Journal of Multivariate Analysis 101 (2010) 811–823

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Residual variance estimation using a nearest neighbor statistic

Elia Liitiäinen *, Francesco Corona, Amaury Lendasse

Department of Information and Computer Science, Helsinki University of Technology, P.O. Box 5400 HUT, Finland

ARTICLE INFO

Article history:
Received 22 May 2008
Available online 13 January 2010

AMS subject classification:
62G05
62G20

Keywords:
Residual variance estimation
Noise variance estimation
Nearest neighbors
Nonparametric

ABSTRACT

In this paper we consider the problem of estimating $E[(Y - E[Y \mid X])^2]$ based on a finite sample of independent, but not necessarily identically distributed, random variables $(X_i, Y_i)_{i=1}^M$. We analyze the theoretical properties of a recently developed estimator. It is shown that the estimator has many theoretically interesting properties, while the practical implementation is simple.

1. Introduction

Let $(Z_i)_{i=1}^M = (X_i, Y_i)_{i=1}^M$ be independent identically distributed (i.i.d.) random variables with values in $\Re^d \times \Re$ and assume that the variables are generated by the model

$$Y = f(X) + r \tag{1}$$

for homoscedastic mean zero noise r independent of X and the residual variance $V = \mathrm{Var}[r]$. Estimating the residual variance V is a well-known problem in statistics and especially for the case $d = 1$ many estimators exist. The most straightforward idea is to first approximate f using a regression estimate with M samples $\hat{f}_M$:

$$\frac{1}{M}\sum_{i=1}^{M}(Y_i - \hat{f}_M(X_i))^2$$

and use the resulting function to approximate the residuals. However, to see the difficulties arising when $d > 2$, notice that because of

$$E[(Y - \hat{f}_M(X))^2] = E[(f(X) - \hat{f}_M(X))^2] + V,$$

the rate of convergence is determined by the estimate of f. Assume now that the variables X and r are bounded. Then a classical result of Stone [1] implies that for any nonparametric regression estimator, there exists a sequence (f_M) such that each function in the sequence has the same Lipschitz constant and

$$\liminf_{M\to\infty} M^{2/(2+d)} E[(f_M(X) - \hat{f}_M(X))^2] > 0,$$

* Corresponding author.
E-mail address: elia.liitiainen@hut.fi (E. Liitiäinen).

0047-259X/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2009.12.020

with $\hat{f}_M$ the approximation of f_M based on the sample $(Z_i)_{i=1}^M$. Thus, already for $d = 3$, the error is of order $M^{-2/5}$. However, in [2] it has been shown that the Lipschitz continuity implies that the rate $M^{-1/2}$ is achievable when $d \leq 4$.

A better idea is to estimate V directly without the intermediate step of approximating f or use some kind of a bias correction. For $d = 1$, difference based methods are known to obtain low biases [3,4]. Other approaches include the use of U-statistics [5], least squares [6] and kernel estimators [7]. However, the case $d > 1$ is much less studied and for example the generalization of difference based methods for higher dimensional problems with random covariates is not obvious. This problem is addressed elegantly in [8], where a locally linear estimator is derived and shown to achieve the optimal rate $M^{-1/2}$ up to the dimension eight, even though statistical efficiency is not addressed.

A natural generalization of the model (1) is introducing heteroscedasticity in the noise, that is,

$$\mathrm{Var}[Y|X] = V(X)$$

for some variance function $V(x)$. In the literature it has been shown that the conditional variance function $V(x)$ can be estimated efficiently from data and many methods and theoretical results exist both for the cases $d = 1$ and $d > 1$ [9–15]. It is clear that estimating the whole function is a significantly more difficult task than estimating the variance of homoscedastic noise. However, in many application areas, estimating the simpler quantity

$$E[V(X)] = E[(Y - f(X))^2] \tag{2}$$

is of equal interest. The usefulness of this expectation comes from the fact that it is the minimum mean squared error achievable by a regression estimator and thus a natural criterion for example in the feature selection problem [2] and in general to assess how good a prediction of Y is possible with X as the covariate. Even though Eq. (2) is a straightforward generalization of (1), methods designed for direct estimation of homoscedastic noise variance cannot be applied straightforwardly to the generalized problem. Previous work on estimating (2) includes [2,16].

In this work we analyze a method that estimates (2) using a nearest neighbor statistic, which has been introduced in [17,18]. The method is shown to be consistent and the rate of convergence is analyzed. In contrast to the earlier works [2,16], we are able to show under sufficient regularity and the condition $d \leq 4$, the asymptotic bound ($\hat{V}_{M,k}$ denotes the nearest neighbor estimator)

$$\limsup_{M\to\infty} ME\left[\left(\hat{V}_{M,k} - \frac{1}{M}\sum_{i=1}^{M} r_i^2\right)^2\right] \leq g(k,d)\sigma^4 \tag{3}$$

for a universal constant $g(k, d)$ (to be specified later) decreasing in the free parameter $k > 0$. Here σ refers to an upper bound on $E[r_i^4|X_i]$. The results are significant also in the theory of homoscedastic noise variance estimation, as for example the corresponding asymptotic results in [5,7] apply only to the case $d = 1$. Compared to [2], the i.i.d. assumption is relaxed on the covariates. In addition, the practical implementation of the method is straightforward, as it is based on the use of nearest neighbors and contains only one free integer parameter, which does not affect the rate of convergence.

2. Residual variance estimation by nearest neighbors

2.1. A formal statement of the problem

The problem of residual variance estimation can be stated in a general form as estimating the optimal mean squared error given a finite sample of data [2]. Our basic assumption is that the variables $(Z_i)_{i=1}^\infty$ are independent (but not necessarily identically distributed as in the beginning of the introduction) and the stationarity condition that for some measurable function m,

$$m(x) = E[Y_i|X_i = x] \tag{4}$$

for all i. The model (1) still holds formally be setting $r_i = Y_i - E[Y_i|X_i]$, but in general r may well depend on X. Adopting this notation, in this setting the Borel measurable function m minimizes the mean squared cost

$$V_M = \frac{1}{M}\sum_{i=1}^{M} E[(m(X_i) - Y_i)^2] = \frac{1}{M}\sum_{i=1}^{M} E[r_i^2].$$

Then V_M is the residual variance. Our definition contains heteroscedastic noise as a special case and also allows deterministic covariates. For an estimator $\hat{V}_M$, we will be interested in the mean squared deviation given by

$$E[(V_M - \hat{V}_M)^2].$$

In addition, we will address statistical efficiency by proving the asymptotic bound (3) for our estimator.

For the theoretical analysis, we require all the time the moment condition

$$\sup_{i>0} E[Y_i^4|X_i] \leq K_Y^4 \tag{5}$$

for some constant $K_Y > 0$, which implies that $|m(x)| \leq K_Y$ and $E[r_i^4|X_i] \leq \sigma^4 \leq 16K_Y^4$.

2.2. A heuristic derivation of the estimator

The concept of nearest neighbor is well understood in the literature on computational geometry, machine learning and statistics [19–21]. The nearest neighbor of the point X_i is defined simply as the point closest to it with respect to a similarity measure. Using the Euclidean metric, the formal definition is

$$N[i,1] = \operatorname*{arg\,min}_{1\le j\le M, j\neq i} \|X_i - X_j\|.$$

The kth nearest neighbor is defined recursively as

$$N[i,k] = \operatorname*{arg\,min}_{1\le j\le M, j\neq i, N[i,1],\ldots,N[i,k-1]} \|X_i - X_j\|,$$

that is, the closest point after removal of the preceding neighbors. The corresponding distances are defined as

$$d_{i,k,M} = \|X_i - X_{N[i,k]}\|.$$

Notice that without additional assumptions, these definitions are not necessarily unique as it is possible that two points are at the same distance from X_i. In that case one should use for example randomization [19], which leads to some additional theoretical difficulties. To avoid the problem of ties, we make the assumption that for any three distinct indices $i, l, j > 0$

$$P(\|X_i - X_j\| = \|X_i - X_l\|) = 0, \tag{6}$$

which holds for example when the data is sampled from a density with respect to the Lebesgue measure.

A simple, well-known nonparametric estimator of residual variance [2] is

$$V_M \approx \frac{1}{2M}\sum_{i=1}^{M}(Y_i - Y_{N[i,1]})^2. \tag{7}$$

To clarify the logic behind the estimator, let us assume that the sample $(X_i, Y_i)_{i=1}^M$ is generated by the model $Y = f(X) + r$ for a smooth function f and independent noise r. Now it is reasonable to assume that the points X_i and $X_{N[i,1]}$ are close to each other when the number of observations is high enough and we may approximate heuristically

$$V_M \approx \frac{1}{2M}\sum_{i=1}^{M}(r_i - r_{N[i,1]})^2.$$

Using the assumption that the variables $(r_i)_{i=1}^M$ are independent of the variables $(X_i)_{i=1}^M$ and each other, we may furthermore write

$$E[V_M] \approx \frac{1}{2M}\sum_{i=1}^{M}E[r_i^2] + \frac{1}{2M}\sum_{i=1}^{M}E[r_{N[i,1]}^2] = E[r^2],$$

which is the residual variance. Thus clearly it is possible to prove that the estimator (7) is consistent when the output noise is additive and heteroscedastic. A natural question is, if consistency holds also in a more general setting. The following example from [22] shows that the conditions that are required for convergence are unsatisfying.

Example 2.1. Consider the set of univariate covariates consisting of two distinct parts, $(X_i^1)_{i=1}^{M_1}$ and $(X_i^2)_{i=1}^{2M_1}$ with $X_i^1 = \frac{i}{M_1}$, $X_{2i}^2 = X_i^1 - \frac{1}{4M_1}$ and $X_{2i-1}^2 = X_i^1 + \frac{1}{4M_1}$. The regressands Y_i^1 corresponding to the variables X_i^1 are set as zero mean independent noise with unit variance, whereas for X_i^2 the outputs are set to 0. We set $Y_i = Y_i^1$ when $1 \le i \le M_1$ and $Y_i = Y_i^2$ when $M_1 < i \le M_1 + M_2$. In this case, the approximation (7) gives $\frac{1}{2M}\sum_{i=1}^{M}(Y_{N[i,1]} - Y_i)^2 = \frac{1}{2}$. However, the right answer is 1/3 and thus it is clear that the method is not consistent in this example.

The optimal regression function m in Example 2.1 is 0 and thus trivially smooth. Consequently Example 2.1 shows that the consistency of the estimator (7) requires conditions both on the optimal regression function m and the conditional variance function. The difficulties arise from the fact that $r_{N[i,1]}$ is in general not similarly distributed as r_i. It would be possible to extend the estimator for $k > 1$ by

$$V_M \approx \frac{1}{(1 + k_M^{-1})M}\sum_{i=1}^{M}\left(Y_i - \frac{1}{k_M}\sum_{j=1}^{k_M}Y_{N[i,j]}\right)^2, \tag{8}$$

where the assumption $k_M/M \to 0$ as $M \to \infty$ is essential for consistency. However, even though it is possible to show that the approximation (8) is able to give consistent estimates under general assumptions by using the results for k nearest neighbor regression estimators [23,19], it is not without problems. One practical problem is the choice of k_M, which would be difficult, as for example cross-validation inevitably increases variance.

It is of course possible to approximate m with a local polynomial or a neural network model instead of a simple locally constant approximator. An alternative solution based on modified nearest neighbor graphs was introduced in [2]. In this paper we analyze a slightly simpler method (originally proposed in [17,18] for $k = 1$) based on modifying the approximator (7) as

$$\hat{V}_{M,k} = \frac{1}{Mk^2} \sum_{i=1}^{M} \left(\sum_{j_1=1}^{k} (Y_i - Y_{N[i,2j_1]}) \right) \left(\sum_{j_2=1}^{k} (Y_i - Y_{N[i,2j_2-1]}) \right) \tag{9}$$

for a positive integer k. To understand the logic behind the estimator, assume that the function m is continuous and $k = 1$. Then a heuristic approximation and conditional independence yield

$$V_M \approx E[\hat{V}_{M,k}] = \frac{1}{M} \sum_{i=1}^{M} E[(r_i - r_{N[i,1]})(r_i - r_{N[i,2]})] = E\left[\frac{1}{M} \sum_{i=1}^{M} r_i^2 \right],$$

which is the residual variance. Moreover, it can be seen that the quality of the estimate depends only on the smoothness of m and therefore the estimator is able to solve Example 2.1.

In the next section we formalize the discussion and show that the estimator (9) is indeed consistent. Moreover, analysis of the rate of convergence turns out to be relatively easy due to the simplicity of the method. Surprisingly it also turns out that the new estimator tends to have (at least asymptotically) a smaller bias than the estimator (7) and the one in [2] even in the additive heteroscedastic noise case. While the selection of the optimal value for the free parameter k is difficult, the rate of convergence and consistency are obtained for any fixed value, which is an advantage, that many other noise variance estimators do not share.

After observing that in fact,

$$E\left[\left(\sum_{j_1=1}^{k} (Y_1 - Y_{N[1,2j_1]}) \right) \left(\sum_{j_2=1}^{k} (Y_1 - Y_{N[1,2j_2-1]}) \right) \middle| X_1 \right] = E[r_1^2 | X_1], \tag{10}$$

it seems that the estimator (9) provides a simple approximation to the conditional variance function $E[r_1|X_1 = x]$ as well. But unlike (9), the validity of the approximation (10) relies heavily on the choice of k. Thus while the generalization is straightforward, the practical and theoretical problems concerning the estimation are rather different and remain an unexplored topic.

2.3. Convergence properties

As a first case, we address consistency in the special case of i.i.d. covariates as stated in the next assumption.

(A1) The random variables $(Z_i)_{i=1}^{\infty}$ are identically distributed and conditions (4), (5) and hold (6).

In the following theorem we show L^2-convergence under assumption (A1). Despite its generality, our result is not surprising as similar results exist for the nearest neighbor regression estimate, see for example [19].

Theorem 2.2. *Suppose assumption* (A1) *holds and let k be a positive integer. Then the convergence in mean square*

$$E[(\hat{V}_{M,k} - V_M)^2] \to 0$$

holds as $M \to \infty$.

In addition to asymptotic convergence, it is of interest to investigate rates of convergence. It is clear, that some regularity assumptions are needed as arbitrarily slow convergence is otherwise possible [24,2]. A common, sufficient and rather realistic assumption is Holder continuity or the stronger differentiability condition of m.

Definition 2.3. For $0 < \gamma \leq 1$, we define $H(\gamma, c)$ as the class of bounded functions with the property

$$|f(x) - f(y)| \leq c\|x - y\|^{\gamma}. \tag{11}$$

For $1 < \gamma \leq 2$, we require that $m \in H(1, c)$ and $\partial_i m \in H(\gamma - [\gamma], c)$ for $1 \leq i \leq d$. Here $\partial_i m$ refers to the partial derivatives of m.

The following two assumptions summarize the conditions needed for convergence analysis. Condition (A2) on the moments of the covariates is motivated by the theory of nearest neighbor regression estimates [23]. Let us also observe the fact, that the continuity assumption (A3) needs to hold only in an appropriately chosen set of probability one depending on the context.

(A2) For some constants $c_1 > 0$ and $\beta_2 > 2d$,

$$\sup_{i>0} E[\|X_i\|^{\beta_2}] \leq c_1.$$

Moreover, conditions (4)–(6) hold.

(A3) There exist constants $0 < \gamma \leq 2$ and $c_2 > 0$ such that $m \in H(\gamma, c_2)$.

In the rest of the paper, we denote by β a fixed number with $2d < \beta < \beta_2$. As a detail, notice that by Jensen's inequality and the choice $c_1 \geq 1$,

$$\sup_{i>0} E[\|X_i\|^\beta] \leq c_1.$$

Assumption (A3) implies the existence of a bounded gradient $\|\nabla m\| \leq c_2$ when $\gamma \geq 1$.

The following theorem shows that the theoretically optimal rate of convergence $O(M^{-1/2})$ is achieved as long as $d \leq 4$. Moreover, to demonstrate the effect of the parameter k, an upper bound for the constants is calculated when the covariates are bounded. The quantity $L(d)$ in the bound is defined as the minimum amount of cones ($\|e\| = 1$)

$$C(e) = \{x \in \Re^d : e^T x \leq \|x\| \cos 30^\circ\}$$

needed to cover the space $\Re^d$. For example, $L(1) = 2$ and $L(2) = 6$.

Theorem 2.4. *Suppose assumptions* (A2) *and* (A3) *hold, let k be a positive integer and set $\alpha = \min\{d - d^2/\beta, 2\gamma, 2\}$, $\tilde{\gamma} = \min\{d - d^2/\beta, \gamma, 1\}$. Then there exist constants c_4 and c_5 depending on d, k, c_1, c_2, β_2, σ, K_Y and γ such that*

$$E\left[\left(\hat{V}_{M,k} - \frac{1}{M}\sum_{i=1}^{M} r_i^2\right)^2\right]^{1/2} \leq c_3^{1/2} M^{-1/2} + c_4^{1/2} M^{-1/2-\tilde{\gamma}/2d} + c_5 M^{-\alpha/d}$$

with

$$c_3 = (4k^{-3} + 2k^{-2} + 4L(d)k^{-2} + 2L(d)k^{-1})\sigma^4.$$

Moreover, if the variables $(X_i)_{i=1}^M$ take values in the unit cube $[-1/2, 1/2]^d$ with $d > 1$ and $\gamma \leq 1$, then we may choose $\alpha = 2\gamma$, $\tilde{\gamma} = \gamma$,

$$c_4 = 2^{4+\gamma+\gamma/d} d^\gamma k^{\gamma/d} K_Y \sigma^2 c_2 (4k^2 L(d) + 2kL(d) + 2k + 1)$$

and $c_5 = 2^{2\gamma+2\gamma/d} d^\gamma k^{2\gamma/d} c_2^{2\gamma}$. In the case $d = 1$, these bounds hold for $\gamma \leq 1/2$.

We obtain straightforwardly the following corollary.

Corollary 2.5. *Suppose assumptions* (A2) *and* (A3) *hold, let k be a positive integer and set $\alpha = \min\{d - d^2/\beta, 2\gamma, 2\}$. Then if $\alpha > d/2$,*

$$\limsup_{M\to\infty} ME\left[\left(\hat{V}_{M,k} - \frac{1}{M}\sum_{i=1}^{M} r_i^2\right)^2\right] \leq g(k,d)\sigma^4$$

with

$$g(k,d) = 4k^{-3} + 2k^{-2} + 4L(d)k^{-2} + 2L(d)k^{-1}. \tag{12}$$

This corollary implies also that with an appropriate choice of k increasing with respect to M, the estimator is asymptotically normal and behaves similarly as the optimal estimator $\frac{1}{M}\sum_{i=1}^{M} r_i^2$. However, the selection of the optimal k is by no means easy; for a practical solution without analysis of consistency, see for example [5]. Moreover, it seems likely that a central limit theorem can be proven also for a fixed k; see for example [20,8].

Remark 2.6. The proof of Theorem 2.4 shows that

$$|E[\hat{V}_{M,k} - V_M]| \leq c_5 M^{-\alpha/d}.$$

As an example, notice that in the special case of covariates distributed in the unit cube with $k = 1$, $\gamma = 1$ and $d = 2$, we have $\alpha = 2$ and $c_5 = 16c_2^2$. In practical inference problems with a small number of samples available, small bias is important as the variance tends to be much smaller than the actual residual variance and the variance of the regressand.

The rate of convergence in Theorem 2.4 is essentially the same as that obtained in [2]. The constants in Theorem 2.4 are suboptimal due to simplicity; however, in any case the theorem gives useful bounds for the asymptotic variance and the systematic error of the estimator.

Next we show that with bounded covariates, the bias actually goes to zero faster than $M^{-2/d}$ (with a fixed k) when $d \geq 3$. This result is used to generalize Corollary 2.5 to the case $d = 4$. It is interesting, that most estimators based on locally constant approximations, for example (7) and the modification [2], have a bias of order $O(M^{-1/2})$ when $d = 4$, whereas our estimator is able to achieve $o(M^{-1/2})$. The following relatively weak condition is needed.

(A4) The distributions of the variables $(X_i)_{i=1}^\infty$ are absolutely continuous with respect to the Lebesgue measure with a common density p.

Theorem 2.7. *Suppose assumptions* (A1)–(A4) *hold with* $X_i \in [-1/2, 1/2]^d$ *for all* $i > 0$. *If* $\gamma > 1$ *and* $3 \leq d \leq 4$, *then*

$$\limsup_{M\to\infty} M^{2/d} |E[\hat{V}_{M,k}] - V_M| = 0$$

and

$$\limsup_{M\to\infty} ME\left[\left(\hat{V}_{M,k} - \frac{1}{M}\sum_{i=1}^{M} r_i^2\right)^2\right] \leq g(k,d)\sigma^4$$

for the function $g(k, d)$ *defined in Eq.* (12).

It is likely that under sufficient regularity conditions on the covariates, it would be possible to obtain the rate $O(M^{-1/2})$ even in dimension 5. For example, the conditions used in [25] would probably be sufficient. However, this type of restrictive conditions would be hard to verify in practice.

3. Some properties of nearest neighbors

3.1. How many points can share the same nearest neighbors?

In this section, our goal is to shortly investigate some properties of nearest neighbors needed for our theoretical analysis. We start by addressing the question, how many points can share the same k first nearest neighbors. The following upper bound gives a sufficient answer to this problem. In what follows, by $B(x, r)$ we denote the open ball with center x and radius r. A similar proof can be found in [20].

Theorem 3.1. *For any* $k > 0$ *and* $0 < j \leq M$ *the number of points in* $(X_i)_{i=1}^M$ *that have the point* X_j *among their k first nearest neighbors is almost surely bounded by* $kL(d)$.

Proof. Fix the point X_j and for vectors $\|e\| = 1$, define the cones

$$C(e) = \{x \in \Re^d : e^T(x - X_j) \leq \cos 30^\circ \|x - X_j\|\}.$$

Notice that for $z, y \in C(e)$, we have the geometrically intuitive bound

$$(z - X_j)^T(y - X_j) \geq \frac{1}{2}\|z - X_j\|\|y - X_j\|. \tag{13}$$

Let us now make the counterassumption that there exists $k + 1$ points $(X_{j_i})_{i=1}^{k+1} \subset C(e)$ that have X_j among their k nearest neighbors. Recalling Eq. (6), we may assume that $\|X_{j_{k+1}} - X_j\| > \|X_{j_k} - X_j\| > \cdots > \|X_{j_1} - X_j\|$. Then by inequality (13) we have for any $1 \leq i \leq k$,

$$\begin{aligned}\|X_{j_{k+1}} - X_{j_i}\|^2 &\leq \|X_{j_{k+1}} - X_j\|^2 + \|X_{j_i} - X_j\|^2 - \|X_{j_{k+1}} - X_j\|\|X_{j_i} - X_j\| \\ &< \|X_{j_{k+1}} - X_j\|^2\end{aligned}$$

the last inequality being strict. Thus we may conclude that X_j cannot be among the k nearest neighbors of the point $X_{j_{k+1}}$ leading to a contradiction.

As the final step, recall that we may cover the space $\Re^d$ with $L(d)$ cones of degree 30°. Then each point in the sample falls into one of these cones and we may conclude that X_j can be among the k nearest neighbors of at most $kL(d)$ points. □

As the bound involving $L(d)$ is essentially deterministic, it is expected to be conservative in practice. Thus it would be of interest to derive probabilistic bounds, for example under i.i.d. sampling, which might turn out to considerably tighter.

3.2. Bounds on the moments of nearest neighbor distances

In this section we examine the empirical moments

$$\delta_{M,k,\alpha} = \frac{1}{M}\sum_{i=1}^{M} \min\{1, d_{i,k,M}^{\alpha}\}. \tag{14}$$

In [23], a probabilistic technique is used to bound $\delta_{M,k,\alpha}$; however, here we derive a geometric, essentially deterministic bound. Similar, slightly weaker geometric results using a different technique can be found in [26]. Bounds on $\delta_{M,k,\alpha}$ are useful, because, as will be seen later, they are related to the rate of convergence of a class of nonparametric statistical estimators.

In what follows, by $B(x, r)$ we denote the open ball with center x and radius r.

Theorem 3.2. *Suppose that $X_i \in [-1/2, 1/2]^d$ almost surely for any i and let $0 < \alpha \leq d$. Then*

$$\frac{1}{M}\sum_{i=1}^{M} d_{i,k,M}^{\alpha} \leq 2^{\alpha} d^{\alpha/2} k^{\alpha/d} M^{-\alpha/d}. \tag{15}$$

Proof. Denote by S_d the volume of the unit ball. Notice that for $\alpha = d$, Eq. (15) can be written as an integral of a sum of indicator functions:

$$\frac{1}{M}\sum_{i=1}^{M} d_{i,k,M}^{d} \leq \frac{1}{M} 2^d S_d^{-1} \int_{B(0,\sqrt{d})} \sum_{i=1}^{M} I(x \in B(X_i, d_{i,k,M}/2))\mathrm{d}x. \tag{16}$$

The sum inside the integral could be bounded straightforwardly by Theorem 3.1. However, this would lead to unnecessarily bad constants as it can be shown that the sum is actually always at most k. To see this, choose any $x \in \Re^d$ and make the counterassumption that there exists $k+1$ points, denoted by $X_{i_1}, \ldots, X_{i_{k+1}}$ (the indices being distinct), such that $x \in B(X_{i_j}, d_{i_j,k,M}/2)$ for $j = 1, \ldots, k+1$. Let $(i_j, i_{j'})$ be the pair that maximizes the distance $\|X_{i_j} - X_{i_{j'}}\|$. Under these conditions the triangle inequality yields

$$\|X_{i_j} - X_{i_{j'}}\| < \frac{1}{2} d_{i_j,k,M} + \frac{1}{2} d_{i_{j'},k,M}.$$

On the other hand,

$$\begin{aligned}
\|X_{i_j} - X_{i_{j'}}\| &= \frac{1}{2}\|X_{i_j} - X_{i_{j'}}\| + \frac{1}{2}\|X_{i_j} - X_{i_{j'}}\| \\
&= \frac{1}{2} \max_{1 \leq j' \leq k+1} \|X_{i_j} - X_{i_{j'}}\| + \frac{1}{2} \max_{1 \leq j \leq k+1} \|X_{i_j} - X_{i_{j'}}\| \\
&\geq \frac{1}{2} d_{i_j,k,M} + \frac{1}{2} d_{i_{j'},k,M}
\end{aligned}$$

leading to a contradiction and the desired conclusion. Now we have

$$\frac{1}{M}\sum_{i=1}^{M} d_{i,k,M}^{d} \leq 2^d d^{d/2} k M^{-1}.$$

The case $\alpha < d$ follows straightforwardly from the case $\alpha = d$, because Jensen's inequality implies

$$\frac{1}{M}\sum_{i=1}^{M} d_{i,k,M}^{\alpha} \leq \left(\frac{1}{M}\sum_{i=1}^{M} d_{i,k,M}^{d}\right)^{\alpha/d}. \quad \square$$

The following theorem extends Theorem 3.2 to the case of unbounded covariates following the ideas introduced in [23].

Theorem 3.3. *Suppose assumption* (A2) *holds and fix $0 < \alpha < d - d^2/\beta_2$. Then there exists a constant c independent of M and k with*

$$E[\delta_{M,k,\alpha}] \leq c k^{\alpha/d} M^{-\alpha/d}.$$

Proof. The proof is based on the idea of dividing the space $\Re^d$ into bounded sets (for example hypercubes) and then examining the samples in each cube separately. For a vector of integers $a = (a_1, \ldots, a_d)$, define S_a as the cube $S_a = [a_1 - 1/2, a_1 + 1/2] \times \cdots \times [a_d - 1/2, a_d + 1/2]$. Set I_a as the random set of indices $\{0 < i \leq M : X_i \in S_a\}$ and denote by $|I_a|$ its cardinality. Theorem 3.2 yields the upper bound

$$M\delta_{M,k,\alpha} \leq \sum_a \sum_{i \in I_a} \min\{1, d_{i,k,M}^{\alpha}\} \leq c k^{\alpha/d} \sum_a I(\|I_a\| > 0)|I_a|^{1-\alpha/d}$$

for some constant c independent of k and M. By Chebyshev's inequality and assumption (A2) we may estimate for $a \neq 0$

$$\begin{aligned}
E[|I_a|] = \sum_{i=1}^{M} P(X_i \in S_a) &\leq \sum_{i=1}^{M} P(\|X_i\| \geq \|a\|_\infty/2) \\
&\leq 2^{\beta_2} \sum_{i=1}^{M} \frac{E[\|X_i\|^{\beta_2}]}{\|a\|_\infty^{\beta_2}} \leq \frac{2^{\beta_2} c_1 M}{\|a\|_\infty^{\beta_2}}.
\end{aligned} \tag{17}$$

Jensen's inequality implies that $E[|I_a|^{1-\alpha/d}] \leq E[|I_a|]^{1-\alpha/d}$, which together with inequality (17) yields the upper bound

$$ME[\delta_{M,k,\alpha}] \leq ck^{\alpha/d}M^{1-\alpha/d} + 2^{\beta_2-\beta_2\alpha/d}c_1^{1-\alpha/d}ck^{\alpha/d}M^{1-\alpha/d}\sum_{a\neq 0}\|a\|_\infty^{-\beta_2+\beta_2\alpha/d}.$$

Recalling that $\alpha < d - d^2/\beta_2$ we have

$$\sum_{a\neq 0}\|a\|_\infty^{-\beta_2+\beta_2\alpha/d} \leq \sum_{a\neq 0}\|a\|_\infty^{-d-\epsilon} \leq c_d$$

for some constant c_d depending only on the dimension d. Here $\epsilon > 0$ ensures that the sum is finite. □

Next we examine (14) when assumption (A1) holds without assuming the moment condition (A2). In this case it is possible to show convergence to zero, even though the speed of convergence may be arbitrarily slow.

Theorem 3.4. *Suppose assumption* (A1) *holds. Then for any* $\alpha > 0$ *and* $k > 0$,

$$E[\delta_{M,k,\alpha}] \to 0$$

as $M \to \infty$.

Proof. For any $\epsilon > 0$, we may choose $J > 0$ such that for $i > 0$, $P(X_i \in [-J, J]^d) > 1 - \epsilon$. Then we define $(\tilde{X}_i)_{i=1}^{M_J}$ as the sample consisting of those vectors in $(X_i)_{i=1}^M$ that fall in the hypercube $[-J, J]^d$. Correspondingly $\tilde{\delta}_{J,k,\alpha}$ is defined as the average α-moment of the distance to the kth nearest neighbor in this new sample. Then we may estimate

$$E[\delta_{M,k,\alpha}] \leq P(X_i \notin [-J, J]^d) + E[\tilde{\delta}_{J,k,\alpha}] \leq \epsilon + E[\tilde{\delta}_{J,k,\alpha}].$$

However, by Theorem 3.2, the latter term in the right hand side becomes arbitrarily small as $M \to \infty$ and thus the proof is complete. □

3.3. An asymptotic property of nearest neighbor distributions

Consider the point X_i and its nearest neighbor $X_{N[i,1]}$ under i.i.d. sampling. Then the behavior of the unit vector

$$\frac{X_i - X_{N[i,1]}}{\|X_i - X_{N[i,1]}\|} \tag{18}$$

is essentially governed by the behavior of the probability distribution in the neighborhood of X_i. However, given enough samples, it is reasonable to assume that the distribution is locally almost constant. This on the other hand implies that the vector (18) is asymptotically approximately uniformly distributed on the unit circle. Our goal here is to formalize this idea and prove a result, which is needed in the proof of Theorem 2.7. Moreover, it is probable that the rather deep uniformity property has potential applications in other fields of statistical estimation.

Lemma 3.5. *Suppose assumptions* (A1)–(A4) *hold and define*

$$\omega_{x_0}(r) = 1 - \int_{B(x_0,r)} p(x)\mathrm{d}x.$$

Then the distribution of the variables $X_{N[i,1]}, \ldots, X_{N[i,k]}$ *conditional on* X_i *is given by the density*

$$p(x_{i,1}, \ldots, x_{i,k}|X_i) = k!\binom{M-1}{k}\omega_{X_i}(\|X_i - x_{i,k}\|)^{M-k-1}\prod_{j=1}^k p(x_{i,j}) \tag{19}$$

defined for $\|x_{i,1} - X_i\| < \|x_{i,2} - X_i\| < \cdots < \|x_{i,k} - X_i\|$.

Proof. Let $(l_j)_{j=1}^k$ be a set of distinct indices between 1 and M excluding i. Then

$$P(N[i,1] = l_1, \ldots, N[i,k] = l_k|X_i, (X_{l_j})_{j=1}^k = (x_{i,j})_{j=1}^k) = \omega_{X_i}(\|X_i - x_{i,k}\|)^{M-k-1}. \tag{20}$$

Eq. (19) follows by multiplying (20) by $\prod_{j=1}^k p(x_{i,j})$ and taking the sum over the sets $(l_j)_{j=1}^k$. See also [21]. □

The following theorem is the main result in this section. The proof is rather long, but straightforward.

Theorem 3.6. *Suppose assumptions* (A1)–(A4) *hold with* $\gamma > 1$ *in* (A3) *and* $X_i \in [-1/2, 1/2]^d$ *for all* $i > 0$. *Then for fixed* $j_2 > j_1 > 0$, $d \geq 3$ *and any* $i > 0$,

$$\limsup_{M\to\infty} M^{2/d}|E[(m(X_{N[i,j_1]}) - m(X_i))(m(X_{N[i,j_2]}) - m(X_i))]| = 0.$$

Proof. Choose $L \geq 1$, $0 < \epsilon < 1$ and define the events

$$A_M = \{d_{i,j_2,M} \leq LM^{-1/d}\}$$
$$B_M = \left\{p(X_i) \leq L, \int_{B(X_i,LM^{-1/d})} |p(x) - p(X_i)|\mathrm{d}x \leq \epsilon(L+1)^{-2j_2-1}M^{-1}\right\}$$
$$C_M = A_M \cup B_M.$$

It is well known that almost all points in $\Re^d$ are Lebesgue points for p [27]. Thus

$$M \int_{B(X_i,LM^{-1/d})} |p(x) - p(X_i)|\mathrm{d}x \to 0 \tag{21}$$

almost surely as $M \to \infty$ for a fixed L. On the other hand, by Theorem 3.3 we may choose $L \geq 1$ such that $P(d_{i,j_2,M} > LM^{-1/d}) + P(p(X_i) > L) \leq \epsilon$. With such a choice of L, we have by Eq. (21) for the complement of C_M,

$$P(C_M^C) \leq 2\epsilon, \tag{22}$$

when M is large enough. Set $b_{i,j} = m(X_j) - m(X_i)$ and notice that by Lipschitz continuity

$$|b_{i,N[i,j_1]}b_{i,N[i,j_2]}| \leq \min\{4K_Y^2, c_2^2 d_{i,j_2,M}^2\}.$$

By Holder's inequality, Eq. (22), Theorem 3.2 and the previous remark, we can control the behavior at the complement of C_M by (I denotes the indicator function of an event)

$$\begin{aligned} E[|b_{i,N[i,j_1]}b_{i,N[i,j_2]}|I(C_M^C)] &\leq (c_2^2 + 4K_Y^2)E[\min\{1, d_{i,j_2,M}^2\}I(C_M^C)] \\ &\leq 4dj_2^{2/d}(c_2^2 + 4K_Y^2)\epsilon^{1-2/\alpha_2}M^{-2/d}, \end{aligned} \tag{23}$$

where $2 < \alpha_2 < d$. Next, using the gradient of m, define $\Delta(X_i, X_j) = \nabla m(X_i)(X_j - X_i)$ and notice that taking $c_2 \geq 1$, we may use assumption (A3) and the mean value theorem to estimate

$$I(C_M)|b_{i,N[i,j_1]}b_{i,N[i,j_2]} - \Delta(X_i, X_{N[i,j_1]})\Delta(X_i, X_{N[i,j_2]})| \leq 3c_2^2 L^4 M^{-2/d-(\gamma-[\gamma])/d}, \tag{24}$$

which goes to zero faster than $M^{-2/d}$. Thus on C_M, nonlinearities are negligible.

To proceed, define the set $\Xi_x \subset \Re^{d\times j_2}$ by

$$\Xi_x = \{0 < \|x_{i,1} - x\| < \|x_{i,2} - x\| \ldots < \|x_{i,j_2} - x\| \leq LM^{-1/d}\}.$$

The definition of B_M implies the inequality

$$I(B_M)\int_{B(X_i,LM^{-1/d})} p(x)\mathrm{d}x \leq (L^2+1)M^{-1}. \tag{25}$$

Next it is important to notice that the function $\Delta(x, y)$ integrates to zero with respect to the uniform measure on the unit sphere with center x (as a function of y). Using Lemma 3.5, inequality (25) and this remark we have

$$\begin{aligned} &|E[I(C_M)\Delta(X_i, X_{N[i,j_1]})\Delta(X_i, X_{N[i,j_2]})|X_i]| = |E[I(A_M)\Delta(X_i, X_{N[i,j_1]})\Delta(X_i, X_{N[i,j_2]})|X_i]I(B_M)| \\ &\leq j_2!\binom{M-1}{j_2}\left|\int_{\Xi_x}\Delta(X_i, x_{i,j_1})\Delta(X_i, x_{i,j_2})\omega_{X_i}(\|X_i - x_{i,j_2}\|)^{M-j_2-1}p(X_i)\prod_{1\leq j\leq j_2, j\neq j_1} p(x_{i,j})\mathrm{d}x_{i,1:j_2}\right| \\ &\quad + j_2!\binom{M-1}{j_2}c_2^2 I(B_M)L^2M^{-2/d}\left(\int_{B(X_i,LM^{-1/d})} p(x)\mathrm{d}x\right)^{j_2-1}\int_{B(X_i,LM^{-1/d})}|p(x) - p(X_i)|\mathrm{d}x \\ &\leq j_2!c_2^2\epsilon M^{-2/d}, \end{aligned} \tag{26}$$

where the inequality $\binom{M-1}{j_2} \leq M^{j_2}$ is used. Now the proof is finished by combining inequalities (23), (24) and (26) as ϵ can be chosen arbitrarily small. □

4. Conclusion

From the practical point of view, the product estimator is attractive because while the number of neighbors k introduces a degree of freedom, consistency holds even for fixed values. Moreover, the rate of convergence is invariant of the choice of k as long as it is kept fixed. These properties together with consistency for heteroscedastic noise make the method attractive compared to many other residual variance estimator, especially those based on the use of kernels.

It was shown that if k is allowed to increase slowly, the optimal asymptotic variance can be achieved. The fact that this property holds even in four dimensions indicates a faster rate of convergence than the worst-case bounds would indicate. A full analysis of the symmetry argument in Theorem 3.6 is a topic of future research.

Appendix

A.1. Useful lemmas

The rather complicated form of nearest neighbor graphs makes it more difficult to apply the strong law of large numbers as the classical theory does not apply as such. However, the following lemma solves this problem in a satisfying way. The proof is based on Theorem 3.1, which bounds the degree of interaction in the nearest neighbor graph. The notation $E[\cdot|X_1^M]$ means the conditional expectation with respect to the σ-algebra generated by the variables $(X_i)_{i=1}^M$.

Lemma A.1. *For a fixed $k > 0$, let $h(z_1, z_{1,1}, \ldots, z_{i,k})$ be a measurable function and define the random variables $h_i = h(Z_i, Z_{N[i,1]}, \ldots, Z_{N[i,k]})$. Assume that for all $i > 0$, $E[h_i|X_1^M] = 0$, $E[h_i^2|X_1^M] \leq U_1^2$ and $\frac{1}{M}\sum_{i=1}^M E[h_i^2|X_1^M]^{1/2} \leq U_2$ for some random variables U_1, U_2 measurable with respect to the σ-algebra generated by the variables $(X_i)_{i=1}^M$. Then we have*

$$E\left[\left(\frac{1}{M}\sum_{i=1}^M h_i\right)^2 \middle| X_1^M\right] \leq \frac{U_1U_2(k(k+1)L(d)+k+1)}{M}$$

almost surely.

Proof. Set $S(i) = \{i, N[i,1], \ldots, N[i,k]\}$ and define the sets

$$I(j) = \{0 < i \leq M : S(i) \cap S(j) \neq \emptyset\}.$$

Conditioned on the sample $(X_i)_{i=1}^M$, the variables h_i and h_j are independent whenever $i \notin I(j)$. Thus by Holder's inequality,

$$\begin{aligned} E\left[\sum_{i=1}^M h_ih_j \middle| X_1^M\right] &= \sum_{i\in I(j)} E[h_ih_j|X_1^M] \leq \sum_{i\in I(j)} E[h_i^2|X_1^M]^{1/2}E[h_j^2|X_1^M]^{1/2} \\ &\leq |I(j)|U_1E[h_j^2|X_1^M]^{1/2}. \end{aligned}$$

By Theorem 3.1, any point in the set $S(j)$ can be among the k first nearest neighbors for at most $kL(d)$ points. Thus we may conclude that $|I(j)| \leq k(k+1)L(d)+k+1$. Now the proof is finished by writing

$$\begin{aligned} E\left[\left(\frac{1}{M}\sum_{i=1}^M h_i\right)^2 \middle| X_1^M\right] &= \frac{1}{M^2}\sum_{i=1}^M\sum_{j=1}^M E[h_ih_j|X_1^M] \\ &\leq \frac{U_1U_2(k(k+1)L(d)+k+1)}{M}. \quad \square \end{aligned}$$

Next we show that $\hat{V}_{M,k}$ can be divided into a sum of three random variables, which have different convergence properties.

Lemma A.2. *The estimator $\hat{V}_{M,k}$ can be decomposed as $\hat{V}_{M,k} = S_1 + S_2 + S_3$ with*

$$E\left[\left(S_1 - \frac{1}{M}\sum_{i=1}^M r_i^2\right)^2\right] \leq (4k^{-3} + 2k^{-2} + 4L(d)k^{-2} + 2L(d)k^{-1})\sigma^4$$

and (with the notation $b_{i,j} = m(X_i) - m(X_j)$)

$$\begin{aligned} S_2 &= \frac{1}{k^2M}\sum_{i=1}^M\left(\sum_{j_1=1}^k b_{i,N[i,2j_1]}\right)\left(\sum_{j_2=1}^k r_i - r_{N[i,2j_2-1]}\right) + \frac{1}{k^2M}\sum_{i=1}^M\left(\sum_{j_1=1}^k r_i - r_{N[i,2j_1]}\right)\left(\sum_{j_2=1}^k b_{i,N[i,2j_2-1]}\right) \\ S_3 &= \frac{1}{k^2M}\sum_{j_1,j_2=1}^k\sum_{i=1}^M b_{i,N[i,2j_1]}b_{i,N[i,2j_2-1]}. \end{aligned}$$

Proof. We make the following choice of S_1:

$$S_1 = \frac{1}{k^2 M} \sum_{i=1}^{M} \left(\sum_{j_1=1}^{k} r_i - r_{N[i,2j_1]} \right) \left(\sum_{j_2=1}^{k} r_i - r_{N[i,2j_2-1]} \right).$$

The terms in the sum can be expanded as

$$\frac{1}{k^2} \left(\sum_{j_1=1}^{k} r_i - r_{N[i,2j_1]} \right) \left(\sum_{j_2=1}^{k} r_i - r_{N[i,2j_2-1]} \right) = \frac{1}{k^2} \sum_{j_1=1}^{k} \sum_{j_2=1}^{k} (r_i^2 - r_i r_{N[i,2j_1]} - r_i r_{N[i,2j_2-1]} + r_{N[i,2j_1]} r_{N[i,2j_2-1]}). \tag{27}$$

To analyze the sum (27), we define the sets

$$I_{i,j} = \{1 \le l \le M : \{i, j\} \subset \{l, N[l, 1], \ldots, N[l, 2k]\}\}$$

and the random variables $(a_{i,j})_{1 \le i < j \le M}$ by collecting the constants corresponding to each term $r_i r_j$ in the sum (27). Then S_1 can be written in a more simple way as

$$S_1 - \frac{1}{M} \sum_{i=1}^{M} r_i^2 = \frac{1}{k^2} \sum_{1 \le i < j \le M} a_{i,j} r_i r_j. \tag{28}$$

Using Theorem 3.1, the variable $|a_{i,j}|$ can be bounded by $|I_{i,j}| \le 2 + 2kL(d)$. In addition, notice that we must have

$$\sum_{1 \le i < j \le M} |a_{i,j}| \le (2k + k^2)M, \tag{29}$$

because the sum $S_1 - \frac{1}{M} \sum_{i=1}^{M} r_i^2$ has $(2k + k^2)M$ terms, when written in the form of Eq. (27). Conditioned on the variables $(X_i)_{i=1}^{M}$, the variables $a_{i,j}$ are constants and thus we have by conditional independence

$$E[a_{i,j} a_{i',j'} r_i r_j r_{i'} r_{j'} | X_1^M] = 0,$$

when $(i, j) \ne (i', j')$, and by Holder's inequality

$$\begin{aligned} E\left[\left(\frac{1}{k^2} \sum_{1 \le i < j \le M} a_{i,j} r_i r_j \right)^2 \middle| X_1^M \right] &= \frac{1}{k^4} \sum_{1 \le i < j \le M} a_{i,j}^2 E[r_i^2 r_j^2 | X_1^M] \\ &\le \frac{2\sigma^4 + 2\sigma^4 kL(d)}{k^4} \sum_{1 \le i < j \le M} |a_{i,j}| \\ &\le (4k^{-3} + 2k^{-2} + 4L(d)k^{-2} + 2L(d)k^{-1})\sigma^4. \quad \square \end{aligned}$$

A.2. *Proof of* Theorem 2.2

Recall that $|b_{i,j}| \le 2K_Y$, $E[(r_i - r_{N[i,k]})^2 | X_1^M] \le 2\sigma^2$ and

$$E\left[\left(\sum_{j_1=1}^{k} b_{i,N[i,2j_1]} \right) \left(\sum_{j_2=1}^{k} r_i - r_{N[i,2j_2-1]} \right) + \left(\sum_{j_1=1}^{k} r_i - r_{N[i,2j_1]} \right) \left(\sum_{j_2=1}^{k} b_{i,N[i,2j_2-1]} \right) \middle| X_1^M \right] = 0.$$

Using these observations together with Lemma A.1 implies that $E[S_2^2] \to 0$ as $M \to \infty$ in Lemma A.2. Thus we need to show that asymptotically $E[S_3^2] \to 0$, or equivalently, by the boundedness of m, $E[|S_3|] \to 0$. Let us first assume that m is continuous with a compact support and consequently uniformly continuous. Then for any $\epsilon > 0$, we may choose $\delta > 0$ such that $\|x - y\| < \delta$ implies $|m(x) - m(y)| < \epsilon$. Then for any $i > 0$,

$$E[|S_3|] \le 4K_Y^2 P(d_{i,2k,M} > \delta) + \epsilon^2. \tag{30}$$

By Theorem 3.4, $P(d_{i,2k,M} > \delta) \to 0$ as $M \to \infty$ and thus the claim is proven under the continuity assumption. The general case follows by a density argument as we may choose a continuous compactly supported $\tilde{m}$ such that for any $\epsilon >$ and all $i > 0$, $E[(m(X_i) - \tilde{m}(X_i))^2] < \epsilon$.

A.3. Proof of Theorem 2.4

Without losing generality, we assume $\gamma \leq 1$. Under the conditions of the theorem, $|b_{i,N[i,k]}| \leq 2K_Y$ and

$$|b_{i,N[i,k]}| \leq (2K_Y + c_2)\min\{1, d_{i,k,M}^{\gamma}\},$$

which relates the proof to the bounds derived in Section 3.2. The term S_2 can be bounded using Lemma A.1, Theorem 3.3 and the triangle inequality by

$$\begin{aligned} E[S_2^2] &\leq 16(2K_Y + c_2)(4k^2L(d) + 2kL(d) + 2k + 1)K_Y\sigma^2\delta_{M,2k,\gamma}M^{-1} \\ &\leq c_4 M^{-1-\tilde{\gamma}/d}, \end{aligned}$$

for some constant c_4. In a corresponding way,

$$|S_3| \leq (2K_Y + c_2)^2\delta_{M,2k,\alpha} \leq c_5 M^{-\alpha/d}.$$

When the covariates are bounded and $2\gamma \leq d$, we may approximate straightforwardly $|b_{i,N[i,k]}| \leq c_2 d_{i,k,M}^{\gamma}$ and use Theorem 3.2 to estimate

$$|S_3| \leq 2^{2\gamma+2\gamma/d} d^{\gamma} k^{2\gamma/d} c_2^2 M^{-2\gamma/d}$$

with a similar inference for S_2.

A.4. Proof of Theorem 2.7

The first claim is a consequence of Lemma A.2 and Theorem 3.6 because

$$E[S_1] = E[S_2] = 0.$$

To prove the second claim, we need to prove that the variance of $S_3 - E[S_3]$ goes to zero faster than M. This will be done using the well-known Efron–Stein inequality. Let us assume for simplicity that $k = 1$. For the indicator functions $I(d_{i,2,M} > \epsilon)$, we have by Theorem 3.3

$$\frac{1}{M}\sum_{i=1}^{M} |b_{i,N[i,1]}b_{i,N[i,2]}| I(d_{i,2,M} > \epsilon) \leq 4K_Y \frac{1}{M}\sum_{i=1}^{M} I(d_{i,2,M} > \epsilon) = O(M^{-1}).$$

Set

$$\tilde{S}_3^{(1)} = \frac{1}{M}\sum_{i=1}^{M} b_{i,N[i,1]}b_{i,N[i,2]} I(d_{i,2,M} \leq \epsilon)$$

and notice that by the triangle inequality

$$\begin{aligned} E[(S_3 - E[S_3])^2]^{1/2} &\leq E[(\tilde{S}_3^{(1)} - E[\tilde{S}_3^{(1)}])^2]^{1/2} + 2E[(S_3 - \tilde{S}_3^{(1)})^2]^{1/2} \\ &= E[(\tilde{S}_3^{(1)} - E[\tilde{S}_3^{(1)}])^2]^{1/2} + o(M^{-1/2}). \end{aligned}$$

By choosing ϵ small enough, we may assume that each term in the sum $\tilde{S}_3$ is smaller than $\delta > 0$. Next define the variable $\tilde{S}_3^{(2)}$ by replacing X_1 by a similarly distributed independent copy X_1'. In this definition, the variables $(X_i)_{i=2}^M$ are kept intact; thus $\tilde{S}_3^{(2)}$ is similarly distributed as $\tilde{S}_3$.

As the perturbation can affect only those points, which have X_1 (correspondingly X_1') among their two nearest neighbors in the original sample or in the modified sample $\{X_1'\} \cup \{X_i\}_{i=2}^M$, we have by Theorem 3.1

$$(\tilde{S}_3 - \tilde{S}_3^{(2)})^2 \leq C\delta^2 M^{-2},$$

for a constant C depending only on the dimensionality d. Then the well-known Efron–Stein inequality [28] implies that

$$\mathrm{Var}[\tilde{S}_S] \leq C\delta^2 M^{-1}.$$

As δ can be chosen arbitrarily small, the proof is completed.

References

[1] C.J. Stone, Optimal rates of convergence for nonparametric estimators, Annals of Statistics 8 (6) (1980) 1348–1360.
[2] L. Devroye, L. Györfi, D. Schäfer, The estimation problem of minimum mean squared error, Statistics and Decisions 21 (1) (2003) 15–28.
[3] T. Gasser, L. Stroka, C. Jennen-Steinmetz, Residual variance and residual pattern in nonlinear regression, Biometrika 73 (3) (1986) 625–633.
[4] P. Hall, J. Kay, D. Titterington, Asymptotically optimal difference-based estimation of variance in nonparametric regression, Biometrika 77 (3) (1990) 521–528.
[5] U. Müller, A. Schick, W. Wefelmeyer, Estimating the error variance in nonparametric regression by a covariate-matched U-statistic, Statistics 37 (3) (2003) 179–188.
[6] T. Tong, Y. Wang, Estimating residual variance in nonparametric regression using least squares, Biometrika 92 (4) (2005) 821–830.
[7] P. Hall, J. Marron, On variance estimation in nonparametric regression, Biometrika 77 (2) (1990) 415–419.
[8] V. Spokoiny, Variance estimation for high-dimensional regression models, Journal of Multivariate Analysis 82 (1) (2002) 111–133.
[9] J. Fan, Q. Yao, Efficient estimation of conditional variance functions in stochastic regression, Biometrika 85 (3) (1998) 645–660.
[10] D. Ruppert, M.P. Wand, U. Holst, O. Hössjer, Local polynomial variance-function estimation, Technometrics 39 (3) (1997) 262–273.
[11] H.G. Müller, U. Stadtmüller, Estimation of heteroscedasticity in regression analysis, Annals of Statistics 15 (2) (1987) 610–625.
[12] H.G. Müller, U. Stadtmüller, On variance function estimation with quadratic forms, Journal of Statistical Planning and Inference 35 (1) (1993) 126–136.
[13] A. Munk, N. Bissantz, T. Wagner, G. Freitag, On difference based variance estimation in nonparametric regression when the covariate is high dimensional, Journal of the Royal Statistical Society: Series B 67 (1) (2005) 19–41.
[14] L.D. Brown, M. Levine, Variance estimation in nonparametric regression via the difference sequence method, Annals of Statistics 35 (5) (2007) 2219–2232.
[15] T. Cai, M. Levine, L. Wang, Variance function estimation in multivariate nonparametric regression, Journal of Multivariate Analysis 100 (1) (2009) 126–136.
[16] Q. Yao, H. Tong, Nonparametric estimation of ratios of noise to signal in stochastic regression, Statistica Sinica 10 (3) (2000) 751–770.
[17] D. Evans, Estimating the variance of multiplicative noise, in: 18th International Conference on Noise and Fluctuations, ICNF, in: AIP Conference Proceedings, vol. 780, 2005, pp. 99–102.
[18] D. Evans, A.J. Jones, Non-parametric estimation of residual moments and covariance, Proceedings of the Royal Society A 464 (2099) (2008) 2831–2846.
[19] L. Devroye, L. Györfi, A. Krzy zak, G. Lugosi, On the strong universal consistency of nearest neighbor regression function estimates, Annals of Statistics 22 (3) (1994) 1371–1385.
[20] M. Penrose, Random Geometric Graphs, in: Oxford Studies in Probability, no. 5, Oxford University Press, 2003.
[21] R.R. Snapp, S.S. Venkatesh, Asymptotic expansions of the k nearest neighbor risk, Annals of Statistics 26 (3) (1998) 850–878.
[22] E. Liitiäinen, A. Lendasse, F. Corona, On non-parametric residual variance estimation, Neural Processing Letters 28 (3) (2008) 155–167.
[23] M. Kohler, A. Krzyzak, H. Walk, Rates of convergence for partitioning and nearest neighbor regression estimates with unbounded data, Journal of Multivariate Analysis 97 (2) (2006) 311–323.
[24] A. Antos, L. Devroye, L. Györfi, Lower bounds for Bayes error estimates, IEEE Transactions on Patterns Analysis and Machine Intelligence 21 (7) (1999) 643–645.
[25] D. Evans, A.J. Jones, W.M. Schmidt, A proof of the gamma test, Proceedings of the Royal Society A 458 (2027) (2002) 2759–2799.
[26] S.R. Kulkarni, S.E. Posner, Rates of convergence of nearest neighbor estimation under arbitrary sampling, IEEE Transactions on Information Theory 41 (4) (1995) 1028–1039.
[27] W. Rudin, Real and Complex Analysis, in: Higher Mathematics Series, McGraw-Hill Science, 1986.
[28] J.M. Steele, An Efron–Stein inequality for nonsymmetric statistics, Annals of Statistics 14 (2) (1986) 753–758.