Non-parametric Residual Variance Estimation in Supervised Learning

Elia Liitiäinen, Amaury Lendasse, and Francesco Corona

Helsinki University of Technology - Lab. of Computer and Information Science
P.O. Box 5400, FI-2015 HUT - Espoo, Finland

Abstract. The residual variance estimation problem is well-known in statistics and machine learning with many applications for example in the field of nonlinear modelling. In this paper, we show that the problem can be formulated in a general supervised learning context. Emphasis is on two widely used non-parametric techniques known as the Delta test and the Gamma test. Under some regularity assumptions, a novel proof of convergence of the two estimators is formulated and subsequently verified and compared on two meaningful study cases.

1 Introduction

The residual variance estimation problem is well-known in machine learning and statistics under various contexts [1,2]. Residual variance estimation can be viewed as the problem of estimating the variance of the part of the output that cannot be modelled with the given set of input variables. This type of information is valuable and gives elegant methods to do model selection [2].

While there exist numerous applications of residual variance estimators to supervised learning [3], time series analysis [4] and machine learning [5,2,6], it seems that a rigorous and general framework for analysis is still missing. For example, in [2] and [7] the theoretical model assumes additive noise and independent identically distributed (iid) variables.

The principal objective of this paper is to define such a general framework for residual variance estimation by extending its formulation to the non-iid case. The model is chosen to be realistic from the point of view of supervised learning. Secondly, we view two well-known residual variance estimators, the Delta test [8] and the Gamma test [7] in the general setting and we discuss their convergence properties. Based on the theoretical achievements, our general approach seems to open new directions for future research and it appears of fundamental nature.

The paper is organized as follows: in section 2, we formulate the framework for residual variance estimation in supervised learning. In section 3, we discuss nearest neighbors and prove a novel theoretical result for empirical moments of nearest neighbor distances for later use. In sections 4 and 5 we discuss the Delta test and the Gamma test with some theoretical proofs. Sections 6 and 7 complete the presentation illustrating our experimental results and conclusions.

F. Sandoval et al. (Eds.): IWANN 2007, LNCS 4507, pp. 63–71, 2007.

2 Residual Variance Estimation

By residual variance estimation we mean estimating the lowest possible mean squared error (MSE) in a given regression problem based on data. An abstract formulation of the problem is the goal of this section. Our approach is mainly intended for data-derived modeling using stationary models and is a generalization of the formulation discussed in [7].

2.1 Basic Definitions

Before stating the general form of the problem of residual variance estimation, we provide some general definitions that are needed in the subsequent treatment. Our starting point is standard: we 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 $\mathbb{R}^{n+1}$ with distributions given by the joint densities $p_i(x, y)$ (w.r.t. the Lebesgue measure). 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.

In what follows, we will make the technical assumption that the distributions corresponding to the densities p_i are equivalent; that is, almost surely $p_i(x, y) = 0$ implies $p_j(x, y) = 0$ for any pairs (i, j) and (x, y). Justified by the fact that, in practice, most random variables are bounded, we also assume that the vectors (X_i, Y_i) take values in the unit cube $[0, 1]^{n+1}$.

2.2 Statement of the Problem

In this section, we state the problem of residual variance estimation in the general case of independent observations from the point of view of supervised learning. The novelty of our approach is that we do not assume an additive noise model and independent identically distributed inputs, like in [7], for example.

In the regression (supervised learning) problem, 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 goal is to minimize a cost function, typically, the MSE between the model and the outputs. In this case, the problem reduces to finding the function $g : [0, 1]^n \to \mathbb{R}$ that minimizes

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

even though, in practice, the expectations usually have to be estimated by averaging over the samples available.

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 $[0,1]^n$ by $B([0,1]^n)$, formally, the problem consists of computing $V_M = \inf_{g \in B([0,1]^n)} L_M(g)$. 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 MSE one can achieve using a learning machine. It is not difficult to see that an estimate for V_M is very useful, as it gives a bound after which we may conclude that a model is overfitting [3].

The following proposition characterizes the solution of the problem from the theoretical point of view.

Proposition 1. *The function that minimizes the cost in equation 1 is given by*

$$g(x) = \sum_{i=1}^{M} \frac{p_i(x) E[Y_i | X_i = x]}{\sum_{i=1}^{M} p_i(x)}. \tag{2}$$

If the stationarity condition $E[Y_i|X_i = x] = E[Y_j|X_j = x]$ holds for all $i, j > 0$, then $g(x) = E[Y_i|X_i = x]$ 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 g is given by $g(x) = E[\tilde{Y}|\tilde{X} = x]$. It is a well-known fact that the conditional expectation gives the optimal function in the sense of L^2-norm [9]. Hence, starting from the definition of abstract conditional expectations [9], it is possible to show that g is of the form defined in equation 2.

3 Nearest Neighbors

The concept of nearest neighbors [7] 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 in section 2.2.

The definition of the nearest neighbor is based on the use of a proximity measure to determine similarity between points. Here, we choose the Euclidean metric, which is the most widely used choice and natural in absence of prior information. In such a setting, the nearest neighbor of a point is given by

$$N[i,1] = \text{argmin}_{1 \le j \le M, j \ne i} \|X_i - X_j\|. \tag{3}$$

The k-th nearest neighbor is defined recursively as

$$N[i,k] = \text{argmin}_{1 \le j \le M, j \ne i, N[i,1], \ldots, N[i,k-1]} \|X_i - X_j\|, \tag{4}$$

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

$$\delta_{M,\alpha,k} = \frac{1}{M} \sum_{i=1}^{M} d_{i,k,M}^{\alpha} \tag{5}$$

which is the empirical α-moment for the distances to the k-th nearest neighbor. It is worthwhile noticing that the existence of densities for the variables $(X_i)_{i=1}^M$ ensures that the nearest neighbors are uniquely defined, which would not be the case for discrete valued data.

Interestingly, we have the following novel extension of the moment bound in [10], which shows that $\delta_{M,\alpha,k}$ goes to zero with the rate $M^{-\alpha/n}$. This result is also the best rate one can hope for without assumptions on the instrinsic dimensionality of the data; see, for example, the work of Evans on nearest neighbor distributions [7]. In the following proposition, the notation S_n means the volume of the unit ball in $\mathbb{R}^n$ and $B(x,r)$ the open ball of radius r and center x in $\mathbb{R}^n$.

Proposition 2. *With probability one for* $0 < \alpha \leq n$,

$$\delta_{M,\alpha,k} \leq 9^\alpha k^{2\alpha/n} M^{-\alpha/n}. \tag{6}$$

Proof. Our proof is essentially deterministic. We start by fixing a realization of the sample $(X_i)_{i=1}^M$ and a point $x \in [0,1]^n$. Suppose that $x \in B(X_j, d_{j,k,M})$ for some $0 < j \leq M$. Then, if we define the new sample $(\tilde{X}_i)_{i=1}^{M+1}$ as the union of $(X_i)_{i=1}^M$ and x with $\tilde{X}_{M+1} = x$, we know that in this new sample $x = \tilde{X}_{\tilde{N}[j,l]}$ for some $0 < l \leq k$, where the l-th nearest neigbor is taken in the augmented sample. However, for any choice of r, the number of elements in the set

$$I_{x,r} = \{0 < i \leq M : \tilde{X}_{\tilde{N}[i,r]} = x\} \tag{7}$$

is bounded by $3^n r$ (see [11] and [7]). This, on the other hand, implies that the number of elements in the set

$$I_x = \{0 < i \leq M : \tilde{X}_{\tilde{N}[i,r]} = x, \text{ for some } 0 < r \leq k\} = \cup_{r=1}^k I_{x,r} \tag{8}$$

is bounded by (with the notation $|\cdot|$ for cardinality)

$$|I_x| \leq \sum_{r=1}^k |I_{x,r}| \leq \frac{1}{2} k(k+1) 3^n \leq k^2 3^n. \tag{9}$$

Thus, if we pick a point x, it can belong to at most $k^2 3^n$ different k-th nearest neighbor balls $B(X_j, d_{j,k,M})$. Denoting by $I_{B(x,r)}$ the indicator function of the ball $B(x,r)$ and observing that $\delta_{M,\alpha,k}$ can be written as an integral, we have (using $d_{i,k,M} \leq \sqrt{2}$)

$$\begin{aligned} \delta_{M,n,k} &= \frac{S_n^{-1}}{M} \sum_{i=1}^M \int_{\mathbb{R}^n} I_{B(X_i, d_{i,k,M})}(x) dx \\ &= \frac{S_n^{-1}}{M} \int_{B(0,3)} \sum_{i=1}^M I_{B(X_i, d_{i,k,M})}(x) dx \leq \frac{9^n k^2}{M}. \end{aligned} \tag{10}$$

By Jensen's inequality [9] it can be shown that $\delta_{M,\alpha,k} \leq \delta_{M,n,k}^{\alpha/n}$ which implies that $\delta_{M,\alpha,k} \leq 9^\alpha k^{2\alpha/n} M^{-\alpha/n}$ finishing the proof.

4 Delta Test

Delta test is one of the simplest way to solve the residual variance estimation problem of section 2. The main advantages of this method are robustness and intuitivity, which make it an ideal tool for the applier in low dimensional problems. For some applications of this method we refer, for example, to [8,6].

The idea in Delta test is that similar inputs in the input space tend to produce similar outputs, the difference being caused by statistical fluctuations in the output. To state the Delta test in mathematical terms, we define the sums

$$\gamma_{M,k} = \frac{1}{2M}\sum_{i=1}^{M}(Y_i - Y_{N[i,k]})^2. \tag{11}$$

Then, the Delta test approximates the noise variance V_M (see section 2.2) as $V_M \approx \gamma_{M,1}$. Asymptotically, one would expect this approximation to be a good one. Indeed, next we will give a novel proof of asymptotic unbiasedness in a stationary setting.

Proposition 3. *Assume that for $i, j > 0$ and $x \in [0,1]^n$ the following two stationarity conditions hold with the residual variance $V = V_M$ independent of M:*

$$E[Y_i|X_i = x] = E[Y_j|X_j = x] \tag{12}$$
$$E[(Y_i - E[Y_i|X_i = x])^2|X_i = x] = V \tag{13}$$

and also assume that the function $f(x)$ defined by $f(x) = E[Y_1|X_1 = x]$ is continuous. Then for any choice $k > 0$, $E[\gamma_{M,k}] - V_M \to 0$ as $M \to \infty$. In addition, the convergence $\gamma_{M,k} - E[\gamma_{M,k}] \to 0$ holds in probability.

Proof. By independence of the samples: $E[Y_i - f(X_i)|X_i, X_{N[i,k]}, Y_{N[i,k]}] = E[Y_i - f(X_i)|X_i] = 0$. Based on this observation we conclude that

$$\begin{aligned} &E[(Y_i - f(X_i))(Y_{N[i,k]} - f(X_{N[i,k]}))] \\ &= E[(Y_{N[i,k]} - f(X_{N[i,k]}))E[Y_i - f(X_i)|X_i, X_{N[i,k]}, Y_{N[i,k]}]] = 0. \end{aligned} \tag{14}$$

Set $\Delta_{i,k}f = f(X_i) - f(X_{N[i,k]})$ and $Z_{i,k} = (\Delta_{i,k}f)^2 + 2(Y_i - Y_{N[i,k]} - \Delta_{i,k}f)\Delta_{i,k}f$. Then, by algebraic manipulation and equation 14

$$E[(Y_i - Y_{N[i,k]})^2] = E[(Y_i - f(X_i))^2] + E[(Y_{N[i,k]} - f(X_{N[i,k]}))^2] + E[Z_{i,k}]. \tag{15}$$

The first term in the right hand side is V_M. By the assumptions, $E[(Y_{N[i,k]} - f(X_{N[i,k]}))^2] = V_M$ and, thus, we only need to show that $E[Z_{i,k}] \to 0$. By the boundeness of the output, $|Z_{i,k}| \leq 7|\Delta_k f|$. Choose now $\epsilon, \delta > 0$ such that

$\|x - z\| < \epsilon$ implies that $|f(x) - f(z)| < \delta/7$ for any vectors $x, z \in [0,1]^n$. Then, by proposition 2 (with $I(\cdot)$ the indicator function)

$$\frac{1}{M}\sum_{i=1}^{M}|Z_{i,k}| = \delta + \frac{7}{M}\sum_{i=1}^{M} I(d_{i,k,M} > \epsilon) \leq \delta + \frac{7}{M\epsilon}\sum_{i=1}^{M} d_{i,k,M}$$
$$\leq \delta + 63k^{2/n}\epsilon^{-1}M^{-1/n}. \tag{16}$$

Thus, for any $\delta > 0$, $\limsup_{M\to\infty} |E[\gamma_{M,k}] - V_M| \leq \delta$ which concludes the first part of the proof. For the result $\gamma_{M,k} - E[\gamma_{M,k}] \to 0$ we refer to [7], chapter 7 (the proof in [7] can be straightforwardly generalized to the non-iid case). It seems, moreover, possible to prove almost sure convergence using similar techniques, as discussed in [11].

The first question that arises from our proof is the speed of convergence. It has been shown in a more restricted setting that the bias of the estimator is of order $M^{-2/n}$ (see [10]). Based on this result we may conclude that, from the theoretical point of view, the rate of convergence of the Delta test is reasonable up to the dimension four. However, for the reason that it cannot take advantage of linearity in the mapping between the inputs and outputs, we suggest using it with caution in dimensions higher than two.

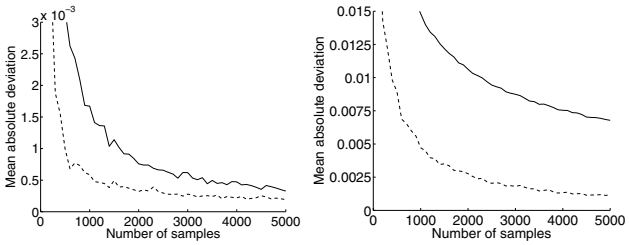
It is worthwhile noticing that replacing condition 13 by $E[(Y_i - E[Y_i|X_i = x])^2|X_i = x] = E[(Y_j - E[Y_j|X_j = x])^2|X_j = x]$ for all i and j (allowing the optimal residual to be place dependent) seems mainly a technical detail.

5 Gamma Test

Because the Delta test is not expected to give accurate estimates of residual variance in dimension higher than four, we discuss in this section an improvement of the method which suits better high dimensional supervised learning problems, the Gamma test. The Gamma test is a well-known method with many applications in machine learning and nonlinear statistics [2]. The convergence has been proven in [7] in a restricted iid setting.

The idea in Gamma test is to assume an approximately linear relationship between $\gamma_{M,k}$ and $\delta_{M,2,k}$ (equations 5 and 11). Then, the estimate for the residual variance V_M is obtained by minimizing the cost function (for some $k > 1$) $C(a,b) = \sum_{l=1}^{k}(\gamma_{M,l} - a - b\delta_{M,2,l})^2$, and taking $V_M \approx a$. The validity of the assumption made when specifying the cost function $C(a,b)$ is by no means trivial. Discussion on this subject can be found in [7] in an iid setting. The next proposition extends the convergence result in [7].

Proposition 4. *Assume that almost surely* $\liminf_{M\to\infty} \delta_{M,2,2}/\delta_{M,2,1} > 1$ *and assumptions of proposition 3 hold. Then the Gamma test estimate converges in probability to* V_M *as* M *goes to infinity.*



(a) (b)

Fig. 1. The experimental results. The solid lines corresponds to the mean absolute deviation from the correct residual variance of Delta test and the dashed to the Gamma test. Figure (a) corresponds to the first experiment and (b) to the second.

Proof. We define $E_k[\delta_{M,2,l}] = \frac{1}{k}\sum_{l=1}^{k}\delta_{M,2,l}$ and $E_k[\gamma_{M,l}]$ in a similar way. Then, the Gamma test estimator can be written in closed form as

$$V_M \approx E_k[\gamma_{M,l}] - \frac{E_k[\delta_{M,2,l}]\sum_{l=1}^{k}(\delta_{M,2,l} - E_k[\delta_{M,2,l}])(\gamma_{M,l} - E_k[\gamma_{M,l}])}{\sum_{l=1}^{k}(\delta_{M,2,l} - E_k[\delta_{M,2,l}])^2}. \quad (17)$$

Denoting the second term in the right hand side by U_k, we notice that, by proposition 3, it is enough to show that $U_k \to 0$. Under the condition $\delta_{M,2,2}/\delta_{M,2,1} > c$ for some $c > 1$, we have the inequality

$$\delta_{M,2,k} - E_k[\delta_{M,2,l}] = \frac{1}{k}\sum_{l=1}^{k}(\delta_{M,2,k} - \delta_{M,2,l}) \geq \frac{1 - c^{-1}}{k}\delta_{M,2,k}. \quad (18)$$

Next, note that $E_k[\delta_{M,2,l}] \leq \delta_{M,2,k}$ and $|\delta_{M,2,l} - E_k[\delta_{M,2,l}]| \leq \delta_{M,2,k}$. We may conclude that $|U_k| \leq C(k)\max_{0<l\leq k}|\gamma_{M,l} - E_k[\gamma_{M,l}]|$ for some constant $C(k)$ which depends only on k. However, by proposition 3, $\gamma_{M,l} - E_k[\gamma_{M,l}] \to 0$ for $0 < l \leq k$, which implies that $U_k \to 0$ (in probability).

The condition $\liminf_{M\to\infty}\delta_{M,2,2}/\delta_{M,2,1} > 1$ seems to hold in practical situations. However, there exists counter-examples where it does not hold and, thus, some assumption on the densities $(p_i)_{i=1}^{\infty}$ is required. Partially this question has been answered in [7], but the non-iid case is still unexplored. Another open question is the speed of convergence of the Gamma test. For discussion see [10], where it is conjectured that the (worst-case) bias of the estimator is of order $M^{-3/n}$, which suggests that in dimensions up to three fast convergence is expected.

6 Experiments

To compare the Delta test and the Gamma test we present two experiments. In the first case, we simulated samples from the highly nonlinear model $Y_i = \sin(\pi X_i^{(1)})\sin(\pi X_i^{(2)}) + \epsilon_i$ with (ϵ_i) independent zero-mean Gaussian noise with

variance 0.01. In the second case, the model is $Y_i = \frac{1}{2}\sin(\pi X_i^{(1)})\sin(\pi X_i^{(2)}) + \frac{1}{2}\sin(X_i^{(3)})\sin(X_i^{(4)}) + \epsilon_i$. In both experiments the mean absolute deviation from the true value is estimated by averaging over 100 simulations. In each experiment the samples $(X_i)_{i=1}^M$ are independent, half of them being sampled from the uniform distribution on $[-1,1]^n$ and the other half from the multidimensional normal distribution (with zero mean and diagonal covariance matrix $\frac{1}{4}\mathrm{I}$) limited to $[-1,1]^n$. For the Gamma test we fix $k = 10$ as proposed in [2].

The results are presented in figure 1. Despite the nonlinearity of the problems, both methods are able to give good estimates in the first experiment, whereas the second one is more challenging due to higher dimensionality of the input space and much more samples are needed for good estimates.

7 Conclusions

In this paper, the residual variance estimation problem is stated in the supervised learning context. Two numerical methods for solving it are presented with proofs of convergence. Clearly, the Gamma test improves the accuracy of the Delta test. However, while the estimators converge rapidly in low dimensional problem, high dimensional nonlinear problems still pose a challenge both from theoretical and practical point of views. Our formulation of the residual variance estimation problem opens new directions for future research. For example, it is of interest to investigate non-stationary systems. Non-parametric residual variance estimators seem to be able to give solutions under relatively weak conditions while at the same time being easy to implement.

References

1. Müller, U., Schik, A., Wefelmeyer, W.: Estimating the error variance in nonparametric regression by a covariate-matched U-statistic. Statistics 37, 179–188 (2003)
2. Jones, A.J.: New tools in non-linear modelling and prediction. Computational Management Science 1, 109–149 (2004)
3. Lendasse, A., Ji, Y., Reyhani, N., Verleysen, M.: LS-SVM hyperparameter selection with a nonparametric noise estimator. In: ICANN 2005. LNCS, vol. 3697, pp. 625–630. Springer, Heidelberg (2005)
4. Kemp, S.E.: Gamma test analysis tools for non-linear time series. PhD thesis, University of Glamorgran (2006)
5. Reyhani, N., Hao, J., Ji, Y., Lendasse, A.: Mutual information and Gamma test for input selection. In: ESANN'2005 proceedings, Bruges (Belgium), pp. 503–508 (27-29 April 2005)
6. Lendasse, A., Corona, F., Hao, J., Reyhani, N., Verleysen, M.: Determination of the Mahalanobis matrix using nonparametric noise estimations. In: ESANN'2006 proceedings, Bruges (Belgium), pp. 227–237 (26-28 April 2006)
7. Evans, D.: Data-derived estimates of noise for unknown smooth models using near-neighbour asymptotics. PhD thesis, Cardiff University (2002)
8. Pi, H., Peterson, C.: Finding the embedding dimension and variable dependencies in time series. Neural Comput 6, 509–520 (1994)

9. Shiryaev, A.N.: Probability. Springer, Heidelberg (1995)
10. Liitiäinen, E., Corona, F., Lendasse, A.: Nearest neighbor distributions and noise variance estimation (accepted for publication). In: ESANN 2007, European Symbosium on Artifical Neural Networks.
11. Devroye, L., Wagner, T. J.: Distribution-free probability inequalities for the deleted and holdout estimates. IEEE Transactions on Information Theory, pp. 202–207, 1979.