Nonparametric estimation and forecasting of time-varying parameter models

Yu Bai
Department of Econometrics and Business Statistics, Monash University
Bin Peng
Department of Econometrics and Business Statistics, Monash University
Shuping Shi
Department of Economics, Macquarie University
Wenying Yao
Melbourne Business School, University of Melbourne

April 27, 2024

Abstract

This paper considers practical issues when local estimator is used in an out-of-sample forecasting context. We first propose an approach to select the optimal bandwidth and prove the asymptotic optimality of the procedure. We then discuss the implications on the choice of kernel weighting functions. The theoretical results are examined through a Monte Carlo study. The methods are illustrated through two empirical applications on bond return predictability and real-time inflation forecasting. We find that our methods provide substantial gains for 12-month overlapping bond returns, and generally perform better than forecasts from constant coefficient models for inflation.

Keywords: Local Estimator; Bandwidth Parameter; Kernel Weighting Function; Bond Return Predictability; Inflation Forecasting

1 Introduction

Many important economic decisions are based on forecasting models that is known to be affected by parameter instability. It is now widely recognized that parameter instability is a crucial source of forecast failure. The empirical evidence of parameter instability has also been well documented, see, for instance, bond return predictability (Gargano et al., 2019; Borup et al., 2023), volatility forecasting (Oh and Patton, 2021) and macroeconomic forecasting (Stock and Watson, 1996; Pettenuzzo and Timmermann, 2017).

Motivated by concerns of parameter instability, forecasters often want to make predictions using the most recent data. They may do this by using a window of recent data, which is the so-called "rolling window" forecast scheme. As rolling window estimator is a special case of the local estimator when a flat kernel weighting function is used (Inoue et al., 2017), forecaster may have alternative choice of weighting functions and need to select the associated bandwidth parameters.

This paper considers practical issues when local estimator is used in an out-of-sample forecasting context. First, we propose an approach to select the bandwidth parameter by minimizing the conditional expected loss at the end of the sample. It is well known that bandwidth parameter plays a crucial role in determining the bias-variance tradeoff for the local estimator. Thus, it also affects forecasting performance. Our approach is similar to the one in Inoue et al. (2017) for rolling window selection, but we show that the asymptotic optimality holds when a generic weighting function is used for local estimation and a general loss function is used for forecast evaluation, which covers the asymmetric loss functions such as those considered in Laurent et al. (2012).

We then discuss the implications on the choice of kernel weighting functions, which has been less addressed in the literature. Our analysis is based on the limiting behavior of the regret risk (Hirano and Wright, 2017). We find that choice of kernel weighting functions is related to bandwidth selection and reflects the usual bias-variance trade-off. We show that, when either estimation variance or estimation bias dominates, there is a simple criterion to select the optimal kernel weighting function. However, when bandwidth is selected based on our approach, both estimation variance and bias are present in the limit, making the choice more involved.

The theoretical analyses are examined through an extensive Monte Carlo study. Using a linear predictive regression model, we find that, local estimator with optimal bandwidth has satisfactory out-of-sample (OOS) forecasting performance. In general, using all but downweighting the data is preferred, particularly for the multi-step ahead forecasts.

We present two empirical applications. As not only parameters but also best performing models may change over time, we also consider methods to deal with model uncertainty. In the first application, we look at the bond return predictability. Treasury bonds 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). However, the forecasting performance from local estimator has not been examined. We find that using local estimator with optimal bandwidth is particularly useful for 12-month overlapping returns. It performs better than rolling window forecasts with fixed window size in most cases. In terms of choice of kernel weighting functions, not using all data is generally preferred. Model uncertainty also plays a role. Using a simple forecast selection rule from individual model forecasts with optimal bandwidth generally has the best forecasting performance.

We look at the real-time inflation forecasting in our second application. Because asset prices are forward-looking, they contain information about future economic development, and are potentially useful predictors of inflation. Parameter instability in individual inflation forecasting model with asset prices has also been well documented in the literature (Stock and Watson, 2003). We find that using local estimator with optimal bandwidth delivers gains compared to forecasts obtained from non-local estimator (constant coefficient models) in most of the cases. Commodity price emerges as the best predictor for short horizon forecasts (1 quarter ahead), and gains are mostly from the post-pandemic period.

The rest of the paper is organized as follows. In Section 2, we describe local estimation methods and briefly discuss assumptions imposed on time-varying parameters. In Section 3, we introduce our bandwidth selection procedure and discuss implications on the choice of kernel weighting functions. Section 4 provides a Monte Carlo study. Section 5 presents two empirical applications on bond return predictability and real-time inflation forecasting, and Section 6 concludes. Technical details on models and assumptions, together with data description and additional empirical results, are provided in the Appendix. Proofs of the auxiliary results and main theorems, as well as additional simulation results are provided in the supplementary material.

2 Estimation under parameter instability

Let (y_t) be the scalar variable of interest and (X_t) be a vector of predictors. We wish to forecast $y_{T+h}(1 \leqslant h < \infty)$, given the knowledge of X_T. The forecast $\hat{y}_{T+h|T}$ is created using a rule: $\hat{y}_{T+h|T}(\theta)$, where θ is a $d \times 1$-dimensional model parameters. The model parameters are estimated via M-estimation minimizing

$$\hat{\theta}_T = \arg\min_{\theta \in \Theta} \frac{1}{T} \sum_{t=1}^{T} \ell_t(\theta), \tag{1}$$

where $\ell_t(\theta) = L(y_{t+h}, \hat{y}_{t+h|t}(\theta))$ is the loss function.

It is well known that parameter instability plagues commonly used forecasting models and predictive content is unstable over time (Rossi, 2013). To handle the instability issues and remain

agnostic on the types of parameter time variation, we take a nonparametric approach[1] and assume that time-varying parameters θ_t are modeled as the function of scaled time point:

$$\theta_t = \theta(t/T), \quad t = 1, 2, \cdots, T. \tag{2}$$

As explained in Robinson (1989), the requirement that time-varying parameter is a function of scaled time point is essential to derive 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.

The population loss at the end of the sample is defined by

$$E_T\left(\ell_{T+h}(\theta_T)\right), \tag{3}$$

where $\theta_T = \theta(T/T)$ is the parameter value at time T.[2] Since θ_T is unknown, we consider a local estimator for θ_T defined by

$$\hat{\theta}_{K,b,T} = \arg\min_{\theta\in\Theta} \frac{1}{Tb} \sum_{t=1}^{T} k_{tT}\ell_t(\theta), \tag{4}$$

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 $T \to \infty$. Different specifications of $K(\cdot)$ lead to different types of forecasting schemes. If $k_{tT} = 1$ for all t, we are back to the non-local estimation as in (1). If $K(u) = \mathbb{1}_{\{-1<u<0\}}$, we are in the rolling forecast scheme with window size $\lfloor Tb \rfloor$ (Giacomini and Rossi, 2009).

Example 1. *Consider the linear predictive regression model:*

$$y_{t+h} = X_t'\theta + \varepsilon_{t+h}, \quad t = 1, 2, \cdots, T-h,$$

where ε_{t+h} is a disturbance term. Then, under mean squared error (MSE) loss: $\ell_t(\theta) = (y_{t+h} - X_t'\theta)^2$, the local estimator for θ_T is given by

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

[1] Alternatively, we can use Bayesian approach, but a specification for the evolution of time-varying parameters is needed. For a comparison of different specifications to accounting for parameter instability in Bayesian context, see, for instance, Pettenuzzo and Timmermann (2017).

[2] As in Example 1, when the parameters are time-varying, the target y_{T+1} depends on θ_{T+1}, which is different from θ_T. However, under Assumption A1, the local time variation is asymptotically negligible and we can treat θ_{T+1} to be approximately equal to θ_T.

Example 2. *Consider the GARCH(1,1) model:*

$$y_t = \sigma_t \varepsilon_t$$
$$\sigma_t^2 = \omega + \alpha y_{t-1}^2 + \beta \sigma_{t-1}^2,$$

where $\varepsilon_t \sim (0, 1)$*. 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, 2021):*

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

The theoretical analysis in the next section requires the asymptotic properties of the local estimator. While a formal technical discussion is postponed to the Appendix A, we highlight the main assumption imposed on the time-varying parameters θ_t. We assume that, $\theta(\cdot) : (0, 1] \longrightarrow \Theta$ is twice continuously differentiable on $(0, 1]$. This condition implies θ_t changes slowly over time.

Remark 1. *Recently, Giraitis et al. (2014) introduce a new class of stochastic time-varying coefficient model in which* θ_t *evolves as bounded random walk process. They show that the local estimator can consistently estimate the paths of the stochastic coefficients. However, it can be easily seen that the paths of some of the stochastic processes they consider are indeed twice continuously differentiable. To see this, suppose that* θ_t *is a realization of a bounded random walk process:* $\frac{1}{T^{H-1/2}}\xi_t$*, where* $\Delta\xi_t = (1 - L)^{1-H} v_t \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$ *and* H *is the memory parameter. Simple algebra gives* $\theta_t = \left(\frac{t}{T}\right)^{H-1/2} \frac{1}{t^{H-1/2}}\xi_t = \left(\frac{t}{T}\right)^{H-1/2} C_t$*, where* $C_t = \frac{1}{t^{H-1/2}}\xi_t = O_p(1)$ *by Theorem 2 in Davydov (1970). This implies that* $\theta_t = \theta(t/T) \propto \left(\frac{t}{T}\right)^{H-1/2}$*, which is twice continuously differentiable.*

The bandwidth selection criteria considered in Section 3.1 depends on the unknown parameter value θ_T. Following Inoue et al. (2017), we consider to replace the unknown θ_T with the local linear estimator. The local linear estimator of θ_T and its first order derivative $\theta_T^{(1)}$ is given by

$$\left(\tilde{\theta}_T, \tilde{\theta}_T^{(1)}\right) = \underset{(\theta, \theta^{(1)}) \in \Theta \times \tilde{\mathbb{R}}^d}{\arg\min} L_T(\theta, \theta^{(1)}), \tag{5}$$

where $\tilde{\mathbb{R}} = [-M, M]$ for some $M > 0$. The objective function $L_T(\theta, \theta^{(1)})$ is defined as

$$L_T(\theta, \theta^{(1)}) = \frac{1}{T\tilde{b}} \sum_{t=1}^{T} \tilde{k}_{tT} \ell_t\left(\theta + \theta^{(1)}(t/T - 1)\right), \tag{6}$$

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 $T \to \infty$.

Remark 2. *We focus on the use of local estimator in* (4) *to create the forecast. First, as the rolling window estimator is a special case when the flat kernel weighting function* $K(u) = \mathbb{1}_{\{-1<u<0\}}$ *is used. Moreover,* (4) *would allow us to discuss whether alternative kernel weighting functions could be better than the flat weighting function. Second, we may use the local linear estimator to create the forecast. However, it requires the support of* $\tilde{K}$ *to be compact (Assumption A4). This again rules out interesting cases such as the Gaussian kernel:* $K(u) \propto e^{-\frac{u^2}{2}}$*, which is found to deliver the best forecasting performance in Kapetanios et al. (2019).*

3 Out-of-sample forecasting

To use the local estimator (4), a forecaster has to choose kernel weighting function K and bandwidth parameter b. In order to understand the implications of selecting K and b, we will analyze the population loss $E_T\left(\ell_{T+h}(\hat{\theta}_{K,b,T})\right)$.

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 (ignoring the smaller order terms):

$$E_T\left(\ell_{T+h}(\hat{\theta}_{K,b,T})\right) \approx E_T\left(\ell_{T+h}\left(\theta_T\right)\right) + E_T\left(\frac{\partial \ell_{T+h}\left(\theta_T\right)}{\partial \theta'}\right)\left(\hat{\theta}_{K,b,T} - \theta_T\right) + \frac{1}{2}\left(\hat{\theta}_{K,b,T} - \theta_T\right)' E_T\left(\frac{\partial^2 \ell_{T+h}\left(\theta_T\right)}{\partial \theta \partial \theta'}\right)\left(\hat{\theta}_{K,b,T} - \theta_T\right). \quad (7)$$

We see that the population loss can be decomposed into three components. The component $E_T\left(\ell_{T+h}\left(\theta_T\right)\right)$ is related to the future risk, which has nothing to do with parameter estimation. As in Hirano and Wright (2017), we define the *regret* as

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

If we further assume that:[3]

$$E_T\left(\frac{\partial \ell_{T+h}\left(\theta_T\right)}{\partial \theta'}\right) = 0, \quad (8)$$

the regret $R_T(K,b)$ simplifies to (ignoring the constant $1/2$)

$$R_T(K,b) = \left(\hat{\theta}_{K,b,T} - \theta_T\right)' E_T\left(\frac{\partial^2 \ell_{T+h}\left(\theta_T\right)}{\partial \theta \partial \theta'}\right)\left(\hat{\theta}_{K,b,T} - \theta_T\right). \quad (9)$$

[3]For the model considered in Example 1, this implies that $E[\varepsilon_{T+h}|\mathcal{F}_T] = 0$. so the forecast error is assumed to be serially uncorrelated.

Thus, minimizing the population loss at the end of the sample is equivalent to minimize $R_T(K, b)$.

3.1 Selection of the bandwidth parameter b

Let us first consider the choice of the bandwidth parameter b. Suppose that the kernel weighting function K is chosen such that $K = \overline{K}$. Write $\hat{\theta}_{\overline{K},b,T} = \hat{\theta}_{b,T}$ and $\omega_T(\theta_T) = E_T\left(\frac{\partial^2 \ell_{T+h}(\theta_T)}{\partial\theta\partial\theta'}\right)$. We consider to choose b by simply minimizing (9) over the choice set I_T:

$$\hat{b} := \underset{b\in I_T}{\arg\min} \left(\hat{\theta}_{b,T} - \theta_T\right)' \omega_T(\theta_T) \left(\hat{\theta}_{b,T} - \theta_T\right). \tag{10}$$

We first derive the rate of the optimal bandwidth parameter implied by (10), which is characterised in the following theorem:

Theorem 1. *Under Assumptions A1, A2, A3 and A4(i) in the Appendix, the optimal bandwidth parameter $\hat{b}$ obtained by minimizing* (10) *is of order $T^{-\frac{1}{3}}$ in probability.*

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

To make (10) feasible, we replace the unknown θ_T with the local linear estimator $\tilde{\theta}_T$ given in (5). This leads to a feasible selection criteria:

$$\hat{b} := \underset{b\in I_T}{\arg\min} \left(\hat{\theta}_{b,T} - \tilde{\theta}_T\right)' \omega_T\left(\tilde{\theta}_T\right) \left(\hat{\theta}_{b,T} - \tilde{\theta}_T\right). \tag{11}$$

The asymptotic optimality of the feasible selection procedure (11) is formally stated in the next theorem.

Theorem 2. *Under Assumptions A1-A6 in the Appendix, choosing $\hat{b}$ by* (11) *is asymptotically optimal in the sense that*

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

where $\tilde{\theta}_T$ is the local linear estimator from (5) *with bandwidth parameter $\tilde{b}$.*

Theorem 2 also provides an extension to the ones in Inoue et al. (2017) by showing that the asymptotic optimality holds for a generic weighting function when using (4) and a general loss function for forecast evaluation. The asymptotic optimality implies that $\hat{b}$ chosen from (11) yields the same forecasts obtained from the true optimal bandwidth parameter by minimizing the infeasible objective function in (10). The key to establish this result is to use the fact that the

asymptotic bias from local linear estimator vanishes at a faster rate than local estimator in (4), which neccitates Assumption A5.[4]

3.2 Implications on the choice of K

Suppose now forecaster has selected b based on (11), does the choice of K matter in terms of forecast accuracy (lower population loss)? Since $R_T(K,b)$ is related to $\hat{\theta}_{K,b,T} - \theta_T$, we first need an expansion for $\hat{\theta}_{K,b,T} - \theta_T$.

Let $L_T(\theta) = \frac{1}{Tb}\sum_{t=1}^{T} k_{tT}\ell_t(\theta)$. As detailed in the proof of Lemma A1((A9) in the Supplementary Material), we have

$$\hat{\theta}_{K,b,T} - \theta_T = -H_T^{-1}(\theta_T)\,(S_T(\theta_t) + B_T)\,, \tag{12}$$

where

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

and $\overline{\theta}_T$ lies between θ_T and θ_t. Then, as $T\to\infty$, suppose that $T^{1/2}b^{1/2} \asymp b^{-1}$, we can obtain the limiting regret risk $E[R_T(K,b)]$ from Lemma A1:

$$E[R_T(K,b)] = \mathrm{Tr}\left(\frac{\partial^2\ell_{T+h}(\theta_T)}{\partial\theta\partial\theta'}\left(b^2\mu_{1,K}^2\theta_T^{(1)}\theta_T^{(1)'} + \frac{\phi_{0,K}\Sigma_T}{Tb}\right)\right), \tag{13}$$

where $\phi_{0,K} = \int_{\mathbb{B}} K^2(u)du$, $\mu_{1,K} = \int_{\mathbb{B}} uK(u)du$ and Σ_T is the asymptotic variance of the estimator with detailed expression given in Lemma A1.

Several issues are worth mentioning. First, the limiting regret risk consists of a component related to estimation bias $b^2\mu_{1,K}^2\theta_T^{(1)}\theta_T^{(1)'}$ (limit of B_T) and the other component related to estimation variance $\frac{\phi_{0,K}\Sigma_T}{Tb}$ (limit of S_T). Both components involve an integral term related to the kernel weighting function K. $\mu_{1,K}$ is associated with the estimation bias, while $\phi_{0,K}$ is related to the estimation variance. From a risk reduction perspective, an optimal kernel should have smallest $\phi_{0,K}$ and $\mu_{1,K}$. Second, the requirement $T^{1/2}b^{1/2} \asymp b^{-1}$ implies that b is of order $T^{-\frac{1}{3}}$ in probability, which is the optimal one we obtain in Theorem 1.

Let us consider the following three candidate choices of $K(u)$:

$$K_1(u) = \mathbb{1}_{\{-1<u<0\}},\quad K_2(u) = \frac{2}{\sqrt{2\pi}}\exp\left(-\frac{u^2}{2}\right)\mathbb{1}_{\{u<0\}},\quad K_3(u) = \frac{3}{2}(1-u^2)\mathbb{1}_{\{-1<u<0\}}. \tag{14}$$

These three kernel weighting functions are commonly used in applied work. $K_1(u)$ leads to

[4]Assumption A5 imposes conditions on two bandwidth parameters involved (b and $\tilde{b}$), which requires b goes to zero at a faster rate than $\tilde{b}$, $T\tilde{b}^5 \to 0$ and $T^{1/2}\tilde{b}^{1/2}b \to \infty$. The condition that $T\tilde{b}^5 \to 0$ ensures that the bias of $\tilde{\theta}_T$ vanish asymptotically, while the condition $T^{1/2}\tilde{b}^{1/2}b \to \infty$ is required for obtaining results in Theorem 2.

a rolling window estimator with window size $\lfloor Tb \rfloor$. $K_2(u)$ imposes an exponential-type downweighting scheme, which has been used in macroeconomic forecasting context (Dendramis et al., 2020). $K_3(u)$ implies a hyperbolic type downweighting scheme, which is recommended in equity premium forecasts as in Farmer et al. (2023).

Table 1 provides numerical values for $\phi_{0,K}$ and $\mu_{1,K}^2$ associated with the kernel weighting functions considered in (14). As an optimal K should have the lowest $\phi_{0,K}$ and $\mu_{1,K}^2$ at the same time, there is no clear winner among $K_1(u)$, $K_2(u)$ and $K_3(u)$ asymptotically. In finite sample, we may expect that, if estimation variance dominates, $K_2(u)$ has the lowest $\phi_{0,K}$ so it may work better: all data should be used and downweighted. However, if the bias term dominates, $K_3(u)$ has the lowest $\mu_{1,K}^2$ so it would be preferred: only recent data should be used and downweighted.

Table 1: Numerical values of $\phi_{0,K}$ and $\mu_{1,K}^2$

	$\phi_{0,K}$	$\mu_{1,K}^2$
$K_1(u)$	1	0.25
$K_2(u)$	0.56	0.64
$K_3(u)$	1.20	0.14

4 Monte Carlo experiments

We now turn to the Monte Carlo experiments. The purpose of this section is to examine the finite sample performance of our bandwidth selection method, as well as the implications on the choice of kernel weighting functions.

4.1 DGPs

The DGPs are based on a bivariate VAR(1) as in Inoue et al. (2017):

$$\begin{bmatrix} y_{t+1} \\ x_{t+1} \end{bmatrix} = \begin{bmatrix} a_t & b_t \\ 0 & \rho_t \end{bmatrix} \begin{bmatrix} y_t \\ x_t \end{bmatrix} + \begin{bmatrix} \varepsilon_{t+1}^y \\ \varepsilon_{t+1}^x \end{bmatrix}, \tag{15}$$

where the error terms $(\varepsilon_{t+1}^y, \varepsilon_{t+1}^x)'$ are generated from $\mathcal{N}(0, I_2)$. 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).

We have 9 different specifications for $(a_t, b_t)'$. For DGPs 1-4, they are generated according to

(1) $a_t = 0.9 - 0.4(t/T)$, $b_t = 1 + (t/T)$;

(2) $a_t = 0.9 - 0.4(t/T)^2$, $b_t = 1 + (t/T)^2$;

(3) $a_t = 0.9 - 0.4 \exp(-3.5t/T)$, $b_t = 1 + \exp(-16(t/T - 0.5)^2)$;

(4) $a_t = 0.55 + 0.4\cos(4\pi(t/T))$, $b_t = 0.8 + \sin(4\pi(t/T))$.

For DGPs 5-9, we first set $a_t = 0.75 - 0.2\sin(3\pi(t/T))$. For b_t, we first generate $v_t = (1-L)^{1-d}\epsilon_t$, where $\epsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(0, 0.1^2)$. Then, we generate ξ_t from the random walk model: $\Delta\xi_t = v_t$. Finally, we set $b_t = \frac{\xi_t}{T^{H-0.5}}$. We consider $H = 0.51, 0.75, 1, 1.25, 1.49$ for DGPs 5-9, respectively. As explained in Remark 1, DGPs 5-9 also satisfy Assumption A1[5].

4.2 Implementations

We consider the following predictive regression model:

$$y_{t+h} = \theta_{0,t} + X_t'\theta_t + \varepsilon_{t+h}, \tag{16}$$

where $X_t = (y_t, x_t)'$. As we do not have the constant term in the DGPs (15), we set $\theta_{0,t} = 0$. The forecast $\hat{y}_{T+h|T} = X_T'\hat{\theta}_{K,b,T}$ is evaluated by the mean squared forecast error (MSFE) loss: $(y_{T+h} - \hat{y}_{T+h|T})^2$.

The model parameters are estimated by the (local) least square (LS):

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

where the weights $k_{tT} = K\left(\frac{t-T}{Tb}\right)$ are computed from a kernel weighting function $K(u)$ with bandwidth parameter b. We consider three different choices of kernel weighting functions as discussed in section 3.2:

$$K_1(u) = \mathbb{1}_{\{-1<u<0\}}, \quad K_2(u) = \frac{2}{\sqrt{2\pi}}\exp\left(-\frac{u^2}{2}\right)\mathbb{1}_{\{u<0\}}, \quad K_3(u) = \frac{3}{2}(1-u^2)\mathbb{1}_{\{-1<u<0\}}.$$

Of course, when $k_{tT} = 1$ for all t, we are back to the non-local LS estimates when parameters in (16) are assumed to be constant over time.

Under (16) and MSFE loss, $R_T(K, b)$ becomes

$$R_T(K, b) = \left(\hat{\theta}_{K,b,T} - \theta_T\right)'\left(X_T X_T'\right)\left(\hat{\theta}_{K,b,T} - \theta_T\right). \tag{17}$$

The true parameters θ_T in (17) are approximated by the local linear estimator, which are the first $d \times 1$ elements of the following:

$$(\tilde{\theta}_T', \tilde{\theta}_T^{'(1)})' = \left(\sum_{t=1}^{T-h} \tilde{k}_{tT} Z_t Z_t'\right)^{-1}\left(\sum_{t=1}^{T-h} \tilde{k}_{tT} Z_t y_{t+h}\right),$$

[5]In the supplementary material, we provide additional simulation results when either parameters are constant over time or they follow a one-time break process (when Assumption A1 is not satisfied).

where $Z_t = \left[X_t', X_t'(\frac{t-T}{T})\right]'$, $\tilde{k}_{tT} = \tilde{K}\left(\frac{t-T}{T\tilde{b}}\right)$ are computed from a kernel weighting function $\tilde{K}(u)$ with tuning parameter $\tilde{b}$. We use the Epanechnikov Kernel: $\tilde{K}(u) = \frac{3}{2}(1-u^2)\mathbb{1}_{\{-1<u<0\}}$ to compute $\tilde{\theta}_T$ and $\tilde{b}$ is set by the rule-of-thumb method: $\tilde{b} = 1.06T^{-1/5}$. For b, we set $b = cT^{-1/3}$ and select c by minimizing $R_T(K, b)$ using a course grid of width 0.1 from 1 to 7.

We consider four different sample size: $T = 150, 300, 450, 600$. We evaluate the out-of-sample (OOS) forecasting performance of y_{T+h} based on 5000 Monte Carlo simulations. Apart from the forecasts from local estimators with optimal bandwidth selection, we also consider rolling window forecasts with fixed window size R_0. We consider both $R_0 = 40$ and $R_0 = 60$. The former is used in applications with quarterly data (Stock and Watson, 2003), while the later is used in applications with monthly data (Farmer et al., 2023). The benchmark is set to be the forecasts obtained from non-local LS estimates. The forecast evaluations are based on the ratios of RMSFEs (square root of MSFEs): $\frac{\sqrt{\sum_{m=1}^{M}(y_{T+h}^{(m)}-\hat{y}_{T+h|T}^{(m)})^2}}{\sqrt{\sum_{m=1}^{M}(y_{T+h}^{(m)}-\tilde{y}_{T+h|T}^{(m)})^2}}$, where $M = 5000$, $\tilde{y}_{t+h|t}^{m}$ is the benchmark forecast and $\hat{y}_{t+h|t}^{m}$ is the forecast from using local estimators. If the ratio of RMSFEs is less than 1, the forecasts generated from local estimator are more accurate than the ones from non-local estimator.

4.3 Simulation results

Table 2 presents the out-of-sample forecasting performance from the simulated dataset. Let us first start by commenting the results for 1-step ahead forecasts ($h = 1$). First, using local estimator with optimal bandwidth improves forecast accuracy in all cases for DGPs 1-3. For DGPs 4,5,8, it also delivers gains as sample size increases. For DGPs 6-7, it is outperformed by the benchmark, even though the differences are generally very small. Second, fixed rolling window forecasts sometimes have the best results, but choices of window size do affect the forecasting performance. Finally, in terms of choices of kernel weighting functions, using K_2 with optimal bandwidth is better than K_1 and K_3 in most of the cases. As explained in Section 3.2, it is more likely that estimation variance dominates the bias in finite sample under our DGPs.

We now move on to the 12-step ahead forecasts. As DGPs are based on VAR(1), forecasts obtained from predictive regression (16) are the directly forecasts (Marcellino et al., 2006). These forecast errors are serially correlated, and thus, (8) does hot hold. It is of practical interests to examine if our bandwidth selection methods still work well and whether the implications on the choices of kernel weighting functions still hold. In all, using K_2 with optimal bandwidth is overall better than others, particularly for $T = 450$ and $T = 600$, since it delivers the best forecasting performance in all cases.

Table 2: Forecasting performance from simulated dataset

DGP	Fixed1	Fixed2	Opt-K_1	Opt-K_2	Opt-K_3	Fixed1	Fixed2	Opt-K_1	Opt-K_2	Opt-K_3
			$h = 1$					$h = 12$		
					T=150					
1	0.971	0.967	0.984	0.967	0.992	0.950	0.942	0.975	0.953	0.983
2	0.873	0.880	0.884	0.880	0.886	0.924	0.919	0.953	0.928	0.959
3	0.891	0.884	0.907	0.895	0.911	0.979	0.975	1.000	0.981	1.006
4	1.175	1.264	1.013	1.008	1.017	0.959	0.957	0.966	0.958	0.969
5	1.016	1.014	1.026	1.006	1.033	0.999	0.984	1.021	0.996	1.032
6	1.021	1.018	1.023	1.006	1.028	0.998	0.988	1.021	0.996	1.029
7	1.029	1.026	1.036	1.016	1.045	1.005	0.994	1.029	1.003	1.041
8	1.037	1.032	1.042	1.020	1.049	0.988	0.978	1.014	0.990	1.023
9	1.026	1.024	1.031	1.013	1.038	0.995	0.986	1.008	0.989	1.019
					T=300					
1	0.956	0.949	0.956	0.948	0.959	0.974	0.961	0.980	0.966	0.984
2	0.861	0.854	0.861	0.859	0.863	0.940	0.927	0.947	0.931	0.951
3	0.879	0.869	0.882	0.874	0.884	0.981	0.975	0.987	0.975	0.988
4	1.005	1.071	0.953	0.963	0.960	0.981	0.977	0.979	0.976	0.983
5	1.017	1.015	1.012	1.003	1.015	0.996	0.982	0.998	0.982	1.006
6	1.025	1.019	1.021	1.009	1.024	1.004	0.990	1.007	0.991	1.016
7	1.018	1.013	1.016	1.006	1.018	1.003	0.991	1.003	0.988	1.011
8	1.023	1.016	1.018	1.007	1.021	1.014	1.001	1.014	0.999	1.023
9	1.025	1.020	1.019	1.009	1.023	0.996	0.981	0.999	0.983	1.005
					T=450					
1	0.961	0.952	0.954	0.949	0.956	0.966	0.952	0.960	0.951	0.965
2	0.847	0.842	0.842	0.845	0.843	0.935	0.922	0.931	0.922	0.934
3	0.871	0.863	0.868	0.864	0.868	0.986	0.977	0.983	0.973	0.984
4	0.948	0.986	0.938	0.951	0.942	1.002	0.994	0.993	0.989	1.000
5	1.013	1.009	1.005	0.998	1.007	1.003	0.982	0.989	0.978	0.996
6	1.014	1.008	1.008	1.001	1.008	1.013	0.998	1.002	0.988	1.007
7	1.022	1.015	1.014	1.006	1.017	1.009	0.996	0.999	0.988	1.005
8	1.021	1.014	1.012	1.005	1.015	1.011	0.994	0.997	0.988	1.004
9	1.023	1.014	1.011	1.004	1.015	1.007	0.996	0.998	0.988	1.003
					T=600					
1	0.962	0.948	0.948	0.944	0.950	0.973	0.958	0.960	0.952	0.963
2	0.826	0.820	0.819	0.823	0.819	0.930	0.920	0.924	0.916	0.925
3	0.884	0.875	0.876	0.872	0.878	0.982	0.972	0.974	0.968	0.975
4	0.923	0.945	0.928	0.937	0.930	1.019	1.009	1.003	0.998	1.011
5	1.016	1.008	1.003	0.997	1.006	0.995	0.978	0.978	0.970	0.982
6	1.019	1.012	1.009	1.002	1.011	1.018	1.001	1.000	0.988	1.005
7	1.020	1.013	1.008	1.003	1.011	1.011	0.995	0.994	0.986	0.999
8	1.015	1.023	1.014	1.005	1.014	1.004	0.990	0.989	0.979	0.993
9	1.016	1.006	1.004	0.999	1.006	1.018	1.002	0.999	0.989	1.005

Note: Fixed1: rolling window estimator with window size equal to 40; Fixed2: rolling window estimator with window size equal to 60; Opt-K_i: local estimator with optimal bandwidth selection, where $K_1(u) = \mathbb{1}_{\{-1<u<0\}}$, $K_2(u) = \frac{2}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) \mathbb{1}_{\{u<0\}}$ and $K_3(u) = \frac{3}{2}(1-u^2)\mathbb{1}_{\{-1<u<0\}}$.

5 Empirical applications

We present two empirical applications on bond return predictability and inflation forecasting. As in the previous section, forecasts are constructed based on the predictive regression model as in (16). The forecasts are evaluated by MSFE loss, and the model parameters are estimated by (local) least square, using three different choices of kernel weighting functions as in (14). Under MSFE loss, the regret $R_T(K, b)$ is given in (17) and local linear estimator is used to approximate the unknown true parameter θ_T. The one-sided Epanchnikov kernel is used to compute the local linear estimator with bandwidth parameter selected by the rule-of-thumb method: $\tilde{b} = 1.06T^{-1/5}$. Once the kernel weighting function is chosen, we set $b = cT^{-1/3}$ and use $\hat{c}$ to obtain our final forecast by minimizing $R(K, b)$ over a coarse grid of c. For the application with monthly observations (bond return predictability), we use a course grid of width 0.1 from 1 to 7 for c. For the application with quarterly observations (forecasting inflation), c ranges from 1 to 5 with step size equal to 0.1.

In addition to parameter instability, forecasters face model uncertainty in real-world applications as the best performing model may change over time. To address this issue, we consider both forecast combination (FC) methods and forecast selection (FS) methods. For the forecast combination methods, we consider the following three different combination schemes to obtain the forecast combination weights $\omega_{i,t}$. We begin by considering the equal-weighted (EW) combinations:

$$\omega_{i,t} = 1/N,$$

where N is the number of candidate models.

We then consider the discounted MSFE (DMSFE) combining method (Stock and Watson, 2004; Rapach et al., 2010), where the weights $\omega_{i,t}$ are the functions of historical forecasting performance of the individual models over the holdout OOS period. The weights are computed according to

$$\omega_{i,t} = \frac{\phi_{i,t}^{-1}}{\sum_{j=1}^{N} \phi_{j,t}^{-1}},$$

where

$$\phi_{i,t} = \sum_{s=T_0}^{t-1} \rho^{t-1-s} (y_{s+h} - \hat{y}_{i,s+h|s})^2.$$

ρ is a discounting factor and $\hat{y}_{i,s+h|s}$ is the forecast from model i. Thus, this method assigns higher weights to individual model forecasts which have lower MSFEs over the holdout OOS period. When $\rho = 1$, there is no discounting and these weights are exactly the ones derived by Bates and Granger (1969) for the case where the individual forecasts are uncorrelated. When $\rho < 1$, higher weights are attached to the recent forecast accuracy for individual models. In both applications, we set $\rho = 0.9$.

The third forecast combination methods we consider is the least square (LS) model averaging. The weights $\omega_t = (\omega_{1,t}, \cdots, \omega_{N,t})$ are computed from the solution of the minimization problem:

$$\hat{\omega}_t = \arg\min_{\omega \in \mathcal{H}} C_t(\omega),$$

where $\mathcal{H}$ is the unit simplex in $\mathbb{R}^N$:

$$\mathcal{H} = \left\{ \omega \in [0,1]^N : \sum_i \omega_i = 1 \right\}.$$

Following Hansen (2008), we consider the Mallows Model Averaging (MMA) criterion, in which $C_t(\omega)$ takes the form

$$C_t(\omega) = \omega' \hat{e}' \hat{e} \omega + 2\omega' D \hat{\sigma}^2.$$

$\hat{e} = [\hat{e}(1), \cdots, \hat{e}(N)]$, where $\hat{e}_i$ is a vector collecting forecast errors from model i over the holdout OOS period. $D = (d(1), \cdots, d(N))'$, where $d(i)$ is the number of parameters have to be estimated from model i. $\hat{\sigma}^2$ is an estimator of innovation variance in (16) from the largest fitted model.

For the forecast selection methods, motivated by Granziera and Sekhposyan (2019), we proceed as follows. Let

$$\Delta L_{i,t-1} = \sum_{s=T_0}^{t-1} \left((y_{s+h} - \hat{y}_{i,s+h|s})^2 - (y_{s+h} - \hat{y}_{0,s+h|s})^2 \right)$$

be the loss differences from model i relative to the benchmark forecasts $\hat{y}_{0,s+h|s}$ over the holdout OOS period. Suppose that there is a vector of state variables S_t which are useful to explain the model's relative forecasting performance. In the first step, we run the regression

$$\Delta L_{i,t-1} = a_i' S_{t-2} + \epsilon_{t-1,i}$$

to obtain $\hat{a}_i$ using non-local estimator. In the second step, we predict y_{t+h} using the forecast from model i with the smallest $\Delta \hat{L}_{i,t} = \hat{a}_i' S_{t-1}$. We fit $\Delta L_{i,t-1}$ using both an AR(1) model ($S_t = [1 \ \Delta L_{i,t-2}]'$) and an AR(1)-X model ($S_t = [1 \ \Delta L_{i,t-2} \ X_{t-2}]'$). We use the 1-step ahead macro uncertainty index U_t developed in Jurado et al. (2015) as the addition exogenous variable, since it has been found to be useful in explaining bond return predictability (Borup et al., 2023).

Finally, to provide a rough gauge of whether differences in accuracy are significantly different, we apply the Diebold and Mariano (1995) (DM) test for equal forecast accuracy with fixed smoothing asymptotics as in Coroneo and Iacone (2020) , which is shown to deliver predictive accuracy tests that are correctly sized even when the number of out-of-sample observations are small.

5.1 Bond return predictability

5.1.1 Data and key variables

We use the following notation for the (log) yield of an n-year bond:

$$y_t^{(n)} = -\frac{1}{n} p_t^{(n)},$$

where $p_t^{(n)}$ is the log price of the n-year zero-coupon bond at time t. The holding-period return from buying an n-year bond at time t and selling it m-period later is

$$r_{t+12m}^{(n)} = p_{t+12m}^{(n-m)} - p_t^{(n)},$$

where m is in years and n can be 2,3,4, or 5 years in our analysis. Our target variable is the excess return

$$rx_{t+12m}^{(n)} = r_{t+12m}^{(n)} - my_t^{(m)},$$

where $y_t^{(m)}$ is the annualized T-bill rate. We consider both overlapping returns ($m = 1$) and one-month excess returns ($m = 1/12$)[6].

Empirical studies have found that forward rates or forward spreads contain information on future excess bond returns. Fama and Bliss (1987) find that forward spread has predictive power on excess bond returns and its forecasting power increases with the forecast horizon. Cochrane and Piazzesi (2005) find that a linear combination of forward rates predicts excess bond returns. Furthermore, Ludvigson and Ng (2009) extract factors from a large panel of macroeconomic variables and show that these factors are useful in predicting future bond excess returns.

Our predictor variables are computed as follows. The forward spreads (FB) is simply defined as

$$FB_t^{(n,m)} = f_t^{(n-m,n)} - my_t^{(m)},$$

where the forward rate $f_t^{(n-m,n)}$ is defined as

$$f_t^{(n-m,n)} = p_t^{(n-m)} - p_t^{(n)}.$$

The Cochrane-Piazzesi (CP) factor is formed as a linear combination of forward rates:

$$CP_t^m = (\hat{\delta}^m)' \boldsymbol{f}_t^m,$$

where $\boldsymbol{f}_t^m = (f_t^{(1-m,1)}, f_t^{(2-m,2)}, f_t^{(3-m,3)}, f_t^{(4-m,4)}, f_t^{(5-m,5)})'$. The coefficient vector $\hat{\delta}^m$ is estimated

[6]While early studies (Ludvigson and Ng, 2009) focus on overlapping returns, the use of one-month excess returns could offer several advantages. For more discussions, see Gargano et al. (2019) and Borup et al. (2023).

from

$$\frac{1}{4}\sum_{n=2}^{5} rx_{t+12m}^{(n)} = \delta_0^m + (\delta^m)' \boldsymbol{f}_t^m + \overline{\varepsilon}_{t+12m}.$$

Suppose we observe a $T \times M$ panel of macroeconomic variables z_{it} generated by a factor model

$$z_{it} = \kappa_i' g_t + \epsilon_{it},$$

where $g_t = (g_{1,t}, \cdots, g_{s,t})$ is a $s \times 1$ vector of common factors. Following Ludvigson and Ng (2009), we first obtain the estimates $\hat{g}_t$ by principle components analysis. Then, we build a single linear combination from a subset of the first eight estimated principle components, $\hat{G}_t = \left(\hat{g}_{1,t}, \hat{g}_{1,t}^3, \hat{g}_{3,t}, \hat{g}_{4,t}, \hat{g}_{8,t}\right)$ to obtain the Ludvigson-Ng (LN) factor[7]:

$$LN_t = (\hat{\lambda}^m)' \hat{G}_t,$$

where $\hat{\lambda}$ is obtained from the regression

$$\frac{1}{4}\sum_{n=2}^{5} rx_{t+12m}^{(n)} = \lambda_0^m + (\lambda^m)' \hat{G}_t + \overline{\varepsilon}_{t+12m}.$$

We study monthly excess bond return predictability for the United States over the period 1961:M6-2023:M12. The yield data are taken from Liu and Wu (2021). To compute the LN factor, we use the FRED-MD data set. The final vintage (2024:M4) data are used. Before extracting factors, each variable is transformed as described in the Appendix of McCracken and Ng (2016).

Panel A of Table 3 presents summary statistics for both 12-month overlapping returns and one-month excess returns when $n = 2, 3, 4, 5$. There are two key differences between 12-month overlapping returns and one-month excess returns. First, 12-month overlapping returns have much stronger persistence (higher serial correlation) than one-month excess returns because of the smoothing effect of using overlapping returns. Second, 12-month overlapping returns are far less leptokurtic (smaller kurtosis) than one-month excess returns.

[7]Ludvigson and Ng (2009) select this combination of factors using the Schwarz information criterion.

Table 3: Summary statistics: bond returns

	12-month overlapping excess returns					One-month excess returns			
	2 years	3 years	4 years	5 years		2 years	3 years	4 years	5 years
Mean	0.429	0.755	1.047	1.169	Mean	0.977	1.370	1.522	1.689
Std.dev.	1.663	3.049	4.264	5.319	Std.dev.	2.731	3.898	4.985	5.948
Skew	0.006	-0.080	-0.055	-0.072	Skew	0.563	0.172	-0.054	0.026
Kurt	4.066	3.886	3.747	3.687	Kurt	16.939	11.125	8.124	7.246
AC(1)	0.931	0.932	0.932	0.929	AC(1)	0.177	0.145	0.112	0.111

Note: This table reports summary statistics (mean, standard deviation, skewness, kurtosis, and first-order autocorrelation (AC(1))) for monthly bond excess returns (two-to five year bond maturities) in our study. The left block is based on 12-month overlapping returns, computed in excess of a 12-month T-bill rate. The right block is based on monthly returns computed in excess of a one-month T-bill rate.

5.1.2 Implementations

The forecasts are constructed based on the predictive regression model as in (16). Specifically, we consider three univariate models: FB ($X_t = FB_t^{(n,m)}$), CP ($X_t = CP_t^m$) and LN($X_t = LN_t$), along with a multivariate model that includes all three predictors FB+CP+LN ($X_t = \left[FB_t^{(n,m)}\ CP_t^m\ LN_t\right]'$) for a total of four models. The benchmark forecasts are obtained from the EH (efficient hypothesis) model which assumes no predictability by letting $\theta_t = 0$ in (16) for all t.

The initial estimation sample runs from 1961:M6 to 1984:M12 and the first available individual forecast is for 1985:M1. We use 5-year holdout OOS (60 observations) to obtain the initial weights for forecast combination and implement the forecast selection procedure. Thus, the forecast evaluation period runs from 1990:M1 to 2023:M12.

5.1.3 Empirical results

The forecasting results are summarized in Tables 4-5. For all entries, they are the ratios of MSFEs relative to the benchmark forecasts. Values below 1 indicate that the corresponding model (method) performs better than the benchmark. Entries shaded in gray indicate the best performing model. Differences in accuracy that are statistically different from zero (using either fixed b-smoothing or fixed m-smoothing asymptotics) are denoted by an asterisk, corresponding to the 5 percent significance level.

Table 4 provides results for 12-month overlapping returns. The results are notably promising, demonstrating sizable and (sometimes) significant gains compared to the benchmark forecasts. Among all the individual models, multivariate specification (FB+CP+LN) with local estimator using K_3 as the kernel weighting function and optimal bandwidth is the best for 2-year and 3-year bond returns, while K_1 with optimal bandwidth yields the best results for 4-year and 5-year

bond returns. Forecast combination further improves forecast accuracy, yet forecast selection consistently outperforms it. In terms of model specification for forecast selection, a simple AR(1) specification with K_3 as the kernel weighting function and optimal bandwidth is the best for 2-year and 3-year bond returns, while K_1 with optimal bandwidth is the best for 4-year bond returns. For 5-year bond returns, adding the macro uncertainty index to the AR(1) specification with K_1 and optimal bandwidth is the best. In general, fixed rolling window forecasts also deliver gains compared to the benchmark, but using either K_1 or K_3 with optimal bandwidth provides additional benefits in most of the cases.

The forecasting results for one-month excess returns are presented in Table 5. These results differ significantly from 12-months overlapping returns, as the EH model turns out to be a tough benchmark. Using local estimator is not useful in most of the cases. While the LS forecast combination with K_2 as the kernel weighting function improves forecast accuracy, the gains are typically very small and insignificant. Forecast selection using a simple AR(1) specification from the non-local estimator consistently outperforms other methods. Although local estimator with K_2 delivers gains for 2-year and 4-year bond returns from forecast selection, they are outperformed by the forecasts from non-local estimator.

To get a better understanding of the source of the gains, we plot in Figure 1 the cumulative sums of MSFEs differences (over the benchmark forecasts) from the best performing methods over the evaluation sample. When MSFEs differences are negative, alternative method is preferred. The left panel is for 12-month overlapping returns while the right panel presents one-month excess returns. Different markers indicate the different bond maturities. Overall, gains from using local estimator with forecast selection increase with the bond maturity for overlapping returns . For one-month excess returns, forecast selection from non-local estimator improves forecast accuracy in most of the evaluation period. While the forecasting performance for 2-year and 3-year bond returns are relatively stable over time, gains for 4-year and 5-year bond returns largely come from the post Great Recession-financial crisis period (2009-2023).

Based on Table 4, results indicate that using non-local estimator severely deteriorates forecast accuracy, especially when macro variables (LN factor) are involved. To examine whether these results are driven by the pandemic period observations (2020:M1-2023:M12), we also provide results over the forecast evaluation period 1990:M1-2019:M12 (shown in Table C1). The overall patterns remain highly similar. Forecast selection consistently produces the most accurate results. In terms of the best performing kernel weighting function, they are almost the same, except that K_3 is the best for 5-year bond returns. One notable difference in the pre-pandemic period is that, apart from the CP factor, using non-local estimator improves forecast accuracy compared to the benchmark forecasts, but it is outperformed by forecasts obtained from the local estimator. The inclusion of pandemic period observations does not appear to have a major impact on forecasting performance for one-month excess returns, as the results from the pre-pandemic period (shown

in Table C2) are quite similar.

Table 4: Out-of-sample forecasting performance: 12-month overlapping bond returns, 1990:M1-2023:M12

n		OLS	Fixed	Opt-K_1	Opt-K_2	Opt-K_3	n		OLS	Fixed	Opt-K_1	Opt-K_2	Opt-K_3
				Individual models							Individual models		
2	FB	1.028	1.075	0.911	0.885	0.906	3	FB	0.978	1.052	0.905	0.893	0.892
	CP	1.072	0.954	0.809	0.838	0.795		CP	1.071	0.924	0.789	0.823	0.773
	LN	5.136	0.738	0.705	0.794	0.695		LN	5.184	0.757	0.716	0.883	0.710
	FB+CP+LN	4.081	0.985	0.628*	0.876	0.588*		FB+CP+LN	4.764	0.894	0.644*	0.947	0.632*
				Forecast combination methods							Forecast combination methods		
	EW	1.766	0.726	0.644*	0.711	0.614*		EW	1.853	0.726*	0.651	0.741	0.627*
	DMSFE	1.050	0.632*	0.590*	0.643*	0.554*		DMSFE	1.081	0.661*	0.602*	0.680*	0.577*
	LS-MMA	2.167	0.935	0.664*	0.830	0.624*		LS-MMA	2.740	0.882	0.683*	0.851	0.646*
				Forecast selection methods							Forecast selection methods		
	AR(1)	0.689	0.513*	0.484*	0.559*	0.457*		AR(1)	0.649*	0.489*	0.495*	0.584*	0.430*
	AR(1)+U	0.695	0.519*	0.478*	0.563*	0.462*		AR(1)+U	0.669*	0.496*	0.492*	0.589*	0.436*
				Individual models							Individual models		
4	FB	0.952	1.011	0.899	0.879	0.856	5	FB	0.911	0.962	0.867	0.858	0.829
	CP	1.074	0.874	0.764	0.798	0.740		CP	1.076	0.829	0.746	0.783	0.718
	LN	5.285	0.786	0.734	0.987	0.730		LN	5.296	0.857	0.766	1.161	0.767
	FB+CP+LN	5.450	0.904	0.660*	1.123	0.711		FB+CP+LN	4.843	1.060	0.735	1.201	0.817
				Forecast combination methods							Forecast combination methods		
	EW	1.934	0.711	0.648*	0.761	0.625*		EW	1.845	0.724	0.658*	0.789	0.643*
	DMSFE	1.121	0.646*	0.600*	0.690*	0.584*		DMSFE	1.135	0.664*	0.619*	0.721	0.605*
	LS-MMA	2.943	0.835	0.679*	0.899	0.665*		LS-MMA	3.039	0.880	0.701*	0.898	0.692*
				Forecast selection methods							Forecast selection methods		
	AR(1)	0.621*	0.496*	0.475*	0.577*	0.549*		AR(1)	0.615*	0.500*	0.505*	0.568*	0.631
	AR(1)+U	0.635*	0.518*	0.483*	0.576*	0.556*		AR(1)+U	0.624*	0.512*	0.494*	0.567*	0.634

Note: This table reports ratios of out-of-sample MSFEs for individual prediction models, as well as combination forecast and forecast selection based on these models. For individual models, we consider FB, CP and LN predictors fitted to monthly bond excess returns, $rx_{t+12}^{(n)}$, measured relative to the (annualized) T-bill rate $y_t^{(1)}$. The benchmark forecasts are based on the EH (efficient hypothesis) model which assumes $\theta_t = 0$ in (16). OLS: non-local least square; Fixed: rolling window estimator with window size equal to 60; Opt-K_i: local estimator with optimal bandwidth selection, where $K_1(u) = \mathbb{1}_{\{-1<u<0\}}$, $K_2(u) = \frac{2}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) \mathbb{1}_{\{u<0\}}$ and $K_3(u) = \frac{3}{2}(1-u^2)\mathbb{1}_{\{-1<u<0\}}$; EW: equal weighted forecast combinations; DMSFE: discount mean square forecast error (DMSFE) combinations; LS-MMA: least square forecast combinations using Mallows Model Averaging (MMA) criterion; AR(1): relative forecasting performance over the holdout OOS $\Delta L_{i,t-1}$ is modeled as an AR(1) process for forecast selections. Differences in accuracy that are statistically different from zero (using either fixed b-smoothing or fixed m-smoothing asymptotics) are denoted by an asterisk, corresponding to the 5 percent significance level.

Table 5: Out-of-sample forecasting performance: one-month bond excess returns, 1990:M1-2023:M12

n		OLS	Fixed	Opt-K_1	Opt-K_2	Opt-K_3	n		OLS	Fixed	Opt-K_1	Opt-K_2	Opt-K_3
				Individual models							Individual models		
2	FB	1.005	1.053	1.020	1.001	1.033	3	FB	0.984	1.046	1.028	0.996	1.032
	CP	1.253*	1.076*	1.041	1.051	1.057		CP	1.187*	1.065*	1.039	1.045*	1.056*
	LN	1.007	1.039	1.011	1.003	1.020		LN	1.003	1.037*	1.016	1.008	1.027*
	FB+CP+LN	1.307*	1.151*	1.118*	1.083*	1.130*		FB+CP+LN	1.197*	1.135*	1.108*	1.066*	1.124*
				Forecast combination methods							Forecast combination methods		
	EW	1.078	1.055	1.026	1.020	1.037		EW	1.047	1.048	1.028	1.014	1.038
	DMSFE	1.052	1.056	1.025	1.018	1.037		DMSFE	1.034	1.048	1.028	1.014	1.039
	LS-MMA	0.994	1.037	1.016	0.997	1.022		LS-MMA	0.984	1.032	1.020	0.996	1.024
				Forecast selection methods							Forecast selection methods		
	AR(1)	0.978	1.077*	1.036	0.999	1.068*		AR(1)	0.979	1.073*	1.046	1.001	1.054
	AR(1)+U	0.985	1.075*	1.031	1.005	1.037		AR(1)+U	0.989	1.059*	1.050	1.007	1.043
				Individual models							Individual models		
4	FB	0.980	1.032	1.025	0.997	1.033	5	FB	0.985	1.036	1.038	1.000	1.040
	CP	1.155*	1.056*	1.038	1.041*	1.052*		CP	1.138*	1.051*	1.037*	1.040*	1.050*
	LN	0.999	1.033	1.020	1.009	1.027		LN	0.997	1.031	1.022	1.009	1.029*
	FB+CP+LN	1.147*	1.118*	1.102*	1.051	1.111*		FB+CP+LN	1.124*	1.112*	1.098*	1.048	1.110*
				Forecast combination methods							Forecast combination methods		
	EW	1.032	1.038	1.026	1.010	1.034		EW	1.026	1.037	1.030	1.012	1.037
	DMSFE	1.024	1.038	1.026	1.010	1.034		DMSFE	1.020	1.038	1.030	1.012	1.038
	LS-MMA	0.979	1.025	1.020	0.996	1.024		LS-MMA	0.985	1.030	1.028	1.000	1.032
				Forecast selection methods							Forecast selection methods		
	AR(1)	0.971	1.060*	1.046*	0.994	1.053*		AR(1)	0.979	1.045*	1.047*	1.007	1.052*
	AR(1)+U	0.981	1.055*	1.055*	0.999	1.044*		AR(1)+U	1.007	1.061*	1.054*	1.009	1.057*

Note: This table reports ratios of out-of-sample MSFEs for monthly bond excess returns, $rx_{t+1}^{(n)}$, measured relative to the (annualized) T-bill rate $(1/12)y_t^{(1/12)}$, over the forecasting evaluation period 1990:M1-2023:M12. See Notes to Table 4.

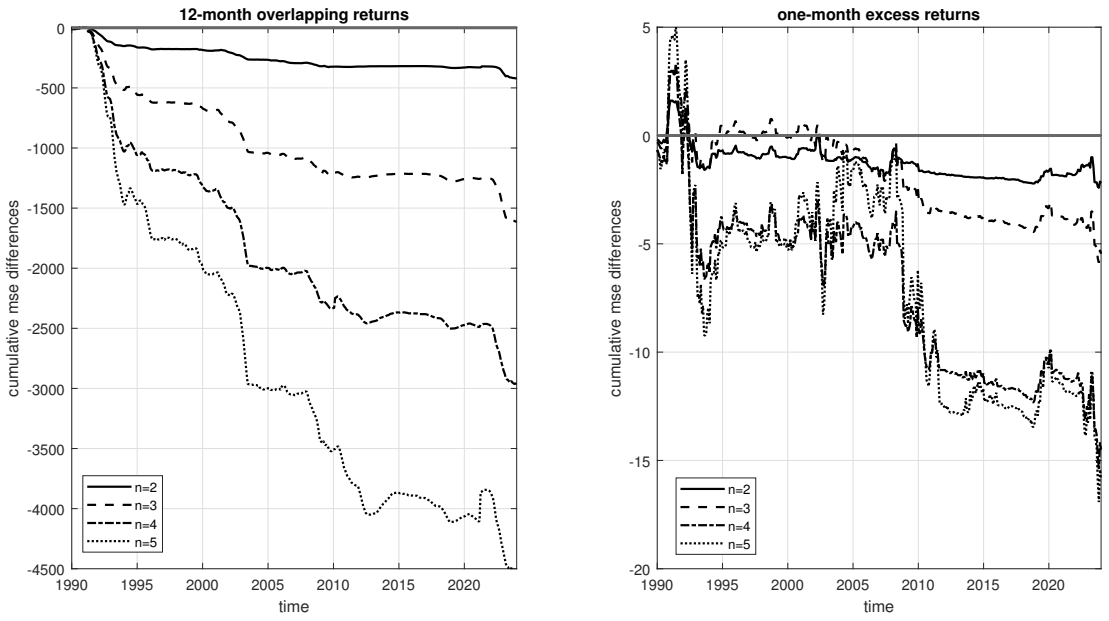

Figure 1: The figure presents cumulative sums of MSFEs differences (relative to the benchmark) for bond returns from the best performing methods.

5.2 Forecasting inflation

5.2.1 Data

We use real-time data for inflation, obtained from the Federal Reserve Bank of Philadelphia's Real-Time Dataset for Macroeconomists (RTDSM), described in Croushore and Stark (2001). Inflation is measured as $400 \ln(P_{t+1}/P_t)$, where P_t is the GDP price index[8].

Unlike GDP price index, asset prices are not revised, and we rely on just currently available time series. The predictors we consider include interest rates, default spread, stock market variables, commodity prices, exchange rates and monetary variables. A detailed data description (mnemonics, data sources and data transformations) is provided in Table B1.

5.2.2 Implementations

The forecasts are constructed based on the ARDL model:

$$y_{t+h} = \theta_{0,t} + \theta_{1,t} y_{t-1} + \theta'_{2,t} X_t + \varepsilon_{t+h},$$

where X_t is a $d \times 1$ vector of exogenous variables and h is the forecast horizon. To examine which predictor is useful in forecasting inflation, we include exogenous predictors one at a time. Thus,

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

we have 10 individual models.

Our analysis of real-time inflation forecasts uses real-time data vintages from 1985:Q1 through 2024:Q1. To measure the forecast accuracy of the different methods, we follow Romer and Romer (2000), among many others, and use the second available (in the RTDSM) estimate as actuals. The initial estimation sample runs from 1959:Q3 to 1984:Q4 and the first available individual forecast is for 1985:Q1. We report results for forecasts at horizons of 1,2, and 4 quarters ahead. We use 10-year holdout OOS (40 observations) to obtain the initial weights for forecast combination and implement the forecast selection procedure. As the 2024:Q1 vintage only contains the first available estimate of 2023:Q4, the forecast evaluation period runs from 1995:Q1 to 2023:Q3.

Finally, in terms of benchmark forecasts, we use the popular UC-SV model (Stock and Watson, 2007):

$$
\begin{aligned}
y_t &= \tau_t + \varepsilon_t^y, \quad \varepsilon_t^y \sim N(0, e^{h_t}), \\
\tau_t &= \tau_{t-1} + \varepsilon_t^\tau, \quad \varepsilon_t^\tau \sim N(0, e^{g_t}), \\
h_t &= h_{t-1} + \varepsilon_t^h, \quad \varepsilon_t^h \sim N(0, \omega_h^2), \\
g_t &= g_{t-1} + \varepsilon_t^g, \quad \varepsilon_t^g \sim N(0, \omega_g^2),
\end{aligned}
$$

where τ_t is the trend inflation and h_t, g_t are the time-varying volatilities associated with cycle (ε_t^y) and trend component (τ_t) of the inflation. The model parameters are ω_h^2, ω_g^2, and the initial conditions τ_0, h_0 and g_0[9] of the associated processes. The model is estimated using Bayesian methods in non-centered parameterization and then transform back to the centered parameterization to perform predictive simulation. The estimation details can be found in Appendix B in Chan (2018).

5.2.3 Empirical results

Table 6 reports the relative MSFEs of each model (method) relative to that of the UC-SV benchmark. As in Tables 4-5, values below 1 indicate that the corresponding model(method) performs better than the benchmark. In all, using local estimator with optimal bandwidth generally improves forecasting performance relative to that of the non-local estimator, as entries are more close to 1 in those cases. For 1 quarter ahead forecasts, using commodity price as the exogenous predictor is clearly preferred, as it is the only case in which using local estimator with optimal bandwidth provide gains relative to the benchmark, no matter which kernel weighting function is used. Yet, K_2 is the best. Forecast selection from AR(1) specification comes the second, and using K_2 again works better than K_1 and K_3. However, gains from using commodity price and forecast selection get lost once we move on to the 2 quarters and 4 quarters ahead forecasts.

[9]We assume Normal priors for all model parameters: $\omega_h \sim N(0, 0.2^2)$, $\omega_g \sim N(0, 0.2^2)$, $h_0 \sim N(0, 10)$, $g_0 \sim N(0, 10)$, and $\tau_0 \sim N(0, 10)$.

To better understand the models' forecasting performance over time, we plot in Figure 2 the cumulative sums of MSFEs differences (over the benchmark forecasts) from the individual model forecasts with commodity price. When MSFEs differences are negative, using commodity price with local estimator is preferred. Different markers indicate the different forecast horizons. For 1 quarter ahead forecasts, gains are from the post-pandemic period (2020:Q1-2023:Q3). For 2 quarters ahead forecasts, forecasting performance from commodity prices improves relative to the benchmark after the Great Recession-financial crisis period (2007:Q1–2009:Q4), but it is outperformed by the benchmark after 2022:Q4.

The results from both Table 6 and Figure 2 again raise the concern that forecasting performance may be driven by the pandemic period observations (2020:M1-2023:M12). In Table C3, we report the relative MSFEs of each model (method) relative to that of the UC-SV benchmark over the evaluation period 1990:Q1-2019:Q4. In all, using local estimator with optimal bandwidth still provide gains relative to that of the non-local estimator. UC-SV model is still a tough benchmark, as in most of the entries are above 1 for 1 quarter and 4 quarters ahead forecasts. For 1 quarter ahead forecasts, using discounted MSFE forecast combination method with K_1 as kernel weighting function and optimal bandwidth is the best. It is quite beneficial from using local estimator with optimal bandwidth for 2 quarters ahead forecasts, as entries are generally less than 1. Using default spread with K_1 and optimal bandwidth is the best. However, as implied from Figure 2, gains are likely to get lost in the recent period.

Table 6: Out-of-sample forecasting performance for inflation

	OLS	Fixed	Opt-K_1	Opt-K_2	Opt-K_3		OLS	Fixed	Opt-K_1	Opt-K_2	Opt-K_3
					$h = 1$						
	Individual models						Individual models				
Fedfunds	1.176*	1.133	1.038	1.045	1.071	EXCAUSx	1.187	1.180	1.173	1.130	1.193
Term spread	1.208*	1.222	1.191	1.147	1.189	EXUSUKx	1.162	1.208	1.175	1.097	1.202
Default spread	1.189*	1.216	1.200	1.102	1.322	Commodity	1.033	1.031	0.981	0.954	0.997
S&P 500	1.230*	1.213	1.196	1.102	1.240	M1real	1.337*	1.680	3.675	2.994	3.404
S&P 500 PE ratio	1.216*	1.221	1.204	1.130	1.203	M2real	1.166*	1.393	1.332	1.155	1.450
	Forecast combination methods						Forecast selection methods				
EW	1.170*	1.157	1.155	1.117	1.170	AR(1)	1.133	1.016	0.976	0.969	1.006
DMSFE	1.170	1.159	1.162	1.120	1.176	AR(1)+U	1.129	1.141	1.084	1.038	1.064
LS-MMA	1.087*	1.085	1.011	1.019	1.026						
					$h = 2$						
	Individual models						Individual models				
Fedfunds	1.699	1.428	1.309	1.350	1.322	EXCAUSx	1.679	1.479	1.430	1.438	1.449
Term spread	1.739	1.558	1.567	1.540	1.560	EXUSUKx	1.681	1.510	1.451	1.421	1.489
Default spread	1.730	1.486	1.393	1.452	1.476	Commodity	1.545	1.284	1.263	1.270	1.257
S&P 500	1.757	1.445	1.377	1.409	1.400	M1real	1.794	1.443	1.388	1.535	1.341
S&P 500 PE ratio	1.762	1.548	1.566	1.515	1.555	M2real	1.661	1.388	1.384	1.395	1.409
	Forecast combination methods						Forecast selection methods				
EW	1.681	1.389	1.337	1.388	1.341	AR(1)	1.655	1.240	1.252	1.391	1.279
DMSFE	1.684	1.382	1.327	1.389	1.330	AR(1)+U	1.698	1.255	1.264	1.390	1.243
LS-MMA	1.652	1.478	1.321	1.381	1.330						
					$h = 4$						
	Individual models						Individual models				
Fedfunds	1.659	1.282	1.289	1.341	1.268	EXCAUSx	1.622*	1.193	1.167	1.232	1.136
Term spread	1.654*	1.278	1.300	1.336	1.248	EXUSUKx	1.631*	1.253	1.241	1.267	1.222
Default spread	1.721*	1.268	1.272	1.296	1.306	Commodity	1.539*	1.151	1.124	1.175	1.086
S&P 500	1.642*	1.265	1.216	1.255	1.200	M1real	1.584*	1.171	1.255	1.314	1.289
S&P 500 PE ratio	1.679*	1.207	1.192	1.280	1.157	M2real	1.610*	1.308	1.304	1.334	1.313
	Forecast combination methods						Forecast selection methods				
EW	1.613*	1.201	1.189	1.259	1.165	AR(1)	1.623*	1.215	1.252	1.200	1.275
DMSFE	1.611*	1.198	1.185	1.258	1.161	AR(1)+U	1.611*	1.239	1.261	1.214	1.295
LS-MMA	1.575*	1.239	1.186	1.201	1.147						

Note: This table reports ratios of out-of-sample MSFEs for inflation forecast using real-time data over the forecast evaluation period 1990:Q1-2023:Q3. See Notes to Table 4.

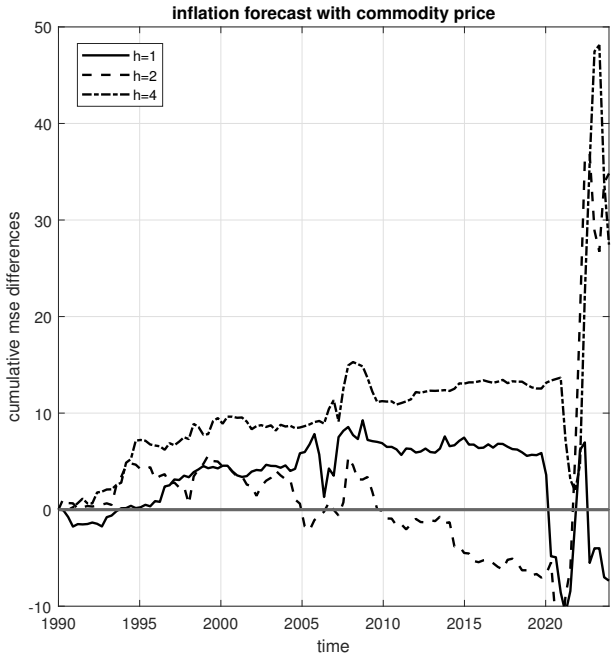

Figure 2: The figure presents cumulative sums of MSFEs differences (relative to the benchmark) for output growth and inflation from the individual model forecasts with commodity price.

6 Conclusion

In this paper, we consider practical issues associated with the use of local estimator in an out-of-sample forecasting context. We first propose an approach to select the bandwidth parameter by minimizing the conditional population loss at the end of the sample. The approach is similar to Inoue et al. (2017) for rolling window selection, but we show that asymptotic optimality holds when a generic weighting function is used for estimation and a general loss function is used for forecast evaluation.

We then move on to the implications on the choice of kernel weighting functions, which has been less addressed in the literature. Our analysis is based on the limiting behavior of the regret risk. We find that choice of kernel weighting function is related to bandwidth selection and reflects the usual bias-variance trade-off. When either estimation variance or estimation bias dominates, the criteria to select the optimal kernel weighting function is quite simple. However, when bandwidth is selected based on our approach, both estimation variance and bias are present in the limit, making the choice more involved.

Our theoretical analyses are evaluated through an extensive Monte Carlo study. Using a linear predictive regression model, we find that local estimator with optimal bandwidth generally improves forecast accuracy under various form of parameter instability. In general, using all but downweighting the data is preferred, particularly for the multi-step ahead forecasts.

We present two empirical applications. In the first application, we look at the bond return predictability. We find that our methods are particularly useful for 12-month overlapping returns. Model uncertainty also plays a role. Using a simple forecast selection rule from individual model forecasts with optimal bandwidth generally has the best forecasting performance. In our second application, we look at the real-time inflation forecasting. We find that using local estimator with optimal bandwidth generally performs better than forecasts from non-local estimator. Commodity price is a powerful predictor and gains are mostly from the post-pandemic period.

References

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

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

Cai, Z. and T. Juhl (2023). The distribution of rolling regression estimators. *Journal of Econometrics 235*(2), 1447–1463.

Chan, J. C. (2018). Specification tests for time-varying parameter models with stochastic volatility. *Econometric Reviews 37*(8), 807–823.

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

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

Croushore, D. and T. Stark (2001). A real-time data set for macroeconomists. *Journal of econometrics 105*(1), 111–130.

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

Davydov, Y. A. (1970). The invariance principle for stationary processes. *Theory of Probability & Its Applications 15*(3), 487–498.

Dendramis, Y., G. Kapetanios, and M. Marcellino (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 R. S. Mariano (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 253–263.

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

Farmer, L. E., L. Schmidt, and A. Timmermann (2023). Pockets of predictability. *The Journal of Finance 78*(3), 1279–1341.

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

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

Giraitis, L., G. Kapetanios, and T. Yates (2014). Inference on stochastic time-varying coefficient models. *Journal of Econometrics 179*(1), 46–65.

Granziera, E. and T. Sekhposyan (2019). Predicting relative forecasting performance: An empirical investigation. *International Journal of Forecasting 35*(4), 1636–1657.

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

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

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

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

Jurado, K., S. C. Ludvigson, and S. Ng (2015). Measuring uncertainty. *American Economic Review 105*(3), 1177–1216.

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

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

Kristensen, D. and Y. J. Lee (2023). Local polynomial estimation of time-varying parameters in nonlinear models. *Mimeo*.

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

Li, D., Z. Lu, and O. Linton (2012). Local linear fitting under near epoch dependence: uniform consistency with convergence rates. *Econometric Theory 28*(5), 935–958.

Liu, Y. and J. C. Wu (2021). Reconstructing the yield curve. *Journal of Financial Economics 142*(3), 1395–1425.

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

Marcellino, M., J. H. Stock, and M. W. Watson (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. and S. Ng (2020). Fred-qd: A quarterly database for macroeconomic research. Technical report, National Bureau of Economic Research.

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

Oh, D. H. and A. J. Patton (2021). Better the devil you know: Improved forecasts from imperfect models. *Mimeo*.

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

Rapach, D. E., J. K. Strauss, and G. Zhou (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 D. H. Romer (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*, Volume 2, pp. 1203–1324. Elsevier.

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

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

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

Stock, J. H. and M. W. Watson (2007). Why has us inflation become harder to forecast? *Journal of Money, Credit and banking 39*, 3–33.

Zhou, Z. and W. B. Wu (2010). Simultaneous inference of linear models with time varying coefficients. *Journal of the Royal Statistical Society Series B: Statistical Methodology 72*(4), 513–531.

Appendix

NOTATION: $\|\cdot\|$ is the Euclidean norm. $\|\cdot\|_p$ is the L_p norm. $x_n = O_p(y_n)$ states that the vector of random variables x_n is at most of order y_n in probability, and $x_n = o_p(y_n)$ is of smaller order than y_n in probability. $x_n \asymp y_n$ states that $x_n/y_n = O_p(1)$ or $x_n/y_n = O(1)$. The operator $\xrightarrow{p}$ denotes convergence in probability, and $\xrightarrow{d}$ denotes convergence in distribution. $E_T[\cdot] = E[\cdot|\mathcal{F}_T]$ is the conditional expectation operator, where $\mathcal{F}_T$ is the information set available at time T.

A The model and assumptions

We consider time series models of the following form

$$y_{t+h,T} = G\left(y_{t,T}, X_{t,T}, \varepsilon_t; \theta_{t,T}\right), \quad \theta_{t,T} = \theta\left(t/T\right), \ \ t = 1, 2, \cdots, T, \tag{A1}$$

where $G(y, x, \varepsilon; \theta)$ is a known function, $X_{t,T}$ contains exogenous predictors and ε_t is a sequence of errors and $1 \leqslant h < \infty$ is the forecast horizon. Collect $Z_{t,T} = (y_{t+h,T}, y_{t,T}, X'_{t,T})'$. Then, given the specification of G and the property of ε_t, we can obtain the corresponding loss: $\ell_{t,T}\left(\theta(t/T)\right) = \ell(Z_{t,T}; \theta(t/T))$.

Under certain regularity conditions on G and ε_t, it can be shown that[10], for each $u \in [0, 1]$, the stationary solution to the model (A1) exists and takes the following form:

$$y^*_{t+h}(u) = G\left(y^*_t(u), X^*_t(u), \varepsilon_t; \theta(u)\right). \tag{A2}$$

[10]For details, see Dahlhaus et al. (2019), Karmakar et al. (2022) and Kristensen and Lee (2023).

Before stating formally the technical assumptions, we introduce the following two definitions.

Definition A1. *A triangular array of processes $W_{t,T}(\theta)$, $\theta \in \Theta$, $t = 1, 2, \cdots, T$, $T = 1, 2, \cdots$ 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 $\sup_T E(|C_T|^\eta) < \infty$ for some $\eta > 0$.

Note that this definition follows from Kristensen and Lee (2023) to let an additional term ρ^t appear 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 A2. *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*

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

We are now ready to state the regularity conditions imposed for the derivations of all theoretical results.

Assumption A1. *$\theta_{t,T} = \theta(t/T)$, $\theta(\cdot) : (0, 1] \longrightarrow \Theta$ is twice continuously differentiable and $\Theta \subseteq \mathbb{R}^k$.*

Assumption A2. *(i) $\ell_{t,T}(\theta)$ is measurable and three-times continuously differentiable w.r.t. θ;*

(ii) $\ell_{t,T}(\theta)$ is locally stationary with stationary approximation $\tilde{\ell}_{u,t}(\theta)$ for each rescaled time point $u \in (0, 1]$;

(iii) $\ell^{(1)}_{t,T}(\theta) = \frac{\partial \ell_{t,T}(\theta)}{\partial \theta}$ is locally stationary with stationary approximation $\tilde{\ell}^{(1)}_{u,t}(\theta) = \frac{\partial \tilde{\ell}_{u,t}(\theta)}{\partial \theta}$ for each rescaled time point $u \in (0, 1]$;

(iv) For each $j = 1, 2, \cdots, k$, $\ell^{(2j)}_{t,T}(\theta) = \frac{\partial^2 \ell_{t,T}(\theta)}{\partial \theta \partial \theta_j}$ is locally stationary with stationary approximation $\tilde{\ell}^{(2j)}_{u,t}(\theta) = \frac{\partial^2 \tilde{\ell}_{u,t}(\theta)}{\partial \theta \partial \theta_j}$ for each rescaled time point $u \in (0, 1]$.

Assumption A3. *For the rescaled time point $u = 1$,*

(i) $\tilde{\ell}_{1,t}(\theta)$ is ergodic and L_1-continuous w.r.t θ; $E\left[\tilde{\ell}_{1,t}(\theta)\right]$ is uniquely minimized at θ_T;

(ii) $\tilde{\ell}_{1,t}^{(1)}(\theta)$ *is ergodic and central limit theorem (CLT) holds (as* $Tb \to \infty$*):*

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

where $\phi_{0,K} = \int_{\mathbb{B}} K^2(u)du$ *and* $\Lambda_T = \text{Var}\left(\frac{\partial \tilde{\ell}_{1,t}(\theta_T)}{\partial \theta'}\right)$*;*

(iii) For each $j = 1, 2, \cdots, k$, $\tilde{\ell}_{1,t}^{(2j)}(\theta)$ *is ergodic and all the eigenvalues of* $\tilde{\ell}_{1,t}^{(2)}(\theta) = \frac{\partial^2 \tilde{\ell}_{1,t}(\theta)}{\partial \theta \partial \theta'}$ *are uniformly bounded (below and above) over* $\theta \in \mathbb{R}^k$*.*

Assumption A4. *Let* $K(\cdot)$ *and* $\tilde{K}(\cdot)$ *be the kernel weighting functions for* (4) *and* (5)*, respectively:*

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

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

Assumption A5. *The bandwidth parameters* b *and* $\tilde{b}$ *are such that: (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 A6. *The choice set for* b *is* I_T*, where* $I_T \subset [\underline{b}, \overline{b}]$*. Also, the cardinality of* I_T*, denoted by* $\#I_T$*, satisfies* $\#I_T = \overline{b}^{\tau}$ *for some* $0 < \tau < 1$*.*

Assumption A1 impose conditions on the time-varying parameters. This includes cases when $\theta(\cdot)$ is modeled as smooth deterministic function of t/T and when $\theta(\cdot)$ is the path of persistent and bounded stochastic process (Remark 1).

Assumption A2 imposes conditions on the loss, its score and Hessian. We do not assume stationarity, but require the existence of stationary approximation for the scaled time point $u = 1$. 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 (A1). More details can be found in Dahlhaus et al. (2019) and Karmakar et al. (2022). Note that, the conditions are also model specific. Karmakar et al. (2022) provide analysis on both recursive defined time series (tvARMA or tvARCH models) and time-varying GARCH model.

Assumption A3 imposes conditions on the approximated stationary process for the rescaled time point $u = 1$. These conditions ensure that certain weak law of large numbers (WLLN) and CLT can be directly applied in the proof of Lemmas A1 and A2. Traditionally, this assumption can be verified by primitive conditions such as mixing conditions on the process. However, as explained in Li et al. (2012), mixing conditions may lead to some undesirable properties in time-varying parameter models. We can follow Inoue et al. (2017) by assuming that the process is near-epoch dependence. Our assumption follows closely from Cai and Juhl (2023), which make the use of the characterizations of processes from Zhou and Wu (2010).

Assumption A4 introduces conditions for the weighting function. As explained in Kristensen and Lee (2023), when local linear estimator is used, support of the kernel weighting function $\mathbb{B}$ should be compact. This rules out the use of certain weighting function, such as $K_2(u)$. Assumption A5 imposes conditions on the two bandwidth parameters which again ensures the asymptotic optimality of the bandwidth parameter selection procedure.

Assumption A6 implies the number of elements in 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 Hardle and Marron (1985).

B Data appendix

Table B1: Data description and variable transformation for predictors used in Section 5.2

Mnemonic	Description	Source	Transformation
Fedfunds	Effective Federal Funds Rate (Percent)	FRED-QD	y_t
Term spread	10-Year Treasury Constant Maturity Minus 3-Month Treasury Bill, secondary market (Percent)	FRED-QD	y_t
Default Spread	BAA-rated corporate bond yields Minus AAA-rated corporate bond yields (Percent)	FRED-QD	y_t
S&P 500	S&P's Common Stock Price Index: Composite	FRED-QD	$100\Delta \ln y_t$
S&P PE ratio	S&P's Composite Common Stock: Price-Earnings Ratio	FRED-QD	$100\Delta \ln y_t$
EXCAUSx	Canada / U.S. Foreign Exchange Rate	FRED-QD	$100\Delta \ln y_t$
EXUSUKx	U.S. / U.K. Foreign Exchange Rate	FRED-QD	$100\Delta \ln y_t$
Commodity	Moody's commodity price index	GFD	$100\Delta \ln y_t$
M1real	Real M1 Money Stock (Billions of 1982-84 Dollars), deflated by CPI	FRED-QD	$100\Delta \ln y_t$
M2real	Real M2 Money Stock (Billions of 1982-84 Dollars), deflated by CPI	FRED-QD	$100\Delta \ln y_t$

Note: FRED-QD refers to the database developed in McCracken and Ng (2020) and maintained by the Federal Reserve Bank of St. Louis. GFD refers to the Global Financial Database.

C Additional empirical results

Table C1: Out-of-sample forecasting performance: 12-month overlapping bond returns, 1990:M1-2019:M12

n		OLS	Fixed	Opt-K_1	Opt-K_2	Opt-K_3	n		OLS	Fixed	Opt-K_1	Opt-K_2	Opt-K_3
				Individual models							Individual models		
2	FB	1.005	1.224	0.990	0.921	1.006	3	FB	0.948	1.159	0.945	0.892	0.972
	CP	1.087	1.066	0.847	0.851	0.872		CP	1.082	0.995	0.801	0.814	0.815
	LN	0.837	0.730	0.680	0.664	0.679		LN	0.785	0.745	0.688	0.673	0.695
	FB+CP+LN	0.814	0.571*	0.605*	0.582	0.570*		FB+CP+LN	0.790	0.598*	0.598*	0.606*	0.583*
				Forecast combination methods							Forecast combination methods		
	EW	0.787	0.735*	0.632*	0.643*	0.624*		EW	0.773	0.727	0.620*	0.646*	0.625*
	DMSFE	0.691	0.624	0.560*	0.575*	0.544*		DMSFE	0.689	0.640*	0.558*	0.590*	0.562*
	LS-MMA	0.876	0.716	0.655	0.661	0.632		LS-MMA	0.847	0.733	0.661	0.685*	0.636*
				Forecast selection methods							Forecast selection methods		
	AR(1)	0.560*	0.500*	0.460*	0.483*	0.433*		AR(1)	0.520*	0.479*	0.441*	0.496*	0.422*
	AR(1)+U	0.567*	0.507*	0.454*	0.490*	0.440*		AR(1)+U	0.544*	0.489*	0.439*	0.508*	0.428*
				Individual models							Individual models		
4	FB	0.910	1.080	0.905	0.859	0.918	5	FB	0.865	1.002	0.866	0.825	0.868
	CP	1.078	0.908	0.753	0.770	0.755		CP	1.079	0.840	0.724	0.744	0.718
	LN	0.760	0.744	0.691	0.670	0.701		LN	0.756	0.756	0.708	0.681*	0.712
	FB+CP+LN	0.775	0.599*	0.587*	0.590*	0.574*		FB+CP+LN	0.775	0.616*	0.621*	0.605*	0.604*
				Forecast combination methods							Forecast combination methods		
	EW	0.751	0.692*	0.603*	0.626*	0.609*		EW	0.747	0.675	0.605*	0.626*	0.609*
	DMSFE	0.667*	0.619*	0.547*	0.578*	0.555*		DMSFE	0.668*	0.615*	0.558*	0.585*	0.565*
	LS-MMA	0.818	0.698	0.640*	0.658*	0.618*		LS-MMA	0.813	0.704	0.658*	0.668*	0.638*
				Forecast selection methods							Forecast selection methods		
	AR(1)	0.496*	0.482*	0.441*	0.479*	0.428*		AR(1)	0.503*	0.474*	0.442*	0.475*	0.426*
	AR(1)+U	0.513*	0.498*	0.438*	0.487*	0.432*		AR(1)+U	0.513*	0.483*	0.427*	0.474*	0.423*

Note: This table reports ratios of out-of-sample MSFEs for monthly bond excess returns, $rx_{t+12}^{(n)}$, measured relative to the (annualized) T-bill rate $y_t^{(1)}$, over the forecasting evaluation period 1990:M1-2019:M12. See Notes to Table 4.

Table C2: Out-of-sample forecasting performance: one-month bond excess returns, 1990:M1-2019:M12

n		OLS	Fixed	Opt-K_1	Opt-K_2	Opt-K_3	n		OLS	Fixed	Opt-K_1	Opt-K_2	Opt-K_3
				Individual models							Individual models		
2	FB	0.999	1.059	1.022	0.998	1.038	3	FB	0.980	1.047	1.026	0.995	1.035
	CP	1.304*	1.090*	1.054	1.063	1.071		CP	1.220*	1.074*	1.049	1.052*	1.067*
	LN	1.014	1.060*	1.024	1.015	1.036		LN	1.007	1.052*	1.026	1.016	1.038*
	FB+CP+LN	1.379*	1.165*	1.129*	1.101*	1.145*		FB+CP+LN	1.235*	1.144*	1.116*	1.079*	1.134*
				Forecast combination methods							Forecast combination methods		
	EW	1.100	1.070	1.035	1.028	1.051		EW	1.060	1.057	1.034	1.020	1.048
	DMSFE	1.068	1.070	1.034	1.026	1.050		DMSFE	1.045	1.057	1.034	1.020	1.048
	LS-MMA	0.993	1.052	1.025	0.998	1.035		LS-MMA	0.981	1.039	1.025	0.996	1.032
				Forecast selection methods							Forecast selection methods		
	AR(1)	0.981	1.088	1.045	1.003	1.081*		AR(1)	0.986	1.077*	1.042	1.003	1.061
	AR(1)+U	0.986	1.086*	1.039	1.009	1.044		AR(1)+U	0.993	1.061*	1.045	1.007	1.049
				Individual models							Individual models		
4	FB	0.977	1.028	1.021	0.994	1.033	5	FB	0.984	1.032	1.036	0.999	1.041
	CP	1.179*	1.060*	1.045*	1.045*	1.059*		CP	1.157*	1.055*	1.043*	1.042*	1.056*
	LN	1.002	1.044	1.028	1.014	1.035		LN	0.998	1.039	1.028	1.012	1.035
	FB+CP+LN	1.172*	1.125*	1.107*	1.061	1.117*		FB+CP+LN	1.143*	1.117*	1.104*	1.058	1.116*
				Forecast combination methods							Forecast combination methods		
	EW	1.041	1.042	1.030	1.015	1.040		EW	1.033*	1.041	1.035	1.016	1.043
	DMSFE	1.031	1.042	1.030	1.014	1.040		DMSFE	1.026*	1.041	1.035	1.015	1.043
	LS-MMA	0.977	1.025	1.022	0.995	1.027		LS-MMA	0.985	1.029	1.031	1.000	1.036
				Forecast selection methods							Forecast selection methods		
	AR(1)	0.976	1.053	1.043	0.994	1.052		AR(1)	0.985	1.040	1.050*	1.006	1.051
	AR(1)+U	0.982	1.046*	1.055*	0.999	1.040		AR(1)+U	1.011	1.057*	1.055*	1.011	1.055*

Note: This table reports ratios of out-of-sample MSFEs for monthly bond excess returns, $rx_{t+1}^{(n)}$, measured relative to the (annualized) T-bill rate $(1/12)y_t^{(1/12)}$, over the forecasting evaluation period 1990:M1-2019:M12. See Notes to Table 4.

Table C3: Out-of-sample forecasting performance for inflation: 1990:Q1-2019:Q4

	OLS	Fixed	Opt-K_1	Opt-K_2	Opt-K_3		OLS	Fixed	Opt-K_1	Opt-K_2	Opt-K_3
					$h = 1$						
		Individual models						Individual models			
Fedfunds	1.256*	1.096	1.033	1.037	1.080	EXCAUSx	1.333	1.079	1.084	1.077	1.117
Term spread	1.344*	1.111	1.075	1.077	1.077	EXUSUKx	1.302*	1.110	1.089	1.056	1.142
Default spread	1.327*	1.130	1.167	1.064	1.370	Commodity	1.230*	1.129	1.087	1.071	1.120
S&P 500	1.401*	1.132	1.119	1.072	1.203	M1real	1.345*	1.066	1.011	1.045	1.055
S&P 500 PE ratio	1.366*	1.059	0.990	1.031	1.020	M2real	1.330*	1.084	1.033	1.038	1.065
		Forecast combination methods						Forecast selection methods			
EW	1.303*	1.036	0.984	1.018	1.016	AR(1)	1.258	1.059	1.106	1.068	1.073
DMSFE	1.304*	1.034	0.982	1.017	1.016	AR(1)+U	1.302*	1.181	1.143	1.050	1.162
LS-MMA	1.296*	1.098	1.031	1.056	1.053						
					$h = 2$						
		Individual models						Individual models			
Fedfunds	1.287	1.015	1.012	0.962	1.021	EXCAUSx	1.314	0.935	0.900	0.887	0.918
Term spread	1.425	1.020	1.030	1.012	1.026	EXUSUKx	1.351	0.940	0.933	0.943	0.972
Default spread	1.355	0.851	0.784	0.944	0.806	Commodity	1.255	0.947	0.914	0.892	0.923
S&P 500	1.443	0.981	0.928	0.996	0.931	M1real	1.368	0.878	0.798	0.887	0.802
S&P 500 PE ratio	1.418	0.996	0.968	0.984	0.983	M2real	1.369	0.834	0.905	0.890	0.914
		Forecast combination methods						Forecast selection methods			
EW	1.328	0.857	0.813	0.890	0.812	AR(1)	1.413	0.835	0.915	1.110	0.944
DMSFE	1.329	0.850	0.795	0.893	0.792	AR(1)+U	1.418	0.908	0.893	1.130	0.858
LS-MMA	1.353	1.006	0.833	0.962	0.789						
					$h = 4$						
		Individual models						Individual models			
Fedfunds	2.277*	1.454	1.450	1.602*	1.418	EXCAUSx	2.296*	1.400	1.327	1.413	1.227
Term spread	2.523*	1.479	1.499	1.600*	1.381	EXUSUKx	2.306*	1.471*	1.409*	1.454*	1.368
Default spread	2.482*	1.404	1.288	1.381*	1.451	Commodity	2.198*	1.355	1.245	1.293	1.210
S&P 500	2.284*	1.419	1.258	1.390*	1.213	M1real	2.402*	1.421	1.300	1.490*	1.245
S&P 500 PE ratio	2.421*	1.514*	1.453*	1.553*	1.444	M2real	2.355*	1.462	1.472	1.522*	1.407
		Forecast combination methods						Forecast selection methods			
EW	2.283*	1.345	1.240	1.403*	1.173	AR(1)	2.395*	1.209	1.310	1.383	1.406
DMSFE	2.276*	1.332	1.229	1.398*	1.164	AR(1)+U	2.375*	1.292	1.381	1.411	1.485
LS-MMA	2.247*	1.404	1.323	1.349	1.224						

Note: This table reports ratios of out-of-sample MSFEs for inflation forecast using real-time data over the forecast evaluation period 1990:Q1-2019:Q4. See Notes to Table 4.

Online Appendix: Nonparametric estimation and forecasting of time-varying parameter models

The online appendix is organized as follows. Section A presents axillary results. Section B provides proofs of main theorems 1-2. Section C presents some additional simulation results.

NOTATION: $\|\cdot\|$ is the Euclidean norm. $\|\cdot\|_p$ is the L_p norm. $x_n = O_p(y_n)$ states that the vector of random variables x_n is at most of order y_n in probability, and $x_n = o_p(y_n)$ is of smaller order than y_n in probability. $x_n \asymp y_n$ states that $x_n/y_n = O_p(1)$ or $x_n/y_n = O(1)$. The operator $\xrightarrow{p}$ denotes convergence in probability, and $\xrightarrow{d}$ denotes convergence in distribution. $E_T[\cdot] = E[\cdot|\mathcal{F}_T]$ is the conditional expectation operator, where $\mathcal{F}_T$ is the information set available at time T.

A Auxiliary results

Lemma A1. *Suppose that Assumptions A1(i), A2, A3 and A4(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:

$$\left\|\hat{\theta}_{K,b,T} - \theta_T\right\| = O_p\left(r_{T,b}\right),$$

where $r_{T,b} = (Tb)^{-1/2} + b$;

(iii) CLT: 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 $\Sigma_T = H_T^{-1}\Lambda_T H_T^{-1}$, $\mu_{1,K} = \int_{\mathcal{B}} uK(u)du$, $\phi_{0,K} = \int_{\mathcal{B}} K^2(u)du$, $\Lambda_T = \mathrm{Var}\left(\frac{\partial\tilde{\ell}_{1,t}(\theta_T)}{\partial\theta'}\right)$ and $H_T = E\left[\frac{\partial^2\tilde{\ell}_{1,t}(\theta_T)}{\partial\theta\partial\theta'}\right]$.

Proof. Recall the definition of local estimator:

$$\hat{\theta}_{K,b,T} = \arg\min_{\theta\in\Theta} \frac{1}{Tb}\sum_{t=1}^{T} k_{tT}\ell_{t,T}(\theta), \tag{A1}$$

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

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 A2 and Definition A1, we have

$$\begin{aligned}\sup_{\theta\in\Theta}\left|L_T(\theta)-\tilde{L}_T(\theta)\right| &\leqslant \sup_{\theta\in\Theta}\frac{1}{Tb}\sum_{t=1}^{T}k_{tT}\left|\ell_{t,T}(\theta)-\tilde{\ell}_{1,t}(\theta)\right| \\ &\leqslant \frac{1}{Tb}\sum_{t=1}^{T}k_{tT}(T^{-1}+\rho^t)=O(T^{-1})+O\left((Tb)^{-1/2}\right)=o(1), \qquad \text{(A2)}\end{aligned}$$

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 (A1) 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 MacFadden (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 A3(i));

(2) Θ is compact (assumed in Assumption A1);

(3) $\tilde{L}_T(\theta)$ is continuous (implied by Assumption A2(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 A3(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

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

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

$$\begin{aligned}
\hat{\theta}_{K,b,T} - \theta_T &= -\left(\frac{\partial^2 L_T\left(\overline{\theta}_T\right)}{\partial\theta\partial\theta'}\right)^{-1}\left(\frac{\partial L_T(\theta_T)}{\partial\theta}\right) \\
&= -\left(\frac{\partial^2 L_T(\theta_T)}{\partial\theta\partial\theta'}\right)^{-1}\left(\frac{\partial L_T(\theta_T)}{\partial\theta}\right) + \left[\left(\frac{\partial^2 L_T(\theta_T)}{\partial\theta\partial\theta'}\right)^{-1} - \left(\frac{\partial^2 L_T\left(\overline{\theta}_T\right)}{\partial\theta\partial\theta'}\right)^{-1}\right]\frac{\partial L_T(\theta_T)}{\partial\theta} \\
&= -\left(\frac{\partial^2 L_T(\theta_T)}{\partial\theta\partial\theta'}\right)^{-1}\left(\frac{\partial L_T(\theta_T)}{\partial\theta}\right) + \left(\frac{\partial^2 L_T(\theta_T)}{\partial\theta\partial\theta'}\right)^{-1}\left[\frac{\partial^2 L_T\left(\overline{\theta}_T\right)}{\partial\theta\partial\theta'} - \frac{\partial^2 L_T(\theta_T)}{\partial\theta\partial\theta'}\right] \\
&\qquad \times\left(\frac{\partial^2 L_T\left(\overline{\theta}_T\right)}{\partial\theta\partial\theta'}\right)^{-1}\frac{\partial L_T(\theta_T)}{\partial\theta}, \\
&:= -H_T^{-1}(\theta_T)\, S_T(\theta_T) + H_T^{-1}(\theta_T)\left[H_T\left(\overline{\theta}_T\right) - H_T(\theta_T)\right]H_T^{-1}\left(\overline{\theta}_T\right)S_T(\theta_T) \qquad \text{(A3)}
\end{aligned}$$

We will show that

$$\left\|H_T^{-1}(\theta_T)\right\| = O_p(1), \qquad \text{(A4)}$$

$$\|S_T(\theta_T)\| = O_p\left((Tb)^{-1/2} + b\right), \qquad \text{(A5)}$$

$$\left\|H_T\left(\overline{\theta}_T\right) - H_T(\theta_T)\right\| = o_p(1). \qquad \text{(A6)}$$

These bounds together with (A3) implies the consistency rate in A1(i).

Proof of (A4). It follows similarly from (A2) that

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

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

$$\begin{aligned}
\tilde{H}_T(\theta_T) &= \frac{1}{Tb}\sum_{t=1}^{T} k_{tT} E\left[\frac{\partial^2\tilde{\ell}_{1,t}(\theta_T)}{\partial\theta\partial\theta'}\right] + \frac{1}{Tb}\sum_{t=1}^{T} k_{tT}\left(\frac{\partial^2\tilde{\ell}_{1,t}(\theta_T)}{\partial\theta\partial\theta'} - E\left[\frac{\partial^2\tilde{\ell}_{1,t}(\theta_T)}{\partial\theta\partial\theta'}\right]\right) \\
&= \tilde{H}_T^*\left(I_k + \tilde{\Delta}_T\right), \qquad \text{(A7)}
\end{aligned}$$

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 A3(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 A3(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 (A5). We have that

$$\begin{aligned} S_T(\theta_T) = \frac{\partial L_T(\theta_T)}{\partial\theta} &= \frac{1}{Tb}\sum_{t=1}^{T} k_{tT}\frac{\partial \ell_{t,T}(\theta_T)}{\partial\theta} \\ &= \frac{1}{Tb}\sum_{t=1}^{T} k_{tT}\frac{\partial \ell_{t,T}(\theta_t)}{\partial\theta} + \frac{1}{Tb}\sum_{t=1}^{T} k_{tT}\frac{\partial^2 \ell_{t,T}\left(\overline{\theta}_T\right)}{\partial\theta\partial\theta'}\,(\theta_T - \theta_t) \\ &:= S_T(\theta_t) + B_T, \end{aligned} \tag{A8}$$

where the second line follows from mean-value theorem. Let us first consider $S_T(\theta_t)$. Using the similar argument as in (A2), 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 A3(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 A1, 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

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

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

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

By Assumption A3(iii),

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

and

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

where 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 (A5).

Proof of (A6). This follow immediately by the consistency: $\hat{\theta}_{K,b,T} \xrightarrow{p} \theta_T$.

Back to (A3), under the condition $b = O(T^{-1/3})$, we have

$$\sqrt{Tb}\left(\hat{\theta}_{K,b,T} - \theta_T + \tilde{H}_T^{-1}\left(\theta_T\right)\tilde{B}_T\right) = -\tilde{H}_T^{-1}\left(\theta_T\right)\ \sqrt{Tb}\tilde{S}_T(\theta_T). \tag{A9}$$

As the dominating term of the asymptotic bias is given by

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

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

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

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 $\Lambda_T = Var\left(\frac{\partial\tilde{\ell}_{1,t}(\theta_T)}{\partial\theta'}\right)$. □

Lemma A2. *Suppose that Assumptions A1(ii), A2, A3 and A4(ii) hold with $\tilde{b} \to 0$ and $T\tilde{b} \to \infty$. Then, it holds that*

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

Proof. 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 (A3), we have that

$$\begin{pmatrix}\tilde{\theta}_T - \theta_T \\ \tilde{\theta}_T^{(1)} - \theta_T^{(1)}\end{pmatrix} = -\begin{bmatrix}\frac{1}{T\tilde{b}}\sum_{t=1}^{T}\tilde{k}_{tT}\frac{\partial^2\ell_{t,T}(\beta)}{\partial\theta\partial\theta'} & \frac{1}{T\tilde{b}}\sum_{t=1}^{T}\tilde{k}_{tT}\frac{\partial^2\ell_{t,T}(\beta)}{\partial\theta\partial\theta'^{(1)}}\left(\frac{t-T}{T}\right) \\ \frac{1}{T\tilde{b}}\sum_{t=1}^{T}\tilde{k}_{tT}\frac{\partial^2\ell_{t,T}(\beta)}{\partial\theta^{(1)}\partial\theta'}\left(\frac{t-T}{T}\right) & \frac{1}{T\tilde{b}}\sum_{t=1}^{T}\tilde{k}_{tT}\frac{\partial^2\ell_{t,T}(\beta)}{\partial\theta^{(1)}\partial\theta'^{(1)}}\left(\frac{t-T}{T}\right)^2\end{bmatrix}^{-1}\begin{bmatrix}\frac{1}{T\tilde{b}}\sum_{t=1}^{T}\tilde{k}_{tT}\frac{\partial\ell_{t,T}(\beta)}{\partial\theta} \\ \frac{1}{T\tilde{b}}\sum_{t=1}^{T}\tilde{k}_{tT}\frac{\partial\ell_{t,T}(\beta)}{\partial\theta^{(1)}}\left(\frac{t-T}{T}\right)\end{bmatrix} + o_p(1). \quad (A10)$$

Using similar arguments for the proofs of (A4)-(A5), 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 (A4)-(A5), we have

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

and

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

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

$$\begin{pmatrix}\tilde{\theta}_T - \theta_T\\ \tilde{\theta}_T^{(1)} - \theta_T^{(1)}\end{pmatrix} = -\begin{bmatrix}O_p(1) & O_p(\tilde{b})\\ O_p(\tilde{b}) & O_p(\tilde{b}^2)\end{bmatrix}^{-1}\begin{bmatrix}O_p\left((T\tilde{b})^{-1/2}\right) + O_p\left(\tilde{b}^2\right)\\ O_p((T\tilde{b})^{-1/2}\tilde{b}) + O_p(\tilde{b}^3)\end{bmatrix} + o_p(1) \tag{A11}$$

$$= \begin{bmatrix}O_p\left((T\tilde{b})^{-1/2} + \tilde{b}^2\right)\\ O_p\left(T^{-1/2}\tilde{b}^{-3/2} + \tilde{b}\right)\end{bmatrix} \tag{A12}$$

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

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

□

Lemma A3. *Suppose that Assumptions A1, A2, A3 and A4(i) hold with $b \to 0$ and $Tb \to \infty$. Then, for some $0 < \delta < \frac{1}{2}$, it holds that*

$$\sup_{b\in I_T}\left\|\hat{\theta}_{K,b,T} - \theta_T\right\| = O_p(r_{T,b,\delta}), \tag{A13}$$

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

Proof. Given the kernel weighting function $\overline{K}$, write $\hat{\theta}_{\overline{K},b,T} = \hat{\theta}_{b,T}$. As in (A3), the estimator can be decomposed as

$$\begin{aligned}\hat{\theta}_{b,T} - \theta_T &= -H_T(\theta_T)S_T(\theta_T) + o_p(1)\\ &= -H_T(\theta_T)\left(S_T(\theta_t) + B_T\right) + o_p(1),\end{aligned} \tag{A14}$$

where

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

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

$$\sup_{b\in I_T} \left\|T^{1/2}b^{1/2+\delta}S_T(\theta_t)\right\| = O_p(1), \tag{A15}$$
$$\sup_{b\in I_T} \left\|H_T(\theta_T)^{-1}\right\| = O_p(1), \tag{A16}$$
$$\sup_{b\in I_T} \left\|b^{\delta}B_T\right\| = O_p(b), \tag{A17}$$

for some $0 < \delta < 1/2$. These bounds together with (A14) prove (A13).

Proof of (A15). By Boole's inequality and Chebyshev's inequality, we have, for any $\varepsilon > 0$,

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

where the third inequality follows from the proof of (A5) 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 A6.

Proof of (A16). Recall that

$$\begin{aligned}\tilde{H}_T = \frac{1}{Tb}\sum_{t=1}^{T} k_{tT}\frac{\partial^2 \tilde{\ell}_{1,t}(\theta_T)}{\partial\theta\partial\theta'} &= \frac{1}{Tb}\sum_{t=1}^{T} k_{tT}E\left[\frac{\partial^2 \tilde{\ell}_{1,t}(\theta_T)}{\partial\theta\partial\theta'}\right] + \frac{1}{Tb}\sum_{t=1}^{T} k_{tT}\left(\frac{\partial^2 \tilde{\ell}_{1,t}(\theta_T)}{\partial\theta\partial\theta'} - E\left[\frac{\partial^2 \tilde{\ell}_{1,t}(\theta_T)}{\partial\theta\partial\theta'}\right]\right)\\ &:= \tilde{H}_{T,1}\left(I_k + \tilde{\Delta}_T\right),\end{aligned} \tag{A18}$$

where $\tilde{\Delta}_T = \left(\tilde{H}_{T,1}\right)^{-1}\left(\tilde{H}_T - \tilde{H}_{T,1}\right)$. First, (A4) holds uniformly over b:

$$\sup_{b\in I_T}\left\|\tilde{H}_{T,1}^{-1}\right\|_{sp} = O_p(1). \tag{A19}$$

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 Cheby-

shev's inequality, we have

$$\mathbb{P}\left(\sup_{b\in I_T}\left\|\frac{1}{Tb}\sum_{t=1}^{T}k_{tT}\tilde{\Delta}_t\right\|>\varepsilon\right)\leqslant\sum_{b\in I_T}\mathbb{P}\left(\left\|\frac{1}{Tb}\sum_{t=1}^{T}k_{tT}\tilde{\Delta}_t\right\|>\varepsilon\right)$$
$$\leqslant\underbrace{\#I_T}_{O(b^\delta)}\times\underbrace{\sup_{b\in I_T}\mathbb{P}\left(\left\|\frac{1}{Tb}\sum_{t=1}^{T}k_{tT}\tilde{\Delta}_t\right\|>\varepsilon\right)}_{o(1)}=o(1). \tag{A20}$$

To sum up, we continue from (A7):

$$\sup_{b\in I_T}\left\|\tilde{H}_T^{-1}\right\|_{sp}\leqslant\underbrace{\sup_{b\in I_T}\left\|\tilde{H}_{T,1}^{-1}\right\|_{sp}}_{O_p(1)\text{ by (A19)}}\left(1-\underbrace{\sup_{b\in I_T}\left\|\tilde{\Delta}_T\right\|_{sp}}_{o_p(1)\text{ by (A20)}}\right)^{-1}=O_p(1).$$

This also implies (A16).

Proof of (A17). Recall that the stationary approximation of B_T is $\tilde{B}_T$, where $\tilde{B}_T=\tilde{B}_{T,1}+\tilde{B}_{T,2}$:

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

For $\tilde{B}_{T,1}$, again, similarly as in (A15), 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 (A17). □

Lemma A4. *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 A1-A5 hold, we have*

$$\sup_{b\in I_T}\left|\frac{L(b)-A(b)}{L(b)}\right| = o_p(1). \tag{A21}$$

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 A2 and Assumption A5(i), we have

$$\left\|\tilde{\theta}_T-\theta_T\right\| = O_p\left((T\tilde{b})^{-1/2}\right). \tag{A22}$$

We will show that

$$\sup_{b\in I_T}\left|\frac{D_1(b)}{L(b)}\right| = o_p(1), \ \sup_{b\in I_T}\left|\frac{D_2(b)}{L(b)}\right| = o_p(1), \ \sup_{b\in I_T}\left|\frac{D_3(b)}{L(b)}\right| = o_p(1), \tag{A23}$$

$$\sup_{b\in I_T}\left|\frac{D_1'}{L(b)}\right| = o_p(1), \ \sup_{b\in I_T}\left|\frac{D_2'}{L(b)}\right| = o_p(1). \tag{A24}$$

These bounds together with triangular inequality imply (A21).

Proof of (A23). First, by Lemma A3 and Assumption A3(iii), we have $\|\omega_T(\theta_T)\|_{sp} = O_p(1)$ and

$$\sup_{b\in I_T}|L(b)| \leqslant \sup_{b\in I_T}\left\|\hat{\theta}_{b,T}-\theta_T\right\| \|\omega_T(\theta_T)\|_{sp} \sup_{b\in I_T}\left\|\hat{\theta}_{b,T}-\theta_T\right\| = O_p\left(r^2_{T,b,\delta}\right), \tag{A25}$$

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

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

where the second and third line follow from the fact that $\omega^{(1)}_T(\theta_T)$ involves $\tilde{\theta}_T-\theta_T$ so the order of $\left\|\omega^{(1)}_T(\theta_T)\right\|_{sp} = O_p(\tilde{r}_{T,\tilde{b}})$, 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 A5. 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

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

where $\frac{\tilde{r}^2_{T,\tilde{b}}}{r_{T,b,\delta}}\to 0$ is again guaranteed by Assumption A5.

Proof of (A24). First, it is straightforward to show that

$$\left|D'_1\right| = O_p(\tilde{r}^2_{T,\tilde{b}}),\ \ \left|D'_2\right| = O_p(\tilde{r}^3_{T,\tilde{b}}).$$

Together with (A25) and following the same reasoning above, we have

$$\sup_{b\in I_T}\left|\frac{D'_1}{L(b)}\right| = O_p\left(\frac{\tilde{r}^2_{T,\tilde{b}}}{r^2_{T,b,\delta}}\right) = o_p(1),\ \ \sup_{b\in I_T}\left|\frac{D'_2}{L(b)}\right| = O_p\left(\frac{\tilde{r}^3_{T,\tilde{b}}}{r^2_{T,b,\delta}}\right) = o_p(1).$$

□

B Proofs of the theorems

B.1 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 A1 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.

B.2 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 B4, 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 (A23) and (A24), 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 estabilish that

$$I_1(\hat{b}) \xrightarrow{p} 1. \tag{B1}$$

For any $b, b' \in I_T$, it follows immediately from Lemma B4 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 (B1).

C Additional simulation results

This section presents additional simulation results using different specifications for $(a_t, b_t)'$ in (15). In particular, we consider the case when the parameters are constant over time $(a_t, b_t)' = (a, b)$ for all t, as well as cases when $(a_t, b_t)'$ have a one-time break point. We also consider the case when $(a_t, b_t)'$ follows a random coefficient model as in Nyblom (1989).

We have 8 different specifications for $(a_t, b_t)'$. For DGP D1, $(a_t, b_t)' = (0.9, 1)'$ for all t. For DGPs D2-D7, they are generated according to

(1) D2-D4: $a_t = 0.9 - \frac{1}{T^{0.2}} \mathbb{1}(t \geqslant \pi T + 1)$, $b_t = 1 + \frac{1}{T^{0.2}} \mathbb{1}(t \geqslant \pi T + 1)$, where $\pi = 0.25, 0.5, 0.75$, respectively;

(2) D5-D7: $a_t = 0.9 - \frac{1}{T^{0.5}} \mathbb{1}(t \geqslant \pi T + 1)$, $b_t = 1 + \frac{1}{T^{0.5}} \mathbb{1}(t \geqslant \pi T + 1)$, where $\pi = 0.25, 0.5, 0.75$, respectively.

Notice that, the main differences between DGPs D5-D7 and D2-D4 are that the break size is relatively larger in the previous group. Finally, for DGP D8, we set $a_t = a_{t-1} + (0.05/\sqrt{T})\epsilon_t$, $b_t = b_{t-1} + (0.1/\sqrt{T})\epsilon_t$, where $\epsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$, $a_0 = 0.9$, and $b_0 = 1$.

Table D1 presents the out-of-sample forecasting performance from the simulated dataset. Let us start by commenting the results obtained from DGPs D1-D7. Overall, for 1-step ahead forecast, the local estimator with optimal bandwidths selection is quite useful when break size is relatively small. For 12-step ahead forecast, all local estimators perform better than non-local

estimator even if the true parameters are constant over time. Overall, rolling window forecasts with window size equal to 60 are more likely to be the best, but the performance from using $K_2(u)$ with optimal bandwidth selection improves as sample size increases, particularly for 12-step ahead forecast. For DGP D8, as the time variations in parameters are very small, performance from local estimators are very similar compared to the non-local estimator.

References

Newey, W. and D. MacFadden (1994). Large sample estimation and hypothesis testing, chapter 36. *Handbook of Econometrics Vol 4*.

Nyblom, J. (1989). Testing for the constancy of parameters over time. *Journal of the American Statistical Association 84*(405), 223–230.

Table D1: Forecasting performance from simulated dataset

DGP	Fixed1	Fixed2	Opt-K_1	Opt-K_2	Opt-K_3	Fixed1	Fixed2	Opt-K_1	Opt-K_2	Opt-K_3
			$h = 1$					$h = 12$		
					T=150					
D1	1.035	1.022	1.049	1.019	1.058	0.889	0.871	0.917	0.887	0.926
D2	0.919	0.909	0.944	0.927	0.947	0.909	0.902	0.942	0.917	0.947
D3	0.829	0.820	0.852	0.849	0.856	0.710	0.703	0.738	0.723	0.741
D4	0.838	0.899	0.834	0.874	0.835	0.710	0.718	0.719	0.718	0.720
D5	1.026	1.012	1.043	1.013	1.053	0.882	0.868	0.913	0.882	0.926
D6	1.017	1.005	1.035	1.006	1.045	0.829	0.820	0.860	0.838	0.870
D7	1.007	1.002	1.020	0.998	1.027	0.830	0.825	0.850	0.832	0.858
D8	1.030	1.015	1.031	1.019	1.041	0.999	1.000	0.999	0.999	0.999
					T=300					
D1	1.034	1.023	1.030	1.015	1.036	0.914	0.894	0.916	0.900	0.926
D2	0.923	0.912	0.928	0.918	0.929	0.907	0.895	0.916	0.901	0.920
D3	0.826	0.816	0.831	0.826	0.834	0.723	0.716	0.732	0.725	0.734
D4	0.830	0.822	0.829	0.833	0.830	0.721	0.710	0.727	0.722	0.730
D5	1.032	1.019	1.027	1.012	1.031	0.907	0.888	0.912	0.896	0.920
D6	1.022	1.006	1.018	1.003	1.024	0.857	0.847	0.865	0.853	0.870
D7	1.021	1.012	1.020	1.004	1.022	0.852	0.839	0.856	0.846	0.863
D8	1.003	1.001	0.999	0.999	0.999	1.000	1.000	1.000	0.999	0.999
					T=450					
D1	1.032	1.023	1.025	1.012	1.027	0.906	0.887	0.895	0.887	0.901
D2	0.911	0.902	0.908	0.902	0.907	0.882	0.869	0.878	0.871	0.881
D3	0.845	0.837	0.842	0.838	0.844	0.728	0.718	0.728	0.723	0.731
D4	0.845	0.837	0.840	0.838	0.841	0.725	0.716	0.722	0.716	0.724
D5	1.031	1.023	1.024	1.012	1.026	0.901	0.883	0.894	0.883	0.897
D6	1.029	1.015	1.017	1.007	1.021	0.867	0.849	0.860	0.853	0.865
D7	1.022	1.013	1.013	1.003	1.017	0.863	0.843	0.854	0.846	0.860
D8	0.990	0.992	0.992	0.995	0.993	1.001	1.001	1.000	1.000	1.000
					T=600					
D1	1.037	1.025	1.023	1.013	1.025	0.921	0.899	0.901	0.892	0.907
D2	0.898	0.889	0.891	0.888	0.889	0.877	0.866	0.871	0.864	0.871
D3	0.864	0.852	0.857	0.854	0.858	0.730	0.720	0.725	0.721	0.727
D4	0.855	0.847	0.847	0.842	0.848	0.727	0.714	0.717	0.713	0.719
D5	1.030	1.020	1.018	1.009	1.020	0.903	0.882	0.886	0.877	0.891
D6	1.024	1.016	1.013	1.005	1.015	0.869	0.848	0.856	0.849	0.858
D7	1.019	1.010	1.009	1.000	1.009	0.876	0.858	0.864	0.858	0.866
D8	1.0409	1.060	0.999	0.999	1.001	1.003	1.002	1.002	1.002	1.003

Note: Fixed1: rolling window estimator with window size equal to 40; Fixed2: rolling window estimator with window size equal to 60; Opt-K_i: local estimator with optimal bandwidth selection, where $K_1(u) = \mathbb{1}_{\{-1<u<0\}}$, $K_2(u) = \frac{2}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) \mathbb{1}_{\{u<0\}}$ and $K_3(u) = \frac{3}{2}(1-u^2)\mathbb{1}_{\{-1<u<0\}}$.