A New Nonparametric Combination Forecasting with Structural Breaks*

Zongwu Cai[a], Gunawan[b], and Yuying Sun[c]

[a]Department of Economics, University of Kansas, Lawrence, KS 66045, USA

[b]Faculty of Economics and Business, Universitas Gadjah Mada, Yogyakarta 55281, Indonesia

[c]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

September 23, 2024

Abstract: This paper proposes a new nonparametric forecasting procedure based on a weighted local linear estimator for a nonparametric model with structural breaks. The proposed method assigns a weight based on both the distance of observations to the predictor covariates and their location in time and the weight is chosen using multifold forward-validation to account for time series data. We investigate the asymptotic properties of the proposed estimator and show that the weight estimated by the multifold forward-validation is asymptotically optimal in the sense of achieving the lowest possible out-of-sample prediction risk. Additionally, a nonparametric method is adopted to estimate the break date and the proposed approach allows for different features of predictors before and after break. A Monte Carlo simulation study is conducted to provide evidence for the forecasting outperformance of the proposed method over the regular nonparametric post-break and full-sample estimators. Finally, an empirical application to volatility forecasting compares several popular parametric and nonparametric methods, including the proposed weighted local linear estimator, demonstrating its superiority over other alternative methods.

JEL Classification: C14, C22, C53

Keywords: Combination Forecasting; Model Averaging; multifold forward-validation; Nonparametric Model; Structural Break Model; Weighted Local Linear Fitting.

*Contact information: Z. Cai (caiz@ku.edu), Gunawan (gunawan.lee@ugm.ac.id), and Y. Sun (sunyuying@amss.ac.cn). Authors thank a partial support from Natural Sciences Foundation of China grants (72322016, 72073126, 72091212, 71973116, 71631004, 72033008) and Young Elite Scientists Sponsorship Program by CAST[YESS20200072].

1 Introduction

Forecasting time series data often assumes stationarity, and therefore the constancy of model parameters over time, such as mean, variance, frequency, trend, or combined. In practice, these parameters may change over time. For example, the US industrial production experienced slowdown during the financial crisis between 2007 and 2008, as well as the Covid-19 pandemic between 2020 and 2023, while it experienced expansion in other time periods. Therefore, investigating structural instability has been a long-standing issue in time series. These two different regimes are regarded as a consequence of either parameter shifts or parameters varying smoothly over time. For the latter case, the reader is referred to the papers by Cai (2007), Sun, Hong, Lee, Wang, and Zhang (2021), and references therein. The point at which the regime change occurs is called a change point or structural break in the statistics and econometrics literature, whereas the associated models are called to be models with structural break. In practice, breaks in the parameters of a forecasting model are caused by events, economic policies, or treatments that are essentially unknowable ex-ante and may be triggered by various factors, such as institutional, political, social, financial, legal, or technological changes, which may precipitate these breaks. Such breaks are understood better retrospectively rather than at the time of their occurrence. Typically, it is assumed that the modeler does not have knowledge of the process determining the break as addressed in Clements and Hendry (2011).

Structural breaks pose statistical challenges for forecasting exercise. In a time series model with a structural break in the conditional mean and/or conditional variance, a conventional OLS estimator based on full-sample observations might be inconsistent. A consistent estimator can be computed using post-break observations only if the post-break sample is sufficiently large. However, such forecasts may not be optimal or efficient in terms of the mean squared forecast error (MSFE), as the relatively small post-break sample size may induce large estimation uncertainty, especially, for linear models, as addressed by Pesaran and Pick (2011), Pesaran, Pick, and Pranovich (2013), Rossi (2013), Boot and Pick (2020), Lee, Parsaeian, and Ullah (2022a,b), Parsaeian (2023), and references therein. Especially, Boot and Pick (2020) provided a test to determine whether modeling a structural break improves forecast accuracy. Therefore, pre-break observations may still be useful for forecast improvement depending on the magnitude of the break size. If there is no break, the usual full-sample estimator is optimal. If the break is strong, the post-break estimator may be optimal (efficient). If the break is weak or moderate, a combined estimator of the full-sample estimator or the pre-break estimator and the post-break estimator would be optimal, where a combination weight between 0 and 1 is chosen in a way that optimizes the trade-off between the bias and variance efficiency of the full-sample estimator. Obviously, the break might cause the distributions of dependent variable and predictors to be different before and after break. Unfortunately, in the aforementioned literature for linear models, it is implicitly assumed that the distribution of predictors is the same before and after break.

The idea of combining information in producing the forecast could be considered as frequentist model averaging, since we average the pre-break and post break estimators as in Hjort and Claeskens (2003), Hansen (2007), Hansen (2008), Hansen and Racine (2012), Sun et al. (2021), Lee et al. (2022a,b), Sun, Hong, and Cai (2023), and references therein. In this spirit, there is a vast account of literature on different forecast combination methods, particularly, in the parametric literature, see, to name just a few, Clements and Hendry (2006), Clements and Hendry (2011), Pesaran and Timmermann (2005), Pesaran and Timmermann (2007), Timmermann (2006), Pesaran et al. (2013), Boot and Pick (2020), Lee et al. (2022a,b), and references therein. However, to the best of our knowledge, the literature on nonparametric forecast combination methods capable of handling structural changes, especially structural breaks, remains relatively limited; see, for example, Sun et al. (2021) and Sun et al. (2023).

This paper contributes to the nonparametric forecasting with structural breaks literature by proposing a combined nonparametric method to exploit information contained in the dataset before break occurs. Our proposed estimator, inspired by the model averaging method, assigns a weight to observations before and after break. This weight is additional to the usual nonparametric weights that are given to observations based on how far they are located relative to the predictor covariates. Hence, it is termed as a *weighted local linear estimator*. Also, the asymptotic properties, including the asymptotic bias and variance, of the proposed estimator are investigated and some discussions are provided to show that the asymptotic variance indeed can be smaller than that for the nonparametric estimator using only the post-break observations. Furthermore, we propose a novel multifold forward-validation model averaging (MFVMA) approach for selecting data-driven weights in time series forecasting, and the break date estimation employs the latest nonparametric method from Mohr and Selk (2020). This approach is related to cross-validation as discussed in the model averaging and nonparametric literature such as Cai, Fan, and Yao (2000), Zhang and Liu (2023), Gao, Zhang, Wang, and Zou (2016), Liao, Zong, Zhang, and Zou (2019), Cheng and Hansen (2015), Lee et al. (2022a), and references therin. Unlike the standard cross-validation used in model averaging, our multifold forward-validation captures the temporal ordering of time series forecasting and utilizes only the data available up to the forecast time point. The idea behind multifold forward-validation is to divide the dataset into multiple groups, treating each group as a validation set for evaluating the model. Crucially, the validation set always precedes the training set temporally. The implementation of multifold forward-validation is straightforward and flexible, seldom relying on model structure assumptions, unlike criteria such as Mallows-type or other information criteria which require the derivation of related penalty terms as in Zhu, Wan, Zhang, and Zou (2019), Liu and Okui (2013), Li, Li, Racine, and Zhang (2018), and references therein. Finally, we demonstrate that the selected weight is asymptotically optimal in the sense of minimizing the out-of-sample prediction risk, thereby complementing existing methods that primarily concentrate on minimizing the in-sample squared error loss under structural break scenarios. To establish the asymptotic optimality from the predictive perspective, we propose a novel strategy to bound the discrepancy between the nonparametric-based multifold forward-validation and the out-of-sample prediction risk function, instead of Whittle's inequality as in Li (1987) and Hansen and Racine (2012).

The remainder of the paper is organized as follows. In addition to the model setup, the weighted nonparametric regression predictor is proposed and its asymptotic properties are studied in Section 2, together some practical issues such as the break date estimator, how to choose the tuning parameters, and a straightforward generalization of the proposed method to the multiple breaks case. More importantly, we show that the weight estimated by the multifold forward-validation is asymptotically optimal. Section 3 presents a Monte Carlo simulation study and reports its results. Section 4 illustrates an empirical application, while the detailed theoretical justifications are relegated to Section 5. Finally, Section 6 concludes the paper.

2 Forecasting Procedures

2.1 Model Setup

Let $\{(Y_t, \mathbf{X}_t) : t \in \mathbb{N}\}$ be a weakly dependent stochastic process in $\mathbb{R} \times \mathbb{R}^d$. We consider following the forecasting model

$$Y_{t+\tau} = m_t(\mathbf{X}_t) + u_{t+\tau}, \quad 1 \le t \le T, \tag{1}$$

where $\tau \ge 0$ is the given (known) forecasting horizon (τ-step ahead forecast), and the idiosyncratic error $u_{t+\tau}$ satisfies $\mathbb{E}[u_{t+\tau}|\mathcal{F}_t] = 0$ almost surely for the σ-field $\mathcal{F}_t = \sigma(u_{j-1}, \mathbf{X}_j : j \le t)$. It is assumed that there exists a change point at time T_1 with $1 \le T_1 \le T$, in the prediction function such that

$$m_t(\mathbf{x}) = m_{(1)}(\mathbf{x})\mathbb{1}\,(t \le T_1) + m_{(2)}(\mathbf{x})\mathbb{1}\,(t > T_1) = m_{(1)}(\mathbf{x}) - \lambda(\mathbf{x})d_t, \tag{2}$$

where $m_{(1)}(\mathbf{x}) \not\equiv m_{(2)}(\mathbf{x})$ and $\lambda(\mathbf{x}) = m_{(1)}(\mathbf{x}) - m_{(2)}(\mathbf{x})$, the break size function, the break point T_1 might be unknown, $d_t = \mathbb{1}\,(t > T_1)$, and both functions $m_{(1)}(\mathbf{x})$ and $m_{(2)}(\mathbf{x})$ are assumed to be continuous and satisfy some regularity conditions to ensure that $\{(Y_t, \mathbf{X}_t) : t \in \mathbb{N}\}$ is a (or piece wise) stationary α-mixing time series. Here, $\mathbf{X}_t$ is allowed to include some lags of Y_t.[1] If so, the distributions of $\mathbf{X}_t$ might be different before and after break. Without loss of generality, it is assumed that $\mathbf{X}_t$ for $1 \le t \le T_1$ (before break) is stationary with its density $f_b(\cdot)$ and $\mathbf{X}_t$ for $T_1 + 1 \le t \le T$ (after break) is also stationary with its density $f_a(\cdot)$. But, $f_b(\cdot)$ and $f_a(\cdot)$ might not be exactly same. Define $\delta(\mathbf{x}) = f_a(\mathbf{x})/f_b(\mathbf{x})$, which is called the covariate shift[2] function in the machine learning literature for causal inferences, to capture different features $\mathbf{X}_t$ before and after break since $f_b(\cdot)$ and $f_a(\cdot)$ are allowed to be different. Throughout the paper, it is assumed that $T_1 = \lfloor Ts_0 \rfloor$ with $0 \le s_0 \le 1$, the portion of the pre-break observations, so that $T_2 = T - T_1 = T - \lfloor Ts_0 \rfloor$, the portion of the post-break observations. Clearly, for two extreme cases, $s_0 = 0$ means that there is no break and $s_0 = 1$ implies that there is no observation in the post break period. Therefore, without loss of generality, it is assumed throughout the paper that $0 < s_0 < 1$. Finally, note that the expression in the right hand side of (2) would be regarded as a special case of a functional coefficient time series model proposed in Cai et al. (2000) if d_t would be known.

[1] If $\mathbf{X}_t$ contains some lags of Y_t, there is an issue regarding to the stationarity of Y_t. For this aspect, the reader is referred to the paper by Cai and Masry (2000) for details on the conditions on $m_t(\mathbf{X}_t)$ and the theoretical justifications.

[2] The reader is referred to the paper by Bickel, Brückner, and Scheffer (2009) and the book by Sugiyama, Suzuki, and Kanamori (2012) for details on this topic.

Remark 1. *In the literature for linear models, see, for example, Pesaran and Pick (2011), Pesaran et al. (2013), Rossi (2013), and Lee et al. (2022a,b), it is implicitly assumed that both density functions $f_b(\mathbf{x})$ and $f_a(\mathbf{x})$ are same, which is different from our setting here. A structural change can be regarded as an event study and may be caused by an economic policy change, or an intervention (such as COVID-19), or a treatment (some programs), so that $f_b(\mathbf{x})$ and $f_a(\mathbf{x})$ are commonly assumed to be different in the causal inference literature to capture different features $\mathbf{X}_t$. For more details on this aspect, the reader is referred to the paper by Cai, Fang, Lin, and Wu (2023) and references therein, although the main focus in the causal inference is somewhat different from the setting here.*

It is clear that when $m_t(\mathbf{x}) = \beta_t^\top \mathbf{x}$ in (1) with β_t changing smoothly over time, the model in (1) becomes the models studied by Cai (2007) for estimation and forecasting and Sun et al. (2021) for a model averaging. Furthermore, when β_t has structural change, the model in (2) was investigated by Pesaran et al. (2013) and Lee et al. (2022a,b) for the weighted generalized least squares (WGLS) estimators for a conventional structural change linear model to combine the information from both pre-break and post-break. As argued in Pesaran et al. (2013) and Lee et al. (2022a,b), the WGLS estimators proposed in Pesaran et al. (2013) and Lee et al. (2022a,b) have an ability to reduce MSFE under the structural breaks by using the full-sample observations instead of using only the post-break observations, by deriving the optimal weight for the pre-break proportion of the full-sample. Note that for simplicity, our focus is on (2) with only one break, and it is easy to generalize the model in (2) to the multiple breaks case, briefly discussed in Section 2.5.2.

2.2 Weighted Local Linear Estimation

In this subsection, we propose an estimator for nonparametric model with structural break, where break may occur in the mean function and error variance. In particular, we are interested in estimating the mean function after break by partly using information contained in the pre-break observations. Our starting point is the following nonparametric local linear regression problem. For $\mathbf{X}_t$ in a neighborhood of $\mathbf{x}$, a given grid point from the data domain, we can approximate locally the mean function by $m(\mathbf{X}_t) \approx \beta_0(\mathbf{x}) + \beta_1(\mathbf{x})^\top(\mathbf{X}_t - \mathbf{x})$ by ignoring the higher order term, where $\beta_0(\mathbf{x}) = m(\mathbf{x})$ and $\beta_1(\mathbf{x}) = m'(\mathbf{x})$, the first order derivative of $m(\mathbf{x})$. Then, for the given data $\{(Y_t, \mathbf{X}_t)\}_{t=1}^T$, the locally weighted least squares is given by

$$\min_{\beta_0,\beta_1} \sum_{t=1}^{T-\tau} w_t(\gamma, \mathbf{x}) \left(Y_{t+\tau} - \beta_0 - \beta_1^\top (\mathbf{X}_t - \mathbf{x})\right)^2, \tag{3}$$

where

$$w_t(\gamma, \mathbf{x}) = \gamma \mathbb{1}\,(t \le T_1) K_{h_1}(\mathbf{x} - \mathbf{X}_t) + \mathbb{1}\,(t > T_1) K_{h_2}(\mathbf{x} - \mathbf{X}_t) \tag{4}$$

for some $0 \le \gamma \le 1$, and $K_h(u) = K(u/h)/h^d$ with $K(\cdot)$ being a kernel function. To capture different features of $f_b(\cdot)$ and $f_a(\cdot)$, two bandwidths h_1 and h_2 are used: h_1 is for $m_{(1)}(\cdot)$ and h_2 is for $m_{(2)}(\cdot)$. If both $m_{(1)}(\cdot)$ and $m_{(2)}(\cdot)$ have the same degree of smoothness, then, h_1 and h_2 should be the same, denoted by h, so that $w_t(\gamma, \mathbf{x}) = [\gamma \mathbb{1}\,(t \le T_1) + \mathbb{1}\,(t > T_1)] K_h(\mathbf{x} - \mathbf{X}_t)$. As mentioned in Cai et al. (2000), the estimation procedure and its asymptotic theory for the d-dimensional case are the same for the case that $\mathbf{X}_t$ is the univariate case. Therefore, for ease notation, in what follows, the presentation is only for one-dimensional case; that is $d = 1$, so that $\mathbf{X}_t$ and $\mathbf{x}$ become to be X_t and x, respectively.

Equation (4) takes care of both break and smoothnesses of $m_{(1)}(\cdot)$ and $m_{(2)}(\cdot)$ so that the weighting scheme $w_t(\gamma, x)$ assigns a weight to the observations before break, and assigns a weight on each observation based on how close X_t is to the grid point x. Based on (4), post-break observations receive a weight of 1, while a weight of $\gamma \in [0, 1]$ is assigned to pre-break observations. If $\gamma = 0$, the estimator is based on only the post-break observations, whereas γ is close to zero, then, the estimator is heavily weighted on the post-break observations with a small part of information before break. If $\gamma = 1$, then, a structural break is ignored and a full-sample is used to produce a full sample estimator. In other cases where $\gamma \in (0, 1)$, a combination of pre- and post-break observations for the estimator is obtained.

The minimizer of (3) is denoted by $\widehat{\beta}(x) = (\widehat{\beta}_0(x), \widehat{\beta}_1(x))^\top$, which gives $\widehat{m}(x) = \widehat{\beta}_0(x)$, the estimator of $m(x)$, and $\widehat{m}'(x) = \widehat{\beta}_1(x)$, the estimator of $m'(x)$, respectively. To express the estimator in matrix form, we introduce the following notations. Let $Y^\top = (Y_{(1)}^\top, Y_{(2)}^\top)$ be a $(T - \tau) \times 1$ vector of the dependent variable with $Y_{(1)} = (Y_{1+\tau}, \ldots, y_{T_1+\tau})^\top$ and $Y_{(2)} = (y_{T_1+\tau+1}, \ldots, y_T)^\top$, and $\mathbf{X}^\top = (\mathbf{X}_{(1)}^\top, \mathbf{X}_{(2)}^\top)$ be a $(T - \tau) \times 2$ matrix

$$\mathbf{X}_{(1)}^\top = \begin{pmatrix} 1 & \cdots & 1 \\ (X_1 - x) & \cdots & (X_{T_1} - x) \end{pmatrix} \quad \text{and} \quad \mathbf{X}_{(2)}^\top = \begin{pmatrix} 1 & \cdots & 1 \\ (X_{T_1+1} - x) & \cdots & (X_{T-\tau} - x) \end{pmatrix}.$$

Now, define $\mathbf{W}(\gamma)$ as follows: $\mathbf{W}(\gamma) = \mathbf{W}_\gamma \mathbf{W}_k$, where $\mathbf{W}_\gamma = \text{diag}\{\gamma \mathbf{I}_{T_1}, \mathbf{I}_{T-T_1-\tau}\}$ and $\mathbf{W}_k = \text{diag}\{\mathbf{W}_{(1)}, \mathbf{W}_{(2)}\}$ with $\mathbf{W}_{(1)} = \text{diag}(K_{h_1}(x - X_1), \ldots, K_{h_1}(x - X_{T_1}))$ and $\mathbf{W}_{(2)} =$

$\mathrm{diag}(K_{h_2}(x - X_{T_1+1}), \ldots, K_{h_2}(x - X_{T-\tau}))$ as well as $\mathbf{I}_\ell$ denoting an $\ell \times \ell$ identity matrix. Thus, the minimizer of (3) is given by

$$\widehat{\beta}(x) = (\widehat{\beta}_0(x), \widehat{\beta}_1(x))^\top = (\mathbf{X}^\top \mathbf{W}(\gamma) \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}(\gamma) Y. \tag{5}$$

In particular, the weighted local linear (WLL) estimator for the mean function is given by

$$\widehat{m}_{\mathrm{wll}}(x) = \widehat{\beta}_0(x) = \mathbf{e}^\top \widehat{\beta}(x), \tag{6}$$

where $\mathbf{e}^\top = (1, 0)$, and it reduces to the local linear estimator of $m_{(2)}(x)$ based on the observations after break if $\gamma = 0$. Further, equation (5) can be rewritten as

$$\begin{aligned} \widehat{\beta}(x) &= \left[\mathbf{X}^\top \mathbf{W}(\gamma) \mathbf{X}\right]^{-1} \left(\gamma \mathbf{X}_{(1)}^\top \mathbf{W}_{(1)} Y_{(1)} + \mathbf{X}_{(2)}^\top \mathbf{W}_{(2)} Y_{(2)}\right) \\ &= \Gamma \widehat{\beta}_{(1)}(x) + (\mathbf{I}_2 - \Gamma) \widehat{\beta}_{(2)}(x) = \Theta \widehat{\beta}_{\mathrm{full}}(x) + (\mathbf{I}_2 - \Theta) \widehat{\beta}_{(2)}(x), \end{aligned} \tag{7}$$

where $\Gamma = \Gamma(x, \gamma) = \gamma \left[\mathbf{X}^\top \mathbf{W}(\gamma) \mathbf{X}\right]^{-1} \left(\mathbf{X}_{(1)}^\top W_{(1)} \mathbf{X}_{(1)}\right)$, $\widehat{\beta}_{(1)}(x)$ is the local linear estimator using the observations before break, and $\widehat{\beta}_{(2)}(x)$ is the estimator using the observations after break.[3] Further, $\widehat{\beta}_{\mathrm{full}}(x)$ is the local linear estimator using the full sample[4] and $\Theta = \theta(x, \gamma) = \gamma \left[\mathbf{X}^\top \mathbf{W}(\gamma) \mathbf{X}\right]^{-1} \left(\mathbf{X}^\top \mathbf{W}(1) \mathbf{X}\right)$. Therefore, equation (7) can be viewed as the combined estimator of the pre-break and the post-break estimators, i.e., a combination of $\widehat{\beta}_{(1)}(x)$ and $\widehat{\beta}_{(2)}(x)$ with the weight Γ. Alternatively, it can be regarded as the combined estimator from the full sample estimator and the post-break estimator, i.e., a combination of $\widehat{\beta}_{\mathrm{full}}(x)$ and $\widehat{\beta}_{(2)}(x)$ with the weight Θ. Clearly, $\widehat{\beta}(x)$ in (7) involves two bandwidths h_1 and h_2 and weight γ.

2.3 Asymptotic Properties

Before embarking on deriving the asymptotic results, we now give some regularity conditions that are sufficient for the consistency and asymptotic normality of the proposed estimators, although they might not be the weakest ones possible. As pointed out by Cai et al. (2000), the conditions list below are standard and they are satisfied for many applications; see, for instance, the paper by Cai et al. (2000) for details. Then, we present the sketch proofs of the asymptotic properties in Section 5.

[3] $\widehat{\beta}_{(1)}(x) = \left[\mathbf{X}_{(1)}^\top \mathbf{W}_{(1)} \mathbf{X}_{(1)}\right]^{-1} \left(\mathbf{X}_{(1)}^\top \mathbf{W}_{(1)} Y_{(1)}\right)$ and $\widehat{\beta}_{(2)}(x) = \left[\mathbf{X}_{(2)}^\top \mathbf{W}_{(2)} \mathbf{X}_{(2)}\right]^{-1} \left(\mathbf{X}_{(2)}^\top \mathbf{W}_{(2)} Y_{(2)}\right)$.

[4] $\widehat{\beta}_{\mathrm{full}}(x) = (\mathbf{X}^\top \mathbf{W}_k \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}_k Y = \left[\mathbf{X}_{(1)}^\top \mathbf{W}_{(1)} \mathbf{X}_{(1)} + \mathbf{X}_{(2)}^\top \mathbf{W}_{(2)} \mathbf{X}_{(2)}\right]^{-1} \left(\mathbf{X}_{(1)}^\top \mathbf{W}_{(1)} Y_{(1)} + \mathbf{X}_{(2)}^\top \mathbf{W}_{(2)} Y_{(2)}\right)$.

2.3.1 Conditions

Condition A:

A1. The second order derivatives of both mean functions $m_{(1)}(x)$ and $m_{(2)}(x)$ are continuously differentiable.

A2. Both functions $f_b(x)$ and $f_a(x)$ are continuous and positive within the support.

A3. The condition density of $Y_{t+\tau}$ given X_t is bounded and satisfies the Lipschitz condition.

A4. The kernel function $K(\cdot)$ is symmetric and has a compact support, say $[-1, 1]$.

A5. The time series $\{(Y_t, X_t) : t \in \mathbb{N}\}$ is α-mixing with the coefficient $\alpha(k)$ satisfying $\sum_{k=1}^{\infty} k^{c_0} \alpha^{1-2/\delta_0}(k)$ for some $\delta_0 > 2$ and $c_0 > 1 - 2/\delta_0$.

A6. Assume that $h_1 \to 0$, $h_2 \to 0$, $T_1 h_1 \to \infty$, and $T_2 \, h_2 \to \infty$. Also, $\lim_{T\to\infty} h_2/h_1 = h_c$ for some $0 < h_c < \infty$.

Condition B:

B1. Assume that

$$\mathbb{E}\left[Y_{t+\tau}^2 + Y_{t+s+\tau}^2 \,|\, X_t = x_1, X_{t+s} = x_2\right] \le M < \infty$$

for any t, $s \geq 1$, x_1 and x_2.

B2. Assume that there exists a sequence of positive integers $\{s_T\}$ such that $s_T \to \infty$, $s_T = o((Th)^{1/2})$ and $(T/s_T)^{1/2}\alpha(s_T) \to 0$, as $T \to \infty$.

B3. There exists $\delta^* > \delta_0$, where δ_0 is given in Assumption A(5) such that $\alpha(k) = O(k^{-\theta})$, where $\theta > \delta_0\delta^*/[2(\delta^* - \delta_0)]$.

B4. Both h_1 and h_2 satisfy $T_j^{1/2-\delta_0/4} h_j^{\delta_0/\delta^*-1/2-\delta_0/4} = O(1)$ for $j = 1$ and 2.

2.3.2 Asymptotic Theory

Now, we investigate the asymptotic properties of $\widehat{m}_{\text{wll}}(x)$. First, we evaluate Γ. To do so, consider $\mathbf{X}_{(1)}^\top \mathbf{W}_{(1)} \mathbf{X}_{(1)}$. For $j \geq 0$, let $S_j(x) = \frac{1}{T_1} \sum_{t=1}^{T_1} K_{h_1}(X_t - x) \left((X_t - x)/h_1\right)^j$. It is easy to see that

$$\mathbf{X}_{(1)}^\top \mathbf{W}_{(1)} \mathbf{X}_{(1)} = T_1 \mathbf{H_1} \begin{pmatrix} S_0(x) & S_1(x) \\ S_1(x) & S_2(x) \end{pmatrix} \mathbf{H_1},$$

where $\mathbf{H_1} = \text{diag}\{1, h_1\}$. Under Assumptions A1 - A5, it follows from (17) in Section 5 that as $T \to \infty$, $S_j(x) \xrightarrow{p} \mu_j\, f_b(x)$, where $\mu_j = \int K(u)u^j du$ for $j \geq 0$. Therefore, $\mathbf{X}_{(1)}^\top \mathbf{W}_{(1)} \mathbf{X}_{(1)} = f_b(x) T_1 \mathbf{H_1}\, \mu\, \mathbf{H_1}(1+o_p(1))$, where $\mu = \text{diag}\{1, \mu_2\}$. Similarly, $\mathbf{X}_{(2)}^\top \mathbf{W}_{(2)} \mathbf{X}_{(2)} = f_a(x) T_2 \mathbf{H_2}\, \mu\, \mathbf{H_2}(1+o_p(1))$, where $\mathbf{H_2} = \text{diag}\{1, h_2\}$. Hence, $\mathbf{X}^\top \mathbf{W}(\gamma)\mathbf{X} = [\gamma f_b(x) T_1 \mathbf{H}_1 \mu \mathbf{H}_1 + f_a(x) T_2 \mathbf{H_2}\, \mu\, \mathbf{H_2}](1+o_p(1))$, which implies that $\Gamma = \text{diag}\{s_b, s_a\}(1 + o_p(1))$, where $s_b = s_b(\gamma, s_0, x) = \gamma s_0\, [\gamma\, s_0 + (1 - s_0)\delta(x)]^{-1}$ with $\delta(x)$ being the covariate shift function, and $s_a = s_a(\gamma, s_0, x, h_c) = \gamma s_0\, [\gamma\, s_0 + (1 - s_0)\delta(x) h_c^2]^{-1}$ with $h_c = \lim_{T\to\infty}(h_2/h_1)$.[5] Clearly, s_b depends on both γ and s_0 as well as the covariate shift function $\delta(x)$. Note that if $\delta(x) = 1$, both s_b and s_a do not depend on x. Finally, it is easy to see that $0 \leq s_b \leq 1$.

Next, we evaluate the asymptotic bias for $\widehat{m}_{\text{wll}}(x)$. For this purpose, (7) is re-expressed as $\widehat{\beta}(x) = \widehat{\beta}_{(2)}(x) + \Gamma\left[\widehat{\beta}_{(1)}(x) - \widehat{\beta}_{(2)}(x)\right]$, so that $\widehat{m}_{\text{wll}}(x) \approx \widehat{\beta}_{0,(2)}(x) + s_b\left[\widehat{\beta}_{0,(1)}(x) - \widehat{\beta}_{0,(2)}(x)\right]$, where $\widehat{\beta}_{0,(1)}(x)$ and $\widehat{\beta}_{0,(2)}(x)$ are the first component of $\widehat{\beta}_{(1)}(x)$ and $\widehat{\beta}_{(2)}(x)$, respectively. Indeed, $\widehat{\beta}_{0,(1)}(x)$ is the local linear estimator for $m_{(1)}(x)$ using only the pre-break observations and $\widehat{\beta}_{0,(2)}(x)$ is the local linear estimator for $m_{(2)}(x)$ using only the post-break observations, denoted by $\widehat{m}_{(2)}(x)$. Also, we show in Section 5 that the asymptotic biases for $\widehat{\beta}_{0,(1)}(x)$ and $\widehat{\beta}_{0,(2)}(x)$ are $B_1(x) = h_1^2 m''_{(1)}(x)\mu_2/2$ and $B_2(x) = h_2^2 m''_{(2)}(x)\mu_2/2$, respectively. Therefore, the asymptotic bias for $\widehat{m}_{\text{wll}}(x)$ is

$$B_{\text{wll}}(x) = s_b \lambda(x) + s_b B_1(x) + (1 - s_b) B_2(x), \tag{8}$$

where $\lambda(x)$ is defined in (2). Clearly, the first term in the right hand side of $B_{\text{wll}}(x)$ is extra by comparing with that for $\widehat{\beta}_{0,(2)}(x)$ due to the weighted estimation procedure and it is negative if $\lambda(x) < 0$ by ignoring the higher order term. Finally, one can see that for a linear model ($m_t(X_t) = \beta_t^\top X_t$), $B_{\text{wll}}(x)$ reduces to $s_b\lambda(x)$, which is similar to those in Pesaran et al. (2013) and Lee et al. (2022a), so that the results in Pesaran et al. (2013) and Lee et al. (2022a) can be regarded as a special case of (8).

Finally, in addition to the asymptotic bias given in (8), we consider the asymptotic variance of $\widehat{m}_{\text{wll}}(x)$. To this end, we express

$$\mathbf{X}^\top \mathbf{W}(\gamma) U = \gamma \sum_{t=1}^{T_1} K_{h_1}(X_t - x) \begin{pmatrix} 1 \\ X_t - x \end{pmatrix} u_{t+\tau} + \sum_{t=T_1+1}^{T-\tau} K_{h_2}(X_t - x) \begin{pmatrix} 1 \\ X_t - x \end{pmatrix} u_{t+\tau} = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix},$$

[5] According to the asymptotic theory for the kernel estimation for nonparametric regression models, see, for example, Fan and Gijbels (1996) and Fan and Yao (2003), the optimal bandwidth for h_1 is $h_{1,opt} = O_p(T_1^{-1/(4+d)})$ and the one for h_2 is $h_{2,opt} = O_p(T_2^{-1/(4+d)})$. Therefore, h_c exists and $0 < h_c < \infty$.

where U is defined in the same way as Y, which is the main term that contributes to the asymptotic variance of $\widehat{m}_{\rm wll}(x)$, and A_1 and A_2 are defined in a clear manner. Clearly,

$$C_0(\gamma) = \sqrt{\frac{h_2}{T}}\,A_1 = \sqrt{\frac{h_2}{T}}\left[\gamma\sum_{t=1}^{T_1}K_{h_1}(X_t-x)u_{t+\tau} + \sum_{t=T_1+1}^{T-\tau}K_{h_2}(X_t-x)u_{t+\tau}\right]$$
$$\approx \gamma\sqrt{h_c\,s_0}\,C_1 + \sqrt{1-s_0}\,C_2,$$

where

$$C_1 = \sqrt{\frac{h_1}{T_1}}\sum_{t=1}^{T_1}K_{h_1}(X_t-x)u_{t+\tau} \quad \text{and} \quad C_2 = \sqrt{\frac{h_2}{T_2}}\sum_{t=T_1+1}^{T-\tau}K_{h_2}(X_t-x)u_{t+\tau}$$

One can show in Section 5 that under Assumptions B1 - B4,

$$C_1 \xrightarrow{d} N\left(0, \sigma^2_{m,1}(x)\right) \quad \text{and} \quad C_2 \xrightarrow{d} N\left(0, \sigma^2_{m,2}(x)\right),$$

where $\xrightarrow{d}$ denotes the convergence in distribution, $\sigma^2_{m,1}(x) = \nu_0\sigma_1^2(x)f_b(x)$ and $\sigma^2_{m,2}(x) = \nu_0\sigma_2^2(x)f_a(x)$ with $\nu_j = \int u^{2j}K^2(u)du$ $(j \geq 0)$, $\sigma_1^2(x) = \mathbb{E}\left(u^2_{t+\tau}|\, X_t = x\right)$ for $t \leq T_1$ and $\sigma_2^2(x) = \mathbb{E}\left(u^2_{t+\tau}|\, X_t = x\right)$ for $t \geq T_1$, if the conditional variance of $u_{t+\tau}$ given $X_t = x$ has the same break date as the mean function. Also, it is not difficult to show that $\text{Cov}(C_1,\, C_2) \to 0$ as $T \to \infty$. Therefore, it follows from the Cramér-Wold device that

$$\begin{pmatrix} C_1 \\ C_2 \end{pmatrix} \xrightarrow{d} N\left(0, \Sigma_c(x)\right) \tag{9}$$

with $\Sigma_c(x) = \text{diag}\{\sigma^2_{m,1}(x), \sigma^2_{m,2}(x)\}$, which implies that $C_0(\gamma) \xrightarrow{d} N\left(0, \sigma^2_{m,0}(x)\right)$, where $\sigma^2_{m,0}(x) = \nu_0\left[s_0\gamma^2h_c^2\sigma_1^2(x)f_b(x) + (1-s_0)\sigma_2^2(x)f_a(x)\right]$. Hence, we have the following the asymptotic normality for $\widehat{m}_{\rm wll}(x)$ with its detailed discussions given in Section 5.

Theorem 1. *Suppose that Conditions A - B hold. Then, as $T \to \infty$,*

$$\sqrt{Th_2}\left[\widehat{m}_{\rm wll}(x) - m_{(2)}(x) - B_{\rm wll}(x) + o_p(h_1^2 + h_2^2)\right] \xrightarrow{d} N\left(0, \sigma^2_{\rm wll}(x)\right), \tag{10}$$

where $\sigma^2_{\rm wll}(x) = \sigma^2_{m,0}(x)[\gamma\,s_0f_b(x) + (1-s_0)f_a(x)]^{-2}$, which is regarded as the asymptotic variance of $\widehat{m}_{\rm wll}(x)$.

If there is no break in the variance function; that is, $\sigma^2(x) = \mathbb{E}\left(u^2_{t+\tau}|\, X_t = x\right) = \sigma_1^2(x) = \sigma_2^2(x)$, then, it is reduced to $\sigma^2_{\rm wll}(x) = \nu_0\,s_{\rm wll}\sigma^2(x)/f_a(x)$, where $s_{\rm wll} = [\gamma^2s_0h_c^2/\delta(x) + (1-s_0)]/[\gamma\,s_0/\delta(x) + (1-s_0)]^2$. By the same token, it is not difficult to derive the asymptotic variance of $\widehat{m}_{(2)}(x)$, which is $\sigma^2_{(2)}(x) = \nu_0 s_{(2)}\sigma^2(x)/f_a(x)$, where $s_{(2)} = 1/[1-s_0]$. Evidently,

$s_{\text{wll}} < s_{(2)}$ so that the asymptotic variance for $\widehat{m}_{\text{wll}}(x)$ is smaller than that for $\widehat{m}_{(2)}(x)$ in the asymptotic sense. Note that when γ is consistently estimated, denoted by $\hat{\gamma}$, we still have

$$C_0(\hat{\gamma}) = C_0(\gamma) + (\hat{\gamma} - \gamma)\sqrt{s_0}\, C_1 = C_0(\gamma) + o_p(1) \ \xrightarrow{d}\ N\left(0, \sigma^2_{m,0}(x)\right)$$

by Slutsky theorem and (9) and (10), where $\sigma^2_{m,0}(x)$ is defined in (9), which indicates that the asymptotic normality for $\widehat{m}_{\text{wll}}(x)$ is the same for both known γ and the consistent estimate $\hat{\gamma}$, as long as γ can be consistently estimated (see Section 2.4).

Finally, it is clear from (8) and (10) that the mean squared error (MSE) of $\widehat{m}_{\text{wll}}(x)$ is given by

$$\text{MSE}\left(\widehat{m}_{\text{wll}}(x)\right) = B^2_{\text{wll}}(x) + \sigma^2_{wll}(x)/(Th_2), \tag{11}$$

where the asymptotic bias term $B_{\text{wll}}(x)$ is given in (8) and the asymptotic variance term $\sigma^2_{wll}(x)$ can be found in (10), which provides a criterion for choosing the optimal bandwidths and γ simultaneously, described as follows. Therefore, (11) provides a formulation to balance the bias-variance trade-off.

2.3.3 Bandwidth Selection

Various existing bandwidth selection techniques for nonparametric regression can be adapted for the above estimation; see, e.g., Fan and Gijbels (1996) and Fan and Yao (2003). But, as pointed out by Shao (1993) and Cai et al. (2000), the conventional leave-one-out cross-validation method might fail for time series data, since adjacent points might be highly dependent. Therefore, we adapt a simple and quick method proposed by Cai et al. (2000) to select bandwidth h_1 and h_2, described below. It can be regarded as a modified multifold forward-validation criterion that is attentive to the structure of stationary time series data.

To choose the optimal bandwidths $\{h_i\}_{i=1,2}$ from the data, we describe the procedure in detail. For simplicity, our focus here is on choosing $\hat{h}_1$ in a data-driven fashion. To this end, let m and Q be two given positive integers such that $T_1 > mQ$. The idea is first to use Q sub-series of lengths $T_1 - qm$ $(q = 1, \ldots, Q)$ to estimate the unknown mean functions and then compute the one-step forecasting errors of the next section of the time series of length m based on the estimated models. More precisely, we choose the optimal bandwidth that minimize the following AMS error

$$\text{AMS}(h_1) = \frac{1}{Qm} \sum_{q=1}^{Q} \sum_{t=T_1-\tau-qm+1}^{T_1-\tau-qm+m} \left[Y_{t+\tau} - \widehat{m}^{[-q]}(X_t)\right]^2, \tag{12}$$

where $\{\widehat{m}^{[-q]}(\cdot)\}$ is the local linear mean estimate from the sample $\{(Y_{t+\tau}, X_t), 1 \leq t \leq T_i - \tau - qm\}$ with $i = 1$ here. Ten candidate values for each bandwidth are chosen to be equidistant within the range $[10^{-2}, 10] \cdot \tilde{h}_1$ to find the optimal h_1, denoted by $\hat{h}_1$, where $\tilde{h}_1$ represents an initial bandwidth for the first subsample under Gaussian kernel. Note that the theoretically optimal bandwidth $h_{1,opt} \propto T_1^{-1/(4+d)}$, where d represents the dimension of covariates. By the same token, we can choose $\hat{h}_2$ using a similar procedure.

2.4 Consistency of the Multifold Forward-Validation

Now, we choose the optimal weight γ for the pre-break data. Similar to (12), it is to minimize the following empirical MSE based on (11) over the post-break period, for given $\hat{h}_1$ and $\hat{h}_2$ from the above. To choose the optimal γ in $\widehat{m}_{\text{wll}}(X_t)$, we propose a novel multifold forward-validation criterion as follows:

$$\text{MFV}(\gamma) = \frac{1}{Qm} \sum_{q=1}^{Q} \sum_{t=T-\tau-qm+1}^{T-\tau-qm+m} \left[Y_{t+\tau} - \widetilde{m}^{[-q]}(X_t)\right]^2 = \frac{1}{Qm} ||\widetilde{\mathbf{Y}} - \widetilde{\mathbf{m}}_{\text{wll}}(\gamma)||^2, \qquad (13)$$

where $\widetilde{\mathbf{Y}} = (Y_{T-Qm+1}, \cdots, Y_T)^\top$, $\widetilde{\mathbf{m}}_{\text{wll}}(\gamma) = \left(\widetilde{m}^{[-Q]}\left(X_{T-\tau-Qm+1}\right), \cdots, \widetilde{m}^{[-1]}(X_{T-\tau})\right)^\top$, and $||\cdot||$ is the Euclidean norm. We use the multifold forward-validation criterion to select the weight γ as follows:

$$\widehat{\gamma} = \arg\min_{\gamma \in \mathcal{H}} \text{MFV}(\gamma),$$

where $\mathcal{H} = [0, 1]$. Then, the τ-step-ahead MFVMA prediction of $Y_{T+\tau}$ is

$$\widehat{Y}_{T+\tau}(\widehat{\gamma}) \equiv \widehat{m}_{\text{wll}}(X_T) = \mathbf{e}^\top \widehat{\Gamma} \widehat{\beta}_{(1)}(X_T) + \mathbf{e}^\top (\mathbf{I}_2 - \widehat{\Gamma}) \widehat{\beta}_{(2)}(X_T),$$

where $\widehat{\Gamma} = \widehat{\gamma}\left[\mathbf{X}^\top \mathbf{W}(\widehat{\gamma}) \mathbf{X}\right]^{-1} \left(X_{(1)}^\top \mathbf{W}_{(1)} X_{(1)}\right)$.

To evaluate the performance of the MFVMA method, we consider the following quadratic prediction risk function

$$R_{T+\tau}(\gamma) = \mathbb{E}\left\{Y_{T+\tau} - \widehat{Y}_{T+\tau}(\gamma)\right\}^2 - \sigma_{T+\tau}^2,$$

where $\sigma_{t+\tau}^2$=Var$(u_{t+\tau})$ denotes the variance of $u_{t+\tau}$. Intuitively, one would aim to select the model weight γ to minimize the out-of-sample prediction risk function $R_{T+\tau}(\gamma)$ subject to the constraint $0 \leq \gamma \leq 1$. However, this is infeasible due to the expectation depending on the unknown conditional probability density function. Instead of directly minimizing the infeasible risk $R_{T+\tau}(\gamma)$, we select data-driven weights by minimizing the multi-fold forward-validation criterion. We will demonstrate the asymptotic optimality in the sense that the out-of-sample prediction achieves the lowest possible prediction risk as the sample size approaches infinity. For this purpose, define $R^*_{T+\tau}(\gamma) = \mathbb{E}[Y_{T+\tau} - Y^*_{T+\tau}(\gamma)]^2 - \sigma^2_{T+\tau}$, and $\xi^*_{T+\tau} = \inf_{\gamma\in\mathcal{H}} R^*_{T+\tau}(\gamma)$, where $Y^*_{T+\tau}(\gamma) = m^*_{\text{wll}}(X_T)$, $m^*_{\text{wll}}(x) = s_b\beta^*_{0,(1)}(x) + (1-s_b)\beta^*_{0,(2)}(x)$, where $\beta^*_{0,(1)}(x)$ and $\beta^*_{0,(2)}(x)$ are well-defined limits of $\widehat{\beta}_{0,(1)}(x)$ and $\widehat{\beta}_{0,(2)}(x)$ for any given x. We state the requisite conditions for asymptotic optimality, wherein all limiting behaviors are considered as the sample size T tends to infinity.

Condition C:

C1. Assume that $\xi^{*-1}_{T+\tau}\sup_{\gamma\in\mathcal{H}}\{[Y_{T+\tau} - \widehat{Y}_{T+\tau}(\gamma)]^2 - [Y_{T+\tau} - Y^*_{T+\tau}(\gamma)]^2\}\}^2$ is uniformly integrable.

C2. For any given x, $Qm = O(Th_2)$, $T^{-1/2}h_1^{-1/2}\xi^{*-1}_{T+\tau} = o(1)$, $T^{-1/2}h_2^{-1/2}\xi^{*-1}_{T+\tau} = o(1)$, $h_1^2\xi^{*-1}_{T+\tau} = o(1)$, and $h_2^2\xi^{*-1}_{T+\tau} = o(1)$.

C3. The fourth moment of $Y_{t+\tau}$ exists and so, do X_t and $u_{t+\tau}$.

Condition C is a mild technical condition that is commonly employed in the model averaging literature. Specifically, Condition C1 aligns with Condition 7 in Hu and Zhang (2023). Condition C2 elucidates the relationships among $\xi^*_{T+\tau}$, h_1, h_2, and T. Analogous conditions in the literature include Condition 7 of Ando and Li (2014), Condition C.6 of Zhang, Yu, Zou, and Liang (2016), and Condition C.6 of Sun et al. (2023). Condition C3 represents the regularity conditions of the central limit theorem for dependent processes, which is similar to Assumption 4 in Zhang and Liu (2023).

Theorem 2. *Suppose that Conditions A - C hold. Then, as $T \to \infty$,*

$$\frac{R_{T+\tau}(\widehat{\gamma})}{\inf_{\gamma\in\mathcal{H}} R_{T+\tau}(\gamma)} \longrightarrow 1$$

in probability for any $\tau \geq 0$.

Theorem 2 demonstrates that the proposed model averaging prediction attains asymptotic optimality in the sense of realizing the minimum attainable out-of-sample prediction risk. However, in contrast to most existing works that establish asymptotic optimality based on an in-sample squared error loss function, such as Hansen (2007), Wan, Zhang, and Zou (2010), Lee et al. (2022a), and Racine, Li, Yu, and Zheng (2023), the proposed procedure is constructed by utilizing multifold historical data sets, and the asymptotic optimality is established based on the out-of-sample prediction risk function, rendering it more applicable to model averaging for predictive purposes. It is noteworthy that our result of asymptotic optimality holds irrespective of whether the correct models are included in the candidate models with known break dates and bandwidths.

2.5 Practical Implementations

2.5.1 Estimation of Break Date

When the break date T_1 is unknown, it can be estimated using the method proposed by Mohr and Selk (2020). The objective is to estimate the rescaled change point s_0. The estimator itself is based on a Kolmogorov-Smirnov functional of the marked empirical process of residuals; that is

$$\hat{\mathcal{T}}_T(s,z) = \frac{1}{T}\sum_{t=1}^{\lfloor Ts \rfloor} \left(Y_{t+\tau} - \hat{m}_T(X_t)\right)\omega_T(X_t)\mathbb{1}\left(X_t \le z\right)$$

for $s \in [0,1]$, where $x \le y$ is short for $x_j \le y_j$ for all $j = 1, \ldots, d$, $\omega_T(\bullet) = \mathbb{1}\{\bullet \in [-(\log T)^{\frac{1}{d+1}}, (\log T)^{\frac{1}{d+1}}]^d\}$ and for simplicity, $\hat{m}_T(\cdot)$ is the Nadaraya-Watson estimator[6], namely

$$\hat{m}_T(x) = \frac{\sum_{t=1}^{T-\tau} K_h(x - X_t)Y_t}{\sum_{t=1}^{T-\tau} K_h(x - X_t)}.$$

The truncation of the domain of X_t to a compact set within $\mathbb{R}$ by the function $\omega_T(\bullet)$ is motivated by the fact that kernel estimators only perform well in regions where there are many observations and rather poorly on the edges and outside of the sample space. Therefore, the nice asymptotic properties cannot be expected on the whole domain of $\mathbb{R}^d$. Then, s_0 is estimated by

$$\hat{s}_T := \min\left\{ s : \sup_{z\in\mathbb{R}} |\hat{\mathcal{T}}_T(s,z)| = \sup_{\bar{s}\in[0,1]} \sup_{z\in\mathbb{R}} |\hat{\mathcal{T}}_T(\bar{s},z)| \right\}. \tag{14}$$

Note that $\hat{s}_T = \lfloor T\hat{s}_T \rfloor / T$. Under some regularity conditions; see, for instance, Assumptions I - TX.2 in Mohr and Selk (2020), it follows from Mohr and Selk (2020) that $\hat{s}_T$ is a consistent estimate of s_0 with the convergence rate T. The reader is referred to the paper by Mohr and Selk (2020) for details. Therefore, $\hat{s}_T$ in (14) is used in our simulation and empirical studies conducted in Sections and 3 and 4, respectively.

[6] Of course, one can use the local linear fitting scheme.

2.5.2 Extension to Multiple Breaks

The main focus in the previous subsections is on the case of having a single break. However, in practice a time series model may be subject to multiple breaks. The case of multiple breaks is a straightforward extension of the previous sections. The weighted local linear estimator can be similarly defined as the combination of the full-sample estimator and the estimator using observations after the most recent break point, described below. For example, consider a nonparametric model in (1) with two breaks (three periods) so that (2) can be generalized to the following

$$m_t(x) = m_{(1)}(x)\mathbb{1}\,(t \leq S_1) + m_{(2)}(x)\mathbb{1}\,(S_1 < t \leq S_2) + m_{(3)}(x)\mathbb{1}\,(t > S_2),$$

where two break points are at S_1 and S_2 with $1 < S_1 < S_2 < T$. Similar to the estimation procedure as in (7), for simplicity, by following the same idea in (7) (see the last equation in (7)), we adopt the following combined local linear estimator

$$\widehat{m}_{\mathrm{wll},2}(x) = \theta_2\,\widehat{m}_{\mathrm{full}}(x) + (1-\theta_2)\,\widehat{m}_{(3)}(x), \tag{15}$$

where $0 \leq \theta_2 \leq 1$ is the weight, similar to Θ in (7), $\widehat{m}_{\mathrm{full}}(x)$ is the local linear estimator based on the full sample, and $\widehat{m}_{(3)}(x)$ is the local linear estimator based on the observations from the last period ($S_2 < t \leq T$). Similar to the asymptotic analyses presented in Section 2.3, it is not difficult to obtain the asymptotic properties for $\widehat{m}_{\mathrm{wll},2}(x)$. By the same token, one can get a consistent estimate of θ_2 by following the similar procedure outlined in (13) in Section 2.4. Finally, note that the theoretical derivations for $\widehat{m}_{\mathrm{wll},2}(x)$ similar to those for $\widehat{m}_{\mathrm{wll}}(x)$ for single break case, so omitted and available upon request.

For a model with two breaks, other combined estimators are possible, for example, the combination of the full-sample estimator and the two subsample estimators based on the second and third periods. However, this subsample estimator is not consistent for $m_{(3)}(x)$. Also, because the full-sample estimator is the most efficient one, the efficiency of the combined estimator cannot be enhanced by combining with this inconsistent subsample estimator using the second and third subsamples. Therefore, this combined estimator does not balance the trade-off between the bias and variance efficiency. For more discussions, the reader is referred to the paper by Lee et al. (2022b) for linear models. Following the same idea as in (15), it is not difficult to extend to a nonparametric model with three or more than three breaks.

3 Monte Carlo Simulation Studies

In order to evaluate the finite sample performance of our proposed estimator, we consider three data generating processes; that is

(IID) $Y_{t+1} = m_t(X_t) + \sigma_t \varepsilon_t$, where $\varepsilon_t \sim \mathcal{N}(0,1)$ i.i.d.

(TS) $Y_{t+1} = m_t(X_t) + \sigma_t \varepsilon_t$, where $\varepsilon_t \sim \mathcal{N}(0,1)$ i.i.d.

(AR) $Y_{t+1} = m_t(Y_t) + \sigma_t \varepsilon_t$, where $\varepsilon_t \sim \mathcal{N}(0,1)$ i.i.d.

In this study, we simulate distinct distributions for X_t before and after the structural break. In the IID case, we generate samples $X_t \sim \mathcal{N}(0, \sqrt{0.1})$ for the pre-break period $1 \le t \le T_1$, and $X_t \sim \mathcal{N}(1, \sqrt{0.5})$ for the post-break period $T_1 + 1 \le t \le T$. In the TS case, we generate samples from the following distributions for X_t: $X_t = 0.4X_{t-1} + \upsilon_{1,t}$ for $1 \le t \le T_1$ and $X_t = 0.5X_{t-1} + \upsilon_{2,t}$ for $T_1 + 1 \le t \le T$, where $\upsilon_{1,t} \sim \mathcal{N}(0, \sqrt{0.1})$ and $\upsilon_{2,t} \sim \mathcal{N}(1, \sqrt{0.5})$ iid. Further, we introduce a break in variance and a shift in distribution of the error term such that $\sigma_t \varepsilon_t = \sqrt{0.1}\varepsilon_{1,t} \cdot \mathbb{1}(t \le T_1) + \sqrt{0.2}\varepsilon_{2,t} \cdot \mathbb{1}(t > T_1)$, where both $\varepsilon_{1,t} \sim \mathcal{N}(0, \sqrt{0.1})$ and $\varepsilon_{2,t} \sim \mathcal{N}(1, \sqrt{0.5})$ are independently and identically distributed. The mean function is modeled as follows

$$m_t(x) = \sin(x)\mathbb{1}(t \le T_1) + (1-b)\sin(x)\mathbb{1}(t > T_1),$$

where b takes four values as 0.1, 0.3, 0.6, and 1, so that the break size function $\lambda(x) = b\sin(x)$ is characterized by b. The pre-break sample size is defined as a proportion of the full-sample, $T_1 = \lfloor Ts_0 \rfloor$ with $s_0 \in \{0.2, 0.5, 0.8\}$, with sample sizes of $T \in \{500, 1000\}$. The simulation is repeated $M = 1000$ times.

We shall evaluate whether the size of the break in both the mean and variance influences the forecasting performance of our proposed estimator. We distinguish the cases when s_0 is known, or unknown and estimated by $\hat{s}_T$ using (14). We use the Gaussian kernel for estimating the mean function $\widehat{m}(\cdot)$, together with the bandwidth $\{h_i\}_{i=1,2}$ and the weight γ determined by multifold forward-validation as described in Sections 2.3.3 and 2.4, respectively. Note that for simplicity, we adopt $m = [0.1T_i]$ and $Q = 4$ for $\{\hat{h}_i\}_{i=1,2}$ as recommended in Cai et al. (2000).

In order to evaluate forecasting performance, we employ the mean squared forecasting error of one-step-ahead forecast by comparing our weighted local linear estimator ("wll") to post-break estimator ("pb") as well as full sample estimator ("fs"). One-step ahead forecast for Y_t computed at time T using method i is denoted as $\widehat{Y}_{i,T+1}$, and i = wll, pb, or fs. In this simulation exercise, these forecasts are conditional on X_{T+1}, or precisely

$$\widehat{Y}_{\mathrm{wll},T+1} = \widehat{m}_{\mathrm{wll,c}}(X_T),$$

where $\widehat{m}_{\mathrm{wll,c}}(\cdot)$ is computed using (9), while $\widehat{Y}_{\mathrm{pb},T+1}$ is based on local linear estimator using post-break observations only, and $\widehat{Y}_{\mathrm{fs},T+1}$ uses the entire sample. In the case of a known s_0, we use the date T_1 as the break date. In the case of an estimated s_0, we use the estimated break date $\hat{T}_1 = \lfloor T\hat{s}_T \rfloor$ to determine the post-break sample for both the post-break and weighted local linear estimators. Further, We use a fixed estimation window from $t = 1, \dots, T$. The MSFE for each method is calculated as

$$\mathrm{MSFE}_i = \frac{1}{1000}\sum_{m=1}^{1000}\left(Y_{i,T+1}^{(m)} - \widehat{Y}_{i,T+1}^{(m)}\right)^2,$$

where $\widehat{Y}_{i,T+1}^{(m)}$ is the forecasted value for Y_{T+1} computed using method i for the m-th replication.

Tables 1, 2, and 3 display simulation results for $\mathrm{MSFE}_1/\mathrm{MSFE}_2$ and $\mathrm{MSFE}_3/\mathrm{MSFE}_2$ for the IID, TS, and AR data generating process scenarios, respectively. Across all scenarios,

Table 1: MSFE for weighted local linear (WLL) and full-sample (FS) estimator relative to the post-break estimator in the IID data generating process scenario. Sample size $T = 500$ (the left panel) and $T = 1000$ (the right panel) with $M = 1000$ Monte-Carlo replications.

s_0	b	$T = 500$				$T = 1000$			
		s_0 known		s_0 estimated		s_0 known		s_0 estimated	
		WLL	FS	WLL	FS	WLL	FS	WLL	FS
0.2	0.1	0.953	1.762	0.967	3.724	0.969	1.121	0.957	7.939
	0.3	0.956	1.418	0.949	1.476	0.958	1.465	0.946	2.290
	0.6	0.957	1.217	0.952	1.974	0.949	1.100	0.927	11.590
	1.0	0.951	1.399	0.919	1.559	0.956	1.345	0.942	1.252
0.5	0.1	0.950	1.383	0.950	2.517	0.966	1.201	0.948	2.632
	0.3	0.961	1.377	0.959	3.086	0.961	1.851	0.956	1.923
	0.6	0.959	3.163	0.958	1.576	0.946	10.302	0.947	4.896
	1.0	0.957	2.386	0.947	71.359	0.949	1.34e4	0.941	7.229
0.8	0.1	0.954	97.425	0.935	6.544	0.939	32.810	0.945	2.579
	0.3	0.950	33.227	0.955	1.942	0.955	4.239	0.952	1.909
	0.6	0.963	4.116	0.940	8.829	0.941	2.48e3	0.947	8.561
	1.0	0.953	35.396	0.939	3.887	0.941	13.891	0.949	3.929

we observe that our proposed WLL estimator consistently outperforms the conventional post-break estimator, as evidenced by relative MSFEs less than 1. This demonstrates that

Table 2: MSFE for weighted local linear (WLL) and full-sample (FS) estimator relative to the post-break estimator in the TS data generating process scenario. Sample size $T = 500$ (the left panel) and $T = 1000$ (the right panel) with $M = 1000$ Monte-Carlo replications.

s_0	b	$T = 500$				$T = 1000$			
		s_0 known		s_0 estimated		s_0 known		s_0 estimated	
		WLL	FS	WLL	FS	WLL	FS	WLL	FS
0.2	0.1	0.973	1.029	0.958	1.534	0.985	1.000	0.972	1.050
	0.3	0.974	0.991	0.961	1.088	0.985	1.007	0.973	1.171
	0.6	0.976	1.476	0.941	3.112	0.981	1.018	0.959	1.076
	1.0	0.978	1.185	0.963	1.869	0.978	1.011	0.971	1.290
0.5	0.1	0.962	1.081	0.939	1.316	0.971	1.189	0.963	1.199
	0.3	0.963	1.100	0.958	1.436	0.966	1.024	0.968	2.130
	0.6	0.953	1.056	0.964	1.301	0.966	1.111	0.977	3.324
	1.0	0.947	1.151	0.970	14.310	0.962	1.882	0.969	2.883
0.8	0.1	0.959	1.210	0.953	1.408	0.958	1.273	0.959	1.299
	0.3	0.959	1.372	0.952	1.459	0.962	1.172	0.961	1.111
	0.6	0.967	1.111	0.955	1.106	0.967	1.301	0.969	415.13
	1.0	0.965	1.932	0.964	7.422	0.969	1.959	0.970	2.318

the WLL estimator successfully improves the forecast by taking into account the pre-break observations using an optimal weight, rather than relying solely on post-break observations for the forecast. On the other hand, we also observe that the FS estimator yields relative MSFE that is far greater than 1, which means that ignoring structural breaks, using full sample observations for forecast may be unstable and lead to severe bias.

4 An Empirical Example

Volatility forecasting has become a very prominent area of research during the last few decades and several authors have come out with path breaking studies in this area which have helped both the academicians and practitioners in the financial market. There is a vast amount of literature on forecasting volatility with structural breaks; see, for example, the paper by Karlsson (2016) for details. In this empirical example, by following the literature on volatility forecasting with structural break, our goal is to forecast volatility using return

Table 3: MSFE for weighted local linear (WLL) and full-sample (FS) estimator relative to the post-break estimator in the AR data generating process scenario. Sample size $T = 500$ (the left panel) and $T = 1000$ (the right panel) with $M = 1000$ Monte-Carlo replications.

s_0	b	$T = 500$				$T = 1000$			
		s_0 known		s_0 estimated		s_0 known		s_0 estimated	
		WLL	FS	WLL	FS	WLL	FS	WLL	FS
0.2	0.1	0.947	85.960	0.964	3.275	0.949	10.679	0.922	2.658
	0.3	0.937	3.883	0.914	18.228	0.946	101.475	0.927	17.694
	0.6	0.957	2.076	0.944	2.808	0.961	2.127	0.939	2.430
	1.0	0.949	107.269	0.951	4.593	0.966	3.525	0.952	57.703
0.5	0.1	0.942	5.737	0.888	8.383	0.946	10.070	0.948	94.249
	0.3	0.944	21.129	0.929	4.144	0.944	5.916	0.940	30.545
	0.6	0.935	11.995	0.912	7.358	0.954	5.663	0.925	2.857
	1.0	0.960	6.019	0.927	39.593	0.941	5.597	0.917	8.193
0.8	0.1	0.934	20.001	0.972	9.508	0.910	3.69e7	0.922	31.787
	0.3	0.951	22.659	0.954	9.669	0.956	16.597	0.936	11.176
	0.6	0.942	34.159	0.985	28.637	0.956	11.205	0.959	1.44e3
	1.0	0.924	16.167	0.921	18.951	0.943	41.608	0.949	78.915

data. To this end, we consider the following forecasting model with structural break

$$\mathrm{VIX}_{t+1} = m_1(\mathrm{Ret}_t)\mathbb{1}(t \leq T_1) + m_2(\mathrm{Ret}_t)\mathbb{1}(t > T_1) + \varepsilon_t, \tag{16}$$

where VIX_t denotes the volatility index VIX, Ret_t represents the log daily return of the S&P 500 index, T_1 is the break point, and ε_t is the idiosyncratic error term. The variable VIX originates from the Chicago Board Options Exchange's CBOE Volatility Index, sourced from Yahoo Finance. Data on S&P 500 index were obtained from Federal Reserves Economic Data (FRED). The sample period spans from January 1, 2020 to June 30, 2024, encompassing a total of 1,129 daily data points after excluding non-trading days and calculating the log returns. Figure 1 displays a scatterplot of VIX_t against Ret_t. Visual inspection suggests a nonlinear relationship, better captured by locally weighted scatterplot smoothing (LOESS, shown as orange dots) than by ordinary least squares (OLS, represented by the yellow line). To substantiate this observation, we conduct forecasting of VIX_t using various methods from the literature.

We iteratively calculate forecasts using the local linear postbreak estimator with multifold forward-validation as a benchmark, along nine other alternative methods, which are depicted

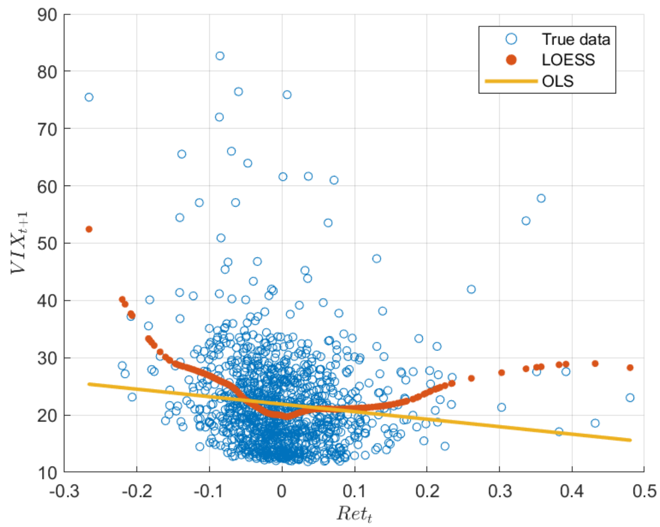


Figure 1: Scatterplot of VIX_{t+1} against Ret_t

.

in Table 4. The dataset is split into two segments: the initial T observations constitute the in-

Table 4: Forecasting methods used in the empirical analysis

No.	Method	Description
1.	WLL	Bias-corrected weighted local linear estimator
2.	PBLL	Local linear postbreak estimator
3.	FSLL	Local linear full sample estimator
4.	PBOLS	OLS postbreak estimator
5.	FSOLS	OLS full sample estimator
6.	PPP	Linear regression using estimated optimal weights (Pesaran et al., 2013)
7.	optW ($\underline{b} = 0, \bar{b} = 1$)	Linear regression using robust optimal weights (Pesaran et al., 2013)
8.	optWd	Linear regression using optimal window (Pesaran and Timmermann, 2007)
9.	aveW	Forecast averaging across estimation windows (Pesaran and Pick, 2011)

sample estimation period, and the remaining observations serve as the pseudo out-of-sample evaluation period. Forecasts are generated step by step during the out-of-sample period, using only the information available at each forecast point. As we widen the estimation window, we re-estimate all model parameters, such as the break points and the optimal weight in the case of our WLL estimator. Subsequently, we evaluate the forecasts for horizons $\tau = 1, 2, 3, 4$, and 5 days using the Diebold-Mariano (DM) test proposed by Diebold and Mariano (1995). The out-of-sample analysis covers the period between June 1, 2023 and

June 30, 2024.

The DM test is considerably more versatile than any alternative test of equality of forecast performance, and is likely to be widely used in empirical evaluation studies. However, the test was found to be quite seriously over-sized for moderate numbers of sample observations. In addition, the long-run variance can frequently be negative when computing standard DM tests as argued by Harvey, Leybourne, and Newbold (1997) and Harvey, Leybourne, and Whitehouse (2017). Therefore, we use a modified version of the DM test in the following. Let $e_{i,t} = Y_t - \widehat{Y}_{i,t}$ and $e_{j,t} = Y_t - \widehat{Y}_{j,t}$ be the forecast errors for method i and j, respectively, and choose the loss differential $d_t = e_{i,t}^2 - e_{j,t}^2$. Denote $\bar{d} = T^{-1}\sum_{t=1}^{T} d_t$ as the the sample mean of the loss differential, or simply $\mathrm{MSFE}_i - \mathrm{MSFE}_j$, and ω^2 the long-run variance of d_t, i.e., $\omega^2 = \sum_{j=-\infty}^{\infty} \Upsilon_j$, with $\Upsilon_j = \mathrm{Cov}\left(d_t, d_{t-j}\right)$. Then, the modified Diebold-Mariano (MDM) test is defined as follows

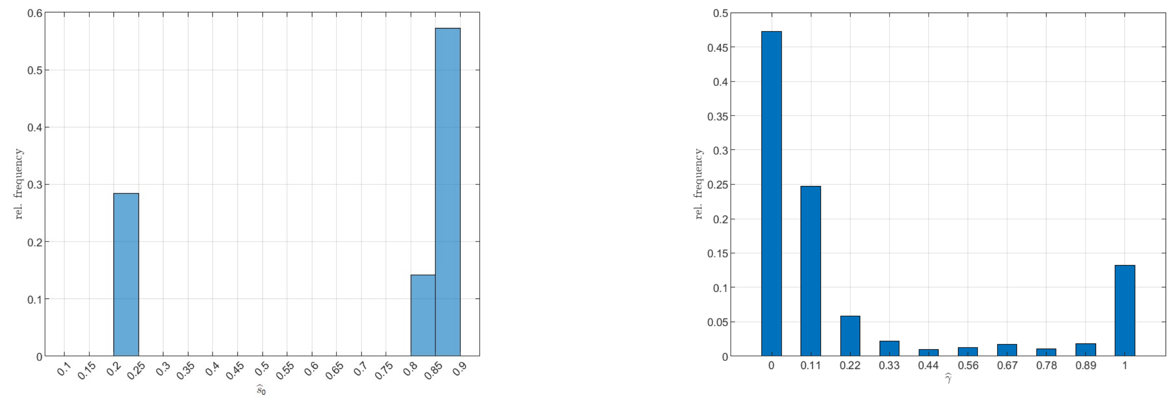
$$\mathrm{MDM} = \begin{cases} \sqrt{T + 1 - 2h + T^{-1}h(h-1)}\left(\frac{\bar{d}}{\hat{\omega}}\right) & \text{if } \hat{\omega} > 0 \\ \sqrt{T}\left(\frac{\bar{d}}{\hat{\omega}_{\text{Bart}}}\right) & \text{otherwise} \end{cases},$$

where $\hat{\omega}^2 = \hat{\gamma}_0 + 2\sum_{j=1}^{\tau-1} \hat{\Upsilon}_j$ and $\hat{\Upsilon}_j = T^{-1}\sum_{t=j+1}^{T}\left(d_t - \bar{d}\right)\left(d_{t-j} - \bar{d}\right)$ is the associated sample autocovariance. The critical values are computed from Student's distribution t_{T-1}. The formula for $\hat{\omega}^2$ makes use of a long-run variance estimator, which is a weighted sum of $\tau - 1$ lags of sample auto-covariances. This approach is motivated by the fact that optimal τ-step-ahead forecast errors are at most $(\tau-1)$-dependent. The magnitude, however, can take a negative value. In such cases, we opt for a Bartlett long variance estimator, defined as follows:

$$\hat{\omega}_{\text{Bart}}^2 = \hat{\Upsilon}_0 + 2\sum_{j=1}^{\tau-1}\left(1 - \frac{j}{\tau}\right)\hat{\Upsilon}_j.$$

To assess the statistical significance of the improved predictive performance achieved by method j, we conduct a hypothesis test comparing it to method i, where method i serves as the benchmark estimator. The null hypothesis (H_0) asserts that there is no significant difference in MSFE between the two methods, specifically $H_0 : \mathrm{MSFE}_i = \mathrm{MSFE}_j$. In contrast, the alternative hypothesis (H_a) posits that method j outperforms method i, i.e., $H_a : \mathrm{MSFE}_i > \mathrm{MSFE}_j$.

Figures 2a and 2b show the estimated rescaled break date $\widehat{s}_0$ and the estimated weights $\widehat{\gamma}$, respectively. Based on these figures, we conclude that the break date is estimated to lie at the 40th percentile of the data around 70% of the time, while the rest is around the 10th percentile of the data. The estimated optimal weights for WLL estimator appear to be bi-modally distributed, with around 47% of the time the estimated optimal weight being zero. Around 25% of the time, we have $\widehat{\gamma} = 0.11$, which means we put relatively low weight on the prebreak sample to improve the forecast. Around 14% of the time, we have $\widehat{\gamma} = 1.0$, which means we put full weight on the prebreak sample, thus run a full-sample estimation. Table 5 displays the MSFE of different forecasting methods outlined in Table



(a) Estimated rescaled breakdate $\widehat{s}_0$

(b) Estimated weight $\widehat{\gamma}$

Figure 2: Distribution of estimated WLL parameters of model (16).

4, calculated at forecast horizon $\tau = 1, 2, 3, 4$, and 5 day. Compared to the benchmark,

Table 5: MSFE of estimators described in Table 4 at forecast horizon $\tau = 1, 2, 3, 4$ and 5 day. ***,**, and * indicate significance of DM test at 1%, 5%, and 10% level, respectively. Benchmark is local linear postbreak estimator (PBLL) using multifold forward-validation for the tuning parameters.

τ	WLL	PBLL	FSLL	PBOLS	FSOLS	PPP	optW	optWd	aveW
1	19.950***	41.863	68.620	44.144	75.655	214.417	215.649	214.132	33.505***
2	19.846***	41.945	68.393	44.354	76.065	214.240	215.330	213.893	33.845***
3	19.834***	42.409	68.414	44.514	76.401	214.173	214.888	213.769	34.135***
4	19.573***	42.748	68.988	44.600	76.686	213.791	214.606	213.442	34.441***
5	19.357***	42.836	68.807	44.583	77.001	213.505	214.687	213.204	34.688**

the WLL, and aveW estimators yield significantly lower MSFEs across all forecast horizons. Our WLL estimator yields the lowest MSFEs, followed by aveW. This indicates that our proposed estimator produced more accurate forecasts compared to other methods. We also observe that nonparametric local linear models (WLL, PBLL, FSLL) yield lower MSFEs compared to the linear models (PBOLS, FSOLS, PPP, optW, optWd). The aveW method, although linear in nature, proves to yield more accurate forecasts than the benchmark model, primarily due to averaging the produced forecasts across a number of estimation windows.

5 Theoretical Proofs

Proof of Theorem 1. The proof of Theorem 1 consists of two parts: the first part derives the asymptotic bias given in (8), and the second part investigates the asymptotic normality provided in (10).

Proof of (8): To establish (8), first, we need to show that under Condition A,

$$S_j(x) \quad \xrightarrow{p} \quad \mu_j\, f_b(x). \tag{17}$$

Indeed, it is easy to show that $E[S_j(x)] \to \mu_j\, f_b(x)$ and $T_1 h_1 \mathrm{Var}(S_j(x)) \to f_b(x)\nu_j$, by following the same idea as in the proof of Theorem 1 in Cai et al. (2000). Next, it is easy to see that in view of (17), the asymptotic bias term of $\hat{\beta}_{0,(1)}(x)$ can be asymptotically expressed as

$$\begin{aligned} B_1(x) &\approx \frac{1}{T_1}\sum_{t=1}^{T_1} K_{h_1}(X_t - x)\left\{m_{(1)}(X_t) - m_{(1)}(x) - m'_{(1)}(x)(X_t - x)\right\}/f_b(x) \\ &\approx \frac{m''_{(1)}(x)}{2}\frac{1}{T_1}\sum_{t=1}^{T_1} K_{h_1}(X_t - x)(X_t - x)^2/f_b(x) \approx m''_{(1)}(x)\mu_2 h_1^2/2 \end{aligned}$$

by Taylor expansion and following the same proof of (17). Similarly, $B_2(x)$, the asymptotic bias for $\hat{\beta}_{0,(2)}(x)$, can be obtained easily. Therefore, (8) is established.

Proof of (10): To establish (10), first, we show that $C_1 \xrightarrow{d} N\left(0, \sigma^2_{m,1}(x)\right)$ and $C_2 \xrightarrow{d} N\left(0, \sigma^2_{m,2}(x)\right)$. To this end, let $Z_t = K_{h_1}(X_t - x)u_{t+\tau}\sqrt{h_1/T_1}$. Then, $C_1 = \sum_{t=1}^{T_1} Z_t$. By following the same procedures as in the proof of Lemma A.1 in Cai et al. (2000), it is not difficult to show that under Conditions A and B, $\mathrm{Var}(C_1) \to \sigma^2_{m,1}(x)$ as $T_1 \to \infty$. To establish the asymptotic normality of C_1, we employ the small-block and large-block technique — namely, $C_1 = Q_l + Q_s + Q_r$, to show that Q_l, the sum of the large-blocks converges a normal distribution in distribution, Q_s, the sum of the small-blocks, can be ignored in probability, Q_r, the sum of the remainder terms, converges to zero in probability, and the large-blocks are asymptotically independent. Also, we prove that for Q_l, the Lindeberg's condition is satisfied. Then, by the Lindeberg's central limit theorem, the asymptotic normal of C_1 is established. By the same token, we can establish the asymptotic normality for C_2. Finally, by following the same steps as used in proving Lemma A.1 in Cai et al. (2000), it is easy to show that $\mathrm{Cov}(C_1, C_2) \to 0$ as $T \to \infty$. This completes the proof of (10). $\square$

Proof of Theorem 2. Note that

$$\begin{aligned}
\widetilde{m}^{[-q]}(x) &= \mathbf{e}^\top \widetilde{\Gamma}^{[-q]} \widetilde{\beta}^{[-q]}_{(1)}(x) + \mathbf{e}^\top (\mathbf{I}_2 - \widetilde{\Gamma}^{[-q]}) \widetilde{\beta}^{[-q]}_{(2)}(x), \\
&= s_b \widetilde{\beta}^{[-q]}_{0,(1)}(x) + (1 - s_b) \widetilde{\beta}^{[-q]}_{0,(2)}(x) + O_p(T^{-1/2} h_1^{-1/2} + T^{-1/2} h_2^{-1/2}),
\end{aligned}$$

where $\widetilde{\cdot}^{[-q]}$ represents the related estimators using the sample $\{(Y_{t+\tau}, X_t), 1 \le t \le T - \tau - qm\}$. Let $\mathrm{MFV}^*(\gamma) = \mathrm{MFV}(\gamma) - \sigma^2_{T+\tau}$, where $\sigma^2_{T+\tau}$ is a constant and unrelated to γ. The selected weight

$$\widehat{\gamma} = \arg\min_{\gamma \in \mathcal{H}} \mathrm{MFV}(\gamma) = \arg\min_{\gamma \in \mathcal{H}} \mathrm{MFV}^*(\gamma).$$

From Lemma1 of Gao, Zhang, Wang, Chong, and Zou (2019), Theorem 2 is valid if

$$\sup_{\gamma \in \mathcal{H}} \left| \frac{R_{T+\tau}(\gamma)}{R^*_{T+\tau}(\gamma)} - 1 \right| = o(1) \quad \text{and} \quad \sup_{\gamma \in \mathcal{H}} \left| \frac{\mathrm{MFV}^*(\gamma) - R^*_{T+\tau}(\gamma)}{R^*_{T+\tau}(\gamma)} \right| = o_p(1).$$

Based on Conditions A - B, for any given x, it is observed that

$$\max_{1 \le i \le 2} ||\widehat{\beta}_{0,(i)}(x) - \beta^*_{0,(i)}(x)|| = O_p(h_1^2 + h_2^2 + T^{-1/2} h_1^{-1/2} + T^{-1/2} h_2^{-1/2}).$$

Again, define $L(\gamma) = \left[Y_{T+\tau} - \widehat{Y}_{T+\tau}(\gamma)\right]^2 - \sigma^2_{T+\tau}$ and $L^*(\gamma) = \left[Y_{T+\tau} - Y^*_{T+\tau}(\gamma)\right]^2 - \sigma^2_{T+\tau}$. We have

$$\begin{aligned}
& \sup_{\gamma \in \mathcal{H}} |L(\gamma) - L^*(\gamma)| \\
= & \sup_{\gamma \in \mathcal{H}} \left\{ (Y^*_{T+\tau}(\gamma) - \widehat{Y}_{T+\tau}(\gamma))(2Y_{T+\tau} - \widehat{Y}_{T+\tau}(\gamma) - Y^*_{T+\tau}(\gamma)) \right\} \\
= & \sup_{\gamma \in \mathcal{H}} \left\{ [m^*_{\mathrm{wll}}(X_T) - \widehat{m}_{\mathrm{wll}}(X_T)](2Y_{T+\tau} - \widehat{Y}_{T+\tau}(\gamma) - Y^*_{T+\tau}(\gamma)) \right\} \\
= & \sup_{\gamma \in \mathcal{H}} \left\{ [m^*_{\mathrm{wll}}(X_T) - \widehat{m}_{\mathrm{wll}}(X_T)] \left[2(Y_{T+\tau} - Y^*_{T+\tau}(\gamma)) - (\widehat{Y}_{T+\tau}(\gamma) - Y^*_{T+\tau}(\gamma)) \right] \right\} \\
= & O_p(h_1^2 + h_2^2 + T^{-1/2} h_1^{-1/2} + T^{-1/2} h_2^{-1/2}).
\end{aligned}$$

Thus,

$$\begin{aligned}
&\sup_{\gamma\in\mathcal{H}} \frac{\left|R_{T+\tau}(\gamma) - R^*_{T+\tau}(\gamma)\right|}{R^*_{T+\tau}(\gamma)} \le \xi^{*-1}_{T+\tau} \sup_{\gamma\in\mathcal{H}} \left|R_{T+\tau}(\gamma) - R^*_{T+\tau}(\gamma)\right| \\
\le\ & \mathbb{E}\left\{\xi^{*-1}_{T+\tau} \sup_{\gamma\in\mathcal{H}} |L(\gamma) - L^*(\gamma)|\right\} = O(\xi^{*-1}_{T+\tau}(h_1^2 + h_2^2 + T^{-1/2}h_1^{-1/2} + T^{-1/2}h_2^{-1/2})),
\end{aligned}$$

where the second step follows from Condition C1 and the last step is due to Condition C2.

Next, define $\mathbf{m}^*(\gamma) = (m^*_{\mathrm{wll}}(X_{T-\tau-Qm+1}), \cdots, m^*_{\mathrm{wll}}(X_{T-\tau}))'$ and $\mathbf{m}(\gamma) = (m(X_{T-\tau-Qm+1}), \cdots, m(X_{T-\tau}))'$. Then, we observe that

$$\begin{aligned}
& \sup_{\gamma\in\mathcal{H}} \left|\frac{\mathrm{MFV}^*(\gamma) - R^*_{T+\tau}(\gamma)}{R^*_{T+\tau}(\gamma)}\right| \\
=\ & \xi^{*-1}_{T+\tau} \sup_{\gamma\in\mathcal{H}} \left|Q^{-1}m^{-1}(\widetilde{\mathbf{Y}} - \widetilde{\mathbf{m}}(\gamma))'(\widetilde{\mathbf{Y}} - \widetilde{\mathbf{m}}(\gamma)) - \sigma^2_{T+\tau} - \mathbb{E}L^*(\gamma)\right| \\
\le\ & \xi^{*-1}_{T+\tau} \sup_{\gamma\in\mathcal{H}} Q^{-1}m^{-1} \left|(\widetilde{\mathbf{Y}} - \widetilde{\mathbf{m}}(\gamma))'(\widetilde{\mathbf{Y}} - \widetilde{\mathbf{m}}(\gamma)) - (\widetilde{\mathbf{Y}} - \mathbf{m}^*(\gamma))'(\widetilde{\mathbf{Y}} - \mathbf{m}^*(\gamma))\right| \\
& + \xi^{*-1}_{T+\tau} \sup_{\gamma\in\mathcal{H}} Q^{-1}m^{-1} \left|(\widetilde{\mathbf{Y}} - \mathbf{m}^*(\gamma))'(\widetilde{\mathbf{Y}} - \mathbf{m}^*(\gamma)) - \mathbb{E}(\widetilde{\mathbf{Y}} - \mathbf{m}^*(\gamma))'(\widetilde{\mathbf{Y}} - \mathbf{m}^*(\gamma))\right| \\
& + \xi^{*-1}_{T+\tau} \sup_{\gamma\in\mathcal{H}} Q^{-1}m^{-1} \left|\mathbb{E}(\widetilde{\mathbf{Y}} - \mathbf{m}^*(\gamma))'(\widetilde{\mathbf{Y}} - \mathbf{m}^*(\gamma)) - Qm\mathbb{E}\{Y_{T+\tau} - Y^*_{T+\tau}(\gamma)\}^2\right| \\
\equiv\ & \xi^{*-1}_{T+\tau} \sup_{\gamma\in\mathcal{H}} \Gamma_1(\gamma) + \xi^{*-1}_{T+\tau} \sup_{\gamma\in\mathcal{H}} \Gamma_2(\gamma) + \xi^{*-1}_{T+\tau} \sup_{\gamma\in\mathcal{H}} \Gamma_3(\gamma).
\end{aligned}$$

Consequently, we need to prove the following equations

$$\xi^{*-1}_{T+\tau} \sup_{\gamma\in\mathcal{H}} \Gamma_1(\gamma) = o_p(1), \quad \xi^{*-1}_{T+\tau} \sup_{\gamma\in\mathcal{H}} \Gamma_2(\gamma) = o_p(1), \quad \text{and} \quad \xi^{*-1}_{T+\tau} \sup_{\gamma\in\mathcal{H}} \Gamma_3(\gamma) = o_p(1). \tag{18}$$

To prove the first assert in (18), we have the decomposition

$$\begin{aligned}
& ||\widetilde{\mathbf{Y}} - \widetilde{\mathbf{m}}(\gamma)||^2 \\
=\ & (\widetilde{\mathbf{Y}} - \mathbf{m}^*(\gamma) + \mathbf{m}^*(\gamma) - \widetilde{\mathbf{m}}(\gamma))'(\widetilde{\mathbf{Y}} - \mathbf{m}^*(\gamma) + \mathbf{m}^*(\gamma) - \widetilde{\mathbf{m}}(\gamma)) \\
=\ & (\widetilde{\mathbf{Y}} - \mathbf{m}^*(\gamma))'(\widetilde{\mathbf{Y}} - \mathbf{m}^*(\gamma)) + (\mathbf{m}^*(\gamma) - \widetilde{\mathbf{m}}(\gamma))'(\mathbf{m}^*(\gamma) - \widetilde{\mathbf{m}}(\gamma)) \\
& + 2(\widetilde{\mathbf{Y}} - \mathbf{m}^*(\gamma))'(\mathbf{m}^*(\gamma) - \widetilde{\mathbf{m}}(\gamma)) \\
=\ & (\widetilde{\mathbf{Y}} - \mathbf{m}^*(\gamma))'(\widetilde{\mathbf{Y}} - \mathbf{m}^*(\gamma)) + (\mathbf{m}^*(\gamma) - \widetilde{\mathbf{m}}(\gamma))'(\mathbf{m}^*(\gamma) - \widetilde{\mathbf{m}}(\gamma)) \\
& + 2(\widetilde{\mathbf{Y}} - \mathbf{m}(\gamma) + \mathbf{m}(\gamma) - \mathbf{m}^*(\gamma))'(\mathbf{m}^*(\gamma) - \widetilde{\mathbf{m}}(\gamma)) \\
=\ & (\widetilde{\mathbf{Y}} - \mathbf{m}^*(\gamma))'(\widetilde{\mathbf{Y}} - \mathbf{m}^*(\gamma)) + (\mathbf{m}^*(\gamma) - \widetilde{\mathbf{m}}(\gamma))'(\mathbf{m}^*(\gamma) - \widetilde{\mathbf{m}}(\gamma)) \\
& + 2(\widetilde{\mathbf{Y}} - \mathbf{m}(\gamma))'(\mathbf{m}^*(\gamma) - \widetilde{\mathbf{m}}(\gamma)) + \mathbf{2}(\mathbf{m}(\gamma) - \mathbf{m}^*(\gamma))(\mathbf{m}^*(\gamma) - \widetilde{\mathbf{m}}(\gamma)).
\end{aligned}$$

For any given x, we have

$$\begin{aligned}
& \sup_{\gamma\in\mathcal{H}} [\widetilde{\mathbf{m}}(\gamma) - \mathbf{m}^*(\gamma)]' [\widetilde{\mathbf{m}}(\gamma) - \mathbf{m}^*(\gamma)] \\
= & \ O_p(Qm||\widetilde{\beta}_{0,(1)}(x) - \beta^*_{0,(1)}(x)||^2 + Qm||\widetilde{\beta}_{0,(2)}(x) - \beta^*_{0,(2)}(x)||^2) \\
= & \ O_p(Qm(h_1^4 + h_2^4 + T^{-1}h_1^{-1} + T^{-1}h_2^{-1})) = O_p(1)
\end{aligned}$$

with Conditions A - B. Also, with Conditions A - B, it is shown that

$$\sup_{\gamma\in\mathcal{H}} |(\mathbf{m}^*(\gamma) - \mathbf{m}(\gamma))'(\widetilde{\mathbf{m}}(\gamma) - \mathbf{m}^*(\gamma))| = O_p(Q^{1/2}m^{1/2}).$$

Then, we have

$$\xi^{*-1}_{T+\tau} \sup_{\gamma\in\mathcal{H}} \Gamma_1(\gamma) = O_p(\xi^{*-1}_{T+\tau}(Qm)^{-1} + \xi^{*-1}_{T+\tau}(Qm)^{-1/2}) = o_p(1),$$

where the last step is obtained from Condition C. Therefore, the first assert in (18) is obtained. For simplicity, let $w^1 = s_b$ and $w^2 = 1 - s_b$. These two candidate models are $Y_{t+\tau} = m_{(1)}(X_t) + u^{(1)}_{t+\tau}$ and $Y_{t+\tau} = m_{(2)}(X_t) + u^{(2)}_{t+\tau}$ for $1 \le t \le T$. Define $u^{(1)*}_{t+\tau} = Y_{t+\tau} - \beta^*_{0,(1)}(X_t)$ and $u^{(2)*}_{t+\tau} = Y_{t+\tau} - \beta^*_{0,(2)}(X_t)$. To verify the second assertion in (18), for any given x, we have

$$\begin{aligned}
& \xi^{*-1}_{T+\tau} \sup_{\gamma\in\mathcal{H}} \Gamma_2(\gamma) \\
= & \ \xi^{*-1}_{T+\tau} \Bigg[(Qm)^{-1} \sup_{\gamma\in\mathcal{H}} | \sum_{t=T-\tau-Qm+1}^{T-\tau} \sum_{i=1}^{2}\sum_{j=1}^{2} w^i w^j \{u_t^{(i)*} u_t^{(j)*} - \mathbb{E} u_t^{(i)*} u_t^{(j)*}\}| \\
& + O_p(T^{-1/2}h_1^{-1/2} + T^{-1/2}h_2^{-1/2}) \Bigg] \\
\le & \ \xi^{*-1}_{T+\tau} \sup_{\gamma\in\mathcal{H}} \sum_{i=1}^{2}\sum_{j=1}^{2} w^i w^j |(Qm)^{-1} \sum_{t=T-\tau-Qm+1}^{T-\tau} \{u_t^{(i)*} u_t^{(j)*} - \mathbb{E} u_t^{(i)*} u_t^{(j)*}\}| + o_p(1) \\
\le & \ \xi^{*-1}_{T+\tau} \sup_{\gamma\in\mathcal{H}} \sum_{i=1}^{2}\sum_{j=1}^{2} |(Qm)^{-1} \sum_{t=T-\tau-Qm+1}^{T-\tau} \{u_t^{(i)*} u_t^{(j)*} - \mathbb{E} u_t^{(i)*} u_t^{(j)*}\}| + o_p(1) \\
= & \ \frac{1}{\xi^*_{T+\tau}\sqrt{Qm}} \sum_{i=1}^{2}\sum_{j=1}^{2} \Psi_T(i,j) + o_p(1),
\end{aligned}$$

where $\Psi_T(i,j) = |\frac{1}{\sqrt{Qm}} \sum_{t=T-\tau-Qm+1}^{T-\tau} \{u_t^{(i)*} u_t^{(j)*} - \mathbb{E} u_t^{(i)*} u_t^{(j)*}\}|$. By Theorem 3.49 and 5.20 in White (1984) and Conditions A-C, we obtain $\Psi_T(i,j) = O_p(1)$ for any i and j. Thus,

$\sum_{i=1}^{2}\sum_{j=1}^{2}\Psi_T(i,j)=O_p(1)$. Therefore, we have

$$\xi_{T+\tau}^{*-1}\sup_{\gamma\in\mathcal{H}}\Gamma_2(\gamma)=O_p(\xi_{T+\tau}^{*-1}(Qm)^{-1/2})+o_p(1)=o_p(1),$$

where the last step is due to Condition C. Finally, we consider the last assert in (18). We observe that

$$\begin{aligned}
\xi_{T+\tau}^{*-1}\sup_{\gamma\in\mathcal{H}}\Gamma_3(\gamma) &\leq \xi_{T+\tau}^{*-1}\Bigg[\sup_{\gamma\in\mathcal{H}}\sum_{i=1}^{2}\sum_{j=1}^{2}w^i w^j\left|Q^{-1}m^{-1}\sum_{t=T-\tau-Qm+1}^{T-\tau}\mathbb{E}u_t^{(i)*}u_t^{(j)*}-\mathbb{E}u_T^{(i)*}u_T^{(j)*}\right| \\
&\quad +O_p(T^{-1/2}h_1^{-1/2}+T^{-1/2}h_2^{-1/2})\Bigg] \\
&= O(\xi_{T+\tau}^{*-1}(Qm)^{-1/2})+o(1)=o(1),
\end{aligned}$$

where the last step is due to Condition C and $\{Y_t, X_t\}$ is a (or piece wise) stationary α-mixing time series in Condition A. Therefore, the proof of Theorem 2 is completed. □

6 Conclusions

When forecasting time series data, structural breaks can present a significant challenge. Existing literature has proposed several methods to handle structural breaks, but they tend to be (semi-)parametric in nature. Typically, these methods incorporate information from the pre-break period by assigning weights between 0 and 1 to the relevant observations. Building on this idea, this paper proposes a similar nonparametric estimator which offers the advantage of not requiring any specific functional form. Our proposed weighted local linear estimator has been shown in previous studies to outperform the usual post-break estimator in parametric cases.

However, our study only considers a single break and a low dimensional covariate, say less than 5, due to the so-called curse of dimensionality. This could be problematic in more complex situations, such as longer time series data with multiple breaks or with missing relevant covariates or with large or ultra large d (either $d\to\infty$ and $d/n\to 0$ or $d\gg n$). To overcome these difficulties, one might use some dimension reduction approaches such as functional coefficient model as in Cai et al. (2000), additive model as in Cai and Masry (2000), and semiparametric model as in Fan, Yao, and Cai (2003) and Cai, Juhl, and Yang (2015), and references therein. Of course, various machine learning methods can be used to estimate these functionals. These extensions warrant further investigation as future research topics.

Finally, it is worth to point out that in real-world applications, where the break date is unknown, an accurate estimation of break dates is essential. To address this issue, future research could explore robust nonparametric methods for testing and estimating multiple breaks in time series data. Such efforts would help to further improve the accuracy and reliability of time series forecasting in the presence of structural breaks. In addition, future research could focus on determining which covariates to include in the model, as well as the optimal number of covariates, in order to create a more powerful forecasting model. Therefore, developing model selection criteria is essential and deserves further investigation. Furthermore, the present study does not cover how to construct prediction intervals of the generated forecasts. In order to address objective, a potential avenue for future research could involve a quantile regression approach that takes into account structural breaks.

Disclosure Statement

We confirm that this work is original and has not been published elsewhere, nor is it currently under consideration for publication elsewhere, and, also, we declare that we have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Ando, T. and K.-C. Li (2014). A model-averaging approach for high-dimensional regression. *Journal of the American Statistical Association 109*(505), 254–265.

Bickel, S., M. Brückner, and T. Scheffer (2009). Discriminative learning under covariate shift. *Journal of Machine Learning Research 10*(September), 2137–2155.

Boot, T. and A. Pick (2020). Does modeling a structural break improve forecast accuracy? *Journal of Econometrics 215*(1), 35–59.

Cai, Z. (2007). Trending time-varying coefficient time series models with serially correlated error. *Journal of Econometrics 136*(1), 163–188.

Cai, Z., J. Fan, and Q. Yao (2000). Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association 95*(451), 941–956.

Cai, Z., Y. Fang, M. Lin, and Z. Wu (2023). A quasi synthetic control method for nonlinear models with high-dimensional covariates. *Statistica Sinica*, DOI: https://doi.org/10.5705/ss.202023.0271.

Cai, Z., T. Juhl, and B. Yang (2015). Functional index coefficient models with variable selection. *Journal of Econometrics 189*(2), 272–284.

Cai, Z. and E. Masry (2000). Nonparametric estimation of additive nonlinear arx time series: Local linear fitting and projections. *Econometric Theory 16*(4), 465–501.

Cheng, X. and B. E. Hansen (2015). Forecasting with factor-augmented regression: A frequentist model averaging approach. *Journal of Econometrics 186*(2), 280–293.

Clements, M. P. and D. F. Hendry (2006). Forecasting with breaks, In G. Elliott, C. Granger and A. Timmermann (eds.). Volume 1 of *Handbook of Economic Forecasting*, pp. 605–657. Elsevier.

Clements, M. P. and D. F. Hendry (2011). Forecasting from misspecified Models in the presence of unanticipated location shifts, In, M.P. Clements and D. F. Hendry (eds.). The Oxford Handbook of Economic Forecasting, pp. 271–314. Oxford University Press.

Diebold, F. X. and R. S. Mariano (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics 13*(3), 253–263.

Fan, J. and I. Gijbels (1996). *Local Polynomial Modeling and Its Applications*. Chapman & Hall.

Fan, J. and Q. Yao (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. Spinger-Verlag, New York.

Fan, J., Q. Yao, and Z. Cai (2003). Adaptive varying-coefficient linear models. *Journal of the Royal Statistical Society, series B 65*(1), 57–80.

Gao, Y., X. Zhang, S. Wang, T. Chong, and G. Zou (2019). Frequentist model averaging for threshold models. *Annals of the Institute of Statistical Mathematics 71*(2), 275–306.

Gao, Y., X. Zhang, S. Wang, and G. Zou (2016). Model averaging based on leave-subject-out cross-validation. *Journal of Econometrics 192*(1), 139–151.

Hansen, B. E. (2007). Least squares model averaging. *Econometrica 75*(4), 1175–1189.

Hansen, B. E. (2008). Least-squares forecast averaging. *Journal of Econometrics 146*(2), 342–350.

Hansen, B. E. and J. S. Racine (2012). Jackknife model averaging. *Journal of Econometrics 167*(1), 38–46.

Harvey, D., S. Leybourne, and P. Newbold (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting 13*(2), 281–291.

Harvey, D. I., S. J. Leybourne, and E. J. Whitehouse (2017). Forecast evaluation tests and negative long-run variance estimates in small samples. *International Journal of Forecasting 33*(4), 833–847.

Hjort, N. L. and G. Claeskens (2003). Frequentist model average estimators. *Journal of the American Statistical Association 98*(464), 879–899.

Hu, X. and X. Zhang (2023). Optimal parameter-transfer learning by semiparametric model averaging. *Journal of Machine Learning Research 24*(358), 1–53.

Karlsson, O. (2016). Volatility forecasting under structural breaks. *Working Paper*, Department of Statistics, Uppsala University.

Lee, T.-H., S. Parsaeian, and A. Ullah (2022a). Forecasting under structural breaks using improved weighted estimation. *Oxford Bulletin of Economics and Statistics 84*(6), 1485–1501.

Lee, T.-H., S. Parsaeian, and A. Ullah (2022b). Optimal forecast under structural breaks. *Journal of Applied Econometrics 37*(5), 965–987e.

Li, C., Q. Li, J. S. Racine, and D. Zhang (2018). Optimal model averaging of varying coefficient models. *Statistica Sinica 28*(4), 2795–2809.

Li, K.-C. (1987). Asymptotic optimality for C_p, C_l, cross-validation and generalized cross-validation: Discrete index set. *Annals of Statistics 15*(4), 958–975.

Liao, J., X. Zong, X. Zhang, and G. Zou (2019). Model averaging based on leave-subject-out cross-validation for vector autoregressions. *Journal of Econometrics 209*(1), 35–60.

Liu, Q. and R. Okui (2013). Heteroscedasticity-robust cp model averaging. *The Econometrics Journal 16*(3), 463–472.

Mohr, M. and L. Selk (2020). Estimating change points in nonparametric time series regression models. *Statistical Papers 61*, 1437–1463.

Parsaeian, S. (2023). Structural breaks in seemingly unrelated regression models. *Macroeconomic Dynamics*, 1–24.

Pesaran, M. and A. Pick (2011). Forecast combination across estimation windows. *Journal of Business & Economic Statistics 29*(2), 307–318.

Pesaran, M. H., A. Pick, and M. Pranovich (2013). Optimal forecasts in the presence of structural breaks. *Journal of Econometrics 177*(2), 134–152.

Pesaran, M. H. and A. Timmermann (2005). Small sample properties of forecasts from autoregressive models under structural breaks. *Journal of Econometrics 129*(1-2), 183–217.

Pesaran, M. H. and A. Timmermann (2007). Selection of estimation window in the presence of breaks. *Journal of Econometrics 137*(1), 134–161.

Racine, J. S., Q. Li, D. Yu, and L. Zheng (2023). Optimal model averaging of mixed-data kernel-weighted spline regressions. *Journal of Business & Economic Statistics 41*(4), 1251–1261.

Rossi, B. (2013). Advances in forecasting under instability, In G. Elliott, C. Granger and A. Timmermann (eds.). *Handbook of Economic Forecasting Vol. 2*, 1203–1324.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association 88*(422), 486–494.

Sugiyama, M., T. Suzuki, and T. Kanamori (2012). *Density Ratio Estimation in Machine Learning*. Cambridge University Press, New York.

Sun, Y., S. Hong, and Z. Cai (2023). Optimal local model averaging for divergent-dimensional functional-coefficient regressions. *Working Paper,* Department of Economics, University of Kansas, https://econpapers.repec.org/paper/kanwpaper/202309.htm.

Sun, Y., Y. Hong, T.-H. Lee, S. Wang, and X. Zhang (2021). Time-varying model averaging. *Journal of Econometrics 222*(2), 974–992.

Timmermann, A. (2006). Forecast combinations, In G. Elliott, C. Granger and A. Timmermann (eds.). Volume 1 of *Handbook of Economic Forecasting*, pp. 135–196. Elsevier.

Wan, A. T., X. Zhang, and G. Zou (2010). Least squares model averaging by mallows criterion. *Journal of Econometrics 156*(2), 277–283.

White, H. (1984). *Asymptotic Theory for Econometricians*. Academic Press.

Zhang, X. and C.-A. Liu (2023). Model averaging prediction by K-fold cross-validation. *Journal of Econometrics 235*(2), 280–301.

Zhang, X., D. Yu, G. Zou, and H. Liang (2016). Optimal model averaging estimation for generalized linear models and generalized linear mixed-effects models. *Journal of the American Statistical Association 111*(516), 1775–1790.

Zhu, R., A. T. Wan, X. Zhang, and G. Zou (2019). A mallows-type model averaging estimator for the varying-coefficient partially linear model. *Journal of the American Statistical Association 114*(526), 882–892.