Q}_{T,H}^{(apr)} - E\tilde{Q}_{T,H}^{(apr)}) + O_H(T^{-2}) \\ &= \hat{\sigma}_{T,u}^2 + \Gamma_{T,H}^{(i)} + o_H(r_{T,H}^{(i)}) + O_H(T^{-2}), \\ \omega_{T,H} &= \sigma_u^2 + \tilde{\Gamma}_{T,H}^{(i)} + o_H(r_{T,H}^{(i)}). \end{aligned} \tag{A.4}$$

Functions $\Gamma_{T,H}^{(i)}$ and $r_{T,H}^{(i)}$ $i = 1, \cdots, 6$, are as follows.

$$\begin{array}{lll} \Gamma_{T,H}^{(1)} = q_{u,H}, & r_{T,H}^{(1)} = H^{-1}, & \lambda_{T,H}^{(1)} = \lambda_u H^{-1}; \\ \Gamma_{T,H}^{(2)} = q_{\beta,H}^{(2)} + q_{u,H}, & r_{T,H}^{(2)} = H, & \lambda_{T,H}^{(2)} = \lambda_\beta^{(2)} H; \\ \Gamma_{T,H}^{(3)} = T^{-1}q_{\beta,H}^{(3)} + q_{u,H}, & r_{T,H}^{(3)} = HT^{-1} + H^{-1}, & \lambda_{T,H}^{(3)} = \lambda_\beta^{(3)} HT^{-1} + \lambda_u H^{-1}; \\ \Gamma_{T,H}^{(4)} = q_{\beta,H}^{(4)} + q_{u,H}, & r_{T,H}^{(4)} = H^2, & \lambda_{T,H}^{(4)} = \lambda_\beta^{(4)} H^2; \\ \Gamma_{T,H}^{(5)} = T^{-2}q_{\beta,H}^{(5)} + q_{u,H}, & r_{T,H}^{(5)} = (H/T)^2 + H^{-1}, & \lambda_{T,H}^{(5)} = \lambda_\beta^{(5)}(H/T)^2 + \lambda_u H^{-1}; \\ \Gamma_{T,H}^{(6)} = q_{\beta,TH}^{(6)} + q_{u,H}, & r_{T,H}^{(6)} = HT^{-1} + H^{-1}, & \lambda_{T,H}^{(6)} = G_{\tau,H} HT^{-1} + \lambda_u H^{-1}. \end{array}$$

We define $\tilde{\Gamma}_{T,H}^{(i)} = \Gamma_{T,H}^{(i)}$, $i = 1, 2, 3$, $\tilde{\Gamma}_{T,H}^{(4)} = \delta_g q_{\beta,H}^{(4)} + q_{u,H}$, $\tilde{\Gamma}_{T,H}^{(5)} = T^{-2}\delta_g' q_{\beta,H}^{(5)} + q_{u,H}$ and $\tilde{\Gamma}_{T,H}^{(6)} = \tilde{\lambda}_{T,H}^{(6)}$, where δ_g, δ_g' and $\lambda_{T,H}^{(6)}$ are the same as in Theorem 3 (i-iii). We will use the functions $\lambda_{T,H}^{(i)}$ in Lemma A.2(ii) to describe the asymptotics

$$\Gamma_{T,H}^{(i)} = \lambda_{T,H}^{(i)} + o_H(r_{T,H}^{(i)}), \quad H \to \infty, \quad i = 1, \cdots, 6.$$

Observe that $T^{-2} = o(r_{T,H}^{(i)})$, uniformly in $H \in I_T$, $i = 1, \cdots, T$.
Hence, the proof of Theorems 1-3, in the case (bi), $i = 1, \cdots, 6$ reduces to the verification of approximations (A.2)–(A.3) with corresponding $\Gamma_{T,H}^{(i)}$, $\tilde{\Gamma}_{T,H}^{(i)}$ and $r_{T,H}^{(i)}$, obtained in Lemmas A.2 and A.3.

Proof of Theorem 1. In case (b1), Lemmas A.1-A.3 imply (A.1)–(A.3), which in turn imply (A.4):

$$Q_{T,H} = \hat{\sigma}^2_{T,u} + q_{u,H} + o_H(H^{-1}), \quad \omega_{T,H} = \sigma^2_u + q_{u,H} + o_H(H^{-1}).$$

In addition, in Lemma A.2(ii) it is shown that $q_{u,H} = \lambda_u H^{-1} + o(H^{-1})$, as $H \to \infty$. This proves that $Q_{T,H}$ and $\omega_{T,H}$ have properties (2.12). □

Proof of Theorem 2. In cases (b2) and (b3), use Lemmas A.1-A.3 to verify (A.1)–(A.3), which imply (A.4) that reads

$$Q_{T,H} = \hat{\sigma}^2_{T,u} + q^{(2)}_{\beta,H} + q_{u,H} + o_H(H), \quad \omega_{T,H} = \sigma^2_u + q^{(2)}_{\beta,H} + q_{u,H} + o_H(H), \quad \text{in (b2)},$$
$$Q_{T,H} = \hat{\sigma}^2_{T,u} + T^{-1}q^{(3)}_{\beta,H} + q_{u,H} + o_H(r^{(3)}_{T,H}), \quad \omega_{T,H} = \sigma^2_u + T^{-1}q^{(3)}_{\beta,H} + q_{u,H} + o_H(r^{(3)}_{T,H}), \quad \text{in (b3)},$$

where $r^{(3)}_{T,H} = HT^{-1} + H^{-1}$. In addition, by Lemma A.2(ii), as $H \to \infty$, $q^{(2)}_{\beta,H} + q_{u,H} = \lambda^{(2)}_\beta H(1+o(1))$ and $T^{-1}q^{(3)}_{\beta,H} + q_{u,H} = \{\lambda^{(3)}_\beta HT^{-1} + \lambda_u H^{-1}\}(1 + o(1))$, which proves that $Q_{T,H}$ and $\omega_{T,H}$ have properties (2.17) and (2.18), respectively. □

Proof of Theorem 3. In cases (b4) and (b5), (A.1)–(A.3) of Lemmas A.1-A.3 imply (A.4):

$$Q_{T,H} = \hat{\sigma}^2_{T,u} + q^{(4)}_{\beta,H} + q_{u,H} + o_H(H^2), \quad \omega_{T,H} = \sigma^2_u + \delta_g q^{(4)}_{\beta,H} + q_{u,H} + o_H(H^2), \quad \text{in (b4)},$$
$$Q_{T,H} = \hat{\sigma}^2_{T,u} + T^{-2}q^{(5)}_{\beta,H} + q_{u,H} + o_H(r^{(5)}_{T,H}), \quad \omega_{T,H} = \sigma^2_u + \delta'_g T^{-1}q^{(5)}_{\beta,H} + q_{u,H} + o_H(r^{(5)}_{T,H}), \quad \text{in (b5)},$$

where $r^{(5)}_{T,H} = (H/T)^2 + H^{-1}$. In addition, by Lemma A.2(ii), as $H \to \infty$, $q^{(4)}_{\beta,H} + q_{u,H} = \lambda^{(4)}_\beta H^2(1 + o(1))$ and $T^{-2}q^{(5)}_{\beta,H} + q_{u,H} = \{\lambda^{(5)}_\beta (H/T)^2 + \lambda_u H^{-1}\}(1+o(1))$, and by (A.33), $q_{u,H} = \lambda_u H^{-1}(1+o(1))$. This proves that $Q_{T,H}$ and $\omega_{T,H}$ satisfy (2.24), (2.25), respectively.

Finally, in case (b6), the results (A.1)–(A.3) of Lemmas A.1-A.3 lead to (A.4):

$Q_{T,H} = \hat{\sigma}^2_{T,u} + q^{(6)}_{\beta,TH} + q_{u,H} + o_H(HT^{-1} + H^{-1})$, $\omega_{T,H} = \sigma^2_u + \tilde{\lambda}^{(6)}_{T,H} + o_H(HT^{-1} + H^{-1})$.

By Lemma A.2(ii), as $H \to \infty$, $q^{(6)}_{\beta,TH} + q_{u,H} = \{G_{\tau,H}HT^{-1} + \lambda_u H^{-1}\}(1 + o(1))$, which completes the proof of (2.26). □

Proof of Corollary 1. Suppose that $q_{u,H}$ reaches its minimum $c_0 = q_{u,H_0}$ at some finite H_0. Then (2.12) implies that $Q_{T,\hat{H}} = c_0 + o(1)$, $\omega_{T,H_{opt}} = c_0 + o(1)$, which in turn implies $\omega_{T,\hat{H}} = Q_{T,\hat{H}} + o(1) = c_0 + o(1)$. Hence, $\omega_{T,\hat{H}} = \omega_{T,H_{opt}} + o(1)$ and $Q_{T,\hat{H}} = \omega_{T,\hat{H}} + o(1)$. This proves (2.13).

If $q_{u,H}$ reaches its minimum at infinity, then by (2.12) $\hat{H} \sim H_{max}$. Recall that by definition H_{max} is of larger order than $T^{1/2}$. Then (2.12) implies that $Q_{T,\hat{H}} = \hat{\sigma}^2_u + O_p(H^{-1}_{max}) = \sigma^2_u + O_p(T^{-1/2})$, noting that $\hat{\sigma}^2_u = \sigma^2_u + O_p(T^{-1/2})$ by (2.9). Similarly, $\omega_{T,H_{opt}} = \sigma^2_u + O(H^{-1}_{opt}) = \sigma^2_u + o(T^{-1/2})$. Hence, $\omega_{T,\hat{H}} = Q_{T,\hat{H}} + O_p(\hat{H}^{-1}) = \sigma^2_u + O_p(T^{-1/2})$, showing that $\omega_{T,\hat{H}} = \omega_{T,H_{opt}} + O(T^{-1/2})$ and $Q_{T,\hat{H}} = \omega_{T,\hat{H}} + O_p(T^{-1/2})$, which verifies (2.14). □

Proof of Corollary 2. For β_t as in (b2), property (2.17) of $Q_{T,H}$ shows that $\hat{H}$ stays bounded. Then (2.19) follows by the same argument as in Corollary 1.

For β_t as in (b3), if $q_{u,H}$ reaches its minimum $c_0 = q_{u,H_0}$ at some finite H_0, then (2.18) implies that $\hat{H}$ stays bounded, and $Q_{T,\hat{H}} = c_0 + o(1)$, $\omega_{T,H_{opt}} = c_0 + o(1)$ and $\omega_{T,\hat{H}} = Q_{T,\hat{H}} + o(1) = c_0 + o(1)$. This yields (2.20).

If $q_{u,H}$ reaches its minimum at infinity, then the relation $T^{-1}q^{(3)}_{\beta,H} + q_{u,H} \sim \lambda^{(3)}_\beta HT^{-1} + \lambda_u H^{-1}$ derived in Theorem 2(ii), implies that $\hat{H} \sim \text{argmin}_H\big(\lambda^{(3)}_\beta HT^{-1} + \lambda_u H^{-1}\big) \sim (\lambda_u/\lambda^{(3)}_\beta)^{1/2}T^{1/2}$. Then

(2.18) implies that $Q_{T,\hat{H}} = \hat{\sigma}_u^2 + O_p(\hat{H}^{-1}) = \sigma_u^2 + O_p(T^{-1/2})$, and similarly, $\omega_{T,H_{opt}} = \sigma_u^2 + O(T^{-1/2})$. Hence, $\omega_{T,\hat{H}} = Q_{T,\hat{H}} + O_p(\hat{H}^{-1}) = \sigma_u^2 + O_p(T^{-1/2})$, showing that $\omega_{T,\hat{H}} = \omega_{T,H_{opt}} + O(T^{-1/2})$ and $Q_{T,\hat{H}} = \omega_{T,\hat{H}} + O_p(T^{-1/2})$, which proves the last two claims of Corollary 2. □

Proof of Corollary 3. Result (2.24) of Theorem 3(i) shows that $q^{(4)}_{\beta,H} + q_{u,H} \sim \lambda^{(4)}_\beta H^2$ as $H \to \infty$, which implies that $\hat{H}$ stays bounded. For a linear trend $\beta_t = ct$, we have $\delta_g = 1$. Then, the approximations of $Q_{T,H}$ and $\omega_{T,H}$ in (2.24) coincide, and (2.27) follows using a similar argument as in the proof of Corollary 1.

For β_t as in (b5), if $q_{u,H}$ reaches its minimum $c_0 = q_{u,H_0}$ at some finite H_0, then (2.25) implies that $\hat{H}$ stays bounded, and $Q_{T,\hat{H}} = c_0 + o(1)$, $\omega_{T,H_{opt}} = c_0 + o(1)$ and $\omega_{T,\hat{H}} = Q_{T,\hat{H}} + o(1) = c_0 + o(1)$. This verifies (2.28).

If $q_{u,H}$ reaches its minimum at infinity, then, the relation $(H/T)^2 q^{(5)}_{\beta,H} + q_{u,h} \sim \lambda^{(5)}_\beta (H/T)^2 + \lambda_u H^{-1}$, derived in Theorem 3(ii), implies that $\hat{H} \sim \text{argmin}_H \big(\lambda^{(5)}_\beta (H/T)^2 + \lambda_u H^{-1}\big) \sim cT^{2/3}$, $c > 0$. Then by (2.25), $Q_{T,\hat{H}} = \hat{\sigma}_u^2 + O_p(T^{-2/3}) = \sigma_u^2 + O_p(T^{-1/2})$, and similarly, $\omega_{T,H_{opt}} = \sigma_u^2 + O(T^{-2/3})$. Hence, $\omega_{T,\hat{H}} = Q_{T,\hat{H}} + O_p(T^{-1/2}) = \sigma_u^2 + O_p(T^{-1/2})$, $\omega_{T,\hat{H}} = \omega_{T,H_{opt}} + O(T^{-1/2})$ and $Q_{T,\hat{H}} = \omega_{T,\hat{H}} + O_p(T^{-1/2})$. This implies the last two claims of Corollary 3. □

Proof of Corollary 4. (i) Suppose that $\tau/\sqrt{T} \to \infty$. To show (2.29), set $\tau' = (T\tau)^{1/4}$ and split the minimization interval I_T into $I'_T = [\alpha, \tau']$ and $I''_T = [\tau', H_{max}]$. Notice that $\tau' = o(\tau)$ and $\tau'/\sqrt{T} \to \infty$. By Theorem 3(iii), minimization of $Q_{T,H}$ in $H \in I'_T$ reduces to minimization of $G_{\tau,H} HT^{-1} + \lambda_u H^{-1}$. Since for $H \le \tau'$, $\tau/H \to \infty$ and $G_{\tau,H} \to \lambda_\beta$, it further reduces to minimization of $\lambda_\beta HT^{-1} + \lambda_u H^{-1}$, which implies (2.29) for the minimizer $\hat{H}$ in I'_T. Then, $\inf_{H \in I'_T} \big(G_{\tau,H} HT^{-1} + \lambda_u H^{-1}\big) \sim cT^{-1/2}$ with $c > 0$. Next, we show that the minimum in I_T is reached in I'_T, because

$$\inf_{H \in I''_T} \big(G_{\tau,H} HT^{-1} + \lambda_u H^{-1}\big) >> T^{-1/2}.$$

To verify the latter, select $\epsilon > 0$ such that $\int_\epsilon^\infty K(v)dv \ge 1/2$. Then, for $H \ge \tau'$, $G_{\tau,H} \ge G_{\tau',H} \ge G_{\tau'\epsilon,H} \ge \Delta^2(\tau'\epsilon/H)/4$. Hence, $\sqrt{T}(G_{\tau,H} HT^{-1}) \ge (\tau'/\sqrt{T})\Delta^2\epsilon/4 \to \infty$ which proves (2.29).

To complete the proof of (i), we need to evaluate $\omega_{T,\hat{H}}$ given by (2.26). By Assumption 1 $K(x) \le C\exp(-c|x|)$, and therefore $\big(\sum_{j=T+1-t_0}^T w_{j,H}\big)^2 \le C(H^{-1}\int_\tau^\infty K(x/H)dx)^2 \le Ce^{-2c\tau/H}$. Since for i.i.d. noise $q_{u,H} \sim \sigma_u^2 H^{-1}$, applying this in (2.26) gives $\omega_{T,\hat{H}} = \sigma_u^2 + O(T^{-1/2} + e^{-2c\tau/\sqrt{T}})$. Obviously, $\omega_{T,H_{opt}}$ satisfies the same relation which yields the last two claims in (i).

(ii) Now, let $\tau/\sqrt{T} = o(1)$. Set $\tau^* = \tau\log(\sqrt{T}/\tau)$ and write $I_T = [\alpha, \tau^*] \cup [\tau^*, H_{max}] =: I^*_T \cup I^{**}_T$. Note that $\tau = o(\tau^*)$. By (2.26), $\inf_{H \in I^*_T} \big(q^{(6)}_{\beta,TH} + q_{u,H}\big) \ge \inf_{H \in I^*_T} q_{u,H} = \inf_{H \in I^*_T} \lambda_u H^{-1}(1 + o(1)) = \lambda_u \tau^{*-1}(1 + o(1))$. Next we show that the minimum in I^{**}_T is of smaller order than in I^*_T. Let $H \ge \tau^*$. Then $\tau/H \to 0$, which combining with $\int_x^\infty K(v)dv = 1 - \int_0^x K(v)dv = 1 + O(x)$, gives $G_{\tau,H} = \Delta^2 \int_0^{\tau/H}(1 + O(x))^2 dx = \Delta^2\tau H^{-1} + O((\tau/H)^2)$. In turn, this implies $G_{\tau,H} HT^{-1} = \Delta^2\tau T^{-1} + o_H(H^{-1})$ because $(\tau/H)^2 HT^{-1} = H^{-1}\tau^2 T^{-1} = o(H^{-1})$. Thus, $\inf_{H \in I^{**}_T} \big(q^{(6)}_{\beta,TH} + q_{u,H}\big) = \inf_{H \in I^{**}_T} \big(\Delta^2\tau T^{-1} + \lambda_u H^{-1} + o_H(H^{-1})\big) = \Delta^2\tau T^{-1} + O_p(H^{-1}_{max}) = o(\tau^{*-1})$, noting that $\tau T^{-1} = \tau^{*-1}\big(\tau^2 T^{-1}\log(\sqrt{T}/\tau)\big) = o(\tau^{*-1})$ when $\tau = o(\sqrt{T})$, while $H^{-1}_{max} = o(\tau^{*-1})$. This shows that the minimum in I^{**}_T is reached by largest possible H and is smaller then the minimum in I^*_T. Hence $\hat{H}$ is not affected by the break, $\hat{H}/\sqrt{T} \to \infty$, and $Q_{T,\hat{H}} = \hat{\sigma}^2_{T,u} + \Delta^2\tau T^{-1} + O_p(\hat{H}^{-1}) = \hat{\sigma}^2_{T,u} + o_p(T^{-1/2})$.

Since $\tau/\hat{H} \to 0$, (2.26) implies that $\omega_{T,\hat{H}} = \sigma_T^2 + \Delta^2(\int_{\tau/\hat{H}}^\infty K(x)dx)^2 + o(1) = \sigma_T^2 + \Delta^2 + o(1)$.

This completes the proof of the corollary. □

A.2 Lemmas

This section includes three lemmas used to prove Theorems 1-3.

Lemma A.1 *If weights w_{tj} and u_t's satisfy Assumptions 1 and 2, and β_t is as in (b1)-(b6), then*

$$E\big[\sup_{H\in I_T}\big|Q_{T,H}-Q_{T,H}^{(apr)}\big|\big]=O(T^{-2}),\qquad \sup_{H\in I_T}\big|\omega_{T,H}-\omega_{T,H}^{(apr)}\big|=O(T^{-2}). \tag{A.5}$$

Proof. We start with the first claim in (A.5). Recall that $w_{tk,H}\le 1$ and $w_{j,H}\le 1$. Note that for $T_0\le t\le T$ and $j,k\le t-1$, $|w_{tj,H}w_{tk,H}-w_{j,H}w_{k,H}I(j,k\le T_1)|\le |w_{tj,H}w_{tk,H}-w_{j,H}w_{k,H}|+|w_{j,H}w_{k,H}-w_{j,H}w_{k,H}I(j,k\le T_1)|\le |w_{tj,H}-w_{j,H}|+|w_{tk,H}-w_{k,H}|+w_{j,H}I(j\ge T_1)+w_{k,H}I(k\ge T_1)\le CT^{-6}$ by (A.30) and (A.29). Hence,

$$\begin{aligned}|Q_{T,H}-Q_{T,H}^{(apr)}|&\le T_n^{-1}\textstyle\sum_{t=T_0}^{T}\sum_{j,k=1}^{t-1}\big|w_{tj,H}w_{tk,H}-w_{j,H}w_{k,H}I(j,k\le T_1)\big||(y_t-y_{t-j})(y_t-y_{t-k})|\\&\le CT_n^{-1}\textstyle\sum_{t=T_0}^{T}\sum_{j,k=1}^{t-1}T^{-6}|(y_t-y_{t-j})(y_t-y_{t-k})|=:j_T,\end{aligned}$$

where j_T does not depend on H. Notice that β_t of (b1)-(b6) satisfies $\max_{t\le T}E|y_t-y_{t-j}|^2\le CT^2$. Hence $E|(y_t-y_{t-j})(y_t-y_{t-k})|\le CT^2$ and $Ej_T\le CT^{-6}T_n^{-1}\sum_{t=T_0}^{T}\sum_{j,k=1}^{t-1}T^2\le CT^{-2}$ which implies (A.5).

To show the second claim in (A.5), note that in models (b1)-(b6), $E|(y_{T+1}-y_{T+1-j})(y_{T+1}-y_{T+1-k})|\le CT^2$, and $|w_{tj,H}w_{tk,H}-w_{j,H}w_{k,H}|\le |w_{tj,H}-w_{j,H}|w_{tk,H}+w_{j,H}|w_{tk,H}-w_{k,H}|\le |w_{tj,H}-w_{j,H}|+|w_{tk,H}-w_{k,H}|\le CT^{-6}$ by (A.30). Then $|\omega_{T,H}-\omega_{T,H}^{(apr)}|\le CT^2\sum_{j,k=1}^{T}|w_{tj,H}w_{tk,H}-w_{j,H}w_{k,H}|\le CT^2\sum_{j,k=1}^{T}T^{-6}\le CT^{-2}$. □

Below T_1 is the same as in $Q_{T,H}^{(apr)}$.

Lemma A.2 *For β_t as in (bi), $i=1,\cdots,6$, as $T\to\infty$, it holds*

$$\text{(i)}\quad E\tilde{Q}_{T,H}^{(apr)}=\Gamma_{T,H}^{(i)}+o_H(r_{T,H}^{(i)});\qquad \text{(ii)}\quad \Gamma_{T,H}^{(i)}=\lambda_{T,H}^{(i)}+o_H(r_{T,H}^{(i)}),\quad H\to\infty, \tag{A.6}$$

$$\text{(iii)}\quad \omega_{T,H}^{(apr)}=\sigma_u^2+\tilde{\Gamma}_{T,H}^{(i)}+o_H(r_{T,H}^{(i)}). \tag{A.7}$$

Proof. Observe the following general facts. Since β_t and u_t are mutual independent and $Eu_j=0$,

$$E\tilde{Q}_{T,H}^{(apr)}=E[Q_{T,H}^{(apr)}-\hat{\sigma}_{TH}^2]=m_{\beta,TH}+m_{u,TH},\quad \omega_{T,H}^{(apr)}=v_{\beta,TH}+v_{u,TH},$$

where $m_{\beta,TH}:=T_n^{-1}\sum_{t=T_0}^{T}E\Big(\sum_{j=1}^{T_1}w_{j,H}(\beta_t-\beta_{t-j})\Big)^2$, and by stationarity of u_t,

$$\begin{aligned}m_{u,TH}&:=T_n^{-1}\textstyle\sum_{t=T_0}^{T}E\Big(\sum_{j=1}^{T_1}w_{j,H}(u_t-u_{t-j})\Big)^2-\sigma_u^2=E\Big(\sum_{j=1}^{T_1}w_{j,H}(u_0-u_{-j})\Big)^2-\sigma_u^2,\\v_{\beta,TH}&:=E\big(\textstyle\sum_{j=1}^{T}w_{j,H}(\beta_{T+1}-\beta_{T+1-j})\big)^2,\quad v_{u,TH}:=E\big(\textstyle\sum_{j=1}^{T}w_{j,H}(u_{T+1}-u_{T+1-j})\big)^2.\end{aligned}$$

First we analyze $m_{u,TH}$ and $v_{u,TH}$. By definition, $q_{u,H}=E\Big(\sum_{j=1}^{\infty}w_{j,H}(u_0-u_{-j})\Big)^2-\sigma_u^2$. Since $E|(u_0-u_{-j})(u_0-u_{-k})|\le 4Eu_0^2<\infty$, $\sum_{k=1}^{\infty}w_{k,H}=1$ and $\sum_{j=T_1}^{\infty}w_{j,H}=O(T^{-6})$ by (A.29), then

$$|m_{u,TH}-q_{u,H}|\le\textstyle\sum_{j=T_1}^{\infty}w_{j,H}=O(T^{-2}),\quad |v_{u,TH}-q_{u,H}-\sigma_u^2|\le\sum_{j=T}^{\infty}w_{j,H}=O(T^{-2}).$$

Hence, uniformly in $H \in I_T$,

$$E\tilde{Q}^{(apr)}_{T,H} = m_{\beta,TH} + q_{u,H} + O(T^{-2}); \quad \omega^{(apr)}_{T,H} = v_{\beta,TH} + \sigma_u^2 + q_{u,H} + O(T^{-2}), \tag{A.8}$$

$$q_{u,H} = \lambda_u H^{-1} + o(H^{-1}), \quad H \to \infty, \tag{A.9}$$

where (A.9) holds by (A.34).

Case (b1). $\beta_t = const$ implies $m_{\beta,TH} = 0$. Hence, (A.8)-(A.9) imply (A.6) (i, ii) and (A.7).

Case (b2). Let $\beta_t \sim I(1)$. We will show that

$$m_{\beta,TH} = q^{(2)}_{\beta,H} + O(T^{-2}), \qquad v_{\beta,TH} = q^{(2)}_{\beta,H} + O(T^{-2}); \quad H^{-1}q^{(2)}_{\beta,H} \to \lambda^{(2)}_\beta, \ H \to \infty \tag{A.10}$$

which together with (A.8) and (A.9) imply (A.6) and (A.7).

Notice that $\zeta_t := \nabla\beta_t = \beta_t - \beta_{t-1} \sim I(0)$ is a stationary process and $\gamma_{\beta,jk} := E[(\beta_t - \beta_{t-j})(\beta_t - \beta_{t-k})] = E[(\beta_0 - \beta_{-j})(\beta_0 - \beta_{-k})]$, $0 \le j, k \le t$ does not depend on t. Since $\beta_j - \beta_0 = \sum_{l=1}^{j} \zeta_l$, then (see e.g. Proposition 4.4.1 in Giraitis, Koul, and Surgailis (2012)), $\gamma_{\beta,jj} \sim j s^2_{\nabla\beta}$, $j \to \infty$,

$$|\gamma_{\beta,jk}| \le C(jk)^{1/2}; \qquad \gamma_{\beta,jk} = s^2_{\nabla\beta}(j \wedge k) + o(j \wedge k), \quad j, k \to \infty. \tag{A.11}$$

Hence, $m_{\beta,TH} = E\Big(\sum_{j=1}^{T_1} w_{j,H}(\beta_0 - \beta_{-j})\Big)^2$. Since $q^{(2)}_{\beta,H} = E\Big(\sum_{j=1}^{\infty} w_{j,H}(\beta_0 - \beta_{-j})\Big)^2$, then

$$\big|m_{\beta,H} - q^{(2)}_{\beta,H}\big| \le CH \textstyle\sum_{j=T_1}^{\infty} w_{j,H}(j/H)^{1/2} \sum_{k=1}^{\infty} w_{k,H}(k/H)^{1/2} = O(HT^{-6}) = O(T^{-2}),$$

uniformly in $H \in I_T$, by (A.11), (A.29) and (A.33). This proves the first claim of (A.10), while the second follows similarly.

To show the third claim, use (A.11) and (A.33), to obtain $q^{(2)}_{\beta,H} \le C\big(\sum_{j=1}^{\infty} w_{j,H} j^{1/2}\big)^2 \le CH$, and $H^{-1}q^{(2)}_{\beta,H} \to s^2_{\nabla\beta} \int\int_0^\infty K(x)K(y)(x \vee y)dxdy = \lambda^{(2)}_\beta$, $H \to \infty$, which completes the proof of (A.10).

Case (b3). Let $\beta_t = T^{-1/2}\tilde{\beta}_t$ where $\tilde{\beta} \sim I(1)$. Since $m_{\beta,TH} = T^{-1}m_{\tilde{\beta},TH}$, then (A.10) implies $m_{\beta,TH} = T^{-1}q^{(3)}_{\beta,H} + O(T^{-2})$, $v_{\beta,TH} = T^{-1}q^{(3)}_{\beta,H} + O(T^{-2})$ and $H^{-1}q^{(3)}_{\beta,H} \to \lambda^{(3)}_\beta$, $H \to \infty$, which together with (A.8) and (A.9) proves (A.6) and (A.7).

We present Case (b5) first as it provides results for Case (b4).

Case (b5). Let $\beta_t = g(t/T)$. We will verify that

$$m_{\beta,TH} = \frac{1}{T^2}q^{(5)}_{\beta,H} + o_H\big(\frac{H^2}{T^2}\big), \ \ v_{\beta,TH} = \frac{1}{T^2}\delta'_g q^{(5)}_{\beta,H} + o_H\big(\frac{H^2}{T^2}\big); \ \ \frac{1}{H^2}q^{(5)}_{\beta,H} \to \lambda^{(5)}_\beta, \ H \to \infty, \tag{A.12}$$

which together with (A.8) and (A.9) implies (A.6) and (A.7).

We approximate $m_{\beta,TH}$ by $m'_{\beta,TH} = T_n^{-1}\sum_{t=T_0}^{T}\big(\sum_{j=1}^{T_1} w_{j,H}\dot{g}(t/T)(j/T)\big)^2$, showing that as $T \to \infty$, (a) $(T/H)^2|m_{\beta,TH} - m'_{\beta,TH}| = o_H(1)$ and $(T/H)^2|m'_{\beta,TH} - T^{-2}q^{(5)}_{\beta,H}| = o_H(1)$, which proves the first claim in (A.12). To show (a), recall that g has two bounded derivatives. Thus, by Taylor expansion $|g(t/T) - g((t-j)/T) + \dot{g}(t/T)(j/T)| \le C(j/T)^2 \le C(j/T)T^{-\delta/2}$ for $j \le T_1$, since $T_1/T \le T^{-\delta/2}$. Hence, $\beta_t - \beta_{t-j} \equiv g(t/T) - g((t-j)/T) = -\dot{g}(t/T)(j/T) + (j/T)o_H(T^{-\delta/2})$. Since by (A.31), $H^{-2}|w_{j,H}w_{k,H}jk| \le C(jk)^{-1}$, we obtain $(T/H)^2 w_{j,H}w_{k,H}\big|(\beta_t - \beta_{t-j})(\beta_t - \beta_{t-k}) - \dot{g}^2(t/T)(j/T)(k/T)\big| \le CH^{-2}w_{j,H}w_{k,H}(jk)^{-1}T^{-\delta/2} \le C(jk)^{-1}T^{-\delta/2}$. Therefore, $(T/H)^2|m_{\beta,TH} - m'_{\beta,TH}| \le CT^{-\delta/2}\sum_{j,k=1}^{T_1}(jk)^{-1} \le C(\log T)^2 T^{-\delta/2} \to 0$, $T \to \infty$, proving (a).

To show (b), notice that $(T/H)^2|m'_{\beta,TH} - T^{-2}q^{(5)}_{\beta,H}| = \big|T_n^{-1}\sum_{t=T_0}^{T}\dot{g}(t/T)^2\big(\sum_{j=1}^{T_1} w_{j,H}(j/H)\big)^2 - c'(g)\big(\sum_{j=1}^{\infty} w_{j,H}(j/H)\big)^2\big| = o_H(1)$, because $T_n^{-1}\sum_{t=T_0}^{T}\dot{g}^2(t/T) \to \int_0^1 \dot{g}^2(x)dx = c'(g)$, while by

(A.29) and (A.33), $\sum_{j=1}^{T_1} w_{j,H}(j/H) \to \int_0^\infty K(x)xdx$ and $\sum_{j=1}^{\infty} w_{j,H}(j/H) \to \int_0^\infty K(x)xdx$. This proves (b).

The second claim of (A.12) follows using a similar argument. The third claim follows noting that $(\sum_{j=1}^{\infty} w_{j,H}(j/H))^2 \to \kappa_3$, by (A.33).

Case (b4). Let $\beta_t = tg(t/T)$. Applying (A.12) to $\beta_t' = (t/T)g(t/T)$ we obtain, $m_{\beta,TH} = q_{\beta,H}^{(4)} + o_H\big(H^2\big)$, $v_{\beta,TH} = \delta_g q_{\beta,H}^{(4)} + o_H\big(H^2\big)$, and $H^{-2}q_{\beta,H}^{(4)} \to \lambda_\beta^{(4)}$ as $H \to \infty$, which together with (A.8) and (A.9) implies (A.6) and (A.7).

Case (b6). We will verify that

$$m_{\beta,TH} = q_{\beta,H}^{(6)} + o_H\big(T^{-1}\big), \;\; v_{\beta,TH} = \Delta^2\big(\sum_{j=T+1-t_0}^{T} w_{j,H}\big)^2; \;\; TH^{-1}q_{\beta,H}^{(6)} \to G_{\tau,\beta}, \;\; H \to \infty, \quad \text{(A.13)}$$

which together with (A.8) and (A.9) imply (A.6) and (A.7).

The first claim follows from definition of $m_{\beta,TH}$ and $q_{\beta,H}^{(6)}$, observing that for $t_0 \geq T_0$ one has $\beta_t - \beta_{t-j} = 0$ if $t < t_0$ or $j < t - t_0$, and taking into account (A.29). The same argument and definition of $v_{\beta,TH}$ imply the equality of the second claim. To show the third claim, observe that $t_0 \geq T_1$ and (A.29) imply

$$\begin{aligned} TH^{-1}q_{\beta,H}^{(6)} &\sim \Delta^2 H^{-1}\textstyle\sum_{t=t_0}^{T}\big(\sum_{j=t-t_0}^{\infty} w_{j,H}\big)^2 \sim \Delta^2 H^{-1}\sum_{s=0}^{\tau}\big(\sum_{j=s}^{\infty} K(j/H)H^{-1}\big)^2 \\ &\sim \Delta^2\textstyle\int_0^{\tau/H}\big(\int_x^\infty K(v)dv\big)^2dx = G_{\tau,H}, \end{aligned}$$

which proves (A.13) and completes the proof of the lemma. □

Lemma A.3 *Under the assumptions of Lemma A.1,*

$$E\sup_{H\in I_T}(r_{T,H}^{(i)})^{-1}\big|\tilde{Q}_{T,H}^{(apr)} - E\tilde{Q}_{T,H}^{(apr)}\big| \to 0. \quad \text{(A.14)}$$

Proof. Denote $\beta_{tj} = \beta_t - \beta_{t-j}$, $u_{tj} = u_t - u_{t-j}$. Then, $\sum_{j=1}^{T_1} w_{j,H}(y_t - y_{t-j}) = \sum_{j=1}^{T_1} w_{j,H}\beta_{tj} + \sum_{j=1}^{T_1} w_{j,H}u_{tj}$, and we can write

$$Q_{T,H}^{(apr)} \;\; = \;\; T_n^{-1}\sum_{t=T_0}^{T}\Big(\sum_{j=1}^{T_1} w_{j,H}(y_t - y_{t-j})\Big)^2 = J_{\beta\beta,TH} - 2J_{\beta u,TH} + J_{uu,TH}, \quad \text{(A.15)}$$

where $J_{\beta\beta,TH} = T_n^{-1}\sum_{t=T_0}^{T}\Big(\sum_{j=1}^{T_1} w_{j,H}\beta_{tj}\Big)^2$, $J_{uu,TH} = T_n^{-1}\sum_{t=T_0}^{T}\big(\sum_{j=1}^{T_1} w_{j,H}u_{tj}\big)^2$ and $J_{\beta u,TH} = T_n^{-1}\sum_{t=T_0}^{T}(\sum_{j=1}^{T_1} w_{j,H}\beta_{tj})(\sum_{k=1}^{T_1} w_{k,H}u_{tk})$.

We will show that in cases (bi), $i = 2, \cdots, 6$,

$$E\sup_{H\in I_T}(r_{T,H}^{(i)})^{-1}E|J_{\beta\beta,TH} - EJ_{\beta\beta,TH}| \;\; \to \;\; 0, \quad \text{(A.16)}$$

$$E\sup_{H\in I_T}(r_{T,H}^{(i)})^{-1}E|J_{\beta u,TH} - EJ_{\beta u,TH}| \;\; \to \;\; 0, \quad \text{(A.17)}$$

and, in addition,

$$E\sup_{H\in I_T} H\big|J_{uu,TH} - \hat{\sigma}_{T,u}^2 - E(J_{uu,TH} - \hat{\sigma}_{T,u}^2)\big| \to 0. \quad \text{(A.18)}$$

Recall that $\tilde{Q}_{T,H}^{(apr)} = Q_{T,H}^{(apr)} - \hat{\sigma}_{T,u}^2$. Then, applying (A.16)-(A.18) in (A.15) yields (A.14).

First, we establish the following general fact. Consider the sum,

$$J_{T,H} := r_{T,H}^{-1} \sum_{j,k=1}^{T_1} w_{j,H} w_{k,H} S_{T,jk}, \tag{A.19}$$

where the random variables $S_{T,jk}$ do not depend on H and $r_{T,H}$ are real numbers such that:

(i) for some $a, b \geq 0$, uniformly in $H \in I_T$, $r_{T,H}^{-1}|w_{j,H} w_{k,H}| \leq j^{-1-a} k^{-1-b} \delta_T$,

(ii) $E|S_{T,jk} - ES_{T,jk}| \leq j^a k^b \delta'_T$, where δ_T, δ'_T depend only on T and satisfy $\delta_T \delta'_T = o((\log T)^{-2})$. Then,

$$E \sup_{H \in I_T} |J_{T,H} - EJ_{T,H}| \to 0, \quad T \to \infty. \tag{A.20}$$

To verify claim, bound $|J_{T,H} - EJ_{T,H}| \leq C\delta_T \sum_{j,k=1}^{T_1} j^{-1-a} k^{-1-b} |S_{T,jk} - ES_{T,jk}| =: \tilde{J}_T$, and note that $E\tilde{J}_T \leq C\delta_T \delta'_T \sum_{j,k=1}^{T_1} j^{-1} k^{-1} \leq C\delta_T \delta'_T \log^2 T \to 0$. Notice that (A.20) remains valid replacing in (A.19) $w_{j,H}$ by any other weights $w'_{j,H}$ satisfying (i).

Proof of (A.16). Write $J_{\beta\beta,TH} = \sum_{j,k=1}^{T_1} w_{j,H} w_{k,H} S_{\beta\beta,T,jk}$ where $S_{\beta\beta,T,jk} := T_n^{-1} \sum_{t=T_0}^{T} \beta_{tj} \beta_{tk}$. In view of (A.20), to prove (A.16), it suffices to show that in cases (b2)-(b6), the sum $J_{T,H} \equiv (1/r_{T,H}^{(i)}) J_{\beta\beta,TH}$ satisfies conditions (i) and (ii).

Case (b2). Here $r_{T,H}^{(2)} = H$. Then $r_{T,H}^{-1} w_{j,H} w_{k,H} \leq C(jk)^{-3/2}$ by (A.31), while by (A.24) of Lemma A.4, $E|S_{\beta\beta,T,jk}| \leq \delta'_T (jk)^{1/2}$ with $\delta'_T = o(\log^{-2} T)$, which verifies (i) and (ii).

Case (b3). Here $\beta_t = T^{-1/2} \tilde{\beta}_t$ and $r_{T,H}^{(3)} \geq HT^{-1}$. Then $(T/H) J_{\beta\beta,TH} = H^{-1} J_{\tilde{\beta}\tilde{\beta},TH}$ and (A.16) follows by the same argument as in the case (b2).

Case (b4)-(b6). Here (A.16) trivially holds because β_t is non-random.

Proof of (A.17). Write $J_{\beta u,TH} = \sum_{j,k=1}^{T_1} w_{j,H} w_{k,H} S_{\beta u,T,jk}$ where $S_{\beta u,T,jk} := T_n^{-1} \sum_{t=T_0}^{T} \beta_{tj} u_{tk}$. Since $Eu_t = 0$, and β_t and u_t are mutually independent, then $ES_{\beta u,T,jk} = 0$. First, we show that

$$E|S_{\beta u,T,jk}| \leq CT^{-1} D_j, \qquad D_j := (\textstyle\sum_{t=T_0}^{T} E\beta_{tj}^2)^{1/2}. \tag{A.21}$$

Indeed, $ES_{\beta u,T,jk}^2 \leq T_n^{-2} \sum_{t,s=T_0}^{T} E[\beta_{tj}\beta_{sj}] E[u_{tk} u_{sk}]$. Bound $|E[\beta_{tj}\beta_{sj}]| \leq E\beta_{tj}^2 + E\beta_{sj}^2$, and note that $|Eu_{tk}u_{sk}| \leq 2|\gamma_u(t-s)| + |\gamma_u(t-s+k)| + |\gamma_u(t-s-k)|$. Since $\sum_k |\gamma_u(k)| < \infty$, then $ES_{\beta u,T,jk}^2 \leq CT^{-2} \sum_{t=T_0}^{T} E\beta_{tj}^2 \sum_{s\in\mathbb{Z}} |\gamma_u(s)| \leq CT^{-2} D_j^2$, which implies (A.21).

To prove (A.17), it remains to show that in each case (b2)-(b6), the sum $(1/r_{T,H}^{(i)}) J_{\beta u,TH}$ satisfies conditions (i) and (ii), yielding (A.20).

Case (b2). Here, $r_{T,H}^{(2)} = H$. Then $(r_{T,H}^{(2)})^{-1} w_{j,H} w_{k,H} \leq (H^{-1} w_{j,H}) w_{k,H} \leq Cj^{-2} k^{-1}$ by (A.31). By (A.11), $E\beta_{tj}^2 \leq Cj$. Therefore, $D_j^2 \leq CTj$, and $E|S_{\beta u,T,jk}| \leq CT^{-1} D_j \leq CT^{-1/2} j^{1/2}$, which verifies (i) and (ii).

Case (b3). Observe that $r_{T,H}^{(3)} \geq HT^{-1} + H^{-1/2} \geq T^{-1/2}$, because $|a| + |b| \geq |ab|^{1/2}$. Hence, $(1/r_{T,H}^{(3)}) w_{j,H} w_{k,H} \leq T^{1/2} w_{j,H} w_{k,H} \leq CT^{1/2} (jk)^{-1}$. Next, since $\beta_t = T^{-1/2} \tilde{\beta}_t$, then $E|S_{\beta u,T,jk}| = T^{-1/2} E|S_{\tilde{\beta} u,T,jk}| \leq T^{-1} j^{1/2}$ by the same argument as in (b2). Since $T_1/T \leq T^{-\delta/2}$ for $j \leq T_1$, and $T^{-1} j^{1/2} \leq T^{-1/2} (T_1/T)^{1/2} \leq T^{-1/2-\delta/4}$, this verifies conditions (i) and (ii).

Case (b4). Here, $r_{T,H}^{(4)} = H^2$. Hence, by (A.31), $(1/r_{T,H}^{(4)}) w_{j,H} w_{k,H} = (H^{-2} w_{j,H}) w_{k,H} \leq Cj^{-3} k^{-1}$. In addition, by the mean value theorem, $|\beta_{tj}| = |tg(t/T) - (t-j) g((t-j)/T)| \leq Cj$. Thus, $D_j^2 \leq Cj^2 T$ and by (A.21), $E|S_{\beta u,T,jk}| \leq CT^{-1} (j^2 T)^{1/2} = CjT^{-1/2}$, which verifies conditions (i) and (ii).

Case (b5). Observe that $r^{(5)}_{T,H} = (H/T)^2 + H^{-1} \geq H^{1/2}T^{-1}$. Hence, by (A.31), $(1/r^{(5)}_{T,H})w_{j,H}w_{k,H} \leq T(H^{-1/2}w_{j,H})w_{k,H} \leq CTj^{-3/2}k^{-1}$. In addition, by the mean value theorem, $|\beta_{tj}| = |g(t/T) - g((t-j)/T)| \leq CjT^{-1}$, so $D_j^2 \leq Cj^2T^{-1}$ and by (A.21), $E|S_{\beta u,T,jk}| \leq CT^{-1}(j^2/T)^{1/2} = CjT^{-3/2} \leq Cj^{1/2}T^{-1}(T_1/T)^{1/2} \leq Cj^{1/2}T^{-1-\delta/4}$ for $j \leq T_1$, verifying (i) and (ii).

Case (b6). Here, $r^{(6)}_{T,H} = (H/T) + H^{-1} \geq T^{-1/2}$. Hence, by (A.31), $(1/r^{(6)}_{T,H})w_{j,H}w_{k,H} \leq T^{1/2}(jk)^{-1}$, while $D_j^2 = \sum_{t=T_0}^{T} \beta_{tj}^2 \leq \sum_{t=t_0}^{t_0+j}(\beta_t - \beta_{t-j})^2 \leq \Delta^2 j$. Then, by (A.21), $E|S_{\beta u,T,jk}| \leq CT^{-1}j^{1/2} \leq CT^{-1/2}(T_1/T)^{1/2} \leq CT^{-1/2}T^{-\delta/4}$, verifying (i) and (ii). This completes the proof of (A.17).

Proof of (A.18). Let $w'_{j,H} := w_{j,H} - w_{j+1,H}$, $j = 1, \cdots T_1 - 1$, $w'_{T_1,H} := w_{T_1,H}$, $\beta'_{tj} = \sum_{s=1}^{j} u_{t-s}$, $j = 1, \cdots, T_1$ and $h_T := \sum_{j=1}^{T_1} w_{j,H}$. Using summation by parts, write $\sum_{j=1}^{T_1} w_{j,H}u_{t-j} = \sum_{j=1}^{T_1} w'_{j,H}\beta'_{tj}$.

Then, $\sum_{j=1}^{T_1} w_{j,H}u_{tj} = h_T u_t - \sum_{j=1}^{T_1} w_{j,H}u_{t-j} = h_T u_t - \sum_{j=1}^{T_1} w'_{j,H}\beta'_{tj}$, and

$$J_{uu,TH} = T_n^{-1}\sum_{t=T_0}^{T}\big(h_T u_t - \sum_{j=1}^{T_1} w'_{j,H}\beta'_{tj}\big)^2 = J'_{\beta'\beta',TH} - 2h_T J'_{\beta' u,TH} + h_T^2\hat{\sigma}^2_{T,u},$$

where $J'_{\beta'\beta',TH} = T_n^{-1}\sum_{t=T_0}^{T}(\sum_{j=1}^{T_1} w'_{j,H}\beta'_{tj})^2$ and $J'_{\beta' u,TH} = T_n^{-1}\sum_{t=T_0}^{T}(\sum_{j=1}^{T_1} w'_{j,H}\beta'_{tj})u_t$. To prove (A.18), it suffices to verify that

$$E\sup_{H\in I_T} H\big|J'_{\beta'\beta',TH} - EJ'_{\beta'\beta',TH}\big| \to 0, \tag{A.22}$$

$$E\sup_{H\in I_T} H\big|J'_{\beta' u,TH} - EJ'_{\beta' u,TH}\big| \to 0, \qquad E\sup_{H\in I_T} H|1 - h_T^2|\hat{\sigma}^2_{T,u} \to 0. \tag{A.23}$$

To show (A.22), we use the same argument as in verifying (A.16) in case (b2). Write $J'_{\beta'\beta',TH} = \sum_{j,k=1}^{T_1} w'_{j,H}w'_{k,H}S_{\beta'\beta',T,jk}$, where $S_{\beta'\beta',T,jk} := T_n^{-1}\sum_{t=T_0}^{T}\beta'_{tj}\beta'_{tk}$. In view of (A.20), to prove (A.22), it suffices to verify conditions (i) and (ii) of (A.19) for the weights $w'_{j,H}$. We can bound $|H^{1/2}w'_{j,H}H^{1/2}w'_{k,H}| \leq C(jk)^{-3/2}$ which follows using (A.31) for $j,k < T_1$ and (A.29) for $j = k = T_1$. Moreover, since $\beta'_{tj} = \sum_{s=t-j}^{t-1} u_s$, then by (A.24) of Lemma A.4(ii), $E\big|S_{\beta'\beta',T,jk} - ES_{\beta'\beta',T,jk}\big| \leq \delta'_T(jk)^{1/2}$ with $\delta'_T = o(\log^{-2}T)$, which verifies (i) and (ii).

To show (A.23), we use the bound $H|w'_{j,H}| \leq Cj^{-1}$ which follows from (A.31) and (A.29). Then $H|J'_{\beta' u,TH} - EJ'_{\beta' u,TH}\big| \leq \sum_{j=1}^{T_1} H|w'_{j,H}|T_n^{-1}\big|\sum_{t=T_0}^{T}(\beta'_{tj}u_t - E\beta'_{tj}u_t)\big| \leq C\sum_{j=1}^{T_1} j^{-1}T^{-1}\big|\sum_{t=T_0}^{T}(\beta'_{tj}u_t - E\beta'_{tj}u_t)\big| =: \tilde{J}_T$. By (A.25) of Lemma A.4(ii), $E\big(T^{-1}\sum_{t=T_0}^{T}\{\beta'_{tj}u_t - E\beta'_{tj}u_t\}\big)^2 \leq CT^{-1}j$. Hence, $E\tilde{J}_T \leq C\sum_{j=1}^{T_1} j^{-1}(T^{-1}j)^{1/2} \leq C(T_1/T)^{1/2} \to 0$, by definition of T_1, which proves the first claim of (A.23). To show the second claim, notice that $E\hat{\sigma}^2_{T,u} = \sigma_u^2$, and $1 - h_T^2 \leq 2(1 - h_T) \leq 2\sum_{j=T_1}^{\infty} w_{j,H} = O(T^{-6})$ by (A.29), which implies (A.23) and completes the proof of the lemma. □

Lemma A.4 *Let $u_t \sim I(0)$. (i) For $1 \leq j \leq t$, set $\beta_{tj} = \sum_{i=t-j+1}^{t} u_i$, $P_{T,jk} := T_n^{-1}\sum_{t=T_0}^{T}\beta_{tj}\beta_{tk}$ and $P_{Tj} := T_n^{-1}\sum_{t=T_0}^{T}\beta_{tj}u_t$. Then,*

$$\max_{1\leq j,k\leq T_1} E(P_{T,jk} - EP_{T,jk})^2 \leq \delta'^2_T jk, \qquad \delta'_T = o(\log^{-2}T), \tag{A.24}$$

$$\max_{1\leq j\leq T_1} E(P_{Tj} - EP_{Tj})^2 \leq CT^{-1}j. \tag{A.25}$$

(ii) Bounds (A.24) and (A.25) remains valid if in $P_{T,jk}$ and $P_{T,j}$ β_{tj} are replaced by $\beta'_{tj} = \sum_{i=t-j}^{t-1} u_i$.

Proof. A short memory process u_t can be written $u_t = \sum_{s=0}^{\infty} a_s \varepsilon_{t-s}$, see (2.10), where ε_j is an $i.i.d.(0, \sigma_\varepsilon^2)$ noise, and $\sum_{k\in\mathbb{Z}} |\gamma_u(k)| < \infty$. Set for simplicity, $a_j = 0$, $j \le -1$, so $u_t = \sum_{s\in\mathbb{Z}} a_{t-s}\varepsilon_s$. Then $\beta_{tj} = \sum_{l=t-j+1}^{t} u_l = \sum_{s\in\mathbb{Z}} (\sum_{l=t-j+1}^{t} a_{l-s})\varepsilon_s$. Hence,

$$P_{T,jk} = T_n^{-1} \sum_{t=T_0}^{T} \beta_{tj}\beta_{tk} = \sum_{s,i\in\mathbb{Z}} B_{T,si}\varepsilon_s\varepsilon_i, \quad B_{T,si} := T^{-1} \sum_{t=T_0}^{T} (\sum_{l=t-j+1}^{t} a_{l-s})(\sum_{v=t-k+1}^{t} a_{v-i}).$$

By Lemma 4.5.1 of Giraitis, Koul, and Surgailis (2012), if an i.i.d. noise ε_t has finite forth moment, then a quadratic form $P_T := \sum_{s,i\in Z} \theta_{si}\varepsilon_s\varepsilon_i$ with weights θ_{si} satisfies $E(Q_T - EQ_T)^2 \le C \sum_{s,i\in\mathbb{Z}} \theta_{si}^2$, where C does not depend on θ_{si}'s. Hence,

$$E(P_{T,jk} - EP_{T,jk})^2 \le C \sum_{s,i\in\mathbb{Z}} B_{T,si}^2 = CT^{-2} \sum_{t',t=T_0}^{T} E[\beta_{t'j}\beta_{tj}]E[\beta_{t'k}\beta_{tk}] =: Cq_T. \tag{A.26}$$

Write $q_T = T^{-2} \sum_{t',t=T_0:\,|t'-t|\ge T_1+T^\epsilon}^{T} [\cdots] + T^{-2} \sum_{t',t=T_0:\,|t'-t|<T_1+T^\epsilon}^{T} [\cdots] = q_{1,T} + q_{2,T}$, where $\epsilon > 0$ is a small number. To prove (A.24), it suffices to show that $q_{i,T} \le \delta_T'^2 jk$, $i = 1, 2$.

To bound $q_{1,T}$, notice that

$$|E\beta_{t'j}\beta_{tj}| \le \delta_T' j, \quad \text{if } t' - t \ge T_1 + T^\epsilon. \tag{A.27}$$

Indeed, for $t' - j \le i' \le t$ and $t - j \le i \le t$ it holds $i' - i \ge t' - j - t \ge T^\epsilon + T_1 - j \ge T^\epsilon$. Then, $|E\beta_{t'j}\beta_{tj}| \le \sum_{i'=t'-j}^{t'} \sum_{i=t-j}^{t} |\gamma_u(i'-i)| \le Cj\delta_T'$, where $\delta_T' := \sum_{v\ge T^{\delta/2}} |\gamma_u(v)| \le C \sum_{v\ge T^\epsilon} |\gamma_u(v)| = o(\log^{-2} T)$, by Assumption 2. From (A.27) and (A.26) it follows $q_{1,T} \le \delta_T'^2 jk$.

To bound $q_{2,T}$, note that by (A.11), $|E\beta_{t'j}\beta_{tj}| \le Cj$, and observe that $T_1 + T^\epsilon = T_0 T^{-\delta/2} + T^\epsilon \le 2T^{1-\delta/2}$, where $\epsilon > 0$ is chosen small. Therefore, $q_{2,T} \le Cjk\, T^{-2} \sum_{t',t=T_0:\,|t'-t|\le T_1+T^\epsilon}^{T} 1 \le CjkT^{-\epsilon} =: \delta_T'^2 jk$, which completes the proof of (A.24).

To prove (A.25), use (A.26), $|E\beta_{t'j}\beta_{tj}| \le Cj$ and equality $\beta_{t,1} = u_t$, to obtain

$$\begin{aligned} E(P_{Tj} - EP_{Tj})^2 &\le CT^{-2} \sum_{t',t=T_0}^{T} E[\beta_{t'j}\beta_{tj}]E[\beta_{t',1}\beta_{t,1}] \\ &\le CT^{-2} \sum_{t',t=T_0}^{T} j|\gamma_u(t-t')| \le CT^{-1} j \sum_{t\in\mathbb{Z}} |\gamma_u(t)| \le CT^{-1} j. \end{aligned}$$

This completes the proof of (A.25) and part (i) of the lemma.

In part (ii) of the lemma, (A.24) and (A.25) follow using the same argument as in (i). □

A.3 Auxiliary results

Denote $v_{t,H} := \sum_{j=1}^{t-1} k_{j,H}$, $t \ge 1$ and $v_H := \sum_{j=1}^{\infty} k_{j,H}$. Recall definitions $w_{tj,H} = k_{j,H}/v_{t,H}$ and $w_{j,H} = k_{j,H}/v_H$. Below $q_{u,H}$ is as in (2.11) and T_1 as in definition of $= Q_{T,H}^{(apr)}$.

Lemma 1 *Under Assumption 1, uniformly in $H \in I_T$, $T \ge 1$, the following holds.*
(i) There exists $c > 0$, $C > 0$ such that for $0 \le \gamma \le 2$,

$$v_H \ge cH, \qquad w_{j,H} \le C(j \vee H)^{-1}, \quad j \ge 1; \tag{A.28}$$

$$w_{j,H} \le CT^{-6}, \quad j \ge T_1, \quad \sum_{j=T_1}^{\infty} w_{j,H}(j/H)^\gamma = O(T^{-6}); \tag{A.29}$$

$$|w_{tj,H} - w_{j,H}| \le CT^{-6}, \quad T_0 \le t \le T, \ \ 1 \le j \le t-1; \tag{A.30}$$

$$w_{j,H}(j/H)^\gamma \le Cj^{-1}, \quad |w_{j,H} - w_{j+1,H}|H^\gamma \le Cj^{-2+\gamma}, \quad j \ge 1. \tag{A.31}$$

(ii) As $H \to \infty$,

$$H^{-1}v_H \to 1, \quad H\sum_{j=1}^{\infty} w_{j,H}^2 \to \int_0^\infty K^2(x)dx, \qquad Hw_{0,H} \to K(0), \tag{A.32}$$

$$\sum_{j=1}^{\infty} w_{j,H}(j/H)^\gamma \to \int_0^\infty K(x)x^\gamma dx, \qquad 0 \le \gamma \le 2, \tag{A.33}$$

$$Hq_{u,H} \to \lambda_u, \tag{A.34}$$

where λ_u is as in Theorem 1.

Proof (i) To prove the first claim of (A.28), with $\epsilon > 0$ bound $v_H \ge \sum_{j=1}^{[\epsilon H]} K(j/H) \ge \delta[\epsilon H]$ where $\delta := \inf_{0\le u\le\epsilon} K(u)$. Notice that $\delta > 0$ when $\epsilon > 0$ is sufficiently small, because $K(u) \to K(0) > 0$, $u \to 0$ by Assumption 1. This implies $v_H \ge cH$ as $H \to \infty$, with $c = \delta\epsilon/2$.

To prove the second claim of (A.28), notice that by (2.3), $K(x) \le C(x^{-1} \wedge 1)$, which together with the first claim implies $w_{j,H} = k_{j,H}/v_H \le CK(j/H)H^{-1} \le C(j \vee H)^{-1}$.

To show (A.29), note that $H_{max}/T_1 \le T^{-\delta/2}$. By (2.3), one can bound $K(x) \le C|x|^{-(m+4)}$, with any $m > 0$. Choose $m\delta/2 \ge 6$. Then, for $j \ge T_1$ and $H \le H_{max}$, $k_{j,H} \le C(H/j)^{m+4} \le C(H_{max}/T_1)^m(H/j)^4 \le CT^{-6}(H/j)^4$. So, $w_{j,H} \le CT^{-6}$, since $j \ge H$. In addition, $\sum_{j=T_1}^{\infty} w_{j,H}(j/H)^\gamma \le CT^{-6}H^{-1}\sum_{j=T_1}^{\infty}(H/j)^2 \le CT^{-6}HT_1^{-1} \le CT^{-6}$, proving (A.29).

To show (A.30), first we verify that

$$|v_H - v_{t,H}| \le CT^{-6}H, \qquad t \ge T_0,$$

for some $c > 0$. By definition of H_{max} and T_0, $H_{max}/T_0 \le T^{-\delta}$. Using (2.3), bound $k_{s,H} \le C(H/s)^{m+1}$ choosing m such that $\delta m \ge 6$. Then, $v_H - v_{t,H} = \sum_{s=t}^{\infty} k_{s,H} \le C\sum_{j=t}^{\infty}(H/j)^{m+1} \le C(H/t)^m H \le C(H_{max}/T_0)^m H \le CT^{-6}H$ when $H \le H_{max}$ and $t \ge T_0$. Together with the bound $v_H^{-1} \le (cH)^{-1}$ of (A.28), this implies $|w_{tj,H} - w_{j,H}| = k_{j,H}|v_{t,H}^{-1} - v_H^{-1}| = w_{tj,H}|v_H - v_{t,H}|v_{t,H}^{-1} \le |v_H - v_{t,H}|v_H^{-1} \le CT^{-6}$.

To prove the first part of (A.31), notice that by (2.3), $K(x)x^\gamma \le Cx^{-1}$, which together with $v_H^{-1} \le CH^{-1}$ of (A.28) implies $w_{j,H}(j/H)^\gamma = k_{j,H}(j/H)^\gamma/v_H \le Cj^{-1}$. To verify the second part of (A.31), bound $H^\gamma|w_{j+1,H} - w_{j,H}| = H^\gamma v_H^{-1}|K(j+1/H) - K(j/H)| \le CH^{-2+\gamma}|\dot{K}(\xi)|$ for some $\xi \in [jH^{-1}, (j+1)H^{-1}]$. By (2.3), $|\dot{K}(\xi)| \le C\xi^{-2+\gamma} \le C(H/j)^{2-\gamma}$, which implies (A.31).

(ii) *Proof of (A.32) and (A.33).* Under (2.3), these claims follow using standard arguments of approximation of a sum by an integral.

Proof of (A.34). Write

$$q_{u,H} = \sum_{s\in\mathbb{Z}}\Big(\sum_{k=1}^{\infty} w_{k,H}w_{k+|s|,H} - w_{|s|,H}\Big)\gamma_u(s) + w_{0,H}\gamma_u(0) =: \sum_{s\in\mathbb{Z}} \nu_{s,H}\gamma_u(s) + w_{0,H}\gamma_u(0).$$

Since $w_{k+|s|,H} \le CH^{-1}$, then $H|\nu_{|s|,H}| \le C(\sum_{k=1}^{\infty} w_{k,H} + 1) = 2C$, where C does not depend on H and s. Moreover, since $|\dot{K}|$ is a bounded function, then $|w_{k+|s|,H} - w_{k,H}| = v_H^{-1}\big|K\big((k+|s|)/H\big) - K\big(k/H\big)\big| \le C|s|H^{-2}\sup_x|\dot{K}(x)| \le C|s|H^{-2}$, where C does not depend on k, s and H. Therefore, for any fixed s, $H|\nu_{s,H} - \nu_{0,H}| \le C|s|H^{-1}(\sum_{k=1}^{\infty} w_{k,H} + 1) = C|s|H^{-1} \to 0$, as $H \to \infty$. Since, $\sum_{s\in\mathbb{Z}}|\gamma_u(s)| < \infty$, by theorem of dominated convergence, $H\sum_{s\in\mathbb{Z}}|(\nu_{s,H} - \nu_{0,H})\gamma_u(s)| \to 0$. This, together with (A.32) implies (A.34):

$$\begin{aligned} HQ_{u,H} &= H\nu_{0,H}\sum_{s\in\mathbb{Z}}\gamma_u(s) + Hw_{0,H}\gamma_u(0) + o(1) = H\big(\sum_{k=1}^{\infty} w_{k,H}^2 - w_{0,H}\big)s_u^2 + Hw_{0,H}\sigma_u^2 + o(1) \\ &\to \Big(\int_0^\infty K^2(x)dx - K(0)\Big)s_u^2 + K(0)\sigma_u^2 = \lambda_u, \quad H \to \infty. \end{aligned}$$

□

References

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

Bai, J., and P. Perron (1998): "Estimating and Testing Linear Models with Multiple Structural Changes," *Econometrica*, 66, 47–78.

Castle, J. L., N. Fawcett, and D. Hendry (2011): "Forecasting Breaks and Forecasting During Breaks," in *Oxford Handbook of Economic Forecasting*, ed. by M. P. Clements, and D. F. Hendry, pp. 315–354. Oxford University Press.

Chow, G. (1960): "Tests of equality between sets of coefficients in two linear regressions," *Econometrica*, 28, 591–603.

Clements, M. P., and D. F. Hendry (1998a): *Forecasting economic time series*. CUP, Cambridge.

Clements, M. P., and D. F. Hendry (1998b): "Intercept corrections and structural change," *Journal of Applied Econometrics*, 11, 475–94.

Clements, M. P., and D. F. Hendry (2006): "Forecasting with Breaks," in *Handbook of Economic Forecasting*, ed. by G. Elliott, C. W. J. Granger, and A. Timmermann, pp. 605–657. Elsevier, North-Holland.

Dahlhaus, R. (1996): "Fitting Time Series Model to Nonstationary Processes," *Annals of Statistics*, 25, 1–37.

Eklund, J., G. Kapetanios, and S. Price (2010): "Forecasting in the presence of recent structural change," Bank of England Working Paper 406.

Elliott, G., and A. Timmermann (2008): "Economic Forecasting," *Journal of Economic Literature*, 46(1), 3–56.

Giacomini, R., and H. White (2005): "Tests of conditional predictive ability," *Econometrica*, 74, 1545–1578.

Giraitis, L., G. Kapetanios, and T. Yates (2011): "Inference on stochastic time-varying coefficient models," Queen Mary, University of London Working Paper no. 540.

Giraitis, L., H. Koul, and D. Surgailis (2012): *Large Sample Inference for Long memory Processes*. Imperial College Press, London.

Hendry, D. F. (2000): "On detectable and non-detectable structural change," *Structural Change and Economic Dynamics*, 11, 45–65.

Kapetanios, G. (2007): "Estimating Deterministically Time Varying Variances in Regression Models," *Economics Letters*, 97(2), 97–104.

Kapetanios, G., V. Labhard, and S. Price (2006): "Forecasting using predictive likelihood model averaging," *Economics Letters*, 91, 373–379.

Orbe, S., E. Ferreira, and J. Rodriguez-Poo (2005): "Nonparametric Estimation of Time Varying Parameters under Shape Restrictions," *Journal of Econometrics*, 126, 53–77.

Pesaran, M. H., and A. Timmermann (2007): "Selection of estimation window in the presence of breaks," *Journal of Econometrics*, 137, 134–61.

Rossi, B. (2012): "Advances in Forecasting Under Instability," in *Handbook of Economic Forecasting*, ed. by G. Elliott, and A. Timmermann. Elsevier-North Holland.

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