Nonparametric estimation and forecasting of time-varying parameter models

Yu Bai
Department of Econometrics and Business Statistics, Monash University

January 25, 2024

Abstract

This paper addresses issues of using local estimator in a forecasting model affected by parameter instability. We analyse the choices of kernel weighting function and bandwidth parameter associated with the local estimator. We prove the asymptotic optimality of the bandwidth selection procedure and provide an analytical criterion on the choice of kernel weighting function. The theoretical results are examined through an extensive Monte Carlo study and an empirical application on bond return predictability.

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

1 Introduction

Many important economic decisions are based on a forecasting model 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 choices of weighting functions and need to select the associated bandwidth parameter.

This paper aims to address issues of using local estimator in an out-of-sample forecasting context. First, 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. We propose to select the bandwidth parameter by simply minimizing the conditional expected loss at the end of the sample. This 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 criterion we use for bandwidth selection. We find that choice of kernel weighting functions is related to the bandwidth selection, which reflects the usual bias-variance trade-off. We show that, when estima-

tion variance from the local estimator dominates, the rescaled criterion converges in distribution to a random variable. In this case, there is a simple criterion on the optimal kernel weighting function, which does not depend on the properties of the time-varying parameters. In other cases, the estimation bias is present in the limit. The optimal kernel weighting function is related to the property of parameter instability, 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 selection performs better than the fixed rolling window forecast under various types of parameter instability. In general, using all but downweighting the data is preferred.

We present an empirical application on 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 local estimator with optimal bandwidth selection is particularly useful for 12-month overlapping returns. It generally performs better than rolling window forecasts with fixed window size. Choice of kernel weighing function does affect forecasting performance, but combining forecasts from local estimators with different kernel weighting functions further improves forecast accuracy.

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 discuss selection of the bandwidth parameter and implications on the choice of kernel weighting functions. Section 4 provides Monte Carlo study on the theoretical analyses. Section 5 presents an empirical application on bond return predictability, and Section 6 concludes. Some technical details and auxiliary lemmas are provided in the Appendix. Proofs of the lemmas and main theorems are provided in the supplementary material.

NOTATION: $\|\cdot\|$ is the Euclidean norm. $|\cdot|$ denotes the associated norm when $\cdot$ is one dimensional. $x_n = O_p(y_n)$ states that the vector of random variables x_n is at most of order y_n in probabil-

ity, 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.

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 $k \times 1$-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.

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, OLS estimator is obtained when $\ell_t(\cdot)$ is the mean squared error loss: $\ell_t(\theta) = (y_{t+h} - X_t'\theta)^2$.

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

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

where $\varepsilon_t \sim (0, 1)$. Then, the quasi-maximum likelihood estimation of $\theta = (\omega, \alpha, \beta)'$ is equivalent

to minimizing the in-sample QLIKE loss function (Oh and Patton, 2021):

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

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, we often need a specification for the time-varying parameters. Evolutions of parameters can be either discrete and abrupt, such as in Markov Switching and change-point models, or continuous and smooth, such as in random coefficient models in which parameters typically evolve as random walk processes[1].

To remain agnostic on the types of parameter time variation, we take a nonparametric approach and assume that time-varying parameters $\theta_t = (\theta_{1,t}, \theta_{2,t}, \cdots, \theta_{k,t})'$ $(t = 1, 2, \cdots, T)$ are modeled as the function of scaled time point:

$$\theta_{\ell,t} = \theta_\ell(t/T), \quad \ell = 1, 2, \cdots, k. \tag{2}$$

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.

Since the forecasts and their evaluations are based on $\theta(1)$[2], we consider a local estimator for $\theta(1)$ defined by

$$\hat{\theta}_{K,b,T} = \arg\min_{\theta \in \Theta} \frac{1}{Tb} \sum_{t=1}^{T} k_{tT} \ell_t(\theta), \tag{3}$$

where $k_{tT} = K((t-T)/(Tb))$, $K(\cdot)$ is a kernel function, and $b = b_T > 0$ is a bandwidth parameter

[1]For a comparison of different approaches to accounting for parameter instability in the context of macroeconomic forecasting models, see, for instance, Pettenuzzo and Timmermann (2017).

[2]As in Example 1, when the parameters are time-varying, the target y_{T+h} depends on $\theta(1+1/T)$, which is different from $\theta(1)$. However, under Assumption B1, the local time variation is asymptotically negligible: $\theta(1 + 1/T) \cong \theta(1)$.

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).

The theoretical analysis in the next section requires the asymptotic properties of the local estimator. While a formal technical discussion is left in the Appendix A, we highlight the main assumption imposed on the time-varying parameters θ_t. We assume that, for each element $\ell = 1, 2, \cdots, k$ in (2), the following

$$|\theta_\ell(t/T) - \theta_\ell(s/T)| \leqslant c_\ell \left(\frac{|t-s|}{T}\right)^\gamma, \quad t, s = 1, 2, \cdots, T, \tag{4}$$

holds for some $0 < \gamma \leqslant 1$ and c_ℓ is a positive bounded constant. This condition implies that time variation in θ_t is bounded and θ_t changes slowly over time. When γ is equal 1, $\theta_\ell(\cdot)$ becomes Lipschitz-continuous.

(4) is first introduced in Chen and Hong (2016) for nonparametric estimation of time-varying GARCH models. Studies such as Robinson (1989) and Cai (2007) assume that $\theta_\ell(\cdot)$ is twice continuous differentiable. This satisfies (4) when $\gamma = 1$. As explained in Chen and Hong (2016), this condition allows for a set of a finite number of points where $\theta_\ell(\cdot)$ is discontinuous on $(0, 1]$.

Implementation of the bandwidth selection approach in Section 3.1 also requires the use of local linear estimator. The local linear sample loss function is defined as

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

where $\theta^{(1)}(\cdot)$ denotes the first order derivative of $\theta(\cdot)$ and 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 $T \to \infty$. Let $\Theta^{(1)} = [-M, M]^k$ with some $M > 0$, the local linear estimator of $\theta(1)$, $\theta^{(1)}(1)$ is given

by

$$(\tilde{\theta}(1), \tilde{\theta}^{(1)}(1)) = \underset{(\theta,\theta^{(1)})\in\Theta\times\Theta^{(1)}}{\arg\min} \; L_T(\theta, \theta^{(1)}). \tag{6}$$

Remark 1. *We focus on the use of local estimator in* (3) *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.* (3) *would allow us to discuss whether alternative kernel weighting functions could be better than the flat weighting function. Second, consistency of the local linear estimator requires the support of* $\tilde{K}$ *to be compact (Assumption A4). This again rules out interesting cases such as the EWMA (exponential weighted moving average) forecast (*$K(u) \propto e^{-\frac{u^2}{2}}$*). In addition, consistency of the local estimator requires a weaker condition on* (4) *compared to the local linear estimator, since differentiability is not required (Assumption A1).*

3 Out-of-sample forecasting

To use the local estimator (3), a forecaster faces a concrete decision problem as she 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 $E_T(\ell_{T+h}(\hat{\theta}_{K,b,T})) = E_T\Big(L(y_{T+h}, \hat{y}_{T+h|t}(\hat{\theta}_{K,b,T}))\Big)$ at the end of the sample. Provided that $\ell_{T+h}(\cdot)$ is twice continuously differentiable w.r.t. θ, a second-order Taylor series expansion around the true $\theta(1)$ gives (ignoring the smaller order terms):

$$\ell_{T+h}(\hat{\theta}_{K,b,T}) \approx \ell_{T+h}(\theta(1)) + \frac{\partial \ell_{T+h}(\theta(1))}{\partial \theta'} (\hat{\theta}_{K,b,T} - \theta(1)) + \frac{1}{2}(\hat{\theta}_{K,b,T} - \theta(1))' \frac{\partial^2 \ell_{T+h}(\overline{\theta}(1))}{\partial\theta\partial\theta'} (\hat{\theta}_{K,b,T} - \theta(1)), \tag{7}$$

where $\overline{\theta}(1)$ lies between $\hat{\theta}_{K,b,T}$ and $\theta(1)$. Taking conditional expectations on both sides we then find

$$E_T(\ell_{T+h}(\hat{\theta}_{K,b,T})) \approx \underbrace{E_T(\ell_{T+h}(\theta(1)))}_{R_T^1} + \underbrace{E_T\Big(\frac{\partial \ell_{T+h}(\theta(1))}{\partial \theta'}\Big)(\hat{\theta}_{K,b,T} - \theta(1))}_{R_T^2}$$
$$+ \frac{1}{2}(\hat{\theta}_{K,b,T} - \theta(1))' E_T\Big(\frac{\partial^2 \ell_{T+h}(\overline{\theta}(1))}{\partial \theta \partial \theta'}\Big)(\hat{\theta}_{K,b,T} - \theta(1)). \quad (8)$$

We see that the population loss can be decomposed into three components. The component R_T^1 is related to the future risk, which has nothing to do with parameter estimation. If we further assume that[3]:

$$E_T\Big(\frac{\partial \ell_{T+h}(\theta(1))}{\partial \theta'}\Big) = 0, \quad (9)$$

the term R_T^2 also drops out. Define the third term in (8) as (ignoring the constant $1/2$)

$$R_T(K,b) = (\hat{\theta}_{K,b,T} - \theta(1))' E_T\Big(\frac{\partial^2 \ell_{T+h}(\overline{\theta}(1))}{\partial \theta \partial \theta'}\Big)(\hat{\theta}_{K,b,T} - \theta(1)). \quad (10)$$

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. Write $\hat{\theta}_{\overline{K},b,T} = \hat{\theta}_{b,T}$ and $\omega_T(\overline{\theta}(1)) = E_T\Big(\frac{\partial^2 \ell_{T+h}(\overline{\theta}(1))}{\partial \theta \partial \theta'}\Big)$. We consider to choose b by simply minimizing (10) over the choice set I_T:

$$\hat{b} := \underset{b \in I_T}{\arg\min}\ (\hat{\theta}_{b,T} - \theta(1))' \omega_T(\overline{\theta}(1))(\hat{\theta}_{b,T} - \theta(1)). \quad (11)$$

[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.

Notice that, the cardinality of the set $|I_T|$ must shrink to zero as $T \to \infty$, since the consistency of $\hat{\theta}_{b,T}$ requires $b \to 0$.

We first drive the rate of the optimal bandwidth parameter implied by (11), which is characterised in the following theorem:

Theorem 1. *Under Assumptions A1(i), A2, A3 and A4(i), the optimal bandwidth parameter* $\hat{b}$ *obtained by minimizing* (11) *is of order* $T^{-\frac{1}{2\gamma+1}}$ *in probability for some* $0 < \gamma \leqslant 1$.

Theorem 1 shows that, the optimal bandwidth parameter $\hat{b}$, or the optimal effective number of observations $\lfloor Tb \rfloor$, is inversely related to γ. When $\gamma = 1$, $\lfloor Tb \rfloor$ is of order $T^{2/3}$ in probability, which is the maximum allowable rate[4]. As γ decreases, the upper bound on the amount of changes from θ_t to θ_s implied from (2) also increases. This means that overall time variation in θ_t is likely to be more frequent. The effective number of observations $\lfloor Tb \rfloor$ should also be lower.

By the consistency of $\hat{\theta}_{b,T}$(Lemma B1), we have $\overline{\theta}(1) \xrightarrow{p} \theta(1)$. To make (11) feasible, we replace the unknown $\theta(1)$ with the local linear estimator $\tilde{\theta}(1)$ given in (6). This leads to a feasible selection criteria:

$$\hat{b} := \underset{b \in I_T}{\arg\min} \, (\hat{\theta}_{b,T} - \tilde{\theta}(1))' \, \omega_T(\tilde{\theta}(1)) \, (\hat{\theta}_{b,T} - \tilde{\theta}(1)). \tag{12}$$

The asymptotic optimality of the feasible selection procedure (12) is formally stated in the next theorem.

Theorem 2. *Under Assumptions A1-A5, choosing* $\hat{b}$ *by* (12) *is asymptotically optimal in the sense that*

$$(\hat{\theta}_{b,T} - \tilde{\theta}(1))' \, \omega_T(\tilde{\theta}(1)) \, (\hat{\theta}_{b,T} - \tilde{\theta}(1)) \asymp \inf_{b \in I_T} \, (\hat{\theta}_{b,T} - \theta(1))' \, \omega_T(\theta(1)) \, (\hat{\theta}_{b,T} - \theta(1))$$

where $\tilde{\theta}(1)$ *is the local linear estimator from* (6) *with bandwidth parameter* $\tilde{b}$.

Theorem 2 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 (3) and a general loss function

[4]Inoue et al. (2017) obtain the same results for the rolling window estimator under Assumption A1(ii) for $\theta(\cdot)$.

for forecast evaluation. The asymptotic optimality implies that $\hat{b}$ chosen from (12) yields the same forecasts obtained from the true optimal bandwidth parameter by minimizing the infeasible objective function in (11). 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 (3). The requirements for two bandwidth parameters involved (b and $\tilde{b}$) are rather intuitive. We could first let $T\tilde{b}^5 \to 0$ to let the smoothing bias of $\tilde{\theta}(1)$ vanish asymptotically. Then, the remaining conditions hold when b goes to zero at a faster rate than $\tilde{b}$.

Remark 2. *The condition for the time-varying parameters θ_t imposed on Theorem 2 is stronger than Theorem 1 since it requires that $\theta(\cdot)$ is twice continuously differentiable. However, this condition is not that restrictive as it covers particular the ones considered in Giraitis et al. (2014) and Dendramis et al. (2021), where* (3) *is used to estimate a path of the stochastic time-varying coefficients. To see this, suppose that $\theta_{\ell,t}$ is a realization of a bounded random walk process: $\frac{1}{\sqrt{T}}\xi_t$, where $\Delta\xi_t = v_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. Simple algebra gives $\theta(\frac{t}{T}) = \sqrt{\frac{t}{T}}\frac{1}{\sqrt{t}}\xi_t = \sqrt{\frac{t}{T}}C_t$, where $C_t = \frac{1}{\sqrt{t}}\xi_t = O_p(1)$. This implies that $\theta(\frac{t}{T}) \propto \sqrt{\frac{t}{T}}$, which is twice continuously differentiable.*

3.2 Implications on the choice of K

Suppose now forecaster has selected b based on (12), does the choice of K matter in terms of forecast accuracy (lower population loss)? We shall discuss this based on the limiting behavior of $R_T(K,b)$. Since $R_T(K,b)$ is related to $\hat{\theta}_{K,b,T} - \theta(1)$, we first need an expansion for $\hat{\theta}_{K,b,T} - \theta(1)$.

Let $L_T(\theta) = \frac{1}{Tb}\sum_{t=1}^{T} k_{tT}\ell_t(\theta)$. As $\hat{\theta}_{K,b,T}$ satisfies the first-order condition $\frac{\partial L_T(\hat{\theta}_{K,b,T})}{\partial\theta} = 0$, by mean-value theorem, we have

$$0 = \frac{\partial L_T(\hat{\theta}_{K,b,T})}{\partial\theta} = \frac{\partial L_T(\theta(1))}{\partial\theta} + \frac{\partial^2 L_T(\bar{\theta}(1))}{\partial\theta\partial\theta'}(\hat{\theta}_{K,b,T} - \theta(1)), \tag{13}$$

where $\overline{\theta}(1)$ lies between $\hat{\theta}_{K,b,T}$ and $\theta(1)$. By applying mean-value theorem on $\frac{\partial L_T(\theta(1))}{\partial \theta}$, we have

$$\frac{\partial L_T(\theta(1))}{\partial \theta} = \frac{1}{Tb}\sum_{t=1}^{T} k_{tT}\frac{\partial \ell_t(\theta(1))}{\partial \theta} = \frac{1}{Tb}\sum_{t=1}^{T} k_{tT}\frac{\partial \ell_t(\theta(t/T))}{\partial \theta} + \frac{1}{Tb}\sum_{t=1}^{T} k_{tT}\frac{\partial^2 \ell_t(\theta^*(t/T))}{\partial \theta \partial \theta'}(\theta(1) - \theta(t/T))$$
$$= S_{1,T} + B_{2,T}, \tag{14}$$

where $\theta^*(t/T)$ lies between $\theta(1)$ and $\theta(t/T)$ for each t. Let $H_{1,T} = \frac{1}{Tb}\sum_{t=1}^{T} k_{tT}\frac{\partial^2 \ell_t(\theta)}{\partial \theta \partial \theta'}$, by plugging (14) back to (13) and rearranging terms, we have

$$\hat{\theta}_{K,b,T} - \theta_1 = -H_{1,T}^{-1}(S_{1,T} + B_{2,T}). \tag{15}$$

Thus, $\hat{\theta}_{K,b,T} - \theta(1)$ can be decomposed into a variance term $H_{1,T}^{-1}S_{1,T}$ and a bias term $H_{1,T}^{-1}B_{2,T}$. The limiting behavior of $R_T(K,b)$ shall be determined by whether the variance term or the bias term dominates. This is formally stated in the next theorem.

Theorem 3. *Suppose that Assumptions A1(i), A2, A3 and A4(i) hold with $b \to 0$ and $Tb \to \infty$. Then, it holds that*

(i) If $T^{1/2}b^{1/2+\gamma} \to 0$, we have

$$Tb \cdot R_T(K,b) \xrightarrow{d} \phi_{0,K}\Sigma^{1/2}(1)Z'\omega_T(\theta(1))Z\Sigma^{1/2}(1),$$

where $\phi_{0,K} = \int_{\mathcal{B}} K^2(u)du$, $Z \sim \mathcal{N}(0, I_k)$ and $\Sigma(1)$ is defined as in Lemma B1;

(ii) If $T^{1/2}b^{1/2+\gamma} \to \infty$, we have

$$b^{-2\gamma} \cdot R_T(K,b) \xrightarrow{p} \mu_{\gamma,K}^2 C'\omega_T(\theta(1))C,$$

where $\mu_{\gamma,K} = \int_{\mathcal{B}} u^\gamma K(u)du$ and $C = (c_1, \cdots, c_k)'$ is a collection of Hölder constant given in Assumption A1(i);

(iii) If $T^{1/2}b^{1/2} \asymp b^{-\gamma}$*, we have*

$$Tb \cdot \Big(R_T(K,b) + b^{2\gamma}\mu_{\gamma,K}^2 C'\omega_T(\theta(1))C\Big) \xrightarrow{d} \phi_{0,K}\Sigma^{1/2}(1)Z'\omega_T(\theta(1))Z\Sigma^{1/2}(1),$$

where $\mu_{\gamma,K}$*,* C *and* $\phi_{0,K}$ *are defined as in (i) and (ii).*

The limiting behavior of $R_T(K,b)$ is closely related to the bandwidth parameter b and reflects the usual bias-variance trade-off. Consider first when $T^{1/2}b^{1/2+\gamma} \to 0$. In this case, the bias introduced by (3) vanishes asymptotically. The rescaled $R_T(K,b)$ converges in distribution to a random variable. When $\omega_T(\theta(1))$ is idempotent matrix, the asymptotic distribution has a more elegant expression, since $Z'\omega_T(\theta_1)Z \sim \chi^2$, where the degree of freedom is given by trace$\Big(\omega_T(\theta(1))\Big)$. As $\phi_{0,K}$ affects the scale of the limiting distribution, we clearly prefer a kernel weighting function which has smallest $\phi_{0,K}$.

If $T^{1/2}b^{1/2+\gamma} \to \infty$, estimation bias from (3) dominates and $R_T(K,b)$ converges to a non-stochastic term which is not 0. Since we do not know γ, b may be set improperly so we are in case (ii) described in Theorem 3. In this case, $\mu_{\gamma,K} = \int u^\gamma K(u)du$ plays a role so we clearly want to choose a kernel weighting function which has smallest $\mu_{\gamma,K}$. In the third case, $(Tb)^{1/2}$ and $b^{-\gamma}$ diverge at the same rate. Notice that, in this case, b is of order $T^{-\frac{1}{2\gamma+1}}$ in probability, which is the optimal one we obtain in Theorem 1. Both bias and variance are present, and thus, both $\phi_{0,K}$ and $\mu_{\gamma,K}$ play a role and we clearly prefer a kernel weighting function with smallest $\phi_{0,K}$ and $\mu_{\gamma,K}$.

Let us illustrate more on Theorem 3 by considering the following three candidate choices of $K(u)$:

$$K_1(u) = \mathbb{1}_{\{-1<u<0\}}, \quad K_2(u) = \frac{2}{\sqrt{2\pi}}\exp\big(-\frac{u^2}{2}\big)\mathbb{1}_{\{u<0\}}, \quad K_3(u) = \frac{3}{2}(1-u^2)\mathbb{1}_{\{-1<u<0\}}. \tag{16}$$

These three kernel weighting functions are commonly used in applied work. $K_1(u)$ leads to a rolling window estimator with window size $\lfloor Tb \rfloor$. $K_2(u)$ imposes an exponential-type down-

weighting 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).

Suppose that $\gamma = 1$ and we set $b = cT^{-1/(3+\delta)}$. If $\delta > 0$, we are in Case (i). By computing $\phi_{0,K}$, we have $\phi_{0,K_1} = 1$, $\phi_{0,K_2} \approx 0.5642$ and $\phi_{0,K_3} = 1.2$. In this case, there is a clear winner. $K_2(u)$ is preferred: all data should be used and downweighted. If $\delta < 0$, we are in Case (ii). By computing $\mu^2_{\gamma,K}$, we have $\mu^2_{1,K_1} = 0.25$, $\mu^2_{1,K_2} \approx 0.637$ and $\mu^2_{1,K_3} = 0.141$. Then, we may expect that, if the bias term dominates, $K_3(u)$ would be preferred: only recent data should be used and downweighted. If $\delta = 0$, we are in Case (iii). As both $\phi_{0,K}$ and $\mu_{\gamma,K}$ are present, there is no clear winner among $K_1(u)$, $K_2(u)$ and $K_3(u)$.

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):

$$\begin{bmatrix} y_{t+1} \\ x_{t+1} \end{bmatrix} = \begin{bmatrix} a_t & b_t \\ 0 & \rho_t \end{bmatrix} \begin{bmatrix} y_t \\ x_t \end{bmatrix} + \begin{bmatrix} \varepsilon^y_{t+1} \\ \varepsilon^x_{t+1} \end{bmatrix}, \tag{17}$$

where the error terms $(\varepsilon^y_{t+1}, \varepsilon^x_{t+1})'$ 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 consider the cases in Giraitis et al. (2014) and Dendramis et al. (2021). We first generate $v_{it} = (1-L)^{1-d}\epsilon_{it}$, where $\epsilon_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0, 0.1)$. Then, we generate ξ_{it} from the random walk model: $\Delta\xi_{it} = v_{it}$. Finally, we set $a_t = 0.9\frac{\xi_{1t}}{\max_{1\leqslant j\leqslant t}|\xi_{1j}|}$ and $b_t = \xi_{2t}/\sqrt{T}$. We consider $d = 0.51, 0.75, 1, 1.25, 1.49$ for DGPs 5-9, respectively. Notice that, as explained in Remark 2, DGPs 5-9 also satisfy both Assumption A1 (i) and (ii)[5].

4.2 Forecasting models, implementation and forecast evaluation

We consider the following predictive regression model:

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

where $X_t = (y_t, x_t)'$ and $\theta_t = (a_t, b_t)'$. 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):

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

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

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

section 3.2:

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

Of course, when $k_{tT} = 1$ for all t, we are back to the non-local LS estimates.

Under (18) and MSFE loss, $R_T(K, b)$ becomes

$$R_T(K, b) = (\hat{\theta}_{K,b,T} - \theta(1))' \left(X_T X_T'\right) (\hat{\theta}_{K,b,T} - \theta(1)). \tag{19}$$

The true parameters $\theta(1)$ in (19) are approximated by the local linear estimator, which are the first $k \times 1$ elements of the following:

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

where $Z_t = [X_t', X_t'(\frac{t-T}{T})]'$, $\tilde{k}_{tT} = \tilde{K}\Big(\frac{t-T}{T\tilde{b}}\Big)$ 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 $\tilde{\theta}(1)$ and $\tilde{b}$ is selected by the rule-of-thumb method: $\tilde{b} = 1.06T^{-1/5}$. For b, as all DGPs are in the case when $\gamma = 1$, 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 forecasting performance of y_{T+h} based on 5000 Monte Carlo simulations. Apart from the forecasts from local estimator 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 1 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, except for a few cases when $T = 150$, all local estimators have better forecasting performance compared to the benchmark, and gains generally increase with sample size. Second, rolling window forecasts with fixed window size are sometimes the best, but window size does matter. As sample size increases, using alternative weighting functions with optimal bandwidth selection performs better compared to the fixed rolling window forecast. Finally, in terms of choices of kernel weighting functions, $K_2(u)$ is more likely to be better than others, as we see that it is the best in 6 out of 9 DGPs when $T = 600$. Notice that, we set $b = cT^{-1/3}$ and the true γ is equal to 1. Although we are in Case (iii) in Theorem 3, the finite sample performance is more similar to Case (i). As $K_2(u)$ has the smallest $\phi_{0,K}$, it generally delivers the best results.

We now move on to the 12-step ahead forecasts. As DGPs are based on VAR(1), forecasts obtained from predictive regression (18) are the directly forecasts (Marcellino et al., 2006). These forecast errors are serially correlated, and thus, (9) does hot hold. We would like to examine if our bandwidth selection methods still work well and whether the implication on the choice of kernel weighting functions still holds. In all, $K_2(u)$ with optimal bandwidth selection is overall better than others, as it is again the best in 6 out of 9 DGPs when $T = 450$ and $T = 600$. However, in DGPs 5-7 when innovations used to generate paths of time-varying parameters have short memory (Surgailis et al., 2012), forecasts from local estimators are outperformed by the

benchmark.

5 Application to bond return predictability

5.1 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. The excess return is

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

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. Therefore, we consider predictors based on forward spreads (FB) as in Fama and Bliss (1987),

[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).

Table 1: 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	1.019	1.010	1.026	1.006	1.035	1.023	1.011	1.040	1.014	1.050
6	0.980	0.983	0.983	0.973	0.987	1.025	1.013	1.042	1.017	1.058
7	0.971	0.970	0.979	0.964	0.984	1.045	1.023	1.079	1.040	1.098
8	0.924	0.927	0.929	0.920	0.933	1.015	0.996	1.052	1.012	1.074
9	0.744	0.766	0.739	0.748	0.737	0.982	0.965	1.004	0.979	1.022
					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	1.012	1.003	1.009	0.996	1.013	1.021	1.015	1.021	1.009	1.025
6	0.976	0.976	0.976	0.968	0.977	1.040	1.027	1.039	1.019	1.046
7	0.965	0.962	0.965	0.957	0.965	1.054	1.029	1.048	1.023	1.060
8	0.902	0.898	0.901	0.896	0.902	1.048	1.022	1.045	1.019	1.063
9	0.605	0.613	0.606	0.619	0.603	0.980	0.957	0.983	0.963	0.997
					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	1.012	1.008	1.008	0.998	1.009	1.024	1.013	1.016	1.007	1.019
6	0.989	0.985	0.984	0.976	0.986	1.039	1.024	1.026	1.013	1.032
7	0.958	0.951	0.952	0.950	0.954	1.051	1.030	1.033	1.016	1.043
8	0.886	0.884	0.883	0.881	0.884	1.034	1.013	1.018	0.999	1.029
9	0.532	0.536	0.535	0.548	0.531	1.008	0.986	0.990	0.978	1.002
					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	1.010	1.004	1.003	0.996	1.004	1.025	1.018	1.016	1.007	1.019
6	0.985	0.974	0.973	0.968	0.973	1.032	1.022	1.021	1.009	1.024
7	0.942	0.936	0.936	0.933	0.936	1.049	1.030	1.025	1.011	1.032
8	0.894	0.889	0.888	0.888	0.887	1.040	1.013	1.007	0.993	1.019
9	0.488	0.490	0.492	0.503	0.488	1.006	0.974	0.973	0.964	0.984

Notes: 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 $K_3(u) = \frac{3}{2}(1-u^2)\mathbb{1}_{\{-1<u<0\}}$.

Cochrane-Piazzesi (CP) factor as in Cochrane and Piazzesi (2005), and a combination of both FB and CP factor.

The 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 from

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

We study monthly excess bond return predictability for the United States over the period 1961:06-2022:12. The yield data are taken from Liu and Wu (2021). Panel A of Table 2 presents summary statistics for both 12-month overlapping returns and one-month excess returns when $n = 2, 3, 4, 5$. A key difference between 12-month overlapping returns and one-month excess returns is that the former has much stronger persistence (higher serial correlation) because of the smoothing effect of using overlapping returns. Panel B of Table 2 presents summary statistics for the predictors based on FB and CP. We find that FBs and CP based on $m = 1$ have higher means and are more volatile (higher standard deviations), but have thinner tails (lower kurtoses) compared to those when $m = 1/12$. Both predictors are very persistent, even though the persistence

Table 2: Summary statistics

Panel A: Excess 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

Panel B: Predictors

	Fama-Bliss					Fama-Bliss				
	2 years	3 years	4 years	5 years	CP	2 years	3 years	4 years	5 years	CP
Mean	0.428	0.749	1.027	1.134	1.705	0.085	0.117	0.129	0.143	0.295
Std.dev.	0.632	0.954	1.178	1.338	1.534	0.092	0.108	0.121	0.131	0.246
Skew	-0.385	-0.368	-0.146	-0.014	0.385	-0.315	-0.332	-0.228	-0.172	-0.340
Kurt	3.723	3.883	3.457	2.973	5.344	4.356	4.423	3.131	3.027	5.572
AC(1)	0.926	0.940	0.946	0.955	0.861	0.911	0.903	0.925	0.934	0.843

Notes: This table reports summary statistics for monthly bond returns and the predictor variables used in our study. Panel A report the mean, standard deviation, skewness, kurtosis, and first-order autocorrelation (AC(1)) of bond excess returns for two-to five year bond maturities. 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. Panel B reports the same summary statistics for the predictors: the Fama-Bliss (FB) forward spreads (two, three, four, and five years), and Cochrane-Piazzesi (CP) factor. The left block is again based on 12-month holding period while the right block is based on one-month holding period.

level is slightly lower when $m = 1/12$.

5.2 Forecasting model, implementation and forecast evaluation

The forecasts are constructed based on the following predictive regression model:

$$rx_{t+12m}^{(n)} = \begin{bmatrix} \theta_{0,t} & \theta_{1,t}' \end{bmatrix} \begin{bmatrix} 1 \\ X_t \end{bmatrix} + \varepsilon_{t+12m}, \tag{20}$$

where X_t is a $k \times 1$ vector of predictors, $\theta_{1,t}$ is a $k \times 1$ vector of time-varying parameters and $\theta_{0,t}$ is the time-varying intercept term. Our analysis considers two predictor variables described in the previous subsection. Specifically, we consider two univariate models: FB ($X_t = FB_t^{(n,m)}$) and CP ($X_t = CP_t^m$), along with a multivariate model that includes both predictors ($X_t = [FB_t^{(n,m)}\ CP_t^m]'$)

for a total of three models.

We consider two different benchmark models. For 12-month overlapping returns, we consider benchmark forecasts from the yield curve: $X_t = [PC_{1t}\ PC_{2t}\ PC_{3t}]'$, where PC_{1t}, PC_{2t}, and PC_{3t} are the first three principal components of 1-month to 5-year bond yields. For one-month excess returns, benchmark forecasts are obtained from the EH (efficient hypothesis) model which assumes no predictability by letting $\theta_{1,t} = 0$ in (20) for all t. All benchmark forecasts are obtained from the non-local LS estimates.

As in the Monte-Carlo experiments, the forecasts are evaluated by the MSFE loss. $R_T(K, b)$ is given by 19. We consider forecasts obtained from non-local estimator, rolling window estimator with window size equal to 60 (5 years of observations) and local estimator using three different kernel weighting functions as in section 3.2 with bandwidth selected by letting $b = cT^{-1/3}$. The key difference from the Monte-Carlo experiments is that we do not know γ a priori in the empirical application. By assuming $\gamma = 1$, we may end up in Case (ii) in Theorem 3. The rescaled $R_T(K, b)$ converges to a non-stochastic term which is not 0. Thus, apart from the individual forecasts constructed from local estimators, we also consider forecast combinations. Let $\hat{f}_t^{K_1}$, $\hat{f}_t^{K_2}$, $\hat{f}_t^{K_3}$ be the forecasts from using three kernel weighting functions discussed in section 3.2 and $\omega_{i,t}$ be the combination weights. We consider the following four different combination schemes:

(1) Equal weighted (EW) combinations: $\omega_{i,t} = 1/3$;

(2) Discount mean square forecast error (DMSFE) combinations:

$$\omega_{i,t} = \frac{\phi_{i,t}^{-1}}{\sum_{j=1}^{3} \phi_{j,t}^{-1}},$$

where

$$\phi_{i,t} = \sum_{s=T_0}^{t-1} \rho^{t-1-s} (rx_{s+12m}^{(n)} - \hat{rx}_{s+12m|s}^{(n)})^2,$$

and $\rho = 0.9$ is a discounting factor;

(3) Least square (LS) combinations: $\omega_{i,t}$ are determined by the LS regression coefficients from

$$rx_{s+12m}^{(n)} = \omega_{0,s} + \sum_{j=1}^{3} \omega_{j,s} \hat{f}_s^{K_j} + e_{s+12m},$$

and the forecasts are given by $\hat{\omega}_{0,t} + \sum_{j=1}^{3} \hat{\omega}_{j,t} \hat{f}_t^{K_j}$;

(4) Least Absolute Deviation (LAD) combinations: same as (3), but the coefficients are estimated by Least Absolute Deviation (LAD).

The EW forecast combination is highly robust and widely used in economic forecasting (Stock and Watson, 2004). DMSFE forecast combinations are first introduced in Stock and Watson (2004) and have been found to work pretty well in equity premium prediction (Rapach et al., 2010). Granger and Ramanathan (1984) inspired more research effort in the direction of LS forecast combinations. The LAD forecast combinations are first introduced by Elliott and Timmermann (2004) to address several unappealing features from LS forecast combinations.

The initial estimation sample runs from 1961:06 to 1974:12 and the first available individual forecast is for 1975:01. We use 5-year training sample (60 observations) to estimate the initial forecast combination weights. Thus, the forecast evaluation period runs from 1980:01 to 2022:12. 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).

5.3 Empirical results

The forecasting results are summarized in Table 3. The upper panel (Panel A) considers 12-month overlapping returns, while the bottom panel (Panel B) is for one-month excess returns. For all entries, they are the ratios of MSFEs relative to the benchmark forecasts. Values below 1 indicate

that the corresponding specification performs better than the benchmark[7]. Entries shaded in gray indicate the best performing model.

Consider first the 12-month overlapping returns. Overall, the results are very promising, particularly for combination forecasts, as they deliver sizable and (sometimes) significant gains compared to the benchmark forecasts. Except for the 2-year excess bond returns, by combining FB and CP factor delivers the best forecasts, with either LS or LAD forecast combinations. For the individual forecasts, we find that models with both FB and CP factor generally perform better than the univariate specifications. The best choice of kernel weighting functions does depend on maturity and model specification. Forecasts from non-local estimator (OLS) and rolling window estimator with fixed window size are always outperformed by the benchmark when FB factor is used as a predictor.

The results are quite different when we move on to the one-month excess returns. The EH model turns out to be a tough benchmark, as there is only one case in which forecasts from non-local estimator perform slightly better for 2-year excess bond returns when FB factor is used. Using local estimator is not useful, and we do not see improvement from forecast combinations. The reason is likely due to the persistency mismatch between target variables and the predictors (Table 2). The serial correlation for one-month excess return is very low, but the levels for the predictors are still relatively high when $m = 1/12$.

6 Conclusion

In this paper, we analyse the choice of bandwidth parameter and kernel weighting function associated with the local estimator in an out-of-sample forecasting context. We first propose to select the bandwidth parameter by minimizing the 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 asymp-

[7] As the serial correlations of 12-month overlapping returns are high, EH model is unlikely to hold. We have also tried using EH model as the benchmark and found that the gains are larger compared to forecasts from EH model.

Table 3: Out-of-sample forecasting performance of bond returns

	Model	Individual forecasts					Forecast combinations			
		Panel A: 12-month overlapping excess returns								
		OLS	Fixed	Opt-K_1	Opt-K_2	Opt-K_3	EW	DMSFE	LS	LAD
2 years	FB	1.055	1.223	1.043	0.983	1.009	0.996	0.986	0.962	1.007
	CP	0.972	0.991	0.839	0.858	0.833	0.830	0.823	0.826	0.846
	FB+CP	0.977	0.941	0.876	0.852	0.846	0.840	0.832	0.829	0.855
3 years	FB	1.048	1.183	1.013	0.990	0.988	0.984	0.975	0.935	0.997
	CP	0.965	0.946	0.810*	0.831*	0.812	0.806*	0.800*	0.812	0.800
	FB+CP	0.967	0.882	0.839	0.823	0.782	0.794	0.784*	0.738*	0.737*
4 years	FB	1.039	1.128	1.021	0.988	0.979	0.983	0.972	0.913	0.938
	CP	0.977	0.910	0.801*	0.820*	0.799*	0.796*	0.790*	0.801	0.783*
	FB+CP	0.977	0.845	0.802*	0.808*	0.761*	0.770*	0.756*	0.741*	0.742*
5 years	FB	1.079	1.139	1.023	1.021	0.995	1.000	0.989	0.944	0.954
	CP	1.003	0.903	0.810*	0.829*	0.809*	0.805*	0.799*	0.815	0.800
	FB+CP	0.986	0.846	0.812	0.818*	0.776*	0.784*	0.773*	0.759*	0.756*
		Panel B: one-month excess returns								
		OLS	Fixed	Opt-K_1	Opt-K_2	Opt-K_3	EW	DMSFE	LS	LAD
2 years	FB	0.997	1.058	1.065	1.028	1.085	1.054	1.054	1.135	1.056
	CP	1.059	1.101	1.103	1.089	1.153	1.110	1.110	1.078	1.051
	FB+CP	1.071	1.132	1.115	1.080	1.134	1.101	1.102	1.260	1.041
3 years	FB	1.011	1.090	1.102	1.051	1.112	1.081	1.081	1.126	1.084
	CP	1.051	1.076	1.069	1.063	1.115	1.077	1.078	1.057	1.078
	FB+CP	1.051	1.083	1.066	1.042	1.078*	1.054	1.055	1.180	1.054
4 years	FB	0.993	1.052	1.060	1.022	1.074	1.046	1.045	1.096	1.113
	CP	1.039	1.052	1.042	1.039	1.079	1.048	1.048	1.055	1.128
	FB+CP	1.034	1.074*	1.044*	1.018	1.055*	1.031*	1.032*	1.129	1.119
5 years	FB	0.988	1.039	1.049	1.014	1.063	1.037	1.037	1.067	1.044
	CP	1.033	1.039	1.035	1.029	1.065	1.038	1.039	1.064	1.077
	FB+CP	1.027	1.070*	1.051*	1.025	1.072*	1.043*	1.043*	1.103	1.073

Notes: This table reports ratios of out-of-sample MSFEs for three prediction models based on the FB and CP predictors fitted to monthly bond excess returns, $rx^{(n)}_{t+12m}$, measured relative to the annualized T-bill rate. Panel A shows the results for overlapping returns ($m = 1$), while Panel B is based on the results for one-month excess returns ($m = 1/12$). For overlapping returns, the benchmark forecasts are based on the yield curve (measured by the first three principal components of 1-month to 5-year bond yields). For one-month excess returns, the benchmark forecasts are based on the EH (efficient hypothesis) model which assumes $\theta_1 = 0$ in (20). 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 $K_1(u) = \mathbb{1}_{\{-1<u<0\}}$, $K_2(u) = \frac{2}{\sqrt{2\pi}} \exp\big(-\frac{u^2}{2}\big)\mathbb{1}_{\{u<0\}}$ 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: LS combinations; LAD: least absolute deviation combinations. 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.

totic 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 function, which has been less addressed in the literature. Our analysis is based on the limiting behavior of the criterion we use to select the optimal bandwidth. We find that choice of kernel weighting function is related to the bandwidth selection, which reflects the usual bias-variance trade-off. When the estimation variance dominates, the criteria to select the optimal kernel weighting function is quite simple. In other cases, the criteria is more involved as it depends on the property of parameter time variation.

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 selection generally improves forecast accuracy under various form of parameter instability, compared to the rolling window forecast with fixed window size. In general, using all data and downweighting them is preferred.

We present an empirical application on bond return predictability. We find that our methods are particularly useful for 12-month overlapping returns. The optimal bandwidth selection procedure produces better forecasts compared to the rolling window forecasts with fixed window size. Choice of kernel weighting functions does matter, but combining forecasts from different kernel weighting functions further improves forecast accuracy.

One caveat of our analyses is that we require the loss function is smooth. In the supplementary material, we provide implementation details when the loss function is not smooth, using quantile predictive regression model as an example. However, this implies that loss function used for forecast evaluation is different from the one for the selection criterion. In addition, an additional bandwidth parameter is involved to smooth the loss function. We leave the detailed analysis in this case for future research.

References

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

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

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

Chen, B. and Y. Hong (2016). Detecting for smooth structural changes in garch models. *Econometric Theory 32*(3), 740–791.

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.

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

Dendramis, Y., L. Giraitis, and G. Kapetanios (2021). Estimation of time-varying covariance matrices for large datasets. *Econometric Theory 37*(6), 1100–1134.

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.

Elliott, G. and A. Timmermann (2004). Optimal forecast combinations under general loss functions and forecast error distributions. *Journal of Econometrics 122*(1), 47–79.

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.

Granger, C. W. and R. Ramanathan (1984). Improved methods of combining forecasts. *Journal of forecasting 3*(2), 197–204.

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.

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.

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.

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.

Surgailis, D., H. L. Koul, and L. Giraitis (2012). *Large sample inference for long memory processes*. World Scientific Publishing Company.

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

A The model, assumptions and discussions

We consider time series models of the following form

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

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: $\ell_{t,T}(\theta(t/T)) = \ell(Z_{t,T}; \theta(t/T))$.

Under certain regularity conditions on G and ε_t, it can be shown that[8], for each $u \in [0, 1]$, the stationary solution to the model (A1) exists and takes the following form:

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

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

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

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* $W_{t/T,t}(\theta)$ *for each rescaled time point* $t/T \in [0, 1]$, *such that for some* $0 < \rho < 1$ *and all* T,

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

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. $\theta_{t,T} = \theta(t/T) = \theta(u)$, $u = t/T$, $\theta(\cdot) : (0, 1] \longrightarrow \Theta$ *and* Θ *is compact. Let* $\theta_\ell(u)$ *(*$\ell = 1, 2, \cdots, k$*) be the* ℓ*th elements in* $\theta(u)$.

(i) For each point $u \in (0, 1]$, *there exists some* $0 < \gamma \leqslant 1$ *and* $c_\ell < \infty$, *such that*

$$|\theta_\ell(u) - \theta_\ell(v)| \leqslant c_\ell\,|u - v|^\gamma,$$

where $v \in N_\varepsilon(u)$, a small neighborhood containing u;

(ii) $\theta_\ell(\cdot)$ *is twice continuously differentiable on* $(0, 1]$.

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* $\ell_{u,t}(\theta)$ *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* $\ell^{(1)}_{u,t}(\theta) = \frac{\partial \ell_{u,t}(\theta)}{\partial \theta}$ *for each rescaled time point* $u \in (0, 1]$*;*

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

Assumption A3. *For each rescaled time point* $u \in (0, 1]$,

(i) $\ell_{u,t}(\theta)$ *is ergodic and* L_1*-continuous w.r.t* θ*;* $E[\ell_{u,t}(\theta)]$ *is uniquely minimized at* $\theta(u)$*;*

(ii) $\ell^{(1)}_{u,t}(\theta)$ *is ergodic and satisfies* $E(\ell^{(1)}_{u,t+h}(\theta)|\mathcal{F}_t) = 0$*, where* $\mathcal{F}_t = \sigma(y^*_s(u), X^*_s(u), s \leqslant t)$*.* $(y^*_s(u), X^*_s(u))$ *are the stationary solution of the model given in* (A2)*; central limit theorem (CLT) holds (as* $Tb \to \infty$*):*

$$\frac{1}{\sqrt{Tb}} \sum_{t=1}^{T} k_{tT} \frac{\partial \ell_{u,t}(\theta_u)}{\partial \theta'} \xrightarrow{d} \mathcal{N}(0, \phi_{0,K} \Lambda_u),$$

where $\phi_{0,K} = \int_C K^2(u)du$ *and* $\Lambda_u = Var\left(\frac{\partial \ell_{u,0}(\theta_u)}{\partial \theta'}\right)$*;*

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

Assumption A4. *Let* $K(\cdot)$ *and* $\tilde{K}(\cdot)$ *be the kernel weighting functions for* (3) *and* (6)*, respectively:*

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

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

Assumption A5. *The tuning parameters* b *and* $\tilde{b}$ *are such that: (i)* $T\tilde{b}^5 \to 0$*; (ii)* $b/\tilde{b} \to 0$*; (iii)* $T^{1/2}\tilde{b}^{1/2}b^{\gamma} \to \infty$ *for some* $0 < \gamma \leqslant 1$.

Assumption A1 impose conditions on the time-varying parameters. While (i) is more general than (ii) and is sufficient for the consistency of the local estimator, for the asymptotic optimality of the bandwidth parameter selection, we do require differentiability. However, as explained in Remark 2, this condition is not restrictive as the cases considered in Giraitis et al. (2014) are included.

Assumption A2 imposes conditions on the loss, its score and Hessian. We do not assume stationarity, but require the existence of stationary approximation for each scaled time point $u \in (0, 1]$. 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 each rescaled time point $u \in (0, 1]$. These conditions ensure that certain weak law of large numbers (WLLN) and CLT can be directly applied in the proof of Lemmas B1 and B2. 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. Alternative, we can follow 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.

B Auxiliary results

This section presents Lemmas needed for the proofs of Theorems 1 and 2. Proofs of all lemmas and theorems are provided in the supplementary material.

Lemma B1. *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: $\hat{\theta}_{K,b,T} \xrightarrow{p} \theta(1)$;

(ii) Consistency rate: for some $0 < \gamma \leqslant 1$, *we have*

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

(iii) CLT: if $T^{1/2}b^{1/2+\gamma} \to 0$, *we have*

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

where $\Sigma(1) = H^{-1}(1)\Lambda(1)H^{-1}(1)$, $\phi_{0,K} = \int_{\mathcal{B}} K^2(u)du$, $\Lambda(1) = Var\left(\frac{\partial \ell_{1,0}(\theta(1))}{\partial \theta'}\right)$ *and* $H(1) = E\left[\frac{\partial^2 \ell_{1,0}(\theta(1))}{\partial \theta \partial \theta'}\right]$.

Lemma B2. *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*

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

Lemma B3. *Suppose that Assumptions A1(i), A2, A3 and A4(i) hold with* $b \to 0$ *and* $Tb \to \infty$. *Then, for some* $0 < \delta < \frac{1}{2}$ *and* $0 < \gamma \leqslant 1$, *it holds that*

$$\sup_{b \in I_T} \left\| \hat{\theta}_{\overline{K},b,T} - \theta(1) \right\| = O_p(r_{T,b,\delta,\gamma}), \tag{B1}$$

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

Lemma B4. *Define*

$$L(b) = (\hat{\theta}_{b,T} - \theta(1))' \, \omega_T(\theta(1)) \, (\hat{\theta}_{b,T} - \theta(1)),$$

$$A(b) = (\hat{\theta}_{b,T} - \tilde{\theta}(1))' \, \omega_T(\tilde{\theta}(1)) \, (\hat{\theta}_{b,T} - \tilde{\theta}(1)),$$

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*

$$\sup_{b \in I_T} \left| \frac{L(b) - A(b)}{L(b)} \right| = o_p(1). \tag{B2}$$