Journal of Machine Learning Research 16 (2015) 1863-1877 Submitted 6/15; Published 9/15

On the Asymptotic Normality of an Estimate of a Regression Functional

László Györfi * GYORFI@CS.BME.HU
Department of Computer Science and Information Theory
Budapest University of Technology and Economics
Magyar Tudósok körútja 2., H-1117 Budapest, Hungary

Harro Walk WALK@MATHEMATIK.UNI-STUTTGART.DE
Department of Mathematics
University of Stuttgart
Pfaffenwaldring 57, D-70569 Stuttgart, Germany

Editor: Alex Gammerman and Vladimir Vovk

Abstract

An estimate of the second moment of the regression function is introduced. Its asymptotic normality is proved such that the asymptotic variance depends neither on the dimension of the observation vector, nor on the smoothness properties of the regression function. The asymptotic variance is given explicitly.

Keywords: nonparametric estimation, regression functional, central limit theorem, partitioning estimate

1. Introduction

This paper considers a histogram-based estimate of second moment of the regression function in multivariate problems. The interest in the second moment is motivated by the fact that by estimating it one obtains an estimate of the best possible achievable mean squared error, a quantity of obvious statistical interest. It is shown that the estimate is asymptotically normally distributed. It is remarkable that the asymptotic variance only depends on moments of the regression function but neither on its smoothness, nor on the dimension of the space. The proof relies on a Poissonization technique that has been used successfully in related problems.

Let Y be a real valued random variable with $\mathbb{E}\{Y^2\} < \infty$ and let $X = (X^{(1)}, \dots, X^{(d)})$ be a d-dimensional random observational vector. In regression analysis one wishes to estimate Y given X, i.e., one wants to find a function g defined on the range of X so that $g(X)$ is "close" to Y. Assume that the main aim of the analysis is to minimize the mean squared error :

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

∗. This research has been partially supported by the European Union and Hungary and co-financed by the European Social Fund through the project TMOP-4.2.2.C-11/1/KONV-2012-0004 - National Research Center for Development and Market Introduction of Advanced Information and Communication Technologies.

As is well-known, this minimum is achieved by the regression function $m(x)$, which is defined by

$$m(x) = \mathbb{E}\{Y \mid X = x\}. \tag{1}$$

For each measurable function g one has

$$\begin{aligned}\mathbb{E}\{(g(X) - Y)^2\} &= \mathbb{E}\{(m(X) - Y)^2\} + \mathbb{E}\{(m(X) - g(X))^2\} \\ &= \mathbb{E}\{(m(X) - Y)^2\} + \int \mid m(x) - g(x) \mid^2 \mu(dx),\end{aligned}$$

where μ stands for the distribution of the observation X.

It is of great importance to be able to estimate the minimum mean squared error

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

accurately, even before a regression estimate is applied: in a standard nonparametric regression design process, one considers a finite number of real-valued features $X^{(i)}$, $i \in I$, and evaluates whether these suffice to explain Y. In case they suffice for the given explanatory task, an estimation method can be applied on the basis of the features already under consideration, if not, more or different features must be considered. The quality of a subvector $\{X^{(i)}, i \in I\}$ of X is measured by the minimum mean squared error

$$L^*(I) := \mathbb{E}\left(Y - \mathbb{E}\{Y|X^{(i)}, i \in I\}\right)^2$$

that can be achieved using the features as explanatory variables. $L^*(I)$ depends upon the unknown distribution of $(Y, X^{(i)} : i \in I)$. The first phase of any regression estimation process therefore heavily relies on estimates of L^* (even *before* a regression estimate is picked).

Concerning dimension reduction the related testing problem is on the hypothesis

$$L^* = L^*(I).$$

This testing problem can be managed such that we estimate both L^* and $L^*(I)$, and accept the hypothesis if the two estimates are close to each other. (Cf. De Brabanter et al. (2014).)

Devroye et al. (2003), Evans and Jones (2008), Liitiäinen et al. (2008), Liitiäinen et al. (2009), Liitiäinen et al. (2010), and Ferrario and Walk (2012) introduced nearest neighbor based estimates of L^*, proved strong universal consistency and calculated the (fast) rate of convergence.

Because of

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

and $\mathbb{E}\{Y^2\} < \infty$, estimating L^* is equivalent to estimating the second moment S^* of the regression function:

$$S^* = \mathbb{E}\{m(X)^2\} = \int m(x)^2 \mu(dx).$$

In this paper we introduce a partitioning based estimator of S^*, and show its asymptotic normality. It turns out that the asymptotic variance depends neither on the dimension of the observation vector, nor on the smoothness properties of the regression function. The asymptotic variance is given explicitly.

2. A Splitting Estimate

We suppose that the regression estimation problem is based on a sequence

$$(X_1, Y_1), (X_2, Y_2), \dots$$

of i.i.d. random vectors distributed as (X, Y). Let

$$\mathcal{P}_n = \{A_{n,j}, j = 1, 2, \dots\}$$

be a cubic partition of $\mathbb{R}^d$ of size $h_n > 0$.

The partitioning estimator of the regression function m is defined as

$$m_n(x) = \frac{\nu_n(A_{n,j})}{\mu_n(A_{n,j})} \text{ if } x \in A_{n,j}, \tag{2}$$

(interpreting $0/0 = 0$) with

$$\nu_n(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{X_i \in A\}} Y_i$$

and

$$\mu_n(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{X_i \in A\}}.$$

(Here $\mathbb{I}$ denotes the indicator function.)

If for cubic partition

$$nh_n^d \to \infty \quad \text{and} \quad h_n \to 0 \tag{3}$$

as $n \to \infty$, then the partitioning regression estimate (2) is weakly universally consistent, which means that

$$\lim_{n\to\infty} \mathbb{E}\left\{\int (m_n(x) - m(x))^2 \mu(dx)\right\} = 0 \tag{4}$$

for any distribution of (X, Y) with $\mathbb{E}\{Y^2\} < \infty$, and for bounded Y it holds

$$\lim_{n\to\infty} \int (m_n(x) - m(x))^2 \mu(dx) = 0 \tag{5}$$

a.s. (Cf. Theorems 4.2 and 23.1 in Györfi et al. (2002).)

Assume splitting data

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

and

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

such that $(X_1, Y_1), \dots, (X_n, Y_n), (X_1', Y_1'), \dots, (X_n', Y_n')$ are i.i.d.

The splitting data estimate of S^* is defined as

$$S_n := \frac{1}{n}\sum_{i=1}^{n} Y_i' m_n(X_i') = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{\infty} \mathbb{I}_{\{X_i' \in A_{n,j}\}} Y_i' \frac{\nu_n(A_{n,j})}{\mu_n(A_{n,j})}.$$

Put

$$\nu'_n(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{X'_i \in A\}} Y'_i,$$

then S_n has the equivalent form

$$S_n = \sum_{j=1}^{\infty} \nu'_n(A_{n,j}) \frac{\nu_n(A_{n,j})}{\mu_n(A_{n,j})}. \tag{6}$$

Theorem 1 *Assume (3) and that μ is non-atomic and has bounded support. Suppose that there is a finite constant C such that*

$$\mathbb{E}\{|Y|^3 \mid X\} < C. \tag{7}$$

Then

$$\sqrt{n}\left(S_n - \mathbb{E}\{S_n\}\right)/\sigma \stackrel{\mathcal{D}}{\to} N(0,1),$$

where

$$\sigma^2 = 2\int M_2(x)m(x)^2\mu(dx) - \left(\int m(x)^2\mu(dx)\right)^2 - \int m(x)^4\mu(dx),$$

with

$$M_2(X) = \mathbb{E}\{Y^2 \mid X\}.$$

The estimation problem is motivated by the above mentioned dimension reduction such that one estimates S^* for the original observation vector and for the observation vector where some components are left out. If the two estimates are "close" to each other, then we decide that the left out components are ineffective. Theorem 1 is on the random part of the estimates. Therefore there is a further need to study the difference of the biases of the estimates. Under (3) we have

$$\lim_{n\to\infty} \mathbb{E}\{S_n\} = S^*$$

and for Lipschitz continuous m the rate of convergence can be of order $n^{-1/d}$ for suitable choice of h_n. (Cf. Devroye et al. (2013).) Similarly to De Brabanter et al. (2014) we conjecture that this difference of the biases has universally a fast rate of convergence.

Obviously, there are several other possibilities for defining partitioning based estimates and proving their asymptotic normality, for example,

$$\frac{1}{n}\sum_{i=1}^{n} m_n(X'_i)^2$$

or

$$\sum_{j=1}^{\infty} \frac{\nu_n(A_{n,j})^2}{\mu_n(A_{n,j})}.$$

Notice that both estimates have larger bias and variance than our estimate (6) has.

The proof of Theorem 1 works without any major modification for consistent k_n nearest neighbor (k_n-NN) estimate m_n if $k_n \to \infty$ and $k_n/n \to 0$. A delicate and important research problem is the case of non-consistent 1-NN estimate m_n, because for 1-NN estimate m_n the bias is smaller. We conjecture that even in this case one has a CLT.

We prove Theorem 1 in the next section.

3. Proof of Theorem 1

Introduce the notations

$$U_n = \sqrt{n}\left(S_n - \mathbb{E}\{S_n \mid Z_n\}\right)$$

and

$$V_n = \sqrt{n}\left(\mathbb{E}\{S_n \mid Z_n\} - \mathbb{E}\{S_n\}\right),$$

then

$$\sqrt{n}\left(S_n - \mathbb{E}\{S_n\}\right) = U_n + V_n.$$

We prove Theorem 1 by showing that for any $u, v \in \mathbb{R}$

$$\mathbb{P}\{U_n \le u, V_n \le v\} \to \Phi\left(\frac{u}{\sigma_1}\right)\Phi\left(\frac{v}{\sigma_2}\right) \tag{8}$$

where Φ denotes the standard normal distribution function, and

$$\sigma_1^2 = \int M_2(x)m(x)^2\mu(dx) - \left(\int m(x)^2\mu(dx)\right)^2 \tag{9}$$

and

$$\sigma_2^2 = \int M_2(x)m(x)^2\mu(dx) - \int m(x)^4\mu(dx). \tag{10}$$

Notice that V_n is measurable with respect to Z_n, therefore

$$\begin{aligned}
&\left|\mathbb{P}\{U_n \le u, V_n \le v\} - \Phi\left(\frac{u}{\sigma_1}\right)\Phi\left(\frac{v}{\sigma_2}\right)\right| \\
&= \left|\mathbb{E}\{\mathbb{I}_{\{V_n \le v\}}\mathbb{P}\{U_n \le u \mid Z_n\}\} - \Phi\left(\frac{u}{\sigma_1}\right)\Phi\left(\frac{v}{\sigma_2}\right)\right| \\
&\le \left|\mathbb{E}\left\{\mathbb{I}_{\{V_n \le v\}}\left(\mathbb{P}\{U_n \le u \mid Z_n\} - \Phi\left(\frac{u}{\sigma_1}\right)\right)\right\}\right| \\
&+ \left|\left(\mathbb{P}\{V_n \le v\} - \Phi\left(\frac{v}{\sigma_2}\right)\right)\Phi\left(\frac{u}{\sigma_1}\right)\right| \\
&\le \mathbb{E}\left\{\left|\mathbb{P}\{U_n \le u \mid Z_n\} - \Phi\left(\frac{u}{\sigma_1}\right)\right|\right\} + \left|\mathbb{P}\{V_n \le v\} - \Phi\left(\frac{v}{\sigma_2}\right)\right|.
\end{aligned}$$

Thus, (8) is satisfied if

$$\mathbb{P}\{U_n \le u \mid Z_n\} \to \Phi\left(\frac{u}{\sigma_1}\right) \tag{11}$$

in probability and

$$\mathbb{P}\{V_n \le v\} \to \Phi\left(\frac{v}{\sigma_2}\right). \tag{12}$$

Proof of (11).
Let's start with the representation

$$\begin{aligned}
U_n &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i' m_n(X_i') - \mathbb{E}\{Y_i' m_n(X_i') \mid Z_n\})\right) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Y_i' m_n(X_i') - \mathbb{E}\{Y_i' m_n(X_i') \mid Z_n\}).
\end{aligned}$$

Because of (7) and the Jensen inequality, for any $1 \le s \le 3$, we get

$$M_s(X) := \mathbb{E}\{|Y|^s \mid X\} = (\mathbb{E}\{|Y|^s \mid X\}^{1/s})^s \le (\mathbb{E}\{|Y|^3 \mid X\}^{1/3})^s \le C^{s/3}, \tag{13}$$

especially, for $s = 1$

$$M_1(X) = |m(X)| \le C^{1/3}$$

and

$$\mathbb{E}\{|Y|^3\} \le C.$$

Next we apply a Berry-Esseen type central limit theorem (see Theorem 14 in Petrov (1975)). It implies that

$$\left|\mathbb{P}\{U_n \le u \mid Z_n\} - \Phi\left(\frac{u}{\sqrt{\mathbb{V}ar(Y_1' m_n(X_1') \mid Z_n)}}\right)\right| \le \frac{c}{\sqrt{n}} \frac{\mathbb{E}\{|Y_1' m_n(X_1')|^3 \mid Z_n\}}{\sqrt{\mathbb{V}ar(Y_1' m_n(X_1') \mid Z_n)}^3}$$

with the universal constant $c > 0$. Because of

$$\mathbb{E}\{Y_1' m_n(X_1') \mid Z_n\} = \int m(x) m_n(x) \mu(dx),$$

we get that

$$\begin{aligned}\mathbb{V}ar(Y_1' m_n(X_1') \mid Z_n) &= \mathbb{E}\{Y_1'^2 m_n(X_1')^2 \mid Z_n\} - \mathbb{E}\{Y_1' m_n(X_1') \mid Z_n\}^2 \\ &= \int M_2(x) m_n(x)^2 \mu(dx) - \left(\int m(x) m_n(x) \mu(dx)\right)^2.\end{aligned}$$

Now (4), together with the boundedness of M_2 by (13), implies that

$$\mathbb{V}ar(Y_1' m_n(X_1') \mid Z_n) \to \sigma_1^2$$

in probability, where σ_1^2 is defined by (9). Further

$$\mathbb{E}\{|Y_1' m_n(X_1')|^3 \mid Z_n\} \le C \int |m_n(x)|^3 \mu(dx).$$

Put

$$A_n(x) = A_{n,j} \text{ if } x \in A_{n,j}.$$

Again, applying the Jensen inequality we get

$$|m_n(x)|^3 \le \left|\frac{\sum_{i=1}^n \mathbb{I}_{\{X_i \in A_n(x)\}} |Y_i|^{3/2}}{\sum_{i=1}^n \mathbb{I}_{\{X_i \in A_n(x)\}}}\right|^2,$$

the right hand side of which is the square of the regression estimate, where Y is replaced by $|Y|^{3/2}$. Thus, (4) together with $\mathbb{E}\{|Y|^3\} < \infty$ implies that

$$\int \left|\frac{\sum_{i=1}^n \mathbb{I}_{\{X_i \in A_n(x)\}} |Y_i|^{3/2}}{\sum_{i=1}^n \mathbb{I}_{\{X_i \in A_n(x)\}}}\right|^2 \mu(dx) \to \mathbb{E}\{\mathbb{E}\{|Y|^{3/2} \mid X\}^2\} < C$$

in probability. These limit relations imply (11).

Proof of (12).
Assuming that the support S of μ is bounded, let l_n be such that $S \subset \cup_{j=1}^{l_n} A_{n,j}$. Also we re-index the partition so that

$$\mu(A_{n,j}) \geq \mu(A_{n,j+1}),$$

with $\mu(A_{n,j}) > 0$ for $j \leq l_n$, and $\mu(A_{n,j}) = 0$ otherwise. Then,

$$S_n = \sum_{j=1}^{l_n} \nu_n'(A_{n,j}) \frac{\nu_n(A_{n,j})}{\mu_n(A_{n,j})}, \tag{14}$$

and

$$l_n \leq \frac{c}{h_n^d}.$$

The condition $nh_n^d \to \infty$ implies that

$$l_n/n \to 0.$$

Because of (14) we have that

$$\begin{aligned} V_n &= \sqrt{n} \sum_{j=1}^{l_n} \mathbb{E}\{\nu_n'(A_{n,j}) \mid Z_n\} \left(\frac{\nu_n(A_{n,j})}{\mu_n(A_{n,j})} - \mathbb{E}\left\{ \frac{\nu_n(A_{n,j})}{\mu_n(A_{n,j})} \right\} \right) \\ &= \sqrt{n} \sum_{j=1}^{l_n} \nu(A_{n,j}) \left(\frac{\nu_n(A_{n,j})}{\mu_n(A_{n,j})} - \mathbb{E}\left\{ \frac{\nu_n(A_{n,j})}{\mu_n(A_{n,j})} \right\} \right), \end{aligned}$$

where

$$\nu(A) = \mathbb{E}\{\nu_n(A)\}.$$

Observe that we have to show the asymptotic normality for a finite sum of dependent random variables. In order to prove (12), we follow the lines of the proof in Beirlant and Györfi (1998) and use a Poissonization argument. With this we introduce a modification M_n of V_n such that

$$\Delta_n := V_n - M_n \to 0,$$

the proof of which follows, starting from (23).

Now we proceed arguing for M_n. Introduce the notation N_n for a Poisson(n) random variable independent of $(X_1, Y_1), (X_2, Y_2), \ldots$. Moreover put

$$n\tilde{\nu}_n(A) = \sum_{i=1}^{N_n} I_{\{X_i \in A\}} Y_i$$

and

$$n\tilde{\mu}_n(A) = \sum_{i=1}^{N_n} I_{\{X_i \in A\}}.$$

The key result in this step is the following property:

Proposition 2 *(Beirlant and Mason (1995), Beirlant et al. (1994).) Put*

$$\tilde{M}_n = \sqrt{n} \sum_{j=1}^{l_n} \nu(A_{n,j}) \left(\frac{\tilde{\nu}_n(A_{n,j})}{\tilde{\mu}_n(A_{n,j})} - \mathbb{E}\left\{ \frac{\tilde{\nu}_n(A_{n,j})}{\tilde{\mu}_n(A_{n,j})} \right\} \right), \tag{15}$$

and

$$M_n = \sqrt{n} \sum_{j=1}^{l_n} \nu(A_{n,j}) \left(\frac{\nu_n(A_{n,j})}{\mu_n(A_{n,j})} - \mathbb{E}\left\{ \frac{\tilde{\nu}_n(A_{n,j})}{\tilde{\mu}_n(A_{n,j})} \right\} \right). \tag{16}$$

Assume that

$$\Phi_n(t,v) = \mathbb{E}\left(\exp\left(it\tilde{M}_n + iv\frac{N_n - n}{\sqrt{n}} \right) \right) \to e^{-(t^2\rho^2+v^2)/2}$$

for a constant $\rho > 0$, where $i = \sqrt{-1}$. Then

$$M_n/\rho \xrightarrow{\mathcal{D}} N(0,1).$$

Put

$$T_n = t\tilde{M}_n + v\frac{N_n - n}{\sqrt{n}},$$

for which a central limit result is to hold:

$$T_n \xrightarrow{\mathcal{D}} N\left(0, t^2\rho^2 + v^2\right) \tag{17}$$

as $n \to \infty$. Remark that

$$\mathbb{V}ar(T_n) = t^2\mathbb{V}ar(\tilde{M}_n) + 2tv\mathbb{E}\left\{ \tilde{M}_n \frac{N_n - n}{\sqrt{n}} \right\} + v^2.$$

For a cell $A = A_{n,j}$ from the partition with $\mu(A) > 0$, let $Y(A)$ be a random variable such that

$$\mathbb{P}\{Y(A) \in B\} = \mathbb{P}\{Y \in B | X \in A\},$$

where B is an arbitrary Borel set.

Introduce the notations

$$q_{n,k} = \mathbb{P}\{n\mu_n(A) = k\} = \binom{n}{k} \mu(A)^k (1-\mu(A))^{n-k}$$

and

$$\tilde{q}_{n,k} = \mathbb{P}\{n\tilde{\mu}_n(A) = k\} = \frac{(n\mu(A))^k}{k!} e^{-n\mu(A)}.$$

Concerning the expectation, with $(Y_1(A), Y_2(A), \dots)$ an i.i.d. sequence of random variables distributed as $Y(A)$ we find that

$$\begin{aligned}
\mathbb{E}\left\{ \frac{\tilde{\nu}_n(A)}{\tilde{\mu}_n(A)} \right\} &= \sum_{k=0}^{\infty} \mathbb{E}\left\{ \frac{\tilde{\nu}_n(A)}{\tilde{\mu}_n(A)} \mid n\tilde{\mu}_n(A) = k \right\} \mathbb{P}\{n\tilde{\mu}_n(A) = k\} \\
&= \sum_{k=1}^{\infty} \mathbb{E}\left\{ \frac{\sum_{i=1}^k Y_i(A)}{k} \right\} \tilde{q}_{n,k} \\
&= \mathbb{E}\left\{ Y_1(A) \right\} (1 - \tilde{q}_{n,0}) \\
&= \frac{\nu(A)}{\mu(A)} (1 - \tilde{q}_{n,0}),
\end{aligned} \tag{18}$$

further, by (24)

$$\mathbb{E}\left\{\frac{\nu_n(A)}{\mu_n(A)}\right\} = n\mathbb{E}\left\{\frac{Y_n(A)}{1+(n-1)\mu_{n-1}(A)}\right\} = \frac{\nu(A)}{\mu(A)}(1-(1-\mu(A))^n)), \tag{19}$$

Moreover,

$$\begin{aligned}
\mathbb{E}\left\{\frac{\tilde{\nu}_n(A)^2}{\tilde{\mu}_n(A)^2}\right\} &= \sum_{k=0}^{\infty}\mathbb{E}\left\{\frac{\tilde{\nu}_n(A)^2}{\tilde{\mu}_n(A)^2} \mid n\tilde{\mu}_n(A)=k\right\}\mathbb{P}\{n\tilde{\mu}_n(A)=k\} \\
&= \sum_{k=1}^{\infty}\mathbb{E}\left\{\frac{\left(\sum_{i=1}^{k} Y_i(A)\right)^2}{k^2}\right\}\tilde{q}_{n,k} \\
&= \sum_{k=1}^{\infty}\frac{k\mathbb{E}\left\{Y_1(A)^2\right\}+k(k-1)\mathbb{E}\left\{Y_1(A)\right\}^2}{k^2}\tilde{q}_{n,k} \\
&= \mathbb{V}ar\left(Y_1(A)\right)\sum_{k=1}^{\infty}\frac{1}{k}\tilde{q}_{n,k}+\mathbb{E}\left\{Y_1(A)\right\}^2(1-\tilde{q}_{n,0}),
\end{aligned}$$

and

$$\begin{aligned}
\sum_{k=1}^{\infty}\frac{1}{k}\tilde{q}_{n,k} &= \sum_{k=1}^{\infty}\frac{1}{k}\frac{(n\mu(A))^k}{k!}e^{-n\mu(A)} \\
&= \sum_{k=1}^{\infty}\frac{1}{k+1}\frac{(n\mu(A))^k}{k!}e^{-n\mu(A)}+\sum_{k=1}^{\infty}\frac{1}{k(k+1)}\frac{(n\mu(A))^k}{k!}e^{-n\mu(A)} \\
&\leq \frac{1}{n\mu(A)}(1-\tilde{q}_{n,0})+\frac{3}{n^2\mu(A)^2}(1-\tilde{q}_{n,0}).
\end{aligned}$$

The independence of the Poisson masses over different cells leads to

$$\begin{aligned}
\mathbb{V}ar(\tilde{M}_n) &= n\sum_{j=1}^{l_n}\nu(A_{n,j})^2\mathbb{V}ar\left(\frac{\tilde{\nu}_n(A_{n,j})}{\tilde{\mu}_n(A_{n,j})}\right) \\
&\leq n\sum_{j=1}^{l_n}\nu(A_{n,j})^2\Big(\mathbb{V}ar\left(Y_1(A_{n,j})\right)\big(\frac{1}{n\mu(A_{n,j})}(1-e^{-n\mu(A_{n,j})}) \\
&\quad+\frac{3}{n^2\mu(A_{n,j})^2}(1-e^{-n\mu(A_{n,j})})\big) \\
&\quad+\mathbb{E}\left\{Y_1(A_{n,j})\right\}^2(1-e^{-n\mu(A_{n,j})})-\mathbb{E}\left\{Y_1(A_{n,j})\right\}^2(1-e^{-n\mu(A_{n,j})})^2\Big) \\
&\leq \sum_{j=1}^{l_n}\frac{\nu(A_{n,j})^2}{\mu(A_{n,j})^2}\mathbb{V}ar\left(Y_1(A_{n,j})\right)\mu(A_{n,j}) \\
&\quad+\sum_{j=1}^{l_n}\frac{3\mathbb{V}ar\left(Y_1(A_{n,j})\right)\nu(A_{n,j})^2}{n\mu(A_{n,j})^2} \\
&\quad+n\sum_{j=1}^{l_n}\nu(A_{n,j})^2\mathbb{E}\left\{Y_1(A_{n,j})\right\}^2e^{-n\mu(A_{n,j})}\Big)
\end{aligned}$$

such that the bounding error in these inequalities is of order $O(l_n/n)$. (4) together with the boundedness of M_2 and m implies that

$$\begin{aligned}
&\sum_{j=1}^{l_n} \frac{\nu(A_{n,j})^2}{\mu(A_{n,j})^2} \mathbb{V}ar\left(Y_1(A_{n,j})\right) \mu(A_{n,j}) \\
&= \int \frac{\int_{A_n(x)} M_2(z)\mu(dz)}{\mu(A_n(x))} \left(\frac{\int_{A_n(x)} m(z)\mu(dz)}{\mu(A_n(x))} \right)^2 \mu(dx) - \int \left(\frac{\int_{A_n(x)} m(z)\mu(dz)}{\mu(A_n(x))} \right)^4 \mu(dx) \\
&= \sigma_2^2 + o(1),
\end{aligned}$$

where σ_2^2 is defined by (10). Moreover,

$$\sum_{j=1}^{l_n} \frac{3\mathbb{V}ar\left(Y_1(A_{n,j})\right) \nu(A_{n,j})^2}{n\mu(A_{n,j})^2} \leq \frac{3C^{4/3} l_n}{n} \to 0.$$

Then

$$\begin{aligned}
& n \sum_{j=1}^{l_n} \nu(A_{n,j})^2 \mathbb{E}\left\{Y_1(A_{n,j})\right\}^2 e^{-n\mu(A_{n,j})} \\
&= \sum_{j=1}^{l_n} \frac{\nu(A_{n,j})^2}{\mu(A_{n,j})^2} \mathbb{E}\left\{Y_1(A_{n,j})\right\}^2 n\mu(A_{n,j}) e^{-n\mu(A_{n,j})} \mu(A_{n,j}) \\
&\leq C^{4/3} \sum_{j=1}^{l_n} n\mu(A_{n,j})^2 e^{-n\mu(A_{n,j})} \\
&\leq C^{4/3} (\max_{z>0} z^2 e^{-z}) l_n/n \to 0.
\end{aligned}$$

So we proved that

$$\mathbb{V}ar(\tilde{M}_n) \to \sigma_2^2.$$

To complete the asymptotics for $\mathbb{V}ar(T_n)$, it remains to show that

$$\mathbb{E}\left\{ \tilde{M}_n \frac{N_n - n}{\sqrt{n}} \right\} \to 0 \text{ as } n \to \infty.$$

Because of

$$N_n = n \sum_{j=1}^{l_n} \tilde{\mu}_n(A_{n,j})$$

and

$$n = n \sum_{j=1}^{l_n} \mu(A_{n,j}),$$

we have that

$$
\begin{aligned}
&\mathbb{E}\left\{\tilde{M}_n \frac{N_n - n}{\sqrt{n}}\right\} \\
&= n\sum_{j=1}^{l_n} \mathbb{E}\left\{\frac{\tilde{\nu}_n(A_{n,j})}{\tilde{\mu}_n(A_{n,j})}\nu(A_{n,j})(\tilde{\mu}_n(A_{n,j}) - \mu(A_{n,j}))\right\} \\
&= n\sum_{j=1}^{l_n} \nu(A_{n,j})\left(\mathbb{E}\left\{\tilde{\nu}_n(A_{n,j})\right\} - \mathbb{E}\left\{\frac{\tilde{\nu}_n(A_{n,j})}{\tilde{\mu}_n(A_{n,j})}\right\}\mu(A_{n,j})\right) \\
&= n\sum_{j=1}^{l_n} \nu(A_{n,j})\left(\nu(A_{n,j}) - \frac{\nu(A_{n,j})}{\mu(A_{n,j})}(1 - e^{-n\mu(A_{n,j})})\mu(A_{n,j})\right) \\
&= n\sum_{j=1}^{l_n} \nu(A_{n,j})^2 e^{-n\mu(A_{n,j})} \\
&\leq C^{2/3}(\max_{z>0} z^2 e^{-z}) l_n/n \to 0.
\end{aligned}
$$

To finish the proof of (17) by Lyapunov's central limit theorem, it suffices to prove that

$$
n^{3/2}\sum_{j=1}^{l_n}\mathbb{E}\Big\{\Big|t\left(\frac{\tilde{\nu}_n(A_{n,j})}{\tilde{\mu}_n(A_{n,j})} - \mathbb{E}\left\{\frac{\tilde{\nu}_n(A_{n,j})}{\tilde{\mu}_n(A_{n,j})}\right\}\right)\nu(A_{n,j}) + v\left(\tilde{\mu}_n(A_{n,j}) - \mu(A_{n,j})\right)\Big|^3\Big\} \to 0
$$

or, by invoking the c_3 inequality $|a+b|^3 \leq 4(|a|^3 + |b|^3)$, that

$$
n^{3/2}\sum_{j=1}^{l_n}\mathbb{E}\left\{\left|\frac{\tilde{\nu}_n(A_{n,j})}{\tilde{\mu}_n(A_{n,j})} - \mathbb{E}\left\{\frac{\tilde{\nu}_n(A_{n,j})}{\tilde{\mu}_n(A_{n,j})}\right\}\right|^3\right\}\nu(A_{n,j})^3 \to 0 \tag{20}
$$

and

$$
n^{3/2}\sum_{j=1}^{l_n}\mathbb{E}\left\{|\tilde{\mu}_n(A_{n,j}) - \mu(A_{n,j})|^3\right\} \to 0. \tag{21}
$$

In view of (20), because of (13) it suffices to prove

$$
D_n := n^{3/2}\sum_{j=1}^{l_n}\mathbb{E}\left\{\left|\frac{\tilde{\nu}_n(A_{n,j})}{\tilde{\mu}_n(A_{n,j})} - \mathbb{E}\left\{\frac{\tilde{\nu}_n(A_{n,j})}{\tilde{\mu}_n(A_{n,j})}\right\}\right|^3\right\}\mu(A_{n,j})^3 \to 0 \tag{22}
$$

For a cell A, (18) implies that

$$
\begin{aligned}
\mathbb{E}\left\{\left|\frac{\tilde{\nu}_n(A)}{\tilde{\mu}_n(A)} - \mathbb{E}\left\{\frac{\tilde{\nu}_n(A)}{\tilde{\mu}_n(A)}\right\}\right|^3\right\} &\leq 4\mathbb{E}\left\{\left|\frac{\tilde{\nu}_n(A)}{\tilde{\mu}_n(A)} - \frac{\nu(A)}{\mu(A)}(1-\tilde{q}_{n,0})\mathbb{I}_{\{\tilde{\mu}_n(A)>0\}}\right|^3\right\} \\
&\quad + 4\mathbb{E}\left\{\left|\frac{\nu(A)}{\mu(A)}(1-\tilde{q}_{n,0})\mathbb{I}_{\{\tilde{\mu}_n(A)>0\}} - \frac{\nu(A)}{\mu(A)}(1-\tilde{q}_{n,0})\right|^3\right\}.
\end{aligned}
$$

On the one hand, (18), (13) and (25) imply that, for a constant K,

$$\begin{aligned}
&\mathbb{E}\left\{\left|\frac{\tilde{\nu}_n(A)}{\tilde{\mu}_n(A)}-\frac{\nu(A)}{\mu(A)}(1-\tilde{q}_{n,0})\mathbb{I}_{\{\tilde{\mu}_n(A)>0\}}\right|^3\right\}\\
&=\sum_{k=0}^{\infty}\mathbb{E}\left\{\left|\frac{\tilde{\nu}_n(A)}{\tilde{\mu}_n(A)}-\frac{\nu(A)}{\mu(A)}(1-\tilde{q}_{n,0})\mathbb{I}_{\{\tilde{\mu}_n(A)>0\}}\right|^3 \mid n\tilde{\mu}_n(A)=k\right\}\mathbb{P}\{n\tilde{\mu}_n(A)=k\}\\
&=\sum_{k=1}^{\infty}\mathbb{E}\left\{\frac{\left|\sum_{i=1}^{k}(Y_i(A)-\mathbb{E}\{Y_i(A)\})\right|^3}{k^3}\right\}\tilde{q}_{n,k}\\
&\leq K\sum_{k=1}^{\infty}\frac{1}{k^{3/2}}\tilde{q}_{n,k}\\
&\leq c_1\frac{1}{n^{3/2}\mu(A)^{3/2}},
\end{aligned}$$

where we applied the Marcinkiewicz and Zygmund (1937) inequality for absolute central moments of sums of i.i.d. random variables. On the other hand

$$\mathbb{E}\left\{\left|\frac{\nu(A)}{\mu(A)}(1-\tilde{q}_{n,0})\mathbb{I}_{\{\tilde{\mu}_n(A)>0\}}-\frac{\nu(A)}{\mu(A)}(1-\tilde{q}_{n,0})\right|^3\right\}\leq C\tilde{q}_{n,0}.$$

Therefore

$$\begin{aligned}
D_n &\leq n^{3/2}c_2\sum_{j=1}^{l_n}\left(\frac{1}{n^{3/2}\mu(A_{n,j})^{3/2}}+e^{-n\mu(A_{n,j})}\right)\mu(A_{n,j})^3\\
&\leq c_2\left(\sum_{j=1}^{l_n}\mu(A_{n,j})^{3/2}+\sum_{j=1}^{l_n}n^{3/2}e^{-n\mu(A_{n,j})}\mu(A_{n,j})^3\right)\\
&\leq c_2\sum_{j=1}^{l_n}\mu(A_{n,j})^{3/2}\left(1+\max_{z>0}z^{3/2}e^{-z}\right)\\
&= c_3\int\mu(A_n(x))^{1/2}\mu(dx)\\
&\to 0,
\end{aligned}$$

where we used the assumption that μ is non-atomic. Thus, (20) is proved.

The proof of (21) is easier. Notice that (21) means

$$F_n := n^{-3/2}\sum_{j=1}^{l_n}\mathbb{E}\left\{\left|\sum_{i=1}^{N_n}\mathbb{I}_{\{X_i\in A_{n,j}\}}-n\mu(A_{n,j})\right|^3\right\}\to 0.$$

One has

$$
\begin{aligned}
&\mathbb{E}\left\{\left|\sum_{i=1}^{N_n}\mathbb{I}_{\{X_i\in A_{n,j}\}}-n\mu(A_{n,j})\right|^3\right\}\\
&\le 4\mathbb{E}\left\{\left|\sum_{i=1}^{N_n}(\mathbb{I}_{\{X_i\in A_{n,j}\}}-\mu(A_{n,j}))\right|^3\right\}+4\mathbb{E}\left\{|(N_n-n)\mu(A_{n,j})|^3\right\}\\
&\le c_4\left(\sum_{k=1}^{\infty}k^{3/2}\mu(A_{n,j})^{3/2}e^{-n}\frac{n^k}{k!}+\mathbb{E}\left\{|N_n-n|^3\right\}\mu(A_{n,j})^3\right)\\
&\le c_5\left(n^{3/2}\mu(A_{n,j})^{3/2}+n^{3/2}\mu(A_{n,j})^3\right).
\end{aligned}
$$

Therefore

$$
F_n\le 2c_5\sum_{j=1}^{l_n}\mu(A_{n,j})^{3/2}\to 0,
$$

and so (21) is proved, too.

The remaining step in the proof of (12) is to show that

$$
\Delta_n:=V_n-M_n=n^{1/2}\sum_{j=1}^{l_n}\left(\mathbb{E}\left\{\frac{\tilde{\nu}_n(A_{n,j})}{\tilde{\mu}_n(A_{n,j})}\right\}-\mathbb{E}\left\{\frac{\nu_n(A_{n,j})}{\mu_n(A_{n,j})}\right\}\right)\nu(A_{n,j})\to 0. \tag{23}
$$

By (18) and (19) have that

$$
\begin{aligned}
|\Delta_n|&=\left|n^{1/2}\sum_{j=1}^{l_n}\frac{\nu(A_{n,j})}{\mu(A_{n,j})}(e^{-n\mu(A_{n,j})}-(1-\mu(A_{n,j}))^n)\nu(A_{n,j})\right|\\
&=n^{1/2}\sum_{j=1}^{l_n}\frac{\nu(A_{n,j})^2}{\mu(A_{n,j})^2}(e^{-n\mu(A_{n,j})}-(1-\mu(A_{n,j}))^n)\mu(A_{n,j})\\
&\le C^{2/3}n^{1/2}\sum_{j=1}^{l_n}(e^{-n\mu(A_{n,j})}-(1-\mu(A_{n,j}))^n)\mu(A_{n,j}).
\end{aligned}
$$

For $0\le z\le 1$, using the elementary inequalities

$$
1-z\le e^{-z}\le 1-z+z^2
$$

we have that

$$
e^{-nz}-(1-z)^n=(e^{-z}-(1-z))\sum_{k=0}^{n-1}e^{-kz}(1-z)^{n-1-k}\le nz^2e^{-(n-1)z},
$$

and thus we get that

$$\begin{aligned}
|\Delta_n| &\leq C^{2/3} n^{1/2} \sum_{j=1}^{l_n} (e^{-n\mu(A_{n,j})} - (1-\mu(A_{n,j}))^n)\mu(A_{n,j}) \\
&\leq C^{2/3} n^{1/2} \sum_{j=1}^{l_n} n\mu(A_{n,j})^3 e^{-(n-1)\mu(A_{n,j})} \\
&\leq \frac{C^{2/3}}{n^{1/2}} \sum_{j=1}^{l_n} \mu(A_{n,j}) \left([n\mu(A_{n,j})]^2 e^{-n\mu(A_{n,j})} \right) e \\
&\leq \frac{C^{2/3}}{n^{1/2}} \sum_{j=1}^{l_n} \mu(A_{n,j}) \max_{z\geq 0}(z^2 e^{-z}) e \\
&\to 0.
\end{aligned}$$

This ends the proof of (12) and so the proof of Theorem 1 is complete. ∎

Next we give two lemmas, which are used above.

Lemma 3 *If $B(n,p)$ is a binomial random variable with parameters (n,p), then*

$$\mathbb{E}\left\{ \frac{1}{1+B(n,p)} \right\} = \frac{1-(1-p)^{n+1}}{(n+1)p}. \tag{24}$$

Lemma 4 *If $Po(\lambda)$ is a Poisson random variable with parameter λ, then*

$$\mathbb{E}\left\{ \frac{1}{Po(\lambda)^3} \mathbb{I}_{\{Po(\lambda)>0\}} \right\} \leq \frac{24}{\lambda^3}. \tag{25}$$

References

J. Beirlant and L. Györfi. On the asymptotic L_2-error in partitioning regression estimation. *Journal of Statistical Planning and Inference*, 71:93–107, 1998.

J. Beirlant and D. Mason. On the asymptotic normality of l_p-norms of empirical functionals. *Mathematical Methods of Statistics*, 4:1–19, 1995.

J. Beirlant, L. Györfi, and G. Lugosi. On the asymptotic normality of the l_1- and l_2- errors in histogram density estimation. *Canadian J. Statistics*, 22:309–318, 1994.

K. De Brabanter, P. G. Ferrario, and L. Györfi. Detecting ineffective features for nonparametric regression. In J. A. K. Suykens, M. Signoretto, and A. Argyriou, editors, *Regularization, Optimization, Kernels, and Support Vector Machines*, pages 177–194. Chapman & Hall/CRC Machine Learning and Pattern Recognition Series, 2014.

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

L. Devroye, P. Ferrario, L. Györfi, and H. Walk. Strong universal consistent estimate of the minimum mean squared error. In B. Schölkopf, Z. Luo, and V. Vovk, editors, *Empirical Inference - Festschrift in Honor of Vladimir N. Vapnik*, pages 143–160. Springer, Heidelberg, 2013.

D. Evans and A. J. Jones. Non-parametric estimation of residual moments and covariance. *Proceedings of the Royal Society*, A 464:2831–2846, 2008.

P. G. Ferrario and H. Walk. Nonparametric partitioning estimation of residual and local variance based on first and second nearest neighbors. *Journal of Nonparametric Statistics*, 24:1019–1039, 2012.

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

E. Liitiäinen, F. Corona, and A. Lendasse. On nonparametric residual variance estimation. *Neural Processing Letters*, 28:155–167, 2008.

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

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

J. Marcinkiewicz and A. Zygmund. Sur les fonctions indépendantes. *Fundamenta Mathematicae*, 29:60–90, 1937.

V. V. Petrov. *Sums of Independent Random Variables*. Springer-Verlag, Berlin, 1975.