ESANN'2007 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2007, d-side publi., ISBN 2-930307-07-2.

Nearest Neighbor Distributions and Noise Variance Estimation

Elia Liitiäinen[1], Francesco Corona[1,2] and Amaury Lendasse[1]

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

2- Università di Cagliari - Dept. of Chemical Engineering and Materials
Piazza d'Armi 1, I-9123 - Cagliari, Italy
{elia,fcorona,lendasse}@cis.hut.fi

Abstract. In this paper, we address the problem of deriving bounds for the moments of nearest neighbor distributions. The bounds are formulated for the general case and specifically applied to the problem of noise variance estimation with the Delta test and the Gamma test. For this problem, we focus on the rate of convergence and the bias of the estimators and validate the theoretical achievements with experimental results.

1 Introduction

Many statistical estimators extensively used in machine learning exploit the properties of nearest neighbor distributions [1]. For instance, the estimators of mutual information by Kraskov *et al.* [3] and the estimators of noise variance proposed by Pi and Peterson [5] and Stefansson *et al.* [6] (the Delta test and the Gamma test, respectively) are based on such properties and are commonly applied in recurrent tasks like model and variable selection.

In this paper, we focus on some theoretical issues concerning arbitrary moments of nearest neighbor distance distributions; in details, we propose rigorous formulation for lower and upper bounds of the moments. The proposed theory is formulated under general and practical assumptions. An application of the theory is examined on the forementioned estimators of noise variance. As for the Delta test, we show that the estimator is asymptotically unbiased with an expected rate of convergence that can be very slow, and we derive the bounds for such rate. As for the Gamma test, the rate of convergence is conjectured under clear hypothesis and the proof is stated as an open problem.

The paper is organized as follows: in Section 2, the moments of nearest neighbor distributions are addressed to derive theoretical bounds; in Section 3, to illustrate our results, we concentrate on noise variance estimation and analyze the rate of convergence of the two estimators. In Section 4, the conclusions are validated with two synthetic experiments.

2 Nearest neighbor distributions

Suppose the random variables $(X_i)_{i=1}^M$ are independent and identically distributed (i.i.d.) according to some bounded probability density (i.e., $p \leq \|p\|_\infty$) with

M describing the number of samples. We assume that the range of the variables is included in a compact set $C \subset \mathbb{R}^n$. By $N[i,k]$ we denote the k-th nearest neighbor of X_i and $d_{i,k}$ is $\|X_i - X_{N[i,k]}\|$ in the Euclidean norm. Let $B(x_0, r) = \{x \in \mathbb{R}^n | \|x - x_0\| < r\}$ denote neighborhood balls in $\mathbb{R}^n$ and define the function

$$\omega_{x_0}(r) = \int_{B(x_0,r)} p(x)dx, \tag{1}$$

corresponding to the probability that a point from C is contained in $B(x_0, r)$.

In the following, we derive theoretical bounds for arbitrary moments of the nearest neighbor distance distribution in order to examine the connection between the *curse of dimensionality* and methods based on such distributions.

2.1 A lower bound

To demonstrate a lower bound for the moment, we state a general proposition.

Proposition 2.1. *For every* $\alpha > 0$

$$E[\omega_{X_i}(d_{i,k})^\alpha | X_i] = \frac{\Gamma(k+\alpha)\Gamma(M)}{\Gamma(k)\Gamma(M+\alpha)}. \tag{2}$$

In Equation 2, $\Gamma(\cdot)$ is the Gamma function and α refers to the α-th moment of nearest neighbor distance distributions [1, 3]. The proof is based on the properties of the Beta function given in [1], where it is also shown that

$$\frac{\Gamma(M)}{\Gamma(M+\alpha)} = M^{-\alpha} + O(M^{-\alpha-1}). \tag{3}$$

Denoting the constant term $\Gamma(k+\alpha/n)/\Gamma(k)$ as $c(k,\alpha,n)$ and V_n the Lebesgue measure of the unit ball in $\mathbb{R}^n$ (i.e., its volume), we demonstrate

Proposition 2.2. *For the constant* $c(k,\alpha,n) \geq 1$,

$$E[d_{i,k}^\alpha] \geq c(k,\alpha,n) V_n^{-\alpha/n} \|p\|_\infty^{-\alpha/n} \frac{\Gamma(M)}{\Gamma(M+\alpha/n)}. \tag{4}$$

Proof. An algebraic manipulation of $E[d_{i,k}^\alpha | X_i]$ yields

$$E[d_{i,k}^\alpha | X_i] \geq L(X_i)^{-\alpha/n} E[\omega_{X_i}(d_{i,k})^{\alpha/n} | X_i], \tag{5}$$

where $L(x) = \sup_{0<r<\infty} \omega_x(r)/r^n$. By equation 2

$$E[\omega_{X_i}(d_{i,k})^{\alpha/n} | X_i] = c(k,\alpha,n) \frac{\Gamma(M)}{\Gamma(M+\alpha/n)}. \tag{6}$$

From $L(X_i) \leq \|p\|_\infty V_n$ it follows that

$$E[d_{i,k}^\alpha | X_i] \geq c(k,\alpha,n) V_n^{-\alpha/n} \|p\|_\infty^{-\alpha/n} \frac{\Gamma(M)}{\Gamma(M+\alpha/n)}. \tag{7}$$

The fact that $E[d_{i,1}]$ is of order $M^{-1/n}$ shows that once the dimension of the input space grows, the average distance to the nearest neighbor goes very slowly to zero when the number of samples is increased. Thus, a slow rate of convergence is expected for many methods based on using nearest neighbors.

2.2 Upper bounds

We derive upper bounds for the α-moment of nearest neighbor distributions from deterministic considerations. Our demonstration stems from the work of Kulkarni and Posner [4]. Because of the less general approach that we are proposing, the derived bounds are different as a less general setting allows tighter bounds. The first proposition is stated assuming $k = 1$ and $0 \in C$, even though it is possible to derive a similar bound for $k > 1$.

Let $\lambda(dx)$ denote the Lebesgue measure on $\mathbb{R}^n$ and define $D(C) = \sup_{x \in C} \|x\|$.

Proposition 2.3. *For any* $0 \leq \alpha \leq n$,

$$\frac{1}{M} \sum_{i=1}^{M} d_{i,1}^{\alpha} \leq 4^{\alpha} D(C)^{\alpha} M^{-\alpha/n}. \tag{8}$$

Proof. First notice that $\|X_i - X_j\| \geq \frac{1}{2}(d_{i,1} + d_{j,1})$. This implies that

$$\|x - X_i\| + \|x - X_j\| \geq \|X_i - X_j\| \geq \frac{1}{2}(d_{i,1} + d_{j,1}) \quad \forall x \in C. \tag{9}$$

Now, $x \in B(X_i, d_{i,1}/2) \cap B(X_j, d_{j,1}/2)$ would lead to a contradiction with equation 9 and thus $B(X_i, d_{i,1}/2) \cap B(X_j, d_{j,1}/2) = \emptyset$. Then, directly from the definition of measure, it follows that

$$\frac{1}{M} \sum_{i=1}^{M} d_{i,1}^{n} = \frac{1}{M} V_n^{-1} 2^n \sum_{i=1}^{M} \lambda(B(X_i, d_{i,1}/2)) \leq 4^n D(C)^n M^{-1}. \tag{10}$$

For any $\alpha \leq n$, Jensen's inequality yields

$$\frac{1}{M} \sum_{i=1}^{M} d_{i,1}^{\alpha} \leq \Big(\frac{1}{M} \sum_{i=1}^{M} d_{i,1}^{n} \Big)^{\alpha/n} \leq 4^{\alpha} D(C)^{\alpha} M^{-\alpha/n}. \tag{11}$$

As in Subsection 2.1, we also state a probabilistic upper bound. The proof is omitted because it can be easily derived by analogy to Proposition 2.1.

Proposition 2.4. *Fix* $\alpha, K > 0$. *Then*

$$E[d_{i,k}^{\alpha}] \leq c(k, \alpha, n) E[S(X_i)^{\alpha/n}] \frac{\Gamma(M)}{\Gamma(M + \alpha/n)} + o(M^{-\alpha/n}) \tag{12}$$

with $S(x) = \sup_{0<r<K} r^n / \omega_x(r)$.

It is possible to prove that asymptotically the remainder term $o(M^{-\alpha/n})$ goes to zero rapidly for any choice $K, \alpha > 0$.

3 Noise variance estimation

In function estimation, in addition to the inputs $(X_i)_{i=1}^M$ we have the scalar outputs $(Y_i)_{i=1}^M$. In the additive noise setting, the functionality between the inputs and outputs is assumed to be $Y_i = f(X_i) + r_i$, where r_i is i.i.d. noise with zero mean. Estimating the variance of noise $\mathrm{Var}[r]$ is an important problem in machine learning and nonlinear statistics [2]. Under the regularity hypothesis assumed in Section 2, we examine two well-known noise variance estimators based on nearest neighbor distributions: the Delta and the Gamma test.

3.1 Delta test and bias

The nearest neighbor formulation of the Delta test estimates $\mathrm{Var}[r]$ by

$$\mathrm{Var}[r] \approx \gamma_1 = \frac{1}{2M}\sum_{i=1}^{M}(Y_i - Y_{N[i,1]})^2, \text{ with } \mathrm{Var}[\gamma_1] \to 0 \text{ for } M \to \infty. \quad (13)$$

Under the hypothesis that f is smooth, we show that the Delta test is asymptotically unbiased, but the convergence may be very slow. By Taylor expansion

$$\begin{aligned} & E[(Y_{N[i,k]} - Y_i)^2] = 2\mathrm{Var}[r] + E[(\nabla f(X_i)\cdot(X_{N[i,k]} - X_i))^2] \\ + \quad & E[R(d_{i,k})]. \end{aligned} \quad (14)$$

The Cauchy-Schwarz inequality and inequality in Proposition 2.3 yields

$$E[(\nabla f(X_i)\cdot(X_{N[i,1]} - X_i))^2] \le 16 \sup_{x\in C}\|\nabla f(x)\|^2 D(C)^2 M^{-2/n}. \quad (15)$$

Proposition 2.4 would give an alternative upper bound. For higher order terms

$$|R(d_{i,1})| \le \frac{1}{2}d_{i,1}^3 \sup_{x\in C}\|\nabla f(x)\|\|D^2 f(x)\| + \frac{1}{4}d_{i,1}^4 \sup_{x\in C}\|D^2 f(x)\|^2. \quad (16)$$

The worst case for rate of convergence is demonstrated for a simple problem. Consider $Y_i = w^T X_i$ with inputs (X_i) uniformly distributed on $[-\frac{1}{2}, \frac{1}{2}]^n$. By symmetry for any function g defined on $\mathbb{R}^{n\times M}$, we have

$$\begin{aligned} & 2^n E[g(X_1^{(1)}, X_1^{(2)}, \ldots, X_2^{(1)}, \ldots, X_N^{(n-1)}, X_N^{(n)})] \\ = \quad & \sum_{\{j_i\}_{i=1}^n \subset \{0,1\}^n} E[g((-1)^{j_1}X_1^{(1)}, \ldots, (-1)^{j_1}X_2^{(1)}, \ldots, (-1)^{j_n}X_N^{(n)})], \end{aligned} \quad (17)$$

which implies the covariance $E[(X_i - X_{N[i,1]})(X_i - X_{N[i,1]})^T] = \frac{1}{n}d_{i,1}^2 I$. From proposition 2.2 we get the following lower bound which is illustrated in the experimental section:

$$\begin{aligned} & E[(\nabla f(X_i)\cdot(X_{N[i,1]} - X_i))^2] \\ = \quad & \frac{1}{n}\|w\|^2 E[d_{i,1}^2] \ge \frac{1}{n}\|w\|^2 V_n^{-2/n}\frac{\Gamma(M)}{\Gamma(M+2/n)}. \end{aligned} \quad (18)$$

3.2 Gamma test and bias

The Gamma test improves the Delta test and it is formulated from the terms

$$\delta_j = \frac{1}{M}\sum_{i=1}^{M} \|X_i - X_{N[i,j]}\|^2 \quad \text{and} \quad \gamma_j = \frac{1}{2M}\sum_{i=1}^{M}(Y_i - Y_{N[i,j]})^2. \tag{19}$$

The estimates $\mathrm{Var}[r] \approx a$ are obtained by setting $k > 0$, and choosing $a, b > 0$ that minimize the function $\sum_{j=1}^{k}(\gamma_j - a - b\delta_j)^2$. For proof of convergence of the estimator, see [1].

The rate of convergence is still an open problem: assuming that the Gamma test maintains the term of order $d_{i,k}^3$ but reduces the terms with lower order (see Equation 14), we state the following conjecture.

Conjecture 3.1. *Under sufficient regularity assumptions, the bias of the Gamma test is at most of order* $M^{-3/n}$.

Even if we show experimental evidence of correctness, the conjecture is to be understood as an open question yet to be rigorously addressed.

4 Experimental results

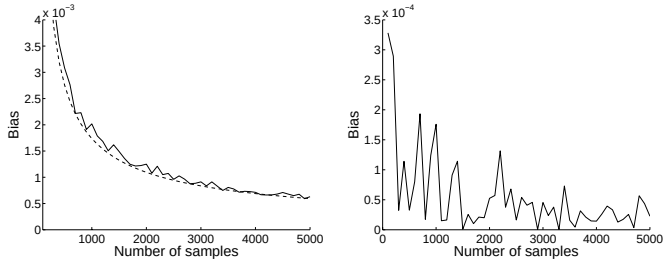
In this Section, we validate our theoretical formulations with experimental results on the problem of estimating the variance of the noise with the Delta and the Gamma test. The experimental setup is synthetic and consists of inputs that are uniformly distributed on $[0, 1]^n$. As suggested in the literature [2], for the Gamma test we fixed $k = 10$.

The first study case we examine is based on a simple linear model of the form $Y = X_1 + X_2 + X_3 + \epsilon$ with $\epsilon \sim N(0, 0.04)$ and $n = 3$. The results are calculated for the number of samples varying between 100 and 5000, with 1000 repetitions to estimate the expectation of the estimator for each number of samples. The curves of the absolute values of the bias are depicted in Figure 1(a) and 1(b). As for the Delta test, the theoretical lower bound is calculated by equation 18.

As for the second study case, we consider a nonlinear setting where the data are generated by the model $Y = \sin(2\pi X_1)\sin(2\pi X_2) + \sin(2\pi X_3)\sin(2\pi X_4) + \sin(2\pi X_5) + \epsilon$, with $\epsilon \sim N(0, 0.1)$ and $n = 5$. The results are reported in Figure 2(a) and 2(b). For this problem, we did not calculate theoretical bounds as the formulation of tight bounds is still investigated. Instead we fitted the curve $aM^{-3/5}$ to the experimental bias curve of the Gamma test.

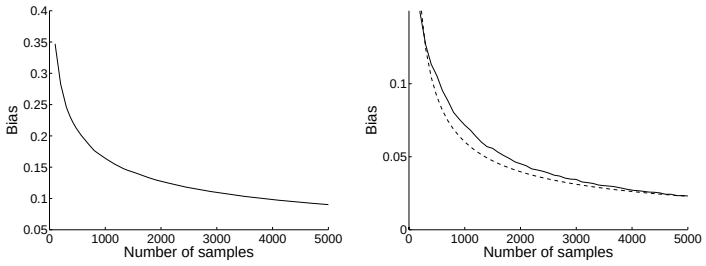
5 Conclusion

In this paper, theoretical bounds are derived for nearest neighbor distributions. Our results are promising and tighter general bounds are still under investigation. The treatment is developed in the general context and, in order to support the presentation, we are focusing on noise variance estimation with the Delta



(a) (b)

Fig. 1: Case Study 1: The bias of the Delta (a) and the Gamma test (b). The dotted line in (a) is the theoretical lower bound. Notice the different orders of magnitude in the diagrams.



(a) (b)

Fig. 2: Case Study 2: The bias of the Delta (a) and the Gamma test (b). In (b) the dotted line is the curve $aM^{-3/5}$ fitted to the results of the experiment.

and the Gamma test. Error bounds for the Delta test are derived and the rate of convergence is illustrated by the experiments. We also conjecture the rate of convergence of the Gamma test and the simulation results support the conjecture.

References

[1] D. Evans. *Data-derived estimates of noise for unknown smooth models using near-neighbour asymptotics.* PhD thesis, Cardiff University, 2002.

[2] A. J. Jones. New tools in non-linear modelling and prediction. *Computational Management Science*, 2004.

[3] A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Phys. Rev. E*, 69:066138, 2004.

[4] S. R. Kulkarni and S. E. Posner. Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE T. Inform. Theory*, 41(4), 1995.

[5] H. Pi and C. Peterson. Finding the embedding dimension and variable dependencies in time series. *Neural Comput.*, 6:509–520, 1994.

[6] A. Stefánsson, N. Koncar, and A. J. Jones. A note on the Gamma test. *Neural Comput. Appl.*, 5:131–133, 1997.