Statistica Sinica Preprint No: SS-2015-0365.R2	
Title	On the Efficiency of Online Approach to Nonparametric Smoothing of Big Data
Manuscript ID	SS-2015-0365.R2
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202015.0365
Complete List of Authors	Efang Kong and Yingcun Xia
Corresponding Author	Yingcun Xia
E-mail	staxyc@nus.edu.sg

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0365

On the Efficiency of Online Approach to Nonparametric Smoothing of Big Data

Efang Kong[1] and Yingcun Xia[2,1]
[1]University of Electronic Science and Technology, China
[2]National University of Singapore, Singapore

Abstract: The online updating approach (ONLINE) has been commonly used for the analysis of big data and online transient data. We consider in this paper how to improve its efficiency for various ONLINE kernel-based nonparametric estimators. Our findings include: (i) the optimal choice concerning the bandwidth and how it differs from that for the classical estimators; (ii) the optimal choice among a general class of sequential updating schemes; (iii) that the relative efficiencies of ONLINE Parzen-Rosenblatt density estimation or Nadaraya-Waston (N-W) regression estimation change with the dimension p of covariate in a nonlinear manner, and (iv) that while the classical local-linear fitting renders the estimators design-adaptive, their ONLINE counterparts still depend on the design of covariates in its leading terms of bias, they are still preferred over the ONLINE N-W estimators.

Keywords: Big data; kernel density estimation; N-W estimation; online updating estimation; varying coefficient model.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0365

1 Introduction

The concept of 'big data' has become a dominating topic in nearly all academic disciplines as well as in applied fields. In a broad sense, big data is characterized by its massive scale of volume, complexity, variety, velocity, variability, and veracity (Hilbert (2015)), and it is for this reason that most classical statistical methods are not suited for analysis of big data. Various new procedures have been proposed, which include, among others, subsampling-based approaches (Kleiner et al. (2014); Liang et al. (2013); and Ma et al.(2013)), the divide-and-conquer approach (Lin and Xi (2011); Chen and Xie (2014)), and the sequential updating approach (Wang et al. (2015), and the references therein). Most of these works, however, were carried out in a parametric set-up, that often fails to capture the complexity that is inherent to the data.

Our focus in this paper is on big data, the processing of which in one go exceeds the capacity of a single computer due to the high volume (amount of data) and high velocity (the arriving speed of data). With small demand on storage and being capable of real time updating, the sequential updating approach has been shown to be particularly useful in handling massive data with high velocity. Hereinafter, we refer to approaches of such nature as ONLINE and, in contrast to those classical estimation methods (estimators) as OFFLINE, which requires massive storage space and astronomical computational effort between updates. Wang et al. (2015) discussed in detail the ONLINE adaptation of estimation in linear regression models. Similar ideas have been applied to nonparametric settings; see, for example, Aggarwal et al. (2003). For Parzen-Rosenblatt density estimation, Lambert et al. (1999) proposed an algorithm based on multipole techniques; Cai et al. (2003) proposed a M-kernel approach, with further refinement made in Heinz and Beeger (2006); Kristan et al. (2010, 2011) considered online kernel density estimation based on Gaussian mixture models. What these algorithms have in common is their comparable estimation efficiency to their OFFLINE counterparts. These comparisons have been based on numerical experiments, and no theoretical properties have been established.

It is thus the purpose of this paper to provide a systematic study of the adaptation of a wide variety of kernel-based nonparametric estimators for the analysis of big (online)

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0365

data, and the asymptotic properties of the resulting ONLINE estimators. The adaptation is realized via a sequential updating scheme coupled with index-specific bandwidths. We examine in details the asymptotic efficiencies of these ONLINE estimators and how they depend on the tuning parameters, that in this case, refer to the sequential updating scheme and the index-specific bandwidths. In particular, as far as the N-W type of nonparametric estimator is concerned, we prove that the constant of proportionality associated with index-specific bandwidths should be smaller than that associated with the OFFLINE N-W estimator; we identify the optimal choice among a very general class of sequential updating schemes; and we demonstrate that the relative efficiency of ONLINE N-W estimator changes with the dimension (p) of covariate in a nonlinear manner. It is well known that the local linear estimator is design-adaptive (Fan and Gijbels (1996)) in that its leading bias term does not depend on the design density of covariate. We show that the ONLINE adaptation of local linear estimators is still susceptible to the designs, but to a less extent compared to the ONLINE N-W estimators. For the same reason, general results do not exist on the relative efficiencies of ONLINE local linear estimators against their OFFLINE counterparts.

The rest of the paper is organized as follows. In Section 2, we introduce the general principle of ONLINE adaptation in the context of kernel density estimation. In Sections 3 and 4, we discuss the ONLINE adaptation of estimators in nonparametric regression, and in the varying coefficient models. Section 5 contains some simulation studies where the numerical performances of various ONLINE estimators are evaluated and compared with their respective OFFLINE counterparts.

2 Nonparametric density estimation

Suppose we have IID copies $\mathbf{X}_1, ..., \mathbf{X}_n$ of $\mathbf{X}$, a random $p-$dimensional vector with probability density function $f(.)$. The Parzen-Rosenblatt estimator of $f(.)$ is defined as

$$\hat{f}_n(\mathbf{x}) = n^{-1} \sum_{i=1}^{n} K_{h_n}(\mathbf{X}_{ix}), \tag{2.1}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0365

where $\mathbf{X}_{ix} \stackrel{def}{=} \mathbf{X}_i - \mathbf{x}$, $K(.)$ is a kernel density function in R^p, $K_{h_n}(\mathbf{u}) = K(\mathbf{u}/h_n)/h_n^p$ and $h_n(> 0)$ is the smoothing parameter (bandwidth). As h_n is usually chosen based on the total number of observations n, so whenever n increases (new observations arrive), the summands in (2.1) have to be re-evaluated with the new bandwidth. With data streaming or online data, the evaluation of such classical kernel-based nonparametric estimators thus requires massive computing power and storage capacity.

To discuss the mean squared error (MSE) of $\hat{f}_n(\mathbf{x})$, let $R_2(K) = \int K^2(\mathbf{u})d\mathbf{u}$, and $\mathbf{I}_p$ denote the $p \times p$ identity matrix; for any $\mathbf{u} \in R^p$, $\mathbf{u}^\top$ stands for its transpose. The following conditions are assumed throughout.

(A1) The kernel function $K(.)$ is symmetric and $\int \mathbf{u}\mathbf{u}^\top K(\mathbf{u})d\mathbf{u} = \mathbf{I}_p$.

(A2) The density function $f(.)$ has bounded third order derivatives.

The existence of bounded third order derivatives of $f(.)$ is assumed purely for convenience; it could certainly be relaxed at the expense of more complicated technical analysis, but it does not affect the main findings of this paper.

Hereinafter, for a multivariate function, $f(.)$ say, denote by $\nabla_f(.)$, its gradient vector and by $\mathcal{H}_f(.)$, its Hessian (second-order partial derivatives) matrix. For univariate functions, we revert to the traditional and more convenient notations $f^{(1)}(.)$ and $f^{(2)}(.)$.

Lemma 2.1 *Suppose (A1) and (A2) hold and that* $h_n \to 0,\ nh_n/\log n \to \infty$. *Then*

$$E\{|\hat{f}_n(\mathbf{x}) - f(\mathbf{x})|^2\} = AMSE(f_n(\mathbf{x})|h_n) + o(h_n^4 + (nh_n^p)^{-1}),$$

where

$$AMSE(\hat{f}_n(\mathbf{x})|h_n) \stackrel{def}{=} \frac{1}{4}[tr\{\mathcal{H}_f(\mathbf{x})\}]^2 h_n^4 + \frac{1}{nh_n^p} f(\mathbf{x}) R_2(K);$$

the asymptotically optimal bandwidth that minimizes $AMSE(\hat{f}_n(\mathbf{x})|h_n)$ *is given by*

$$h_{n,opt} \stackrel{def}{=} \Big[\frac{f(\mathbf{x})R_2(K)}{tr\{\mathcal{H}_f(\mathbf{x})\}^2}\Big]^{1/(4+p)} (p/n)^{1/(4+p)}, \tag{2.2}$$

and, correspondingly,

$$AMSE(\hat{f}_n(\mathbf{x})|h_{n,opt}) = \frac{(p+4)}{4(p^p n^4)^{1/(4+p)}} [f(\mathbf{x})R_2(K)]^{4/(4+p)} [tr\{\mathcal{H}_f(\mathbf{x})\}]^{2p/(4+p)}.$$

The proofs can be found in, e.g., Wand and Jones (1995).

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0365

2.1 ONLINE Parzen-Rosenblatt density estimation

This is implemented as follows. With only $\mathbf{X}_1$, the estimate is a simple $K_{\tilde{h}_1}(\mathbf{X}_1 - \mathbf{x})$ for some predetermined bandwidth $\tilde{h}_1$; once $\mathbf{X}_2$ is available, we update the estimate as

$$(1-\beta_2)K_{\tilde{h}_1}(\mathbf{X}_{1x}) + \beta_2 K_{\tilde{h}_2}(\mathbf{X}_{2x}),$$

where $\beta_2 \in (0,1)$ is some pre-specified constant, and $\tilde{h}_2$ is yet another bandwidth chosen independent of $\mathbf{X}_1$. In general, suppose $\tilde{f}_{n-1}(\mathbf{x}), n \geq 2$, is the current estimate after $\mathbf{X}_1, ..., \mathbf{X}_{n-1}$ have been observed. Once $\mathbf{X}_n$ arrives, the estimate is then updated as a weighted sum of $\tilde{f}_{n-1}(\mathbf{x})$ and $K_{\tilde{h}_n}(\mathbf{X}_{nx})$:

$$\tilde{f}_n(\mathbf{x}) \stackrel{def}{=} (1-\beta_n)\tilde{f}_{n-1}(\mathbf{x}) + \beta_n K_{\tilde{h}_n}(\mathbf{X}_{nx}), \tag{2.3}$$

where $\beta_n \in (0,1)$ is a pre-specified constant, and the bandwidth $\tilde{h}_n$ is again chosen independent of all the preceding observations. To highlight the fact that different bandwidths are used in each step leading up to $\tilde{f}_n(\mathbf{x})$, and also its dependence on the weighting sequence $\{\beta_i,\, i = 1, \cdots, n\}$, we rewrite $\tilde{f}_n(\mathbf{x})$ as $\tilde{f}_n(\mathbf{x}|\tilde{h}_n, \beta_n)$, and thus (2.3) is

$$\tilde{f}_n(\mathbf{x}|\tilde{h}_n, \beta_n) = (1-\beta_n)\tilde{f}_{n-1}(\mathbf{x}|\tilde{h}_{n-1}, \beta_{n-1}) + \beta_n K_{\tilde{h}_n}(\mathbf{X}_{nx}). \tag{2.4}$$

This formulation lays the foundation for our ONLINE adaptation of kernel based estimation. Many well-known estimators in nonparametric or semiparametric models are formed based on statistics of forms similar to (2.3); more examples can be found Section 3 and Section 4.

The asymptotic properties of $\tilde{f}_n(\mathbf{x}|\tilde{h}_n, \beta_n)$ depend on sequences of tuning parameters: the sequence of bandwidths $\{\tilde{h}_n\}$, and the sequence of weights $\{\beta_n\}$. In this paper, we mainly focus on the case where $\beta_n = n^{-1}$, which is optimal in the sense that the corresponding $\tilde{f}_n(\mathbf{x}|\tilde{h}_n, \beta_n)$ with $\beta_n = n^{-1}$ has the smallest AMSE amongst a general class of weighting series; see Section 2.3. For ease of exposition, $\tilde{f}_n(\mathbf{x}|\tilde{h}_n, n^{-1})$ is simply written as $\tilde{f}_n(\mathbf{x}|\tilde{h}_n)$,

$$\tilde{f}_n(\mathbf{x}|\tilde{h}_n) = \frac{1}{n}\sum_{i=1}^{n} K_{\tilde{h}_i}(\mathbf{X}_{ix}). \tag{2.5}$$

To investigate how the sequence of bandwidths $\{\tilde{h}_n\}$ affects the efficiency of the ONLINE estimator (2.5), we start with the index-specific bandwidths

$$\tilde{h}_i = ci^{-\alpha}, \quad i = 1, 2, \cdots, \quad \text{for some constants } c > 0,\ \alpha > 0.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0365

Lemma 2.2 *Under (A1) and (A2), we have*

$$AMSE(\tilde{f}_n(\mathbf{x}|\tilde{h}_n)) = \frac{c^4}{4(1-2\alpha)^2}[tr\{\mathcal{H}_f(\mathbf{x})\}]^2 n^{-4\alpha} + \frac{1}{(1+p\alpha)c^p} f(\mathbf{x}) R_2(K) n^{p\alpha-1},$$

which is minimized at

$$\alpha = 1/(4+p), \quad c = \Big(\frac{p(p+2)}{2(p+4)}\Big)^{1/(4+p)} \Big(\frac{R_2(K)f(\mathbf{x})}{[tr\{\mathcal{H}_f(\mathbf{x})\}]^2}\Big)^{1/(4+p)},$$

with minimum

$$\frac{(p+4)^2}{8(p+2)}\Big(\frac{p(p+2)}{2(p+4)}\Big)^{-p/(4+p)} n^{-4/(4+p)} [f(\mathbf{x})R_2(K)]^{4/(4+p)} [tr\{\mathcal{H}_f(\mathbf{x})\}]^{2p/(4+p)}.$$

From Lemma 2.1, for OFFLINE Parzen-Rosenblatt estimator with bandwidth $h_n = cn^{-\alpha}$, the optimal values for α and c are

$$\alpha = 1/(p+4), \quad c = \Big\{\frac{f(\mathbf{x})R_2(K)}{tr\{\mathcal{H}_f(\mathbf{x})\}^2}\Big\}^{1/(4+p)} p^{1/(4+p)}. \tag{2.6}$$

So, while the optimal choices of α are identical for OFFLINE and ONLINE, their respective optimal choices for the coefficient c do differ, with the ratio given by

$$\frac{\text{optimal } \tilde{h}_n}{\text{optimal } h_n} = \left(\frac{p+2}{2(p+4)}\right)^{1/(p+1)}, \tag{2.7}$$

which is always less than 1. A ratio of less than one is expected, for otherwise, the use of index-specific bandwidths $\tilde{h}_i \propto i^{-\alpha}$ will result in too large a bias for ONLINE, which cannot be compensated for by the accompanying reduction in the variance.

The relative efficiency of the ONLINE Parzen-Rosenblatt estimator against its OFFLINE counterpart is

$$\frac{AMSE(\hat{f}_n(\mathbf{x})|\text{optimal } h_n)}{AMSE(\tilde{f}_n(\mathbf{x})|\text{optimal } \tilde{h}_n)} = 2^{4/(4+p)}\Big(\frac{p+2}{p+4}\Big)^{(2p+4)/(p+4)}. \tag{2.8}$$

This starts at 0.9432 for $p = 1$, drops to its lowest level of 0.9186 at $p = 4$, and then slowly increases to 1 as $p \to \infty$; see Figure 1.

The non-linear pattern in the relative efficiency (2.8) versus dimensionality is a result of several confounding factors. These include, for example, the use of a common bandwidth by OFFLINE in contrast to the use of index-specific bandwidths by ONLINE. Also, that the OFFLINE has smaller bias but larger variance, yet the exact opposite holds true for ONLINE. For both OFFLINE and ONLINE, as the dimension increases, the so-called 'curse of

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0365

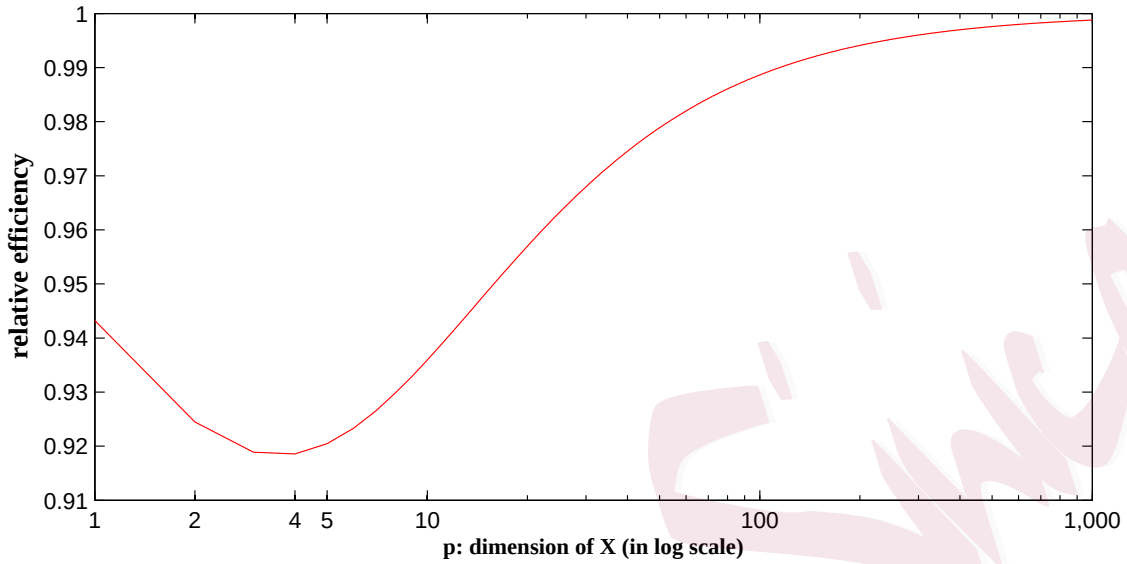

Figure 1: Relatively efficiencies of against the OFFLINE Parzen-Rosenblatt estimator (2.1).

dimensionality' quickly overtakes the limited amount of improvement achieved with the use of an optimal bandwidth, so that their performances become increasingly indistinguishable from each other (equally bad) with a growing dimension.

2.2 Another performance measurement for ONLINE

Our focus has been the performance of $\tilde{f}_n(\mathbf{x}|\tilde{h}_n)$ for a given n. Yet, over the course of which the data stream, the sequence of estimates $\{\tilde{f}_n(\mathbf{x}|\tilde{h}_n),\ n = 1, 2, \cdots,\}$ is obtained, and any of which when available, could be used as a substitute for the true but unknown $f(.)$ for purposes of statistical inference. It is thus relevant to examine their collective performance up to the present stage, $n = N$, say. We define the congregated MSE (CMSE) as

$$CMSE(N, \alpha, c) \equiv \frac{1}{N}\sum_{n=1}^{N} E\{|\tilde{f}_n(\mathbf{x}|\tilde{h}_n) - f(\mathbf{x})|^2\}, \tag{2.9}$$

which is a function of N and constants α and c that specify the sequence of the index-specific bandwidths $\{\tilde{h}_i = ci^{-\alpha},\ i = 1, 2, \cdots\}$.

Lemma 2.3 *If (A1) and (A2) hold,*

$$CMSE(N, \alpha, c) = \frac{c^4[tr\{\mathcal{H}_f(\mathbf{x})\}]^2}{4(1-2\alpha)^2(1-4\alpha)}N^{-4\alpha} + \frac{R_2(K)f(\mathbf{x})}{c^p p\alpha(p\alpha+1)}N^{p\alpha-1} + o(N^{-4\alpha} + N^{p\alpha-1});$$

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0365

the sum of the first two terms is minimized when

$$\alpha = 1/(4+p), \quad c = \Big[\frac{p(p+2)}{2(p+4)}\Big]^{1/(4+p)} \Big(\frac{R_2(K) f(\mathbf{x})}{[tr\{\mathcal{H}_f(\mathbf{x})\}]^2}\Big)^{1/(4+p)}.$$

In view of Lemma 2.2, the sequence of the index-specific bandwidths that minimizes the MSE of an individual $\tilde{f}_n(\mathbf{x}|\tilde{h}_n)$, is also the optimal choice for the congregated MSE (2.9), .

2.3 Optimum of ONLINE Parzen-Rosenblatt density estimators

We have been focused on the ONLINE Parzen-Rosenblatt estimator (2.4) with the weighting sequence $\beta_n = 1/n$, and the index-specific bandwidths taking the form of $\tilde{h}_i = ci^{-\alpha}$ for some constants c and α. In this section, we find out whether such an ONLINE estimator can be improved if the index-specific bandwidths $\tilde{h}_i$ and the weighting series β_n are allowed to take on more general forms. The ONLINE estimator (2.5) has a larger bias that cannot be fully compensated by the reduction in its variance. On one hand, due to the use of index-specific bandwidths $\tilde{h}_i \propto i^{-\alpha}$, the summands $K_{\tilde{h}_i}(\mathbf{X}_{ix})$ have larger biases than their counterparts using a 'universal' bandwidth $h_n \propto n^{-\alpha}$. That the optimal choice for coefficient c in $\tilde{h}_i$ is smaller than the optimal coefficient for the universal h_n (see (2.7)) is not enough to fully correct the inflation in the bias. We investigate whether the situation would improve if the coefficient in the index-specific bandwidths is allowed to change with i as well. Here (2.4) with $\beta_n = n^{-1}$, or its equivalent (2.5), has an equal weight of n^{-1} assigned to each summand $K_{\tilde{h}_i}(\mathbf{X}_{ix})$, $i = 1, 2, \cdots, n$, taking no consideration of the fact that these estimates behave differently. Hence we look at alternative weighting series β_n.

Consider a general formulation of index-specific bandwidths with varying coefficient, such that

$$\tilde{h}_i = c(1-\theta_i) i^{-\alpha} \quad \text{for some } c > 0,\ \alpha > 0,\ \text{and } \theta_i \downarrow 0, \quad \text{as } i \to \infty. \tag{2.10}$$

Examples of such θ_i include $\theta_i = 1 - \exp(-a\, i^{-b})$ or $\theta_i = a(\log i)^{-b}$, for any given values of $a > 0$, $b > 0$.

Lemma 2.4 *Under the conditions of Lemma 2.2, the AMSE of $\tilde{f}_n(\mathbf{x}|\tilde{h}_n)$ of (2.5) with varying coefficient bandwidths (2.10) is identical to that of $\tilde{f}_n(\mathbf{x}|\tilde{h}_n)$ with index-specific bandwidths $\tilde{h}_i = ci^{-\alpha}$.*

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0365

Having varying coefficients in the index-specific bandwidth $\tilde{h}_i$ thus does not improve the asymptotic efficiency of the ONLINE estimator.

For alternative weighting we consider that $\beta_n \downarrow 0$, as $n \to \infty$. For any positive integers n and N such that $n < N$, let

$$w_{N,n} := \beta_n \prod_{k=n+1}^{N} (1-\beta_k), \quad S_N := \sum_{n=1}^{N} w_{N,n} n^{-2\alpha}, \quad \tilde{S}_N := \sum_{n=1}^{N} w_{N,n}^2 n^{p\alpha}.$$

Lemma 2.5 *For ONLINE estimator $\tilde{f}_N(\mathbf{x}|\tilde{h}_N, \beta_N)$ (2.3) with weighting series $\{\beta_n\}$ and index-specific bandwidths $\tilde{h}_n = cn^{-\alpha}$ for some constants $c > 0$ and $\alpha > 0$, we have*

$$AMSE(\tilde{f}_N(\mathbf{x}|\tilde{h}_N, \beta_N)) = \frac{c^4}{4}[tr\{\mathcal{H}_f(\mathbf{x})\}]^2 S_N^2 + c^{-p} f(\mathbf{x}) R_2(K) \tilde{S}_N.$$

If $\beta_n \propto n^{-1}$, we have $S_N \propto N^{-2\alpha}$ and $\tilde{S}_N \propto N^{p\alpha-1}$; if $n\beta_n \to \infty$, we have $S_N \propto N^{-2\alpha}$ and $\tilde{S}_N / N^{p\alpha-1} \to \infty$; if $n\beta_n \to 0$, we have $N^a S_N \to \infty$ and $N^a \tilde{S}_N \to \infty$, for any $a > 0$.

Therefore, an admissible weighting series must be such that $\beta_n \propto n^{-1}$. We find that the AMSE of the ONLINE estimator is minimal with $n\beta_n \to 1$.

Lemma 2.6 *If $\tilde{h}_n = cn^{-\alpha}$ for some constants $\alpha > 0$ and $c > 0$, among the ONLINE estimators of form (2.3) with weighting series $\beta_n \propto n^{-1}$, AMSE is minimized if and only if $n\beta_n \to 1$.*

Thus, as far as the minimization of AMSE is concerned, the ONLINE estimator (2.5) with index-specific bandwidths $\tilde{h}_i = ci^{-\alpha}$ is optimal. In the following discussions of the ONLINE adaptation of other nonparametric estimators, we use the weighting series $\beta_n = 1/n$.

3 Nonparametric regression

Suppose IID observations $(Y_i, \mathbf{X}_i)$ are generated according to $Y_i = m(\mathbf{X}_i) + \varepsilon_i,\ i = 1, \cdots, n$, where $\mathbf{X}_i \in R^p$ with probability density function $f(.)$, $E(\varepsilon_i|\mathbf{X}_i) = 0,\ Var(\varepsilon_i|\mathbf{X}_i) = \sigma_\varepsilon^2$, and $m(.)$ is some unknown function $m(.)$. We are interested in the estimation of $m(.)$ for any given $\mathbf{x}$ in the support of $f(.)$.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0365

3.1 Local constant estimator

The Nadaraya-Watson (local constant) estimate of $m(.)$ is

$$\hat{m}_{nw}(\mathbf{x}|h_n) = \frac{\sum_{i=1}^n K_{h_n}(\mathbf{X}_{ix})Y_i}{\sum_{i=1}^n K_{h_n}(\mathbf{X}_{ix})}, \tag{3.11}$$

for some bandwidth h_n. This is the ratio of two N-W type of estimate we have seen in Section 1. Its ONLINE version is

$$\tilde{m}_{nw}(\mathbf{x}|\alpha, c) \;=\; \frac{\sum_{i=1}^n K_{h_i}(\mathbf{X}_{ix})Y_i}{\sum_{i=1}^n K_{h_i}(\mathbf{X}_{ix})}, \tag{3.12}$$

again with index-specific bandwidths $h_i = ci^{-\alpha}$, $i = 1, 2\cdots, n$, for some constants $c > 0$, $\alpha > 0$. The asymptotic properties of the two estimators closely resemble what we have seen in the previous section. Write $\tilde{R}_2(K) := \int \mathbf{u}\mathbf{u}^\top K^2(\mathbf{u})d\mathbf{u}$.

Lemma 3.1 *Suppose (A1) and (A2) hold. If $m(.)$ has bounded third order derivatives, then*

$$\begin{aligned} AMSE\{\hat{m}_{nw}(\mathbf{x}|h_n)\} &= \frac{1}{4}[tr\{\mathcal{H}_m(\mathbf{x})\} + 2\nabla m^\top(\mathbf{x})\nabla f(\mathbf{x})/f(\mathbf{x})]^2 h_n^4 + \frac{\tilde{R}_2(K)\sigma_\varepsilon^2}{nh_n^p f(\mathbf{x})}, \\ AMSE\{\tilde{m}_{nw}(\mathbf{x}|\alpha, c)\} &= \frac{c^4}{4(1-2\alpha)^2}[tr\{\mathcal{H}_m(\mathbf{x})\} + 2\nabla m^\top(\mathbf{x})\nabla f(\mathbf{x})/f(\mathbf{x})]^2 n^{-4\alpha} \\ &\quad + \frac{n^{p\alpha-1}\tilde{R}_2(K)\sigma_\varepsilon^2}{(1+p\alpha)c^p f(\mathbf{x})}. \end{aligned}$$

Along the line of the derivations used to obtain (2.6) and (2.7), one finds that the optimal choice for α is still $1/(p+4)$; the ratio between the optimal bandwidths that minimize the AMSE of the N-W type of estimators ($\hat{m}_{nw}(\mathbf{x}|h_n)$ and $\tilde{m}_{nw}(\mathbf{x}|\alpha, c)$) is identical to that give in (2.7); and the relative efficiency of $\tilde{m}_{nw}(\mathbf{x}|\alpha, c)$ against $\hat{m}_{nw}(\mathbf{x}|h_n)$ is, as given in (2.8),

$$\frac{AMSE\{\hat{m}_{nw}(\mathbf{x}|\text{optimal } h_n)\}}{AMSE\{\tilde{m}_{nw}(\mathbf{x}|\text{optimal } \tilde{h}_n)\}} = 2^{4/(4+p)}\Big(\frac{p+2}{p+4}\Big)^{(2p+4)/(p+4)}. \tag{3.13}$$

3.2 Local linear estimator

In the case of a smooth enough $m(.)$, its estimation can also be based on the local linear approximation of $m(.)$. Consider the minimization of

$$\sum_{i=1}^n K_{h_n}(\mathbf{X}_{ix})\{Y_i - \mathbf{b}^\top \mathbf{X}_{in}(\mathbf{x})\}^2, \tag{3.14}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0365

with respect to $\mathbf{b} \in R^{p+1}$, where $\mathbf{X}_{in}(\mathbf{x}) := [1, \mathbf{X}_{ix}^\top/h_n]^\top$. The minimizer $\hat{m}_n(\mathbf{x})$, is an estimate of $m_n(\mathbf{x}) = [m(\mathbf{x}), h_n\nabla^\top m(\mathbf{x})]^\top$. The first element of $\hat{m}_n(\mathbf{x})$, $\hat{m}_l(\mathbf{x}|h_n)$, is referred to as the local linear estimator of $m(\mathbf{x})$. $\hat{m}_l(\mathbf{x}|h_n)$ has the same asymptotic variance as the local constant estimate $\hat{m}_{nw}(\mathbf{x}|h_n)$, but its bias term admits a much simpler form.

Lemma 3.2 *If conditions of Lemma 3.1 hold,*

$$E[\hat{m}_l(\mathbf{x}|h_n)] = m(\mathbf{x}) + \frac{1}{2}h_n^2[tr\{\mathcal{H}_m(\mathbf{x})\}] + O(h_n^4),$$
$$Var[\hat{m}_l(\mathbf{x}|h_n)] = (nh_n^p)^{-1}\tilde{R}_2(K)\sigma_\varepsilon^2/f(\mathbf{x}) + o((nh_n^p)^{-1}).$$

The bias of $\hat{m}_l(\mathbf{x}|h_n)$ only depends on the Hessian matrix of $m(.)$ while, for the N-W estimator $\hat{m}_{nw}(\mathbf{x}|h_n)$, the bias also depends on the first order derivatives of both $f(.)$ and $m(.)$ and $f^{-1}(.)$. Thus, the local linear estimator is design-adaptive (Fan and Gijbels (1996)). We look at whether the ONLINE local linear estimator is also design-adaptive.

The minimizer of (3.14) has the analytic form

$$\begin{aligned}\hat{m}_n(\mathbf{x}) &= \Big[\sum_{i=1}^n K_{h_n}(\mathbf{X}_{ix})\mathbf{X}_{in}(\mathbf{x})\mathbf{X}_{in}^\top(\mathbf{x})\Big]^{-1}\sum_{i=1}^n K_{h_n}(\mathbf{X}_{ix})\mathbf{X}_{in}(\mathbf{x})Y_i \\ &\stackrel{def}{=} [n\mathcal{S}_n(\mathbf{x})]^{-1}n\mathcal{S}_n(\mathbf{x}, Y).\end{aligned}$$

For $\mathcal{S}_n(\mathbf{x})$ and $\mathcal{S}_n(\mathbf{x}, Y)$, we cab take their ONLINE counterparts as

$$\tilde{\mathcal{S}}_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n K_{\tilde{h}_i}(\mathbf{X}_{ix})\tilde{\mathbf{X}}_{in}(\mathbf{x})\tilde{\mathbf{X}}_{in}^\top(\mathbf{x}), \quad \tilde{\mathcal{S}}_n(\mathbf{x}, Y) = \frac{1}{n}\sum_{i=1}^n K_{\tilde{h}_i}(\mathbf{X}_{ix})\tilde{\mathbf{X}}_{in}(\mathbf{x})Y_i,$$

with $\tilde{\mathbf{X}}_{in}(\mathbf{x}) = [1, \mathbf{X}_{ix}^\top/\tilde{h}_i]^\top$ and $\tilde{h}_i = ci^{-\alpha}$, as before. Let

$$\tilde{m}_n(\mathbf{x}) \quad = \quad [\tilde{\mathcal{S}}_n(\mathbf{x})]^{-1}\tilde{\mathcal{S}}_n(\mathbf{x}, Y); \tag{3.15}$$

the ONLINE local linear estimator, $\tilde{m}_l(\mathbf{x}|\alpha, c)$, is thus given by the first element of $\tilde{m}_n(\mathbf{x})$.

Lemma 3.3 *If teh conditions of Lemma 3.1 hold,*

$$\begin{aligned}Bias(\tilde{m}_l(\mathbf{x}|\alpha, c)) &= \frac{c^2\alpha^2 n^{-2\alpha}}{(1-2\alpha)(1-\alpha)^2}\frac{\nabla^\top m(\mathbf{x})\nabla f(\mathbf{x})}{f(\mathbf{x})} + \frac{c^2 n^{-2\alpha}}{2(1-2\alpha)}tr\{\mathcal{H}_m(\mathbf{x})\} + O(n^{-3\alpha}), \\ \mathrm{Var}(\tilde{m}_l(\mathbf{x}|\alpha, c)) &= \frac{\tilde{R}_2(K)\sigma_\varepsilon^2}{f(\mathbf{x})c^p(1+p\alpha)}n^{p\alpha-1}(1+o(1)).\end{aligned} \tag{3.16}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0365

Comparing the bias term here with that of the ONLINE local constant estimator $\tilde{m}_{nw}(\mathbf{x}|\alpha, c)$ given in Lemma 3.1, one sees that $\tilde{m}_l(\mathbf{x}|\alpha, c)$ is still susceptible to large bias caused by clustering in the design of X, albeit to a less degree, reflected in the coefficient being brought down from $\frac{1}{2(1-2\alpha)}$, in $\tilde{m}_{nw}(\mathbf{x}|\alpha, c)$, to $\frac{\alpha^2}{(1-2\alpha)(1-\alpha)^2}$ in $\tilde{m}_l(\mathbf{x}|\alpha, c)$ with $\alpha = 1/(p+4)$. This does not mean that $\tilde{m}_l(\mathbf{x}|\alpha, c)$ is categorically less efficient than its OFFLINE counterpart $\hat{m}_l(\mathbf{x}|h_n)$, because the bias of the former is not necessarily larger than the latter: it not only depends on the inverse of the density function $f(.)$, but also on other quantities such as first order derivatives of $m(.)$ and $f(.)$.

4 The varying coefficient regression model

This model has been extensively studied; see, e.g., Fan and Zhang (1999) and, more recently, Park et al. (2015) and the references therein. Suppose we have IID observations $(Y_i, \mathbf{X}_i, U_i), i = 1, 2, \cdots$, generated according to

$$Y_i = g_1(U_i)x_{i1} + \cdots + g_q(U_i)x_{iq} + \varepsilon_i,$$

with $\mathbf{X}_i = (x_{i1}, \cdots, x_{iq})^\top$, $E(\varepsilon_i|\mathbf{X}_i, U_i) = 0$, $Var(\varepsilon_i|\mathbf{X}_i, U_i) = \sigma^2$ and $g(.)$, $k = 1, \cdots, q$, are unknown functions. To include an intercept term, we set $x_{i1} \equiv 1$. Our interest is in the estimation of $\mathbf{g}(u_0) = [g_1(u_0), \cdots, g_q(u_0)]^\top$ for any given u_0 in the support of $f(.)$, the density function of U_i. For any given $u \in R$, let $\nu(u) = E(\mathbf{X}_i\mathbf{X}_i^\top|U_i = u)$ and write $\nu(u)f_U(u)$ as $(\nu.f)(u)$. Suppose both $\nu(.)$ and $f(.)$ have bounded third order derivatives.

4.1 The N-W estimator

Write U_{i0} for $U_i - u_0$. Consider the minimization of

$$\sum_{i=1}^{n}(Y_i - \mathbf{X}_i^\top\mathbf{g})^2 K_{h_n}(U_{i0})$$

with respect to vector $\mathbf{g} \in R^q$. The minimum is achieved at

$$\hat{\mathbf{g}}_{nw}(u_0|h_n) := \Big[\sum_{i=1}^{n} K_{h_n}(U_{i0})\mathbf{X}_i\mathbf{X}_i\Big]^{-1}\sum_{i=1}^{n} K_{h_n}(U_{i0})\mathbf{X}_iY_i. \tag{4.17}$$

The ONLINE version of (4.17) is

$$\tilde{\mathbf{g}}_{nw}(u_0|c) := \Big[\sum_{i=1}^{n} K_{\tilde{h}_i}(U_{i0})\mathbf{X}_i\mathbf{X}_i^\top\Big]^{-1}\sum_{i=1}^{n} K_{\tilde{h}_i}(U_{i0})\mathbf{X}_iY_i,$$

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0365

where $\tilde{h}_i = ci^{-1/5},\ i = 1, 2, \cdots$, for some constant $c > 0$. Write

$$\mathbf{g}^{(1)}(u_0) = [g_1^{(1)}(u_0), \cdots, g_q^{(1)}(u_0)]^\top, \quad \mathbf{g}^{(2)}(u_0) = [g_1^{(2)}(u_0), \cdots, g_q^{(2)}(u_0)]^\top.$$

Lemma 4.1 *If $\nu(u_0)$ is positive definite, the $g_k(.),\ k = 1, \cdots, q$, have bounded third order derivatives, $h_n \propto n^{-1/5}$, and $\tilde{h}_i = ci^{-1/5}$, $i \geq 1$,*

$$\hat{\mathbf{g}}_{nw}(u_0|h_n) = \mathbf{g}(u_0) + h_n^2 B(u_0) + [n(\nu.f)(u_0)]^{-1} \sum_{i=1}^{n} K_{h_n}(U_{i0})\mathbf{X}_i\varepsilon_i + o_p(n^{-1/2}), \quad (4.18)$$

$$\tilde{\mathbf{g}}_{nw}(u_0|c) = \mathbf{g}(u_0) + \frac{5c^2}{3}n^{-2/5}B(u_0) + [n(\nu.f)(u_0)]^{-1} \sum_{i=1}^{n} K_{\tilde{h}_i}(U_{i0})\mathbf{X}_i\varepsilon_i + o_p(n^{-1/2}), \quad (4.19)$$

where $B(u_0) = [(\nu.f)(u_0)]^{-1}[(\nu.f)^{(1)}(u_0)]\mathbf{g}^{(1)}(u_0) + \frac{1}{2}\mathbf{g}^{(2)}(u_0)$.

For the vector-valued estimate $\hat{\mathbf{g}}_{nw}(u_0)$, we define its MSE as $E\|\hat{\mathbf{g}}_{nw}(u_0) - \mathbf{g}(u_0)\|^2$, where $\|.\|$ stands for the Euclidean norm. Based on Lemma 4.1,

$$AMSE(\hat{\mathbf{g}}_{nw}(u_0|h_n)) = h_n^4\|B(u_0)\|^2 + (nh_n)^{-1}\sigma^2 tr\{[(\nu.f)(u_0)]^{-1}\},$$
$$AMSE(\tilde{\mathbf{g}}_{nw}(u_0|c)) = \frac{25c^4}{9}n^{-4/5}\|B(u_0)\|^2 + \frac{5\sigma^2}{6c}n^{-4/5}tr\{[(\nu.f)(u_0)]^{-1}\}.$$

The relative efficiency of these two estimators, when evaluated for their respective optimal bandwidths are the same as those given in (2.8) with $p = 1$. While U is one dimensional, here general results also hold for when U is p-dimensional, with the ratio of their respective optimal bandwidths as given in (2.7) and the relative efficiency of the two estimators given by (2.8),

$$\frac{AMSE\{\hat{\mathbf{g}}_{nw}(u|\text{optimal } h_n)\}}{AMSE\{\tilde{\mathbf{g}}_{nw}(u|\text{optimal } \tilde{h}_n)\}} = 2^{4/(4+p)}\Big(\frac{p+2}{p+4}\Big)^{(2p+4)/(p+4)}. \quad (4.20)$$

4.2 The local linear estimator

Write $G_n(u_0) = [g_1(u_0), h_n g_1^{(1)}(u_0), \cdots, g_q(u_0), h_n g_q^{(1)}(u_0)]^\top$, so

$$g_k(U_i) = g_k(u_0) + g_k^{(1)}(u_0)U_{i0} + \frac{1}{2}g_k^{(2)}(u_0)U_{i0}^2 + O(|U_{i0}|^3),\ k = 1, \cdots, q.$$

Take $\mathbf{X}_{n,i}(u_0) = \mathbf{X}_i \otimes [1, U_{i0}/h_n]^\top$, where $\otimes$ stands for the Kronecker product. Then the local linear estimate of $G_n(u_0)$ is obtained via the minimization of the function

$$\sum_{i=1}^{n}(Y_i - \mathbf{X}_{n,i}(u_0)^\top\mathbf{g})^2 K_{h_n}(U_{i0}) \quad (4.21)$$

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0365

with respect to $\mathbf{g} \in R^{2q}$. It is realized at $\hat{G}_n(u_0) \equiv \Sigma_n^{-1}\mathcal{S}_n$, where

$$\mathcal{S}_n = n^{-1}\sum_{i=1}^{n} K_{h_n}(U_{i0})Y_i\mathbf{X}_{n,i}(u_0); \quad \Sigma_n = n^{-1}\sum_{i=1}^{n} K_{h_n}(U_{i0})\mathbf{X}_{n,i}(u_0)\mathbf{X}_{n,i}^{\top}(u_0).$$

The local linear estimate of $\mathbf{g}(u_0|h_n)$ is $\hat{\mathbf{g}}_l(u_0) := \mathbf{I}_{q,2q}\hat{G}_n(u_0)$, where $\mathbf{I}_{q,2q} = \mathbf{I}_q \otimes [1,0]$.

Lemma 4.2 *If the conditions in Lemma 4.1 hold,*

$$\hat{\mathbf{g}}_l(u_0|h_n) = \mathbf{g}(u_0) + \frac{1}{2}h_n^2\mathbf{g}^{(2)}(u_0) + [(\nu.f)(u_0)]^{-1}\frac{1}{n}\sum_{i=1}^{n} K_{h_n}(U_{i0})\mathbf{X}_i\varepsilon_i + O_p(n^{-1/2}). \quad (4.22)$$

Comparing this with the local constant estimator (4.18), we can see that the two share the same stochastic error term, but the local linear estimator is again 'design-adaptive': there is no presence of $f^{-1}(.)$ in its bias term.

For the ONLINE adaptation of $\hat{\mathbf{g}}_l(u_0|h_n)$, write $\tilde{\mathbf{X}}_{n,i}(u_0) = \mathbf{X}_i \otimes [1, U_{i0}/\tilde{h}_i]^{\top}$, with $\tilde{h}_i = c\, i^{-1/5},\ i = 1, 2, \cdots$. The ONLINE version of $\hat{G}_n(u_0)$ is given by

$$\tilde{G}_n(u_0|c) = \Big[\sum_{i=1}^{n} K_{\tilde{h}_i}(U_{i0})\tilde{\mathbf{X}}_{n,i}(u_0)\tilde{\mathbf{X}}_{n,i}^{\top}(u_0)\Big]^{-1}\sum_{i=1}^{n} K_{\tilde{h}_i}(U_{i0})\tilde{\mathbf{X}}_{n,i}(u_0)Y_i. \quad (4.23)$$

Let $\mu_4(K)$ be the fourth moment of $K(.)$.

Lemma 4.3 *If the assumptions of Lemma 4.1 hold, then*

$$\begin{aligned}\tilde{G}_n(u_0|c) = {} & \mathbf{g}(u_0) \otimes \begin{bmatrix}1\\0\end{bmatrix} + \frac{5c}{4}\mathbf{g}^{(1)}(u_0) \otimes \begin{bmatrix}0\\1\end{bmatrix} + \frac{5c^2}{6}n^{-2/5}\mathbf{g}^{(2)}(u_0) \otimes \begin{bmatrix}1\\0\end{bmatrix} \\ & + \frac{5c^2}{48}n^{-2/5}\Big([(\nu.f)(u_0)]^{-1}[(\nu.f)^{(1)}(u_0)]\mathbf{g}^{(1)}(u_0)\Big) \otimes \begin{bmatrix}1\\0\end{bmatrix} \\ & + \frac{5c^3}{4}n^{-3/5}\Big([(\nu.f)(u_0)]^{-1}[(\nu.f)^{(1)}(u_0)]\mathbf{g}^{(2)}(u_0)\Big) \otimes \begin{bmatrix}0\\\mu_4(K)\end{bmatrix} \\ & + \frac{1}{n}\sum_{i=1}^{n} K_{\tilde{h}_i}(U_{i0})\Big([(\nu.f)(u_0)]^{-1}\mathbf{X}_i\Big) \otimes [1, U_{i0}/\tilde{h}_i]^{\top}\varepsilon_i + o_p(n^{-1/2}).\end{aligned} \quad (4.24)$$

Apparently, for the estimate of the gradient vector to be asymptotically unbiased, it is necessary that $c = 4/5$. Yet, as far as the functions $\mathbf{g}(u_0)$ are concerned, their estimate, $\tilde{\mathbf{g}}(u_0) := \mathbf{I}_{q,2q}\tilde{G}_n(u_0)$, is indeed asymptotically unbiased, with an asymptotic variance identical to that of $\tilde{\mathbf{g}}_{nw}(u_0|c)$, the ONLINE local constant estimator. The asymptotic bias of $\tilde{\mathbf{g}}(u_0)$ is two terms:

$$\frac{5c^2}{48}n^{-2/5}[(\nu.f)(u_0)]^{-1}[(\nu.f)^{(1)}(u_0)]\mathbf{g}^{(1)}(u_0) + \frac{5c^2}{6}n^{-2/5}\mathbf{g}^{(2)}(u_0).$$

Again, the local linear ONLINE estimator is not fully design-adaptive; it only manages to limit the extent to which its bias is affected by the design density $f(.)$ of $\mathbf{X}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0365

5 Simulation study

In this section, we present numerical results regarding the efficiency of some ONLINE estimators against their OFFLINE counterparts. With this, there is the fact that the computation in the case of OFFLINE is time consuming, while that for ONLINE can be realized almost real-time.

The first example concerns the estimation of probability density function

$$f(\mathbf{x}) = \frac{0.7}{(2\pi\sigma_1^2)^{p/2}} \exp\Big(- \frac{\sum_{i=1}^p (x_i - \mu_1)^2}{2\sigma_1^2}\Big) + \frac{0.3}{(2\pi\sigma_2^2)^{p/2}} \exp\Big(- \frac{\sum_{i=1}^p (x_i - \mu_2)^2}{2\sigma_2^2}\Big)$$

where $\mathbf{x} = (x_1, ..., x_p)^\top, \mu_1 = 1.5, \mu_2 = -1.5, \sigma_1^2 = \sigma_2^2 = p$. Random samples of size n between 10^3 to 10^9 were drawn from the distribution. As this (normal) density function varies greatly from location to location, the problem is more pronounced if density functions (and their estimators) of different dimensions (p) are to be compared with each other. To take this fact into account, for any given estimate $\hat{f}(.)$ we define its standardized mean squared error (MSE) as

$$sMSE = \frac{1}{\#S} \sum_{(s_1,\cdots,s_p)} \{\hat{f}(s_1, ..., s_p) - f(s_1, ..., s_p)\}^2 / \sum_{(s_1,...,s_p)\in S} \{f(s_1, ..., s_p)\}^2$$

where the summation is over all points in R^p with coordinates $(s_1, \cdots, s_p)$, such that $s_k \in S = \{-5 + i/10 : i = 1, ..., 100\}$. The empirical relative efficiency of ONLINE against OFFLINE is defined by the ratio of the estimation errors of the latter against the former. In our calculation, the bandwidths were chosen based on (2.7), such that

$$h_n = cn^{-1/(p+4)}, \quad \tilde{h}_n = \tilde{c}n^{-1/(p+4)} \tag{5.25}$$

with

$$\tilde{c} = c\left(\frac{p+2}{2(p+4)}\right)^{1/(p+1)}.$$

and c set as in (2.6), the optimal choice for the OFFLINE estimator. This should work to the advantage of the OFFLINE estimator. Based on 200 replications, the average of sMASE of the estimators and relative efficiency are listed in Table 1 and Table 2, respectively. For ease of reading, 10 times of square-root of sMAVE is reported in the table. The results basically support our conclusions as shown in Figure 1.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0365

Table 1: Average $sMSE^{1/2} \times 10$ of estimators of density functions and $MSE^{1/2} \times 10$ for the varying coefficient models based on 200 replications

function	method	sample size						
		10^3	10^4	10^5	10^6	10^7	10^8	10^9
$f(\mathbf{x}), p = 1$	OFFLINE	0.7619	0.1225	0.1228	0.0497	0.0203	0.0081	0.0032
	ONLINE	0.7788	0.1262	0.1264	0.0513	0.0210	0.0084	0.0033
$f(\mathbf{x}), p = 2$	OFFLINE	1.3695	0.3088	0.3025	0.1482	0.0702	0.0321	0.0153
	ONLINE	1.3919	0.3200	0.3141	0.1531	0.0729	0.0333	0.0159
$f(\mathbf{x}), p = 4$	OFFLINE	3.1965	1.1961	1.2123	0.6886	0.4052	0.2334	0.1326
	ONLINE	3.2348	1.2296	1.2425	0.7136	0.4230	0.2436	0.1383
$f(\mathbf{x}), p = 10$	OFFLINE	7.9315	5.8123	5.8227	4.7199	3.7517	2.9093	2.2187
	ONLINE	7.9702	5.8840	5.9008	4.7928	3.8270	2.9744	2.2729
$f(\mathbf{x}), p = 20$	ONLINE	9.9223	9.7139	0.0097	0.0095	0.0092	0.0088	0.0084
	OFFLINE	9.9249	9.7233	0.0097	0.0095	0.0092	0.0089	0.0084
$g_0(U)$	OFFLINE	1.3129	0.4527	0.1557	0.1313	0.0922	0.0512	0.0307
	ONLINE	1.3858	0.4711	0.1594	0.1339	0.0946	0.0526	0.0315
$g_1(U)$	OFFLINE	1.1815	0.4752	0.2081	0.0857	0.0431	0.0186	0.0080
	ONLINE	1.3612	0.5174	0.2240	0.0903	0.0453	0.0195	0.0084

Table 2: Empirical relative efficiency of ONLINE estimators against OFFLINE estimators

function	sample size						
	10^3	10^4	10^5	10^6	10^7	10^8	10^9
$f(\mathbf{x}), p = 1$	0.9570	0.9426	0.9429	0.9384	0.9384	0.9435	0.9514
$f(\mathbf{x}), p = 2$	0.9681	0.9311	0.9271	0.9370	0.9277	0.9269	0.9269
$f(\mathbf{x}), p = 4$	0.9765	0.9462	0.9519	0.9312	0.9176	0.9178	0.9187
$f(\mathbf{x}), p = 10$	0.9903	0.9758	0.9737	0.9698	0.9610	0.9567	0.9528
$f(\mathbf{x}), p = 20$	0.9995	0.9981	0.9982	0.9970	0.9951	0.9927	0.9903
$g_0(U)$	0.8976	0.9234	0.9532	0.9622	0.9500	0.9478	0.9512
$g_1(U)$	0.7534	0.8437	0.8632	0.9012	0.9043	0.9101	0.9078

Next, we considered the varying coefficient model

$$y = g_0(U) + g_1(U)X + \varepsilon,$$

where $U \sim$ Uniform$(0, 1)$, $X \sim N(0, 1)$ and independent of U, $g_0(u) = \sin(2\pi u)$, and $g_1(u) = 4(u - 0.5)^2$. For any estimate $\hat{g}_k(.)$, its estimation error is

$$MSE = \frac{1}{11} \sum_{U=0,0.1,\ldots,1} (g_k(U) - \hat{g}_k(U))^2, \quad k = 0, 1;$$

and the empirical relative efficiency of ONLINE against OFFLINE is given by the ratio of their respective estimator errors. Bandwidths were as specified in the previous example.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0365

Based on 200 replications, the average MSE of the estimators and the relative efficiency against OFFLINE N-W estimator are tabulated in Tables 1 and 2. The results are largely in line with our theoretical conclusion. We also find that when the local linear estimator is used, the ONLINE estimator can indeed be more efficient in some cases than the OFFLINE.

Supplementary Materials: Proofs of results in this paper are included in the online supplemental materials.

Acknowledgements: The authors thank Professor Zhiliang Ying and a referee for their thoughtful comments which lead to substantial improvement of the paper. YC Xia's research is partially supported by National Natural Science Foundation of China: 71371095, and MOE grant of Singapore: MOE2014-T2-1-072.

References

Aggarwal, C. Han, J., Wang, J., and Yu, P. (2003) A framework for clustering evolving data streams. In *Proceeding of VLDB* 29, 81-92. VLDB Endowment 2003 table of contents ISBN:0-12-722442-4

Cai, Z., Qian, W., Wei, L., and Zhou, A. (2003) M-kernel merging: towards density estimation over data streams. In *Proceedings of Database Systems for Advanced Applications*.

Chen, X. and Xie, M. (2014) A Split-and-Conquer approach For analysis of extraordinarily large data. *Statistica Sinica* 24, 1655-1684.

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

Fan, J. and Zhang, W. (1999) Statistical estimation in varying coefficient models. *Annals of Statistics* 27, 1491-1518.

Heinz, C. and Seeger, B. (2006) Towards kernel density estimation over streaming data. *Proceedings of the 13th International Conference on Management of Data*. Tata McGraw-Hill Publishing.

Hilbert, M. (2015) Big data for development: a review of promises and challenges. Development Policy Review (forthcoming). Accessible at martinhilbert.net/big-data-for-development.

Kleiner, A., Talwalkar, A., Sarkar, P., and Jordan, M. I. (2014) A scalable bootstrap for massive data. *Journal of the Royal Statistical Society: Series B.* 76, 795-816.

Kristan, M., Leonardis, A., and Skočaj, D. (2011) Multivariate online kernel density estimation with Gaussian kernels. *Pattern Recognition* 44, 2630-2642.

Kristan, M., Skočaj, D., and Leonardis, A. (2010) Online kernel density estimation for interactive learning. *Journal of Image and Vision Computing* 28, 1106-1116.

Lambert, C., Harrington, S., Harvey, C., and Glodjo, A. (1999) Efficient online nonparametric kernel density estimation. *Algorithmica* 25, 37-57.

Liang, F., Cheng, Y., Song, Q., Park, J. and Yang, P. (2013) A resampling-based stochastic approximation method for analysis of large geostatistical data. *Journal of the American Statistical Association* 108, 325-339.

Lin, N. and Xi, R. (2011) Aggregated estimating equation estimation, *Statistics and Its Interface* 4, 73-83.

Ma, P., Mahoney, M. W., and Yu, B. (2013) A statistical perspective on algorithmic leveraging. ArXiv preprint: 1306.5362.

Park, B. U., Mammen, E., Lee, Y. K., and Lee, E. R. (2015) Varying coefficient regression models: a review and new developments. *International Statistical Review* 83: 3664.

Wand, M. P. and Jones, M.C. (1995) *Kernel Smoothing.* London: Chapman and Hall: London.

Wang, C., Chen, M-H, Schifano, E., Wu, J., and Yan, J. (2015) Statistical methods and computing for big data. ArXiv preprint: 1502.07989.