Neural Process Lett (2008) 28:155–167
DOI 10.1007/s11063-008-9087-8

On Nonparametric Residual Variance Estimation

Elia Liitiäinen · Francesco Corona · Amaury Lendasse

Published online: 26 October 2008

Abstract In this paper, the problem of residual variance estimation is examined. The problem is analyzed in a general setting which covers non-additive heteroscedastic noise under non-iid sampling. To address the estimation problem, we suggest a method based on nearest neighbor graphs and we discuss its convergence properties under the assumption of a Hölder continuous regression function. The universality of the estimator makes it an ideal tool in problems with only little prior knowledge available.

Keywords Residual variance estimation · Noise variance · Nearest neighbor · Nonparametric

1 Introduction

The problem of residual variance estimation consists of estimating the minimum mean squared generalization error obtainable by a nonlinear model [3,12]. The residual variance is a natural measure of relevance in the context of data-derived modeling and can be profitably exploited in common tasks like input and model structure selection for neural networks as shown in [8]. In many cases, it offers a viable alternative to information theoretic measures of dependency.

The problem originates from statistics, where it is often called noise variance estimation [21,4,6,14]. Many studies on the topic analyze the additive and homoscedastic noise case under the independent identically distributed (iid) sampling assumption, or fixed (equispaced

E. Liitiäinen (✉) · F. Corona · A. Lendasse
Department of Computer Science and Engineering, Helsinki University of Technology,
P.O. Box 5400, 2015 Espoo, Finland
e-mail: elia.liitiainen@hut.fi

F. Corona
e-mail: fcorona@james.hut.fi

A. Lendasse
e-mail: lendasse@hut.fi

or univariate) design setting. For example, statistically efficient difference-based methods are derived in [17,7] and kernel-based estimators in [6]. From the general data analysis point of view, however, the multivariate random covariates setting is more interesting and challenging; in fact, while kernel-based estimators have straightforward extensions [7,14], difference-based methods do not have clear counterparts in the random multivariate design case. One possible generalization is discussed, for instance, in [21], where local linear regression is used to derive an estimator that is unbiased for linear problems.

Despite the usefulness of noise variance estimators in supervised learning and relevance estimation, there has been less research on the topic in machine learning. Previous work using near neighbor statistics includes [4,12], whereas the differogram is used in [15]. Applications of these estimators include model selection for support vector machines and multilayer neural networks [8,15,11] and input selection [8]. However, again the forementioned estimators are analyzed assuming noise with constant variance, which is a strong assumption taking into account that small amount of prior knowledge is often available.

Thus, an important step is to examine the case of heteroscedastic noise, which means that the noise variance is a function of the covariates. The one-dimensional case has been examined, for example, in [1], but much less effort has been devoted to examine the multivariate case. Methods based on the use of local linear regression have been developed and analyzed in [22,19,9]; however, all of these methods contain free parameters, which are not always easy to estimate.

Instead of estimating the whole variance function, we concentrate on estimating its expectation over the sample space in a general non-iid setting. This alternative approach is thoroughly investigated in [3], where a modified nearest neighbor graph combined with a locally constant estimator is used to generalize the nonparametric first nearest neighbor noise variance estimator to the heteroscedastic noise case. One interesting fact is that the convergence properties of the estimator are independent of the smoothness of the variance function, which allows general convergence properties while, at the same time, not assuming additive noise.

Stemming from such a recognition, we suggest an alternative estimator also based on the use of nearest neighbor graphs but characterized by a slight but important modification that allows a simpler formulation but similar convergence properties. The method is fully nonparametric with no free parameters. The convergence is proven using a similar technique as that developed in [12] which leads to different proofs than those in [3]. We also conjecture that in practice the speed of convergence is expected to be faster than our theoretical worst-case bound.

In the first two sections of this paper, the problem of residual variance estimation is formalized (Sect. 2), the concept of nearest neighbors is briefly overviewed and a theoretical upper bound is derived (Sect. 3). In Sects. 4 and 5, we introduce the nonparametric residual variance estimator and a convergence result is proven. In order to support the presentation and demonstrate the properties of the estimator in practice, the results on numerical experiments are illustrated in Sect. 6.

2 Residual Variance Estimation

By residual variance estimation, we mean estimating the lowest possible mean squared generalization error in a given regression problem based on given data. Our approach is mainly intended for data-derived modeling using stationary models and is a generalization of the formulation discussed in [4].

Before stating the general form of the problem, we provide some general definitions that are needed in the subsequent treatment.

Our starting point is standard: let us assume that $(\Omega, \mathcal{F}, \mathcal{P})$ is a probability space with the σ-algebra $\mathcal{F}$ of events and the probability measure $\mathcal{P}$. The random vectors $(Z_i)_{i=1}^{\infty} = (X_i, Y_i)_{i=1}^{\infty}$ are independently distributed taking values in the product space $X \times \mathbb{R}$, where (X, ρ) is a metric space with the distance ρ. Notice that our model covers also the case of a fixed design, where the variables $(X_i)_{i=1}^{\infty}$ are chosen in some deterministic fashion.

The joint distribution is given by the joint density $p_i(x, y)$ (w.r.t. a dominating measure λ) from which we get the marginal density $p_i(x)$ be integrating out y. The scalar variables (Y_i) model the output of a system, whereas (X_i) describe the input; in practice, only a finite sample $(X_i, Y_i)_{i=1}^{M}$ is available and the number of samples M is the critical quantity when performing any statistical inference. Justified by the fact that, in practice, most random variables are bounded, we assume that $\sup_{x,y \in X} \rho(x, y) \leq 1$.

2.1 Problem Statement

Without assuming an additive noise model and independent identically distributed input, we state the problem of residual variance estimation in the general case of independent observations from the point of view of regression. In regression, the goal is to build a model between the variables (X_i) and (Y_i) given a finite sample $(X_i, Y_i)_{i=1}^{M}$; this can be done in diverse ways including linear models and neural networks.

The model is chosen by minimizing a cost function, typically, the mean squared error (MSE) between the model and the outputs. In this case, the problem reduces at estimating the function $g : X \to \mathbb{R}$ that minimizes the expectation

$$L_M(g) = \frac{1}{M} \sum_{i=1}^{M} E[(Y_i - g(X_i))^2], \tag{1}$$

even though, in practice, the expectations usually have to be estimated by averaging over the samples available and it is necessary to restrict the complexity of the function g to avoid overfitting. In statistics the assumption of iid random variables is common; in that case the simplification $L_M(g) = E[(Y_1 - g(X_1))^2]$ is possible.

The estimation of the residual variance is the inverse of this problem: the goal is to find the minimum value that the cost L_M can achieve on the set of bounded measurable functions. Denoting the set of bounded and measurable functions on X by $B(X)$, formally, the problem consists of computing

$$V_M = \inf_{g \in B(X)} L_M(g). \tag{2}$$

The value V_M is the variance of the residual and it describes the magnitude of the part of the output that remains unexplained with the theoretically optimal model. From the data-derived modelling point of view, the quantity V_M is the best possible generalization error one can achieve using a learning machine.

The following theorem characterizes the theoretically optimal solution of the regression problem.

Theorem 1 *The function that minimizes the cost* (1) *is given by*

$$m(x) = \sum_{i=1}^{M} \frac{p_i(x) E[Y_i | X_i = x]}{\sum_{i=1}^{M} p_i(x)}. \tag{3}$$

If the stationarity condition $E[Y_i|X_i = x] = E[Y_j|X_j = x]$ *holds for all* $i, j > 0$*, then* $m(x) = E[Y_i|X_i = x]$ *for any* $i > 0$.

Proof Define the density function $q(x, y) = M^{-1}\sum_{i=1}^{M} p_i(x, y)$ and assume that the random variable $(\tilde{X}, \tilde{Y})$ is distributed according to q. Then, it can be seen that $L_M(g) = E[(\tilde{Y} - g(\tilde{X}))^2]$, which implies that the optimal function m is given by $m(x) = E[\tilde{Y}|\tilde{X} = x]$ as it is a well-known fact that the conditional expectation gives the optimal function in the sense of L^2-norm [20]. Hence, starting from the definition of abstract conditional expectations [20], it is possible to show that g is of the form defined in (3). □

3 Nearest Neighbors

The concept of nearest neighbors (see for example [4]) has found its applications in various fields including non-parametric regression and classification. Our goal is to use nearest neighbors based estimators to approximatively solve the problem of residual variance estimation presented above.

3.1 Basic Definition

The definition of the nearest neighbor is based on the use of a proximity measure to determine similarity between points.

The nearest neighbor of a point X_i is

$$N[i, 1] = \text{argmin}_{1\le j\le M, j\ne i}\,\rho(X_i, X_j). \tag{4}$$

Possible ties are solved by taking the minimal index. The k-th nearest neighbor is defined recursively as

$$N[i, k] = \text{argmin}_{1\le j\le M, j\ne i, N[i,1],\ldots,N[i,k-1]}\,\rho(X_i, X_j), \tag{5}$$

that is, the closest point after removal of the preceeding neighbors. The corresponding distances are defined as $d_{i,k,M} = \rho(X_i, X_{N[i,k]})$.

We also set

$$\delta_{M,k,\alpha} = \frac{1}{M}\sum_{i=1}^{M} d_{i,k,M}^{\alpha} \tag{6}$$

which is the empirical α-moment for the distances to the k-th nearest neighbor.

3.2 Moment Bound Under a Dimensionality Constraint

In [12], it is shown that under the assumption $X = \mathbb{R}^n$, the quantity $M^{\alpha/n}\delta_{M,\alpha,k}$ is bounded by a universal constant for $\alpha \le n$. This result is very useful, as it holds for all points sets and is, thus, of deterministic nature. In this section, we show that a corresponding result holds in a more general context.

To proceed, a constraint on the dimensionality of the metric space (X, ρ) is required. There exists many possible definitions for the dimensionality of a fractal or metric space including the Hausdorff dimension, capacity dimension and correlation dimension [5]. In this work, we will instead use the concept of packing numbers [10], which is related to the study of nearest neighbors as it is able to give an upper bound for the empirical moments.

It is worthwhile noticing that it remains largely an open question how other definitions of dimension are related to the average distance of the nearest neighbors; even though, it seems that, for example, the Hausdorff dimension is too weak a concept to provide geometric upper bounds.

A set $\mathcal{A} \subset X$ is an ϵ-packing, if for every distinct points $x, y \in \mathcal{A}$, $\rho(x, y) > \epsilon$. For $\epsilon > 0$, we define the packing numbers as the cardinality of the maximal packing, that is:

$$N_{packing}(\epsilon) = \sup_{\mathcal{A} \text{ is an } \epsilon\text{-packing}} |\mathcal{A}|. \tag{7}$$

Note that if X is a bounded subset of $\mathbb{R}^n$ then $N_{packing}(\epsilon) \leq C_n \epsilon^{-n}$ for $0 < \epsilon < 1$ and some constant C_n depending only on n and X. However, the situation where X is a low dimensional manifold in a high dimensional space is also common, in such a case the dimensionality of the metric space X is smaller than that of the space in which it is embedded. In fact, if X is a bounded manifold in the space $\Re^{n'}$ locally parametrizable by n coordinates with $n < n'$, then under some moderate regularity conditions on X and ρ, we have $N_{packing}(\epsilon) \leq C_{n'} \epsilon^{-n}$ for $0 < \epsilon < 1$ and some constant $C_{n'}$.

The next theorem is slightly weaker than the one in [12] in terms of an additional logaritmic factor due to the weaker assumption made about the space X. The bound will be used in the analysis of the bias of our residual variance estimator. See [10] for corresponding results in the context of classification. Being based on a geometric argument, the bound is independent of the underlying distribution.

Theorem 2 *Assume that for some constants $C_n, n > 0$, $N_{packing}(\epsilon) \leq C_n \epsilon^{-n}$ when $0 < \epsilon < 1$.*

Then, for $0 < \alpha < n$ and $M \geq kC_n$,

$$\delta_{M,k,\alpha} \leq \frac{n}{n - \alpha} k^{\alpha/n} C_n^{\alpha/n} M^{-\alpha/n} - \frac{\alpha k C_n M^{-1}}{n - \alpha}. \tag{8}$$

For $\alpha = n$, we have the bound

$$\delta_{M,k,n} \leq k C_n M^{-1} \left(1 + \log\left(\frac{M}{kC_n}\right)\right). \tag{9}$$

Proof Choose arbitrarily $t > 0$ and $0 < \alpha \leq n$ and define the set of indices

$$I_t = \{i : d_{i,k,M} > t\}. \tag{10}$$

Choose $i_1 \in I_t$ and define the set $I_{t,1} = I_t \setminus \{N[i_1, 1], \ldots, N[i_1, k-1]\}$. Then pick up $i_2 \neq i_1$ ($i_2 \in I_{t,1}$) and set $I_{t,2} = \{i_1\} \cup I_{t,1} \setminus \{N[i_2, 1], \ldots, N[i_2, k-1]\}$. Correspondingly,

$$I_{t,3} = \{i_1, i_2\} \cup I_{t,2} \setminus \{N[i_3, 1], \ldots, N[i_3, k-1]\} \tag{11}$$

with $i_3 \neq i_1, i_2$. By repeating the forementioned procedure as long as possible, we construct the sets $\{I_{t,j}\}_{j=1}^{L}$ for some $L \geq |I_t|/k$. Notice that by construction each index in the sequence $(i_j)_{j=1}^{L}$ is in $I_{t,L}$. Thus, in each iteration, a point is chosen from the active set and its nearest neighbors are removed up to the index $k-1$ (excluding the previously chosen points). Then, this chosen point is added to the set $\{i_j\}$.

Choose now $i, j \in I_{t,L}$ with $i \neq j$ and notice that from the properties of $I_{t,L}$ it follows that $\rho(X_i, X_j) \geq t$ and $|I_{t,L}| \leq N_{packing}(t)$, consequently. On the other hand, $I_{t,L}$ contains by construction exactly L points which implies that the cardinality $|I_t|$, is bounded by

$$|I_t| \leq kL \leq kN_{packing}(t) \leq kC_n t^{-n}. \tag{12}$$

Under the assumption that $M \geq kC_n$, we have (see [18], Theorem 8.16):

$$\begin{aligned}\delta_{M,k,\alpha} &= \int_0^1 \alpha t^{\alpha-1} M^{-1} |I_t| dt \leq \int_0^1 \alpha \min\left(kC_n t^{-n} M^{-1}, 1\right) t^{\alpha-1} dt \\ &= k^{\alpha/n} C_n^{\alpha/n} M^{-\alpha/n} + \int_{M^{-1/n} k^{1/n} C_n^{1/n}}^1 \alpha k C_n M^{-1} t^{\alpha-1-n} dt \\ &= \frac{n}{n-\alpha} k^{\alpha/n} C_n^{\alpha/n} M^{-\alpha/n} + \frac{\alpha k C_n M^{-1}}{\alpha - n}. \qquad (13)\end{aligned}$$

In the case that $\alpha = n$ and $M \geq kC_n$, we have:

$$\begin{aligned}\delta_{M,k,n} &\leq kC_n M^{-1} + \int_{M^{-1/n} k^{1/n} C_n^{1/n}}^1 n k C_n M^{-1} t^{-1} dt \\ &= kC_n M^{-1} \left(1 + \log\left(\frac{M}{kC_n}\right)\right). \qquad (14)\end{aligned}$$

□

4 Nonparametric Residual Variance Estimation

The concept of local continuity can be exploited to derive a nonparametric nearest neighbor estimator of residual variance.

Denoting by V_M the minimum of the cost in (3), a reasonable nonparametric estimator would be [3]:

$$V_M \approx \frac{1}{2M} \sum_{i=1}^{M} (Y_{N[i,1]} - Y_i)^2. \qquad (15)$$

Analysis about these methods can be found for example in [12,4], where it has been shown that the estimator has good properties under some stationarity conditions. Based on the simple and intuitive formulation of the estimator, one would expect the method to have good convergence properties in most situations. However, the next example shows that the estimator (15) is not necessarily consistent in the heteroscedastic noise case and thus it is not satisfying from the theoretical point of view.

Example 1 Let us consider that the set of univariate inputs $(X_i)_{i=1}^M$ consists of two distinct parts (containing M_1 and $2M_1$ variables) denoted by $(X_i^1)_{i=1}^{M_1}$ and $(X_i^2)_{i=1}^{2M_1}$ respectively. Furthermore, to construct our counterexample, we set $X_i^1 = \frac{i}{M_1}$, $X_{2i}^2 = X_i^1 - \frac{1}{4M_1}$ and $X_{2i-1}^2 = X_i^1 + \frac{1}{4M_1}$. The outputs 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. In this case, the expectation value of the approximation in (15) is $E\left[\frac{1}{2M}\sum_{i=1}^M (Y_{N[i,1]} - Y_i)^2\right] = \frac{1}{2}$. However, the right answer in this case is $1/3$ and, thus, it is clear that the method is not consistent in this example.

The above problem was also noticed in [3], where a solution based on modified nearest neighbor graphs was proposed. Our proposal to avoid the problem is to modify (15) to get

$$\hat{V}_M = \frac{1}{M} \sum_{i=1}^{M} (Y_i - Y_{N[i,1]})(Y_i - Y_{N[i,2]}), \qquad (16)$$

which, despite the non-intuitive formulation is shown to have much better properties than the original estimator.

In the rest, we show that the novel estimator converges regardless of the smoothness of the conditional variance function (and thus is able to solve the counterexample above).

5 Properties of the Estimator

In this section, we analyze the theoretical properties of the proposed estimator. We show that the estimator is asymptotically consistent in a general statistical setting. The analysis is done assuming that the conditions of Theorem 2 hold and that the absolute values of the outputs Y_i are bounded by some constant (this assumption is not the weakest possible, but simplifies the analysis). It is also necessary to require some smoothness of the function m in (3).

We have the following theorem on the rate of convergence of the estimator. The main point in the proof is the fact that no smoothness assumptions are needed on the conditional variance functions $E[(Y_i - m(X_i))^2 | X_i = x]$. Moreover, the distributions of the covariates $(X_i)_{i=1}^M$ do not affect the rate of convergence.

Theorem 3 *Assume that the continuity condition*

$$|m(x) - m(y)| \leq C_m \rho(x, y)^\gamma \tag{17}$$

holds for some constants $C_m > 0$, $0 < \gamma \leq 1$ *and* $m(x) = E[Y_1 | X_1 = x] = E[Y_i | X_i = x]$ *for all* $i > 0$ *and* $x, y \in X$.

Then, the bias of the estimator given in (16) *is bounded by*

$$|E[\hat{V}_M] - V_M| \leq C_m^2 E[\delta_{M,2,2\gamma}]. \tag{18}$$

Proof The proof is based on conditionalization with respect to the sample $(X_i)_{i=1}^M$ (denoted by $E[\cdot | X_1^M]$). The treatment relies on the basic properties of abstract conditional expectations, see for example [20]. We make the definitions

$$b_{i,j} = m(X_i) - m(X_j) \tag{19}$$

$$r_i = Y_i - m(X_i). \tag{20}$$

Then, we write

$$\begin{aligned} &E[(Y_i - Y_{N[i,1]})(Y_i - Y_{N[i,2]})] \\ &\quad = E[(b_{i,N[i,1]} + r_i - r_{N[i,1]})(b_{i,N[i,2]} + r_i - r_{N[i,2]})] \\ &\quad = E[b_{i,N[i,1]} b_{i,N[i,2]}] + E[b_{i,N[i,1]}(r_i - r_{N[i,2]})] + E[b_{i,N[i,2]}(r_i - r_{N[i,1]})] \\ &\qquad + E[(r_i - r_{N[i,1]})(r_i - r_{N[i,2]})]. \end{aligned} \tag{21}$$

Now using the fact that $E[r_i|X_1^M]=0$

$$E[r_{N[i,k]}|X_1^M]=\sum_{j=1}^M E[r_{N[i,k]}|X_1^M]I(N[i,k]=j)$$
$$=\sum_{j=1}^M E[r_j I(N[i,k]=j)|X_1^M]$$
$$=\sum_{j=1}^M E[r_j|X_1^M]I(N[i,k]=j)=0, \tag{22}$$

and observing that $b_{i,N[i,k]}$ is a function of the variables $(X_i)_{i=1}^M$, we have using the properties of conditional expectations

$$E[b_{i,N[i,1]}(r_i-r_{N[i,2]})]+E[b_{i,N[i,2]}(r_i-r_{N[i,1]})]$$
$$=E[b_{i,N[i,1]}E[r_i-r_{N[i,2]}|X_1^M]]+E[b_{i,N[i,2]}E[r_i-r_{N[i,1]}|X_1^M]]$$
$$=0 \tag{23}$$

and $|b_{i,N[i,1]}b_{i,N[i,2]}|\le C_m^2 d_{i,2,M}^{2\gamma}$.

Next by the independence of the samples,

$$E[r_i r_{N[i,1]}|X_1^M]=E[r_i r_{N[i,2]}|X_1^M]=E[r_{N[i,2]}r_{N[i,1]}|X_1^M]=0. \tag{24}$$

This follows from the properties of conditional expectations:

$$E[r_{N[i,2]}r_{N[i,1]}|X_1^M]=\sum_{j=1}^M\sum_{l=1}^M E[r_l r_j|X_1^M]I(N[i,2]=j)I(N[i,1]=l)$$
$$=\sum_{j=1}^M\sum_{l=1}^M E[Y_l Y_j-m(X_l)m(X_j)|X_1^M]I(N[i,2]=j)I(N[i,1]=l)$$
$$=0 \tag{25}$$

and for this reason $E[(r_i-r_{N[i,1]})(r_i-r_{N[i,2]})]=E[r_i^2]$ leading to the conclusion

$$|E[\hat{V}_M]-V_M|\le M^{-1}C_m^2 E\left[\sum_{i=1}^M d_{i,2,M}^{2\gamma}\right]. \tag{26}$$

□

Corollary 1 *Under the assumptions of Theorems 2 and 3 with $\gamma=1$, we have for $n\le 2$,*

$$|E[\hat{V}_M]-V_M|\le 2C_n C_m^2 M^{-1}\left(1+\log(\frac{M}{2C_n})\right) \tag{27}$$

and, for $n>2$,

$$|E[\hat{V}_M]-V_M|\le \frac{2^{2/n}n}{n-2}C_n^{2/n}C_m^2 M^{-2/n}+\frac{4C_n C_m^2 M^{-1}}{2-n}. \tag{28}$$

Proof The corollary follows from Theorem 2 by noticing that $E[d_{i,k,M}^2]\le E[d_{i,k,M}^n]$ for $n\le 2$. □

Example 2 To demonstrate Corollary 1, let us choose $X = \left[0, 1/\sqrt{2}\right]^2 \subset \Re^2$ with ρ the Euclidean metric and

$$m(x) = \omega^T x \tag{29}$$

for some vector $\omega \in \Re^2$. Then we may choose $n = 2$, $C_m = \|\omega\|$ (the Euclidean norm of ω), $\gamma = 1$ and $C_n = 1$. With these choices, inequality (27) gives

$$|E[\hat{V}_M] - V_M| \leq 2\|\omega\|^2 M^{-1} \left(1 + \log(\frac{M}{2})\right). \tag{30}$$

Based on Corollary 1, it can be concluded that fast convergence is expected when $n \leq 2$, whereas for a higher dimension the rate of convergence decreases. Example 2 demonstrates the application of Corollary 1 when m is linear. Notice that for a nonlinear m, C_m can be chosen as the upper bound for the norm of the gradient of m in case it exists.

If the stationarity condition on the conditional expectations $E[Y_i|X_i]$ can be removed is an interesting question for future research. We would like to note that the weaknesses and strongpoints of the method are the same as for many other nonparametric regression methods including the Nadaraya-Watson and k-NN estimators. The simplicity of the method makes it a good choice in low dimensional problems, even though more sophisticated method obtain better rates of convergence (see for example [21]).

In this section, the bias of the method was examined. Another important question is the variance. However, the variance of nearest neighbor based estimators is relatively well understood [2,4]. We state the following theorem and give a short proof.

Theorem 4 *Assuming that the variables $(X_i, Y_i)_{i=1}^M$ possess a density with respect to the Lebesgue measure on $\Re^n \times \Re$, we have*

$$\sup_{M>0} ME[(\hat{V}_M - E[\hat{V}_M])^2] < \infty. \tag{31}$$

Proof (sketch) Choose $l > 0$ and define a new sample $(\tilde{X}_i, \tilde{Y}_i)_{i=1}^M$ by taking $(X_i, Y_i) = (\tilde{X}_i, \tilde{Y}_i)$ when $i \neq l$ and $(\tilde{X}_l, \tilde{Y}_l)$ as an independent random variable distributed similarly as (X_l, Y_l). Thus we have simply replaced one of the original random variables by an independent copy. We define $\hat{V}_{M,l}$ as the residual variance estimator in this new sample. Now, because the absolute values of the variables $(Y_i)_{i=1}^M$ are assumed to be bounded by some constant c (we may choose $c = 1$) we have

$$|(Y_i - Y_{N[i,1]})(Y_i - Y_{N[i,2]})| \leq 4 \tag{32}$$

and (see [4], chapter 6.4 for a similar argument)

$$M|\hat{V}_M - \hat{V}_{M,l}| \leq 8 + 8\sum_{j=1}^{2} |\{i :\ X_{N[i,j]} = X_l\}| + 8\sum_{j=1}^{2} |\{i :\ \tilde{X}_{\tilde{N}[i,j]} = \tilde{X}_l\}|. \tag{33}$$

However, for example by lemma 4.1 in [4] (the existence of densities ensures that ties do not occur),

$$8 + 8\sum_{j=1}^{2} |\{i : X_{N[i,j]} = X_l\}| + 8\sum_{j=1}^{2} |\{i : \tilde{X}_{\tilde{N}[i,j]} = \tilde{X}_l\}| \leq K(n) \tag{34}$$

for some constant $K(n)$ depending on the dimension n but not on M. But now the well-known Efron-Stein inequality [13] implies

$$E[(\hat{V}_M - E[\hat{V}_M])^2] \leq \frac{1}{M^2}\sum_{l=1}^{M} E[(\hat{V}_M - \hat{V}_{M,l})^2] \leq \frac{1}{M}K(n) \tag{35}$$

and the proof is complete. We would like to remark, that corresponding results for slightly different estimators can be found for example in [2] and [4]; here we shortly demonstrated how to adapt such result to our special case using a classical concentration inequality. □

The variance goes in general slowly to zero, but in practical problems the major difficulty is the bias of the algorithm, as the variance tends to be small compared to the variance of the output. Once we take into account both the bias and variance parts, we can see that the rate of convergence obtained in [3] is similar to the rate suggested by our analysis in this section.

Finally, we would like to further comment on the accuracy of the algorithm. It can be seen from (21) that the term $b_{i,N[i,1]}b_{i,N[i,2]}$ causes the finite sample bias of the method. For the estimator in (15) the corresponding term is of the form $b^2_{i,N[i,1]}$ as can be seen by a similar calculation as in (21). However, moving to the linear case with $m(x) = w^T x$ (the general case could be analyzed with a Taylor expansion), we may write

$$b_{i,N[i,k]} = w^T(X_i - X_{N[i,k]}), \tag{36}$$

which is related to the angle between the two vectors. Then, it is reasonable to assume, that asymptotically the terms $b_{i,N[i,1]}$ and $b_{i,N[i,2]}$ become uncorrelated leading to a low bias for the method, as the angle between w and $X_i - X_{N[i,k]}$ is in general asymptotically uniformly and independently distributed [4,16]. Based on this observation, we conjecture that actually the novel method improves the original algorithm also in terms of rate of convergence, a fact for which we do not yet have a formal proof. We believe that the actual rate of convergence is $M^{-3/n}$ or even $M^{-4/n}$, of course depending on the regularity of the underlying system. As will be seen in the next section, our hypothesis is confirmed by experimental results, which demonstrate a significant improvement in accuracy compared to [3]. However, a rigorous theoretical analysis remains a topic of future research with potential applications in other estimation problems as well.

Finally, one should remark that the proposed method is very simple. It could be possible to obtain improvement by combining our idea with a more sophisticated tool such as the Gamma test [4] and the local linear estimator in [21]. However, possible benefit probably comes at the cost of decreased robustness in real life applications.

6 Experiments

In the experiments we show, that the theoretical considerations lead to a practical algorithm by comparing the second (modified nearest neighbor) estimator in [3] and our method (16) in three different test problems. Notice that comparison with methods like the Gamma test [4] is not meaningful, because estimators designed for homoscedastic noise variance estimation do not necessarily address heteroscedasticity as demonstrated in Example 1.

6.1 Linear Problems

In the first two experiments the estimators are tested on two linear cases. The results of the experiment are plotted in Figs. 1a and b.

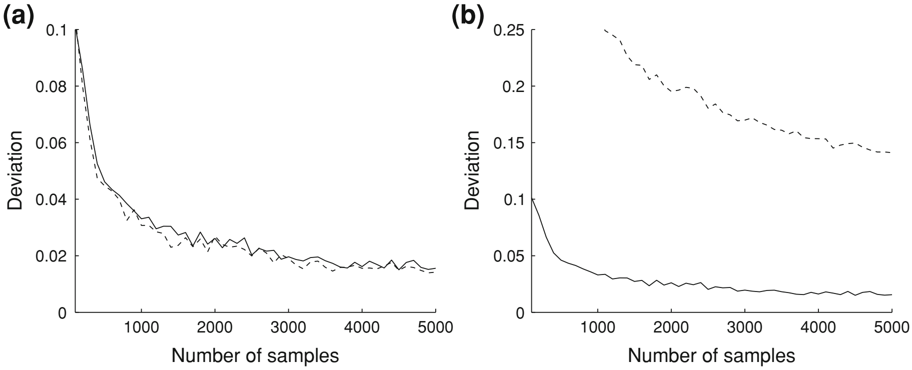

Fig. 1 Results of the linear models with the first experiment in (**a**) and the second in (**b**). The dotted line is the mean absolute deviation of the estimator in [3] and the solid that of the estimator (16)

In the first one, the observations are related to the inputs by

$$Y = X_1 + 3X_2 + \sin(4\pi X_1)\epsilon, \tag{37}$$

where (X_1, X_2) is sampled from the uniform distribution on $[0, 1]^2$ and $\epsilon \sim N(0, 1)$ is independent Gaussian noise. The variance of residual is in this case 0.5 and the variance of the output 10.5. The experiment is repeated 100 times with the number of samples ranging from 100 to 5000 and the mean absolute deviation from the real noise variance is calculated.

The second linear experiment is made to test the methods in a higher dimensional case. In this case the model is

$$Y = X_1 + X_2 + X_3 + X_4 + 3X_5 + \epsilon \tag{38}$$

with $(X_1, X_2, X_3, X_4, X_5) \in [0, 1]^5$ and $\epsilon \sim N(0, 1)$ the number of samples varying between 100 and 5000. Observe that variable X_5 has more weight than the others making the problem more challenging for methods using the Euclidean distance.

In the first experiment, the methods are approximately equivalent, whereas in the second one our method is more accurate. Especially the second problem is challenging for both methods, as the problem is relatively high dimensional, whereas in the first experiment the error is mainly caused by statistical fluctuation of the estimators around their expectations. However, even in the second case the novel estimator achieves reasonable estimates.

6.2 Nonlinear Problems

The third experiment is a highly nonlinear product of sinusoids. The model is

$$Y = \sin(2\pi X_1)\sin(2\pi X_2)\sin(2\pi X_3) + 0.2\sin(4\pi X_1)\epsilon \tag{39}$$

with $\epsilon \sim N(0, 1)$ and again (X_1, X_2, X_3) uniformly distributed. The variance of the residual is 0.02. Again, the mean absolute deviations are calculated over the sample size varying from 100 to 10000 with the results in Fig. 2.

The result with the novel residual variance estimator is again asymptotically good. However, in this problem both methods perform badly with a small number of samples due to the nonlinearity of the data. Convergence is nevertheless approached with a much better result for the proposed novel method.

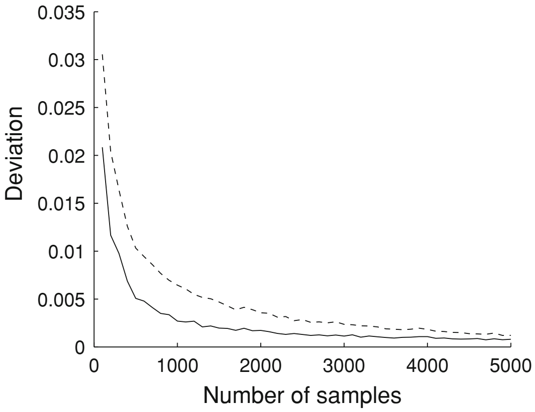

Fig. 2 Results of the parity function experiment. The dotted line is the mean absolute deviation of the estimator in [3] and the solid that of the estimator (16)

7 Conclusion

In this paper, a novel method for residual variance estimation is presented in the context of supervised learning. It seems that despite the usefulness of residual variance estimators they are not well-known in the machine learning community and most work has been done in the context of statistics. Thus one of the goals of this paper is introducing a novel tool for model selection and data analysis.

The theoretical bounds derived for the proposed method imply that the method has good asymptotic properties in low dimensional spaces. The experiments show that in mildly nonlinear problems fast convergence is expected, whereas in highly nonlinear problems a large number of samples may be required. Interestingly, the results strongly support the conjecture that the novel method has better convergence properties than the original method on which it is based with clear practical implications.

In the future, it is of interest to extend the idea to locally linear estimators of residual variance [21]. In this case, better rates of convergences would be obtained with the price of added complexity and thus possibly reduced robustness. This type of a method would fit well in data-sparse high dimensional applications and is thus an interesting topic for future research.

Another important topic is further examination of the properties of the novel method. An interesting open question is, what are the minimal regularity conditions required to obtain the fast rates of convergence discussed at the end of Sect. 5. This type of a theory would also have interesting applications in the field of nonparametric statistics in general.

References

1. Brown LD, Levine M (2006) Variance estimation in nonparametric regression via the difference sequence method. Technical report, Purdue University
2. Devroye L, Wagner TJ (1979) Distribution-free probability inequalities for the deleted and holdout estimates. IEEE Trans Inf Theory 25(2):202–207
3. Devroye L, Schäfer D, Györfi D, Walk H (2003) The estimation problem of minimum mean squared error. Stat Decis 21(1):15–28
4. Evans D (2002) Data-derived estimates of noise for unknown smooth models using near-neighbour asymptotics. Ph.D. thesis, Cardiff University
5. Falconer KJ (1985) The geometry of fractal sets. Cambridge University Press, London

6. Hall P, Marron J (1990) On variance estimation in nonparametric regression. Biometrika 77(2)
7. Hall P, Kay J, Titterington W (1990) Asymptotically optimal difference-based estimation of variance in nonparametric regression. Biometrika 77(3):521–528
8. Jones AJ (2004) New tools in non-linear modelling and prediction. Comput Manage Sci 1(2):109–149
9. Kohler M (2006) Nonparametric regression with additional measurement errors in the dependent variable. J Stat Plann Infer 136(10):3339–3361
10. Kulkarni SR, Posner SE (1995) Rates of convergence of nearest neighbor estimation under arbitrary sampling. IEEE Trans Infor Theory 41(4)
11. Lendasse A, Ji Y, Reyhani N, Verleysen M (2005) LS-SVM hyperparameter selection with a nonparametric noise variance estimator. In: *ICANN 2005, international conference on artifical neural networks*. pp 625–630
12. Liitiäinen E, Lendasse A, Corona F (2007) Non-parametric residual variance estimation in supervised learning. In: *IWANN 2007, international work-conference on artifical neural networks*. pp 63–71
13. Massart P (2007) *Concentration inequalities and model selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003*, vol 1896 of *Lecture notes in mathematics / Ecole d'Eté Probabilités de Saint-Flour*. Springer, Dordrecht
14. Müller U, Schik A, Wefelmeyer W (2003) Estimating the error variance in nonparametric regression by a covariate-matched U-statistic. Statistics 37(3):179–188
15. Pelckmans K, Brabanter JD, Suykens J, Moor BD (2004) The differogram: nonparametric noise variance estimation and its use for model selection. Neurocomputing 69(1–3):100–122
16. Penrose MD (2008) Laws of large numbers in stochastic geometry with statistical applications. Bernoulli 13(4):1124–1150
17. Rice J (1984) Bandwidth choice for nonparametric regression. Ann Stat 12(4):1215–1230
18. Rudin W (1986) Real and complex analysis, Higher Mathematics Series. McGraw-Hill Science, NY
19. Ruppert D, Wand M, Holst U, Hossjer O (1997) Local polynomial variance-function estimation. Technometrics 39(3):262–273
20. Shiryaev AN (1995) Probability. Springer, Dordrecht
21. Spokoiny V (2002) Variance estimation for high-dimensional regression models. J Multivariate Anal 10(4):465–497
22. Yao Q, Tong H (2000) Nonparametric estimation of ratios of noise to signal in nonparametric regression. Statistica Sinica 10(3):751–770