Nonparametric forecasting with one-sided kernel adopting pseudo one-step ahead data☆

Jungwoo Kim[a,*], Joocheol Kim[a]

[a]*School of Economics, Yonsei University, South Korea*

Abstract

A new nonparametric forecasting using one-sided kernel is proposed via adopting pseudo one-step ahead data. Adopting pseudo one-step data is inspired from the difference between training error and test error, which motivates us to reduce test error minimization problem to training error minimization problem. The theoretical basis and the numerical justification of the new approach are presented.

Keywords: Nonparametric methods, Time series, One-sided kernel, Local regression, Exponential smoothing

1. Introduction

Since local constant regression or kernel regression presented by Nadaraya (1964) and Watson (1964) opened a new horizon of nonparametric statistics and local polynomial regression by Cleveland and Devlin (1988) widened its methodological view, it has shown huge development in the applied statistical method. Since it does not any restriction on a statistical model excluding a researcher's arbitrary perspective, it can lead to the more flexible approaches.

From the beginning of local regression, its main focus has been fitting a curve to a given data as usually seen in field of machine learning or signal processing. Thus, there are relatively fewer studies in the field of forecasting than those on the fitting data problem. However, there are some time-series studies adopting a nonparametic approach, not only focusing on forecasting, among which the study of Härdle et al. (1997) is circulated one summarizing over the wide range of statistical use of nonparametric approaches associated with time-series analysis.

Unlike the fitting or interpolation problem, in the forecasting problem, we can not get any data in the future. This gives a critical implication in terms of using an ordinary kernel, guiding us to use *one-sided* kernel. Because of the absence of the data in the future, it is natural to use one-sided kernel for forecasting problem which assigns the highest weight to the most recent data point like exponential smoothing presented by Brown (1959, p.195). Meanwhile, when using an ordinary or two-sided kernel, we can obtain a smoothing curve around a certain point by the data on both sides of a given point, thus an interpolation would work well. However, with one-sided kernel it is hard to avoid a wiggle fitted-curve due to the absence of the data on the right-hand side or in the future.

One of the previous studies explicitly dealing with forecasting problem purely based on one-sided kernel approach may be the work of Gijbels et al. (1999), whose contribution is also the investigation on the relation between exponential smoothing and kernel regression. Li and Heckman (2003) also developed the one-sided kernel approach for extrapolation in an arbitrarily small region with the simulation and empirical examination.

As for other uses of one-side kernel, Fan et al. (2003) applied various types of one-sided kernels in the fitting problem and Hansen (1995) used one-sided kernel when heteroscedasticity exists at the threshold of some interval of the domain.

Most distinct and important parameter in the nonparametric estimation is *bandwidth* parameter usually termed as h. Numerous studies across the many fields of study are carried out to find the bandwidth for accurately fitting. Of course, a setting a constant bandwidth over a given domain is the basic approach, and more recently time-adaptive bandwidth or variable bandwidth scheme are drawing attention; see Brockmann et al. (1993) and Ye et al. (2006).

For the case using one-sided kernel, selecting a bandwidth parameter to a given data is also critical subject. However, as mentioned, the absence of the future data should be an obstacle as the conventional method. The intuition leads us to consider some way to obtain the bandwidth optimal to the future. This will be clear from the later context.

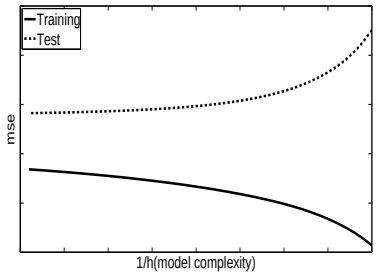
☆This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

*Corresponding author at: School of Economics, Yonsei University, South Korea

Email addresses: orignkik1@gmail.com (Jungwoo Kim), joocheol@yonsei.ac.kr (Joocheol Kim)

To assess a prediction performance of a statistical method, the prediction error is mainly used for its mathematical convenience. Also, there are two types of error, training error and test error. Training error $\mathrm{E}[(y_t - m(X_t))^2|\mathcal{T}]$($\mathcal{T}$ is a set of pair (x_t, y_t) for $t = 1, ..., T$) typically underestimates test error $\mathrm{E}[(y_{T+1} - m(X_{T+1}))^2|\mathcal{T}]$; see Friedman et al. (2001, p.38). Therefore, a model *biased* only to explain the current events from the data until the present(overfitting), it more easily fails to predict the future value. For instance, if we take a reciprocal of bandwidth as the model complexity as Hansen (2009, p.2), we can easily check this difference between training error and test error as the model complexity increases.

Figure 1: Training error vs test error



However, in many statistical studies, regardless of whether they are of nonparametric or of parametric statistics, the optimal parameter obtained from a data until the current time is applied to forecast the future value without concerning this limitation. This inevitable limitation is one of the essential reasons for the difference between training error and test error. Indeed, applying the bandwidth optimally derived from the data up to T to forecasting a value at $T+1$, it potentially implies $h_T = \mathrm{E}(h_{T+1}|\mathcal{F}_T)$, i.e., a martingale process, a quite strong condition to a given data-generating process.

However, if we can reduce test error minimization problem to training error minimization problem, we can expect to obtain less test error. Considering test error becomes training error after the realization of the next period we have,

$$\mathrm{E}[(y_{T+1} - m(X_{T+1}))^2|\mathcal{T}+1] \leq \mathrm{E}[(y_{T+1} - m(X_{T+1}))^2|\mathcal{T}] \tag{1.1}$$

Accordingly, how to reduce test error minimization problem to training error minimization problem must be the key.

As repeatedly mentioned, the inevitable limitation of a forecasting problem is due to an absence of the data in the future. This problem is similar to the *boundary effect* of kernel estimation. To come up with some solution to this problem, several novel methods presuming *pseudo data* was proposed(For these boundary corrections, see Karunamuni and Alberts (2005)).

The remainder of this paper is organized as follows. Section 2 contains the theoretical basis, the test error improvement of our new approach and its statistical features. Section 3 is devoted to examining our new approach by simulation data. Section 4 is for empirical examination. Section 5 summarizes and gives some future works.

2. Methodology

The data-generating model is

$$y_t = m(x_t) + e_t$$

where x_t is an equally time spaced predictor and y_t is a response variable, and e_t is an independent and identically distributed random variable with zero mean and unit variance and is independent of the x_t.

For the generalized estimator form, consider the local polynomial estimator of $m(x_t)$,

$$\hat{m}(x; h) = e_1^T (X_x^T W_x X_x)^{-1} X_x^T W_x Y \tag{2.1}$$

where,

$$X_x = \begin{pmatrix} 1 & x_1 - x & \cdots & (x_1 - x)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{T+1} - x & \cdots & (x_{T+1} - x)^p \end{pmatrix}, \qquad W_x = \underset{1 \leq t \leq T}{\mathrm{diag}} \{K\Big(\frac{x_t - x}{h}\Big)\}, \qquad Y = [y_1, ..., y_T]^T$$

and e_1 is the column vector of 1 in 1st position and 0's in other positions, $K(\cdot)$ is a kernel function and h is a bandwidth parameter

The estimator (2.1) was first introduced by Stone (1977), this type of formulation has an advantage as it suffices only to concern the intercept term of $\hat{m}$ at a given point x.

In this section, we introduce a new one-sided kernel forecasting approach adopting *pseudo one-step ahead data*, and verify its improvement in terms of test error and present statistical properties.

2.1. Test error improvement

In forecasting problem, a goal is to minimize test error or prediction error(PE),

$$\mathrm{PE}(\hat{m}) = \mathrm{E}[(y_{T+1} - \hat{m}(x_{T+1}; h))^2 | \mathcal{T}] \tag{2.2}$$

where $\mathcal{T}$ is a training set $\{(x_t, y_t) : t = 1, ..., T\}$.

For a mathematical convenience the expected prediction error(EPE) is more useful,

$$\mathrm{EPE}(m) = \mathrm{E}((y_{T+1} - \hat{m}(x_{T+1}; h))^2 \tag{2.3}$$

rewriting this conditioning on x_o,

$$= \mathrm{E}_x \mathrm{E}_{y|x}((y - \hat{m}(x; h))^2 | x_o) \tag{2.4}$$

In this study we deal with a discrete stochastic process, then (2.3) is approximated by the average squared residuals(ASR),

$$\mathrm{ASR}(h) = T^{-1} \sum_{i=1}^{T} (y_t - \hat{m}(x_t; h))^2 \tag{2.5}$$

Denote the minimizing parameter as $\hat{h}_o$,

$$\hat{h}_o = \underset{h \geq 0}{\mathrm{argmin}}(\mathrm{ASR}(h)) \tag{2.6}$$

Using $\hat{h}_o$, for example, we get a local constant estimated forecast $\hat{y}_{T+1}$,

$$\hat{y}_{T+1} = \sum_{t=1}^{T} K(\frac{x_t - x_{T+1}}{\hat{h}_o}) y_t \Big/ K(\frac{x_t - x_{T+1}}{\hat{h}_o}) = \hat{m}(x_{T+1}; h) \tag{2.7}$$

rewriting PE for the forecasting expression,

$$\mathrm{PE}(\hat{m}) = \mathrm{E}(y_{T+1} - \hat{y}_{T+1}) | \mathcal{T})^2 \tag{2.8}$$

This is the basic scheme of the ordinary forecasting method using one-sided kernel as shown in the previous studies as Gijbels et al. (1999) and Li and Heckman (2003). However, as mentioned earlier in the introduction section, if we reduce test error minimization problem to training error minimization problem we can expect the lower test error.

To do this completely, of course, we need to know the actual one-step ahead data in advance, but it is impossible. Instead, we present a *pseudo one-step ahead data* $\tilde{y}_{T+1}$ in substitution for y_{T+1} to obtain a *pseudo-training error* defined as below, from which the pseudo-optimal bandwidth for forecasting $\tilde{y}_{T+1}$ is obtained. It is natural that the closer $\tilde{y}_{T+1}$ is to the actual y_{T+1}, the more optimal bandwidth can be obtained, that is, the closer bandwidth to the bandwidth gained from minimizing the actual training error up to $T + 1$.

Accordingly, unless a given variable follows a martingale $\mathrm{E}(y_{T+1} | \mathcal{T}) = y_T$, it is better to approximate y_{T+1} by $\mathrm{E}(y_t | \mathcal{T})$ not by y_T.

Lemma 2.1. $\mathrm{E}[\|\mathrm{E}(y_{T+1}) - \mathrm{E}(y_t | \mathcal{T})]\| \leq \mathrm{E}[\|\mathrm{E}(y_{T+1}) - y_T\|]$

Proof. See appendix. □

To express training error from using pseudo one-step ahead data, define the pseudo-training error(PTE),

$$\mathrm{PTE}(\hat{m}) = \mathrm{E}(y_{T+1} - \hat{m}(x_{T+1})|\tilde{\mathcal{T}}_p)^2 \tag{2.9}$$

where $\tilde{\mathcal{T}}_p$ is a pseudo-training set, $\mathcal{T} \cup \{(x_{T+1},\, \tilde{y}_{T+1})) : \tilde{y}_{T+1} = g(\mathcal{T})\}$.

A real-valued function $g(\mathcal{T})$ approximates y_{T+1}, of course, how accurately $g(\mathcal{T})$ to well approximate y_{T+1} should determine how smaller test error we can obtain.

Also, a reasonable alternative for $g(\mathcal{T})$ may be a *neighboring mean*, $1\big/(T-s+1)\sum_{t=s}^{T} y_t$ rather than a simple mean $\mathrm{E}(y_T|\mathcal{T})$.

Then, remaining problem is how to select a range from s to T in order to well approximate y_{T+1}. Accordingly, we implement a new nonparametric parameter r denoting this *range*. Then,

$$\tilde{y}_{T+1} = g(\mathcal{T}) := r^{-1} \sum_{t=T-r+1}^{T} y_t \tag{2.10}$$

With the parameter r the pseudo-ASR(or PTE) including $\tilde{y}_{T+1}$ becomes,

$$\mathrm{ASR}(h, r) = \frac{1}{(T+1)}\{\sum_{t=1}^{T}(y_t - \hat{m}(x_t; h))^2 + (\tilde{y}_{T+1}(r) - \hat{m}(x_{T+1}; h))^2\} \tag{2.11}$$

getting the expected value of this,

$$\mathrm{MASR}(h, r) = \mathrm{E}(\mathrm{ASR}(h, r)) = \mathrm{E}(y_t - \hat{m}(x_t; h))^2 + \mathrm{E}(\tilde{y}_{T+1}(r) - \hat{m}(x_{T+1}; h))^2 \tag{2.12}$$

Denote the minimizing parameters as $\hat{h}_p$, $\hat{r}_p$

$$\hat{h}_p = \underset{h\geq 0}{\mathrm{argmin}}(\mathrm{MASR}(h, r)), \quad \hat{r}_p = \underset{r\geq 0}{\mathrm{argmin}}(\mathrm{MASR}(h, r)) \tag{2.13}$$

Then (2.8) can be converted into a sort of training error(termed as PTE above), since it uses a value up to $T+1$, $\tilde{y}_{T+1}$ being approximated to y_{T+1} in the pseudo manner. Rewriting (2.2) with this parameter r to compare with ordinal PE,

$$\mathrm{PE}(\hat{m}) = \begin{cases} \mathrm{E}(y_{T+1} - \hat{m}(X_{T+1}; h)|\mathcal{T})^2 & \text{for } y_{t=1,\dots T}\text{, (i.e., ordinal PE)} \\ \mathrm{E}(y_{T+1} - \hat{m}(X_{T+1}; h)|\tilde{\mathcal{T}}_p)^2 & \text{for } y_{t=1,\dots T} \text{ and } \tilde{y}_{T+1}(r) \end{cases} \tag{2.14}$$

Considering that MASR(h, r) has an extra parameter dimension relative to MASR(h) and $\mathcal{T} \subset \tilde{\mathcal{T}}_p$, we obtain a theorem as below

Theorem 2.1. *Let* $\mathrm{PE}_o = \mathrm{E}(y_{T+1} - \hat{m}(X_{T+1}; h)|\mathcal{T})^2$ *and* $\mathrm{PE}_p = \mathrm{E}(y_{T+1} - \hat{m}(X_{T+1}; h)|\tilde{\mathcal{T}}_p)^2$. *Then,* $\mathrm{PE}_p \leq \mathrm{PE}_o$.

Proof. See appendix. □

2.2. Statistical properties

For the generalization to the univariate and pth order polynomial regression case, let

$$X_x = \begin{pmatrix} 1 & x_1 - x & \cdots & (x_1 - x)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{T+1} - x & \cdots & (x_{T+1} - x)^p \end{pmatrix}, \tag{2.15}$$

, and use the higher order extension of kernel $K_{(p)}$ in accordance with the degree of a polynomial regression as suggested by Lejeune and Sarda (1992),

$$K_{(p)} = \{|M_p(u)|/|N_p|\}\, K(u),$$

where N_p is the $(p+1)\times(p+1)$ matrix whose (i, j) entry is $\int u^{i+j-2}K(u)\, du$ and M_p is the same as N_p except for the first column equal to $(1, u, ..., u^p)^T$. Then bias and variance are as follows.

Theorem 2.2. *Suppose that assumption 1 and assumption 2(in appendix) hold, that* $h \to 0, nh \to \infty$ *as* $n \to \infty$ *and* $r \leq T$. *Recalling (2.1), where* X_x *is given by (2.15). Then,*

$$\mathrm{E}\left\{\hat{m}(x;h) - m(x) | x_1, ..., x_{T+1}\right\} = \Big\{ \int u^{p+1} K_{(p)}(u)\, du \Big\} \Big\{ \frac{m^{p+1}(x)}{(p+1)!} \Big\} h^{p+1} + o(h^{p+1})$$

and

$$\mathrm{Var}\left\{\hat{m}(x;h) | x_1, ..., x_{T+1}\right\} = \Big\{ \int K_{(p)}(u)^2\, du \Big\} \{(Th+h)^{-1}(1+o(1))\} \tag{2.16}$$

Proof. See appendix. □

Note that our minimizing problem includes up to $\tilde{y}_{T+1}$ by implementing a new parameter r, thus, $T+1$ is in the expression of variance rather than T .

To obtain MSE, let

$$\mathrm{B}^2 = \Big\{ \int u^{p+1} K_{(p)}(u)\, du \Big\}^2 \Big\{ \frac{m^{p+1}(x)}{(p+1)!} \Big\}^2 \tag{2.17}$$

and

$$\mathrm{V} = \Big\{ \int K_{(p)}(u)^2\, du \Big\} \tag{2.18}$$

Since $\mathrm{MSE}(\hat{m}) = \mathrm{Bias}(\hat{\mathrm{m}})^2 + \mathrm{Var}(\hat{m})$, following corollary as Gijbels et al. (1999) is derived.

Corollary 2.1. *Under the assumptions and results of theorem 2.2,*

$$\mathrm{MSE(h)} = \mathrm{B}^2 h^{2p+2} + \mathrm{V}(Th+h)^{-1} + o(h^{2p+2} + (Th+h)^{-1}) \tag{2.19}$$

Recalling $\mathrm{PE} = \mathrm{MSE} + \sigma_e^2$ and minimizing (2.12) gives the optimal bandwidth and range,

Theorem 2.3. *Let* $\tilde{e}$ *be an error from* $\tilde{y}$ *and be independent to* $\hat{e}$ *from* $\hat{m}(x)$,

$$h_{MASR} = h_{MSE} = \mathrm{C}_{MSE}(T+1)^{-1/(2p+3)}\{1+o(1)\} \tag{2.20}$$

where $\mathrm{C}_{MSE} = \{V_p/(2p+2)N_p^2\}^{1/(2p+3)}$

$$r_{MASR} = (M(x_T) - M(x_{T-r}))m(x_{T-r})^{-1} + o(T^{-\alpha}), \qquad \alpha > 0 \tag{2.21}$$

where $M(x)$ *is the intergral of* $m(x)$.

Proof. See appendix. □

Remark 2.1. *The last term of (2.12) in the brace is,*

$$(\tilde{y}_{T+1} - \hat{m}(x_{T+1};h))^2 = (\tilde{y}_{T+1} - y_{T+1} + y_{T+1} - \hat{m}(x_{T+1};h,r))^2 \tag{2.22}$$

By $\tilde{e}$ *independent to* $\hat{e}$, *we have (2.20) and (2.21) up to* $T+1$. *Also, note that the rate of MSE and bandwidth* $(T+1)^{-1}$ *is faster than* $(T)^{-1}$.

3. Simulation

For simulation experiment, we use local constant regression and local linear regression for predicting a future value.

Simulation data sets of sample sizes 100 and 300 are generated from three types of functions, and the number of simulation is 100 for each sample size. In each simulation the last data point is forecasted, and our new forecasting approach using pseudo one-step ahead data(PSEUDO) is compared to the ordinal forecasting approach with one-sided kernel(exponential weighted moving average used in Gijbels et al. (1999), abbreviated as EWMA) which uses $\hat{h}_T$ from using data up to T to forecast the future value.

3.1. Local constant regression

Let $Y = (y_1, ..., y_T)$ be a stochastic process at equally spaced points $(x_1, ..., x_T)$. The goal is to forecast y_{T+1}. A local constant estimator is

$$\hat{y}_{T+1} = \sum_{t=1}^{T} K_c(\frac{x_t - x_{T+1}}{h}) y_t \Big/ K_c(\frac{x_t - x_{T+1}}{h}) \tag{3.1}$$

where K_c is one-sided kernel function, denoted for $c = e$ an exponentially weighted kernel and for $c = g$ a Gaussian kernel,

$$K_c(u) = \begin{cases} \exp(u)\mathbb{1}_{u\leq 0} & \text{for} \quad c = e \\ 1/\sqrt{2\pi}\exp(-u^2/2)\mathbb{1}_{u\leq 0} & \text{for} \quad c = g \end{cases} \tag{3.2}$$

Before we proceed to show simulation results, we note one preliminary fact. Though the case using actual $T+1$ value data(ACTUAL) should give the least test error by the optimal bandwidth to forecast $T+1$ value, test error of PSEUDO can be less than test error of the case using actual $T+1$ value. Since the optimal bandwidth is obtained from minimizing a *sum* of squared error in a given data, it does not always guarantee the minimum squared error of the very last forecast. Consider a function below for instance.

$$y_t = \sin(x_t/3) + (2/3)\log(x_t) + e_t$$

where $e_t \sim iid.N(0, 1)$.

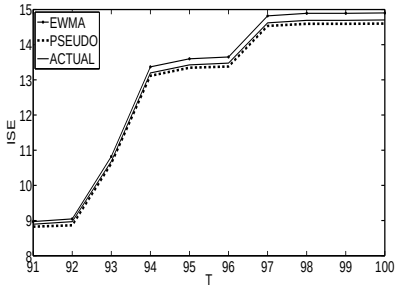
The result of the comparison of test error is shown as below. As seen in the table and figure, test error from PSEUDO approach is less than test error from ACTUAL. This is the encouraging result in the sense that without actual $T+1$ value we can expect a less test error by PSEUDO approach.

Table 1: Test error comparison

T	100		
Method	EWMA	PSEUDO	ACTUAL
Test error	1.202	1.173	1.181

$*$ Last 50's points forecasted

Figure 2: Test error less than the actual data



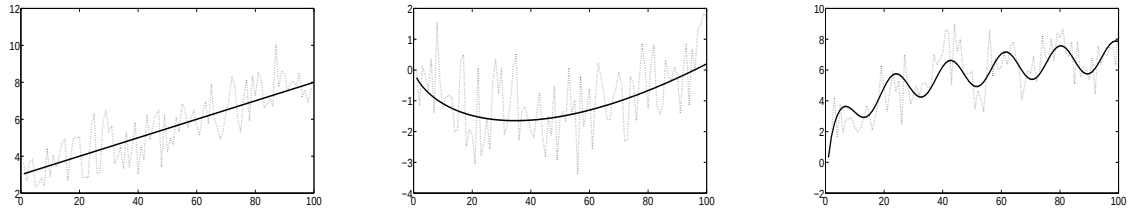
For simulation experiments, we examine three types of functions, each of which can represent a linear, polynomial and periodic function respectively.

$$\text{Model 1} : y_t = 3 + 0.05x_t + e_t$$
$$\text{Model 2} : y_t = 0.02(1.5x_t^{5/4} - 15x_t^{3/4} + x_t^{1/2}) + e_t$$
$$\text{Model 3} : y_t = \sin(x_t/3) + (2/3)\log(x_t) + e_t$$

where $e_t \sim iid.N(0,1)$.

Figure 3: Simulation function

(a) Model 1 (b) Model 2 (c) Model 3



From table 3 to table 8, the exponentially weighted kernel function is denoted as Exp. and h denotes the optimal bandwidth obtained from each method. Note that we present a standard deviation($S.D.$) of the bandwidth instead of a standard error which is simply obtained by dividing a standard deviation by the root of simulation times as Li and Heckman (2003).

As shown in the tables, test error of PSEUDO approach is less than test error of EWMA. Note that optimal bandwidths are greater than those from EWMA in many cases. It is explained by the fact that the more model complexity(here, $1/h$) increases the test error increases as in the figure 1. In other words, the bandwidth obtained from PSEUDO approach tends to utilize more observations, to be less biased against the future value, than EWMA. Differences between two kernels are not evident in terms of test error and the bandwidth.

Table 2: Model 1.

Kernel	Exp.				Gaussian			
T	100		300		100		300	
	EWMA	PSEUDO	EWMA	PSEUDO	EWMA	PSEUDO	EWMA	PSEUDO
Test error	1.461	1.438	1.321	1.301	1.237	1.209	1.141	1.114
h	4.568	4.619	4.557	4.609	5.655	5.753	5.607	5.686
$S.D.$	(0.602)	(0.642)	(0.530)	(0.561)	(0.829)	(0.831)	(0.817)	(0.827)

Table 3: Model 2.

Kernel	Exp.				Gaussian			
T	100		300		100		300	
	EWMA	PSEUDO	EWMA	PSEUDO	EWMA	PSEUDO	EWMA	PSEUDO
Test error	1.124	1.093	1.298	1.264	1.150	1.110	1.170	1.138
h	7.631	7.522	7.646	7.519	8.766	8.684	8.801	15.36
$S.D.$	(2.583)	(2.524)	(2.314)	(2.202)	(2.571)	(2.459)	(2.443)	(48.12)

Table 4: Model 3.

Kernel	Exp.				Gaussian			
T	100		300		100		300	
	EWMA	PSEUDO	EWMA	PSEUDO	EWMA	PSEUDO	EWMA	PSEUDO
Test error	1.018	0.978	1.335	1.286	1.182	1.142	1.350	1.305
h	1.904	1.919	1.897	1.928	2.094	2.128	2.150	2.238
$S.D.$	(0.522)	(0.496)	(0.519)	(0.512)	(0.373)	(0.392)	(0.391)	(0.709)

As a visualization of test error comparison, the integrated sum of squared error(ISE) in the below figure shows the

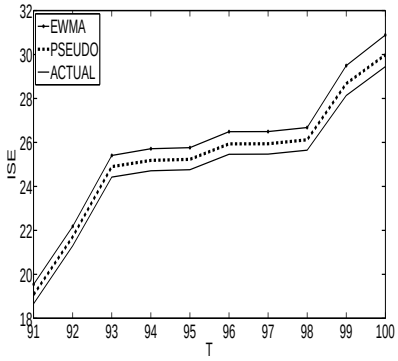
difference of test error between PSEUDO and EWMA. The test error from the case using the actual y_{T+1} is also included.

$$\text{ISE} = \int (y - \hat{m}(x))^2 \mathrm{d}x \approx \sum_{i=2}^{t} (y_i - \hat{m}(x_i))^2 \quad \text{for } t = 2, ..., T \tag{3.3}$$

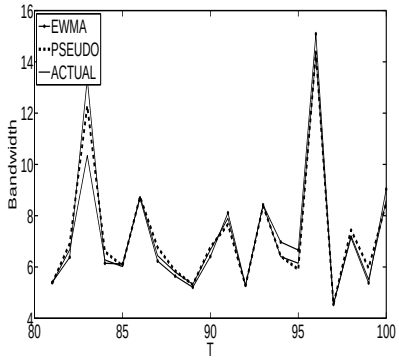
The ISE of PSEUDO is between EWMA and the case using actual data on the left figure. Meanwhile, this is not obvious in the bandwidth plot except that the bandwidth from PSEUDO is between EWMA and the case using actual data in several intervals. As for the range parameter, extreme values seem to be dominant, that is, the range of a whole sample or the range only referring some nearest points.

Figure 4: ISE, bandwidth comparison and range

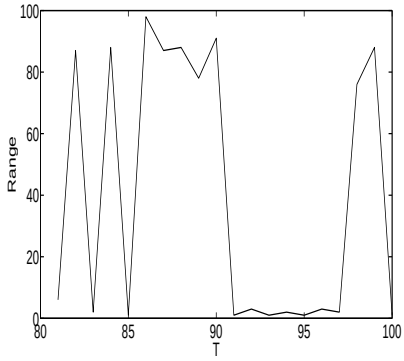
(a) ISE



(b) Bandwidth



(c) Range



3.2. Local linear regression

Applying (3.1) to a local linear estimator, the natural extension as in Gijbels et al. (1999) is,

$$\hat{y}_{t+1} = e_1^T \{X(x_{T+1})^T W_h(x) X(x_{T+1})\}^{-1} X(x_{T+1})^T W_h(x) Y_T \tag{3.4}$$

where

$$X(x) = \begin{pmatrix} 1 & x_1 - x \\ \vdots & \vdots \\ 1 & x_T - x \end{pmatrix}, \qquad W_h(x) = \underset{1 \le t \le T}{\text{diag}} \{K\Big(\frac{X_t - x}{h}\Big)\} \tag{3.5}$$

and e_1 is the column vector of 1 in 1st position and 0's in other positions and $K(\cdot)$ is a kernel function.

The below tables show that the results of PSEUDO approach outperform those from EWMA again. Likewise, the bandwidths are mainly greater in the case of PSEUDO approach than the bandwidths in the case of EWMA. Two types of the kernel do not show significantly different results as the previous local constant regression case. We can also find that the test error difference between EWMA and PSEUDO approach is mainly greater in the case of local linear regression than in the case of local constant regression, which may be relevant to the shortcoming known as *boundary effect* of local constant regression;see Hastie and Loader (1993).

Table 5: Model 1.

Kernel	Exp.				Gaussian			
T	100		300		100		300	
	EWMA	PSEUDO	EWMA	PSEUDO	EWMA	PSEUDO	EWMA	PSEUDO
Test error	0.917	0.901	1.024	0.999	0.975	0.955	0.935	0.885
h	695.2	679.0	671.4	677.5	682.9	704.0	642.5	646.6
$S.D.$	(432.3)	(427.9)	(432.9)	(434.0)	(448.1)	(435.8)	(463.8)	(454.0)

Table 6: Model 2.

Kernel	Exp.				Gaussian			
T	100		300		100		300	
	EWMA	PSEUDO	EWMA	PSEUDO	EWMA	PSEUDO	EWMA	PSEUDO
Test error	1.150	1.127	1.070	1.047	0.983	0.958	1.098	1.065
h	26.90	24.30	362.6	356.8	20.35	20.47	20.82	22.48
$S.D.$	(98.46)	(69.06)	(219.5)	(216.6)	(5.073)	(4.916)	(6.534)	(14.96)

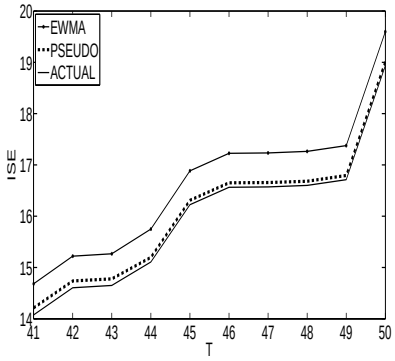
Table 7: Model 3.

Kernel	Exp.				Gaussian			
T	100		300		100		300	
	EWMA	PSEUDO	EWMA	PSEUDO	EWMA	PSEUDO	EWMA	PSEUDO
Test error	1.963	1.841	1.598	1.479	1.237	1.209	1.141	1.114
h	7.174	7.721	6.558	7.178	13.34	21.71	16.22	24.91
$S.D.$	(8.855)	(10.16)	(6.515)	(7.039)	(10.47)	(5.745)	(32.16)	(28.52)

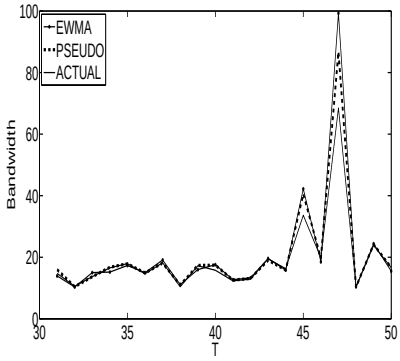
The figures below describe ISE, bandwidths and range comparison in the local linear case. The result on ISE is more distinct than the previous result from the local constant case, which supports that local polynomial regression usually outperforms local constant regression as claimed by Fan and Gijbels (1996).

Figure 5: ISE, bandwidth comparison and range

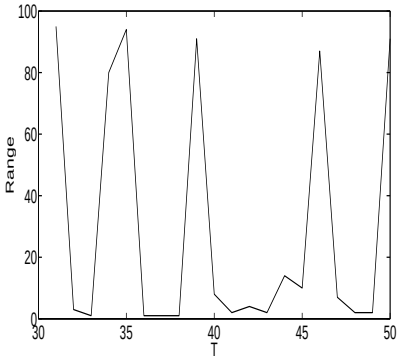
(a) ISE



(b) Bandwidth



(c) Range



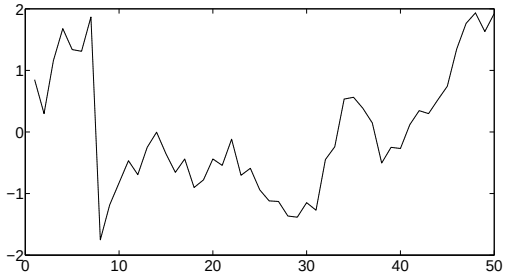
4. Empirical result

For the empirical study, we examine three types of data, exchange rate, bond price and commodity future price to embrace different types of economic data, and each data is standardized.

4.1. Exchange rate(EUR/USD)

Firstly, exchange rate(EUR/USD) of 50 days(from Jun.15.2016 to Aug.31.2016) is used.

Figure 6: Exchange rate(EUR/USD)



Total 20 exchange rates(from 31st to 50th) are forecasted and used to obtain test error. The test error comparison is presented in the table including the result of the case using an actual one-step ahead value(ACTUAL). As shown on the left of the table, we could see that test error of ACTUAL is smallest, the second smallest PSEUDO, being consistent with the results from simulations. However, on the right side of the table shows that test error of PSEUDO can be smallest as mentioned earlier.

Table 8: Test error

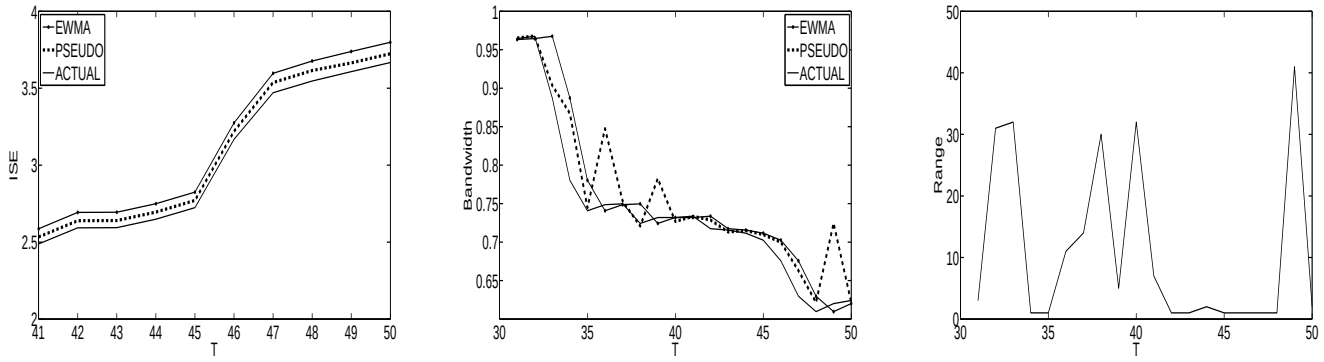
	Local constant			Local linear		
Method	EWMA	PSEUDO	ACTUAL*	EWMA	PSEUDO	ACTUAL*
Test error	0.190	0.186	0.183	0.222	0.204	0.205

*ACTUAL: Using an actual data.

The below figures describe ISE, bandwidth, and range across the three methods, and they come from the local constant case, The plot of ISE is consistent with the simulation result. Also, the plot of bandwidths shows that the bandwidths from PSEUDO approach are mainly greater than other bandwidths as before. Note that the plot of EWMA is lagged behind ACTUAL exactly by one step. The range plot shows that the extreme values are dominant across the interval, implying pseudo one-step data referring points in the whole sample, or nearest points. Also, it shows a slightly increasing trend as the number of data increases.

Figure 7: ISE, bandwidth and range

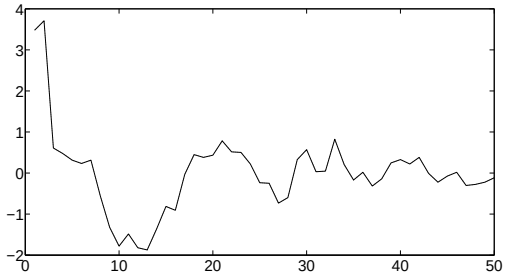
(a) ISE (b) Bandwidth (c) Range



4.2. U.S. 30 Year Treasury Bond

Secondly, U.S. 30 Year Treasury Bond price of 50 days(from Jun.15.2016 to Aug.31.2016) is used as empirical data. As seen in the below figure, the price of the bond shows a bit less wiggle shape than the case of the exchange rate.

Figure 8: U.S. 30 Year Treasury Bond



Total 20 bond prices(from 31st to 50th) are forecasted and used for test error. Unlike the previous result, the results of test error across three methods are almost same in the local constant case, which may be relevant to the less wiggle feature of this data.(we could examine by a simulation that the noise-free model gives the same result in the local constant case). In the local linear case, test error from PSEUDO is smallest as before, of course, this is not a general case.

Table 9: Test error

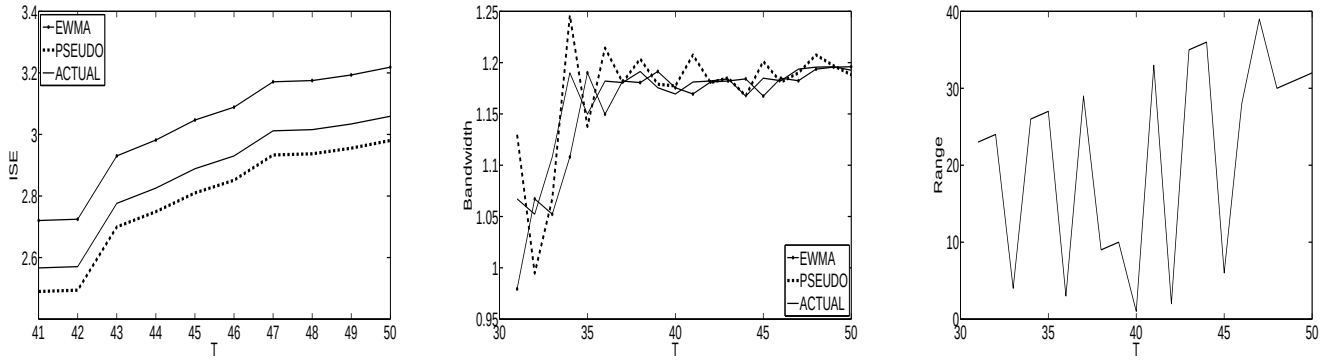
	Local constant			Local linear		
Method	EWMA	PSEUDO	ACTUAL*	EWMA	PSEUDO	ACTUAL*
Test error	0.107	0.107	0.107	0.161	0.149	0.153

*ACTUAL: Using an actual data.

The below figure describes ISE, bandwidth, and range. The figures come from the local linear case, ISE is smallest in PSEUDO. As for the result of bandwidth, the bandwidth from PSEUDO is more variable and greater than others.

Figure 9: ISE, bandwidth and range

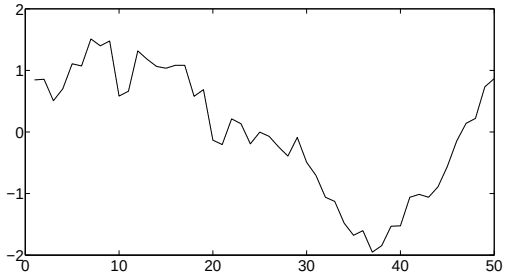
(a) ISE (b) Bandwidth (c) Range



4.3. Crude Oil Futures

Lastly, crude oil future price of 50 days(from Jun.15.2016 to Aug.31.2016) is used. The forecasting interval is from 31st to 50th. The data shows not as much wiggle as the previous examples, however including a sharp decrease just before the 40th data.

Figure 10: Crude Oil Futures



The test error comparison is presented in the table. In both local constant and local linear cases, the results are similar as previous ones.

Table 10: Test error

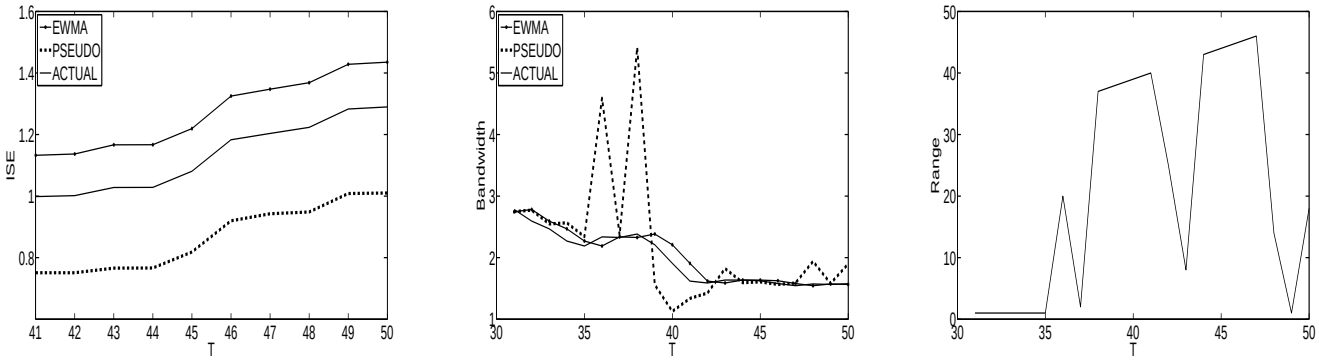
	Local constant			Local linear		
Method	EWMA	PSEUDO	ACTUAL*	EWMA	PSEUDO	ACTUAL*
Test error	0.083	0.081	0.080	0.072	0.051	0.065

*ACTUAL: Using an actual data.

The figures come from the local linear case, ISE is smallest in PSEUDO. Bandwidth shows a more variable shape than the others, and the range plot are similar as before.

Figure 11: ISE, bandwidth and range

(a) ISE (b) Bandwidth (c) Range



5. Concluding remarks

Motivated by the difference between training error and test error, our study presents a new nonparametric forecasting approach adopting pseudo one-step ahead data in the minimization of sum of squared error. Its theoretical basis and the justification from simulation data and empirical data are provided. From the simulation and empirical results, using local polynomial regression seems to give a better forecast than using local constant regression.

Future works are as follows. Considering the correlated environment as Gijbels et al. (1999) and Hart (1991) in which the parameter r may be differently obtained from the effect of autoregressive error terms and the heteroscedastic environment as Muller and Stadtmuller (1987) and Ruppert and Wand (1994), to expand our new method's applicability to these environments should be followed.

Secondly, expanding the domain space from the discrete space to the continuous space can be worthy of work, similar to Li and Heckman (2003). Most of the cases, time-series data is a discrete stochastic process, however, if this work would be made it can be also applicable to the spatial data analysis based on Kolmogorov existence theorem.

Lastly, it will be an interesting try to apply our 'pseudo data' scheme to parametric statistical analysis. That is, when forecasting a future value based on a parametric model, an optimization including pseudo one-step ahead data may be able to give a better performance.

References

Brockmann, M., Gasser, T., & Herrmann, E. (1993). Locally adaptive bandwidth choice for kernel regression estimators. *Journal of the American Statistical Association*, *88*(424), 1302–1309.

Brown, R. G. (1959). *Statistical Forecasting for Inventory Control*. McGraw-Hill.

Cleveland, W. S. & Devlin, S. J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, *83*(403), 596–610.

Fan, J. & Gijbels, I. (1996). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, volume 66. CRC Press.

Fan, J., Jiang, J., Zhang, C., & Zhou, Z. (2003). Time-dependent diffusion models for term structure dynamics. *Statistica Sinica*, 965–992.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin.

Gijbels, I., Pope, A., & Wand, M. P. (1999). Understanding exponential smoothing via kernel regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *61*(1), 39–50.

Hansen, B. E. (1995). Regression with nonstationary volatility. *Econometrica: Journal of the Econometric Society*, 1113–1132.

Hansen, B. E. (2009). Lecture notes on nonparametrics. *Lecture notes*.

Härdle, W., Lütkepohl, H., & Chen, R. (1997). A review of nonparametric time series analysis. *International Statistical Review*, *65*(1), 49–72.

Hart, J. D. (1991). Kernel regression estimation with time series errors. *Journal of the Royal Statistical Society. Series B (Methodological)*, 173–187.

Hastie, T. & Loader, C. (1993). Local regression: Automatic kernel carpentry. *Statistical Science*, 120–129.

Karunamuni, R. J. & Alberts, T. (2005). On boundary correction in kernel density estimation. *Statistical Methodology*, *2*(3), 191–212.

Lejeune, M. & Sarda, P. (1992). Smooth estimators of distribution and density functions. *Computational Statistics & Data Analysis*, *14*(4), 457–471.

Li, X. & Heckman, N. E. (2003). Local linear extrapolation. *Journal of Nonparametric Statistics*, *15*(4-5), 565–578.

Muller, H.-G. & Stadtmuller, U. (1987). Estimation of heteroscedasticity in regression analysis. *The annals of statistics*, 610–625.

Nadaraya, E. A. (1964). On Estimating Regression. *Theory of Probability & Its Applications*, *9*(1), 141–142.

Ruppert, D. & Wand, M. P. (1994). Multivariate locally weighted least squares regression. *The annals of statistics*, 1346–1370.

Stone, C. J. (1977). Consistent nonparametric regression. *The annals of statistics*, 595–620.

Watson, G. S. (1964). Smooth regression analysis. *Sankhy: The Indian Journal of Statistics, Series A*, 359–372.

Ye, A., Hyndman, R. J., & Li, Z. (2006). Local linear multivariate regression with variable bandwidth in the presence of heteroscedasticity. Technical report.

A. Appendix

A.1. Proof of Lemma 2.1.

PROOF. Rewrite each term as a squared form. Then, RHS is

$$\mathrm{E}[m(x_{T+1})^2 + \mu_T^2 - 2m(x_{T+1})\mu_T)]$$

and LHS is

$$\mathrm{E}[m(x_{T+1})^2 + m(x_T)^2 + e_T^2 - 2m(x_{T+1})m(x_T) - 2m(x_{T+1})e_T + 2m(x_T)e_T]$$

cancelling common terms

$$0 \leq \sigma^2$$

A.2. Proof of Theorem 2.1.

PROOF. Denoting filtration up to y_T by $\mathcal{F}_T$ and up to $\tilde{y}_{T+1}$ by $\mathcal{F}_p$. Then,

$$\mathcal{F}_T \subseteq \mathcal{F}_p \text{ and } \mathrm{PE}_o \text{ is measurable-}\mathcal{F}_p.$$

It is also obvious that

$$\inf_{h\geq 0, r\geq 0}\{\mathrm{PE}_p\} \leq \inf_{h\geq 0}\{\mathrm{PE}_p\}.$$

and recalling that training error is less than test error

$$\inf_{h\geq 0}\{\mathrm{PE}_p\} \leq \inf_{h\geq 0}\{\mathrm{PE}_o\}$$

A.2. Proof of Theorem 2.2.

PROOF. The assumption and proof come from the theorem 4.1 of Ruppert and Wand (1994). Here, we sketch the proof with several notices.

Assumption 1. The kernel function K is compactly supported, and its second moment $\int u^T u K(u)\, du = \mu_2 \neq 0$ and all odd-order moments of K vanish.

Assumption 2. The point x is in $\mathrm{supp}(f)$, $f(x)$ is density of x. At x, $f(x)$ is continuously differentiable and all fourth-order derivatives of $m(x)$ are continuous.

Using (2.1) and expanding Y to $(p+2)$th order of Taylor series gives

$$\mathrm{E}\{\hat{m}(x;h) - m(x)|x_1, ..., x_{T+1}\} = e_1^T(n^{-1}X_x^T W_x X_x)^{-1}(S_x + R_x). \tag{A.1}$$

where

$$S_x = n^{-1}X_x^T W_x \times \left\{ \frac{m^{p+1}(x)}{(p+1)!} \begin{bmatrix} (X_1 - x)^{p+1} \\ \vdots \\ (X_{T+1} - x)^{p+1} \end{bmatrix} + \frac{m^{p+2}(x)}{(p+2)!} \begin{bmatrix} (X_1 - x)^{p+2} \\ \vdots \\ (X_{T+1} - x)^{p+2} \end{bmatrix} \right\} \tag{A.2}$$

which is $(p+2)$th order Taylor expansion, and R_x is a vector of Taylor remainder terms. Let $A = \mathrm{diag}(1, h, ..., h^p)$, Q_p be the $(p+1)\times(p+1)$ matrix whose (i,j) entry is μ_{i+j-1} and $\mathbf{1}$ is a matrix whose every entry is one. Then, we obtain

$$n^{-1}X_x^T W_x X_x = A\{f(x)N_p + hf'(x)Q_p\}A + o(hA\mathbf{1}A) \tag{A.3}$$

The second term in the curly bracket, which comes from the Taylor expansion of $n^{-1}X_x^T W_x X_x$ is included for the case when p is even so that the some elements of the first term vanish.

And for $k = 1, ...,$ according to standard results from kernel density estimation being used for (A.2),

$$A^{-1}n^{-1}X_x^T W_x \begin{bmatrix} (X_1 - x)^k \\ \vdots \\ (X_{T+1} - x)^k \end{bmatrix} = \left\{ h^k f(x) \begin{bmatrix} \mu_k \\ \mu_{k+1} \\ \vdots \\ \mu_{k+p} \end{bmatrix} + h^{k+1} f'(x) \begin{bmatrix} \mu_{k+1} \\ \mu_{k+2} \\ \vdots \\ \mu_{k+p+1} \end{bmatrix} + o(h^{k+1}) \right\} \tag{A.4}$$

Substituting (A.2) with (A.4) and (A.3) into (1), and after rearranging terms considering that some moments in N_p and Q_p vanish, we have

$$E\{\hat{m}(x;h) - m(x)|x_1, ..., x_{T+1}\} = \left\{ \sum_{j=1}^{p+1} (N_p^{-1})_{1p} \mu_{p+2+j} \right\} \left\{ \frac{m^{p+1}(x) f'(x)}{f(x)(p+1)!} + \frac{m^{p+2}(x)}{(p+2)!} \right\} h^{p+2} + o(h^{p+2}) \tag{A.5}$$

In particular, since odd-order kernel moments do not vanish in the one-sided kernel we can omit the last term in the second curly bracket and it suffices to consider just $p + 1$ instead $p + 2$ as Gijbels et al. (1999).

For the conditional variance in the theorem, first note that

$$\mathrm{Var}\{\hat{m}(x;h)|x_1, ..., x_{T+1}\} = e_1^T (X_x^T W_x X_x)^{-1} X_x^T W_x V W_x X_x (X_x^T W_x X_x)^{-1} e_1 \tag{A.6}$$

where $V = \sigma_e^2 I_{T+1}$.

Let T_p be the $(p + 1) \times (p + 1)$ matrix having its (i, j) entry equal to $\int u^{i+j-2} K(u)\, du$. Then,

$$n^{-1} X_x^T W_x^2 X_x = h^{-1} A f(x) T_p A + o(h^{-1} A \mathbf{1} A). \tag{A.7}$$

Combining (A.3) and (A.7) into (A.6) we have

$$\mathrm{Var}\{\hat{m}(x;h)|x_1, ..., x_{T+1}\} = (e_1^T N_p^{-1} T_p N_p^{-1} e_1)(Th + h)^{-1}(1 + o(1)) \tag{A.8}$$

The first factor on the right hand side of (A.5) and (A.8) is termed as the *kernel dependent constants* and proven as equivalent to $\int u^{p+2} K_{(p)}(u)\, du$ and $\int K_{(p)}(u)^2\, du$ respectively in Ruppert and Wand (1994).

Since our model assumes homoscedasticity like the model of Li and Heckman (2003), the heteroscedastic function in Ruppert and Wand (1994) does not show in the variance expression.

A.2. Proof of Theorem 2.3.

PROOF. Assuming $\tilde{e}$ independent to $\hat{e}$ as in Remark 2.1., obtaining the optimal bandwidth is straightforward.

To obtain optimal r, let x_T be a continuous variable denoted as t for a mathematical convenience, then the last term of (2.12) becomes

$$\mathrm{E}[r^{-1} \int_{T-r}^{T} y_t\, dt - \hat{m}_{T+\delta}]^2 = \mathrm{E}[r^{-1} \int_{T-r}^{T} m_t + e_t\, dt - \hat{m}_{T+\delta}]^2, \qquad \delta > 0 \tag{A.9}$$

Derivative with respect to r is

$$2 \int (r^{-1} \int_{T-r}^{T} m_t + e_t\, dt - \hat{m}_{T+\delta}) \frac{\partial}{\partial r} (r^{-1} \int_{T-r}^{T} m_t + e_t\, dt - \hat{m}_{T+\delta}) f(e)\, de \tag{A.10}$$

Suppose that m_t and e_t are integrable as M_t and E_t with respect to t,

$$2 \int (r^{-1} \int_{T-r}^{T} m_t + e_t\, dt - \hat{m}_{T+\delta}) \frac{\partial}{\partial r} \{r^{-1}(M_T - M_{T-r} + E_T - E_{T-r})\} f(e)\, de \tag{A.11}$$

Rearranging and collecting like terms gives,

$$2\int M^{(1)}_{T-r} + E^{(1)}_{T-r} - r^{-1}(M_T - M_{T-r} + E_T - E_{T-r})f(e)\,de \tag{A.12}$$

$$=2\int M^{(1)}_{T-r} - r^{-1}(M_T - M_{T-r})f(e)\,de + o(T^{-\alpha}), \qquad \alpha > 0 \tag{A.13}$$

$$=m_{T-r} - r^{-1}(M_T - M_{T-r}) + o(T^{-\alpha}) = 0 \tag{A.14}$$