Oberwolfach
Preprints

OWP 2012 - 08

LÁSZLÓ GYÖRFI, HARRO WALK

Rate of Convergence of the Density Estimation of Regression Residual

Mathematisches Forschungsinstitut Oberwolfach gGmbH
Oberwolfach Preprints (OWP) ISSN 1864-7596

Imprint:

Mathematisches Forschungsinstitut Oberwolfach gGmbH (MFO)
Schwarzwaldstrasse 9-11
77709 Oberwolfach-Walke
Germany

Tel   +49 7834 979 50
Fax   +49 7834 979 55
Email   admin@mfo.de
URL   www.mfo.de

The Oberwolfach Preprints (OWP, ISSN 1864-7596) are published by the MFO.

Statistics & Risk Modeling 29, 1–17 (2012)

Rate of convergence of the density estimation of regression residual

László Györfi, Harro Walk

Received: Month-1 99, 2003; Accepted: Month-2 99, 2004

Summary: Consider the regression problem with a response variable Y and with a d-dimensional feature vector X. For the regression function $m(x) = \mathbb{E}\{Y|X = x\}$, this paper investigates methods for estimating the density of the residual $Y - m(X)$ from independent and identically distributed data. If the density is twice differentiable and has compact support then we bound the rate of convergence of the kernel density estimate. It turns out that for $d \leq 3$ and for partitioning regression estimates, the regression estimation error has no influence in the rate of convergence of the density estimate.

1 Introduction

Let Y be a real valued random variable and let $X = (X^{(1)}, \dots, X^{(d)})$ be a d-dimensional random vector. The coordinates of X may have different types of distributions, some of them may be discrete (for example binary), others may be absolutely continuous. In the sequel we do not assume anything about the distribution of X. The task of regression analysis is to estimate Y given X, i.e., one aims to find a function F defined on the range of X such that $F(X)$ is "close" to Y. Typically, closeness is measured in terms of the *mean squared error* of F,

$$\mathbb{E}\{(F(X) - Y)^2\}.$$

It is well-known that the mean squared error is minimized by the regression function m with

$$m(x) = \mathbb{E}\{Y \mid X = x\} \tag{1.1}$$

and a minimum mean squared error is

$$L^* := \mathbb{E}\{(Y - m(X))^2\} = \min_F \mathbb{E}\{(Y - F(X))^2\},$$

since, for each measurable function F, the mean squared error can be decomposed into

$$\mathbb{E}\{(F(X) - Y)^2\} = \mathbb{E}\{(m(X) - Y)^2\} + \int_{\mathbb{R}^d} (m(x) - F(x))^2 \mu(dx),$$

AMS 2000 subject classifications. Primary: 62G08; Secondary: 62G09.
Key words and phrases: regression residual, kernel density estimation, partitioning, kernel and nearest neighbor regression estimation, rate of convergence.

where μ denotes the distribution of X. The second term on the right hand side is called *excess error* or integrated squared error of the function F. Clearly, the mean squared error of F is close to its minimum if and only if the excess error $\int_{\mathbb{R}^d}(m(x)-F(x))^2\mu(dx)$ is close to zero.

The regression function cannot be calculated as long as the distribution of (X,Y) is unknown. Assume, however, that we observed data

$$D_n=\{(X_1,Y_1),\dots,(X_n,Y_n)\} \tag{1.2}$$

consisting of independent and identically distributed copies of (X,Y). D_n can be used to produce an estimate $m_n=m_n(\cdot,D_n)$ of the regression function m. Since m arises from L_2 considerations, it is natural to study $L_2(\mu)$ convergence of the regression estimate m_n to m.

It is of great importance to be able to estimate the various characteristics of the residual

$$Y-m(X).$$

For nonparametric estimates of the minimum mean squared error $L^*=\mathbb{E}\{(Y-m(X))^2\}$ see, e.g., Dudoit and van der Laan [12], Kohler [18], Liitiäinen et al. [19], [20], Liitiäinen et al. [21], Müller and Stadtmüller [22], Neumann [24], Pelckmans et al. [26], Stadtmüller and Tsybakov [28] and the literature cited there. Devroye et al. [9] proved that without any tail and smoothness condition L^* cannot be estimated with guaranteed rate of convergence, and showed a first nearest neighbor based estimate, which for Lipschitz continuous m has faster rate of convergence than that of the usual plug-in estimators. Müller, Schick and Wefelmeyer [23] estimated L^* as the variance of an independent measurement error Z in the model

$$Y=m(X)+Z \tag{1.3}$$

such that $\mathbb{E}\{Z\}=0$, and X and Z are independent. Sometimes it is called additive noise model or homeoscedastic regression model.

2 The rate of convergence of the kernel density estimate

In this paper we deal with the problem how to estimate the density f of the residual $Y-m(X)$ assuming that the density f exists. Our aim is to estimate f from i.i.d. data (1.2).

Under some smoothness conditions on the density f, Ahmad [1], Cheng [3], [4], Efromovich [13], [14], Akritas and Van Keilegom [2], Neumeyer and Van Keilgom [25] studied the estimate the density of the residual. Under the additive noise model (1.3), Devroye et al. [5] introduced a density estimate of the residual, and proved its universal (density free) strong consistency in L_1.

Next we introduce a data splitting scheme. Assume that we are given two independent samples:

$$D_n=\{(X_1,Y_1),\dots,(X_n,Y_n)\}$$

and

$$D'_n = \{(X'_1, Y'_1), \dots, (X'_n, Y'_n)\}.$$

From sample D_n we generate a strongly universally consistent regression estimate m_n. Let K be a density on $\mathbb{R}$, called kernel. For a bandwidth $h > 0$, introduce the notation

$$K_h(z) = \frac{1}{h} K(z/h).$$

Then the kernel density estimate is defined by

$$f_n(z) := \frac{1}{n} \sum_{i=1}^{n} K_{h_n}(z - Z_i), \tag{2.1}$$

where in the i-th term we plug-in the approximation of the i-th residual

$$Z_i := Y'_i - m_n(X'_i),$$

and $\{h_n\}$ is a bandwidth sequence. Given D_n, the common density of Z_i's is g_n.

Under the additive noise model (1.3), Devroye et al. [5] proved the density-free strong consistency of f_n.

An important problem is to bound the rate of convergence of

$$\mathbb{E}\left\{ \int_{\mathbb{R}} |f_n(z) - f(z)| dz \right\},$$

where f_n is the kernel estimate (2.1). The main question is the size of degradation with respect to the case when using an oracle that $Y_i - m(X_i)$ is available, i.e., what is the influence of the regression estimate in the rate of convergence of the density estimate.

Theorem 2.1 *Under the model of additive noise (1.3), assume that the density f is twice differentiable and has a compact support contained in the interval I. Moreover, suppose that the kernel K is symmetric ($K(x) = K(-x)$), bounded and has compact support. Then*

$$\begin{aligned} \mathbb{E}\left\{ \int_{\mathbb{R}} |f_n(z) - f(z)| dz \right\} \le\ & c_1 h_n^2 + \frac{c_2}{\sqrt{n h_n}} \\ & + c_3 \mathbb{E}\left\{ \left| \int_{\mathbb{R}^d} m_n(x) \mu(dx) - \mathbb{E}\{m(X)\} \right| \right\} \\ & + c_4 \mathbb{E}\left\{ \int_{\mathbb{R}^d} (m_n(x) - m(x))^2 \mu(dx) \right\}. \end{aligned} \tag{2.2}$$

PROOF. For given $X = x$ and for given $(X_1, Y_1), \dots, (X_n, Y_n)$, the approximate residual

$$Y - m_n(X) = Y - m(X) + m(X) - m_n(X)$$

has the conditional density $f(z + m_n(x) - m(x))$ and so the density $g_n(z)$ of $Y - m_n(X)$ can be calculated as follows:

$$g_n(z) = \int_{\mathbb{R}^d} f(z + m_n(x) - m(x)) \mu(dx).$$

The triangle inequality implies that

$$\begin{aligned}&\int_{\mathbb{R}} |f_n(z)-f(z)|dz\\ &\le \int_{\mathbb{R}} |\mathbb{E}\{f_n(z)\mid D_n\}-f(z)|dz+\int_{\mathbb{R}} |\mathbb{E}\{f_n(z)\mid D_n\}-f_n(z)|dz.\end{aligned}$$

Concerning the conditional bias term, we have that

$$\begin{aligned}&\int_{\mathbb{R}} |\mathbb{E}\{f_n(z)\mid D_n\}-f(z)|dz\\ &=\int_{\mathbb{R}}\left|\int_{\mathbb{R}} K_{h_n}(z-u)g_n(u)du-f(z)\right|dz\\ &\le \int_{\mathbb{R}}\left|\int_{\mathbb{R}} K_{h_n}(z-u)f(u)du-f(z)\right|dz+\int_{\mathbb{R}}\int_{\mathbb{R}} K_{h_n}(z-u)|g_n(u)-f(u)|dudz\\ &=\int_{\mathbb{R}}\left|\int_{\mathbb{R}} K_{h_n}(z-u)f(u)du-f(z)\right|dz+\int_{\mathbb{R}}|g_n(u)-f(u)|du.\end{aligned}$$

For the conditional variation term, we get that

$$\begin{aligned}&\mathbb{E}\left\{\int_{\mathbb{R}} |\mathbb{E}\{f_n(z)\mid D_n\}-f_n(z)|dz\mid D_n\right\}\\ &\le \int_I \mathbb{E}\left\{|\mathbb{E}\{f_n(z)\mid D_n\}-f_n(z)|\mid D_n\right\}dz+2\int_{I^c}\mathbb{E}\{f_n(z)\mid D_n\}dz\\ &\le \int_I \sqrt{\mathbb{E}\left\{|\mathbb{E}\{f_n(z)\mid D_n\}-f_n(z)|^2\mid D_n\right\}}dz\\ &\quad +2\int_{\mathbb{R}}|\mathbb{E}\{f_n(z)\mid D_n\}-f(z)|dz+2\int_{I^c}f(z)dz\\ &\le \frac{\|K\|_2\sqrt{|I|}}{\sqrt{nh_n}}+2\int_{\mathbb{R}}|\mathbb{E}\{f_n(z)\mid D_n\}-f(z)|dz.\end{aligned}$$

Thus,

$$\begin{aligned}&\mathbb{E}\left\{\int_{\mathbb{R}}|f_n(z)-f(z)|dz\right\}\\ &\le \mathbb{E}\left\{\int_{\mathbb{R}}|\mathbb{E}\{f_n(z)\mid D_n\}-f(z)|dz\right\}+\mathbb{E}\left\{\int_{\mathbb{R}}|\mathbb{E}\{f_n(z)\mid D_n\}-f_n(z)|dz\right\}\\ &\le 3\int_{\mathbb{R}}\left|\int_{\mathbb{R}} K_{h_n}(z-u)f(u)du-f(z)\right|dz+3\mathbb{E}\left\{\int_{\mathbb{R}}|g_n(z)-f(z)|dz\right\}\\ &\quad +\frac{\|K\|_2\sqrt{|I|}}{\sqrt{nh_n}}\\ &\le c_1h_n^2+\frac{c_2}{\sqrt{nh_n}}+3\mathbb{E}\left\{\int_{\mathbb{R}}|g_n(z)-f(z)|dz\right\},\end{aligned}$$

where we applied Lemma 5.4 in Devroye, Györfi [7]. The sum of the first and the second term in the right hand side is the same as that of the rate of convergence of the standard kernel estimate (cf. Theorem 5.1 in Devroye, Györfi [7]), so the influence of the excess error can be bounded by $\mathbb{E}\left\{\int_{\mathbb{R}}|g_n(z)-f(z)|dz\right\}$. For twice differentiable density f, let's calculate the second order Taylor expansion of $f(z+m_n(x)-m(x))$ at z:

$$f(z+m_n(x)-m(x)) = f(z)+f'(z)(m_n(x)-m(x))+\frac{f''(z_{n,x})}{2}(m_n(x)-m(x))^2$$

with some $z_{n,x}$. Then

$$\begin{aligned}
&\int_{\mathbb{R}}|g_n(z)-f(z)|dz\\
&=\int_{\mathbb{R}}\Big|\int_{\mathbb{R}^d}f(z+m_n(x)-m(x))\mu(dx)-f(z)\Big|dz\\
&=\int_{\mathbb{R}}\Big|\int_{\mathbb{R}^d}(f'(z)(m_n(x)-m(x))+\frac{f''(z_{n,x})}{2}(m_n(x)-m(x))^2)\mu(dx)\Big|dz\\
&\leq |I|\max_z|f'(z)|\Big|\int_{\mathbb{R}^d}(m_n(x)-m(x))\mu(dx)\Big|\\
&\quad+|I|\max_z|f''(z)|\int_{\mathbb{R}^d}(m_n(x)-m(x))^2\mu(dx)\\
&=c_3\Big|\int_{\mathbb{R}^d}m_n(x)\mu(dx)-\mathbb{E}\{m(X)\}\Big|+c_4\int_{\mathbb{R}^d}(m_n(x)-m(x))^2\mu(dx).
\end{aligned}$$

□

If $h_n=c_5n^{-1/5}$ then

$$c_1h_n^2+\frac{c_2}{\sqrt{nh_n}}=c_6n^{-2/5}.$$

If the regression function m is Lipschitz continuous and X and Y are bounded then the partitioning, the kernel and the nearest neighbor regression estimates have the rate of convergence

$$\mathbb{E}\left\{\int_{\mathbb{R}^d}(m_n(x)-m(x))^2\mu(dx)\right\}\leq c_7n^{-2/(d+2)}, \tag{2.3}$$

(cf. Chapters 4, 5, 6 in Györfi et al [17]). Next we show that under some situations,

$$\mathbb{E}\left\{\Big|\int_{\mathbb{R}^d}m_n(x)\mu(dx)-\mathbb{E}\{m(X)\}\Big|\right\}\leq c_8n^{-2/(d+2)}+c_9n^{-1/2}, \tag{2.4}$$

which would imply that

$$\mathbb{E}\left\{\int_{\mathbb{R}}|f_n(z)-f(z)|dz\right\}\leq c_6n^{-2/5}+c_7n^{-2/(d+2)},$$

and so for $d\leq 3$ the rate of convergence is the same as that of standard kernel estimate. It means that for $d\leq 3$ and for partitioning regression estimates, the regression estimation error has no influence in the rate of convergence of the density estimate.

3 Application for partitioning regression estimation

Stone [29] first pointed out that there exist universally consistent estimators. He considered local averaging estimates, i.e., estimates of the form

$$m_n(x) = \sum_{i=1}^n W_{ni}(x; X_1, \dots, X_n) Y_i = \sum_{i=1}^n W_{ni}(x) Y_i,$$

where $W_{ni}(x)$ are the data-dependent weights governing the local averaging about x. For local averaging regression estimates,

$$\int_{\mathbb{R}^d} m_n(x)\mu(dx) = \sum_{i=1}^n \int_{\mathbb{R}^d} W_{ni}(x)\mu(dx) Y_i =: \sum_{i=1}^n V_{ni} Y_i$$

such that $V_{ni} \geq 0$, $i = 1, \dots, n$ and $\sum_{i=1}^n V_{ni} = 1$. Noticing

$$\mathbb{E}\left\{ \sum_{i=1}^n V_{ni} Y_i \right\} = \mathbb{E}\left\{ \sum_{i=1}^n V_{ni} m(X_i) \right\},$$

we may apply the decomposition

$$\begin{aligned} &\int_{\mathbb{R}^d} m_n(x)\mu(dx) - \mathbb{E}\{m(X)\} \\ &= \left(\sum_{i=1}^n V_{ni} Y_i - \mathbb{E}\left\{ \sum_{i=1}^n V_{ni} Y_i \right\} \right) + \left(\mathbb{E}\left\{ \sum_{i=1}^n V_{ni} m(X_i) \right\} - \mathbb{E}\{m(X)\} \right), \end{aligned}$$

so in order to show (2.4) we prove that

$$\begin{aligned} \mathbb{E}\left\{ \Big| \sum_{i=1}^n V_{ni} Y_i - \mathbb{E}\left\{ \sum_{i=1}^n V_{ni} Y_i \right\} \Big| \right\} &\leq \sqrt{\mathbb{V}ar\Big(\sum_{i=1}^n V_{ni} Y_i \Big)} \\ &\leq \frac{c_{10}}{\sqrt{n}} \end{aligned} \tag{3.1}$$

and

$$\Big| \mathbb{E}\left\{ \sum_{i=1}^n V_{ni} m(X_i) \right\} - \mathbb{E}\{m(X)\} \Big| \leq c_{11} n^{-2/(d+2)}. \tag{3.2}$$

The *partitioning regression estimate* is defined by a partition $\mathcal{P}_n = \{A_{n,1}, A_{n,2} \dots\}$ of $\mathbb{R}^d$ and

$$m_n(x) = \frac{\sum_{i=1}^n Y_i I_{\{X_i \in A_n(x)\}}}{\sum_{i=1}^n I_{\{X_i \in A_n(x)\}}},$$

where $A_n(x)$ denotes the cell $A_{n,j}$ into which x falls, and $0/0 = 0$, by definition. Results on universal consistency can be found in Devroye and Györfi [6], Györfi [16] and Walk [30].

For partitioning estimate we have that

$$\begin{aligned}V_{ni} &= \int_{\mathbb{R}^d} \frac{I_{\{X_i\in A_n(x)\}}}{\sum_{j=1}^n I_{\{X_j\in A_n(x)\}}}\mu(dx)\\ &= \int_{\mathbb{R}^d} \frac{I_{\{x\in A_n(X_i)\}}}{\sum_{j=1}^n I_{\{X_j\in A_n(X_i)\}}}\mu(dx)\\ &= \frac{\mu(A_n(X_i))}{n\mu_n(A_n(X_i))},\end{aligned}$$

where μ_n denotes the empirical distribution for the samples $X_1,\dots,X_n$. One can check that $\sum_{i=1}^n V_{ni}=1$.

Corollary 3.1 *For the kernel density estimate f_n defined by (2.1), choose $h_n=c_5n^{-1/5}$. Let the regression estimate m_n be the partitioning regression estimate. In addition to the conditions of Theorem 2.1, assume that the partition is cubic with side length*

$$h'_n = c_{13}n^{-1/(d+2)},$$

Y and X are bounded, and m satisfies the Lipschitz condition:

$$|m(x)-m(z)\le C\|x-z\|. \tag{3.3}$$

Then

$$\mathbb{E}\left\{\int_{\mathbb{R}}|f_n(z)-f(z)|dz\right\}\le c_6n^{-2/5}+c_7n^{-2/(d+2)}.$$

PROOF. Theorem 4.3 in Györfi et al [17] implies (2.3), so because of Theorem 2.1, we have to show (3.1) and (3.2). Let L denote a bound of $|Y|$. In view of (3.2) we have

$$\begin{aligned}&\left|\mathbb{E}\left\{\sum_{i=1}^n V_{ni}m(X_i)\right\}-\mathbb{E}\{m(X)\}\right|\\ &= \left|n\mathbb{E}\left\{V_{n1}m(X_1)\right\}-\mathbb{E}\{m(X)\}\right|\\ &= \left|\mathbb{E}\left\{\frac{n\mu(A_n(X_1))}{n\mu_n(A_n(X_1))}m(X_1)\right\}-\mathbb{E}\{m(X)\}\right|\\ &= \Big|\sum_{A\in\mathcal{P}_n}\int_A m(x)\mu(dx)\mathbb{E}\left\{\frac{n\mu(A)}{\sum_{i=2}^n I_{\{X_i\in A\}}+1}\right\}-\sum_{A\in\mathcal{P}_n}\int_A m(x)\mu(dx)\Big|\\ &\le L\sum_{A\in\mathcal{P}_n}\mu(A)\left|\mathbb{E}\left\{\frac{n\mu(A)}{\sum_{i=2}^n I_{\{X_i\in A\}}+1}\right\}-1\right|\\ &= L\sum_{A\in\mathcal{P}_n}\mu(A)(1-\mu(A))^n\\ &\le \frac{Le^{-1}|\mathcal{P}_n|}{n}\\ &\le \frac{c_{14}}{nh'^d_n}=c_{10}n^{-2/(d+2)}.\end{aligned}$$

For (3.1), we set

$$U_{ni} = V_{ni} Y_i.$$

We have to show that

$$\mathbb{V}ar\left(\sum_{i=1}^{n} U_{ni}\right) \leq \frac{c^*}{n}. \tag{3.4}$$

Apply the Efron-Stein inequality ([15]). By symmetry, it suffices to use i.i.d. random vectors $(X_1, Y_1), (X_1', Y_1'), (X_2, Y_2), \ldots, (X_n, Y_n)$ and U_{ni}' obtained from U_{ni} by replacement of (X_1, Y_1) by (X_1', Y_1'). The Efron-Stein inequality yields

$$\mathbb{V}ar\left(\sum_{i=1}^{n} U_{ni}\right) \leq \frac{n}{2}\mathbb{E}\left\{\left(\sum_{i=1}^{n} U_{ni} - \sum_{i=1}^{n} U_{ni}'\right)^2\right\}.$$

Thus,

$$\begin{aligned}\mathbb{V}ar\left(\sum_{i=1}^{n} U_{ni}\right) &\leq \frac{3n}{2}\left(\mathbb{E}\left\{U_{n1}^2\right\} + \mathbb{E}\left\{{U'}_{n1}^2\right\} + \mathbb{E}\left\{\left(\sum_{i=2}^{n}(U_{ni} - U_{ni}')\right)^2\right\}\right)\\ &= 3n\mathbb{E}\left\{U_{n1}^2\right\} + \frac{3n}{2}\mathbb{E}\left\{\left(\sum_{i=2}^{n}(U_{ni} - U_{ni}')\right)^2\right\}.\end{aligned}$$

For a binomial-(n, p)-distributed random variable B $(0 < p \leq 1)$, we notice

$$\mathbb{E}\left\{\frac{1}{(1+B)^r}\right\} \leq \frac{r!}{(n+1)^r p^r}$$

$(r \in \mathbb{N})$. Thus, by independence

$$\mathbb{E}\left\{U_{n1}^2\right\} \leq L^2 \int_{\mathbb{R}^d} \mu(A_n(s))^2 \mathbb{E}\left\{\frac{1}{(1 + \sum_{i=2}^{n} I_{\{X_i \in A_n(s)\}})^2}\right\} \mu(ds) \leq \frac{c_{15}}{n^2}.$$

Further, with

$$N_{i\ell} = \sum_{j \in \{2,\ldots,n\} \setminus \{i\}} I_{\{X_j \in A_n(X_\ell)\}}$$

$(i = 2, \ldots, n, \ell \in \{1, i\})$, we have

$$\begin{aligned}&\sum_{i=2}^{n}(U_{ni} - U_{ni}')\\ &= \sum_{i=2}^{n} \mu(A_n(X_i)) Y_i \frac{I_{\{X_1' \in A_n(X_i)\}} - I_{\{X_1 \in A_n(X_i)\}}}{(1 + I_{\{X_1 \in A_n(X_i)\}} + N_{ii})(1 + I_{\{X_1' \in A_n(X_i)\}} + N_{ii})},\end{aligned}$$

therefore

$$
\begin{aligned}
&\mathbb{E}\left\{\left(\sum_{i=2}^{n}(U_{ni}-U'_{ni})\right)^2\right\}\\
\leq\ & 4L^2\mathbb{E}\left\{\left(\sum_{i=2}^{n}\frac{\mu(A_n(X_i))I_{\{X_1\in A_n(X_i)\}}}{(1+N_{ii})^2}\right)^2\right\}\\
=\ & 4L^2\mathbb{E}\left\{\sum_{i=2}^{n}\frac{\mu(A_n(X_i))^2 I_{\{X_1\in A_n(X_i)\}}}{(1+N_{ii})^4}\right\}\\
&+4L^2\mathbb{E}\left\{\sum_{i,\ell=2,\dots,n,\, i\neq\ell}\frac{\mu(A_n(X_i))\mu(A_n(X_\ell))I_{\{X_1\in A_n(X_i)\}}I_{\{X_1\in A_n(X_\ell)\}}}{(1+N_{ii})^2(1+N_{\ell\ell})^2}\right\}.
\end{aligned}
$$

The identity $I_{\{X_1\in A_n(X_i)\}}=I_{\{X_i\in A_n(X_1)\}}$ implies that

$$
\begin{aligned}
&\mathbb{E}\left\{\left(\sum_{i=2}^{n}(U_{ni}-U'_{ni})\right)^2\right\}\\
\leq\ & 4L^2\mathbb{E}\left\{\sum_{i=2}^{n}\frac{\mu(A_n(X_1))^2 I_{\{X_i\in A_n(X_1)\}}}{(1+N_{i1})^4}\right\}\\
&+4L^2\mathbb{E}\left\{\sum_{i,\ell=2,\dots,n,\, i\neq\ell}\frac{\mu(A_n(X_1))^2 I_{\{X_i\in A_n(X_1)\}}I_{\{X_\ell\in A_n(X_1)\}}}{(1+N_{i1})^2(1+N_{\ell 1})^2}\right\}\\
\leq\ & 4L^2(n-1)\mathbb{E}\left\{\frac{\mu(A_n(X_1))^2 I_{\{X_2\in A_n(X_1)\}}}{(1+\sum_{j=3}^{n}I_{\{X_j\in A_n(X_1)\}})^3}\right\}\\
&+4L^2(n-1)(n-2)\mathbb{E}\left\{\frac{\mu(A_n(X_1))^2 I_{\{X_2\in A_n(X_1)\}}I_{\{X_3\in A_n(X_1)\}}}{(1+\sum_{j=4}^{n}I_{\{X_j\in A_n(X_1)\}})^4}\right\}\\
=\ & 4L^2(n-1)\int_{\mathbb{R}^d}\mu(A_n(s))^3\mathbb{E}\left\{\frac{1}{(1+\sum_{j=3}^{n}I_{\{X_j\in A_n(s)\}})^3}\right\}\mu(ds)\\
&+4L^2(n-1)(n-2)\int_{\mathbb{R}^d}\mu(A_n(s))^4\mathbb{E}\left\{\frac{1}{(1+\sum_{j=4}^{n}I_{\{X_j\in A_n(s)\}})^4}\right\}\mu(ds)\\
\leq\ & \frac{c_{16}}{n^2}.
\end{aligned}
$$

Thus, (3.4) is obtained, and the proof of the corollary is complete. □

4 Application for kernel regression estimation

The *kernel regression estimate* is given by

$$m_n(x) = \frac{\sum_{i=1}^n Y_i K'_{h'_n}(x - X_i)}{\sum_{i=1}^n K'_{h'_n}(x - X_i)},$$

where $h'_n > 0$ is a smoothing factor depending upon n, K' is an absolutely integrable function (the kernel), and $K'_{h'_n}(x) = K'(x/h'_n)$. Under some additional conditions on the kernel and under the conditions

$$h'_n \to 0, \ \ nh'^d_n \to \infty$$

Devroye and Wagner [11], Spiegelman and Sacks [27], Devroye and Krzyżak [10] and Walk [30], [31] proved consistency theorems for the kernel estimate.

For kernel estimate we have that

$$V_{ni} = \int_{\mathbb{R}^d} \frac{K'_{h'_n}(x - X_i)}{\sum_{j=1}^n K'_{h'_n}(x - X_j)} \mu(dx).$$

Obviously, $\sum_{i=1}^n V_{ni} \leq 1$. In the special case of window kernel $K'(x) = I_{\{\|x\| \leq 1\}}$, one has

$$V_{ni} = \int_{\mathbb{R}^d} \frac{I_{\{\|x - X_i\| \leq h'_n\}}}{\sum_{j=1}^n I_{\{\|x - X_j\| \leq h'_n\}}} \mu(dx).$$

Corollary 4.1 *For the kernel density estimate f_n defined by (2.1), choose $h_n = c_5 n^{-1/5}$. Let the regression estimate m_n be the kernel regression estimate with window kernel. In addition to the conditions of Theorem 2.1, assume that*

$$h'_n = c_{17} n^{-1/(d+2)},$$

Y and X are bounded, and m satisfies the Lipschitz condition (3.3). Put

$$\bar{m}_h(x) := \frac{\int_{S_{x,h}} m(s)\mu(ds)}{\mu(S_{x,h})},$$

where $S_{x,h}$ stands for the sphere centered at x and radius h. Then

$$\mathbb{E}\left\{\int_{\mathbb{R}} |f_n(z) - f(z)| dz\right\} \leq c_6 n^{-2/5} + c_7 n^{-2/(d+2)}$$
$$+ \left| \int_{\mathbb{R}^d} \bar{m}_{h'_n}(x)\mu(dx) - \int_{\mathbb{R}^d} m(x)\mu(dx) \right|.$$

PROOF. Theorem 5.2 in Györfi et al [17] implies (2.3), so because of Theorem 2.1, we have to show (3.1) and (3.2) such that $\mathbb{E}\{m(X)\} = \int_{\mathbb{R}^d} m(x)\mu(dx)$ is replaced by

$\int_{\mathbb{R}^d} \bar{m}_{h'_n}(x)\mu(dx)$. (3.1) has been proved in Lemma 3.10, a) of Walk [31]. Concerning (3.2), we have that

$$\begin{aligned}\mathbb{E}\left\{\sum_{i=1}^n V_{ni}m(X_i)\right\} &= n\mathbb{E}\left\{V_{n1}m(X_1)\right\}\\ &= n\mathbb{E}\left\{\int_{\mathbb{R}^d} \frac{I_{\{X_1\in S_{x,h'_n}\}}}{1+\sum_{j=2}^n I_{\{X_j\in S_{x,h'_n}\}}} m(X_1)\mu(dx)\right\}.\end{aligned}$$

Because of independence, we get that

$$\begin{aligned}\mathbb{E}\left\{\sum_{i=1}^n V_{ni}m(X_i)\right\} &= \int_{\mathbb{R}^d}\int_{S_{x,h'_n}} m(s)\mu(ds)\mathbb{E}\left\{\frac{n}{1+\sum_{j=2}^n I_{\{X_j\in S_{x,h'_n}\}}}\right\}\mu(dx)\\ &= \int_{\mathbb{R}^d} \bar{m}_{h'_n}(x)\mathbb{E}\left\{\frac{n\mu(S_{x,h'_n})}{1+\sum_{j=2}^n I_{\{X_j\in S_{x,h'_n}\}}}\right\}\mu(dx)\\ &= \int_{\mathbb{R}^d} \bar{m}_{h'_n}(x)\left(1-(1-\mu(S_{x,h'_n}))^n\right)\mu(dx),\end{aligned}$$

therefore

$$\begin{aligned}&\left|\mathbb{E}\left\{\sum_{i=1}^n V_{ni}m(X_i)\right\} - \int_{\mathbb{R}^d} m(x)\mu(dx)\right|\\ &\le \left|\int_{\mathbb{R}^d} \bar{m}_{h'_n}(x)\mu(dx) - \int_{\mathbb{R}^d} m(x)\mu(dx)\right| + \left|\int_{\mathbb{R}^d} \bar{m}_{h'_n}(x)(1-\mu(S_{x,h'_n}))^n\mu(dx)\right|\\ &\le \left|\int_{\mathbb{R}^d} \bar{m}_{h'_n}(x)\mu(dx) - \int_{\mathbb{R}^d} m(x)\mu(dx)\right| + L\int_{\mathbb{R}^d}(1-\mu(S_{x,h'_n}))^n\mu(dx).\end{aligned}$$

The compact support of X can be covered by $M_n = c\cdot {h'}_n^{-d}$ many balls, with translates of $S_{0,h'_n/2}$ and with centers $x_1,\dots,x_{M_n}$. Thus,

$$\begin{aligned}\int_{\mathbb{R}^d}(1-\mu(S_{x,h'_n}))^n\mu(dx) &\le \sum_{j=1}^{M_n}\int_{S_{x_j,h'_n/2}}(1-\mu(S_{x,h'_n}))^n\mu(dx)\\ &\le \sum_{j=1}^{M_n}\int_{S_{x_j,h'_n/2}}(1-\mu(S_{x_j,h'_n/2}))^n\mu(dx)\\ &= \sum_{j=1}^{M_n}\mu(S_{x_j,h'_n/2})(1-\mu(S_{x_j,h'_n/2}))^n\\ &\le \frac{M_n e^{-1}}{n}\\ &\le \frac{ce^{-1}}{n{h'}_n^d}\\ &= c_{11}n^{-2/(d+2)}.\end{aligned}$$

□

5 Application for nearest neighbor regression estimation

For the *k-nearest neighbor regression estimate*, $W_{ni}(x; X_1, \dots, X_n)$ is chosen to be $1/k_n$ if X_i is one of the k_n nearest neighbors of x among $X_1, \dots, X_n$, and zero otherwise. More formally, we fix $x \in \mathbb{R}^d$, and reorder the data $(X_1, Y_1), \dots, (X_n, Y_n)$ according to increasing values of $\|X_i - x\|$. The reordered data sequence is denoted by

$$(X_{(1,n)}(x), Y_{(1,n)}(x)), \dots, (X_{(n,n)}(x), Y_{(n,n)}(x)).$$

$X_{(k,n)}(x)$ is called the kth nearest neighbor (k-NN) of x. In the sequel we assume that this ordering is unique almost surely for μ almost all x, i.e., tie occurs with probability zero. The k_n-NN regression function estimate is defined by

$$m_n(x) = \frac{1}{k_n} \sum_{i=1}^{k_n} Y_{(i,n)}(x).$$

If

$$k_n \to \infty, \;\; k_n/n \to 0$$

then the consistency of the k-nearest neighbor regression estimate was established by Stone [29] and by Devroye et al. [8].

Let the set $A_{n,i}$ consists of those x's, for which X_i is one of the k_n nearest neighbors of x among $X_1, \dots, X_n$. Then

$$V_{ni} = \frac{\mu(A_{n,i})}{k_n}.$$

Obviously, $\sum_{i=1}^n V_{ni} = 1$.

Corollary 5.1 *For the kernel density estimate f_n defined by (2.1), choose $h_n = c_5 n^{-1/5}$. Let the regression estimate m_n be the k_n-nearest neighbor estimate. In addition to the conditions of Theorem 2.1, assume that*

$$k_n = c_{15} n^{2/(d+2)},$$

Y and X are bounded, tie occurs with probability zero, and m satisfies the Lipschitz condition (3.3). Put

$$\tilde{m}_n(x) = \frac{1}{k_n} \int_{S_{x,\rho_n(x)}} m(s)\mu(ds),$$

where $\rho_n(x)$ is the solution of the equation

$$\frac{k_n}{n} = \mu(S_{x,\rho_n(x)}).$$

Then, for $d \geq 2$,

$$\mathbb{E}\left\{ \int_{\mathbb{R}} |f_n(z) - f(z)| dz \right\} \leq c_6 n^{-2/5} + c_7 n^{-2/(d+2)} + \left| \int_{\mathbb{R}^d} \tilde{m}_n(x)\mu(dx) - \int_{\mathbb{R}^d} m(x)\mu(dx) \right|.$$

PROOF. For $d \geq 2$, Theorem 6.2 in Györfi et al. [17] and Theorem 3.2 in Liitiäinen et al. [20], which generalizes Lemma 6.4 in Györfi et al. [17] from $d \geq 3$ to $d \geq 2$, imply (2.3). So because of Theorem 2.1, we shall show (2.4). For this, we use arguments of Devroye et al. [8] (or from Section 23.3 in Györfi et al. [17]). Define the auxiliary estimate

$$m_n^*(x) = \frac{1}{k_n} \sum_{i=1}^{n} Y_i I_{\{X_i \in S_{x,\rho_n(x)}\}}.$$

Denoting

$$R_n(x) = \|X_{(k_n,n)}(x) - x\|,$$

we have that

$$\begin{aligned}
&\left| \int_{\mathbb{R}^d} (m_n^*(x) - m_n(x))\mu(dx) \right| \\
&= \frac{1}{k_n} \left| \sum_{i=1}^{n} Y_i \int_{\mathbb{R}^d} (I_{\{X_i \in S_{x,\rho_n(x)}\}} - I_{\{X_i \in S_{x,R_n(x)}\}})\mu(dx) \right| \\
&\leq \frac{L}{k_n} \sum_{i=1}^{n} \left| \int_{\mathbb{R}^d} (I_{\{X_i \in S_{x,\rho_n(x)}\}} - I_{\{X_i \in S_{x,R_n(x)}\}})\mu(dx) \right|.
\end{aligned}$$

By considering the cases $\rho_n(x) \leq R_n(x)$ and $\rho_n(x) > R_n(x)$ we obtain that at a change of i, for each x the sign of

$$I_{\{X_i \in S_{x,\rho_n(x)}\}} - I_{\{X_i \in S_{x,R_n(x)}\}}$$

and thus the sign of

$$\int_{\mathbb{R}^d} (I_{\{X_i \in S_{x,\rho_n(x)}\}} - I_{\{X_i \in S_{x,R_n(x)}\}})\mu(dx)$$

remains unaltered. Therefore

$$\begin{aligned}
&\left| \int_{\mathbb{R}^d} (m_n^*(x) - m_n(x))\mu(dx) \right| \\
&\leq L \left| \frac{1}{k_n} \sum_{i=1}^{n} \int_{\mathbb{R}^d} (I_{\{X_i \in S_{x,\rho_n(x)}\}} - I_{\{X_i \in S_{x,R_n(x)}\}})\mu(dx) \right| \\
&= L \left| \int_{\mathbb{R}^d} M_n^*(x)\mu(dx) - \int_{\mathbb{R}^d} 1\mu(dx) \right|,
\end{aligned}$$

where M_n^* is defined as m_n^* with Y_i replaced by constant 1. Then

$$\begin{aligned}
&\left| \int_{\mathbb{R}^d} (m_n(x) - m(x))\mu(dx) \right| \\
&\leq \left| \int_{\mathbb{R}^d} (m_n^*(x) - m_n(x))\mu(dx) \right| + \left| \int_{\mathbb{R}^d} (m_n^*(x) - m(x))\mu(dx) \right| \\
&\leq L \left| \int_{\mathbb{R}^d} M_n^*(x)\mu(dx) - \int_{\mathbb{R}^d} 1\mu(dx) \right| + \left| \int_{\mathbb{R}^d} (m_n^*(x) - m(x))\mu(dx) \right|.
\end{aligned}$$

Therefore obviously it suffices to show (2.4) for m_n^* instead of m_n, where we now have

$$V_{ni} = \frac{1}{k_n} \int_{\mathbb{R}^d} I_{\{X_i \in S_{x,\rho_n(x)}\}} \mu(dx).$$

It is now enough to show (3.1) and (3.2), where $\mathbb{E}\{m(X)\}$ is replaced by $\mathbb{E}\{\tilde{m}_n(X)\}$. But then apparently (3.2) is fulfilled which vanishing right hand side. (3.1) means

$$\mathbb{V}ar\left(\int_{\mathbb{R}^d} m_n^*(x)\mu(dx)\right) \leq \frac{c_{10}^2}{n}.$$

Apply the Efron-Stein inequality ([15]). By symmetry, it suffices to use i.i.d. random vectors $(X_1, Y_1), (X_1', Y_1'), (X_2, Y_2), \ldots, (X_n, Y_n)$ and $m_n'^*$ obtained from m_n^* by replacement of (X_1, Y_1) by (X_1', Y_1'). The Efron-Stein inequality yields

$$\mathbb{V}ar\left(\int_{\mathbb{R}^d} m_n^*(x)\mu(dx)\right) \leq \frac{n}{2}\mathbb{E}\left\{\left(\int_{\mathbb{R}^d} m_n^*(x)\mu(dx) - \int_{\mathbb{R}^d} m_n'^*(x)\mu(dx)\right)^2\right\}.$$

There exist cones $C_1, \ldots, C_{\gamma_d}$, each with top 0 and with angle $\pi/3$ such that

$$\cup_{j=1}^{\gamma_d} C_j = \mathbb{R}^d.$$

Now we argue as in Devroye et al. [8]. $m_n^*(x) - m_n'^*(x)$ is absolutely bounded by $2L/k_n$ and can differ from zero if $X_1 \in S_{x,\rho_n(x)}$ or $X_1' \in S_{x,\rho_n(x)}$. One has $X_1 \in S_{x,\rho_n(x)}$ or $X_1' \in S_{x,\rho_n(x)}$ if and only if $\mu(S_{x,\|x-X_1\|}) \leq k_n/n$ or $\mu(S_{x,\|x-X_1'\|}) \leq k_n/n$. But the μ-measure of the random set of such x's is bounded by $2\gamma_d k_n/n$ (cf. Lemma 6.2 in Györfi et al. [17]). Therefore

$$\left|\int_{\mathbb{R}^d} (m_n^*(x) - m_n'^*(x))\mu(dx)\right| \leq \frac{2L}{k_n}\frac{2\gamma_d k_n}{n} = \frac{4L\gamma_d}{n}.$$

Thus

$$\mathbb{V}ar\left(\int_{\mathbb{R}^d} m_n^*(x)\mu(dx)\right) \leq \frac{8L^2\gamma_d^2}{n}.$$

□

Acknowledgements

This research was supported through the programme ”Research in Pairs” by the Mathematisches Forschungsinstitut Oberwolfach in 2012.

References

[1] Ahmad, I. A. Residuals density estimation in nonparametric regression. *Statistics and Probability Letters*, 14:133–139, 1992.

[2] Akritas, M. G. and Van Keilegom, I. Non-parametric estimation of the residual distribution. *Board of the Foundation of the Scandinavian Journal of Statistics*, Blackwell Publishers Ltd, 28:549–567, 2001.

[3] Cheng, F. Consistency of error density and distribution function estimators in nonparametric regression. *Statistics and Probability Letters*, 59:257–270, 2002.

[4] Cheng, F. Weak and strong uniform consistency of a kernel error density estimator in nonparametric regression. *Journal of Statistical Planning and Inference*, 119:95–107, 2004.

[5] Devroye, L., Felber, T., Kohler, M. and Krzyzak, A. $L1$-consistent estimation of the density of residuals in random design regression models. *Statistics and Probability Letters*, 82:173-179, 2012.

[6] Devroye, L. and Györfi, L. Distribution-free exponential upper bound on the L_1 error of partitioning estimates of a regression function. In *Proceedings of the Fourth Pannonian Symposium on Mathematical Statistics*, eds. Konecny F., Mogyoródi, J. and Wertz, W. , pp. 67-76. Akadémiai Kiadó, Budapest, 1983.

[7] Devroye, L. and Györfi, L. *Nonparametric Density Estimation: The L_1 View*. John Wiley, New York, 1985.

[8] Devroye, L., Györfi, L., Krzyżak, A. and Lugosi, G. On the strong universal consistency of nearest neighbor regression function estimates. *Annals of Statistics*, 22:1371–1385, 1994.

[9] Devroye, L., Schäfer, D., Györfi, L. and Walk, H. The estimation problem of minimum mean squared error. *Statistics and Decisions*, 21:15-28, 2003.

[10] Devroye, L. and Krzyżak, A. An equivalence theorem for L1 convergence of the kernel regression estimate. *Journal of Statistical Planning and Inference*, 23:71–82, 1989.

[11] Devroye, L. and Wagner, T. J. Distribution-free consistency results in nonparametric discrimination and regression function estimation. *Annals of Statistics*, 8:231–239, 1980.

[12] Dudoit, S. and van der Laan, M.J. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. Statistical Methodology, 2:131–154, 2005.

[13] Efromovich, S. Estimation of the density of regression errors. *Annals of Statistics*, 33:2194–2227, 2005.

[14] Efromovich, S. Optimal nonparametric estimation of the density of regression errors with finite support. *AISM*, 59:617–654, 2006.

[15] Efron, B. and Stein, C. On jacknife estimate of variance. *Annals of Statistics*, 9:586–596, 1981.

[16] Györfi, L. Universal consistencies of regression estimate for unbounded regression functions. In *Nonparametric Functional Estimation and Related Topics*, ed. G. Roussas, pp. 329–338. Kluwer Academic Publishers, Dordrecht, 1991.

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

[18] Kohler, M. Nonparametric regression with additional measurement errors in the dependent variable, *Journal of Statistical Planning and Inference*, 136:3339–3361, 2006.

[19] Liitiäinen, E., Corona, F. and Lendasse, A. On nonparametric residual variance estimation. *Neural Processing Letters*, 28:155–167, 2009.

[20] Liitiäinen, E., Corona, F. and Lendasse, A. Residual variance estimation using a nearest neighbor statistic. *Journal of Multivariate Analysis*, 101:811–823, 2010.

[21] Liitiäinen, E., Verleysen, M, Corona, F. and Lendasse, A. Residual variance estimation in machine learning. *Neurocomputing*, 72:3692–3703, 2009.

[22] Müller, H.-G. and Stadtmüller, U. Estimation of heteroscedasticity in regression analysis, *Annals of Statistics*, 15:610–625, 1987.

[23] Müller, U., Schick, A. and Wefelmeyer, W. Estimating the error variance in nonparametric regression by a covariate-matched U-statistic, *Statistics*, 37:179–188, 2003.

[24] Neumann, M.-H. Fully data-driven nonparametric variance estimators, *Statistics*, 25:189–212, 1994.

[25] Neumeyer, N. and Van Keilegom, I. Estimating the error distribution in nonparametric multiple regression with applications to model testing. *Journal of Multivariate Analysis*, 101:1067–1078, 2010.

[26] Pelckmans, K., De Brabanter, J., Suykens, J. A. K. and De Moor, B. The differogram: Non-parametric noise variance estimation and its use for model selection. Neurocomputing, 69:100–122, 2005.

[27] Spiegelman, C. and Sacks, J. Consistent window estimation in nonparametric regression. *Annals of Statistics*, 8:240–246, 1980.

[28] Stadtmüller, U. and Tsybakov, A. Nonparametric recursive variance estimation, *Statistics*, 27:55–63, 1995.

[29] Stone, C. J. Consistent nonparametric regression. *Annals of Statistics*, 5:595–645, 1977.

[30] Walk, H. Almost sure convergence properties of Nadaraya-Watson regression estimates. In *Modeling Uncertainty. An Examination of its Theory, Methods and Applications*, eds. M. Dror, P. L'Ecuyer and F. Szidarovszky, pp. 201-223. Kluwer Academic Publishers, Dordrecht, 2002.

[31] Walk, H. Strong universal consistency of smooth kernel regression estimates. *Ann. Inst. Statist. Math.*, 57:665–685, 2005.

László Györfi
Department of Computer Science and Information Theory
Budapest University of Technology and Economics
1521 Stoczek u. 2, Budapest, Hungary
gyorfi@cs.bme.hu

Harro Walk
Department of Mathematics
Universität Stuttgart
Pfaffenwaldring 57, D-70569 Stuttgart, Germany
walk@mathematik.uni-stuttgart.de