On Asymptotic Properties of Linear Smoothers

Kris De Brabanter, Jos De Brabanter, and Bart De Moor, *Fellow, IEEE*

Abstract

In this paper we generalize some asymptotic results for specific smoothers to that of linear smoothers. Linear smoothers are a general name for the four related paradigms of nonparametric regression i.e., local averaging, local modeling, global modeling (or least squares estimation) and penalized modeling. We develop conditions for asymptotic normality and asymptotic negligibility for these type of estimators. Further, we give uniform and local Berry-Esseen bounds and establish rates of convergence for the kernel regression estimator. Finally, we give a concentration inequality for linear smoothers.

Index Terms

Central limit theorem, linear smoothers, Berry-Esseen bounds, asymptotic negligibility, nonparametric regression, concentration inequality

I. Introduction

ASSUME the data is generated according to

$$Y_k = m(x_k) + e_k, \qquad 1 \leq k \leq n, \tag{1}$$

then the regression function function m is defined as $m(x_k) = \mathbf{E}[Y_k]$. We assume that $\mathbf{E}(e_k) = 0$ and $\mathbf{Var}(e_k) = \sigma_e^2$ for all $k = 1, \ldots, n$. Further, we assume that $x \in \mathbb{R}$ is non-random, e are independent and identically distributed (i.i.d.) random variables and n is fixed. There is an extensive literature describing approximations for the regression function m based on empirical data. Examples are the Nadaraya-Watson (NW) kernel regression estimator, local polynomial regression, splines, orthogonal series estimators, wavelets regression, k-nearest neighbors regression, etc. Asymptotic properties of these smoothers are well understood and asymptotic normality has been established, see [19] and [24] for the NW smoother, [9] for local polynomial regression, [21] for (regression) splines, [1] for series estimators and [15] for k-nearest neighbors regression. All the latter proofs made use of specific properties of the studied smoother. In this paper we will generalize these results to the class of linear smoothers (see Definition 1) which include local averaging, local modeling, global modeling (or least squares estimation) and penalized modeling (see Chapter 2 in [10] and reference therein) and establish conditions under which these type of estimators are asymptotically normal. It will be shown that the weights play a crucial role in the conditions for asymptotic normality as well as in the conditions for asymptotic negligibility. Also in the well-known Berry-Esseen bounds these weights fulfill an important role.

The outline of the paper is as follows: Section II presents results regarding the asymptotic normality for local averaging regression estimators under mild smoothness conditions on the regression function and conditions on the weights. Section III gives an idea of the closeness of the obtained normal approximation and establishes a concentration inequality for linear smoothers. Section IV states some conclusions and gives an outlook on extensions to data-dependent bandwidths or data-dependent smoothing parameters.

Definition 1. *An estimator $\hat{m}_n$ of m is a linear smoother if, for each x, there exists a vector $W(x;\theta_n) = (W_1(x;\theta_n), \ldots, W_n(x;\theta_n))^T$ such that*

$$\hat{m}_n(x) = \sum_{k=1}^{n} W_k(x;\theta_n) Y_k, \tag{2}$$

where the weights $W_k(x;\theta_n) = W_k(x; x_1, \ldots, x_n, \theta_n) \in \mathbb{R}$ are depending on $x_1, \ldots, x_n$ and on θ_n, a nonrandom tuning parameter sequence e.g. bandwidth h_n in kernel methods, smoothing parameter in splines or number of k nearest neighbors.

II. Asymptotic normality of linear smoothers

A. Theoretical results for a general central limit theorem

For notational convenience, we suppress the tuning parameter sequence in the smoother vector. We begin with a lemma which will enable us to calculate the asymptotic distribution for local averaging regression estimators. In what follows, denote by $\|a\|$ the two norm of a vector a.

K. De Brabanter and B. De Moor are with the Department of Electrical Engineering, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium, e-mail: {kris.debrabanter,bart.demoor}@esat.kuleuven.be.

J. De Brabanter is with the Department Industrieel Ingenieur, KaHo Sint Lieven (Associatie KU Leuven), G. Desmetstraat 1, B-9000 Gent, Belgium, e-mail: jos.debrabanter@kahosl.be

Lemma 1. *Let the moment condition* $\mathbf{E}\,|Y|^{2+\delta}$ *with* $\delta > 0$ *be satisfied. Let* $x \in \mathbb{R}$ *be such that m is continuous at the point* x *and put*

$$V_{nk} = W_k(x)(Y_k - m(x)).$$

Then, for

$$\lim_{n\to\infty} \max_{1\le k\le n} \frac{|W_k(x)|^\delta}{\|W(x)\|^\delta} \to 0 \quad \textit{and} \quad \lim_{n\to\infty} \|W(x)\|^{2+\delta} \to 0 \tag{3}$$

it follows that

$$\frac{1}{\|W(x)\|} \sum_{k=1}^{n} (V_{nk} - \mathbf{E}\,V_{nk}) \xrightarrow{d} N(0, \sigma_e^2). \tag{4}$$

Proof: In order to establish (4), by Lyapunov's theorem it suffices to show that (see e.g. p. 30 in [25] and p. 332 in [28])

$$\frac{1}{D_n^{2+\delta}} \sum_{k=1}^{n} \mathbf{E}\,|V_{nk} - \mathbf{E}\,V_{nk}|^{2+\delta} \to 0$$

for $n \to \infty$ with

$$\begin{aligned}
D_n^2 &= \sum_{k=1}^{n} \mathbf{Var}(V_{nk}) && (5)\\
&= \sum_{k=1}^{n} W_k^2(x)\,\mathbf{Var}(Y_k - m(x)) && (6)\\
&= \sum_{k=1}^{n} W_k^2(x)\,\mathbf{Var}(Y_k - m(x_k) + m(x_k) - m(x)) && (7)\\
&= \sigma_e^2 \|W(x)\|^2, && (8)
\end{aligned}$$

since $m(x_k) - m(x)$ is nonrandom. By using the c_r-inequality

$$\begin{aligned}
\frac{1}{D_n^{2+\delta}} \sum_{k=1}^{n} \mathbf{E}\,|V_{nk} - \mathbf{E}\,V_{nk}|^{2+\delta} &\le \frac{2^{2+\delta} \sum_{k=1}^{n} \mathbf{E}\,|V_{nk}|^{2+\delta}}{\sigma_e^{2+\delta} \|W(x)\|^{2+\delta}} && (9)\\
&= \frac{2^{2+\delta} \sum_{k=1}^{n} \mathbf{E}\,|W_k(x)(Y_k - m(x))|^{2+\delta}}{\sigma_e^{2+\delta} \|W(x)\|^{2+\delta}} && (10)\\
&\le \frac{2^{2+\delta} \sum_{k=1}^{n} |W_k(x)|^{2+\delta}\,\mathbf{E}\,|Y_k - m(x)|^{2+\delta}}{\sigma_e^{2+\delta} \|W(x)\|^{2+\delta}} && (11)\\
&\le \frac{2^{2+\delta}\,\mathbf{E}\,|Y - m(x)|^{2+\delta} \max_{1\le k\le n} |W_k(x)|^\delta \|W(x)\|^2}{\sigma_e^{2+\delta} \|W(x)\|^{2+\delta}} && (12)\\
&\to 0 && (13)
\end{aligned}$$

as $n \to \infty$ by assumption and $\mathbf{E}\,|Y|^{2+\delta} < \infty$. The last assumption is needed such that $D_n^{2+\delta} \to 0$ for $n \to \infty$. ■

Next, we show that $\mathbf{E}\,|Y|^{2+\delta}$ with $\delta > 0$ can be further weakened to $\mathbf{E}\,Y^2 < \infty$.

Lemma 2. *Assume* $\mathbf{E}\,Y^2 < \infty$ *and let* $x \in \mathbb{R}$ *be such that m is continuous at the point* x*, then for*

$$\lim_{n\to\infty} \max_{1\le k\le n} \frac{|W_k(x)|^2}{\|W(x)\|^2} \to 0 \tag{14}$$

and

$$\lim_{n\to\infty} \|W(x)\|^2 \to 0 \tag{15}$$

equation (4) *holds.*

Proof: By Lindeberg's condition it suffices that for every $\varepsilon > 0$

$$\frac{1}{D_n^2} \sum_{k=1}^{n} \mathbf{E}\,(V_{nk} - \mathbf{E}\,V_{nk})^2\, I_{(|V_{nk} - \mathbf{E}\,V_{nk}| \ge \varepsilon D_n)} \to 0 \tag{16}$$

for $n \to \infty$. Since,

$$(V_{nk} - \mathbf{E}\,V_{nk})^2 = (W_k(x)(Y_k - m(x)) - W_k(x)(m(x_k) - m(x)))^2 = W_k^2(x) e_k^2 \tag{17}$$

the terms in the sum (16) can be written for any $L > 0$ as

$$\begin{aligned}
\mathbf{E}\left(W_k^2(x)e_k^2 I_{(|W_k(x)e_k|\geq \varepsilon D_n)}\right) &= \mathbf{E}(W_k^2(x)e_k^2(I_{(|e_k|\leq L)} + I_{(|e_k|>L)})I_{(|W_k(x)e_k|\geq \varepsilon D_n)})) && (18)\\
&\leq W_k^2(x)L^2\,\mathbf{E}(I_{(|W_k(x)e_k|\geq \varepsilon D_n)}) + W_k^2(x)\,\mathbf{E}(e_k^2\cdot I_{(|e_k|>L)}\cdot I_{(|W_k(x)e_k|\geq \varepsilon D_n)}) && (19)\\
&\leq \frac{W_k^2(x)L^2}{\varepsilon^2 D_n^2}\sigma_e^2 W_k^2(x) + W_k^2(x)\,\mathbf{E}(e_k^2\cdot I_{(|e_k|>L)}\cdot I_{(|W_k(x)e_k|\geq \varepsilon D_n)}), && (20)
\end{aligned}$$

by Chebyshev's inequality. Hence,

$$\frac{1}{D_n^2}\sum_{k=1}^n \mathbf{E}\left((W_k(x)e_k)^2 I_{(|W_k(x)e_k|\geq \varepsilon D_n)}\right) \leq \frac{L^2}{\varepsilon^2\sigma_e^2}\max_{1\leq k\leq n}\frac{W_k^2(x)}{\|W(x)\|^2} + \frac{1}{\sigma_e^2}\max_{1\leq k\leq n}\mathbf{E}(e_k^2\cdot I_{(|e_k|>L)}\cdot I_{(|W_k(x)e_k|\geq \varepsilon D_n)}) \quad (21)$$

By fixing L and under assumption (14),

$$\limsup_{n\to\infty}\frac{1}{D_n^2}\sum_{k=1}^n \mathbf{E}\left((W_k(x)e_k)^2 I_{(|W_k(x)e_k|\geq \varepsilon D_n)}\right) \leq \frac{1}{\sigma_e^2}\max_{1\leq k\leq n}\mathbf{E}(e_k^2\cdot I_{(|e_k|>L)}\cdot I_{(|W_k(x)e_k|\geq \varepsilon D_n)}). \quad (22)$$

By letting $L\to\infty$, the second term on the right hand side goes to zero since $EY^2 < \infty$. Again, assumption (15) is needed such that $D_n^2 \to 0$ for $n\to\infty$. ■

Remark 1. *Note that in both proofs the bandwidth sequence h_n (or smoothing parameter sequence) is assumed nonrandom and in particular is independent of the responses Y_k. However, in practice the bandwidths are selected via cross-validation or plug-in procedures. Hence, nonparametric estimators are usually constructed using stochastic bandwidths. We refer the reader to [16] for recent developments.*

Remark 2. *Note that in the context of quadratic forms, conditions of the form*

$$\max_{1\leq k\leq n}\frac{|W_k(x)|^2}{\|W(x)\|^2}, \quad (23)$$

have been studied in [26] and [4]. They show that condition (23) *goes to zero, for $n\to\infty$, if the largest eigenvalue of the smoother matrix W i.e., an $n\times n$ matrix with kth row $W(x_k;h_n)^T$, is dominated by all other eigenvalues together. Hence, condition* (23) *for $n\to\infty$ implies that no single component $\xi_k = W_k(x)Y_k$ of $\hat{m}_n(x)$ contributes an excessive proportion to the variance of $\hat{m}_n(x)$. Also, this condition rules out situations where $\hat{m}_n(x)$ depends only on a negligible proportion of the ξ_k's.*

The following technical will be needed to formulate a central limit theorem (CLT) for local averaging regression estimators.

Lemma 3. *Let the assumptions of Lemma 2 be fulfilled and assume in addition that $\sum_{k=1}^n W_k(x) = 1$ and m has continuous derivatives up to order Q in a neighborhood of x. Then*

$$\frac{1}{\|W(x)\|}(\hat{m}_n(x) - m(x)) = \psi_n(x) + \frac{1}{\|W(x)\|}\mu_n(x) \quad (24)$$

where

$$\psi_n(x) \xrightarrow{d} N(0,\sigma_e^2) \quad (25)$$

and

$$\mu_n(x) \xrightarrow{p} \mu(x) \quad (26)$$

with

$$\mu(x) = \sum_{q=1}^Q\sum_{k=1}^n W_k(x)\frac{(x_k - x)^q}{q!}m^{(q)}(x). \quad (27)$$

Proof: By using $\sum_{k=1}^n W_k(x) = 1$, we can make the following decomposition

$$\begin{aligned}
\frac{1}{\|W(x)\|}(\hat{m}_n(x) - m(x)) &= \frac{1}{\|W(x)\|}\sum_{k=1}^n W_k(x)(Y_k - m(x)) && (28)\\
&= \frac{1}{\|W(x)\|}\sum_{k=1}^n (V_{nk} - \mathbf{E}\,V_{nk}) + \frac{1}{\|W(x)\|}\sum_{k=1}^n \mathbf{E}\,V_{nk} && (29)\\
&= \psi_n(x) + \frac{1}{\|W(x)\|}\sum_{k=1}^n \mathbf{E}\,V_{nk}, && (30)
\end{aligned}$$

so that by (4) we only have to show that

$$\sum_{k=1}^{n} \mathbf{E}\, V_{nk} = \sum_{q=1}^{Q} \sum_{k=1}^{n} W_k(x) \frac{(x_k - x)^q}{q!} m^{(q)}(x). \tag{31}$$

As to (31), observe that since $\mu_n(x) \xrightarrow{p} \mu(x)$

$$\mu(x) = \sum_{k=1}^{n} \mathbf{E}\, V_{nk} = \sum_{k=1}^{n} \mathbf{E}(W_k(x)(Y_k - m(x))) \tag{32}$$

$$= \sum_{k=1}^{n} W_k(x) m(x_k) - m(x), \tag{33}$$

since we assumed that $\sum_{k=1}^{n} W_k(x) = 1$. Now, let m be Q times continuously differentiable, then a Taylor series around the fitting point x yields

$$\mu(x) = \sum_{q=1}^{Q} \sum_{k=1}^{n} W_k(x) \frac{(x_k - x)^q}{q!} m^{(q)}(x). \tag{34}$$

■

Theorem 1 (Central limit theorem for linear smoothers at a given point x)**.** *Let the assumptions or Lemma 3 be fulfilled and assume* $\mathbf{E}\, Y^2 < \infty$*. Then for*

$$\lim_{n\to\infty} \max_{1\le k\le n} \frac{|W_k(x)|^2}{\|W(x)\|^2} \to 0 \quad \textit{and} \quad \lim_{n\to\infty} \|W(x)\|^2 \to 0 \tag{35}$$

it follows that

$$\frac{1}{\|W(x)\|}(\hat{m}_n(x) - m(x)) \xrightarrow{d} N\left(\frac{1}{\|W(x)\|} \cdot \mu(x), \sigma_e^2\right) \tag{36}$$

with $\mu(x)$ *defined as in* (27)*.*

Proof: Follows from (24), (25) and (26). ■

In general, the two conditions on the weights in Theorem 1 are needed. However, in case of probability weights, i.e. the weights are nonnegative and sum up to 1, the condition

$$\lim_{n\to\infty} \max_{1\le k\le n} \frac{|W_k(x)|^2}{\|W(x)\|^2} \to 0 \tag{37}$$

implies

$$\lim_{n\to\infty} \|W(x)\|^2 \to 0. \tag{38}$$

This can be seen as follows:

$$\|W(x)\|^2 = \frac{\|W(x)\|^4}{\|W(x)\|^2} = \frac{\left(\sum_{k=1}^{n} W_k^2(x)\right)^2}{\|W(x)\|^2} \le \frac{\left(\max_{1\le k\le n} W_k(x)\right)^2}{\|W(x)\|^2} = \frac{\max_{1\le k\le n} W_k^2(x)}{\|W(x)\|^2}. \tag{39}$$

Letting $n \to \infty$ yields the result.

To use (36) in practice, several quantities need to be estimated from the data, i.e. the bias and the error variance. Consider the following decomposition

$$\frac{\hat{m}_n(x) - m(x)}{\hat{\sigma}_e \|W(x)\|} = \left(\frac{\hat{m}_n(x) - m_n(x)}{\sigma_e \|W(x)\|}\right) \frac{\sigma_e}{\hat{\sigma}_e}. \tag{40}$$

It is clear that besides a consistent bias estimator (see Lemma 3) one also needs an error variance estimator such that $\hat{\sigma}_e^2 \xrightarrow{p} \sigma_e^2$. A $\sqrt{n}$-consistent variance estimator can be found in [11]. As for the bias estimation, consistent estimators can be found in [7], [8], [13] and [2].

B. Some applications of the general central limit theorem

Next, we illustrate the application of Theorem 1 on two well-known nonparametric regression smoothers.

Definition 2 (Nadaraya-Watson (NW) regression estimator)**.** *The Nadaraya-Watson (NW) kernel smoother [19], [30] is defined as*

$$\hat{m}_n(x) = \sum_{k=1}^{n} W_k(x) Y_k \quad \text{with} \quad W_k(x) = \frac{K\left(\frac{x - x_k}{h_n}\right)}{\sum_{k=1}^{n} K\left(\frac{x - x_k}{h_n}\right)} \tag{41}$$

with $K : \mathbb{R} \to \mathbb{R}^+$ a kernel function decreasing on $[0, \infty]$ with bandwidth $h_n > 0$.

Definition 3 (k-nearest neighbors for regression)**.** *In k- neareast neighbors (k-NN) one determines the k nearest x_i's to x in terms of the distance $\|x - x_i\|$ and estimates $m(x)$ by the average of the corresponding Y_i's. More precisely, for $x \in \mathbb{R}$, let*

$$(x_{(1)}(x), Y_{(1)}(x)), \ldots, (x_{(n)}(x), Y_{(n)}(x)) \tag{42}$$

be a permutation of $(x_1(x), Y_1(x)), \ldots, (x_n(x), Y_n(x))$ such that

$$\|x - x_{(1)}(x)\| \leq \ldots \leq \|x - x_{(n)}(x)\|. \tag{43}$$

The k-NN estimate is defined by

$$\hat{m}_n(x) = \frac{1}{k_n} \sum_{i=1}^{k_n} Y_{(i)}(x) \tag{44}$$

Proposition 1 (Central limit theorem for Nadaraya-Watson regression estimator at a given point x)**.** *Assume $\mathbf{E}\, Y^2 < \infty$ and that the following conditions on the kernel K are fulfilled: $\int K(u)\, du = 1$, $\int uK(u)\, du = 0$, $\int u^2 K(u)\, du < \infty$, $\int |K(u)|\, du < \infty$, $\lim_{|u|\to\infty} |uK(u)| = 0$ and $\sup_u |K(u)| < \infty$. Further assume that $h_n \to 0$ as $n \to \infty$ such that $nh_n \to \infty$. Let m be twice continuously differentiable in a fixed point x and let*

$$nh_n^5 \to c^2 \tag{45}$$

for some $0 \leq c < \infty$. Then

$$\sqrt{nh_n}(\hat{m}_n(x) - m(x)) \xrightarrow{d} N\left(c \cdot \frac{m''(x)}{2} \int u^2 K(u)\, du, \sigma_e^2 \Delta \int K^2(u)\, du\right) \tag{46}$$

under the assumption of fixed design with Δ the range of the data.

Proof: In order to establish asymptotic normality we need to check the conditions of Theorem 1 which are valid if $\mathbf{E}\, Y^2 < \infty$. From the definition of the NW estimator it is immediately clear that $\sum_{k=1}^n W_k(x) = 1$.

A simple calculation shows

$$\max_{1\leq k\leq n} \frac{|W_k(x)|^2}{\|W(x)\|^2} = \frac{A^2}{\sum_{k=1}^n K^2\left(\frac{x-x_k}{h_n}\right)}, \tag{47}$$

where $A = \sup_u |K(u)|$. For large n, the sum

$$\sum_{k=1}^n K^2\left(\frac{x - x_k}{h_n}\right) \tag{48}$$

can be replaced by (see e.g. Chapter 3 in [12] or p. 38 in [14])

$$\frac{nh_n}{\Delta} \int K^2(u)\, du + o(nh_n), \tag{49}$$

with Δ the range of the data. Consequently,

$$\max_{1\leq k\leq n} \frac{|W_k(x)|^2}{\|W(x)\|^2} = \frac{A^2\Delta}{nh_n \int K^2(u)\, du + o(nh_n)} \tag{50}$$

and hence

$$\lim_{n\to\infty} \max_{1\leq k\leq n} \frac{|W_k(x)|^2}{\|W(x)\|^2} \to 0 \tag{51}$$

is satisfied for $n \to \infty$ such that $nh_n \to \infty$. Since the NW weights are probability weights condition $\lim_{n\to\infty} \|W(x)\|^2 \to 0$ is also verified.

In order to obtain consistent bias estimates, we further require the consistency of the kernel density estimator, see [22]. This is valid for $h_n \to 0$ as $n \to \infty$ such that $nh_n \to \infty$ and the conditions on the kernel K in the proposition. Next, the factor $1/\|W(x)\|$ can be written as

$$\frac{\sqrt{nh_n} + o(1)}{\sqrt{\Delta}\sqrt{\int K^2(u)\, du + o(1)}}. \tag{52}$$

Second, we assume that m has continuous derivatives up to order 2. Then, (27) results in

$$\mu(x) = \frac{m'(x)}{nh_n f(x)} \sum_{k=1}^n K\left(\frac{x - x_k}{h_n}\right)(x_k - x) + \frac{m''(x)}{2nh_n f(x)} \sum_{k=1}^n K\left(\frac{x - x_k}{h_n}\right)(x_k - x)^2 \tag{53}$$

because of (26) and Cramér-Slutsky and for a regular design generated by a known design density f. A special case of

$f(x) = 1/\Delta$ is when considering equispaced design. In order to obtain a more workable form of μ, the sums have to be replaced by integrals. By using elements from [14] (p. 38), the following holds as $nh_n \to \infty$

$$\mu(x) = h_n^2 \frac{m'(x)f'(x)}{f(x)} \int u^2 K(u)\, du + h_n^2 \frac{m''(x)}{2} \int u^2 K(u)\, du + o(h_n^2) = h_n^2 \frac{m''(x)}{2} \int u^2 K(u)\, du + o(h_n^2) \tag{54}$$

for $\int u^2 K(u)\, du < \infty$. Note that the first term is zero under the assumptions of fixed design. ■

Proposition 2 (Central limit theorem for k-NN for regression at a given point x)**.** *Assume $\mathbf{E}\, Y^2 < \infty$. Further assume that $k \to \infty$ as $n \to \infty$ such that $k_n/n \to 0$. Let m be twice continuously differentiable in a fixed point x. and let*

$$\frac{k_n^5}{n^4} \to c^2 \tag{55}$$

for some $0 \leq c < \infty$. Then

$$\sqrt{k_n}(\hat{m}_n(x) - m(x)) \xrightarrow{d} N\left(c \cdot \frac{m''(x)\Delta^2}{24}, \sigma_e^2\right) \tag{56}$$

under the assumption of fixed design with Δ the range of the data.

Proof: As before, we assume that $\mathbf{E}\, Y^2 < \infty$. Since $W_k(x) = \frac{1}{k_n}$ if x_i is among the k nearest neighbors of x and zero otherwise it follows that $\sum_{k=1}^n W_k(x) = 1$. Further,

$$\max_{1 \leq k \leq n} \frac{|W_k(x)|^2}{\|W(x)\|^2} = \frac{1/k_n^2}{1/k_n} \tag{57}$$

and the first condition for asymptotic normality is satisfied for $k_n \to \infty$ as $n \to \infty$. Since the k-NN weights are probability weights condition $\lim_{n\to\infty} \|W(x)\|^2 \to 0$ is also verified.

To obtain a bias estimate we can work with order statistics [15] or use (27). In this paper we choose the latter to proceed. In this case, we have to slightly rewrite the k-NN estimator since the weights are not $1/k_n$ everywhere. Note that we can rewrite the k-NN as follows

$$\hat{m}_n(x) = \frac{1}{n} \sum_{i=1}^{k_n} W_i(x) Y_i \tag{58}$$

where the weight sequence $W_i(x)$ is defined through the set of indexes

$$J_x = \{i : x_i \text{ is one of the } k \text{ nearest observations to } x\}. \tag{59}$$

With this set of indexes of neighboring observations, the k-NN weight sequence yields $W_i(x) = n/k_n$ if $i \in J_x$ and zero otherwise. Let $k_n = 2nh_n/\Delta$ and $i \in J_x$, then for an x not too close to the boundary and in an equispaced design setting

$$W_i(x) = \frac{n}{k_n} = \frac{\Delta}{2h_n} = \frac{n}{\sum_{i=1}^n I_{\{|x_i - x| \leq h_n\}}} = n \frac{I_{\{|x_i - x| \leq h_n\}}}{\sum_{i=1}^n I_{\{|x_i - x| \leq h_n\}}}. \tag{60}$$

In other words, for this design k-NN weights are equivalent to kernel weights with a uniform kernel $K(u) = \frac{1}{2} I_{\{|u| \leq 1\}}$. Hence, (58) can be written as

$$\hat{m}_n(x) = \sum_{i=1}^n \frac{I_{\{|x_i - x| \leq h_n\}} Y_i}{\sum_{i=1}^n I_{\{|x_i - x| \leq h_n\}}}. \tag{61}$$

In fact, this the NW regression estimator for a uniform kernel K. Therefore, the bias for the k-NN in fixed design immediately yields

$$\mu(x) = \left(\frac{k_n}{n}\right)^2 \frac{m''(x)\Delta^2}{24} + o\left(\frac{k_n}{n}\right)^2. \tag{62}$$

■

The two preceding results regarding asymptotic normality for NW and k-NN for regression improve the classical Theorem 1.3 in [20] and Theorem 3 in [15]. For second order kernels, we cover the case $nh_n^5 \to c^2 > 0$, where [20] assumes $nh_n^5 \to 0$ which is of particular interest since it is well-known that the optimal bandwidth for estimating m at a fixed point x is of order $n^{-1/5}$. In case of k-NN we also cover the case $k_n^5/n^4 \to c^2 > 0$.

C. Asymptotic negligibility condition

An immediate consequence of satisfying the Lindeberg condition is that the uniform asymptotic negligibility condition for linear smoothers is satisfied. In order to prove Theorem 2 we first need the following lemma.

Lemma 4. *If*

$$\lim_{n\to\infty}\max_{1\le k\le n}\frac{|W_k(x)|^2}{\|W(x)\|^2}\to 0 \quad \textit{and} \quad \lim_{n\to\infty}\|W(x)\|^2\to 0 \tag{63}$$

then

$$\lim_{n\to\infty}\max_{1\le k\le n}\frac{\sigma_e^2 W_k^2(x)}{D_n^2}\to 0. \tag{64}$$

Proof: Let $\mathcal{A}=\{z:|z-\mathbf{E}(W_k(x)Y_k)|\ge\varepsilon D_n\}$*, then for* $1\le k\le n$

$$\sigma_e^2 W_k^2(x)\le\int_{\mathcal{A}}(z-\mathbf{E}(W_k(x)Y_k))^2\,dF_k(z)+\varepsilon^2 D_n^2, \tag{65}$$

with $F_k=F_k(z)$ *the distribution function of the random variable* $W_k(x)Y_k$*. Hence,*

$$\max_{1\le k\le n}\frac{\sigma_e^2 W_k^2(x)}{D_n^2}\le\frac{1}{D_n^2}\sum_{k=1}^{n}\int_{\mathcal{A}}(z-\mathbf{E}(W_k(x)Y_k))^2\,dF_k(z)+\varepsilon^2. \tag{66}$$

Since the first term on the right hand side goes to zero for $n\to\infty$ *(Lindeberg condition is satisfied) and the inequality holds for all* $\varepsilon>0$*.* ■

The following theorem is also related to infinitely divisible and stable distributions, see e.g. [28].

Theorem 2. *If*

$$\lim_{n\to\infty}\max_{1\le k\le n}\frac{|W_k(x)|^2}{\|W(x)\|^2}\to 0 \quad \textit{and} \quad \lim_{n\to\infty}\|W(x)\|^2\to 0 \tag{67}$$

then

$$\lim_{n\to\infty}\max_{1\le k\le n}\mathbf{P}\left\{|W_k(x)Y_k-\mathbf{E}(W_k(x)Y_k)|>\varepsilon D_n\right\}\to 0 \tag{68}$$

for each $\varepsilon>0$ *with* $D_n^2=\sigma_e^2\sum_{k=1}^{n}W_k^2(x)$*.*

Proof: By using Chebyshev's inequality it follows that

$$\mathbf{P}\left\{|W_k(x)Y_k-\mathbf{E}(W_k(x)Y_k)|>\varepsilon D_n\right\}\le\frac{\mathbf{Var}(W_k(x)Y_k)}{\varepsilon^2 D_n^2} \tag{69}$$

and hence

$$\begin{aligned}\max_{1\le k\le n}\mathbf{P}\left\{|W_k(x)Y_k-\mathbf{E}(W_k(x)Y_k)|>\varepsilon D_n\right\}\\ \le\frac{1}{\varepsilon^2}\max_{1\le k\le n}\frac{\sigma_e^2 W_k^2(x)}{D_n^2}.\end{aligned} \tag{70}$$

The theorem is proven by taking the limit on both sides and applying Lemma 4. ■

III. Berry-Esseen bounds and a concentration inequality for linear smoothers

A. Uniform and local Berry-Esseen bounds for linear smoothers

Some idea of the closeness of the normal approximation can be obtained from the Berry-Esseen bound [3], [6] to the case of independent random variables for triangular arrays, see e.g. [25]. The sharpest upper bound for the universal constant in the Berry-Esseen inequality for independent random variables is currently due to [27].

Theorem 3. *Let* $\xi_1,\ldots,\xi_n$*, with* $\xi_k=W_k(x)Y_k$*, be a sequence of independent random variables with finite second moments. Let* $\hat{m}_n(x)=\xi_1+\ldots+\xi_n$ *and* $\mathbf{E}(e^2)=\sigma_e^2$*. Further, assume that* $\mathbf{E}\,|e|^3<\infty$*. Then,*

$$\begin{aligned}&\sup_{-\infty\le a\le\infty}\left|\mathbf{P}\left\{\frac{\hat{m}_n(x)-\mathbf{E}\,\hat{m}_n(x)}{\sqrt{\mathbf{Var}(\hat{m}_n(x))}}\le a\right\}-\Phi(a)\right|\\ &\le C\frac{\mathbf{E}\,|e|^3}{\sigma_e^3}\max_{1\le k\le n}\frac{|W_k(x)|}{\|W(x)\|}\end{aligned} \tag{71}$$

where $\Phi(\cdot)$ *denotes the standard normal distribution function and* $0.4097<C\le 0.56$*.*

Proof: Using the Berry-Esseen inequality yields

$$\begin{aligned}&\sup_{-\infty\le a\le\infty}\left|\mathbf{P}\left\{\frac{\hat{m}_n(x)-\mathbf{E}\,\hat{m}_n(x)}{\sqrt{\mathbf{Var}(\hat{m}_n(x))}}\le a\right\}-\Phi(a)\right|\\&\le C\frac{\sum_{k=1}^n\mathbf{E}\,|\xi_k-\mathbf{E}\,\xi_k|^3}{\left\{\sum_{k=1}^n\mathbf{Var}(\xi_k)\right\}^{3/2}},\end{aligned}\tag{72}$$

for some universal constant C and $\xi_k = W_k(x)Y_k$. Hence,

$$\begin{aligned}\mathbf{E}\,|\xi_k-\mathbf{E}\,\xi_k|^3 &= \mathbf{E}\,|W_k(x)Y_k-W_k(x)m(x_k)|^3 &(73)\\&= \mathbf{E}\,|W_k(x)(Y_k-m(x_k))|^3 &(74)\\&= |W_k(x)|^3\,\mathbf{E}\,|e_k|^3, &(75)\end{aligned}$$

and

$$\begin{aligned}\sum_{k=1}^n\mathbf{E}\,|\xi_k-\mathbf{E}\,\xi_k|^3 &= \sum_{k=1}^n|W_k(x)|^3\,\mathbf{E}\,|e_k|^3 &(76)\\&= \mathbf{E}\,|e|^3\sum_{k=1}^n|W_k(x)|^3. &(77)\end{aligned}$$

The last term can be upper bounded by

$$\begin{aligned}\sum_{k=1}^n|W_k(x)|^3 &= \sum_{k=1}^n|W_k(x)|\,|W_k(x)|^2 &(78)\\&\le \max_{1\le k\le n}|W_k(x)|\sum_{k=1}^n|W_k(x)|^2 &(79)\\&= \max_{1\le k\le n}|W_k(x)|\;\|W(x)\|^2. &(80)\end{aligned}$$

Finally,

$$\begin{aligned}&\sup_{-\infty\le a\le\infty}\left|\mathbf{P}\left\{\frac{\hat{m}_n(x)-\mathbf{E}\,\hat{m}_n(x)}{\sqrt{\mathbf{Var}(\hat{m}_n(x))}}\le a\right\}-\Phi(a)\right| &(81)\\\le\;& C\frac{\max_{1\le k\le n}|W_k(x)|\;\|W(x)\|^2\,\mathbf{E}\,|e|^3}{\left\{\sum_{k=1}^n\sigma_e^2W_k^2(x)\right\}^{3/2}} &(82)\\\le\;& C\frac{\max_{1\le k\le n}|W_k(x)|\;\|W(x)\|^2\,\mathbf{E}\,|e|^3}{\sigma_e^3\|W(x)\|^3} &(83)\\=\;& C\frac{\mathbf{E}\,|e|^3}{\sigma_e^3}\max_{1\le k\le n}\frac{|W_k(x)|}{\|W(x)\|}. &(84)\end{aligned}$$

■

An immediate consequence of Theorem 3 is stated as follows for the NW smoother.

Corollary 1. *Assume the conditions of Theorem 1 and Theorem 3. Then, for the optimal asymptotic order of the bandwidth for NW in mean integrated squared error sense (see p.141 in [29]) $h_n = O(n^{-1/5})$,*

$$\sup_{-\infty\le a\le\infty}\left|\mathbf{P}\left\{\frac{\hat{m}_n(x)-\mathbf{E}\,\hat{m}_n(x)}{\sqrt{\mathbf{Var}(\hat{m}_n(x))}}\le a\right\}-\Phi(a)\right| = O(n^{-2/5}),\tag{85}$$

for the NW kernel smoother.

Although the bound given by (71) is valid for all $a\in\mathbb{R}$, it is only useful for values a near the center of the distribution. The reason is that, for a sufficiently large, the difference $\mathbf{P}\left(\frac{\hat{m}_n(x)-\mathbf{E}(\hat{m}_n(x))}{\sqrt{\mathbf{Var}(\hat{m}_n(x))}}\le a\right)-\Phi(a)$ becomes so close to zero that the bound (71) is too crude to be of any practical use. One way to refine the Berry-Esseen bound is to use local bounds which reflect the dependence on a as well as the smoother vector $W(x)$ and hence on n.

Theorem 4. *Assume the conditions of Theorem 3 and* $\mathbf{E}\,|e|^{2+\delta} < \infty$ *and* $\delta > 0$*. Then,*

$$\left| \mathbf{P}\left\{ \frac{\hat{m}_n(x) - \mathbf{E}\,\hat{m}_n(x)}{\sqrt{\mathbf{Var}(\hat{m}_n(x))}} \leq a \right\} - \Phi(a) \right| \leq \frac{D}{1+|a|^{2+\delta}} \frac{\mathbf{E}\,|e|^{2+\delta}}{\sigma_e^{2+\delta}} \max_{1\leq k\leq n} \frac{|W_k(x)|^\delta}{\|W(x)\|^\delta}, \tag{86}$$

with $0 < D < \infty$.

Proof: By using elements from [23] (Theorem 9, p. 121) and [18] in addition to

$$\sum_{k=1}^{n} \mathbf{E}\,|\xi_k - \mathbf{E}\,\xi_k|^{2+\delta} \leq \mathbf{E}\,|e|^{2+\delta} \max_{1\leq k\leq n} |W_k(x)|^\delta \|W(x)\|^2 \tag{87}$$

and

$$\left(\sum_{k=1}^{n} \mathbf{Var}(\xi_k) \right)^{1+\delta/2} = \sigma_e^{2+\delta}\ \|W(x)\|^{2+\delta}, \tag{88}$$

with $\xi_k = W_k(x)Y_k$, completes the proof. ■

B. Concentration inequality for linear smoothers

To conclude, we determine a distribution-free bound for the difference between a random variable and its expected value. A popular inequality for this purpose is Hoeffding's inequality (see e.g. [5]). However, a more elegant way of showing such a bound can be done via the bounded difference inequality [17] which is an extension to Hoeffding's inequality. We need to show that for $g(\xi_1, \ldots, \xi_n) = \sum_{k=1}^n \xi_k = \sum_{k=1}^n W_k(x)Y_k$, the bounded difference assumption

$$\sup_{\xi_1,\ldots,\xi_n,\xi_1'\in\mathcal{A}} |g(\xi_1,\ldots,\xi_n) - g(\xi_1,\ldots,\xi_{k-1},\xi_k',\xi_{k+1},\ldots,\xi_n)| \leq c_i \quad 1 \leq k \leq n, \tag{89}$$

where ξ_k' represents the change in the kth variable of g, is valid for some set $\mathcal{A}$ and a function $g : \mathcal{A}^n \to \mathbb{R}$. It now follows that

$$\begin{aligned} |g(\xi_1,\ldots,\xi_n) - g(\xi_1,\ldots,\xi_{k-1},\xi_k',\xi_{k+1},\ldots,\xi_n)| &= |\xi_k - \xi_k'| & (90)\\ &\leq 2M|W_k(x)| & (91)\end{aligned}$$

with $|Y| \leq M$. According to McDiarmid's bounded difference inequality [17], we have that

$$\begin{aligned} \mathbf{P}\left\{ \left| \sum_{k=1}^{n} W_k(x)Y_k - \mathbf{E}\sum_{k=1}^{n} W_k(x)Y_k \right| \geq \varepsilon \right\} &= \mathbf{P}\left\{ |\hat{m}_n(x) - \mathbf{E}\,\hat{m}_n(x)| \geq \varepsilon \right\} & (92)\\ &\leq 2\exp\left(-\frac{\varepsilon^2}{2M^2\|W(x)\|^2} \right). & (93)\end{aligned}$$

IV. Conclusion and outlook

In this paper we considered several asymptotic properties for the class of linear smoothers. We formulated a general central limit theorem for this broad class of smoothers and illustrated its potential on the Nadaraya-Watson kernel regression estimator and the k-nearest neighbor regression estimator.

Note that all results have been obtained for nonrandom bandwidths $0 < h_n \to 0$ (or smoothing parameters) as the sample size $n \to \infty$. In order to improve the estimators' finite sample performance, bandwidths (or smoothing parameters) are selected by means of data-driven methods. As such, bandwidths (or smoothing parameters) are generally stochastic. Therefore, it is desirable to obtain the aforementioned asymptotic results when for example the bandwidth h_n is data dependent and consequently stochastic. It would be interesting to investigate whether Theorem 1 remains valid with the following extra assumption $\hat{h}_n/h_n - 1 = o_p(1)$ accounting for stochastic bandwidths, where $\hat{h}_n$ denotes the bandwidth obtained by a data-driven tuning procedure.

V. Acknowledgments

Kris De Brabanter is a postdoctoral researcher supported by an FWO fellowship grant. We acknowledge stimulating discussions with Luc Devroye (travel grant supported by Academische Stichting Leuven) and with Gérard Biau. We also thank the editor and the three anonymous reviewers for useful suggestions.

REFERENCES

[1] D.W.K. Andrews, "Asymptotic normality of series estimators for nonparametric and semiparametric regression models", *Econometrica*, vol. 59, no.2, pp. 307–345, 1991.

[2] J. Beran, Y. Feng, and S. Heiler. "Modifying the double smoothing bandwidth selector in nonparametric regression", *Stat. Methodol.*, vol. 6, no. 5, pp. 447-465, 2009.

[3] A.C. Berry, "The accuracy of the Gaussian approximation to the sum of independent variables", *Trans. Amer. Math. Soc.*, vol. 49, no. 1, pp. 122–136, 1941.

[4] R.J. Bhansali, L. Giraitis and P.S. Kokoskza, "Convergence of quadratic forms with nonvanishing diagonal", *Statist. Probab. Lett.*, vol. 77, no. 7, pp. 726–734, 2006.

[5] L. Devroye and G. Lugosi, *Combinatorial methods in density estimation*, Springer-Verlag, 2001.

[6] C.G. Esseen, "On the Lyapounov limit of error in the theory of probability", *Ark. Mat. Astr. Fy.*, vol. 28A, pp. 1–19, 1942.

[7] R.L. Eubank and P.L. Speckman, "Confidence bands in nonparametric regression", *J. Amer. Statist. Assoc.*, vol. 88, no. 424, pp. 1287-1301, 1993.

[8] J. Fan and I. Gijbels, "Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation", *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 57, no. 2, pp. 371–394, 1995.

[9] J. Fan and I. Gijbels, *Local polynomial modelling and its applications*, John Wiley & Sons, 1996.

[10] L. Györfi, M. Kohler, A. Krzyżak and H. Walk, *A Distribution-Free Theory of Nonparametric Regression*, Springer, 2002.

[11] P. Hall, J. Kay and D. Titterington, "Asymptotically optimal difference based estimation of variance in nonparametric regression", *Biometrika*, vol. 77, pp. 521-528, 1990.

[12] W. Härdle, *Applied Nonparametric Regression*, Cambridge University Press, 1990.

[13] W. Härdle, P. Hall and J.S. Marron, "Regression smoothing parameters that are not far from their optimum", *J. Amer. Statist. Assoc.*, vol. 87, no. 417, pp. 227-233, 1992.

[14] C. Loader, *Local Regression and Likelihood*, Springer, 1999.

[15] Y.P. Mack, "Local properties of k-NN regression estimates", *SIAM J. Alg. Disc. Meth.*, vol. 2, no. 3, pp. 311–323, 1981.

[16] C. Martins-Filho and P. Saraiva, "On asymptotic normality of the local polynomial regression estimator with stochastic bandwidths", *Comm. Statist. Theory Methods*, vol. 41, no. 6, pp. 1052–1068, 2012.

[17] C. McDiarmid, "On the method of bounded differences", *Surveys in Combinatorics*, Cambridge University Press, pp. 148–188, 1989.

[18] R. Michel, "On the constant in the nonuniform version of the Berry-Esseen theorem", *Z. Wahr. Verw. Geb.*, vol. 55, no. 1, pp. 109–117, 1981.

[19] E.A. Nadaraya, "On estimating regression", *Theory Probab. Appl.*, vol. 9, no. 1, pp. 141–142, 1964.

[20] E.A. Nadaraya, *Nonparametric Estimation of Probability Densities and Regression Curves*, Kluwer Academic Publishers, 1989.

[21] W.K. Newey, "Convergence rates and asymptotic normality for series estimators", *J. Econometrics*, vol. 79, pp. 147–168, 1997.

[22] E. Parzen, "On estimation of a probability density function and mode", *Ann. Math. Statist.*, vol. 33, pp. 1065–1076, 1962.

[23] V.V. Petrov, *Sums of Independent Random Variables*, Springer-Verlag (Translation from Russian), 1975.

[24] E.F. Schuster, ''Joint asymptotic distribution of the estimated regression function at a finite number of distinct points", *Ann. of Math. Stat.*, vol. 43, no. 1, pp. 84–88, 1972.

[25] R.J. Serfling, *Approximation theorems of mathematical statistics*, John Wiley & Sons, 1980.

[26] J. Shao, *Mathematical statistics*, 2nd ed., Springer, 2003.

[27] I.G. Shevtsova, "An improvement of convergence rate estimates in the Lyapunov theorem", *Dokl. Math.*, vol. 82, no. 3, pp. 862–864, 2010.

[28] A.N. Shiryaev, *Probability*, 2nd ed., Springer, 1996.

[29] J.S. Simonoff, *Smoothing Methods in Statistics*, Springer, 1996.

[30] G.S. Watson, "Smooth regression analysis", *Sankhy$\bar{a}$ Ser. A*, vol. 26, no. 4, pp. 359–372, 1964.