Optimal bandwidth selection for forecasting under parameter instability

Yu Bai* Bin Peng† Shuping Shi‡ Wenying Yao§

November 16, 2025

Abstract

This paper addresses practical issues associated with the use of the local estimator in forecasting models that are affected parameter instability. We propose an approach to select the bandwidth parameter in the context of out-of-sample forecasting. Derived by minimizing the conditional expected end-of-sample loss, the selection procedure is shown to be asymptotically optimal. We also discuss the implications on the choice of kernel functions. The theoretical properties are examined through an extensive Monte Carlo study. Two empirical applications on forecasting excess bond returns and the yield curve demonstrate the superior forecasting performance of the local estimator with the proposed optimal bandwidth selection.

Keywords: Local estimator; Bandwidth selection; Kernel function; Bond return predictability; Yield curve forecasting.

JEL: C14, C51, C53

*Corresponding author: Faculty of Finance, City University of Macau. Edf. Jardim Chu Kuong, 81 Av. Xian Xing Hai, Macau. Email: yubai@cityu.edu.mo

†Department of Econometrics and Business Statistics, Monash University

‡Department of Economics, Macquarie University

§Melbourne Business School, University of Melbourne

1 Introduction

Many important economic decisions are based on forecasting models that are known to be affected by parameter instability (Rossi, 2013). It is widely recognized that parameter instability is a main source of forecast failure. There are ample empirical evidences documenting that failure to take into account parameter instability can lead to poor out-of-sample forecasting performance. See, for example, Stock and Watson (1996) and Pettenuzzo and Timmermann (2017) for macroeconomic forecasting, Welch and Goyal (2008), Gargano et al. (2019), and Borup et al. (2023) for financial return forecasting, and Inoue et al. (2021) and Oh and Patton (2024) for volatility forecasting.

Motivated by concerns about parameter instability, forecasters often estimate the model parameters and make predictions using the more recent data. A common approach is to rely on a fixed number of the most recent observations, known as the "rolling-window" estimation and forecast scheme. The rolling-window estimator can be viewed as a special case of the local estimator in nonparametric estimation when a flat (Uniform) kernel function is used. Inoue et al. (2017) develop a method for selecting the optimal window size for the rolling-window estimator by minimizing the conditional mean squared forecast error. A similar challenge arises with the more general local estimator, where one must determine the bandwidth parameter and the kernel function to estimate time-varying model coefficients and construct the forecasts. For example, Giraitis et al. (2013) propose a method for bandwidth selection in a simple location model with time-varying mean and provide theoretical justification for their approach. Pesaran et al. (2013) take a parametric approach and introduce a weighted least squares estimator for forecasting in the presence of continuous and discrete structural breaks. Their focus is on selecting weights, which depends on the nature of the breaks in the model coefficients.

This paper proposes a bandwidth selection procedure for the local estimator by directly minimizing the conditional expected loss at the end of the sample. The bandwidth parameter is central to the bias-variance trade-off and can significantly affect models' forecasting performance. Our approach extends the result of Inoue et al. (2017) in two important ways. Firstly, the asymptotic optimality of the bandwidth selection procedure applies to any generic kernel function for the local estimator, in contrast to the Uniform kernel used by Inoue et al. (2017). Secondly, we allow for a general loss function for forecast evaluation, which covers cases such as asymmetric loss functions as considered in Laurent et al. (2012). In addition, we discuss the choice of kernel functions by exploring its impact on the expected loss. We show that, when the bandwidth parameter is set to its optimal value—i.e., minimizing the end-of-sample risk—the left-sided triangular kernel, rather than the flat kernel, is optimal among the class of affine kernel functions. This is consistent with the previous literature (Cheng et al., 1997; Smetanina et al., 2025), which finds that the left-sided triangular kernel is optimal for the local linear (polynomial) estimators at the boundary point.

The theoretical analyses are examined through an extensive Monte Carlo study. Using a linear predictive regression model with various types of parameter instability as the data generating processes (DGPs), we find that the local estimator with the proposed optimal bandwidth selection procedure performs well. The gains over the benchmark, which ignores parameter instability, increase with both sample size and forecast horizon. Moreover, using alternative kernel functions generally improves the forecasting performance compared to the Uniform kernel.

We apply the proposed bandwidth selection method to three empirical applications. The first application investigates bond return predictability using data on monthly U.S. government bond yields. Although the out-of-sample predictability of bond returns has been explored in previous studies such as Gargano et al. (2019) and Borup et al. (2023), the fore-

casting performance of models estimated using local estimator has not been examined. Our results show that using the local estimator combined with the optimal bandwidth almost always leads to improved forecasting accuracy. Our second application considers yield curve forecasts using the popular Dynamic Nelson-Siegel (DNS) model as in Diebold and Li (2006). The last application focuses on real-time inflation forecasting using a variety of financial variables. Inoue et al. (2017) have examined the inflation forecast using final vintage data and the rolling-window estimator to account for the parameter instability. The empirical results suggest that using more flexible forms of kernel function delivers further benefits than the rolling-window estimator in improving the forecasting performance.

The rest of the paper is organized as follows. Section 2 introduces the model setup, the estimators and their asymptotic properties. Section 3 proposes the procedure for selecting the bandwidth parameter in terms of end-of-sample risk minimization and establishes its asymptotic optimality. Section 4 discusses the choice of kernel function and derives the optimal affine kernel. Section 5 provides the Monte Carlo study. Section 6 presents two empirical applications on bond return predictability and yield curve forecasting. Section 7 concludes. Proofs of the main theorems are provided in the Appendix. Auxiliary theoretical results, further simulation evidence, and an empirical application on real-time inflation forecasting can be found in the Online Supplement.

Before proceeding, we introduce the notations. Let $\|\cdot\|$ denote the Euclidean norm and $\|\cdot\|_p$ be the L_p norm. The expression $x_n \asymp y_n$ states that $x_n/y_n = O_p(1)$ and $y_n/x_n = O_p(1)$ (or $x_n/y_n = O(1)$ and $y_n/x_n = O(1)$). We use $\xrightarrow{p}$ and $\xrightarrow{d}$ to denote convergence in probability and convergence in distribution, respectively. $E_t[\,\cdot\,] = E[\,\cdot\,|\mathcal{F}_t]$ is the conditional expectation operator, where $\mathcal{F}_t$ is the information set available at time t.

2 Estimation under parameter instability

We consider time series models of the form

$$y_{t+h} = G(X_t, \varepsilon_t; \theta_{t,T}) \quad \text{with} \quad \theta_{t,T} = \theta\left(t/T\right), \tag{1}$$

where y_{t+h} is the scalar target variable of interest, $G(.)$ is a known function, X_t is a vector of predictors, ε_t is a sequence of errors, and $1 \leq h < \infty$ denotes the forecast horizon. $\theta(t/T)$ is a $k \times 1$ vector of time-varying parameters (TVPs), modeled as a function of the scaled time point $t/T \in (0, 1]$.

2.1 Estimators

The objective is to compute an h-step-ahead forecast conditional on the information set $\mathcal{F}_T$, denoted by $\hat{y}_{T+h|T}(\theta_T)$, for the actual outcome y_{T+h}. Since $\theta_T := \theta(1)$ is unknown, estimation becomes necessary. We take a nonparametric approach and introduce two different estimators in this section, the local estimator and the local linear estimator.

The local estimator for θ_T is defined by

$$\hat{\theta}_{K,b,T} = \arg\min_{\theta \in \Theta} \frac{1}{Tb} \sum_{t=1}^{T} k_{tT}\, \ell_t(\theta), \tag{2}$$

where $k_{tT} = K\left((t-T)/(Tb)\right)$, $K(\cdot)$ is a kernel function and $b = b_T > 0$ is a bandwidth parameter satisfying $b \to 0$, $Tb \to \infty$ as $T \to \infty$. The in-sample loss function is defined as $\ell_{t+h}(\theta) := \ell_{t+h,T}(\theta) = L(y_{t+h}, \hat{y}_{t+h|t}(\theta))$. We use the shorthand $\ell_t(\theta)$ for brevity throughout the main text.

Different specifications of $K(\cdot)$ lead to different types of forecasting schemes. For example, $k_{tT} = 1$ for all t leads to the non-local full-sample estimation $\hat{\theta} = \arg\min_{\theta \in \Theta} \frac{1}{T} \sum_{t=1}^{T} \ell_t(\theta)$. When $K(u) = \mathbf{1}_{\{-1<u<0\}}$, we effectively use a rolling-window of size $\lfloor Tb \rfloor$ in the estimation of the

parameter vector θ_t (Giacomini and Rossi, 2009).

Example 1. *Consider the time-varying linear predictive regression model* $y_{t+h} = X_t'\theta_t + \varepsilon_{t+h}$, *where* ε_{t+h} *is a disturbance term. Then, under squared error loss:* $\ell_t(\theta) = (y_{t+h} - X_t'\theta)^2$, *the local estimator for* θ_T *is given by*

$$\hat{\theta}_{K,b,T} = \left(\sum_{t=1}^{T-h} k_{tT} X_t X_t'\right)^{-1} \left(\sum_{t=1}^{T-h} k_{tT} X_t y_{t+h}\right). \tag{3}$$

Example 2. *Consider the time-varying GARCH(1,1) model* $y_t = \sigma_t \varepsilon_t$ *and* $\sigma_t^2 = \omega_t + \alpha_t y_{t-1}^2 + \beta_t \sigma_{t-1}^2$, *where* ε_t *is a white noise with variance one. Then, under the QLIKE loss*

$$L(y_t^2, \sigma_t^2) = \frac{y_t^2}{\sigma_t^2} - \log\left(\frac{y_t^2}{\sigma_t^2}\right) - 1,$$

the local quasi-maximum likelihood estimation of $\theta_T = (\omega_T, \alpha_T, \beta_T)'$ *is equivalent to minimizing the in-sample local QLIKE loss function (Oh and Patton, 2024):*

$$\hat{\theta}_{K,b,T} = \arg\min_{\theta} \frac{1}{Tb} \sum_{t=2}^{T} k_{tT} L(y_t^2, \sigma_t^2).$$

The local linear estimator is based on a local approximation $\theta_t \approx \theta_T + \theta_T^{(1)}(t/T - 1)$, where θ_T is the end-of-sample parameter and $\theta_T^{(1)}$ denotes its first-order derivative. The local linear estimator is given by

$$\left(\tilde{\theta}_T, \tilde{\theta}_T^{(1)}\right) = \arg\min_{\theta, \theta^{(1)}} L_T(\theta, \theta^{(1)}). \tag{4}$$

The in-sample loss function $L_T(\theta, \theta^{(1)})$ is defined as

$$L_T(\theta, \theta^{(1)}) = \frac{1}{T\tilde{b}} \sum_{t=1}^{T} \tilde{k}_{tT} \ell_t\left(\theta + \theta^{(1)}(t/T - 1)\right), \tag{5}$$

where the weights $\tilde{k}_{tT} = \tilde{K}\left(\frac{t-T}{T\tilde{b}}\right)$ are computed using a kernel function $\tilde{K}(\cdot)$ with a bandwidth parameter $\tilde{b}$ such that $\tilde{b} \to 0$ and $T\tilde{b} \to \infty$ as $T \to \infty$.

2.2 Assumptions

We introduce the technical assumptions required to derive the asymptotic properties of the estimators.

Assumption 1. *The true parameter* $\theta(\cdot) : [0,1] \to \Theta \subseteq \mathbb{R}^k$ *is twice continuously differentiable, and* Θ *is compact.*

Assumption 2. *Let* $\ell_t^{(1)}(\theta) = \frac{\partial \ell_{t,T}(\theta)}{\partial \theta}$ *and* $\ell_t^{(2)}(\theta) = \frac{\partial^2 \ell_t(\theta)}{\partial \theta \partial \theta'}$. *Given* θ, *the loss function* $\ell_t(\theta)$ *satisfies:*

(i) $\ell_t(\theta)$ *is three-times continuously differentiable in* θ*;*

(ii) $\{\ell_t(\theta), \ell_t^{(1)}(\theta), \ell_t^{(2)}(\theta)\}_t$ *is locally stationary with stationary approximation* $\{\tilde{\ell}_u(\theta), \tilde{\ell}_u^{(1)}(\theta), \tilde{\ell}_u^{(2)}(\theta)\}_u$ *for each rescaled time point* $u = t/T$*;*

(iii) $E_T\left[\ell_{T+h}^{(1)}(\theta_T)\right] = 0$*;*

The definition of the locally stationary processes can be found in Appendix A.1.

Assumption 3. *(i)* $\tilde{\ell}_1(\theta)$ *is ergodic and* L_1*-continuous w.r.t* θ*;* $E\left[\tilde{\ell}_1(\theta)\right]$ *is uniquely minimized at* $\theta(1)$*, in which the definition of* L_1*-continuous can be found in Appendix A.1;*

(ii) As $(T, b, Tb) \to (\infty, 0, \infty)$*,*

$$\frac{1}{\sqrt{Tb}} \sum_{t=1}^{T} k_{tT} \frac{\partial \tilde{\ell}_1(\theta_t)}{\partial \theta'} \xrightarrow{d} \mathcal{N}\left(0, \phi_{0,K} \Lambda_T\right),$$

where $\phi_{0,K} = \int_{\mathbb{B}} K^2(u) du$ *and* $\Lambda_T = \mathrm{Var}\left(\tilde{\ell}_1^{(1)}(\theta_T)\right)$*;*

(iii) The eigenvalues of $E[\tilde{\ell}_1^{(2)}(\theta)]$ *are uniformly bounded (below and above) over* θ*.*

Assumption 4. *Let* $K(\cdot)$ *and* $\tilde{K}(\cdot)$ *be the left-sided kernel functions for the two estimators* (2) *and* (4)*, respectively:*

(i) $K(u) \geqslant 0$, $u \in \mathbb{B}$ *is a Lipschitz continuous left-sided kernel function and* $\int_{\mathbb{B}} K(u)du = 1$*;*

(ii) $\tilde{K}(u) \geqslant 0$, $u \in \mathbb{B}$ *is a Lipschitz continuous left-sided kernel function,* $\int_{\mathbb{B}} \tilde{K}(u)du = 1$*, and* $\mathbb{B}$ *is compact.*

Assumption 1 imposes conditions on the time-varying parameters. As explained in Robinson (1989), the requirement that θ_t is a function of the scaled time point t/T is essential in deriving the consistency of the nonparametric estimator, since the amount of local information on which an estimator depends has to increase suitably with sample size T.

Assumption 2 imposes conditions on the loss function. We do not assume stationarity, but require the existence of a stationary approximation at the scaled time point $u = 1$, i.e. at the end of the sample. This assumption can be verified from more primitive conditions on G, ε_t and $\theta(\cdot)$, which is also related to the existence of stationary solution of (1). More details can be found in Dahlhaus et al. (2019) and Karmakar et al. (2022).[1]

Remark 1. *While Assumption 2(iii) requires correct model specification, it does not preclude the use of a direct approach (Marcellino et al., 2006) to construct multi-step-ahead forecasts. To see this, consider the data-generating process (DGP):*

$$y_t = a(t/T)\, y_{t-1} + \varepsilon_t, \quad \varepsilon_t \overset{i.i.d.}{\sim} (0, \sigma^2).$$

Suppose we construct a 2-step-ahead forecast using $y_t = \theta\, y_{t-2} + u_t$*. By recursive iteration,*

$$y_t = \beta_t\, y_{t-2} + u_t, \ \text{with } \beta = a(t/T)\, a((t-1)/T) \ \text{and } u_t = \varepsilon_t + a(t/T)\varepsilon_{t-1}.$$

Under MSE loss, the score is

$$\frac{\partial \ell_t(\theta)}{\partial \theta} = -2\, u_t\, y_{t-2}, \ \text{and } E_{t-2}\left[\frac{\partial \ell_t(\theta)}{\partial \theta}\right] = 0.$$

[1] Note that these conditions are also model specific. Karmakar et al. (2022) provide analyses on recursively defined time series (tvARMA or tvARCH models) and time-varying GARCH model.

Hence, Assumption 2(iii) continues to hold.

Assumption 3 imposes conditions on the approximated stationary process at the rescaled time point $u = 1$. These conditions ensure that certain weak law of large numbers and central limit theorem can be directly applied in the proof of Lemma 1 and Lemma 2.

Assumption 4 imposes conditions on the kernel functions $K(\cdot)$ and $\tilde{K}(\cdot)$. As Kristensen and Lee (2023) point out, when the local linear estimator is used, the support of the kernel function $\mathbb{B}$ should be compact. This rules out the use of certain kernel functions with unbounded support for the local linear estimator, such as the Gaussian kernel $K_G(u) = \frac{2}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) \mathbf{1}_{\{u<0\}}$.

2.3 Asymptotic properties

The asymptotic properties of the local estimator (2) and the local linear estimator (4) are given below.

Lemma 1. *Suppose that Assumptions 1, 2, 3 and 4(i) hold with $b \to 0$ and $Tb \to \infty$. Then, it holds that*

(i) Consistency: $\hat{\theta}_{K,b,T} \xrightarrow{p} \theta_T$;

(ii) Consistency rate: $\left\|\hat{\theta}_{K,b,T} - \theta_T\right\| = O_p\left((Tb)^{-1/2} + b\right)$;

(iii) If $b = O(T^{-1/3})$, we have

$$\sqrt{Tb}\left(\hat{\theta}_{K,b,T} - \theta_T - b\theta_T^{(1)}\mu_{1,K}\right) \xrightarrow{d} \mathcal{N}\left(0, \phi_{0,K}\Sigma_T\right),$$

where $\mu_{1,K} = \int_{\mathbb{B}} uK(u)du$, $\phi_{0,K} = \int_{\mathbb{B}} K^2(u)du$, $\Sigma_T = H_T^{-1}\Lambda_T H_T^{-1}$, $\Lambda_T = \text{Var}\left(\tilde{\ell}_1^{(1)}(\theta_T)\right)$ and $H_T = E\left(\tilde{\ell}_1^{(2)}(\theta_T)\right)$.

Lemma 2. *Suppose that Assumptions 1, 2, 3 and 4(ii) hold with* $\tilde{b} \to 0$ *and* $T\tilde{b} \to \infty$. *Then, it holds that*

$$\left\|\tilde{\theta}_T - \theta_T\right\| = O_p\left((T\tilde{b})^{-1/2} + \tilde{b}^2\right).$$

Two issues are worth mentioning. First, Lemma 1(i)-(ii) and Lemma 2 show that the local estimator $\hat{\theta}_{K,b,T}$ and the local linear estimator $\tilde{\theta}_T$ are both consistent, with the rate of convergence depending on b and $\tilde{b}$ respectively. Second, Lemma 1(iii) provides the asymptotic distribution of the local estimator, which serves as the basis for deriving the optimal kernel (Theorem 3 in Section 4).

3 Optimal bandwidth selection fore forecasting

We analyze the expected loss at the end of the sample $E_T\left(\ell_{T+h}(\hat{\theta}_{K,b,T})\right)$ to derive the optimal bandwidth selection. Suppose that $E_T\left(\ell_{T+h}(\hat{\theta}_{K,b,T})\right)$ admits the following Taylor series expansion around an open neighborhood of θ_T:

$$\begin{aligned} E_T\left(\ell_{T+h}(\hat{\theta}_{K,b,T})\right) \approx E_T\left(\ell_{T+h}\left(\theta_T\right)\right) + E_T\left(\frac{\partial \ell_{T+h}\left(\theta_T\right)}{\partial \theta'}\right)\left(\hat{\theta}_{K,b,T} - \theta_T\right) \\ + \frac{1}{2}\left(\hat{\theta}_{K,b,T} - \theta_T\right)' E_T\left(\frac{\partial^2 \ell_{T+h}\left(\theta_T\right)}{\partial \theta \partial \theta'}\right)\left(\hat{\theta}_{K,b,T} - \theta_T\right). \end{aligned} \tag{6}$$

The population loss in (6) can be decomposed into three components. The first term in the expansion, $E_T\left(\ell_{T+h}\left(\theta_T\right)\right)$, only involves the true parameter θ_T and is invariant in the parameter estimation. Following Hirano and Wright (2017), we define the *regret* as

$$R_T(K,b) = E_T\left(\frac{\partial \ell_{T+h}\left(\theta_T\right)}{\partial \theta'}\right)\left(\hat{\theta}_{K,b,T} - \theta_T\right) + \frac{1}{2}\left(\hat{\theta}_{K,b,T} - \theta_T\right)' E_T\left(\frac{\partial^2 \ell_{T+h}\left(\theta_T\right)}{\partial \theta \partial \theta'}\right)\left(\hat{\theta}_{K,b,T} - \theta_T\right).$$

Under Assumption 2(v), $E_T\left(\frac{\partial \ell_{T+h}(\theta_T)}{\partial \theta'}\right) = 0$, the regret $R_T(K,b)$ simplifies to

$$R_T(K,b) = \left(\hat{\theta}_{K,b,T} - \theta_T\right)' E_T\left(\frac{\partial^2 \ell_{T+h}\left(\theta_T\right)}{\partial \theta \partial \theta'}\right)\left(\hat{\theta}_{K,b,T} - \theta_T\right), \tag{7}$$

where the constant 1/2 is omitted. Thus, minimizing the population loss at the end of the sample (6) is equivalent to minimizing $R_T(K,b)$ in (7).

The derivation above gives rise to a procedure of selecting the bandwidth parameter b given the kernel function $K(u)$ with the aim to minimize the expected out-of-sample loss. Denote $E_T\left(\frac{\partial^2 \ell_{T+h}(\theta_T)}{\partial\theta\partial\theta'}\right)$ in (7) as $\omega_T(\theta_T)$, we consider to choose b by minimizing $R_T(K,b)$ in (7) over a choice set I_T:

$$\hat{b} := \underset{b\in I_T}{\arg\min}\ (\hat{\theta}_{K,b,T} - \theta_T)'\,\omega_T(\theta_T)\ (\hat{\theta}_{K,b,T} - \theta_T). \tag{8}$$

The bandwidth parameter selected using (8) is optimal in the sense that it minimizes the end-of-sample risk. This result is formally stated in the following theorem.

Theorem 1. *Under Assumptions 1, 2, 3 and 4(i), the optimal bandwidth parameter $\hat{b}$ obtained by minimizing (8) is of order $T^{-\frac{1}{3}}$ in probability.*

Theorem 1 implies that the optimal effective number of observations $\lfloor Tb \rfloor$, is of order $T^{2/3}$ in probability. This is the same as the result of Inoue et al. (2017) for rolling-window selection in linear predictive regression models, but the framework considered here is more general.

Although the selection criteria (8) is optimal asymptotically, it is infeasible as it involves the unknown θ_T. This problem is solved by replacing θ_T with the local linear estimator $\tilde{\theta}_T$ given in (4). The consistency of $\tilde{\theta}_T$ implies that the asymptotic property of the criterion is not affected by such a substitution. This leads to a feasible selection criterion:

$$\hat{b} := \underset{b\in I_T}{\arg\min}\ \left(\hat{\theta}_{K,b,T} - \tilde{\theta}_T\right)'\,\omega_T\left(\tilde{\theta}_T\right)\,\left(\hat{\theta}_{K,b,T} - \tilde{\theta}_T\right). \tag{9}$$

For the subsequent analysis, we require two additional assumptions.

Assumption 5. *The bandwidth parameters b and $\tilde{b}$ satisfy: (i) $T\tilde{b}^5 \to 0$; (ii) $b/\tilde{b} \to 0$; (iii) $T^{1/2}\tilde{b}^{1/2}b \to \infty$.*

Assumption 6. *Let $I_T \subset [\underline{b}, \overline{b}]$ denote the candidate set for $b \in \mathbb{R}^+$, where $\underline{b}$ and $\overline{b}$ satisfy the conditions imposed on b in Assumption 5. In addition, the Lebesgue measure of I_T, denoted by $\lambda(I_T)$, satisfies $\lambda(I_T) = O(\overline{b}^{\tau})$ for some $\tau \in (0, 1)$.*

Assumption 5 imposes conditions on the two bandwidth parameters b and $\tilde{b}$. It requires b goes to zero at a faster rate than $\tilde{b}$, which is crucial for proving the asymptotic optimality of the proposed bandwidth selection procedure (Theorem 2 in Section 3). The condition that $T\tilde{b}^5 \to 0$ ensures that the bias of $\tilde{\theta}_T$ vanish asymptotically. Assumption 6 implies length of the choice set I_T shrinks at the rate of $\overline{b}^{\tau}$ for some $0 < \tau < 1$. This assumption is useful to derive results uniformly in b, as in Marron (1985) and Härdle and Marron (1985).

Theorem 2 below establishes the asymptotic optimality of the feasible selection criterion (9) compared to the infeasible criterion (8).

Theorem 2. *Under Assumptions 1–6, choosing $\hat{b}$ by (9) is asymptotically optimal in the sense that*

$$\left(\hat{\theta}_{K,\hat{b},T} - \tilde{\theta}_T\right)' \omega_T\left(\tilde{\theta}_T\right) \left(\hat{\theta}_{K,\hat{b},T} - \tilde{\theta}_T\right) \asymp \inf_{b \in I_T} \left(\hat{\theta}_{K,b,T} - \theta_T\right)' \omega_T\left(\theta_T\right) \left(\hat{\theta}_{K,b,T} - \theta_T\right)$$

where $\tilde{\theta}_T$ is the local linear estimator from (4) with bandwidth parameter $\tilde{b}$.

Theorem 2 shows that the approximation error introduced by replacing θ_T with $\tilde{\theta}_T$ is asymptotically negligible. It provides an extension to the results in Inoue et al. (2017) by showing that the asymptotic optimality of the local estimator obtained in (2) holds for a generic kernel function and a generic loss function for forecast evaluation. The asymptotic optimality implies that $\hat{b}$ chosen from (9) yields the same forecasts as what can be obtained from using the true optimal bandwidth parameter by minimizing the infeasible objective function in (8). The key to establish this result is that the asymptotic bias from the local linear estimator vanishes at a faster rate than the local estimator, which necessitates Assumption 5(iii).

4 Choice of kernel functions

The bandwidth selection procedure in the previous section assumes a given kernel $K(\cdot)$. We now explore how different choices of kernel functions affect forecast accuracy. Taking expectations on both sides of (7) gives the regret risk as defined in Hirano and Wright (2017):

$$\begin{aligned} E[R_T(K,b)] =& \text{tr}\left(E_T\left(\tfrac{\partial^2 \ell_{T+h}(\theta_T)}{\partial\theta\partial\theta'}\right) E\left[\left(\hat{\theta}_{K,b,T}-\theta_T\right)\left(\hat{\theta}_{K,b,T}-\theta_T\right)'\right]\right) \\ &+ E\left(\hat{\theta}_{K,b,T}-\theta_T\right)' E_T\left(\tfrac{\partial^2 \ell_{T+h}(\theta_T)}{\partial\theta\partial\theta'}\right) E\left(\hat{\theta}_{K,b,T}-\theta_T\right). \end{aligned}$$

Using Lemma 1(iii), the limit of the regret risk as $T\to\infty$ becomes

$$E[R_T(K,b)] \sim \text{tr}\left(E_T\left(\tfrac{\partial^2 \ell_{T+h}(\theta_T)}{\partial\theta\partial\theta'}\right)\left(b^2\mu_{1,K}^2\theta_T^{(1)}\theta_T^{(1)\prime}+\tfrac{\phi_{0,K}\Sigma_T}{Tb}\right)\right), \tag{10}$$

where $\mu_{1,K}=\int_{\mathbb{B}} uK(u)du$ and $\phi_{0,K}=\int_{\mathbb{B}} K^2(u)du$. Minimizing (10) with respect to b leads to a solution of the optimal bandwidth

$$b_{\text{opt}} = \left\{\frac{\phi_{0,K}\text{tr}\left(E_T\left(\tfrac{\partial^2 \ell_{T+h}(\theta_T)}{\partial\theta\partial\theta'}\right)\Sigma_T\right)}{2\mu_{1,K}^2\text{tr}\left(E_T\left(\tfrac{\partial^2 \ell_{T+h}(\theta_T)}{\partial\theta\partial\theta'}\right)\theta_T^{(1)}\theta_T^{(1)\prime}\right)}\right\}^{1/3} T^{-1/3}, \tag{11}$$

which has the same order as in Theorem 1. By plugging (11) back to (10) and rearranging terms, we get

$$E[R_T(K,b)] \sim \frac{3}{(2T)^{2/3}} Q(K)^{2/3} M(\theta_T),$$

where $Q(K) := -\mu_{1,K}\phi_{0,K}$ is solely determined by the choice of the kernel function and $M(\theta_T) := \left(\text{tr}\left(E_T\left(\tfrac{\partial^2 \ell_{T+h}(\theta_T)}{\partial\theta\partial\theta'}\right)\Sigma_T\right)\right)^{2/3}\left(\text{tr}\left(E_T\left(\tfrac{\partial^2 \ell_{T+h}(\theta_T)}{\partial\theta\partial\theta'}\right)\theta_T^{(1)}\theta_T^{(1)\prime}\right)\right)^{1/3}$. From a risk reduction perspective, we should choose a kernel function that leads to the smallest $Q(K)$. The optimal kernel is therefore defined by

$$\min_{K\in\mathcal{C}_K} Q(K), \tag{12}$$

where $\mathcal{C}_K$ denotes the class of functions satisfying Assumption 4. The following theorem establishes the optimal kernel function among the subset of affine functions $\mathcal{A}_K := \{K(\cdot) \in \mathcal{C}_K : K(u) = a + bu,\ a \in \mathbb{R},\ b \in \mathbb{R}\}$.

Theorem 3. *Consider the affine kernel functions $K(\cdot) \in \mathcal{A}_K$. Under the setup in Theorem 1, the optimal kernel function defined by (12) is given by*

$$K_T(u) = 2\,(1 - |u|)\,\mathbf{1}_{\{-1<u<0\}}.$$

The optimal kernel for the local estimator (2) is the left-sided triangular kernel $K_T(\cdot)$. In a similar context, both Smetanina et al. (2025) and Cheng et al. (1997) show that $K_T(\cdot)$ is optimal for their local polynomial estimators at the left boundary point ($u = 0$). Smetanina et al. (2025) also consider a more flexible specification, $K_s(u) = (1 + s/2 - su)\mathbf{1}_{\{-1<u<0\}}$, and propose method to select s, given an arbitrary choice of b. It is also worth noting that the optimal kernel we derive is obtained under two conditions: (i) the bandwidth parameter b is fixed at the value b_{opt} and (ii) Assumption 2(iii) holds.

5 Monte Carlo experiments

We examine the finite-sample performance of the optimal bandwidth selection procedure with different kernel functions. Following Pesaran and Timmermann (2007) and Inoue et al. (2017), the DGPs are assumed to be bivariate Vector Autoregression (VAR) models of lag one:

$$\begin{bmatrix} y_{t+1} \\ x_{t+1} \end{bmatrix} = \begin{bmatrix} a_t & b_t \\ 0 & \rho_t \end{bmatrix} \begin{bmatrix} y_t \\ x_t \end{bmatrix} + \begin{bmatrix} \varepsilon^y_{t+1} \\ \varepsilon^x_{t+1} \end{bmatrix}, \tag{13}$$

where the error terms $(\varepsilon^y_{t+1}, \varepsilon^x_{t+1})'$ are generated from *i.i.d.* $\mathcal{N}(0, I_2)$. We set $\rho_t = 0.55 + 0.4\sin(4\pi(t/T))$. Thus, $\{x_t\}$ is a locally stationary process (Dahlhaus et al., 2019). The first equation in (13) is the predictive regression of interest with parameters $\theta_t = (a_t, b_t)'$.

We consider seven different specifications for the time-varying parameters $(a_t,\ b_t)'$, which are summarized in Table 1. Assumption 1 is satisfied for all these specifications. DGPs V1–V4 use deterministic time-varying parameters with different forms of time-variation in the parameters. In DGPs V5–V7, a_t still has deterministic time variation, but b_t is the realization of a persistent stochastic process whose degree of smoothness is controlled by the order of integration d.

Table 1: Specification of DGPs: V1–V7.

DGP	a_t	b_t	d
V1	$0.9 - 0.4(t/T)$	$1 + (t/T)$	
V2	$0.9 - 0.4(t/T)^2$	$1 + (t/T)^2$	
V3	$0.9 - 0.4\exp(-3.5t/T)$	$1 + \exp(-16(t/T - 0.5)^2)$	
V4	$0.7 + 0.2\cos(4\pi(t/T))$	$1.5 + 0.5\sin(4\pi(t/T))$	
V5		$\xi_t/t^{d+0.5},\ \Delta\xi_t = v_t,$	0.4
V6	$0.75 - 0.2\sin(3\pi\,(t/T))$	with $v_t = (1-L)^{-d}\epsilon_t,$	0
V7		and $\epsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(0, 0.1^2)$.	-0.3

5.1 Forecasting models

We consider the following predictive regression model to construct the forecasts:

$$y_{t+h} = X_t'\theta_t + \varepsilon_{t+h}, \tag{14}$$

where $X_t = (y_t, x_t)'$. Under the mean squared error (MSE) loss, the model parameters θ_t are estimated using local least squares

$$\hat{\theta}_{K,b,T} = \left(\sum_{t=1}^{T-h} k_{tT} X_t X_t'\right)^{-1} \left(\sum_{t=1}^{T-h} k_{tT} X_t y_{t+h}\right). \tag{15}$$

The regret risk under the MSE loss becomes

$$R_T(K,b) = (\hat{\theta}_{K,b,T} - \theta_T)'\ (X_T X_T')\ (\hat{\theta}_{K,b,T} - \theta_T). \tag{16}$$

We set $b = cT^{-1/3}$ and select c according to the feasible criterion (9) using a coarse grid of width 0.1 from 1 to 7. The true parameters in (16) are approximated by the local linear estimator (4) with the (left-sided) Epanechnikov kernel $\tilde{k}(u) = \frac{3}{2}(1-u^2)\mathbf{1}_{\{-1<u<0\}}$, with the rule-of-thumb bandwidth parameter $\tilde{b} = 1.06T^{-1/5}$.

We consider four different kernel functions for the local estimator:

$$\begin{aligned} K_R(u) &= \mathbf{1}_{\{-1<u<0\}}, \qquad K_G(u) = \frac{2}{\sqrt{2\pi}} \exp\big(-\frac{u^2}{2}\big)\mathbf{1}_{\{u<0\}}, \\ K_E(u) &= \frac{3}{2}(1-u^2)\mathbf{1}_{\{-1<u<0\}}, \qquad K_T(u) = 2\left(1-|u|\right)\mathbf{1}_{\{-1<u<0\}}. \end{aligned} \tag{17}$$

The Uniform kernel $K_R(\cdot)$ assigns equal weights to all observations within a certain window whose size is determined by the bandwidth parameter.[2] Using $K_R(\cdot)$ together with the optimal bandwidth selection procedure is equivalent to the rolling window selection method proposed by Inoue et al. (2017). The Gaussian kernel $K_G(\cdot)$ implies an exponential-type downweighting scheme and all observations are used in the estimation. The Epanechnikov kernel $K_E(\cdot)$ imposes a hyperbolic-type scheme, while the Triangular kernel $K_T(\cdot)$ imposes a linear downweighting scheme. Although $K_R(\cdot)$ has been heavily used in empirical work, there has been a growing interest in other kernel functions. For instance, $K_G(\cdot)$ has been used in macroeconomic forecasting (Kapetanios et al., 2019; Dendramis et al., 2020), and $K_E(\cdot)$ is recommended for equity premium forecasts in Farmer et al. (2022).

5.2 Forecast evaluation

We evaluate the performance of the out-of-sample prediction for y_{T+h} over $M = 5,000$ Monte Carlo simulations for $T = 200$, 400, 800 and $h = 1$, 5. For each simulated sample, the parameter estimation and the bandwidth selection is implemented using the entire sample, and the forecasts are constructed at the end of the sample. The benchmark for forecasting

[2] All kernel functions considered here are left-sided versions of the original kernels. For brevity, we refer to them by their original names.

comparison is the forecasts obtained from full-sample non-local least square estimates, which assume constant coefficients throughout the entire sample period. The forecast evaluations are based on the ratios of MSEs:

$$\frac{\sum_{m=1}^{M}(y_{T+h}^{m}-\hat{y}_{T+h|T}^{m})^{2}}{\sum_{m=1}^{M}(y_{T+h}^{m}-\bar{y}_{T+h|T}^{m})^{2}},$$

where $\bar{y}_{T+h|T}^{m}$ is the benchmark forecast and $\hat{y}_{T+h|T}^{m}$ is the forecast from using local estimators. If the ratio of MSEs is less than 1, the forecasts generated from the local estimator are more accurate than the ones from the non-local estimator.

Table 2 presents the forecasting comparison among the local estimators with four different kernel functions in (17). The top, middle and bottom panels report the results for different sample sizes. Entries shaded in gray represent the best performing model for each scenario. For 1-step ahead forecasts ($h = 1$), local estimators consistently improve the forecast accuracy when the parameters have deterministic time variation (DGPs V1-–V4). For DGPs V5-V7 where the parameters exhibit stochastic variation, the local estimator also improves forecast accuracy as the sample size increases to $T = 800$ when the Gaussian kernel is used as the weighing function.

For longer-horizon forecasts with $h = 5$, we observe improvements from using local estimators in the cases of stochastic time variation, even when the sample size is small. More importantly, the Gaussian kernel outperforms the alternative kernel functions in nearly all cases. The improvements in MSEs by using K_G combined with the proposed bandwidth selection procedure over the benchmark forecasts are substantial when $h = 5$.

We have conducted additional Monte Carlo simulations for DGPs with constant parameters or a one-time break in the parameters. The optimal window selection method (Opt_R) proposed by Inoue et al. (2017) performs best when the structural break occurs later in the sample. However, in general, using K_G with optimal bandwidth selection procedure is

Table 2: Forecasting performance of the local estimators for DGPs V1–V7.

	$h=1$				$h=5$			
DGP	Opt_R	Opt_G	Opt_E	Opt_T	Opt_R	Opt_G	Opt_E	Opt_T
$T=200$								
V1	0.960	0.933	0.971	0.977	0.792	0.778	0.797	0.803
V2	0.768	0.759	0.773	0.777	0.726	0.716	0.730	0.735
V3	0.815	0.788	0.825	0.831	0.711	0.696	0.715	0.718
V4	0.789	0.810	0.784	0.784	1.011	0.987	1.034	1.044
V5	1.055	1.023	1.066	1.075	0.957	0.926	0.972	0.983
V6	1.065	1.033	1.077	1.085	0.951	0.930	0.960	0.966
V7	1.040	1.016	1.052	1.059	0.967	0.937	0.988	0.999
$T=400$								
V1	0.924	0.909	0.930	0.933	0.749	0.742	0.749	0.751
V2	0.713	0.714	0.715	0.717	0.683	0.677	0.681	0.683
V3	0.780	0.767	0.780	0.782	0.700	0.695	0.700	0.701
V4	0.747	0.773	0.745	0.745	1.019	1.005	1.038	1.048
V5	1.033	1.010	1.036	1.040	0.925	0.907	0.932	0.941
V6	1.031	1.016	1.034	1.039	0.958	0.934	0.970	0.978
V7	1.034	1.018	1.043	1.049	0.946	0.926	0.956	0.965
$T=800$								
V1	0.882	0.879	0.886	0.888	0.718	0.716	0.716	0.717
V2	0.705	0.705	0.707	0.708	0.653	0.652	0.650	0.652
V3	0.747	0.741	0.747	0.748	0.706	0.707	0.703	0.706
V4	0.702	0.724	0.695	0.693	1.014	1.006	1.030	1.038
V5	1.008	0.999	1.010	1.012	0.898	0.895	0.907	0.911
V6	1.007	0.999	1.011	1.013	0.913	0.908	0.922	0.927
V7	1.016	1.008	1.019	1.020	0.928	0.920	0.937	0.942

Note: Ratios of MSEs against the benchmark forecasts using full-sample least square estimators. Opt_R: rolling window selection method proposed by Inoue et al. (2017); Opt_G: optimal bandwidth selection with Gaussian kernel; Opt_E: optimal bandwidth selection with Epanechnikov kernel; Opt_T: optimal bandwidth selection with triangular kernel.

preferred. These results can be found in Section S2 in the Online Supplement.

6 Empirical applications

We present two empirical applications in this Section. The first application considers the prediction of excess bond returns. Our second application focuses on forecasting yield curve using the dynamic Nelson–Siegel model of Diebold and Li (2006). A third application of real-time inflation forecasting is presented in the Online Supplement S4. In all three applications, forecasts are constructed either directly or indirectly from the linear regression models of the form

$$y_{t+h} = \theta_{0,t} + X_t'\theta_t + \varepsilon_{t+h}, \tag{18}$$

where parameters are estimated by local least squares as in (15).

We consider the four kernel functions as in (17), with the same optimal bandwidth selection procedure as in the simulation studies in Section 5. Parameter estimation and bandwidth selection are implemented recursively using an expanding window. The bandwidth parameter is set to be $b = cT^{-1/3}$, with c varying from 1 to 7 for monthly data in bond return prediction, from 1 to 10 for the daily data in yield curve forecasting, and from 1 to 5 for quarterly data in real-time inflation forecasting, all in increments of 0.1. The (left-sided) Epanechnikov kernel with fixed bandwidth parameter $\tilde{b} = 1.06T^{-1/5}$ is used for the local linear estimator $\tilde{\theta}_T$. The parameter estimation and the bandwidth selection are carried out recursively with an expanding window and are updated in each period.

Forecasts are evaluated using the MSE loss. The Diebold-Mariano test (Diebold and Mariano, 1995, DM hereafter) is used to verify the statistical significance of the difference in MSEs between the forecasts produced from the competing models and the baseline forecasts. We follow Coroneo and Iacone (2020) and apply the fixed-smoothing asymptotics for the DM

test, which has been shown to deliver predictive accuracy tests that are correctly sized even when the number of out-of-sample observations are small.

6.1 Bond return predictability

Government bond yields are commonly used as proxies for risk-free interest rates, and hence are central in investors' decisions of portfolio allocation. Recent literature has documented evidence of time-variation in the predictability of bond returns (Gargano et al., 2019; Borup et al., 2023). We examine the forecast performance of the proposed local estimator with optimal bandwidth selection in predicting excess bond returns.

Following Cochrane and Piazzesi (2005), we link the yield of an n-year bond $y_t^{(n)}$ and its log-price $p_t^{(n)}$ at time t via $y_t^{(n)} = -1/n \cdot p_t^{(n)}$. We consider $n = 2$, 3, 4, and 5 years in this analysis. The holding-period return of buying an n-year bond at time t and selling it as an $(n-1)$-year bond at time $t+1$ is

$$r_{t+1}^{(n)} = p_{t+1}^{(n-1)} - p_t^{(n)},$$

which leads to the target variable in this forecasting exercise, the excess return:

$$rx_{t+1}^{(n)} = r_{t+1}^{(n)} - y_t^{(1)}.$$

where the yield of the 1-year bond $y_t^{(1)}$ is used as the risk-free rate.

We consider three commonly used predictors in predicting the excess bond return $rx_{t+1}^{(n)}$. The first is the Fama-Bliss (FB) forward spread (Fama and Bliss, 1987):

$$FB_t^{(n)} = f_t^{(n)} - y_t^{(1)},$$

where the forward rate $f_t^{(n)} = p_t^{(n-1)} - p_t^{(n)}$. Fama and Bliss (1987) find that the forward spread has predictive power on excess bond returns which increases with the forecast horizon.

The second predictor is the Cochrane-Piazzesi (CP) factor of forward rates (Cochrane and Piazzesi, 2005):

$$CP_t = \hat{\delta}'\mathbf{f}_t,$$

where $\mathbf{f}_t = (f_t^{(1)}, f_t^{(2)}, f_t^{(3)}, f_t^{(4)}, f_t^{(5)})'$. The coefficient vector $\hat{\delta}$ is estimated from a predictive regression of $\frac{1}{4}\sum_{n=2}^{5} rx_{t+1}^{(n)}$ on $\mathbf{f}_t$.

Lastly, let $\hat{g}_{it}$ be the ith principle component estimated from a panel of macroeconomic variables. The Ludvigson-Ng (LN) factor is computed as a linear combination of the first eight principle components as in Ludvigson and Ng (2009):

$$LN_t = \hat{\lambda}'\hat{G}_t,$$

where $\hat{G}_t = (\hat{g}_{1,t}, \hat{g}_{2,t}, \ldots, \hat{g}_{8,t})$ and $\hat{\lambda}$ is the vector of factor loading obtained from the regression

$$\frac{1}{4}\sum_{n=2}^{5} rx_{t+1}^{(n)} = \lambda_0 + \lambda'\hat{G}_t + \overline{\varepsilon}_{t+1}.$$

The forecasts are constructed from the linear predictive regression model (18). The three predictors, FB, CP, and LN, are first considered one at a time, and then are included a multivariate model together. The benchmark forecasts are obtained from the model implied by the efficient-market hypothesis, which assumes no predictability by setting $\theta_t = 0$ in (18) for all t. In addition to the forecasts constructed using individual predictors, we also provide results based on forecasting combinations with equal weights and with weights based on discounted MSE (DMSE).[3]

We obtain monthly U.S. zero-coupon government bond yield data from Jing Cynthia Wu's website and use these yields to construct the FB and CP factors.[4] The sample period

[3]Details of the forecasting combination methods can be found in the Online Supplement S3.
[4]https://sites.google.com/view/jingcynthiawu/yield-data.

is from June 1961 to December 2024. The FRED-MD data set is used to compute the LN factors. Each variable is transformed as described in the Appendix of McCracken and Ng (2016). The vintage data dated 2025:M4 are used. The initial estimation sample covers the period from 1961:M6 to 1984:M12 and the first available individual forecast is for 1985"M1. We use 60 monthly observations as the out-of-sample hold-out period to obtain the weights for forecast combination based on the DMSE. Thus, the forecast evaluation is conducted using data from 1990:M1 to 2024:M12.

Table 3 presents the ratios of MSEs from the four competing estimators relative to the benchmark forecasts. Value below one suggests that the corresponding model produces better out-of-sample forecasts than the benchmark. Entries shaded in gray represent the best performing models. Statistically significant differences using the DM test at the 5% significance level are denoted by an asterisk.

Table 3 reveals that the local estimators with the optimal bandwidth selection often lead to better forecasting performance than the benchmark across the different kernel functions considered here. In particular, the forecasts using the Uniform kernel and the Epanechnikov kernel are more accurate than the benchmark in all cases. Forecasting combination methods lead to further improvements in forecasting accuracy. Combining all individual model forecasts from local estimator using the triangular kernel K_T with DMSE weights produces the best forecasts for all four maturities considered, and its improvements over the benchmark forecasts are statistically significant. The overall results provide strong empirical evidence on the superior forecasting performance using the local estimator with the proposed optimal bandwidth selection.

Table 3: Out-of-sample forecasting performance for bond returns: January 1990–December 2024.

	Opt_R	Opt_G	Opt_E	Opt_T	Opt_R	Opt_G	Opt_E	Opt_T
	$n = 2$				$n = 3$			
FB	0.923	0.893	0.919	0.880	0.918	0.890	0.908	0.853
CP	0.822	0.836	0.806	0.757	0.801	0.822	0.783	0.732
LN	0.690	0.863	0.700	0.680	0.707	0.989	0.724	0.709
FB+CP+LN	0.653*	1.005	0.634*	0.619*	0.687*	1.114	0.709	0.701
Comb-EW	0.661*	0.745*	0.638*	0.605*	0.673*	0.784*	0.659*	0.623*
Comb-DMSE	0.609*	0.665*	0.581*	0.554*	0.627*	0.707	0.611*	0.581*
	$n = 4$				$n = 5$			
FB	0.910	0.874	0.874	0.832	0.879	0.853	0.842	0.807
CP	0.775	0.797	0.750	0.708	0.758	0.782	0.730	0.690*
LN	0.729	1.128	0.755	0.754	0.794	1.348	0.811	0.836
FB+CP+LN	0.731	1.329	0.830	0.860	0.813	1.401	0.949	1.017
Comb-EW	0.674*	0.813	0.661*	0.636*	0.686*	0.843	0.680*	0.662*
Comb-DMSE	0.630*	0.716	0.618*	0.591*	0.645*	0.746	0.639*	0.616*

Note: Ratios of MSEs against the benchmark forecasts of no predictability. Opt_R: rolling window selection method proposed by Inoue et al. (2017); Opt_G: optimal bandwidth selection with Gaussian kernel; Opt_E: optimal bandwidth selection with Epanechnikov kernel; Opt_T: optimal bandwidth selection with triangular kernel. Differences in forecasting accuracy that are significant at the 5% level using the DM test are marked by an asterisk. The grey-shaded cells denote the best forecasting performance for each group.

6.2 Yield curve forecasting

Let $y_t(\tau)$ be the yield on a bond with maturity τ at time t. The Nelson–Siegel model (Nelson and Siegel, 1987) summarizes the term structure of bond yields using three factors:

$$y_t(\tau) = \beta_{1,t} + \beta_{2,t}\left(\frac{1 - e^{-\lambda_t \tau}}{\lambda_t \tau}\right) + \beta_{3,t}\left(\frac{1 - e^{-\lambda_t \tau}}{\lambda_t \tau} - e^{-\lambda_t \tau}\right) + e_t(\tau), \tag{19}$$

with the level factor $\beta_{1,t}$, the slope factor $\beta_{2,t}$, and the curvature factor $\beta_{3,t}$. The level factor determines the long-term yield as $\tau \to \infty$ and the slope factor $\beta_{2,t}$ captures the difference between the long- and short-term yields. The decay rate λ_t determines the maturity at which the loading on the curvature factor achieves its maximum. Following Diebold and Li (2006), we fix $\lambda_t = 0.0609$, which implies that the loading on the curvature factor peaks at

30-months.

Diebold and Li (2006) propose to model the Nelson–Siegel factors $\{\beta_{1,t},\, \beta_{2,t},\, \beta_{3,t}\}$ as univariate AR(1) processes:

$$\beta_{i,t+1} = \phi_{i0} + \phi_{i1}\beta_{i,t} + \epsilon_{i,t+1}, \qquad i = 1,\, 2,\, 3, \tag{20}$$

where ϕ_{i0} and ϕ_{i1} are set to be fixed parameters in the specifications of Diebold and Li (2006). Equations (19) and (20) jointly define the Dynamic Nelson–Siegel (DNS) model which has been widely used in yield curve modeling and forecasting (see, for example, Diebold et al., 2008; Christensen et al., 2011; Diebold and Rudebusch, 2013).

The yield forecasts $\hat{y}_{t+h|t}(\tau)$ based on the DNS model are constructed as follows. For each $t = 1, \ldots, T$, we first estimate the cross-sectional regression (19) to obtain the fitted factors $\{\hat{\beta}_{1,t},\, \hat{\beta}_{2,t},\, \hat{\beta}_{3,t}\}$. Then a time-varying AR specification is employed to obtain the predicted factors:

$$\hat{\beta}_{i,T+h|T} = \hat{\phi}_{i0,T} + \hat{\phi}_{i1,T}\hat{\beta}_{i,T}, \qquad i = 1,\, 2,\, 3, \tag{21}$$

where the coefficients $\phi_{i0,T}$ and $\phi_{i1,T}$ are estimated using the proposed local estimator with optimal bandwidth selection. Finally, the out-of-sample prediction for bond yield with maturity τ is constructed using (19):

$$\hat{y}_{T+h|T}(\tau) = \hat{\beta}_{1,T+h|T} + \hat{\beta}_{2,T+h|T}\left(\frac{1-e^{-\lambda\tau}}{\lambda\tau}\right) + \hat{\beta}_{3,T+h|T}\left(\frac{1-e^{-\lambda\tau}}{\lambda\tau} - e^{-\lambda\tau}\right).$$

We use local least square estimators based on the kernel functions in (17) with optimal bandwidth selection procedure to estimate the coefficients in (21). The benchmark forecasts are those generated by the non-local least squares estimators.

Daily data on U.S. zero-coupon government bonds during the period from January 2000 to December 2024 are obtained from Jing Cynthia Wu's website. We consider a total of

twelve maturities: three and six months, and one to ten years. The initial estimation sample runs from January 2000 to December 2004 and the first available individual forecast is for the first trading day of 2005. Thus, the forecast evaluation period is from January 2005 to December 2024. We examine one-day and five-day ahead forecasts, i.e. $h = 1$ and 5.

Table 4: Out-of-sample forecasting performance for the yiled curve: January 2005–December 2024.

	$h = 1$				$h = 5$			
τ (years)	Opt_R	Opt_G	Opt_E	Opt_T	Opt_R	Opt_G	Opt_E	Opt_T
0.25	0.999	0.996	0.999	0.999	1.025	1.003	1.027	1.026
0.5	1.006	0.995	1.005	1.005	1.118	1.024	1.118*	1.111
1	1.005	1.010	1.006	1.007	1.044	1.037	1.047	1.048
2	1.008	1.010	1.010	1.010	1.059	1.046	1.064	1.066
3	1.005	1.007	1.008	1.008	1.073	1.040	1.080	1.082
4	0.994	0.987*	0.992	0.991	1.059	1.003	1.059	1.058
5	0.981	0.973*	0.977	0.975*	1.012	0.963	1.007	1.004
6	0.973*	0.971*	0.970*	0.969*	0.965	0.938*	0.962	0.960
7	0.975*	0.975*	0.973*	0.973*	0.955	0.940*	0.951	0.949
8	0.981*	0.981*	0.979*	0.979*	0.971	0.956	0.968	0.966
9	0.985	0.985*	0.983	0.983	0.984	0.966	0.982	0.980
10	0.994	0.994	0.994	0.995	1.006	0.991	1.007	1.006

Note: Ratios of MSEs against the benchmark forecasts obtained using the non-local least square estimates. Opt_R: rolling window selection method proposed by Inoue et al. (2017); Opt_G: optimal bandwidth selection with Gaussian kernel; Opt_E: optimal bandwidth selection with Epanechnikov kernel; Opt_T: optimal bandwidth selection with triangular kernel. Differences in forecasting accuracy that are significant at the 5% level using the DM test are marked by an asterisk. The grey-shaded cells denote the best forecasting performance for each group.

Table 4 summarizes the ratios of MSEs of using the local estimators with four different kernel functions against the benchmark. For $h = 1$, the gains from local estimators are more pronounced for medium-term yields with τ ranging from four to eight years. The proposed optimal bandwidth selection procedure leads to statistically significant forecast improvement in most cases. Using alternative kernel functions also improves forecast accuracy relative to the rolling window selection method proposed by Inoue et al. (2017) (Opt_R). The Gaussian kernel and the triangular kernel K_T deliver the most accurate forecasts. We also observe improvements in forecast performance for $\tau = 0.25$, 0.5, and 10, although the gains are small

and insignificant. The forecasting comparisons for the five-day horizon are generally similar to the results for the one-day horizon. Gains from using local estimators with the optimal bandwidth selection procedure are more evident for the medium- and long-term yields.

7 Conclusion

Parameter instability is pervasive in many forecasting models, and the local estimator is often employed to address this issue. This paper tackles the choice of the bandwidth parameter when the local estimator is used in an out-of-sample forecasting context. We propose a feasible and asymptotically optimal procedure to select the bandwidth parameter by minimizing the conditional expected loss at the end of the sample. This approach generalizes Inoue et al. (2017) for rolling-window selection without restricting the DGP to be linear predictive regression, and allows for more general forms of loss function and kernel density function. In addition, we discuss the choice of kernel functions affects the expected loss. In particular, we derive the optimal kernel function among the class of affine kernel functions and show that it is the left-sided triangular kernel rather than the flat kernel, suggesting that the rolling-window estimator may not always be the best choice.

Our theoretical results are evaluated through an extensive Monte Carlo study. Simulation results show that the local estimator with the proposed optimal bandwidth selection performs well under various form of parameter instability. The empirical analysis considers three applications of forecasting excess bond returns, yield curve and inflation. The results show that when competing against some widely used forecasting models, the local estimator with the proposed optimal bandwidth selection procedure is able to outperform and produce more accurate forecasts in many cases. These results not only highlight the prevalence of parameter instability in forecasting models, but also provide strong evidence of the benefit and usefulness of our proposed estimator.

References

Bates, J. M. and Granger, C. W. (1969), 'The combination of forecasts', *Journal of the operational research society* **20**(4), 451–468.

Borup, D., Eriksen, J. N., Kjær, M. M. and Thyrsgaard, M. (2023), 'Predicting bond return predictability', *Management Science* .

Cheng, M.-Y., Fan, J. and Marron, J. S. (1997), 'On automatic boundary corrections', *The Annals of Statistics* **25**(4), 1691–1708.

Christensen, J. H., Diebold, F. X. and Rudebusch, G. D. (2011), 'The affine arbitrage-free class of nelson–siegel term structure models', *Journal of Econometrics* **164**(1), 4–20.

Cochrane, J. H. and Piazzesi, M. (2005), 'Bond risk premia', *American economic review* **95**(1), 138–160.

Coroneo, L. and Iacone, F. (2020), 'Comparing predictive accuracy in small samples using fixed-smoothing asymptotics', *Journal of Applied Econometrics* **35**(4), 391–409.

Croushore, D. and Stark, T. (2001), 'A real-time data set for macroeconomists', *Journal of econometrics* **105**(1), 111–130.

Dahlhaus, R., Richter, S. and Wu, W. B. (2019), 'Towards a general theory for nonlinear locally stationary processes', *Bernoulli* **25**(2), 1013–1044.

Dendramis, Y., Kapetanios, G. and Marcellino, M. (2020), 'A similarity-based approach for macroeconomic forecasting', *Journal of the Royal Statistical Society Series A: Statistics in Society* **183**(3), 801–827.

Diebold, F. X. and Li, C. (2006), 'Forecasting the term structure of government bond yields', *Journal of econometrics* **130**(2), 337–364.

Diebold, F. X., Li, C. and Yue, V. Z. (2008), 'global yield curve dynamics and interactions: a dynamic nelson–siegel approach', *Journal of Econometrics* **146**(2), 351–363.

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

Diebold, F. X. and Rudebusch, G. D. (2013), *yield curve modeling and forecasting: the dynamic nelson-siegel approach*, Princeton University Press.

Fama, E. F. and Bliss, R. R. (1987), 'The information in long-maturity forward rates', *The American Economic Review* pp. 680–692.

Farmer, L., Schmidt, L. and Timmermann, A. (2022), 'Pockets of predictability', *Journal of Finance, forthcoming* .

Gargano, A., Pettenuzzo, D. and Timmermann, A. (2019), 'Bond return predictability: economic value and links to the macroeconomy', *Management Science* **65**(2), 508–540.

Giacomini, R. and Rossi, B. (2009), 'Detecting and predicting forecast breakdowns', *The Review of Economic Studies* **76**(2), 669–705.

Giraitis, L., Kapetanios, G. and Price, S. (2013), 'Adaptive forecasting in the presence of recent and ongoing structural change', *Journal of Econometrics* **177**(2), 153–170.

Härdle, W. and Marron, J. S. (1985), 'Optimal bandwidth selection in nonparametric regression function estimation', *The Annals of Statistics* pp. 1465–1481.

Hirano, K. and Wright, J. H. (2017), 'Forecasting with model uncertainty: representations and risk reduction', *Econometrica* **85**(2), 617–643.

Inoue, A., Jin, L. and Pelletier, D. (2021), 'Local-linear estimation of time-varying-parameter garch models and associated risk measures', *Journal of Financial Econometrics* **19**(1), 202–234.

Inoue, A., Jin, L. and Rossi, B. (2017), 'Rolling window selection for out-of-sample forecasting with time-varying parameters', *Journal of econometrics* **196**(1), 55–67.

Kapetanios, G., Marcellino, M. and Venditti, F. (2019), 'Large time-varying parameter vars: a nonparametric approach', *Journal of Applied Econometrics* **34**(7), 1027–1049.

Karmakar, S., Richter, S. and Wu, W. B. (2022), 'Simultaneous inference for time-varying models', *Journal of Econometrics* **227**(2), 408–428.

Kristensen, D. and Lee, Y. J. (2023), 'Local polynomial estimation of time-varying parameters in nonlinear models', *arXiv preprint arXiv:1904.05209* .

Laurent, S., Rombouts, J. V. and Violante, F. (2012), 'On the forecasting accuracy of multivariate garch models', *Journal of Applied Econometrics* **27**(6), 934–955.

Ludvigson, S. C. and Ng, S. (2009), 'Macro factors in bond risk premia', *The Review of Financial Studies* **22**(12), 5027–5067.

Marcellino, M., Stock, J. H. and Watson, M. W. (2006), 'A comparison of direct and iterated multistep ar methods for forecasting macroeconomic time series', *Journal of econometrics* **135**(1-2), 499–526.

Marron, J. S. (1985), 'An asymptotically efficient solution to the bandwidth problem of kernel density estimation', *The Annals of Statistics* **13**(3), 1011–1023.

McCracken, M. W. and Ng, S. (2016), 'Fred-md: a monthly database for macroeconomic research', *Journal of Business & Economic Statistics* **34**(4), 574–589.

McCracken, M. W., Ng, S. et al. (2021), 'Fred-qd: a quarterly database for macroeconomic research', *Federal Reserve Bank of St. Louis Review* **103**(1), 1–44.

Nelson, C. R. and Siegel, A. F. (1987), 'Parsimonious modeling of yield curves', *Journal of business* pp. 473–489.

Newey, W. K. and McFadden, D. (1994), 'Large sample estimation and hypothesis testing', *Handbook of econometrics* **4**, 2111–2245.

Oh, D. H. and Patton, A. J. (2024), 'Better the devil you know: Improved forecasts from imperfect models', *Journal of Econometrics* **242**(1), 105767.

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

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

Pettenuzzo, D. and Timmermann, A. (2017), 'Forecasting macroeconomic variables under model instability', *Journal of business & economic statistics* **35**(2), 183–201.

Rapach, D. E., Strauss, J. K. and Zhou, G. (2010), 'Out-of-sample equity premium prediction: combination forecasts and links to the real economy', *The Review of Financial Studies* **23**(2), 821–862.

Robinson, P. M. (1989), *Nonparametric estimation of time-varying parameters*, Springer.

Romer, C. D. and Romer, D. H. (2000), 'Federal reserve information and the behavior of interest rates', *American economic review* **90**(3), 429–457.

Rossi, B. (2013), Advances in forecasting under instability, *in* 'Handbook of economic forecasting', Vol. 2, Elsevier, pp. 1203–1324.

Smetanina, E. K., Timmermann, A. and Zhu, Y. (2025), 'Shaping forecast models for arbitrary choice of bandwidth', *Mimeo* .

Stock, J. H. and Watson, M. W. (1996), 'Evidence on structural instability in macroeconomic time series relations', *Journal of Business & Economic Statistics* **14**(1), 11–30.

Stock, J. H. and Watson, M. W. (2003), 'Forecasting output and inflation: the role of asset prices', *Journal of economic literature* **41**(3), 788–829.

Stock, J. H. and Watson, M. W. (2004), 'Combination forecasts of output growth in a seven-country data set', *Journal of forecasting* **23**(6), 405–430.

Vogt, M. (2012), 'Nonparametric regression for locally stationary time series', *The Annals of Statistics* **40**(5), 2601–2633.

Welch, I. and Goyal, A. (2008), 'A comprehensive look at the empirical performance of equity premium prediction', *The Review of Financial Studies* **21**(4), 1455–1508.

A Appendix: Mathematical Annex

In the forecasting setting, the variables $y_{t+h,T}$, $X_{t,T}$, and the loss function $\ell_{t,T}$, together with its stationary approximation $\tilde{\ell}_{u,T}$, are all triangular arrays (Vogt, 2012; Dahlhaus et al., 2019). For ease of notation, we omit the subscript T in the main text, but retain the triangular array form in the proofs to ensure rigor.

A.1 Definitions

Definition 1. *A triangular array of processes $W_{t,T}(\theta)$, $\theta \in \Theta$, $t = 1, 2, \cdots, T$, and $T = 1, 2, \cdots$ is locally stationary if there exists a stationary process $\tilde{W}_{t/T,T}(\theta)$ for each rescaled time point $t/T \in [0, 1]$, such that for some $0 < \rho < 1$ and all T,*

$$\mathbb{P}\left(\max_{\theta\in\Theta}\max_{1\leqslant t\leqslant T}\left\|W_{t,T}(\theta) - \tilde{W}_{t/T,T}(\theta)\right\| \leqslant C_T(T^{-1} + \rho^t)\right) = 1,$$

where C_T is a measurable process satisfying $\sup_T E\left(\|C_T\|^\eta\right) < \infty$ for some $\eta > 0$.

Note that this definition follows from Kristensen and Lee (2023) with an additional term ρ^t in the approximation error. This ensures that the process $W_{t,T}(\theta)$ can be arbitrarily initialized. The next definition again is borrowed from Kristensen and Lee (2023).

Definition 2. *A stationary process $W_t(\theta)$, $\theta \in \Theta$, is said to be L_p-continuous w.r.t. θ for some $p \geqslant 1$ if*

(i) $\|W_t(\theta)\|_p < \infty$ *for all* $\theta \in \Theta$;

(ii) $\forall \epsilon > 0$, $\exists \delta > 0$, *such that*

$$E\left[\max_{\theta':\|\theta-\theta'\|<\delta}\|W_t(\theta) - W_t(\theta')\|^p\right]^{1/p} < \epsilon.$$

A.2 Proof of Lemma 1

Recall the definition of local estimator:

$$\hat{\theta}_{K,b,T} = \arg\min_{\theta\in\Theta}\ \frac{1}{Tb}\sum_{t=1}^{T} k_{tT}\ell_{t+h,T}(\theta), \tag{A.1}$$

where $\ell_{t+h,T}(\theta) = L(y_{t+h}, \hat{y}_{t+h|t}(\theta))$. Let $L_T(\theta) = \frac{1}{Tb}\sum_{t=1}^{T} k_{tT}\ell_{t+h,T}(\theta)$.

Proof of (i): Write $\tilde{L}_T(\theta) = \frac{1}{Tb}\sum_{t=1}^{T} k_{tT}\tilde{\ell}_{1,T}(\theta)$, where $\tilde{\ell}_{1,T}(\cdot)$ is the stationary approximation of $\ell_{t,T}$. By Assumption 2 and Definition 1, we have

$$\begin{aligned}\sup_{\theta\in\Theta}\left|L_T(\theta) - \tilde{L}_T(\theta)\right| &\leqslant \sup_{\theta\in\Theta}\frac{1}{Tb}\sum_{t=1}^{T} k_{tT}\left|\ell_{t,T}(\theta) - \tilde{\ell}_{1,T}(\theta)\right| \\ &\leqslant O_p(1)\frac{1}{Tb}\sum_{t=1}^{T} k_{tT}(T^{-1} + \rho^t) = O_p(T^{-1}) + O_p\left((Tb)^{-1/2}\right) = o_p(1),\end{aligned} \tag{A.2}$$

where order of the second term follows from Cauchy-Schwarz inequality:

$$\frac{1}{Tb}\sum_{t=1}^{T} k_{tT}\rho^t \leqslant \sqrt{\frac{1}{(Tb)^2}\sum_{t=1}^{T} k_{tT}^2}\sqrt{\sum_{t=1}^{T}\rho^{2t}} = O\left((Tb)^{-1/2}\right).$$

This implies that (A.1) can be viewed as

$$\hat{\theta}_{K,b,T} = \arg\min_{\theta\in\Theta}\ \tilde{L}_T(\theta).$$

In view of Theorem 2.1 in Newey and McFadden (1994), it is sufficient to verify that

(1) $E\left[\tilde{\ell}_{1,T}(\theta)\right]$ is uniquely minimized at θ_T (assumed in Assumption 3(i));

(2) Θ is compact (assumed in Assumption 1);

(3) $\tilde{L}_T(\theta)$ is continuous (implied by Assumption 2(i));

(4) Uniform weak law of large numbers (UWLLN):

$$\sup_{\theta\in\Theta}\left|\frac{1}{Tb}\sum_{t=1}^{T} k_{tT}\tilde{\ell}_{1,T}(\theta) - E\left[\tilde{\ell}_{1,T}(\theta)\right]\right| = o_p(1).$$

What remains is to show (4). The ergodicity assumed in Assumption 3(i) implied that for each $\theta \in \Theta$, we have

$$\left|\frac{1}{Tb}\sum_{t=1}^{T} k_{tT}\tilde{\ell}_{1,T}(\theta) - E\left[\tilde{\ell}_{1,T}(\theta)\right]\right| = o_p(1).$$

Then, uniform consistency result follows if we could show that $\tilde{L}_T(\theta)$ is stochastic equicontinuous, which follows from the fact that $\tilde{\ell}_{u,t}(\theta)$ is L_1 continuous.

Proof of (ii) and (iii): Let us first define the score and the Hessian:

$$S_T(\theta) = \frac{\partial L_T(\theta)}{\partial \theta} = \frac{1}{Tb}\sum_{t=1}^{T} k_{tT}\frac{\partial \ell_{t,T}(\theta)}{\partial \theta}, \quad H_T(\theta) = \frac{\partial^2 L_T(\theta)}{\partial \theta \partial \theta'} = \frac{1}{Tb}\sum_{t=1}^{T} k_{tT}\frac{\partial^2 \ell_{t,T}(\theta)}{\partial \theta \partial \theta'}.$$

By mean value theorem, we have

$$\frac{\partial L_T\left(\theta_T\right)}{\partial \theta} + \frac{\partial^2 L_T\left(\overline{\theta}_T\right)}{\partial \theta \partial \theta'}\left(\hat{\theta}_{K,b,T} - \theta_T\right) = 0,$$

where $\overline{\theta}_T$ lies between θ_T and $\hat{\theta}_{K,b,T}$. By rearranging terms, we have

$$\begin{aligned}
\hat{\theta}_{K,b,T} - \theta_T &= -\left(\frac{\partial^2 L_T\left(\overline{\theta}_T\right)}{\partial \theta \partial \theta'}\right)^{-1}\left(\frac{\partial L_T\left(\theta_T\right)}{\partial \theta}\right) \\
&= -\left(\frac{\partial^2 L_T\left(\theta_T\right)}{\partial \theta \partial \theta'}\right)^{-1}\left(\frac{\partial L_T\left(\theta_T\right)}{\partial \theta}\right) + \left[\left(\frac{\partial^2 L_T\left(\theta_T\right)}{\partial \theta \partial \theta'}\right)^{-1} - \left(\frac{\partial^2 L_T\left(\overline{\theta}_T\right)}{\partial \theta \partial \theta'}\right)^{-1}\right]\frac{\partial L_T\left(\theta_T\right)}{\partial \theta} \\
&= -\left(\frac{\partial^2 L_T\left(\theta_T\right)}{\partial \theta \partial \theta'}\right)^{-1}\left(\frac{\partial L_T\left(\theta_T\right)}{\partial \theta}\right) + \left(\frac{\partial^2 L_T\left(\theta_T\right)}{\partial \theta \partial \theta'}\right)^{-1}\left[\frac{\partial^2 L_T\left(\overline{\theta}_T\right)}{\partial \theta \partial \theta'} - \frac{\partial^2 L_T\left(\theta_T\right)}{\partial \theta \partial \theta'}\right] \\
&\qquad \times \left(\frac{\partial^2 L_T\left(\overline{\theta}_T\right)}{\partial \theta \partial \theta'}\right)^{-1}\frac{\partial L_T\left(\theta_T\right)}{\partial \theta}, \\
&:= -H_T^{-1}\left(\theta_T\right) S_T\left(\theta_T\right) + H_T^{-1}\left(\theta_T\right)\left[H_T\left(\overline{\theta}_T\right) - H_T\left(\theta_T\right)\right] H_T^{-1}\left(\overline{\theta}_T\right) S_T\left(\theta_T\right) \quad \text{(A.3)}
\end{aligned}$$

We will show that

$$\left\|H_T^{-1}\left(\theta_T\right)\right\| = O_p(1), \quad \text{(A.4)}$$

$$\left\|S_T\left(\theta_T\right)\right\| = O_p\left((Tb)^{-1/2} + b\right), \quad \text{(A.5)}$$

$$\left\|H_T\left(\overline{\theta}_T\right) - H_T\left(\theta_T\right)\right\| = o_p(1). \quad \text{(A.6)}$$

These bounds together with (A.3) implies the consistency rate in 1(i).

Proof of (A.4). It follows similarly from (A.2) that

$$\left\|H_T\left(\theta_T\right) - \tilde{H}_T\left(\theta_T\right)\right\| = o_p(1),$$

where $\tilde{H}_T(\theta_T) = \frac{1}{Tb}\sum_{t=1}^{T} k_{tT}\frac{\partial^2 \tilde{\ell}_{1,T}(\theta_T)}{\partial\theta\partial\theta'}$. Write

$$\begin{aligned}\tilde{H}_T(\theta_T) &= \frac{1}{Tb}\sum_{t=1}^{T} k_{tT} E\left[\frac{\partial^2 \tilde{\ell}_{1,T}(\theta_T)}{\partial\theta\partial\theta'}\right] + \frac{1}{Tb}\sum_{t=1}^{T} k_{tT}\left(\frac{\partial^2 \tilde{\ell}_{1,T}(\theta_T)}{\partial\theta\partial\theta'} - E\left[\frac{\partial^2 \tilde{\ell}_{1,T}(\theta_T)}{\partial\theta\partial\theta'}\right]\right) \\ &= \tilde{H}_T^*\left(I_k + \tilde{\Delta}_T\right), \end{aligned} \tag{A.7}$$

where $\tilde{H}_T^* = \frac{1}{Tb}\sum_{t=1}^{T} k_{tT} E\left[\frac{\partial^2 \tilde{\ell}_{1,T}(\theta_T)}{\partial\theta\partial\theta'}\right]$ and $\tilde{\Delta}_T = \left(\tilde{H}_T^*\right)^{-1}\left(\tilde{H}_T - \tilde{H}_T^*\right)$. By Assumption 3(iii), for any $k \times 1$ vector $a = (a_1, \cdots, a_k)'$ such that $\|a\|^2 = 1$, there exists $v > 0$ such that for all $t \geqslant 1$,

$$a' E\left[\frac{\partial^2 \tilde{\ell}_{1,T}(\theta_T)}{\partial\theta\partial\theta'}\right] a \geqslant 1/v > 0.$$

Thus, we have,

$$\min_{\|a\|=1} a'\tilde{H}_{T,1}a = \min_{\|a\|=1}\left(\frac{1}{Tb}\sum_{t=1}^{T} k_{tT} a' E\left[\frac{\partial^2 \tilde{\ell}_{1,T}(\theta_T)}{\partial\theta\partial\theta'}\right] a\right) \geqslant \frac{1}{v}\left(\frac{1}{Tb}\sum_{t=1}^{T} k_{tT}\right) > 0.$$

This means that the smallest eigenvalue of $\tilde{H}_{T,1}$ is not smaller than $1/v > 0$, which further implies that

$$\left\|\left(\tilde{H}_T^*\right)^{-1}\right\|_{sp} = O_p(1),$$

where $\|\cdot\|_{sp}$ denotes the spectral norm. In addition, by Assumption 3(iii), we have

$$\left\|\tilde{H}_T - \tilde{H}_T^*\right\|_{sp} = o_p(1).$$

Then,

$$\left\|\tilde{H}_T^{-1}(\theta_T)\right\|_{sp} \leqslant \left\|\left(\tilde{H}_T^*\right)^{-1}\right\|_{sp}\left(1 - \left\|\tilde{H}_T - \tilde{H}_T^*\right\|_{sp}\right)^{-1} = O_p(1),$$

which implies that $\left\|H_T^{-1}(\theta_T)\right\|_{sp} = O_p(1)$.

Proof of (A.5). We have that

$$\begin{aligned} S_T(\theta_T) = \frac{\partial L_T(\theta_T)}{\partial \theta} &= \frac{1}{Tb}\sum_{t=1}^{T} k_{tT}\frac{\partial \ell_{t,T}(\theta_T)}{\partial \theta} \\ &= \frac{1}{Tb}\sum_{t=1}^{T} k_{tT}\frac{\partial \ell_{t,T}(\theta_t)}{\partial \theta} + \frac{1}{Tb}\sum_{t=1}^{T} k_{tT}\frac{\partial^2 \ell_{t,T}(\overline{\theta}_T)}{\partial \theta \partial \theta'}(\theta_T - \theta_t) \\ &:= S_T(\theta_t) + B_T, \end{aligned} \tag{A.8}$$

where the second line follows from mean-value theorem. Let us first consider $S_T(\theta_t)$. Using the similar argument as in (A.2), we have

$$\left\| S_T(\theta_t) - \tilde{S}_T(\theta_t) \right\| = o_p(1).$$

where $\tilde{S}_T(\theta_t) = \frac{1}{Tb}\sum_{t=1}^{T} k_{tT}\frac{\partial \tilde{\ell}_{1,T}(\theta_t)}{\partial \theta}$. By Assumption 3(ii), we have $\left\| \tilde{S}_T(\theta_t) \right\| = O_p\left(\frac{1}{\sqrt{Tb}}\right)$. For $\tilde{B}_T$, first notice that by Assumption 1, we have

$$\theta_t \approx \theta_T + \theta_T^{(1)}\left(\frac{t-T}{T}\right) + \frac{\theta_T^{(2)}}{2}\left(\frac{t-T}{T}\right)^2.$$

Then

$$\begin{aligned} \tilde{B}_T &= \frac{1}{Tb}\sum_{t=1}^{T} k_{tT}\left(\frac{\partial^2 \tilde{\ell}_{1,T}(\overline{\theta}_T)}{\partial \theta \partial \theta'} - E\left[\frac{\partial^2 \tilde{\ell}_{1,t}(\overline{\theta}_T)}{\partial \theta \partial \theta'}\right] + E\left[\frac{\partial^2 \tilde{\ell}_{1,T}(\overline{\theta}_T)}{\partial \theta \partial \theta'}\right]\right)\left(\theta_T^{(1)}\left(\frac{t-T}{T}\right) + \frac{\theta_T^{(2)}}{2}\left(\frac{t-T}{T}\right)^2\right) \\ &= \frac{1}{Tb}\sum_{t=1}^{T} k_{tT}\left(\frac{\partial^2 \tilde{\ell}_{1,t}(\overline{\theta}_T)}{\partial \theta \partial \theta'} - E\left[\frac{\partial^2 \tilde{\ell}_{1,T}(\overline{\theta}_T)}{\partial \theta \partial \theta'}\right] + E\left[\frac{\partial^2 \tilde{\ell}_{1,T}(\overline{\theta}_T)}{\partial \theta \partial \theta'}\right]\right)\theta_T^{(1)}\left(\frac{t-T}{T}\right) \\ &\quad + \frac{1}{Tb}\sum_{t=1}^{T} k_{tT}\left(\frac{\partial^2 \tilde{\ell}_{1,t}(\overline{\theta}_T)}{\partial \theta \partial \theta'} - E\left[\frac{\partial^2 \tilde{\ell}_{1,T}(\overline{\theta}_T)}{\partial \theta \partial \theta'}\right] + E\left[\frac{\partial^2 \tilde{\ell}_{1,T}(\overline{\theta}_T)}{\partial \theta \partial \theta'}\right]\right)\left(\frac{\theta_T^{(2)}}{2}\left(\frac{t-T}{T}\right)^2\right) \\ &:= \tilde{B}_{T,1} + \tilde{B}_{T,2}. \end{aligned}$$

Consider first $\tilde{B}_{T,1}$. We have

$$\begin{aligned} \tilde{B}_{T,1} =& \frac{1}{Tb}\sum_{t=1}^{T} k_{tT}\left(\frac{\partial^2 \tilde{\ell}_{1,T}(\theta_T)}{\partial \theta \partial \theta'} - E\left[\frac{\partial^2 \tilde{\ell}_{1,T}(\theta_T)}{\partial \theta \partial \theta'}\right]\right)\theta_T^{(1)}\left(\frac{t-T}{T}\right) \\ &+ \frac{1}{Tb}\sum_{t=1}^{T} k_{tT}\left(E\left[\frac{\partial^2 \tilde{\ell}_{1,t}(\theta_T)}{\partial \theta \partial \theta'}\right]\right)\theta_T^{(1)}\left(\frac{t-T}{T}\right). \end{aligned}$$

By Assumption 3(iii),

$$\left\| \frac{1}{Tb} \sum_{t=1}^{T} k_{tT} \left(\frac{\partial^2 \tilde{\ell}_{1,T}(\theta_T)}{\partial\theta\partial\theta'} - E\left[\frac{\partial^2 \tilde{\ell}_{1,t}(\theta_T)}{\partial\theta\partial\theta'} \right] \right) \theta_T^{(1)} \left(\frac{t-T}{T} \right) \right\| = o_p(1)$$

and

$$\left\| \frac{1}{Tb} \sum_{t=1}^{T} k_{tT} \left(E\left[\frac{\partial^2 \tilde{\ell}_{1,T}(\theta_T)}{\partial\theta\partial\theta'} \right] \right) \theta_T^{(1)} \left(\frac{t-T}{T} \right) \right\| \leqslant \mathcal{C} \frac{1}{Tb} \sum_{t=1}^{T} k_{tT} \left(\frac{t-T}{T} \right) \sim b \int_{\mathcal{B}} uK(u)du,$$

where $\mathcal{C}$ is a generic constant.Thus, we have $\left\| \tilde{B}_{T,1} \right\| = O_p(b)$.Similarly, we could show that $\left\| \tilde{B}_{T,2} \right\| = O_p(b^2)$. This implies that the dominating term is $\tilde{B}_{T,1}$ and we thus have $\left\| \tilde{B}_T \right\| = O_p(b)$. This further implies that $\left\| \tilde{S}_T(\theta_T) \right\| \leqslant \left\| \tilde{S}_T(\theta_t) \right\| + \left\| \tilde{B}_T \right\| = O_p\left(\frac{1}{\sqrt{Tb}} + b \right)$, which establishes (A.5).

Proof of (A.6). This follow immediately by the consistency: $\hat{\theta}_{K,b,T} \xrightarrow{p} \theta_T$.

Back to (A.3), under the condition $b = O(T^{-1/3})$, we have

$$\sqrt{Tb} \left(\hat{\theta}_{K,b,T} - \theta_T + \tilde{H}_T^{-1}(\theta_T) \tilde{B}_T \right) = -\tilde{H}_T^{-1}(\theta_T) \sqrt{Tb} \tilde{S}_T(\theta_T). \tag{A.9}$$

As the dominating term of the asymptotic bias is given by

$$\tilde{B}_T = -\frac{1}{Tb} \sum_{t=1}^{T} k_{tT} E\left[\frac{\partial^2 \tilde{\ell}_{1,T}(\theta_T)}{\partial\theta\partial\theta'} \right] \theta_T^{(1)} \left(\frac{t-T}{T} \right) (1 + o_p(1)).$$

It is straightforward to see the asymptotic bias term can be expressed as

$$\tilde{H}_T^{-1}(\theta_T) \tilde{B}_T = b\theta_T^{(1)} \mu_{1,K},$$

where $\mu_{1,K} = \int_{\mathbb{B}} uK(u)du$. By applying CLT on $\sqrt{Tb}\tilde{S}_{1,T}$, together with Slutsky's theorem, we obtain

$$\sqrt{Tb} \left(\hat{\theta}_{K,b,T} - \theta_T - b\theta_T^{(1)} \mu_{1,K} \right) \xrightarrow{d} \mathcal{N}(0, \phi_{0,K} \Sigma_T),$$

where $\Sigma_T = \tilde{\omega}_T^{-1} \Lambda_T \tilde{\omega}_T^{-1}$, $\tilde{\omega}_T = E\left[\frac{\partial^2 \tilde{\ell}_{1,T}(\theta_T)}{\partial\theta\partial\theta'} \right]$ and $\Lambda_T = Var\left(\frac{\partial \tilde{\ell}_{1,t}(\theta_T)}{\partial\theta'} \right)$.

A.3 Proof of Lemma 2

The objective function is given by

$$L_T(\theta, \theta^{(1)}) = \frac{1}{T\tilde{b}} \sum_{t=1}^{T} \tilde{k}_{tT} \ell_{t,T} \left(\theta + \theta^{(1)}(t/T - 1) \right).$$

Define $\beta_T = \theta_T + \theta_T^{(1)}(t/T - 1)$. Similarly as in (A.3), we have that

$$\begin{pmatrix} \tilde{\theta}_T - \theta_T \\ \tilde{\theta}_T^{(1)} - \theta_T^{(1)} \end{pmatrix} = - \begin{bmatrix} \frac{1}{T\tilde{b}} \sum_{t=1}^{T} \tilde{k}_{tT} \frac{\partial^2 \ell_{t,T}(\beta)}{\partial\theta\partial\theta'} & \frac{1}{T\tilde{b}} \sum_{t=1}^{T} \tilde{k}_{tT} \frac{\partial^2 \ell_{t,T}(\beta)}{\partial\theta\partial\theta'^{(1)}} \left(\frac{t-T}{T}\right) \\ \frac{1}{T\tilde{b}} \sum_{t=1}^{T} \tilde{k}_{tT} \frac{\partial^2 \ell_{t,T}(\beta)}{\partial\theta^{(1)}\partial\theta'} \left(\frac{t-T}{T}\right) & \frac{1}{T\tilde{b}} \sum_{t=1}^{T} \tilde{k}_{tT} \frac{\partial^2 \ell_{t,T}(\beta)}{\partial\theta^{(1)}\partial\theta'^{(1)}} \left(\frac{t-T}{T}\right)^2 \end{bmatrix}^{-1} \begin{bmatrix} \frac{1}{T\tilde{b}} \sum_{t=1}^{T} \tilde{k}_{tT} \frac{\partial \ell_{t,T}(\beta)}{\partial\theta} \\ \frac{1}{T\tilde{b}} \sum_{t=1}^{T} \tilde{k}_{tT} \frac{\partial \ell_{t,T}(\beta)}{\partial\theta^{(1)}} \left(\frac{t-T}{T}\right) \end{bmatrix} + o_p(1) \quad \text{(A.10)}$$

Using similar arguments for the proofs of (A.4)-(A.5), we have

$$\left\| \frac{1}{T\tilde{b}} \sum_{t=1}^{T} \tilde{k}_{tT} \frac{\partial^2 \ell_{t,T}(\beta)}{\partial\theta\partial\theta'} \right\| = O_p(1), \quad \left\| \frac{1}{T\tilde{b}} \sum_{t=1}^{T} \tilde{k}_{tT} \frac{\partial^2 \ell_{t,T}(\beta)}{\partial\theta\partial\theta'^{(1)}} \left(\frac{t-T}{T}\right) \right\| = O_p(\tilde{b})$$

$$\left\| \frac{1}{T\tilde{b}} \sum_{t=1}^{T} \tilde{k}_{tT} \frac{\partial^2 \ell_{t,T}(\beta)}{\partial\theta^{(1)}\partial\theta'} \left(\frac{t-T}{T}\right) \right\| = O_p(\tilde{b}), \quad \left\| \frac{1}{T\tilde{b}} \sum_{t=1}^{T} \tilde{k}_{tT} \frac{\partial^2 \ell_{t,T}(\beta)}{\partial\theta^{(1)}\partial\theta'^{(1)}} \left(\frac{t-T}{T}\right)^2 \right\| = O_p(\tilde{b}^2).$$

Moreover, since

$$\theta_t \approx \theta_T + \theta_T^{(1)} \left(\frac{t-T}{T}\right) + \frac{\theta_T^{(2)}}{2} \left(\frac{t-T}{T}\right)^2,$$

following again the proofs of (A.4)-(A.5), we have

$$\begin{aligned} \frac{1}{T\tilde{b}} \sum_{t=1}^{T} \tilde{k}_{tT} \frac{\partial \ell_{t,T}(\beta)}{\partial\theta} &= \frac{1}{T\tilde{b}} \sum_{t=1}^{T} \tilde{k}_{tT} \frac{\partial \ell_{t,T}(\theta_t)}{\partial\theta} + \frac{1}{T\tilde{b}} \sum_{t=1}^{T} \tilde{k}_{tT} \frac{\partial^2 \ell_{t,T}(\overline{\theta}_T)}{\partial\theta\partial\theta'} \left(\theta_T + \theta_T^{(1)} \left(\frac{t-T}{T}\right) - \theta_t\right) \\ &= \frac{1}{T\tilde{b}} \sum_{t=1}^{T} \tilde{k}_{tT} \frac{\partial \ell_{t,T}(\theta_t)}{\partial\theta} + \frac{1}{T\tilde{b}} \sum_{t=1}^{T} \tilde{k}_{tT} \frac{\partial^2 \ell_{t,T}(\overline{\theta}_T)}{\partial\theta\partial\theta'} \frac{\theta_T^{(2)}}{2} \left(\frac{t-T}{T}\right)^2 \\ &= O_p\left((T\tilde{b})^{-1/2}\right) + O_p\left(\tilde{b}^2\right), \end{aligned}$$

and

$$\begin{aligned} \frac{1}{T\tilde{b}} \sum_{t=1}^{T} \tilde{k}_{tT} \frac{\partial \ell_{t,T}(\beta)}{\partial\theta^{(1)}} \left(\frac{t-T}{T}\right) &= \frac{1}{T\tilde{b}} \sum_{t=1}^{T} \tilde{k}_{tT} \frac{\partial \ell_{t,T}(\theta_t)}{\partial\theta^{(1)}} \left(\frac{t-T}{T}\right) + \frac{1}{T\tilde{b}} \sum_{t=1}^{T} \tilde{k}_{tT} \frac{\partial^2 \ell_{t,T}(\overline{\theta}_T)}{\partial\theta^{(1)}\partial\theta'^{(1)}} \frac{\theta_T^{(2)}}{2} \left(\frac{t-T}{T}\right)^3 \\ &= O_p((T\tilde{b})^{-1/2}\tilde{b}) + O_p(\tilde{b}^3) \end{aligned}$$

where $\overline{\theta}_T$ lies between θ_t and $\theta_T + \theta_T^{(1)}\left(\frac{t-T}{T}\right)$. It follows that

$$\begin{pmatrix} \tilde{\theta}_T - \theta_T \\ \tilde{\theta}_T^{(1)} - \theta_T^{(1)} \end{pmatrix} = - \begin{bmatrix} O_p(1) & O_p(\tilde{b}) \\ O_p(\tilde{b}) & O_p(\tilde{b}^2) \end{bmatrix}^{-1} \begin{bmatrix} O_p\left((T\tilde{b})^{-1/2}\right) + O_p\left(\tilde{b}^2\right) \\ O_p((T\tilde{b})^{-1/2}\tilde{b}) + O_p(\tilde{b}^3) \end{bmatrix} + o_p(1) \tag{A.11}$$

$$= \begin{bmatrix} O_p\left((T\tilde{b})^{-1/2} + \tilde{b}^2\right) \\ O_p\left(T^{-1/2}\tilde{b}^{-3/2} + \tilde{b}\right) \end{bmatrix} \tag{A.12}$$

Therefore, we obtain the consistency rate for $\tilde{\theta}_T$:

$$\left\|\tilde{\theta}_T - \theta_T\right\| = O_p\left((T\tilde{b})^{-1/2} + \tilde{b}^2\right).$$

A.4 Auxiliary Lemmas

Here, we present two auxiliary lemmas. See Online Supplement for the proof of Lemma 3.

Lemma 3. *Suppose that Assumptions 1, 2, 3 and 4(i) hold with $b \to 0$ and $Tb \to \infty$. Then, for some $0 < \delta < \frac{1}{2}$, it holds that*

$$\sup_{b \in I_T} \left\|\hat{\theta}_{K,b,T} - \theta_T\right\| = O_p(r_{T,b,\delta}), \tag{A.13}$$

where $r_{T,b,\delta} = T^{-1/2}b^{-1/2+\delta} + b^{1-\delta}$.

Lemma 4. *Define*

$$L(b) = \left(\hat{\theta}_{b,T} - \theta_T\right)' \omega_T(\theta_T) \left(\hat{\theta}_{b,T} - \theta_T\right),$$
$$A(b) = \left(\hat{\theta}_{b,T} - \tilde{\theta}_T\right)' \omega_T\left(\tilde{\theta}_T\right) \left(\hat{\theta}_{b,T} - \tilde{\theta}_T\right),$$

where $\hat{\theta}_{b,T} = \hat{\theta}_{\overline{K},b,T}$ and $\omega_T(\theta) = E_T\left(\frac{\partial^2 \ell_{T+h}(\theta)}{\partial\theta\partial\theta'}\right)$. Suppose that Assumptions 1-5 hold, we have

$$\sup_{b \in I_T} \left|\frac{L(b) - A(b)}{L(b)}\right| = o_p(1). \tag{A.14}$$

Proof. Recall that $\omega_T(\theta) = E\left[\frac{\partial^2 \ell_{T+h}(\theta)}{\partial\theta\partial\theta'}\right]$. Define

$$\omega_T^{(1)}(\theta_T) = \left[\frac{\partial \omega_T(\theta_T)}{\partial [\theta_T]_1} \cdots \frac{\partial \omega_T(\theta_T)}{\partial [\theta_T]_d}\right] \left(\tilde{\theta}_T - \theta_T\right),$$

where $[\theta_T]_s$ denotes the s^{th} elements of the $d \times 1$ vector θ_T. Let us first expand $A(b)$:

$$\begin{aligned}
A(b) &= \left(\hat{\theta}_{b,T} - \tilde{\theta}_T\right)' \omega_T\left(\tilde{\theta}_T\right)\left(\hat{\theta}_{b,T} - \tilde{\theta}_T\right) \\
&= \left(\hat{\theta}_{b,T} - \theta_T + \theta_T + \tilde{\theta}_T\right)' \left(\omega_T(\theta_T) + \omega_T^{(1)}(\theta_T)\right)\left(\hat{\theta}_{b,T} - \theta_T + \theta_T + \tilde{\theta}_T\right) \\
&= L(b) - 2\left(\hat{\theta}_{b,T} - \theta_T\right)' \omega_T(\theta_T)\left(\tilde{\theta}_T - \theta_T\right) + \left(\tilde{\theta}_T - \theta_T\right)' \omega_T(\theta_T)\left(\tilde{\theta}_T - \theta_T\right) \\
&\quad + \left(\hat{\theta}_{b,T} - \theta_T\right)' \omega_T^{(1)}(\theta_T)\left(\hat{\theta}_{b,T} - \theta_T\right) - 2\left(\hat{\theta}_{b,T} - \theta_T\right)' \omega_T^{(1)}(\theta_T)\left(\tilde{\theta}_T - \theta_T\right) \\
&\quad + \left(\tilde{\theta}_T - \theta_T\right)' \omega_T^{(1)}(\theta_T)\left(\tilde{\theta}_T - \theta_T\right) \\
&:= L(b) - 2D_1(b) + D_1' + D_2(b) - 2D_3(b) + D_2',
\end{aligned}$$

where

$$\begin{aligned}
&D_1(b) = \left(\hat{\theta}_{b,T} - \theta_T\right)' \omega_T(\theta_T)\left(\tilde{\theta}_T - \theta_T\right), \quad D_1' = \left(\tilde{\theta}_T - \theta_T\right)' \omega_T(\theta_T)\left(\tilde{\theta}_T - \theta_T\right), \\
&D_2(b) = \left(\hat{\theta}_{b,T} - \theta_T\right)' \omega_T^{(1)}(\theta_T)\left(\hat{\theta}_{b,T} - \theta_T\right), \quad D_3(b) = \left(\hat{\theta}_{b,T} - \theta_T\right)' \omega_T^{(1)}(\theta_T)\left(\tilde{\theta}_T - \theta_T\right), \\
&D_2' = \left(\tilde{\theta}_T - \theta_T\right)' \omega_T^{(1)}(\theta_T)\left(\tilde{\theta}_T - \theta_T\right).
\end{aligned}$$

Then, we have

$$\frac{L(b) - A(b)}{L(b)} = \frac{2D_1(b)}{L(b)} - \frac{D_1'}{L(b)} - \frac{D_2(b)}{L(b)} + \frac{D_3(b)}{L(b)} - \frac{D_2'}{L(b)}.$$

By Lemma 2 and Assumption 5(i), we have

$$\left\|\tilde{\theta}_T - \theta_T\right\| = O_p\left((T\tilde{b})^{-1/2}\right). \tag{A.15}$$

We will show that

$$\sup_{b\in I_T}\left|\frac{D_1(b)}{L(b)}\right| = o_p(1), \ \sup_{b\in I_T}\left|\frac{D_2(b)}{L(b)}\right| = o_p(1), \ \sup_{b\in I_T}\left|\frac{D_3(b)}{L(b)}\right| = o_p(1), \tag{A.16}$$

$$\sup_{b\in I_T}\left|\frac{D_1'}{L(b)}\right| = o_p(1), \ \sup_{b\in I_T}\left|\frac{D_2'}{L(b)}\right| = o_p(1). \tag{A.17}$$

These bounds together with triangular inequality imply (A.14).

Proof of (A.16). First, by Lemma 3 and Assumption 3(iii), $\|\omega_T(\theta_T)\|_{sp} = O_p(1)$ and

$$\sup_{b\in I_T}|L(b)| \leqslant \sup_{b\in I_T}\left\|\hat{\theta}_{b,T} - \theta_T\right\| \ \|\omega_T(\theta_T)\|_{sp} \ \sup_{b\in I_T}\left\|\hat{\theta}_{b,T} - \theta_T\right\| = O_p\left(r_{T,b,\delta}^2\right), \tag{A.18}$$

for some $0 < \delta < 1/2$. Write $\tilde{r}_{T,\tilde{b}} = (T\tilde{b})^{-1/2}$, we also have

$$\begin{aligned}
\sup_{b\in I_T} |D_1(b)| &\leqslant \sup_{b\in I_T} \left\|\hat{\theta}_{b,T} - \theta_T\right\| \ \left\|\omega_T(\theta_T)\right\|_{sp} \ \left\|\tilde{\theta}_T - \theta_T\right\| = O_p\left(r_{T,b,\delta}\tilde{r}_{T,\tilde{b}}\right), \\
\sup_{b\in I_T} |D_2(b)| &\leqslant \sup_{b\in I_T} \left\|\hat{\theta}_{b,T} - \theta_T\right\| \ \left\|\omega_T^{(1)}(\theta_T)\right\|_{sp} \sup_{b\in I_T} \left\|\hat{\theta}_{b,T} - \theta_T\right\| = O_p\left(r_{T,b,\delta}^2\tilde{r}_{T,\tilde{b}}\right), \\
\sup_{b\in I_T} |D_3(b)| &\leqslant \sup_{b\in I_T} \left\|\hat{\theta}_{b,T} - \theta_T\right\| \ \left\|\omega_T^{(1)}(\theta_T)\right\|_{sp} \left\|\tilde{\theta}_T - \theta_T\right\| = O_p\left(r_{T,b,\delta}\tilde{r}_{T,\tilde{b}}^2\right),
\end{aligned}$$

where the second and third line follow from the fact that $\omega_T^{(1)}(\theta_T)$ involves $\tilde{\theta}_T - \theta_T$ so the order of $\left\|\omega_T^{(1)}(\theta_T)\right\|_{sp} = O_p(\tilde{r}_{T,\tilde{b}})$, which is determined by $\left\|\tilde{\theta}_T - \theta_T\right\|$. These bounds imply that

$$\sup_{b\in I_T} \left|\frac{D_1(b)}{L(b)}\right| = O_p\left(\frac{\tilde{r}_{T,\tilde{b}}}{r_{T,b,\delta}}\right) = o_p(1),$$

where $\frac{\tilde{r}_{T,\tilde{b}}}{r_{T,b,\delta}} \to 0$ is guaranteed by Assumption 5. Similarly, we have

$$\sup_{b\in I_T} \left|\frac{D_2(b)}{L(b)}\right| = O_p\left(\tilde{r}_{T,\tilde{b}}\right) = o_p(1),$$

as $T\tilde{b} \to \infty$. Finally, we have

$$\sup_{b\in I_T} \left|\frac{D_3(b)}{L(b)}\right| = O_p\left(\frac{\tilde{r}_{T,\tilde{b}}^2}{r_{T,b,\delta}}\right) = o_p(1),$$

where $\frac{\tilde{r}_{T,\tilde{b}}^2}{r_{T,b,\delta}} \to 0$ is again guaranteed by Assumption 5.

Proof of (A.17). First, it is straightforward to show that

$$|D_1'| = O_p(\tilde{r}_{T,\tilde{b}}^2), \ \ |D_2'| = O_p(\tilde{r}_{T,\tilde{b}}^3).$$

Together with (A.18) and following the same reasoning above, we have

$$\sup_{b\in I_T} \left|\frac{D_1'}{L(b)}\right| = O_p\left(\frac{\tilde{r}_{T,\tilde{b}}^2}{r_{T,b,\delta}^2}\right) = o_p(1), \ \ \sup_{b\in I_T} \left|\frac{D_2'}{L(b)}\right| = O_p\left(\frac{\tilde{r}_{T,\tilde{b}}^3}{r_{T,b,\delta}^2}\right) = o_p(1).$$

□

A.5 Proof of Theorem 1

For a given kernel function $K = \bar{K}$, write $\hat{\theta}_{\overline{K},b,T} = \hat{\theta}_{b,T}$ and $\omega_T(\theta_T) = E_T\left(\frac{\partial^2 \ell_{T+h}(\theta_T)}{\partial\theta\partial\theta'}\right)$. It follows from Lemma 1 that, the infeasible objective function can be written as

$$\left(\hat{\theta}_{b,T} - \theta_T\right)' \omega_T(\theta_T) \left(\hat{\theta}_{b,T} - \theta_T\right) = r_{T,b} q_T,$$

where q_T is a scalar $O_p(1)$ random variable and $r_{T,b} = (Tb)^{-1/2} + b$. The first-order condition of $r_{T,b}$ with respect to b gives $\hat{b} = O_p(T^{-\frac{1}{3}})$. Since the second order derivative of $r_{T,b}$ is always positive, the optimal bandwidth minimize the objective function.

A.6 Proof of Theorem 2

Write $\hat{\theta}_{\overline{K},b,T} = \hat{\theta}_{b,T}$ and $\omega_T(\theta_T) = E_T\left(\frac{\partial^2 \ell_{T+h}(\theta_T)}{\partial\theta\partial\theta'}\right)$. Let

$$\hat{b} := \arg\min_{b \in I_T} (\hat{\theta}_{b,T} - \tilde{\theta}_T)' \omega_T(\tilde{\theta}_T) (\hat{\theta}_{b,T} - \tilde{\theta}_T)$$

be the bandwidth selected according to the feasible criterion. As in the proof of Lemma 4, the decomposition of $A(b)$ implies that

$$A(\hat{b}) = L(\hat{b}) - 2D_1(\hat{b}) + D_1' + D_2(\hat{b}) - 2D_3(\hat{b}) + D_2'.$$

Then, we have

$$\begin{aligned}\frac{A(\hat{b})}{\inf_{b\in I_T} L(b)} &= \frac{L(\hat{b})}{\inf_{b\in I_T} L(b)} - \frac{2D_1(\hat{b})}{\inf_{b\in I_T} L(b)} + \frac{D_2(\hat{b})}{\inf_{b\in I_T} L(b)} - \frac{2D_3(\hat{b})}{\inf_{b\in I_T} L(b)} + \frac{D_1'}{\inf_{b\in I_T} L(b)} + \frac{D_2'}{\inf_{b\in I_T} L(b)} \\ &:= I_1(\hat{b}) + I_2(\hat{b}) + I_3(\hat{b}) + I_4(\hat{b}) + I_5 + I_6.\end{aligned}$$

Following (A.16) and (A.17), we have

$$I_2(\hat{b}) = o_p(1), \;\; I_3(\hat{b}) = o_p(1), \;\; I_4(\hat{b}) = o_p(1), \;\; I_5 = o_p(1), \;\; I_6 = o_p(1).$$

To proof that $A(\hat{b})/\inf_{b\in I_T} L(b) \to^p 1$, it is suffice to establish that

$$I_1(\hat{b}) \xrightarrow{p} 1. \tag{A.19}$$

For any $b, b' \in I_T$, it follows immediately from Lemma 4 that

$$\sup_{b,b' \in I_T} \left| \frac{L(b) - L(b') - (A(b) - A(b'))}{L(b) + L(b')} \right| \leqslant \sup_{b \in I_T} \left| \frac{L(b) - A(b)}{L(b)} \right| + \sup_{b' \in I_T} \left| \frac{L(b') - A(b')}{L(b')} \right| = o_p(1).$$

This implies that for any $\epsilon > 0$,

$$P\left[\frac{L(\hat{b}) - L(\hat{b}') - (A(\hat{b}) - A(\hat{b}'))}{L(\hat{b}) + L(\hat{b}')} \leqslant \epsilon \right] \to 1.$$

Thus, by rearranging terms, we obtain

$$(1 - \epsilon)L(\hat{b}) - (1 + \epsilon)L(\hat{b}') \leqslant A(\hat{b}) - A(\hat{b}')) \leqslant 0 \qquad a.s.$$

Then, we have

$$1 \leqslant \frac{L(\hat{b})}{L(\hat{b}')} \leqslant \frac{1 + \epsilon}{1 - \epsilon} \qquad a.s.$$

This completes the proof of (A.19).

A.7 Proof of Theorem 3

Let $\{P_n(u)\}_{n=0}^{\infty}$ denote the shifted Legendre polynomials on $[-1, 0]$. That is, $P_n(u) = Q_n(2u+1)$, where $\{Q_n(u)\}_{n=0}^{\infty}$ are the standard Legendre polynomials on $[-1, 1]$. For example,

$$P_0(u) = 1, \quad P_1(u) = 2u + 1, \text{ and } P_2(u) = 6u^2 + 6u + 1.$$

Since $\{Q_n(u)\}_{n=0}^{\infty}$ forms an orthogonal basis in the Hilbert space $L^2([-1, 1])$, it follows by a change of variables that, for any $n, m \geq 0$,

$$\int_{-1}^{0} P_n(u) P_m(u) du = \frac{1}{2n + 1} \delta_{nm}, \tag{A.20}$$

where δ_{nm} is the Kronecker delta, equal to one if $m = n$ and zero otherwise. Moreover, since $\int_{-1}^{1} Q_n(u)\, du = 0$ for all $n \geq 1$, we also have

$$\int_{-1}^{0} P_n(u) = 0, \text{ for all } n \geq 1. \tag{A.21}$$

The basis $\{P_n\}_{n=0}^{\infty}$ is now used to expand the kernel function $K(u)$ into orthogonal series,

that is,

$$K(u) = \sum_{n=0}^{\infty} c_n P_n(u),$$

where $K(\cdot) \in L^2([-1,0]) = \{f(u) : \int_{-1}^{0} f^2(u)du < \infty\}$, in which the inner product is given by $\langle f_1, f_2\rangle = \int_{-1}^{0} f_1(u)f_2(u)du$ and the induced form $\|f\|^2 = \langle f, f\rangle$. To extract the coefficient c_n, take the inner product of both sides with $P_n(u)$:

$$\langle K(u), P_n(u)\rangle = \sum_{m=0}^{\infty} c_n \langle P_m(u), P_n(u)\rangle = c_n \langle P_n(u), P_n(u)\rangle,$$

which simplifies to

$$c_n = \frac{\langle K(u), P_n(u)\rangle}{\langle P_n(u), P_n(u)\rangle} = \frac{\int_{-1}^{0} K(u)P_n(u)du}{\int_{-1}^{0} P_n(u)^2 du}.$$

Moreover, from (A.20), $\int_{-1}^{0} P_n(u)^2 du = 1/(2n+1)$ and

$$c_n = (2n+1)\int_{-1}^{0} K(u)P_n(u)du.$$

It follows that $c_0 = \int_{-1}^{0} K(u)P_0(u)du = \int_{-1}^{0} K(u)\,du = 1$ by Assumption 4(i).

Our objective function is

$$\mathcal{L} := Q(K) = -\int_{-1}^{0} uK(u)du \left(\int_{-1}^{0} K^2(u)du\right). \tag{A.22}$$

We first derive properties of the first term. By definition,

$$-\int_{-1}^{0} u\,K(u)\,du = -\int_{-1}^{0} u\left(\sum_{n=0}^{\infty} c_n P_n(u)\right)du = -\sum_{n=0}^{\infty} c_n \int_{-1}^{0} u\,P_n(u)\,du.$$

By the definition of the shifted Legendre polynomials and a change of variables:

$$\begin{aligned}\int_{-1}^{0} u\,P_n(u)\,du &= \int_{-1}^{0} u\,Q_n(2u+1)\,du = \frac{1}{4}\int_{-1}^{1}(x-1)Q_n(x)\,dx \\ &= \frac{1}{4}\left(\int_{-1}^{1} xQ_n(x)\,dx - \int_{-1}^{1} Q_n(x)\,dx\right).\end{aligned}$$

Using orthogonality properties of the standard Legendre polynomials:

$$\int_{-1}^{1} Q_n(x)\,dx = \begin{cases} 0, & n \geq 1, \\ 2, & n = 0, \end{cases} \quad \text{and} \quad \int_{-1}^{1} x\,Q_n(x)\,dx = \begin{cases} 0, & n \neq 1, \\ \dfrac{2}{3}, & n = 1, \end{cases}$$

we obtain:

$$\int_{-1}^{0} u\, P_n(u)\, du = \begin{cases} -\dfrac{1}{2}, & \text{if } n = 0, \\ \dfrac{1}{6}, & \text{if } n = 1, \\ 0, & \text{if } n \geq 2. \end{cases}$$

Consequently,

$$-\int_{-1}^{0} u\, K(u)\, du = \sum_{n=0}^{\infty} c_n \left(-\int_{-1}^{0} u P_n(u)\, du \right) = \frac{1}{2}c_0 - \frac{1}{6}c_1 = \frac{1}{2} - \frac{1}{6}c_1$$

and the objective function

$$\mathcal{L} = \left(1 + \sum_{n=1}^{\infty} \frac{c_n^2}{2n+1} \right) \left(\frac{1}{2} - \frac{c_1}{6} \right).$$

Now, we investigate the property of the second term in the objective function. Since $K(u) \in L^2([-1, 0])$ admits an expansion in terms of the shifted Legendre polynomials,

$$\begin{aligned} \int_{-1}^{0} K^2(u) du &= \int_{-1}^{0} \left(1 + \sum_{n=1}^{\infty} c_n P_n(u) \right)^2 du \\ &= 1 + 2 \int_{-1}^{0} \sum_{n=1}^{\infty} c_n P_n(u) du + \int_{-1}^{0} \left(\sum_{n=1}^{\infty} c_n P_n(u) \right)^2 du. \end{aligned}$$

Now, by the uniform convergence of the summation and from(A.21),

$$2 \int_{-1}^{0} \sum_{n=1}^{\infty} c_n P_n(u) du = 2c_n \sum_{n=1}^{\infty} \int_{-1}^{0} P_n(u) du = 0.$$

Moreover,

$$\begin{aligned} \int_{-1}^{0} \left(\sum_{n=1}^{\infty} c_n P_n(u) \right)^2 du &= \int_{-1}^{0} \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} c_n c_m P_n(u) P_m(u)\, du \\ &= \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} c_n c_m \int_{-1}^{0} P_n(u) P_m(u)\, du \\ &= \sum_{n=1}^{\infty} c_n^2 \int_{-1}^{0} P_n^2(u)\, du = \sum_{n=1}^{\infty} \frac{c_n^2}{2n+1}, \end{aligned}$$

using the result in (A.20). Consequently,

$$\int_{-1}^{0} K^2(u)du = 1 + \sum_{n=1}^{\infty} \frac{c_n^2}{2n+1}.$$

Next, we turn to the minimization problem. Let $A := \sum_{n=2}^{\infty} \frac{c_n^2}{2n+1}$, so the objective becomes:

$$\mathcal{L} = \left(1 + \frac{c_1^2}{3} + A\right)\left(\frac{1}{2} - \frac{c_1}{6}\right).$$

Observe that A is a sum of non-negative terms, i.e., $A \geq 0$. Moreover, the partial derivative of $\mathcal{L}$ with respect to A is:

$$\frac{\partial \mathcal{L}}{\partial A} = \frac{1}{2} - \frac{c_1}{6},$$

which is positive whenever $c_1 < 3$. Hence, for fixed c_1 within this range, increasing A increases $\mathcal{L}$. Therefore, to minimize $\mathcal{L}$, it is optimal to choose $A = 0$, which is achieved by setting $c_n = 0$ for all $n \geq 2$.

Consequently, the kernel function is approximated by the first two polynomials such that

$$K(u) = c_0 P_0(u) + c_1 P_1(u) = 1 + c_1(2u + 1),$$

which is linear in u. By assumption, $K(u) = 1 + c_1(2u + 1) \geq 0$ for all $u \in [-1, 0]$, implying that $c_1 \in [-1, 1]$. The objective function becomes:

$$\mathcal{L} = \left(1 + \frac{c_1^2}{3}\right)\left(\frac{1}{2} - \frac{c_1}{6}\right).$$

The first order derivative with respect to c_1 is given by:

$$\frac{d\mathcal{L}}{c_1} = \frac{1}{6}(2c_1 - 1 - c_1^2) < 0$$

for $c_1 \in [-1, 1]$, so that the minimum is achieved at the boundary $c_1 = 1$. Therefore, the optimal kernel function is given by:

$$K(u) = 1 + (2u + 1) = 2(1 - |u|), \ u \in [-1, 0].$$

Finally, following the similar steps as the proofs of Theorem 2 in Cheng et al. (1997), we can show that solution of the optimization problem (12) exists.

Online Supplement:

Optimal bandwidth selection for forecasting under parameter instability

NOT FOR PUBLICATION

This Online Supplement is organized as follows. Section S1 provides the proof of Lemma 3. Section S2 reports additional simulation results for the structural break case. Section S3 details the implementation of the forecast combination methods used in the applications on bond return predictability and inflation forecasting. Section S4 presents an empirical application to real-time inflation forecasting using financial variables.

S1 Proof of Lemma 3

Given the kernel function $\overline{K}$, write $\hat{\theta}_{\overline{K},b,T} = \hat{\theta}_{b,T}$. As in (A.3), the estimator can be decomposed as

$$\begin{aligned}\hat{\theta}_{b,T} - \theta_T &= -H_T(\theta_T)S_T(\theta_T) + o_p(1) \\ &= -H_T(\theta_T)\left(S_T(\theta_t) + B_T\right) + o_p(1),\end{aligned} \tag{S.1}$$

where

$$S_T(\theta_t) = \frac{1}{Tb}\sum_{t=1}^{T} k_{tT}\frac{\partial \ell_{t,T}\left(\theta_t\right)}{\partial\theta}, H_T(\theta_T) = \left(\frac{1}{Tb}\sum_{t=1}^{T} k_{tT}\frac{\partial^2 \ell_{t,T}\left(\theta_T\right)}{\partial\theta\partial\theta'}\right)^{-1},$$
$$B_T = \frac{1}{Tb}\sum_{t=1}^{T} k_{tT}\frac{\partial^2 \ell_{t,T}\left(\overline{\theta}_T\right)}{\partial\theta\partial\theta'}\left(\theta_T - \theta_t\right),$$

and $\overline{\theta}_T$ lies between θ_T and θ_t. We will show that

$$\sup_{b\in I_T} \left\|T^{1/2}b^{1/2+\delta}S_T(\theta_t)\right\| = O_p(1), \tag{S.2}$$

$$\sup_{b\in I_T} \left\|H_T(\theta_T)^{-1}\right\| = O_p(1), \tag{S.3}$$

$$\sup_{b\in I_T} \left\|b^{\delta}B_T\right\| = O_p(b), \tag{S.4}$$

for some $0 < \delta < 1/2$. These bounds together with (S.1) prove (A.13).

Proof of (S.2). By Boole's inequality and Chebyshev's inequality, we have, for any $\varepsilon > 0$,

$$\begin{aligned}\mathbb{P}\left(\sup_{b\in I_T}\left\|\frac{1}{T^{1/2}b^{1/2-\delta}}\sum_{t=1}^{T}k_{tT}\frac{\partial\ell_{t,T}(\theta_t)}{\partial\theta}\right\|>\varepsilon\right)&\leqslant\sum_{b\in I_T}\mathbb{P}\left(\left\|\frac{1}{T^{1/2}b^{1/2-\delta}}\sum_{t=1}^{T}k_{tT}\frac{\partial\ell_{t,T}(\theta_t)}{\partial\theta}\right\|>\varepsilon\right)\\&\leqslant\lambda(I_T)\times\sup_{b\in I_T}\mathbb{P}\left(\left\|\frac{1}{T^{1/2}b^{1/2-\delta}}\sum_{t=1}^{T}k_{tT}\frac{\partial\ell_{t,T}(\theta_t)}{\partial\theta}\right\|>\varepsilon\right)\\&\leqslant\lambda(I_T)\times O(b^{-\delta})=O(1),\end{aligned}$$

where the third inequality follows from the proof of (A.5) since $\left\|\frac{1}{T^{1/2}b^{1/2}}\sum_{t=1}^{T}k_{tT}\frac{\partial\ell_{t,T}(\theta_t)}{\partial\theta}\right\| = O_p(1)$. The final equality follows from Assumption 6.

Proof of (S.3). Recall that

$$\begin{aligned}\tilde{H}_T=&\frac{1}{Tb}\sum_{t=1}^{T}k_{tT}\frac{\partial^2\tilde{\ell}_{1,t}(\theta_T)}{\partial\theta\partial\theta'}\\=&\frac{1}{Tb}\sum_{t=1}^{T}k_{tT}E\left[\frac{\partial^2\tilde{\ell}_{1,t}(\theta_T)}{\partial\theta\partial\theta'}\right]+\frac{1}{Tb}\sum_{t=1}^{T}k_{tT}\left(\frac{\partial^2\tilde{\ell}_{1,t}(\theta_T)}{\partial\theta\partial\theta'}-E\left[\frac{\partial^2\tilde{\ell}_{1,T}(\theta_T)}{\partial\theta\partial\theta'}\right]\right)\\:=&\tilde{H}_{T,1}\left(I_k+\tilde{\Delta}_T\right),\end{aligned}\tag{S.5}$$

where $\tilde{\Delta}_T=\left(\tilde{H}_{T,1}\right)^{-1}\left(\tilde{H}_T-\tilde{H}_{T,1}\right)$. First, (A.4) holds uniformly over b:

$$\sup_{b\in I_T}\left\|\tilde{H}_{T,1}^{-1}\right\|_{sp}=O_p(1).\tag{S.6}$$

For $\tilde{\Delta}_T$, let $\tilde{\Delta}_t=\frac{\partial^2\tilde{\ell}_{1,t}(\theta_T)}{\partial\theta\partial\theta'}-E\left[\frac{\partial^2\tilde{\ell}_{1,T}(\theta_T)}{\partial\theta\partial\theta'}\right]$. Then, for any $\varepsilon>0$, , by Boole's inequality and Chebyshev's inequality, we have

$$\begin{aligned}\mathbb{P}\left(\sup_{b\in I_T}\left\|\frac{1}{Tb}\sum_{t=1}^{T}k_{tT}\tilde{\Delta}_t\right\|>\varepsilon\right)&\leqslant\sum_{b\in I_T}\mathbb{P}\left(\left\|\frac{1}{Tb}\sum_{t=1}^{T}k_{tT}\tilde{\Delta}_t\right\|>\varepsilon\right)\\&\leqslant\underbrace{|I_T|}_{O(b^{\delta})}\times\underbrace{\sup_{b\in I_T}\mathbb{P}\left(\left\|\frac{1}{Tb}\sum_{t=1}^{T}k_{tT}\tilde{\Delta}_t\right\|>\varepsilon\right)}_{o(1)}=o(1).\end{aligned}\tag{S.7}$$

To sum up, we continue from (A.7):

$$\sup_{b\in I_T}\left\|\tilde{H}_T^{-1}\right\|_{sp}\leqslant\underbrace{\sup_{b\in I_T}\left\|\tilde{H}_{T,1}^{-1}\right\|_{sp}}_{O_p(1)\text{ by (S.6)}}\left(1-\underbrace{\sup_{b\in I_T}\left\|\tilde{\Delta}_T\right\|_{sp}}_{o_p(1)\text{ by (S.7)}}\right)^{-1}=O_p(1).$$

This also implies (S.3).

Proof of (S.4). Recall that the stationary approximation of B_T is $\tilde{B}_T$, where $\tilde{B}_T = \tilde{B}_{T,1} + \tilde{B}_{T,2}$:

$$\tilde{B}_{T,1} = \frac{1}{Tb}\sum_{t=1}^{T} k_{tT}\left(\frac{\partial^2 \tilde{\ell}_{1,T}(\theta_T)}{\partial\theta\partial\theta'} - E\left[\frac{\partial^2 \tilde{\ell}_{1,T}(\theta_T)}{\partial\theta\partial\theta'}\right]\right)(\theta_T - \theta_t),$$
$$\tilde{B}_{T,2} = \frac{1}{Tb}\sum_{t=1}^{T} k_{tT} E\left[\frac{\partial^2 \tilde{\ell}_{1,T}(\theta_T)}{\partial\theta\partial\theta'}\right](\theta_T - \theta_t).$$

For $\tilde{B}_{T,1}$, again, similarly as in (S.2), we have

$$\mathbb{P}\left(\sup_{b\in I_T}\left\|\tilde{B}_{T,1}\right\| > \varepsilon\right) \leqslant \sum_{b\in I_T}\mathbb{P}\left(\left\|\tilde{B}_{T,1}\right\| > \varepsilon\right) \leqslant |I_T| \times \sup_{b\in I_T}\mathbb{P}\left(\left\|\tilde{B}_{T,1}\right\| > \varepsilon\right) = o(1).$$

Moving to $\tilde{B}_{T,2}$, since for some $0 < \delta < 1/2$, we have

$$\mathbb{P}\left(\sup_{b\in I_T}\left\|b^{\delta}\tilde{B}_{T,2}\right\| > \varepsilon\right) \leqslant \sum_{b\in I_T}\mathbb{P}\left(\left\|\tilde{B}_{T,2}\right\| > b^{-\delta}\varepsilon\right) \leqslant |I_T| \times \sup_{b\in I_T}\mathbb{P}\left(\left\|\tilde{B}_{T,2}\right\| > b^{-\delta}\varepsilon\right) = O(b).$$

Thus, we have

$$\sup_{b\in I_T}\left\|\tilde{B}_T\right\| \leqslant \sup_{b\in I_T}\left\|\tilde{B}_{T,1}\right\| + \sup_{b\in I_T}\left\|\tilde{B}_{T,2}\right\| = O_p(b^{1-\delta}),$$

which implies (S.4).

S2 Additional simulation results

The parameters a_t and b_t for the DGPs (13) are summarized in Table 1. DGP C1 uses constant parameter values $a_t = 0.9$ and $b_t = 1$ for all t. DGPs C2–C4 allow for a one-time structural break in the parameters a_t and b_t at different time points: $\tau = T/4$, $T/2$, and $3T/4$. DGP C5–C7 have a one-time break at the same time points but with smaller-sized break in the parameters.

Forecast comparisons for DGPs C1-C7 are presented in Table S2. In general, the Gaussian kernel delivers the best forecasting results. DGP C1 does not have any parameter instability, and hence the benchmark full-sample estimator is expected to perform well. This is the case when $h = 1$, nonetheless when the forecast horizon is long with $h = 5$, all four local estimators lead to forecasts with smaller MSEs than the benchmark no matter the sample size. When there is a one-time structural break in the parameters, the local estimators tend to produce

Table S1: Specification of DGPs: C1–C7.

DGP	a_t	b_t
C1	0.9	1
C2	$0.9 - T^{-0.2}\,\mathbf{1}(t \geqslant T/4+1)$	$1 + T^{-0.2}\,\mathbf{1}(t \geqslant T/4+1)$
C3	$0.9 - T^{-0.2}\,\mathbf{1}(t \geqslant T/2+1)$	$1 + T^{-0.2}\,\mathbf{1}(t \geqslant T/2+1)$
C4	$0.9 - T^{-0.2}\,\mathbf{1}(t \geqslant 3T/4+1)$	$1 + T^{-0.2}\,\mathbf{1}(t \geqslant 3T/4+1)$
C5	$0.9 - T^{-0.5}\,\mathbf{1}(t \geqslant T/4+1)$	$1 + T^{-0.5}\,\mathbf{1}(t \geqslant T/4+1)$
C6	$0.9 - T^{-0.5}\,\mathbf{1}(t \geqslant T/2+1)$	$1 + T^{-0.5}\,\mathbf{1}(t \geqslant T/2+1)$
C7	$0.9 - T^{-0.5}\,\mathbf{1}(t \geqslant 3T/4+1)$	$1 + T^{-0.5}\,\mathbf{1}(t \geqslant 3T/4+1)$

better forecasts when the size of the break is larger (DGPs C2–C4), and when the break point is closer to the end of the sample. Similar to the results shown in Table 2, the improvement of using local estimators compared to the benchmark forecasts is larger as the forecasting horizon h increases.

S3 Forecast combination methods

Let $\omega_{i,t}$ be the combination weight for model i at time t. For equal-weighted (EW) combinations, we set $\omega_{i,t} = 1/N$, where N is the number of candidate models.

For the discounted MSE (DMSE) combining method (Stock and Watson, 2004; Rapach et al., 2010), the weight $\omega_{i,t}$ is computed according to

$$\omega_{i,t} = \frac{\phi_{i,t}^{-1}}{\sum_{j=1}^{N} \phi_{j,t}^{-1}}, \qquad \text{with} \qquad \phi_{i,t} = \sum_{s=T_0}^{t-1} \rho^{t-1-s} (y_{s+h} - \hat{y}_{i,s+h|s})^2,$$

where ρ is a discounting factor, h is the forecast horizon, y_{s+h} is the true value, and $\hat{y}_{i,s+h|s}$ is the forecast from model i. This method assigns higher weight to an individual model whose forecasts have lower MSEs over the holdout out-of-sample period. When $\rho = 1$, there is no discounting and these weights are exactly the same as Bates and Granger (1969) for the case where the forecasts from one given model are uncorrelated. When $\rho < 1$, higher weights are attached to the more recent forecast accuracy measures for each model. In both applications, we set $\rho = 0.9$.

Table S2: Forecasting performance of the local estimators for DGPs C1–C7.

	$h=1$				$h=5$			
DGP	Opt_R	Opt_G	Opt_E	Opt_T	Opt_R	Opt_G	Opt_E	Opt_T
$T=200$								
C1	1.092	1.040	1.105	1.112	0.891	0.878	0.904	0.908
C2	0.883	0.856	0.889	0.893	0.719	0.708	0.721	0.726
C3	0.727	0.707	0.732	0.735	0.533	0.528	0.534	0.535
C4	0.653	0.701	0.656	0.659	0.478	0.501	0.480	0.482
C5	1.063	1.024	1.077	1.087	0.837	0.826	0.848	0.851
C6	1.039	1.001	1.051	1.061	0.798	0.790	0.807	0.810
C7	1.053	1.009	1.063	1.069	0.768	0.771	0.773	0.775
$T=400$								
C1	1.070	1.040	1.075	1.080	0.880	0.874	0.885	0.887
C2	0.827	0.815	0.828	0.831	0.670	0.665	0.667	0.671
C3	0.730	0.718	0.728	0.730	0.498	0.499	0.500	0.500
C4	0.705	0.697	0.708	0.709	0.491	0.497	0.493	0.494
C5	1.050	1.024	1.056	1.059	0.830	0.827	0.830	0.832
C6	1.036	1.014	1.043	1.048	0.799	0.794	0.802	0.804
C7	1.024	1.001	1.028	1.031	0.783	0.782	0.782	0.782
$T=800$								
C1	1.042	1.025	1.047	1.050	0.876	0.874	0.875	0.878
C2	0.834	0.832	0.837	0.839	0.637	0.643	0.633	0.637
C3	0.760	0.753	0.761	0.761	0.508	0.509	0.507	0.508
C4	0.732	0.726	0.731	0.732	0.481	0.483	0.480	0.480
C5	1.035	1.017	1.039	1.041	0.828	0.830	0.826	0.829
C6	1.007	0.997	1.011	1.012	0.805	0.807	0.801	0.805
C7	1.022	1.007	1.027	1.029	0.791	0.794	0.788	0.789

Note: Ratios of MSEs against the benchmark forecasts using full-sample least square estimators. Opt_R: rolling window selection method proposed by Inoue et al. (2017); Opt_G: optimal bandwidth selection with Gaussian kernel; Opt_E: optimal bandwidth selection with Epanechnikov kernel; Opt_T: optimal bandwidth selection with triangular kernel.

S4 Forecasting inflation

Real-time price index data are obtained from the Federal Reserve Bank of Philadelphia's Real-Time Dataset for Macroeconomists (RTDSM), described in more detail by Croushore and Stark (2001). We use quarterly data from 1985:Q1 to 2023:Q4. Inflation at time t is measured as $400 \times \ln(P_t/P_{t-1})$, where P_t is the GDP price index.[5] Following Romer and Romer (2000) among many others, we use the second available estimate in the RTDSM to compute the actual inflation and measure the forecast accuracy.[6]

The forecasts are computed using the auto-regressive distributed lag (ARDL) model with time-varying coefficients:

$$y_{t+h} = \theta_{0,t} + \theta_{1,t} y_{t-1} + \theta_{2,t} x_t + \varepsilon_{t+h}, \tag{S.8}$$

where x_t is a scalar predictor and h is the forecast horizon. The benchmark forecasts are obtained from a simple AR(1) model by setting $\theta_{2,t} = 0$ in (S.8), estimated using full-sample non-local least square. We also consider forecast combinations from models in which each scalar predictor x_t is used one at a time. In addition, we report forecasts computed with AR(1) model estimated using local estimators for comparison.

We consider a set of predictors inspired by Stock and Watson (2003), which includes interest rates, default spread, stock market variables, commodity prices, exchange rates and monetary variables. Unlike GDP price index, asset prices are not revised, hence we rely on the currently available time series. A detailed description of the list of predictors can be found in Table S3. The initial estimation sample is from 1959:Q3 to 1984:Q4, and the first available individual forecast is computed for 1985:Q1. We use 40 observations as the hold-out out-of-sample to obtain the weights for forecast combination based on the DMSE. Therefore, the forecast evaluation period runs from 1995:Q1 to 2023:Q4. We report results for forecasts at one quarter ($h = 1$) and one year ($h = 4$) ahead.

Table S4 reports the ratio of MSEs of each model to that of the benchmark forecasts.

[5] For simplicity, "GDP price index" refers to the price index series for GNP/GDP. For some of the sample the measure is based on GNP and a fixed weight deflator.

[6] For example, the first available estimate for 2019:Q4 price index is in the 2020:Q1 vintage, and the second available estimate for 2019:Q4 price index is in the 2020:Q2 vintage. This is what we use to calculate 2019:Q4 inflation.

Table S3: The list of predictors and variable transformation in forecasting the U.S. inflation.

Variable	Description	Source	Transform
FFR	Effective federal funds rate	FRED-QD	level
TmSpd	10-year minus 3-month Treasury bill rates	FRED-QD	level
DfSpd	BAA- minus AAA-rated corporate bond yields	FRED-QD	level
S&P500	S&P500 composite index	FRED-QD	$100\Delta \ln$
PE	Price-earnings ratio for S&P500 composite stocks	FRED-QD	$100\Delta \ln$
CAD	Canada/U.S. exchange rate	FRED-QD	$100\Delta \ln$
GBP	U.K./U.S. exchange rate	FRED-QD	$100\Delta \ln$
COM	Moody's commodity price index	GFD	$100\Delta \ln$
M1REAL	Real M1 money stock, deflated by CPI	FRED-QD	$100\Delta \ln$
M2REAL	Real M2 money stock, deflated by CPI	FRED-QD	$100\Delta \ln$

Note: The FRED-QD data set is developed by McCracken et al. (2021) and maintained by the Federal Reserve Bank of St. Louis. GFD refers to the Global Financial Database.

Apart from the full-sample least square estimator (OLS), we consider the fixed-rolling window estimator with window size 40 ($R = 40$), optimal rolling window selection method proposed by Inoue et al. (2017) (Opt_R), and the local estimator with optimally selected bandwidth using the the Gaussian kernel (Opt_G), the Epanechnikov kernel (Opt_E), and the Triangular kernel (Opt_T). The first row represents the AR(1) model, rows two through eleven correspond to the model in Equation (S.8) with different predictors, and the final two rows represent the forecast combinations of the forecasts from different predictors given above.

There are several issues worth mentioning. First, using local estimators improves forecast accuracy for the AR(1) model. Gains are always significant, and are larger for one year ahead forecast ($h = 4$). Second, adding additional predictor is not always useful. Choice of predictor really matters. The commodity price index is the most reliable predictor, which delivers the best forecasting performance. The gains also become more evident for $h = 4$. Using Gaussian kernel is the best for $h = 1$, while triangular kernel is preferred for $h = 4$. Finally, forecast combinations improve the forecast accuracy in nearly all cases, except Opt_T for $h = 1$.

Table S5 presents forecasting evaluation results for the period up to 2019:Q4, excluding COVID-19 observations to avoid pandemic-related distortions. The overall conclusions are similar, with a few noticeable differences. First, using local estimators improves forecast accuracy in all cases. Second, DMSE combining method delivers the best results. Inoue

Table S4: Forecasting performance for U.S. inflation: 1985:Q1–2023:Q4.

	$h=1$				$h=4$			
	Opt_R	Opt_G	Opt_E	Opt_T	Opt_R	Opt_G	Opt_E	Opt_T
AR	0.954	0.935	0.948	0.959	0.748*	0.738*	0.738*	
FFR	0.870*	0.876*	0.898	0.897	0.780*	0.811*	0.767*	0.751*
TmSpd	0.998	0.961	0.997	0.996	0.786*	0.808*	0.754*	0.746*
DfSpd	1.006	0.924	1.108	1.117	0.769*	0.784*	0.790*	0.791*
S&P500	1.002	0.924	1.040	1.061	0.735*	0.759*	0.726*	0.713*
PE	1.009	0.947	1.008	1.007	0.721*	0.774*	0.700*	0.691*
CAD	0.983	0.948	1.000	1.014	0.706*	0.745*	0.687*	0.681*
GBP	0.985	0.919	1.007	1.028	0.751*	0.766*	0.739*	0.729*
COM	0.822	0.800*	0.836	0.848	0.680*	0.710*	0.657*	0.651*
M1REAL	3.081	2.509	2.853	3.583	0.759*	0.795*	0.780*	0.756*
M2REAL	1.116	0.968	1.216	1.475	0.788*	0.806*	0.794*	0.797*
Comb-EW	0.968	0.936	0.981	1.017	0.719*	0.761*	0.704*	0.693*
Comb-DMSE	0.974	0.939	0.986	1.025	0.717*	0.761*	0.702*	0.691*

Note: Ratio of MSEs against the benchmark forecasts of AR(1) model estimated using non-local least square. Opt_R: rolling window selection method proposed by Inoue et al. (2017); Opt_G: optimal bandwidth selection with Gaussian kernel; Opt_E: optimal bandwidth selection with Epanechnikov kernel; Opt_T: optimal bandwidth selection with triangular kernel. Differences in forecasting accuracy that are significant at the 5% level using the DM test are marked by an asterisk. The grey-shaded cells denote the best forecasting performance for each group.

et al. (2017)'s method is overall the best for $h = 1$, while using triangular kernel is the best for $h = 4$. However, when we test the equal forecast accuracy between the best performing case and the second best case (AR Opt_R for $h = 1$ and AR Opt_T for $h = 4$), the results are only significant for $h = 1$. This implies that exogenous predictors are not so useful once we control for parameter instability, especially for longer-horizon forecasts.

Table S5: Forecasting performance for U.S. inflation: 1985:Q1–2019:Q4.

	$h = 1$				$h = 4$			
	Opt_R	Opt_G	Opt_E	Opt_T	Opt_R	Opt_G	Opt_E	Opt_T
AR	0.768*	0.789*	0.777	0.784	0.578*	0.628*	0.548*	0.537*
FFR	0.775*	0.779*	0.811	0.826	0.624*	0.689*	0.610*	0.598*
TmSpd	0.808*	0.809*	0.808	0.811	0.645*	0.688*	0.594*	0.578*
DfSpd	0.876	0.799*	1.029	1.031	0.554*	0.594*	0.624*	0.637*
S&P500	0.840	0.805*	0.903	0.937	0.541*	0.598*	0.522*	0.516*
PE	0.744*	0.774*	0.766*	0.777	0.625*	0.668*	0.621*	0.623*
CAD	0.814*	0.809*	0.838	0.863	0.571*	0.608*	0.528*	0.518*
GBP	0.818	0.793*	0.857	0.886	0.606*	0.625*	0.588*	0.583*
COM	0.816	0.804*	0.841	0.868	0.535*	0.556*	0.520*	0.520*
M1REAL	0.759*	0.785*	0.792	0.822	0.559*	0.641*	0.536*	0.519*
M2REAL	0.776	0.779*	0.800	0.813	0.633*	0.655*	0.605*	0.598*
Comb-EW	0.739*	0.764*	0.763	0.780	0.533*	0.604*	0.505*	0.495*
Comb-DMSE	0.737*	0.764*	0.763*	0.781	.529*	0.601*	0.501*	0.492*

Note: Ratio of MSEs against the benchmark forecasts of AR(1) model estimated using non-local least square. Opt_R: rolling window selection method proposed by Inoue et al. (2017); Opt_G: optimal bandwidth selection with Gaussian kernel; Opt_E: optimal bandwidth selection with Epanechnikov kernel; Opt_T: optimal bandwidth selection with triangular kernel. Differences in forecasting accuracy that are significant at the 5% level using the DM test are marked by an asterisk. The grey-shaded cells denote the best forecasting performance for each group.