7-2014

k-NN Regression on Functional Data with Incomplete Observations

Sashank J. Reddi
Carnegie Mellon University

Barnabas Poczos
Carnegie Mellon University, bapoczos@cs.cmu.edu

Published In
Proceedings of Uncertainty in Artificial Intelligence, 2014.

k-NN Regression on Functional Data with Incomplete Observations

Sashank J. Reddi
Machine Learning Department
Carnegie Mellon University
sjakkamr@cs.cmu.edu

Barnabás Póczos
Machine Learning Department
Carnegie Mellon University
bapoczos@cs.cmu.edu

Abstract

In this paper we study a general version of regression where each covariate itself is a functional data such as distributions or functions. In real applications, however, typically we do not have direct access to such data; instead only some noisy estimates of the true covariate functions/distributions are available to us. For example, when each covariate is a distribution, then we might not be able to directly observe these distributions, but it can be assumed that i.i.d. sample sets from these distributions are available. In this paper we present a general framework and a k-NN based estimator for this regression problem. We prove consistency of the estimator and derive its convergence rates. We further show that the proposed estimator can adapt to the local intrinsic dimension in our case and provide a simple approach for choosing k. Finally, we illustrate the applicability of our framework with numerical experiments.

1 INTRODUCTION

Machine learning has undergone a paradigm shift in the recent times. Traditional machine learning techniques focused on *simple* form of data such as features modeled as vectors in $\mathbb{R}^p$. However, with the advent of modern data collection methods datasets have not only become huge but also more complex, often involving objects like distributions, functions, and sets. Consider the example of brain connectivity mapping data. The brain contains billions of neurons with several trillion physical connections. Neuroimaging approaches like Diffusion Spectrum Imaging (DSI) attempt to visualize the underlying anatomical architecture of neural pathways by creating 3D probability distributions of water diffusion along nerve fiber bundles. In this example the input data consists of distributions, instead of simple finite dimensional vectors. Likewise, there are many instances where the training data consists of functions. For example, whenever we encounter with time series data (e.g. time series of commodity's price, patient's health monitor, energy usage data), then we can always think of the instances as functions whose domain is the time.

Unfortunately, our understanding of algorithms for such complex data is still limited. Most of the existing machine learning and statistical techniques cannot handle such data, often resorting to ad-hoc approaches; thereby ignoring the underlying rich structure in the data. This necessitates the development of a different machine learning paradigm where the *true structure* in the complex data can be exploited. The goal of this paper is to further advance our knowledge of such algorithms.

One of the central issues working with complex functional data is that it is typically difficult to obtain the exact data (functions or distributions). Hence, our access to the data is often restricted to some noisy estimate of the data. For example, when the input variables are distributions, then it is more natural to assume that we only have finite samples from the distributions, but the true distributions (such as their pdf or cdf) are unknown to us. The empirical distribution can be viewed as a noisy estimate of the distribution. Similarly, in the case of function regression, we have the function values at some selected points rather than whole the function itself. We use the terms "measurement error" and "error in variable" to emphasize this issue of noise in the data.

Although, there have been a few attempts to tackle the issues of "error in variables" [3], most of the earlier works do not fully exploit the scenario where we have control over the measurement error. This is particularly relevant to the applications we are interested in such as distribution regression, where we can obtain more accurate measurements of the data by obtaining

more samples from the distribution.

While working with complex data, it is often desirable to have a simple yet powerful algorithms. One such algorithm often used in traditional machine learning is the k-Nearest Neighbor (k-NN) regression estimation. k-NN based algorithms are easy to use and robust. Furthermore, thanks to the extensive research on nearest neighbor search, there are many efficient algorithms for finding the nearest neighbors [1, 22, 5]. Additionally, k-NN estimators have the virtue of adapting to the local structure of the data [10]. Due to these factors, k-NN estimators are well-suited for complex data. However, very little is understood about these estimators in the context of functional data. To this end, we study the problem of k-NN regression on functional data with measurement error. We present our results in a rather broad framework since working within a general framework allows us to use the same tools across different settings and understand the underlying principles of k-NN estimators.

Main Contributions: Our contributions can be summarized as follows: (i) We provide a general framework for analysis of k-NN estimators for functional data. (ii) We prove consistency of the estimators under weak assumptions. (iii) We derive convergence rates for the estimators. (iv) We provide probabilistic bounds which exploit the local intrinsic structure of the probability measure. (v) We provide an adaptive procedure to select k by exploiting the local intrinsic structure. (vi) We apply the framework in two interesting settings, namely distribution regression and function regression. Due to space constraints, we relegate few longer proofs to the appendix.

2 RELATED WORK

Our work is related to functional data analysis, a new exciting field of statistics. We refer interested readers to [4, 17] for a comprehensive treatment of the topic. However, note that most of these works assume direct access to the covariates without any measurement error. This does not fit our framework for regression over distributions or functions.

One popular approach to deal with *distribution* covaraites in ML tasks is to first embed the distributions into a reproducing kernel Hilbert space (RKHS) and then solve the learning problem using the standard machinery of kernel methods [20, 6, 18]. There are both parametric and non-parametric methods proposed along these lines. Parametric methods usually fit a parametric model to distributions for estimating inner products [9, 8, 12]. Few non-parametric methods for distributions also exist. For example, set kernels (since the samples from the distributions are represented by sets) or kernels over distributions may be used. In this context, it is worthwhile to note that the representer theorem was recently generalized for the space of probability distributions [13].

More recently, Póczos et al. [16] proposed a kernel regression approach for solving the regression problem with distribution covariates and real-valued responses. Convergence and sample complexity of the estimator were analyzed in the paper. Oliva et al. [14] provided a similar analysis for the case where the response is also a distribution. Function regression has also gained considerable interest recently. Oliva et al. [15] provided a functional analogue to the LASSO and studied the statistical properties of the estimator. The functional output case has been studied in [11]. None of these works, however, provide an adaptive algorithm which exploits the local structure of the data, such as the local intrinsic dimensionality. This is an important issue, because this dimension plays an important role in the convergence rate. To design efficient algorithms, it is important to be able to adapt to the local intrinsic dimensions. Another important difference between these algorithms and the estimator we propose here is that none of these algorithms are based on k-NN.

Our work is also related to the error in variables model [3]. However, unlike the latter case where the error is $O(1)$ and is not decreasing, we have control over the error and hence, can obtain very accurate (but expensive) measurements. This is true in the applications of our interest like distribution and function regression. As we will see later, we can exploit this additional flexibility to obtain faster rates of convergence.

There has been fairly extensive research on k-NN estimators for regression problem. Kpotufe at al. [10] study k-NN regression and show that it adapts to local intrinsic dimension. Furthermore, they also provide a simple method to choose k that nearly achieves the minimax rate. But these works do not address the problem of our concern, namely k-NN estimators for functional data with error in measurement.

Notation: The symbol $\mathbb{P}(E)$ is used to denote the probability of event E. We use $X \sim \mathbb{P}$ to denote that the random variable X has probability distribution $\mathbb{P}$. We use the $[n]$ and $i:j$ to denote the set $\{1,\ldots,n\}$ and $\{i,\ldots,j\}$ respectively. The symbol $\mathbb{E}(X)$ is used to denote the expectation of random variable X. We use $B(P,r)$ to denote a ball of radius r centered around P (where P is a point in some metric space).

3 PRELIMINARIES

We start this section with a formal discussion of k-NN based *regression* estimators. We denote by $(\mathcal{P},\rho)$ a

metric space $\mathcal{P}$ with distance measure ρ. We assume that the space with this metric ρ is bounded, i.e., there exists a $\nu > 0$ such that $\rho(P, Q) \leq \nu$ for all $P, Q \in \mathcal{P}$. In a typical regression setting, we have m i.i.d samples $(\mathbf{P}, \mathbf{Y}) = \{(P_i, Y_i)\}_{i=1}^m$ from some unknown distribution over $(\mathcal{P} \times \mathbb{R})$, which is the space of input-output pairs. For example, many machine learning applications usually deal with finite dimensional Euclidean spaces, i.e., $\mathcal{P} = \mathbb{R}^p$ and ρ is the Euclidean distance. We assume that for our observations $\{Y_i\}$ it holds that

$$Y_i = f(P_i) + \gamma_i, \quad i \in [m],$$

where f is a regression function $f : \mathcal{P} \to \mathbb{R}$, and γ_i's are noise variables with $\mathbb{E}[\gamma_i] = 0$ and variance $\mathbb{E}[\gamma_i^2] = \sigma^2$. We assume that the functional f is L-Lipschitz, i.e., $|f(P) - f(P')| \leq L\rho(P, P')$ for all $P, P' \in \mathcal{P}$. We use μ and μ_m to denote the marginal distribution and the empirical distribution on $\mathcal{P}$ respectively. k-NN based regression is fairly well-understood when $\mathcal{P} = \mathbb{R}^p$ and ρ is the Euclidean distance [7].

In functional data analysis, as mentioned earlier, it is usually not possible to obtain the samples P_i exactly, and hence we have to deal with a *noisy* representation of P_i. Our goal, however, is still the same as in standard regression problems: to recover the function f. The type of representation generally depends on the application. For example, in the case of distribution regression (i.e., $\mathcal{P}$ is the space of continuous distributions), we only have access to the samples from the distributions (and not the distributions themselves).

To formalize the notion of noisy representation of the input data, assume that we have m i.i.d. samples $(\widehat{\mathbf{P}}, \mathbf{Y}) = \{(\widehat{P}_i, Y_i)\}_{i=1}^m$ instead of $(\mathbf{P}, \mathbf{Y})$. Here $\widehat{P}_i$ denotes the empirical estimation of distribution P_i.

In what follows, we will discuss the details of k-NN regression for functional data. We first look at the case of fixed k (given as an input). We will later investigate an approach to adaptively select k. The regression estimate at P (function or distribution) using $(\widehat{\mathbf{P}}, \mathbf{Y})$ is defined as follows:

$$\hat{f}(P, \widehat{P}_1, \ldots, \widehat{P}_m) = \sum_{j=1}^{m} Y_j W_j(P, \hat{P}_1, \ldots, \hat{P}_m), \quad (1)$$

where

$$W_j(P, \hat{P}_1, \ldots, \hat{P}_m) = \begin{cases} \frac{1}{k} & \text{if } \hat{P}_j \text{ is } k\text{-NN of } P \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

For the sake of brevity, we use $\hat{f}(P)$ and W_j to denote $\hat{f}(P, \hat{P}_1, \ldots, \hat{P}_m)$ and $W_j(P, \hat{P}_1, \ldots, \hat{P}_m)$ respectively. It should also be noted that more general approaches, such as the generalized version of k-NN with non-uniform weights can also be used through the means of a kernel function [10]. For simplicity we only analyze the case of uniform weight in this paper, but all of our results can be extended to the aforementioned scenario.

Before we delve into the technical details of the regression estimator in Equation (1), we have to introduce the definition of doubling dimension, which will help us deriving upper bounds on the generalization error of the estimator. We just briefly describe the definition here, and refer interested readers to [10] for more details.

Definition 1. *(Doubling Dimension) The marginal distribution μ on $\mathcal{P}$ has a* doubling dimension *at most d if there exists a constant C such that for all $P \in \mathcal{P}$ and $r \geq 0$, we have $\mu(B(P, r)) \leq C\epsilon^{-d}\mu(B(P, \epsilon r))$.*

To illustrate the concept, it is instructive to look at the simple case of Euclidean space $\mathbb{R}^d$ and uniform measure over a closed ball. In this case, it is easy to see that the doubling dimension is d. While this describes a *global* notion of doubling dimension (since it holds uniformly over all region), we will also define and use a local notion in a later section. With this setting in mind, we now analyze the consistency and convergence rates of our estimator in a rather broad framework.

4 GENERAL FRAMEWORK FOR ANALYSIS OF k-NN ESTIMATORS

In this section, we first analyze consistency and convergence rates of the k-NN based estimator in Equation (1) within a general framework. We will investigate probabilistic bounds which depend on the local intrinsic structure of the measure μ on $\mathcal{P}$. Finally, we develop an approach which adapts to the local intrinsic dimension by carefully choosing k. Upper bounds on the convergence rates will also be derived for this estimator.

In this general framework, we assume certain application specific bounds on the estimation of $\hat{P}_i$. In particular, we assume the following:

(i) $\mathbb{E}[\rho(P_i, \hat{P}_i)] \leq \Delta$.

(ii) $\mathbb{P}(\rho(P_i, \hat{P}_i) - \mathbb{E}[\rho(P_i, \hat{P}_i)] > \epsilon) \leq \psi_i(\epsilon)$.

These bounds Δ and ψ_i will be instantiated for the cases of distribution and function regression in later sections. The first term provides an upper bound on how close P_i is to $\hat{P}_i$ in expected sense, while the second term measures how far the random variable $\rho(P_i, \hat{P}_i)$ is from its expected value.

4.1 CONSISTENCY OF ESTIMATOR

In this section we analyze the L_2-consistency of the k-NN estimator in Equation (1). In order to prove the consistency of our estimator, we assume that $k \to \infty$ but $k/m \to 0$ as $m \to \infty$. This assumption is typical in k-NN like estimators [7]. Additionally, we also assume that $\Delta \to 0$ as $m \to \infty$. Using these assumptions, we prove the following consistency result.

Theorem 1. *Suppose $k \to \infty$ and $\lim_{m\to\infty} k/m = 0$. Furthermore, we assume that $\Delta \to 0$ as $m \to \infty$. Then $\hat{f}$ is consistent, i.e., $\lim_{m\to\infty} \mathbb{E}[|\hat{f}(P) - f(P)|^2] = 0$.*

Proof. The proof is in the appendix. □

4.2 CONVERGENCE RATES

We now turn our attention towards the convergence rates of the estimator. In particular, we prove that if the measure μ has finite doubling dimension, then we can get a nonparametric convergence rate that depends on this dimension. We already know that $\mathbb{E}[\rho(\hat{P}_i, P_i)]$ is bounded by Δ. Lemma 5 in the Supplementary material provides a bound for uniform convergence.

Let $\Omega(\epsilon_0)$ denote the event that $\rho(\hat{P}_i, P_i) \leq \Delta + \epsilon_0$ for all $i \in [m]$. From Lemma 5, we know that this event occurs with at least probability $1 - \sum_i \psi_i(\epsilon_0)$. For feasibility, we assume that ϵ_0 is large enough such that $\sum_i \psi_i(\epsilon_0) < 1$.

Theorem 2. *Let $d \geq 3$, $m' = \lfloor \frac{m}{k} \rfloor$, and Δ, ϵ_0 be such that $\nu^{-d} \leq m' \leq (4(\Delta + \epsilon_0))^{-d}$ and $\sum_i \psi_i(\epsilon_0) \leq \frac{1}{m}\nu^{-d}$. Then the following holds:*

$$\mathbb{E}[|\hat{f}(P) - f(P)|^2] \leq \frac{\sigma^2}{k} + 2C'L^2 m'^{-2/d} + 2L^2\Delta^2.$$

for some constant C'.

Proof. The proof is in the appendix. □

4.3 LOCAL INTRINSIC BOUNDS FOR ESTIMATOR

We establish probabilistic convergence bounds for our estimators in this section, building on the work of [10]. Our rates exploit the local intrinsic dimension of the measure μ.

In order to obtain uniform bounds over $\mathcal{P}$, we assume additional structure in our problem setting. We assume that the VC-dimension of class $\mathcal{B}$ of balls on $(\mathcal{P}, \rho)$ is $\nu_{\mathcal{B}}$. To capture the notion of local intrinsic dimension, let us define the following.

Definition 2. *(Local Doubling Dimension) We say the measure μ has local doubling dimension d_l on $B(P, r)$ if we have $\mu(B(P, r')) \leq C\epsilon^{-d_l}\mu(B(P, \epsilon r'))$ for all $r' \leq r$ and $0 < \epsilon < 1$.*

Additionally, similar to [10], we assume a noise model that has uniformly bounded tails and variance. More formally, we have for all $\delta > 0$, there exists $t > 0$ such that $\sup_{P\in\mathcal{P}} \mathbb{P}_{Y|P}(|Y - f(P)| > t) \leq \delta$. Infimum amongst all such t is denoted by $T(\delta)$. Our goal is to obtain a probabilistic upper bound on $|\hat{f}(P) - f(P)|$.

The proof uses results from [10] with additional complexity arising due to the estimation error in the variables themselves. We capture the notion of local intrinsic dimension by the doubling dimension at the neighborhood of the point P. We have the following result for the consistency of the estimator:

Theorem 3. *Suppose μ has local doubling dimension d_l on $B(P, r)$. Let $\epsilon = (3Ck/m\mu(B(P, r)))^{1/d_l}$ and $\alpha_m = (\nu_{\mathcal{B}} \log(2m) + \log(16/\delta))/m$. Also, let $\delta' = \delta + \sum_i \psi_i(\epsilon_0)$. Suppose $\mu(B(P, \epsilon r)) \geq \epsilon^{d_l}\mu(B(P, r))/C \geq 3k/m$. Then the following holds with uniformly over all $P \in \mathcal{P}$ with probability at least $1 - \delta'$,*

$$|\hat{f}(P) - f(P)|^2 \leq \frac{2\nu_{\mathcal{B}} T^2(\delta/4m)\log(4m/\delta) + 16\sigma^2}{k} + 2L^2(r + 2(\Delta + \epsilon_0))^2 \left(\frac{3Ck}{m\mu(B(P, r))}\right)^{2/d_l}.$$

Proof. Let $\tilde{f}(P) = \sum_{i=1}^m W_i f(P_i)$. Using $\tilde{f}(P)$, we get the following:

$$|\hat{f}(P) - f(P)|^2 \leq 2\underbrace{|\hat{f}(P) - \tilde{f}(P)|^2}_{\text{Variance}} + 2\underbrace{|\tilde{f}(P) - f(P)|^2}_{\text{Bias}}$$

This is obtained by simple application of AM-GM inequality. We first derive an upper bound for the bias and then deal with the variance. We have:

$$|\tilde{f}(P) - f(P)| = \Big|\sum_{i=1}^m W_i(f(P_i) - f(P))\Big| \leq \sum_{i=1}^m W_i|(f(P_i) - f(P))| \leq L\sum_{i=1}^m W_i\rho(P, P_i).$$

The first step follows from the fact that $\sum_{i=1}^m W_i = 1$. The second and third steps follow from triangle inequality and Lipschitz continuity of f respectively. Consider the index set $\mathcal{J} = \{i_1, \ldots, i_k\}$ which represents the nearest neighbors of P amongst $\{\hat{P}_1, \ldots, \hat{P}_m\}$ where i_j is used to denote the index of the j^{th} nearest neighbor of P amongst $\{\hat{P}_1, \ldots, \hat{P}_m\}$. Similarly, let us denote by i'_j and $\mathcal{J}'$, the index of j^{th} nearest neighbor of P amongst $\{P_1, \ldots, P_m\}$ and corresponding index set respectively. Furthermore, we use r_k to

denote $\rho(P, P_{i'_k})$. In order to obtain an upper bound on the bias we need to analyze $\max_{j\in\mathcal{J}} \rho(P, P_j)$.

Let $\mathcal{E}$ represent the event that $\max_i D(P_i, \hat{P}_i) \leq \Delta+\epsilon_0$ for all $i \in [m]$. We know that this event occurs with probability at least $1 - \sum_i \psi_i(\epsilon_0)$. Conditioned on the event $\mathcal{E}$, we have the following:

$$\rho(P, \hat{P}_i) \leq \rho(P, P_i) + \rho(P_i, \hat{P}_i) \leq \rho(P, P_i) + \Delta + \epsilon_0.$$

for all $i \in [m]$. This holds due to triangle inequality and the definition of the event $\mathcal{E}$. Let By similar argument, we also have $\rho(P, P_i) \leq \rho(P, \hat{P}_i) + \Delta + \epsilon_0$.

The rest of the argument is conditioned on the event $\mathcal{E}$. Using the above relation, we get the following: $\max_{i\in\mathcal{J}'} \rho(P, \hat{P}_i) \leq r_k + \Delta + \epsilon_0$. This is due to the fact that $\rho(P, \hat{P}_i) \leq \rho(P, P_i) + \Delta + \epsilon_0$ for all $i \in [m]$ and the definition of r_k. Using the above relation, we get the following inequality:

$$\max_{i\in\mathcal{J}} \rho(P, \hat{P}_i) \leq \max_{i\in\mathcal{J}'} \rho(P, \hat{P}_i) \leq r_k + \Delta + \epsilon_0.$$

The first step holds since $\mathcal{J}$ are the indices for the k nearest neighbors of P amongst $\{\hat{P}_1, \ldots, \hat{P}_m\}$. But we also have $\rho(P, P_i) \leq \rho(P, \hat{P}_i) + (\Delta + \epsilon_0)$ for all $i \in [m]$ (since we condition on $\mathcal{E}$). From above argument, the following holds:

$$\begin{aligned}\max_{i\in\mathcal{J}} \rho(P, P_i) &\leq \max_{i\in\mathcal{J}} \rho(P, \hat{P}_i) + (\Delta + \epsilon_0) \\ &\leq r_k + 2(\Delta + \epsilon_0).\end{aligned}$$

In order to complete our analysis for the bias, we need to bound the distance r_k. To this end, we appeal to the bound obtained in [10]. In particular, since $\mu(B(P, \epsilon r)) \geq 3k/m$, by invoking Lemma 10 we have $\mu_m(B(P, \epsilon r)) \geq k/m$. Therefore, with probability at least $1 - \delta$, we have $r_k \leq \epsilon r$. Finally, by using union bound over the event above and $\mathcal{E}$, we get the required bound on the bias.

To establish a bound on the variance, we resort to the bound from [10]. We derive the bounds here for the sake of completeness. We need to bound the term $|\hat{f}(P) - \tilde{f}(P)| = |\sum_{i=1}^m W_i(Y_i - f(P_i))|$. The key step is to utilize the classical VC-theory to obtain a bound on $|Y_i - f(P_i)|$.

More formally, let us first condition on the $\mathbf{P} = \{P_1, \ldots, P_m\}$. By using the concept of VC-dimension and applying union bound, we can obtain the final result. We further restrict our attention to the event where $|Y_i - f(P_i)| < T(\delta_0)$ for all $i \in [m]$. Note that this event occurs with probability at least $1 - m\delta_0 > 0.5$. This is obtained by and simple application of union bound. From Markov inequality, we have:

$$\begin{aligned}\mathbb{P}(\exists P \text{ s.t. } |\hat{f}(P) - \tilde{f}(P)| > 2\mathbb{E}(|\hat{f}(P) - \tilde{f}(P)|) + \epsilon) \leq \\ \mathbb{P}(\exists P \text{ s.t. } |\hat{f}(P) - \tilde{f}(P)| > \mathbb{E}(|\hat{f}(P) - \tilde{f}(P)|) + \epsilon|\mathcal{E})\end{aligned}$$

Also note that,

$$\begin{aligned}\mathbb{P}(\exists P \text{ s.t. } |\hat{f}(P) - \tilde{f}(P)| > \mathbb{E}(|\hat{f}(P) - \tilde{f}(P)|) + \epsilon|\mathcal{E}) \\ \leq n^{\nu_\mathcal{B}} \exp(-2k\epsilon^2/T^2(\delta_0))\end{aligned}$$

This is due to the following facts: (i) changing any of the Y_i's changes the function $|\hat{f}(P) - \tilde{f}(P)|$ by at most $T(\delta_0)/k$ and (ii) VC-dimension of the class of balls $\mathcal{B}$ over $\mathcal{P}$ is $\nu_\mathcal{B}$. Hence, using McDiamard's inequality and union bound we get the above result. Let us take $\delta_0 = \delta/2m$. Now using a union bound over aforementioned events and rewriting the result using AM-GM inequality, we have with probability at least 1 - δ

$$\begin{aligned}|\hat{f}(P) - \tilde{f}(P)|^2 < 8\mathbb{E}[|\hat{f}(P) - \tilde{f}(P)|^2] \\ + \frac{T^2(\delta/2m)}{k}\left(\nu_\mathcal{B} \log(2m/\delta)\right)\end{aligned}$$

To complete the proof, we need to obtain an upper bound on the expected value $\mathbb{E}[|\hat{f}(P) - \tilde{f}(P)|^2]$. This is obtained in the following manner:

$$\begin{aligned}\mathbb{E}[|\hat{f}(P) - \tilde{f}(P)|^2] &= \mathbb{E}[|\sum_i (W_i Y_i - f(P_i))|^2] \\ &= \sum_i W_i^2 \mathbb{E}[|Y_i - f(P_i)|^2] \leq \sigma^2/k\end{aligned}$$

The second equality is obtained from the fact that $Y_i - f(P_i)$ are i.i.d random variables. The last inequality is obtained from the assumption that variance of $Y|P$ is bounded by σ^2. Combining, the bounds obtained for the bias and variance, we get the required result. □

Note the dependence of bounds on d_l, the local doubling dimension rather than d. When $d_l \ll d$, we have obtain much better rates of convergence locally.

4.4 SELECTION OF k AND ADAPTIVE CONVERGENCE RATES

In the previous section, we obtained convergence guarantees which depend on the local intrinsic dimension for k-NN estimators. A natural question to investigate is whether these bounds provide any principled approach to choose k. Intuitively, we can see that such an approach should respect the local structure at the query point P. We derive an approach to choose k by handling the bias-variance tradeoff. Here, Δ and

Algorithm 1 Adaptive Selection of k

1: Let $\theta \leq \log(4m/\delta)$, Δ and ϵ_0 be given.
2: Let $\hat{r}_i$ be i^{th} nearest neighbor amongst $\{\hat{P}_1, \ldots, \hat{P}_m\}$.
3: $k = \arg\min_i \left(\theta/i + \hat{r}_i^2 + 16(\Delta + \epsilon_0)^2 \hat{r}_i\right)$.

ϵ_0 are parameters to the algorithm. Δ can be obtained through upper bound on the error or through

estimation procedures. Intuitively, the above approach can be seen as choosing k which minimizes our upper bound. We make this intuition more formal by the following result.

Theorem 4. *Suppose μ has local doubling dimension d_l on $B(P,r)$. Suppose k is chosen according to Algorithm 1 for each $P \in \mathcal{P}$ and $\hat{f}(P)$ is the k-NN estimate. Assume $((\nu_{\mathcal{B}} \log(2m) + \log(16/\delta))/\theta < m^{4/(6+3d_l)}$. Let $\delta' = \delta + \sum_i \psi_i(\epsilon_0)$. Furthermore, let $r < R$ and $\mu(B(P,r)) > 6Cm^{-1/3}$. Then the following statement holds with probability at least $1-\delta'$ simultaneously for all $P \in \mathcal{P}$.*

$$|\hat{f}(P) - f(P)|^2 \leq 2\left(\frac{C}{\theta} + L^2\right) \times \left((1+16R^2)\left(\frac{3C\theta}{m\mu(B(P,r))}\right)^{\frac{2}{2+d_l}} + 48(\Delta+\epsilon_0)^2\right)$$

where $C = \nu_{\mathcal{B}} T^2(\delta/4m)\log(4m/\delta) + 8\sigma^2$.

Proof. We have the following bound on $|\hat{f}(P) - f(P)|^2$ holds with probability at least $1-\delta$:

$$\begin{aligned}|\hat{f}(P) - f(P)|^2 &\leq \frac{2C}{k} + 2L^2(r_k + 2(\Delta+\epsilon_0))^2 \\ &\leq \frac{2C}{k} + 2L^2(\hat{r}_k + 4(\Delta+\epsilon_0))^2 \\ &\leq \left(\frac{2C}{\theta} + 2L^2\right)\left(\frac{\theta}{k} + (\hat{r}_k + 4(\Delta+\epsilon_0))^2\right)\end{aligned}$$

The first and second inequalities holds from Theorem 3 and the fact that $|r_k - \hat{r}_k| \leq 2(\Delta+\epsilon_0)$ where r_k and $\hat{r}_k$ denote the distance of k^{th} nearest neighbor of P amongst $\{P_1,\ldots,P_m\}$ and $\{\hat{P}_1,\ldots,\hat{P}_m\}$ respectively. Note that the procedure we use exactly minimizes $\frac{\theta}{k} + (\hat{r}_k + 4(\Delta+\epsilon_0))^2$. In order to complete the proof, we need to derive an upper bound for the estimator, we need to provide a bound on the minimum value of $\frac{\theta}{k} + (\hat{r}_k + 4(\Delta+\epsilon_0))^2$. To this end, we borrow ideas from [10] (Theorem 3), which provides a upper bound on the minimum value by explicitly constructing a k that has low objective value. We provide all the details here for sake of completeness.

Let $\tau = \theta^{d_l/(2+d_l)}\left(\frac{m\mu(B(P,r))}{3C}\right)^{2/(2+d_l)}$ Using our assumption on local doubling dimension, we have $\mu(B(P,r)) > 6C\theta m^{-d_l/(2+d_l)} \geq 6c\tau/m$. Let $\epsilon = \left(\frac{3C\tau}{m\mu(B(P,r))}\right)^{1/d_l}$. It is easy to see that from the above relationship that $\epsilon < 1$. Moreover,we have

$$\begin{aligned}&\mu(B(P,\epsilon r)) \geq \epsilon^{d_l}\mu(B(P,r))/C \geq 3\tau/m, \\ &\alpha_m = (\nu_{\mathcal{B}}\log(2m) + \log(8/\delta))/m \leq \frac{\theta}{m}m^{4/6+3d_l} \\ &\qquad \leq \frac{\theta}{m}m^{4/6+3d_l} \leq \frac{\tau}{m}.\end{aligned}$$

Therefore, using Lemma 10, we have $\mu_m(B(P,\epsilon r)) \geq \frac{\tau}{m}$ with probability at least $1-\delta$. This in turn implies that $r_k \leq \epsilon r$ for all $k \leq \tau$.

The following argument shows a bound on r_k^2. First, observe that if $k \leq \tau$, we have

$$r_k^2 \leq (\epsilon r)^2 \leq \left(\frac{3C\tau}{m\mu(B(P,r))}\right)^{2/d_l} R^2 = \frac{R^2\theta}{\tau} \leq \frac{R^2\theta}{k}.$$

The first and last inequalities holds since $k \leq \tau$. Let k_0 be the highest integer for which the above inequality holds. It can be proved that that either k_0 or k_0+1 is larger than τ. If $k_0 > \tau$, then the above statement is obviously true. For the case of $k_0 \leq \tau$, it is easy to see that $k_0 + 1 > \tau$ since k_0 is the highest integer for which $r_k^2 \leq \frac{R^2\theta}{k}$ and this holds for all $k \leq \tau$.

Suppose $k_0 \leq \tau$. Let $k_1 = k_0 + 1$ then $\theta/(k_1) < \epsilon^2$ since $\theta/k_1 < \theta/\tau = \epsilon^2$ when $k_1 > \tau$. Moreover, $r_{k_1} \leq 2^{1/d_l}\epsilon r$ since $\mu(B(P, 2^{1/d_l}\epsilon r)) \geq 6\tau/m$ which in turn implies $\mu_m(B(P, 2^{1/d_l}\epsilon r)) \geq 2\tau/m \geq k_1/m$ (by Lemma 10 and the fact that $k_1 \leq 2\tau$).

In the other case of $k_0 > \tau$, by similar argument, we can prove that $\theta/k_0 < \theta/\tau = \epsilon^2$ and $r_{k_0}^2 \leq R^2\epsilon^2$.

Therefore, either k_0 or k_1, satisfy the following:

$$\left(\theta/k + 4r_k^2\right) \leq (1+16R^2)\epsilon^2 \tag{3}$$

Since k is chosen in such a way that it minimizes $\left(\theta/k + (\hat{r}_k + 4(\Delta+\epsilon_0))^2\right)$, we have the following bound:

$$\begin{aligned}|\hat{f}(P) - f(P)|^2 &\leq \left(\frac{2C}{\theta} + 2L^2\right)\left(\frac{\theta}{k} + (\hat{r}_k + 4(\Delta+\epsilon_0))^2\right) \\ &\leq \min_{k_0,k_1}\left(\frac{2C}{\theta} + 2L^2\right)\left(\frac{\theta}{k} + (\hat{r}_k + 4(\Delta+\epsilon_0))^2\right) \\ &\leq \min_{k_0,k_1}\left(\frac{2C}{\theta} + 2L^2\right)\left(\frac{\theta}{k} + (2\hat{r}_k^2 + 32(\Delta+\epsilon_0)^2)\right) \\ &\leq \left(\frac{2C}{\theta} + 2L^2\right)\left((1+16R^2)\epsilon^2 + 48(\Delta+\epsilon_0)^2\right)\end{aligned}$$

The first and second inequalities follow from the fact that $|\hat{r}_k - r_k| \leq 2(\Delta+\epsilon_0)$ and the criteria of choosing k. The final inequality follows from Equation (3). This gives us the required result. □

The above result shows adaptive convergence rates for the k-NN estimators. We now proceed towards applications of the general framework we just discussed.

5 APPLICATIONS

We discuss specific applications of the general framework introduced in the previous section. More specifically, we look at *distribution* regression and *function*

regression settings. We will see that by using appropriate instantiation of the bounds Δ and ψ_i in the general framework, the results for both these case follow in a straightforward manner.

5.1 DISTRIBUTION REGRESSION

We describe distribution regression problem in this section. We consider a regression problem where the input variables are from the space of continuous 1−Lipschitz probability distributions (i.e., $|P(x)-P(y)| \leq \|x-y\|$) on a compact subset $\mathcal{K} \subset \mathbb{R}^p$ (denoted by $\mathcal{D}$). In this case, $\mathcal{P} = \mathcal{D}$ and we assume ρ to the L_1 distance between distributions, i.e., $\rho(P,Q) = \|P-Q\|_1 = \int |P(x)-Q(x)|dx$. Note that $f : \mathcal{D} \to \mathbb{R}$. We assume that class of balls $\mathcal{B}$ on $\mathcal{D}$ have finite VC-dimension $\nu_{\mathcal{B}}$. Here, we have $Y_i = f(P_i) + \gamma_i$ for all $i \in [m]$.

The measurement error comes into play due to the fact that we do not have access to the probability distributions P_i directly; rather we observe samples $X_{i1}, \ldots, X_{in_i} \sim P_i$. From these samples, we estimate the probability distributions through one of the several density estimation procedures like kernel density estimation, data clustering. Let $\hat{P}_1, \ldots, \hat{P}_m$ be estimated probability distributions corresponding to $P_1, \ldots, P_m$ respectively. To summarize, we think of observations as $(\mathbf{P}, \mathbf{Y}) = \{(\hat{P}_i, Y_i)\}_{i=1}^m$ and our goal is to infer the function f. For ease of exposition, we assume that the number of samples observed for all the distributions, i.e., $n = n_i$ for all $i \in [m]$. To apply our framework, we need to instantiate the bounds Δ and ψ_i.

Bound Δ: We have the following bound on the expected error of estimation of the distributions.

Lemma 1. *Under above conditions, we have*

$$\mathbb{E}[\rho(\hat{P}_i, P_i)] \leq \tilde{C} n^{-1/(2+p)}$$

where $\tilde{C}$ is a constant.

Using this result we can take $\Delta = \tilde{C} n^{-1/(2+p)}$. Refer [19] for details of the proof.

Bound ψ_i: We obtain the following bound ψ_i by using McDiamard's inequality.

Lemma 2. *Under above conditions, we have*

$$\mathbb{P}(\rho(\hat{P}_i, P_i) > \mathbb{E}[\rho(\hat{P}_i, P_i)] + \epsilon) \leq e^{-n\epsilon^2/2}$$

Therefore, $\psi_i(\epsilon) = \exp(-n\epsilon^2/2)$ (see [2] for details).

By using the above bounds, we present the main results for distribution regression. Our first result is the consistency of the estimator $\hat{f}$. From Theorem 1, we have the following result. We set $\Delta = \tilde{C} n^{-1/(2+p)}$ and $\epsilon_0 = n^{-1/(2+p)}$.

Theorem 5. *(Consistency of Estimator) Suppose $k \to \infty$ and $\lim_{m\to\infty} k/m = 0$. Furthermore, we assume that $n = \Omega(\log^{(2+p)/p}(m))$. Then $\hat{f}$ is consistent, i.e., $\lim_{m\to\infty} \mathbb{E}[|\hat{f}(P) - f(P)|^2] = 0$.*

The next result provides convergence rates for distribution regression by directly appealing to Theorem 2.

Theorem 6. *(Convergence Rate) Let $d \geq 3$ and $m' = \lfloor \frac{m}{k} \rfloor$. Assume $n \geq (2\log(m) + 2d\log(\nu))^{(2+p)/p}$ and $d\log(1/\nu) \leq \log(m') \leq d\log(n)/(2+p) - d(\log(4 + 4\tilde{C})))$. Then the following holds:*

$$\mathbb{E}[|\hat{f}(P) - f(P)|^2] \leq \frac{\sigma^2}{k} + \frac{2C'L^2}{m'^{2/d}} + 2\frac{\tilde{C}^2 L^2}{n^{2/(2+p)}}.$$

for some constant C'.

The following result shows that the convergence rates in fact depend on the local intrinsic dimension of the probability measure μ. This is obtain from Theorem 3 of the general framework.

Theorem 7. *(Adaptive Convergence Rates) Suppose μ has local doubling dimension d_l on $B(P,r)$. Let $\epsilon = (3Ck/m\mu(B(P,r)))^{1/d_l}$ and $\alpha_m = (\nu_{\mathcal{B}}\log(2m) + \log(16/\delta))/m$. Let $\delta' = \delta + m\exp(-n^{p/(p+2)}/2)$. Suppose $\mu(B(P,\epsilon r)) \geq \epsilon^{d_l}\mu(B(P,r))/C \geq 3k/m$. Then the following holds with uniformly over all $P \in \mathcal{P}$ with probability at least $1-\delta'$,*

$$|\hat{f}(P) - f(P)|^2 \leq \frac{2\nu_{\mathcal{B}}T^2(\delta/4m)\log(4m/\delta) + 16\sigma^2}{k} + 2L^2\left(r + 2\frac{(\tilde{C}+1)}{n^{1/(2+p)}}\right)^2\left(\frac{3Ck}{m\mu(B(P,r))}\right)^{2/d_l}.$$

The final result for distribution regression shows that by using Algorithm 1, we obtain reasonable adaptive convergence rates.

Theorem 8. *(k-Selection Convergence Rates) Suppose μ has local doubling dimension d_l on $B(P,r)$. Suppose k is chosen according to Algorithm 1 for each $P \in \mathcal{P}$ and $\hat{f}(P)$ is the k-NN estimate. Assume $((\nu_{\mathcal{B}}\log(2m) + \log(16/\delta))/\theta < m^{4/(6+3d_l)}$. Let $\delta' = \delta + m\exp(-n^{p/(p+2)}/2)$. Furthermore, let $r < R$ and $\mu(B(P,r)) > 6Cm^{-1/3}$. Then the following statement holds with probability at least $1-\delta'$ simultaneously for all $P \in \mathcal{P}$.*

$$|\hat{f}(P) - f(P)|^2 \leq 2\left(\frac{C_0}{\theta} + L^2\right) \times \left((1+16R^2)\left(\frac{3C\theta}{m\mu(B(P,r))}\right)^{\frac{2}{2+d_l}} + \frac{48(\tilde{C}+1)^2}{n^{-2/(2+p)}}\right)$$

where $C_0 = \nu_{\mathcal{B}}T^2(\delta/4m)\log(4m/\delta) + 8\sigma^2$.

5.2 FUNCTION REGRESSION

In this section, we describe another interesting application—function regression. Here, the input variables belong to the class of 1-Lipschitz (w.r.t Euclidean distance) functions on $[0,1]$ (denoted by $\mathcal{F}$). For this case, we use $\rho(P,Q) = \|P-Q\|_2 = \sqrt{\int_0^1 (P(x)-Q(x))^2 dx}$ (norm in L^2 space of functions). We again focus only on the case when class of balls $\mathcal{B}$ on $\mathcal{P} = \mathcal{F}$ have finite VC-dimension $\nu_{\mathcal{B}}$. The model is $Y_i = f(P_i) + \gamma_i$ for all $i \in [m]$ where P_i are functions. Similar to the distribution regression, we usually do not have access to the function themselves but only the ability to obtain noisy estimates of function value at certain points. For simplicity, we assume a deterministic design where we query at the points $\{X_{ij}\}_{j=1}^n$ where $X_{ij} = j/n$ for all $i \in [m]$ (see [21] for more details about deterministic design). Therefore, we have $Z_{ij} = P_i(X_{ij}) + \zeta_{ij}$ where ζ_{ij} is the noise variable with $\mathbb{E}(\zeta_{ij}) = 0$ and variance $\tilde{\sigma}^2$.

We can intuitively think of function regression as 2-stage regression problem. We first estimate the functions themselves and then perform another regression on these functions to obtain the functional of interest f. From $\{(X_{ij}, Z_{ij})\}_{j=1}^n$, we can obtain estimated functions $\hat{P}_1, \ldots, \hat{P}_m$ corresponding to $P_1, \ldots, P_m$ respectively. This model now fits our framework perfectly once we have appropriate bounds Δ and ψ_i. To obtain these bounds, we directly appeal to the following well-known bounds for regression.

Bound Δ: We have the following bound for Δ in the case of function regression.

Lemma 3. *Under the conditions mentioned above, we have*

$$\mathbb{E}[\rho(\hat{P}_i, P_i)] \le \bar{C} n^{-1/3}$$

where $\bar{C}$ is a constant.

Proof. From Jensen's inequality, we have

$$\mathbb{E}[\rho(\hat{P}_i, P_i)]^2 \le \mathbb{E}[\rho^2(\hat{P}_i, P_i)] \le \bar{C}^2 n^{-2/3}.$$

The last inequality follows from Corollary 1.2 of [21]. □

Bound ψ_i: To obtain bound ψ_i, we resort to McDiamard's inequality and obtain the following bound.

Lemma 4. *Under the conditions mentioned above, we have*

$$\mathbb{P}(\rho(\hat{P}_i, P_i) > \mathbb{E}[\rho(\hat{P}_i, P_i)] + \epsilon) \le e^{-n\epsilon^2/2}$$

Proof. The result follows from simple application of McDiamard's inequality. □

We now state the main results for function regression by appealing to the general framework. We set $\Delta = \bar{C} n^{-1/3}$ and $\epsilon_0 = n^{-1/3}$. The following result shows consistency of estimator $\hat{f}$ in case of function regression. This is obtained from Theorem 1.

Theorem 9. *(Consistency of Estimator) Suppose $k \to \infty$ and $\lim_{m\to\infty} k/m = 0$. Furthermore, we assume that $n = \Omega(\log^3(m))$. Then $\widehat{f}$ is consistent, i.e., $\lim_{m\to\infty} \mathbb{E}[|\widehat{f}(P) - f(P)|^2] = 0$.*

The next result provides convergence rates for function regression by directly using Theorem 2.

Theorem 10. *(Convergence Rate) Let $d \ge 3$ and $m' = \lfloor \frac{m}{k} \rfloor$. Assume $n \ge (2\log(m) + 2d\log(\nu))^3$ and $d\log(1/\nu) \le \log(m') \le d\log(n)/3 - d(\log(4+4\bar{C})))$. Then the following holds:*

$$\mathbb{E}[|\hat{f}(P) - f(P)|^2] \le \frac{\sigma^2}{k} + \frac{2C'L^2}{m'^{2/d}} + \frac{2\bar{C}^2 L^2}{n^{2/3}}.$$

for some constant C'.

The following result shows that the convergence rates in fact depend on the local intrinsic dimension of the probability measure μ.

Theorem 11. *(Adaptive Convergence Rates) Suppose μ has local doubling dimension d_l on $B(P,r)$. Let $\epsilon = (3Ck/m\mu(B(P,r)))^{1/d_l}$ and $\alpha_m = (\nu_{\mathcal{B}}\log(2m) + \log(16/\delta))/m$. Let $\delta' = \delta + m\exp(-n^{1/3}/2)$. Suppose $\mu(B(P,\epsilon r)) \ge \epsilon^{d_l}\mu(B(P,r))/C \ge 3k/m$. Then the following holds uniformly over all $P \in \mathcal{P}$ with probability at least $1-\delta'$,*

$$|\hat{f}(P) - f(P)|^2 \le \frac{2\nu_{\mathcal{B}} T^2(\delta/4m)\log(4m/\delta) + 16\sigma^2}{k} + 2L^2\left(r + \frac{2(\bar{C}+1)}{n^{1/3}}\right)^2 \left(\frac{3Ck}{m\mu(B(P,r))}\right)^{2/d_l}.$$

The final result for distribution regression shows that by using Algorithm 1, we obtain reasonable adaptive convergence rates.

Theorem 12. *(k-Selection Convergence Rates) Suppose μ has local doubling dimension d_l on $B(P,r)$. Suppose k is chosen according to Algorithm 1 for each $P \in \mathcal{P}$ and $\hat{f}(P)$ is the k-NN estimate. Assume $((\nu_{\mathcal{B}}\log(2m) + \log(16/\delta))/\theta < m^{4/(6+3d_l)}$. Let $\delta' = \delta + m\exp(-n^{1/3}/2)$. Furthermore, let $r < R$ and $\mu(B(P,r)) > 6Cm^{-1/3}$. Then the following holds with probability at least $1-\delta'$ simultaneously for all $P \in \mathcal{P}$.*

$$|\hat{f}(P) - f(P)|^2 \le 2\left(\frac{C_0}{\theta} + L^2\right) \times \left((1+16R^2)\left(\frac{3C\theta}{m\mu(B(P,r))}\right)^{\frac{2}{2+d_l}} + \frac{48(\bar{C}+1)^2}{n^{-2/3}}\right)$$

where $C_0 = \nu_{\mathcal{B}} T^2(\delta/4m)\log(4m/\delta) + 8\sigma^2$.

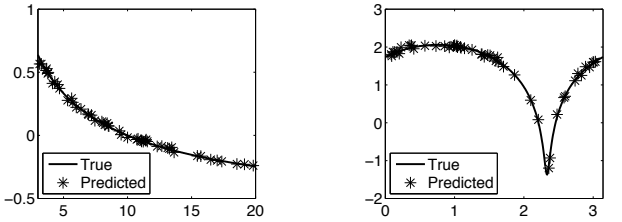
Before ending our discussion of technical results, it is worthy to note two points: (i) Our rates are faster than the logarithmic rates that are sometimes obtained in measurement error nonparametric regression models as in [3]. As mentioned earlier, this is due to the fact that the measurement error corresponds to $\rho(\hat{P}_i, P_i)$ which is not Gaussian for finite n_i and which decreases when n_i increases. (ii) Typically, it is difficult to estimate the distance $\rho(P, Q)$ exactly. This presents additional level of complexity but it can be handled gracefully within our framework by viewing it as another measurement error.

6 EXPERIMENTS

Although the main contribution of our paper is to provide theoretical insights in k-NN based estimation for functional data, we also provide numerical evidence showing the empirical benefits of these estimators. We consider two distribution regression tasks: Beta distribution skewness and Gaussian distribution entropy estimation. In our experiments, we set all the $n, n_1, \ldots, n_m$ set sizes to the same values, which will be specified below. In the first experiment, we generated 325 sample sets from Beta$(a, 3)$ distributions where a was varied between $[3, 20]$ randomly. We constructed $m = 250$ sample sets for training, 25 for validation, and 50 for testing. Each sample set contained $n = 500$ i.i.d samples from Beta$(a, 3)$. Our task in this experiment was to learn the skewness of Beta(a, b) distributions, $f = \frac{2(b-a)\sqrt{a+b+1}}{(a+b+2)\sqrt{ab}}$. We considered the noiseless case, i.e., γ was set to zero. Our estimator is oblivious of the fact that the sample sets are coming from Beta distributions, and it does not know the skewness function values in the test sets either; its values are available only in the training and validation sets.

For obtaining the empirical probability distribution, we use kernel density estimation with Gaussian kernel. The optimal bandwidth is of the kernel is obtained by cross validation. To estimate the L_2 distances between $\hat{p}_i$ and p, we calculated their estimated values in 4096 points on a uniformly distributed grid between the min and max values in the sample sets, and then estimated the integral $\int (p(x) - \hat{p}_i(x))^2 d(x)$ with the rectangle method for numerical integration. To find the appropriate k, we selected the value from $\{1, \ldots, 10\}$ that lead to minimum MSE on validation set. Figure 1(a) displays the predicted values for the 50 test sample sets, and we also show the true values of the skewness functions. As we can see the true and the estimated values are very close to each other.

In the next experiment, our task was to learn the entropy of Gaussian distributions. We chose a 2×2 covariance matrix $\Sigma = AA^T$, where $A \in \mathbb{R}^{2\times 2}$, and A_{ij} was randomly selected from the uniform distribution $U[0, 1]$. Just as in the previous experiments we constructed 325 sample sets from $\{\mathcal{N}(0, R(\alpha_i)\Sigma^{1/2})\}_{i=1}^{325}$. Where $R(\alpha_i)$ is a 2d rotation matrix with rotation angle $\alpha_i = i\pi/325$. From each $\mathcal{N}(0, R(\alpha_i)\Sigma^{1/2})$ distribution we sampled 500 2-dimensional i.i.d. points. Similarly to the previous experiment, 250 points was used for training, 25 for selecting appropriate bandwidth parameters, and 50 for training. Our goal was to learn the entropy of the first marginal distribution: $f = \frac{1}{2}\ln(2\pi e\sigma^2)$, where $\sigma^2 = M_{1,1}$ and $M = R(\alpha_i)\Sigma R^T(\alpha_i) \in \mathbb{R}^{2\times 2}$. μ was zero in this experiment as well. Figure 1(b) displays the learned entropies of the 50 test sample sets. The true and the estimated values are close to each other in this experiment as well.



(a) Skewness of Beta (b) Entropy of Gaussian

Figure 1: (a) Learned skewness of Beta$(a, 3)$ distribution. Axis x: parameter a in $[3, 20]$. Axis y: skewness of Beta$(a, 3)$. (b) Learned entropy of a 1d marginal distribution of a rotated 2d Gaussian distribution. Axes x: rotation angle in $[0, \pi]$. Axis y: entropy. The MSE in two cases are 7.1×10^{-3} and 8.6×10^{-2} respectively.

7 CONCLUSION

We presented a general framework for k-NN estimators for functional data with measurement error. We proved consistency of the estimator and derived upper bounds on the risk. We also analyzed probabilistic bounds capturing the local intrinsic dimension. Furthermore, we presented an algorithm for adaptively choosing k. Two interesting applications of our framework—distribution and function regression—were presented.

In future work, we would like to study lower bounds for the problem and compare our results with the minimax bounds. From practical point of view, it would also be interesting to use these estimators in conjunction with cover trees [1] to obtain fast k-NN estimators. Analyzing the empirical performance on large datasets is another interesting direction.

References

[1] Alina Beygelzimer, Sham Kakade, and John Langford. Cover trees for nearest neighbor. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, 2006.

[2] L. Devroye and G. Lugosi. *Combinatorial methods in density estimation*. Springer, 2001.

[3] J. Fan and Y.K. Truong. Nonparametric regression with errors in variables. *The Annals of Statistics*, pages 1900–1925, 1993.

[4] F. Ferraty and P. Vieu. *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Verlag, 2006.

[5] Alexander Gray and Andrew Moore. 'n-body' problems in statistical learning. In *Advances in Neural Information Processing Systems 13*, pages 521–527. MIT Press, 2000.

[6] S. Grunewalder, G. Lever, L. Baldassarre, S. Patterson, A. Gretton, and M. Pontil. Conditional mean embeddings as regressors. In *ICML*, 2012.

[7] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, New-york, 2002.

[8] T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *NIPS*, pages 487–493. MIT Press, 1998.

[9] T. Jebara, R. Kondor, A. Howard, K. Bennett, and N. Cesa-bianchi. Probability product kernels. *JMLR*, 5:819–844, 2004.

[10] S. Kpotufe. k-nn regression adapts to local intrinsic dimension. *arXiv.org*, 2011.

[11] Nicola Mingotti, Rosa E. Lillo, and Juan Romo. Lasso variable selection in functional regression. Statistics and econometrics working papers, May 2013.

[12] P. Moreno, P. Ho, and N. Vasconcelos. A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications. In *NIPS*, 2004.

[13] Krikamol Muandet, Kenji Fukumizu, Francesco Dinuzzo, and Bernhard Schölkopf. Learning from distributions via support measure machines. In *NIPS*, pages 10–18, 2012.

[14] Junier B. Oliva, Barnabás Póczos, and Jeff G. Schneider. Distribution to distribution regression. In *ICML (3)*, pages 1049–1057, 2013.

[15] Junier B. Oliva, Barnabás Póczos, Timothy Verstynen, Aarti Singh, Jeff G. Schneider, Fang-Cheng Yeh, and Wen-Yih Isaac Tseng. Fusso: Functional shrinkage and selection operator. *CoRR*, abs/1311.2234, 2013.

[16] Barnabás Póczos, Aarti Singh, Alessandro Rinaldo, and Larry A. Wasserman. Distribution-free distribution regression. In *AISTATS*, pages 507–515, 2013.

[17] J.O. Ramsay and B.W Silverman. *Functional data analysis*. Springer, New York, 2nd edition, 2005.

[18] Sashank J. Reddi and Barnabás Póczos. Scale invariant conditional dependence measures. In *Proceedings of the 28th International Conference on Machine Learning*, pages 1355–1363, 2013.

[19] P. Rigollet and R. Vert. Optimal rates for plug-in estimators of density level sets. *Bernoulli*, 15(4):1154–1178, 2009.

[20] A. J. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. volume 4754, pages 13–31. Springer, 2007.

[21] A.B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2010.

[22] Jeffrey K. Uhlmann. Satisfying general proximity/similarity queries with metric trees. *Information Processing Letters*, 40(4):175–179, 1991.