An Affine Invariant k-Nearest Neighbor Regression Estimate

Gérard Biau
Université Pierre et Marie Curie
& Ecole Normale Superieure, France
Email: gerard.biau@upmc.fr

Luc Devroye
School of Computer Science
McGill University
Montreal, Canada
Email: lucdevroye@gmail.com

Vida Dujmović
Carleton University
Ottawa, Canada
Email: vida@scs.carleton.ca

Adam Krzyżak
Department of Computer Science
and Software Engineering
Montreal, Canada
Email: krzyzak@cs.concordia.ca

***Abstract*—We propose a new k-NN regression estimate based on a data-dependent metric in $\mathbb{R}^d$ which is used to define the k-nearest neighbors of a given point. The metric is invariant under all affine transformations. With this metric, the standard k-nearest neighbor regression estimate is asymptotically consistent under the usual conditions on k, and minimal requirements on the input data.**

I. Introduction

The prediction error of standard nonparametric regression methods may be adversely affected by a linear transformation of the coordinate axes. It typically happens with the popular k-nearest neighbor (k-NN) regression estimate defined by Fix and Hodges [8], [9], Cover and Hart [4], Cover [2], [3], where a mere rescaling of the coordinate axes changes the output of the estimate. This is clearly undesirable, especially in applications where data measurements represent different physical items, for instance, weight, blood pressure, sugar level, and the age of the patient. In this example, a simple change in, say, the unit of weight will completely alter the results, and will thus force the statistician to preprocess the input data prior to computing the k-NN estimate.

In this paper, we propose a version of the k-NN regression estimate that is not affected by affine transformations of the coordinate axes. Such a modification could save the user a subjective preprocessing step and would simplify computation of the estimate.

We assume that the data set is a collection of independent and identically distributed $\mathbb{R}^d \times \mathbb{R}$-valued random variables $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$, independent of and with the same distribution as a pair $(\mathbf{X}, Y)$ satisfying $\mathbb{E}|Y| < \infty$. The space $\mathbb{R}^d$ is equipped with the standard Euclidean norm $\|.\|$. For fixed $\mathbf{x} \in \mathbb{R}^d$, our goal is to estimate the regression function $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ from the data $\mathcal{D}_n$. The usual k-NN regression estimate is given by

$$r_n(\mathbf{x}; \mathcal{D}_n) = \frac{1}{k_n} \sum_{i=1}^{k_n} Y_{(i)}(\mathbf{x}),$$

where $(\mathbf{X}_{(1)}(\mathbf{x}), Y_{(1)}(\mathbf{x})), \ldots, (\mathbf{X}_{(n)}(\mathbf{x}), Y_{(n)}(\mathbf{x}))$ is a reordering of the data according to increasing distances $\|\mathbf{X}_i - \mathbf{x}\|$ of the $\mathbf{X}_i$'s to $\mathbf{x}$. (If distance ties occur, a tie-breaking strategy is used. For example, if $\|\mathbf{X}_i - \mathbf{x}\| = \|\mathbf{X}_j - \mathbf{x}\|$, $\mathbf{X}_i$ may be declared "closer" if $i < j$, i.e., the tie-breaking is done by indices.) For simplicity, we will suppress $\mathcal{D}_n$ in the notation and write $r_n(\mathbf{x})$ instead of $r_n(\mathbf{x}; \mathcal{D}_n)$. Stone [29] showed that, for all $p \geq 1$, $\mathbb{E}[r_n(\mathbf{X}) - r(\mathbf{X})]^p \to 0$ for all possible distributions of $(\mathbf{X}, Y)$ with $\mathbb{E}|Y|^p < \infty$, whenever $k_n \to \infty$ and $k_n/n \to 0$ as $n \to \infty$. Thus, the k-NN estimate is asymptotically optimal, for all distributions, or as we say, it is L_p universally consistent.

The drawback of k-NN estimate is that any affine transformation of the coordinate axes affects the k-NN estimate through the norm $\|.\|$. Thus, we want to construct a regression estimate r_n with the following property

$$r_n(\mathbf{x}; \mathcal{D}_n) = r_n(T(\mathbf{x}); \mathcal{D}'_n), \tag{I.1}$$

where $\mathcal{D}'_n = (T(\mathbf{X}_1), Y_1), \ldots, (T(\mathbf{X}_n), Y_n)$. We call r_n affine invariant. In $\mathbb{R}^d$, this invariance of k-NN estimates, it obtained by defining data-dependent affine invariant distance measure. In the next section we define an affine invariant estimation procedure featuring (I.1) which coincides with the k-NN estimate, and discuss its consistency in Section 3.

The literature on affine invariance in nonparametric estimation has been quite rich. The easiest approach is to compute $\hat{M}_n$, the standard estimate of the $d \times d$ covariance matrix of $\mathbf{X}$, and to replace the data $\{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ by $\{(\hat{M}_n^{-1}\mathbf{X}_1, Y_1), \ldots, (\hat{M}_n^{-1}\mathbf{X}_n, Y_n)\}$. Any nonparametric procedure that uses the new data is affine invariant. That method has been discussed in [6] and [14] for pattern recognition and regression, respectively, but it has also been used in kernel density estimation (see, e.g., Samanta [28]). Computational issues aside, this procedure requires at the very least that the second moment of $\mathbf{X}$ exists, which is something we are not willing to assume. Our approach takes ideas from the classical nonparametric literature using concepts such as data depth or multivariate ranks.

There have been attempts to obtain invariance to other transformations. They include invariance under monotone transformations of the coordinate axes, e. g., based on the coordinatewise ranks of the $\mathbf{X}_i$'s. One can show using an L_p norm on the d-vectors of differences between ranks that the classical k-NN regression function estimate is universally consistent in the sense of Stone [29], see Olshen [23] and Devroye [5], Gordon and Olshen [12], [13], Devroye and Krzyżak [7], and Biau and Devroye [1]. Other important examples include the rules based upon statistically equivalent blocks, see Quesenberry and Gessaman [26], Gessaman [10], Gessaman and Gessaman [11], and Devroye, Györfi, and Lugosi [6].

978-1-4673-2579-0/12/$31.00 ©2012 IEEE

II. An affine invariant k-NN estimate

The k-NN estimate discussed in this paper is based upon the notion of empirical distance. We assume that the distribution of $\mathbf{X}$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$ and that $n \geq d$. Thus any collection $\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_d}$ $(1 \leq i_1 < i_2 < \ldots < i_d \leq n)$ of d points among $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are in general position with probability one. Consequently, there exists with probability one a unique hyperplane in $\mathbb{R}^d$ containing these d random points, and we denote it by $\mathcal{H}(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_d})$.

We can now define the empirical distance between d-vectors $\mathbf{x}$ and $\mathbf{x}'$ by

$$\rho_n(\mathbf{x}, \mathbf{x}') = \sum_{1 \leq i_1 < \ldots < i_d \leq n} \mathbf{1}_{\{\text{segment } (\mathbf{x}, \mathbf{x}') \text{ intersects the hyperplane } \mathcal{H}(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_d})\}}.$$

In other words, $\rho_n(\mathbf{x}, \mathbf{x}')$ counts the number of hyperplanes in $\mathbb{R}^d$ passing through d out of the points $\mathbf{X}_1, \ldots, \mathbf{X}_n$, that are separating $\mathbf{x}$ and $\mathbf{x}'$. Intuitively, "near" points have fewer intersections.

This concept of distance based on hyperplanes is known in the multivariate rank tests literature as the empirical lift-interdirection function (Randles [27], Oja [21], Hallin and Paindaveine [15], Oja and Paindaveine [22]). It was introduced in Hettmansperger, Möttönen, and Oja [16] (but not investigated), and independently suggested as an affine invariant alternative to ordinary metrics in the monograph of Devroye, Gÿorfi, and Lugosi [6, Section 11.6]. We use the notion of distance even though, for a fixed sample of size n, ρ_n is only defined with probability one and strictly speaking is not a distance measure. Nevertheless, this empirical distance is invariant under affine transformations $\mathbf{x} \mapsto A\mathbf{x} + \mathbf{b}$, where A is some arbitrary nonsingular linear map and $\mathbf{b}$ any offset vector (see, for instance, Oja and Paindaveine [22, Section 2.4]).

Let us now define our k-NN regression estimate. Pick $\mathbf{x} \in \mathbb{R}^d$ and let $\rho_n(\mathbf{x}, \mathbf{X}_i)$ be the empirical distance between $\mathbf{x}$ and some observation $\mathbf{X}_i$ in the sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$, i. e., the number of hyperplanes in $\mathbb{R}^d$ passing through d out of the observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$, that are cutting the segment $(\mathbf{x}, \mathbf{X}_i)$). Thus our k-NN estimate still takes the familiar form

$$r_n(\mathbf{x}) = \frac{1}{k_n} \sum_{i=1}^{k_n} Y_{(i)}(\mathbf{x}),$$

with the important difference that now the data set $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ is reordered according to increasing values of the empirical distances $\rho_n(\mathbf{x}, \mathbf{X}_i)$, *not* the original Euclidean metric. By construction, the estimate r_n has the desired affine invariance property and, moreover, it coincides with the standard (Euclidean) estimate in dimension $d = 1$. We have the following universal consistency result, where μ denotes the distribution of the random variable $\mathbf{X}$.

Theorem 2.1 (Pointwise L_p consistency): Assume that $\mathbf{X}$ has a probability density, that Y is bounded, and that the regression function r is μ-almost surely continuous. Then, for μ-almost all $\mathbf{x} \in \mathbb{R}^d$ and all $p \geq 1$, if $k_n \to \infty$ and $k_n/n \to 0$,

$$\mathbb{E}\,|r_n(\mathbf{x}) - r(\mathbf{x})|^p \to 0 \quad \text{as } n \to \infty.$$

Theorem 2.1 and the Lebesgue dominated convergence theorem lead to the follwing conclusion.

Corollary 2.1 (Global L_p consistency): Assume that $\mathbf{X}$ has a probability density, that Y is bounded, and that the regression function r is μ-almost surely continuous. Then, for all $p \geq 1$, if $k_n \to \infty$ and $k_n/n \to 0$,

$$\mathbb{E}\,|r_n(\mathbf{X}) - r(\mathbf{X})|^p \to 0 \quad \text{as } n \to \infty.$$

Consistency proof of the standard nearest neighbor estimate given in [29] uses a key result that a given point cannot be the nearest neighbor of more than a constant number other points. This condition does not hold anymore when out transformation is applied, so a single data point may have considerable influence on the estimate. Whether our regression estimate remains universally consistent is an open problem. The consistency can be shown with additional constraints, namely that $\mathbf{X}$ has a density (not necessarily continuous) and that r is μ-almost surely continuous. The outline of the proof is given in the next section. The complete proof will be presented elsewhere.

III. Outline of proof of the theorem

Let $\mathbf{X}$ the random variable with distribution μ and let f be its density with respect to the Lebesgue measure. For every $\varepsilon > 0$, let $\mathcal{B}_{\mathbf{x},\varepsilon} = \{\mathbf{y} \in \mathbb{R}^d : \|\mathbf{y} - \mathbf{x}\| \leq \varepsilon\}$ be the closed Euclidean ball with center at $\mathbf{x}$ and radius ε. Let A^c be the complement of a subset A of $\mathbb{R}^d$. We use Jensen's inequality to bound the L_p error of the regression estimate

$$\begin{aligned}\mathbb{E}\,|r_n(\mathbf{x}) - r(\mathbf{x})|^p &\leq 2^{p-1}\mathbb{E}\left|\frac{1}{k_n}\sum_{i=1}^{k_n}\left[Y_{(i)}(\mathbf{x}) - r\left(\mathbf{X}_{(i)}(\mathbf{x})\right)\right]\right|^p \\ &\quad + 2^{p-1}\mathbb{E}\left[\frac{1}{k_n}\sum_{i=1}^{k_n}\left|r\left(\mathbf{X}_{(i)}(\mathbf{x})\right) - r(\mathbf{x})\right|^p\right] \\ &:= 2^{p-1}\mathbf{I}_n + 2^{p-1}\mathbf{J}_n.\end{aligned}$$

Next we bound $\mathbf{J}_n$ by

$$\begin{aligned}&2^p\zeta^p\,\mathbb{E}\left[\frac{1}{k_n}\sum_{i=1}^{k_n}\mathbf{1}_{\{\|\mathbf{X}_{(i)}(\mathbf{x})-\mathbf{x}\|>\varepsilon\}}\right] \\ &\quad + \left[\sup_{\mathbf{y}\in\mathbb{R}^d:\|\mathbf{y}-\mathbf{x}\|\leq\varepsilon}|r(\mathbf{y}) - r(\mathbf{x})|\right]^p \\ &\quad (\text{since } |Y| \leq \zeta).\end{aligned}$$

The first term converges to 0 by $k_n/n \to 0$ while the second therm can be made arbitrarily small by letting $\varepsilon \to 0$ since r is continuous at $\mathbf{x}$. Thus we have shown that $\mathbf{J}_n \to 0$ as $n \to \infty$.

We next use Marcinkiewicz and Zygmund [18] inequality to bound $\mathbf{I}_n$.

$$\begin{aligned}\mathbf{I}_n &\leq C_p\,\mathbb{E}\left[\frac{1}{k_n^2}\sum_{i=1}^{k_n}\left|Y_{(i)}(\mathbf{x}) - r\left(\mathbf{X}_{(i)}(\mathbf{x})\right)\right|^2\right]^{p/2} \\ &\leq \frac{(2\zeta)^p C_p}{k_n^{p/2}} \qquad (\text{since } |Y| \leq \zeta),\end{aligned}$$

where C_p is some positive constant depending only on p. Consequently, $\mathbf{I}_n \to 0$ as $k_n \to \infty$, and this concludes the proof of the theorem.

References

[1] G. Biau and L. Devroye. On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. *Journal of Multivariate Analysis*, 101:2499–2518, 2010.

[2] T.M. Cover. Estimation by the nearest neighbor rule. *IEEE Transactions on Information Theory*, 14:50–55, 1968.

[3] T.M. Cover. Rates of convergence for nearest neighbor procedures. In *Proceedings of the Hawaii International Conference on Systems Sciences*, pages 413–415, Honolulu, 1968.

[4] T.M. Cover and P.E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13:21–27, 1967.

[5] L. Devroye. A universal k-nearest neighbor procedure in discrimination. In B.V. Dasarathy, editor, *Nearest Neighbor Pattern Classification Techniques*, pages 101–106, Los Alamos, 1991. IEEE Computer Society Press.

[6] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York, 1996.

[7] L. Devroye and A. Krzyżak. New multivariate product density estimators. *Journal of Multivariate Analysis*, 82:88–110, 2002.

[8] E. Fix and J.L. Hodges. *Discriminatory analysis. Nonparametric discrimination: Consistency properties*. Technical Report 4, Project Number 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.
[9] E. Fix and J.L. Hodges. *Discriminatory analysis: Small sample performance*. Technical Report 11, Project Number 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas, 1952.
[10] M. Gessaman. A consistent nonparametric multivariate density estimator based on statistically equivalent blocks. *The Annals of Mathematical Statistics*, 41:1344-1346, 1970.
[11] M. Gessaman and P. Gessaman. A comparison of some multivariate discrimination procedures. *Journal of the American Statistical Association*, 67:468-472, 1972.
[12] L. Gordon and R.A. Olshen. Asymptotically efficient solutions to the classification problem. *The Annals of Statistics*, 6:515–533, 1978.
[13] L. Gordon and R.A. Olshen. Consistent nonparametric regression from recursive partitioning schemes. *Journal of Multivariate Analysis*, 10:611–627, 1980.
[14] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag, New York, 2002.
[15] M. Hallin and D. Paindaveine. Optimal tests for multivariate location based on interdirections and pseudo-Mahalanobis ranks. *The Annals of Statistics*, 30:1103–1133, 2002.
[16] T.P. Hettmansperger, J. Möttönen, and H. Oja. The geometry of the affine invariant multivariate sign and rank methods. *Journal of Nonparametric Statistics*, 11:271–285, 1998.
[17] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.
[18] J. Marcinkiewicz and A. Zygmund. Sur les fonctions indépendantes. *Fundamenta Mathematicae*, 29:60–90, 1937.
[19] C. McDiarmid. On the method of bounded differences. In J. Siemons, editor, *Surveys in Combinatorics, 1989*, London Mathematical Society Lecture Note Series 141, pages 148–188. Cambridge University Press, 1989.
[20] K.S. Miller. *Multidimensional Gaussian Distributions*. Wiley, New York, 1964.
[21] H. Oja. Affine invariant multivariate sign and rank tests and corresponding estimates: A review. *Scandinavian Journal of Statistics*, 26:319–343, 1999.
[22] H. Oja and D. Paindaveine. Optimal signed-rank tests based on hyperplanes. *Journal of Statistical Planning and Inference*, 135:300–323, 2005.
[23] R. Olshen. Comments on a paper by C.J. Stone. *The Annals of Statistics*, 5:632–633, 1977.
[24] K.R. Parthasarathy. *Probability Measures on Metric Spaces*. AMS Chelsea Publishing, Providence, 2005.
[25] V.V. Petrov. *Sums of Independent Random Variables*. Springer-Verlag, Berlin, 1975.
[26] C. Quesenberry and M. Gessaman. Nonparametric discrimination using tolerance regions. *The Annals of Mathematical Statistics*, 39:664–673, 1968.
[27] R.H. Randles. A distribution-free multivariate sign test based on interdirections. *Journal of the American Statistical Association*, 84:1045–1050, 1989.
[28] M. Samanta. A note on uniform strong convergence of bivariate density estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 28:85–88, 1974.
[29] C.J. Stone. Consistent nonparametric regression. *The Annals of Statistics*, 5:595–645, 1977.
[30] R.L. Wheeden and A. Zygmund. *Measure and Integral. An Introduction to Real Analysis*. Marcel Dekker, New York, 1977.