Distribution regression model with a Reproducing Kernel Hilbert Space approach

Thi Thien Trang Bui, Jean-Michel Loubes, Laurent Risser, Patricia Balaresque

Distribution regression model with a Reproducing Kernel Hilbert Space approach

T. Bui [1] & J-M. Loubes [1] & L. Risser [1] & P. Balaresque [2]

[1] *Institut de Mathématiques de Toulouse*
Université Paul Sabatier 118, route de Narbonne F-31062 Toulouse Cedex 9
(tbui, jean-michel.loubes, laurent.risser)@math.univ-toulouse.fr
[2] *Laboratoire d'Anthropologie Moléculaire et Imagerie de Synthèse (AMIS)*
Faculté de Médecine Purpan, 37 allées Jules Guesde, Toulouse
patricia.balaresque@univ-tlse3.fr

Abstract. In this paper, we introduce a new distribution regression model for probability distributions. This model is based on a Reproducing Kernel Hilbert Space (RKHS) regression framework, where universal kernels are built using Wasserstein distances for distributions belonging to $\mathcal{W}_2(\Omega)$ and Ω is a compact subspace of $\mathbb{R}$. We prove the universal kernel property of such kernels and use this setting to perform regressions on functions. Different regression models are first compared with the proposed one on simulated functional data. We then apply our regression model to transient evoked otoascoutic emission (TEOAE) distribution responses and real predictors of the age.

Keywords. Regression, Reproducing kernel Hilbert space, Wasserstein distance, Transient evoked otoscoutic emission.

1 Introduction

Regression analysis is a predictive modeling technique that has been widely studied over the last decades with the goal to investigate relationships between predictors and responses (inputs and outputs) in regression models, see for instance [1, 2] and references therein. When the inputs belong to functional spaces, different strategies have been investigated and used in several application domains about functional data analysis [3, 4]. Extensions of the Reproducing Kernel Hilbert Space (RKHS) framework became recently popular to extend the results of the statistical learning theory in the context of regression of functional data as well as to develop estimation procedures of functional valued functions f [5, 6]. This framework is particularly important in the field of statistical learning theory because of the so-called *Representer theorem*, which states that every function can be written as a linear combination of the kernel function evaluated at training points [7].

In our framework, we aim to solve the regression problem with inputs belonging to probability distribution spaces, whose responses are probability distributions and whose predictors are real values. Specially, we consider the model

$$y_i = f(\mu_i) + \epsilon_i, \tag{1}$$

where $\{\mu_i\}_{i=1}^n$ are probability distributions on $\mathbb{R}$, $\{y_i\}_{i=1}^n$ are real numbers and the ϵ_i represent an independent and identically distributed Gaussian noise. As in classical regression models, this setting estimates an unknown function f from the observations $\{(\mu_i, y_i)\}_{i=1}^n$.

The framework of [8] recently became popular to embed probability distributions into RKHS. It solves the learning problem of distribution regression in a two-stage sampled setting and use the analytical solution of a kernel ridge regression problem to regress from probability measures to real-valued observations. Specifically, the authors embed a distribution to an RKHS $H(k)$ induced by a kernel k which is defined on set of distribution inputs. The regression function is composed of an unknown function f and an element of $H(k)$, where $\mathcal{H}(K)$ is the RKHS induced by kernel K defined on the set of mean embeddings of distributions to RKHS $H(k)$. Whereas the relation between the random distribution and the real number response can be learnt by using directly Representer theorem for the regularized empirical risk over RKHS .

In what follows, we will consider kernels built using the Wasserstein distance. Details on Wasserstein distances and their links with optimal transport problems can be found in [9]. Some kernels with this metric have been developed in [10, 11]. We focus here in the work in [12], in which the authors built a family of positive definite kernel. Within the setting of this paper, we will construct a RKHS corresponding to this kind of kernels to apply the theory of RKHS. More specifically, for an input belonging to Wasserstein spaces, the authors of [12] built a class of positive definite kernels that are functions of Wasserstein distances. More interestingly, the framework of [13], the authors provided a kind of universal kernel tittled the Gaussian-type RBF-kernel. This result is really useful for this paper because from [14, 15] we can build easily a RKHS from a universal kernel. Hence by using the good universal properties, that will be mentioned in Section 3, we will define a new method to construct the RKHS from a universal kernel which is supported by the family of positive definite kernels. Then, we will get a particular estimation from the Representer theorem for an unknown function f in the regression model with distribution inputs.

The paper is structured as follows: In Section 2, we first recall important concepts about kernels on Wasserstein spaces $\mathcal{W}_2(\mathbb{R})$. We then give a brief introduction to Wasserstein spaces on $\mathbb{R}$ and explain how positive definite kernel done are constructed in [12]. Section 3 deals with the proposed setting of distribution regression models. We motivate there the use of universal kernels and build an estimation function for the learning problem. We then assess the numerical performance of this method in Section 4. The tests are first performed on simulated generated data to compare our model with state-of-the-art ones. Then, we study the relationship between the age and hearing sensitivity by using TEOAEs recording that are acquired by stimulating with a very short but strong broadband stimulus. These recordings are then the ear responses by emitting a sound track on a given frequency. More precisely, we predict the age of the subjects, on which they were acquired using the proposed distribution regression model, from TEOAE data. Discussions are finally drawn Section 5.

2 Kernel on Wasserstein space $\mathcal{W}_2(\mathbb{R})$

2.1 The Wasserstein space on $\mathbb{R}$

Let us consider the set $\mathcal{W}_2(\mathbb{R})$ of probability measures on $\mathbb{R}$ with a finite moment of order two. For two μ, ν probability distributions in $\mathcal{W}_2(\mathbb{R})$, we denote $\Pi(\mu, \nu)$ the set of all probability measures π over the product set $\mathbb{R} \times \mathbb{R}$ with first (resp. second) marginal μ (resp. ν).
The transportation cost with quadratic cost function, which we denote quadratic transportation

cost, between two measures μ and ν is defined as:

$$\mathcal{T}_2(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \int |x-y|^2 d\pi(x,y). \tag{2}$$

This transportation cost allows to endow the set $\mathcal{W}_2(\mathbb{R})$ with a metric by defining the quadratic Monge-Kantorovich (or quadratic Wasserstein) distance between μ and ν as

$$W_2(\mu,\nu) = \mathcal{T}_2(\mu,\nu)^{1/2}.$$

A probability measure π in $\Pi(\mu,\nu)$ performing the infimum in (2) is called an optimal coupling. This vocabulary transfers to a random vector (X_1, X_2) with distribution π. We will call $\mathcal{W}_2(\mathbb{R})$ endowed with the distance W_2 the Wasserstein space. More details on Wasserstein spaces and their links with optimal transport problems can be found in [9].

For distributions in $\mathbb{R}$, the Wasserstein distance can be written in a simpler way as follows: For any $\mu \in \mathcal{W}_2(\mathbb{R})$, we denote by F_μ^{-1} the quantile function associated to μ. Given a uniform random variable U on $[0,1]$, $F_\mu^{-1}(U)$ is the random variable with law μ. Then, for every μ and ν the random vector $(F_\mu^{-1}(U), F_\nu^{-1}(U))$ is the optimal coupling (see [9]), where F^{-1} is defined as

$$F_\mu^{-1}(t) = \inf\{u, F_\mu(u) \geq t\}. \tag{3}$$

In this case, the simplest expression for the Wasserstein distance is given in [16]:

$$W_2(\mu,\nu) = \mathbb{E}(F_\mu^{-1}(U) - F_\nu^{-1}(U))^2. \tag{4}$$

Topological properties of Wasserstein spaces are reviewed in [9]. Hereafter, compacity will be required and will be obtained as follows: let $\Omega \subset \mathbb{R}$ be a compact subset, then the Wasserstein space $\mathcal{W}_2(\Omega)$ is also compact. In this paper, we consider Wasserstein spaces $\mathcal{W}_2(\Omega)$, where Ω is a compact subset on $\mathbb{R}$ endowed with the Wasserstein distance W_2. Hence for any $\mu \in \mathcal{W}_2(\Omega)$, we denote $F_\mu|_\Omega : \Omega \to [a,b]$ with $[a,b] \subset [0,1]$ the distribution function restricted on a compact subset Ω. We also define $F_\mu^{-1}|_\Omega$ as:

$$F_\mu^{-1}|_\Omega(t) = inf\{u \in \Omega, F_\mu|_\Omega(u) \geq t\},\ \forall t \in [a,b]. \tag{5}$$

Given a uniform random variable V on $[a,b]$, $F_\mu^{-1}|_\Omega$ is a random variable with law μ. By inheriting properties from $\mathcal{W}_2(\mathbb{R})$ for every μ and ν in $\mathcal{W}_2(\Omega)$, the random vector $\left(F_\mu^{-1}|_\Omega(V), F_\nu^{-1}|_\Omega(V)\right)$ is an optimal coupling. In this case, we consider in this paper the simplest expression for the Wasserstein distance between μ and ν in $\mathcal{W}_2(\Omega)$:

$$W_2(\mu,\nu) = \mathbb{E}(F_\mu^{-1}|_\Omega(V) - F_\nu^{-1}|_\Omega(V))^2. \tag{6}$$

2.2 Kernel

Constructing a positive definite kernel defined on the Wasserstein space is not obvious and was recently done in [12]. For sake of completeness, we recall here briefly this construction.

Theorem 2.1. *Let $k_\Theta : \mathcal{W}_2(\Omega) \times \mathcal{W}_2(\Omega) \to \mathbb{R}$ with the parameter $\Theta := (\gamma, H, l)$ such that $\gamma \neq 0$ and $l > 0$ defined as*

$$k_\Theta(\mu, \nu) := \gamma^2 \exp\left(-\frac{W_2^{2H}(\mu, \nu)}{l}\right). \tag{7}$$

Then for $0 < H \leq 1$, k_Θ is a positive definite kernel.

The proof of this Theorem directly follows Theorem 2.2 and Propositions 2.3. In this paper we use Theorem 2.1 to study the properties of such kernel in the RKHS regression framework.

The following theorem that can be found in [17] or referred to Theorem III.1 in [12], also provides a generic way to construct kernel using completely monotone functions.

Theorem 2.2. ***(Schoenberg)*** *Let $F : \mathbb{R}^+ \to \mathbb{R}^+$ be a completely monotone function, and K a negative definite kernel. Then $(x, y) \mapsto F(K(x, y))$ is a positive definite kernel.*

The following proposition which can be found in [12], finally gives conditions on the exponent H to achieve negative definite kernel using exponents of the Wasserstein distance.

Proposition 2.3. *The function W_2^{2H} is a negative definite kernel if and only if $0 < H \leq 1$.*

Proof. The proof of Theorem 2.1 follows immediately below from Theorem 2.2 and Proposition 2.3. Applying Proposition 2.3, we deduce that $W_2^{2H}(\mu, \nu)$ with $H = 1$ is a negative definite kernel for all μ, ν in $\mathcal{W}_2(\Omega)$.
We can easily see that $e^{-\lambda x}$ with λ positive is a completely monotone function. Let us then consider a mapping as follows:

$$\begin{aligned} F : \mathbb{R}^+ &\to \mathbb{R}^+ \\ x &\mapsto \gamma^2 e^{-\lambda x}, \end{aligned}$$

where $\gamma^2 > 0$, $x = W_2^2(\mu, \nu)$ with $\lambda = \frac{1}{l}$, $l > 0$. Then F is also a completely monotone function. From Theorem 2.2, k_Θ is a positive definite kernel. ■

3 Regression

3.1 Setting

In this section, we aim to define a regression function with distribution inputs. The problem of distribution regression consists in estimating an unknown function $f : \mathcal{W}_2(\Omega) \to \mathbb{R}$ by using observations (μ_i, y_i) in $\mathcal{W}_2(\Omega) \times \mathbb{R}$ for all $i = 1, \cdots, n$. We recall observes in (1) as follows

$$y_i = f(\mu_i) + \epsilon_i. \tag{8}$$

To provide a general form for functions defined on distributions, we will use the RKHS framework. Let $k_\Theta : \mathcal{W}_2(\Omega) \times \mathcal{W}_2(\Omega) \mapsto \mathbb{R}$ be defined in Theorem 2.1. For a fixed valid Θ, we define a space $\mathcal{F}_0$ as follows:

$$\mathcal{F}_0 := \operatorname{span}\left\{k_\Theta(\bullet, \mu) : \mu \in \mathcal{W}_2(\Omega)\right\}.$$

And $\mathcal{F}_0$ is endowed with the inner product

$$\langle f_n, g_m \rangle_{\mathcal{F}_0} = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k_\Theta(\mu_i, \nu_j),$$

where $f_n(\bullet) = \sum_{i=1}^{n} \alpha_i k_\Theta(\bullet, \mu_i)$ and $g_m(\bullet) = \sum_{j=1}^{m} \beta_j k_\Theta(\bullet, \nu_j)$. The norm in $\mathcal{F}_0$ is corresponding to the inner product,

$$\|f_n\|^2_{\mathcal{F}_0} = \sum_{i=1}^{n} \alpha_i^2 k_\Theta(\mu_i, \mu_i). \tag{9}$$

Let $\mathcal{F}$ be the space of all continuous real-valued functions from $\mathcal{W}_2(\Omega)$ to $\mathbb{R}$. The set $\mathcal{F}_0$ consists of all functions in $\mathcal{F}$ which are uniform limits of functions of form f_n. We want to approximate $\mathcal{F}_0$ as well as possible $\mathcal{F}$. Following universal approximating property that is a universal kernel k_Θ has a property that $\mathcal{F}_0 = \mathcal{F}$. Hence we will consider in the following section that a universal kernel of k_Θ to prove that $\mathcal{F}_0$ is dense in $\mathcal{F}$ and that $\mathcal{F}_0 = \mathcal{F}$.
From that for all f, g belong in $\mathcal{F}$, the inner product is well defined as following formula

$$\langle f, g \rangle_{\mathcal{F}} := \lim_{n \to \infty} \langle f_n, g_n \rangle_{\mathcal{F}_0}. \tag{10}$$

Coming back to our problem, we want to estimate the unknown function f by an estimation function $\hat{f}$ obtained by minimizing the regularized empirical risk over the RKHS $\mathcal{F}$. For this consider, we solve the solution of the minimization problem

$$\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \left(\sum_{i=1}^{n} |y_i - f(\mu_i)|^2 + \lambda \|f\|^2_{\mathcal{F}} \right), \tag{11}$$

where $\lambda \in \mathbb{R}^+$ is the regularization parameter. Using the Representer theorem, this leads to the following expression for $\hat{f}$,

$$\hat{f} : \ \mu \mapsto \hat{f}(\mu) := \sum_{j=1}^{n} \hat{\alpha}_j k_\Theta(\mu, \mu_j), \tag{12}$$

where $\{\hat{\alpha}_j\}_{j=1}^{n}$ are parameters typically obtained from training data.

3.2 Universal kernel

First, we recall the definition of a universal kernel and the main theorem to ensure universal properties of positive definite kernels in Theorem 2.1.

Definition 3.1. *Let $C(X)$ be the space of continuous bounded functions on compact domain X. A continuous kernel k on domain X is called universal if the space of all functions induced by k is dense in $C(X)$, i.e, for all $f \in C(X)$ and every $\epsilon > 0$ there exists a function g induced by k with $\|f - g\|_\infty \leq \epsilon$.*

For more information on universal kernel and RKHS, we refer to Chapter 4 in [9] and [14], [15].

Theorem 3.2. *Let choose the parameter Θ in (7) such that $\gamma \neq 0$, $l > 0$ and $H = 1$. The kernel $k_\Theta : \mathcal{W}_2(\Omega) \times \mathcal{W}_2(\Omega) \to \mathbb{R}$ defined in (7) is universal.*

The proof of this Theorem relies on the two following Proposition 3.3 and Proposition 3.4.

Proposition 3.3. *Let $F_\mu|_\Omega : \Omega \to [a, b]$ with $[a, b] \subset [0, 1]$ be the distribution function restricted on a compact subset Ω of $\mathbb{R}$, $F_\mu^{-1}|_\Omega$ be defined by $F_\mu^{-1}|_\Omega(t) = inf\{u \in \Omega, F_\mu|_\Omega(u) \geq t\}$ for all $t \in [a, b]$. Then $F_\mu|_\Omega$ is continuous if and only if $F_\mu^{-1}|_\Omega$ is strictly increasing on $[inf\ ran F_\mu|_\Omega, sup\ ran F_\mu|_\Omega]$. $F_\mu|_\Omega$ is strictly increasing if and only if $F_\mu^{-1}|_\Omega$ is continuous on $ran F_\mu|_\Omega$, where $ran F_\mu|_\Omega := \{F_\mu|_\Omega(x) : x \in \Omega\}$, the range of $F_\mu|_\Omega$.*

See e.g [18] for a proof of Proposition 3.3.

Proposition 3.4. *Let X be a compact metric space and $\mathcal{H}$ be a separable Hilbert space such that there exist a continuous and injective map: $\rho : X \to \mathcal{H}$. For $\sigma > 0$, the Gaussian-type RBF-kernel $k_\sigma : X \times X \to \mathbb{R}$ is the universal kernel, where*

$$k_\sigma(x, x') := \exp(-\sigma^2 \|\rho(x) - \rho(x')\|_{\mathcal{H}}^2), \quad x, x' \in X.$$

See a part of Theorem 2.2 in [13] for proof of Proposition 3.4.

Proof. Proof of Theorem 3.2.
From Proposition 3.3 with the conditions including the distribution restricted on Ω, $F_\mu|_\Omega$ be continuous and $F_\mu|_\Omega$ be strictly increasing on Ω, then there exists a continuous and injective map

$$\begin{aligned} \rho : \mathcal{W}_2(\Omega) &\to \mathbb{L}_2[a, b] \\ \mu &\mapsto \rho(\mu) := F_\mu^{-1}|_\Omega. \end{aligned}$$

$F_\mu^{-1}|_\Omega$ is continuous on $[a, b]$ and strictly increasing on $[\inf \text{ran} F_\mu|_\Omega, \sup \text{ran} F_\mu|_\Omega]$.
We consider a Wasserstein space $\mathcal{W}_2(\Omega)$ metrized by the Wasserstein distance W_2 with Ω be a compact subset on $\mathbb{R}$ and $\mathbb{L}_2[a, b]$ be the usual space of square integrable functions on $[a, b]$. For σ in Proposition 3.4 is exactly defined by $1/\sqrt{l}$ for all $l > 0$. We have

$$k_\Theta(\mu, \nu) = \gamma^2 \exp\left\{-\frac{\|F_\mu^{-1}|_\Omega - F_\nu^{-1}|_\Omega\|_{\mathbb{L}_2[a,b]}^2}{l}\right\}$$

is the universal kernel from Proposition 3.4. We complete the proof of Theorem 3.2. ■

The minimization program in (11) can be solved explicitely using the representer theorem of [19]. Note that Schölkopf and Smola [20] give a simple proof of a more general version of the theorem. Define c_{ij} as follows

$$c_{ij} = \gamma^2 \exp\left(-\frac{W_2^2(\mu_i, \mu_j)}{l}\right)$$

and $\alpha = (\alpha_1, \cdots, \alpha_n)^T$, $Y = (y_1, \cdots, y_n)^T$.
Now we take the matrix formulation of (11) we obtain

$$\min_{\alpha} trace((Y - C\alpha)(Y - C\alpha)^T) + \lambda trace(C\alpha\alpha^T), \tag{13}$$

where the operation trace is defined as

$$trace(A) = \sum_{i=1}^{n} a_{ii}$$

with $A = (a_{ii})_{i=1}^n$.
Taking the derivative of (13) with respect to vector α, we find that α satisfies the system of linear equations

$$(C + \lambda I)\alpha = Y. \tag{14}$$

Hence

$$\hat{f}(\mu) = \sum_{j=1}^{n} \hat{\alpha}_j k_\Theta(\mu, \mu_j), \tag{15}$$

with

$$\hat{\alpha} = (C + \lambda I)^{-1} Y. \tag{16}$$

4 Numerical Simulations and Real data application

4.1 Simulation

4.1.1 Overview of the simulation procedure

In this section, we investigate the regression model for predicting the regression function from distributions. Particularly, we want to estimate the unknown function f in model (8) by using the proposed estimation $\hat{f}$ in (15), so we need to present how we can optimize the parameters in this formula. We then compare the regression model based on RKHS induced by our universal kernel function to more classical kernel functions operating on projections of the probability measures on finite dimensional spaces. We address the input-output map given by

$$f(\nu) = \frac{m_\nu}{0.05 + \sigma_\nu}, \tag{17}$$

where ν is a Gaussian distribution of mean m_ν and variance σ_ν^2. We consider the ground truth function f that we compare with a predicted function $\hat{f}$, such as:

$$\hat{f}(\nu) = \gamma^2 \sum_{j=1}^{n} \hat{\alpha}_j \exp\left[-\frac{W_2^2(\nu, \mu_j)}{l}\right], \tag{18}$$

where the Wasserstein distance between two Gaussian distribution is calculated using:

$$W_2^2(\mu, \nu) = (m_\mu - m_\nu)^2 + (\sigma_\mu - \sigma_\nu)^2,$$

where $\mu \sim \mathcal{N}(m_\mu, \sigma_\mu^2)$ and $\nu \sim \mathcal{N}(m_\nu, \sigma_\nu^2)$.
Each value $\hat{\alpha}_j$ is estimated using Eq. (16) which depends on parameter $\lambda > 0$. Thus our proposed estimation function $\hat{f}$ depends totally on the three parameters $\lambda > 0, \gamma \neq 0$ and $l > 0$. To understand the effects of these parameters on $\hat{f}$, we define reference values of λ, γ and l. We then generate a training set including the normal distributions $\nu_i \sim \mathcal{N}(m_{\nu_i}, \sigma_{\nu_i})$ such that $cor(\nu_i, \nu_j) \neq 0, \forall i, j = 1, \cdots n$, with n be a size of training set. In this simulation, we take $n = 200$.
From the training set $\{(\nu_i, f(\nu_i))\}_{i=1}^{n=200}$, we fit two regression models which we call "Wasserstein" and "Legendre", for which we provide more details below. Then we evaluate the quality of the two regression models on a test set of size n_t of the form $\{(\nu_{t,i}, f(\nu_{t,i}))\}_i^{n_t}$, where $\nu_{t,i}$ is generated in the same way as ν_i above. We consider the following quality criteria, that is the root mean square error (RMSE) to see the qualify of our regression model

$$RMSE^2(\hat{f}, f) = \frac{1}{n_t} \sum_{i=1}^{n_t} \left[f(\nu_{t,i}) - \hat{f}(\nu_{t,i}) \right]^2.$$

4.1.2 Detail on the regression models

We refer our model by *Wasserstein* and introduce briefly *Legendre* regression models. *Wasserstein* model first propose the estimated function as follows

$$\hat{f}(\nu_{t,i}) = \gamma^2 \sum_{j=1}^{n} \hat{\alpha_j} \exp\left[-\frac{(m_{\nu_j} - m_{\nu_{t,i}})^2 + (\sigma_{\nu_j} - \sigma_{\nu_{t,i}})^2}{l} \right], \tag{19}$$

where $\nu_{t,i}, i = 1, \cdots, n_t$ belong to testing set with size n_t, and $\nu_j, j = 1, \cdots, n$ belong to training set with size n. The estimated function $\hat{f}$ depends on three parameters $\gamma \neq 0,\ \lambda > 0$ and $l > 0$.

The *Legendre* model is based on kernel functions operating on finite dimensional linear projections of the distributions. For a Gaussian distribution $\mu \sim \mathcal{N}(m, \sigma^2)$ with density $f_\mu(t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t-m)^2}{2\sigma^2}\right)$ and support $[0, 1]$, we compute for $i = 0, \cdots, \theta - 1$:

$$a_i(\mu) = \int_0^1 \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t-m)^2}{2\sigma^2}\right) p_i(t) dt,$$

where p_i is the i-th normalized Legendre polynomial, with $\int_0^1 p_i^2(t)dt = 1$. The integer θ is called the order of the decomposition. Then k_L operators on the vector $(a_0(\nu), \cdots, a_{\theta-1}(\nu)$ and is of the form

$$k_L(\nu_1, \nu_2) = \gamma^2 \exp\left[-\sum_{i=0}^{\theta-1} \frac{|a_i(\nu_1) - a_i(\nu_2)|}{l_i} \right]. \tag{20}$$

Thus the estimated regression function $\hat{f}$ in this case is calculated by following function

$$\hat{f}(\nu_{t,i}) = \gamma^2 \exp\left[-\sum_{i=0}^{\theta-1} \frac{|a_i(\nu_{t,i}) - a_i(\nu_j)|}{l_i} \right].$$

We just consider the orders of the decomposition 5 and 10. We fix $l_i = l$ for all $i = 1, \cdots, \theta - 1$, this estimated function depends also on three parameters $\gamma \neq 0,\ \lambda > 0$ and $l > 0$.

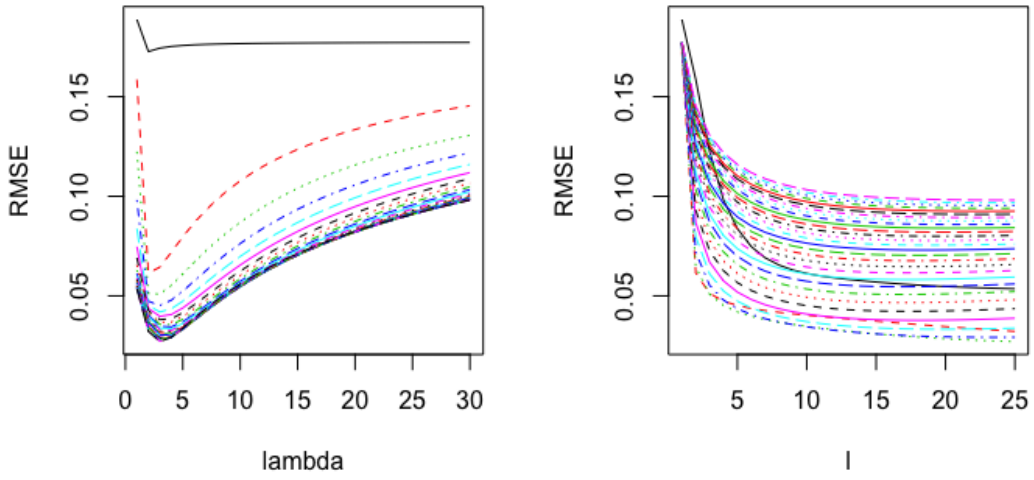

Figure 1: In the case of $n_t = 500$, fixing a value $\gamma = 1/2$, we run $\lambda > 0$ separated by 30 values from 0.005 to 30, $l > 0$ separated by 25 values from 0.005 to 20. We see the two graphs, one follows values of λ in the left side and l in the right side. RMSE will be minimized, in which it is lower than 0.08, with $0 < \lambda < 15$ and l big enough. We note that when $0 < l < 1$ RMSE is quite a big value for all $\lambda > 0$, so we avoid to chose these values of l.

4.1.3 Result

In simulation, we will see the effects of parameters $\lambda > 0$, $\gamma \neq 0$ and $l > 0$ on RMSE between predicted function $\hat{f}(\nu_{i,t})$ and exact function $f(\nu_{i,t})$ through the testing set $\{(\nu_{i,t})\}_{i=1}^{n_t}$. We also take two sizes of testing set $n_t = 500, n_t = 700$ to see the changes of RMSE. We just show the detailed presentation about choosing the optimal parameters on the "Wasserstein" model.

Case of testing set size $n_t = 500$: Now we consider RMSE in the case of $n_t = 500$ under the different fixed values $\gamma = 1/2, 1, 10$ and running $\lambda > 0$ separated by 30 values from 0.005 to 30, $l > 0$ separated by 25 values from 0.005 to 20. Let us see here the values of RMSE with the different cases of γ in following Figure 1, 2, 3.

We realize through three choices of $\gamma = 1/2, 1, 10$ that the values γ give the same impact of RMSE variations, but the smallest RMSE in the case $\gamma = 1$. In following this stimulation, we fix the value of $\gamma = 1$ and run the values of λ and l to see the changes of RMSE in the case of bigger size of testing set.

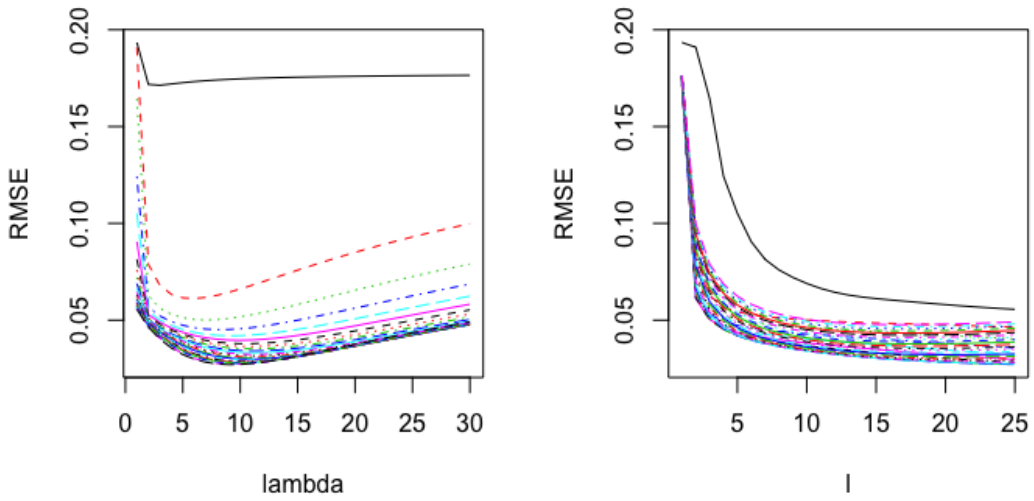

Figure 2: In the case of $n_t = 500$, fixing a value $\gamma = 1$, we run $\lambda > 0$ separated by 30 values from 0.005 to 30, $l > 0$ separated by 25 values from 0.005 to 20. We see the two graphs, one follows values of λ in the left side and l in the right side. The variations of RMSE in this case is not change significantly with the case of $\gamma = 1/2$, however, it looks smaller than the case $\gamma = 1/2$. We also see that RMSE will be minimized by two case: first $0 < \lambda < 1$ and l big enough; second $\lambda > 1$ for all $l > 1$ and RMSE is quite big at $0 < l < 1$ for all $\lambda > 0$.

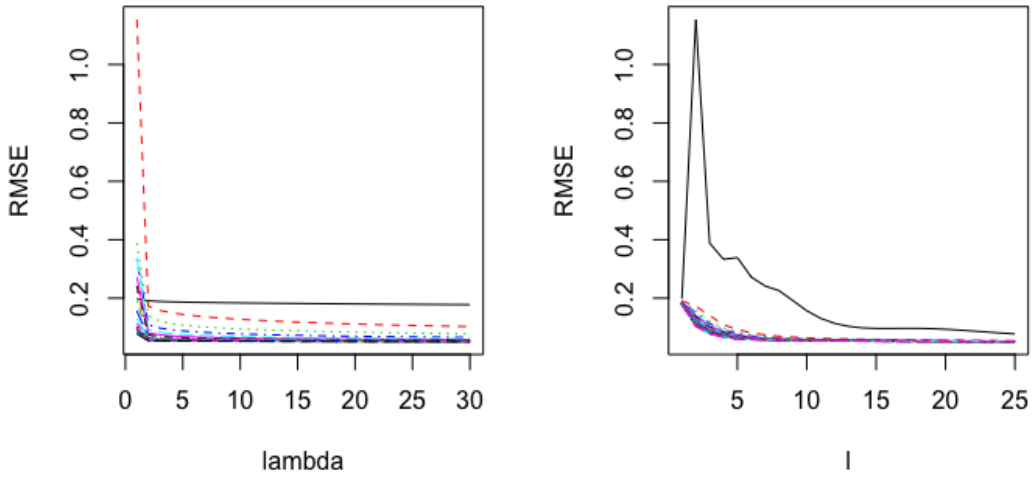

Figure 3: In the case of $n_t = 500$, fixing a value $\gamma = 10$, we run $\lambda > 0$ separated by 30 values from 0.005 to 30, $l > 0$ separated by 25 values from 0.005 to 20. RMSE in this case looks bigger than two above cases of $\gamma = 1/2$ and $\gamma = 1$.

Case of testing set size $n_t = 700$: Now we consider RMSE in the case of $n_t = 700$ in Figure 4 for a fixed value $\gamma = 1$ and running $\lambda > 0$ separated by 30 values from 0.005 to 30, $l > 0$ separated by 25 values from 0.005 to 20. We want to see the affects of testing set size on RMSE. Then we consider directly about the estimated regression function effects under parameters λ and l. As far as we known, there exists oversmoothing and undersmoothing issues which happen sometimes in the learning problem when the error component is small, but the estimated function is oversmooth or undesmooth. See the Figure 5, to more clearly about our regression model with the exact function defined in (17).

And finally, we consider the different RMSE's between "Wasserstein" and "Legendre" model by choosing the values $\gamma = 1$, $\lambda = 5$ and $l = 10$ under considering $n_t = 700$. In Table 1, we show the values of RMSE quality criteria for the "Wasserstein" and "Legendre" distribution regression models. From the values of the RMSE criterion, the "Wasserstein" model clearly outperforms the other models. The RMSE of the "Legendre" models slightly decreases when the order increases, and stay well above the RMSE of the "Wasserstein" model.

Hence from the Figure 5 and Table 1, we can see that by choosing the optimal parameters for γ, l and λ we can obtain a very well estimation function $\hat{f}$ without the under-smoothing and over-smoothing issues in the learning problem. Our regression model stay well above the RMSE criterion.

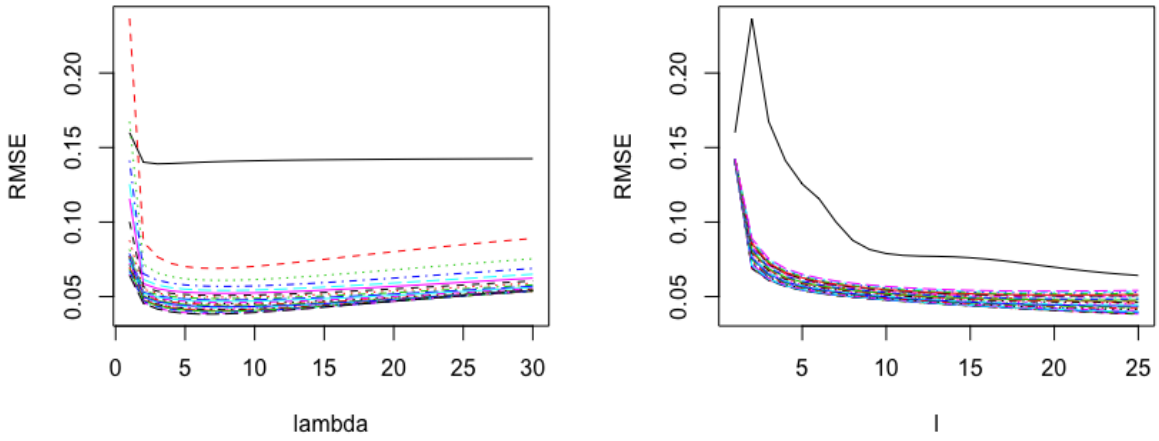

Figure 4: In the case of $n_t = 700$, fixing a value $\gamma = 1$, we run $\lambda > 0$ separated by 30 values from 0.005 to 30, $l > 0$ separated by 25 values from 0.005 to 20. RMSE is almost lower than 0.06 when $\lambda > 1$ and $l > 1$. However for $0 < \lambda < 1$, we can also obtain the small RMSE when l big enough. This figure provides a view about size of testing set, in which for the big enough of testing set size we will obtain the smaller RMSE under of the optimal parameters γ, λ and l.

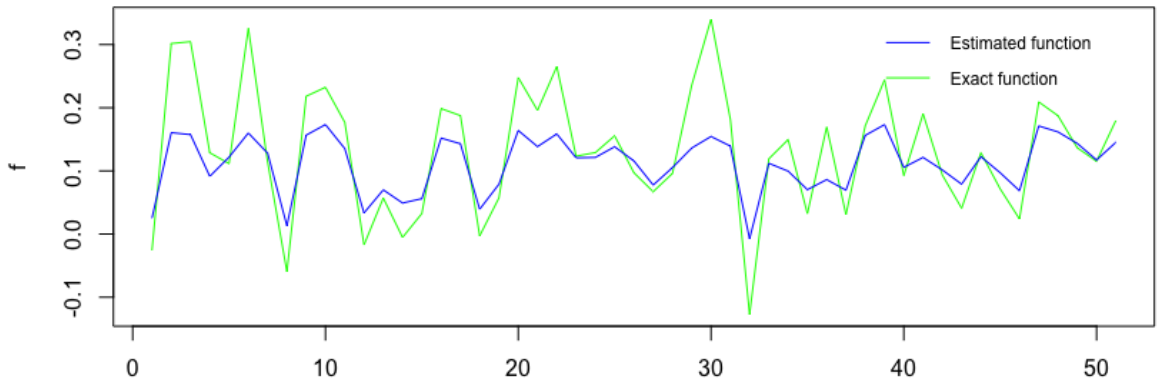

Figure 5: Regression function under exact and estimated function. The green line presents an exact function, which is many more variations, we desire find the optimal parameters to obtain a more smooth curve. From Figure 4, we can chose these parameters following RMSE, however, in some cases it happens over-smoothing and under-smoothing. See in this Figure, the blue line looks like have some desirable properties when we chose the big enough values of λ and l.

model	RMSE
"Wasserstein"	0.04
"Legendre" order 5	0.15
"Legendre" order 10	0.11

Table 1: RMSE values of different quality criteria for the "Wasserstein" and "Legendre" distribution regression models. The "Wasserstein" is based on universal kernel function operating directly on the input Gaussian distributions, while "Legendre" is based on linear projections of the Gaussian distribution inputs on finite-dimensional spaces. For "Legendre", the order value is the dimension of the projection space. The quality criteria is the root mean square error (RMSE) should be minimal. The "Wasserstein" distribution regression model clearly outperforms the "Legendre" with two orders 5 and 10.

Our interpretation for these results is that, because of the nature of the simulated data $(\mu_i, f(\mu_i))$ working directly on distributions and with the Wasserstein distance, is more appropriate than using linear projections. Indeed, in particular, two distributions with similar means and small variances are close to each other with respect to both Wasserstein distance and the value of the output function f. However,the probability density functions of the two distributions are very different from each other with respect to the L^2 distance in the case that the ratio between the two variances is large. Hence linear projections based on probability density functions is inappropriate in the setting considered here.

4.2 Application on evolution of hearing sensitivity

An otoacoustic emission (OAE) is a sound which is generated from within the inner ear. OAEs can be measured with a sensitive microphone in the ear canal and provide a noninvasive measure of cochlear amplification (see Chapter: Hearing basics in [21]). Recording of OAEs has become the main method for newborn and infant hearing screening (see Chapter: Early Diagnosis and Prevention of Hearing Loss in [21]). There are two types of OAEs: spontaneous otoacoustic emissions (SOAEs), which can occur without external stimulation, and evoked otoacoustic emissions (EOAEs), which require an evoking stimulus. In this paper, we consider a type of EOAEs that is Transient-EOAE (TEOAE) (see for instance in [22]), in which the evoked response from a click covers the frequency range up to around 4kHz. More precisely, each TEOAE models the ability of the cochlea to response to some frequencies in order to transform a sound into an information that will be processed by the brain. So to each observation is associated a curve (the Oto-Emission curve) which describes the response of the cochlea at several frequencies to a sound. The level of response depends on each individual and each stimulus should be normalized, but the way each individual reacts is characteristic of its physiological characteristic. Hence to each individual is associated a curve, which after normalization, it is considered as a distribution μ describing the repartition of the responses for different frequencies ranging from 0 to 10 kHz. These distributions are shown in Figure 6 and Table 2.

Name	**Age**	**0(Hz)**	**39.06**	**78. 12**	$\cdots$	**1171.88**	**1210.94**	$\cdots$	**9765.62**	**9804.69**
ABBAS	23	0	0.0006	0.0013	$\cdots$	0.0819	0.0388	$\cdots$	0.0021	0.0015
ADAMS	27	0.0001	0.0010	0.0022	$\cdots$	0.0283	0.0283	$\cdots$	0.0011	0.0006
ADENIYI	30	0.0002	0.0003	0.0014	$\cdots$	0.0231	0.0065	$\cdots$	0.0012	0.0016
DUPLOOY	17	0.0003	0.0005	0.0015	$\cdots$	0.0786	0.1272	$\cdots$	0.0036	0.0031
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
TRIMM	20	0.0005	0.0006	0.0026	$\cdots$	0.0133	0.0215	$\cdots$	0.0002	0.0017
VELE	26	0.0001	0.0005	0.0018	$\cdots$	0.1176	0.0859	$\cdots$	0.0003	0.0005
WALLER	40	0.0001	0.0001	0.0003	$\cdots$	0.0178	0.0210	$\cdots$	0.0014	0.0013
WILLIAM	22	0.0002	0.0003	0.0009	$\cdots$	0.0156	0.0656	$\cdots$	0.0014	0.0018

Table 2: TEOAE data. 48 individuals are considered in human population, recorded in South Africa, with last names in first column, their exact ages in second column and the others describe the responses of the cochlea at several frequencies ranging from 0Hz to 10kHz.

The relationship between age and hearing sensitivity is investigated in [23, 24] The results show that when age increases, the presence of EOAEs by age group and the frequency peak in spectral analysis decreases and EOAE threshold increases. The differences in EOAE have been also reported between age classes in humans. These results convey the idea that the response evolves with age and that the effect of ages in hearing issues is deeply related to the changes of the cochlear properties. Hence our model uses as input these distributions μ and try to build

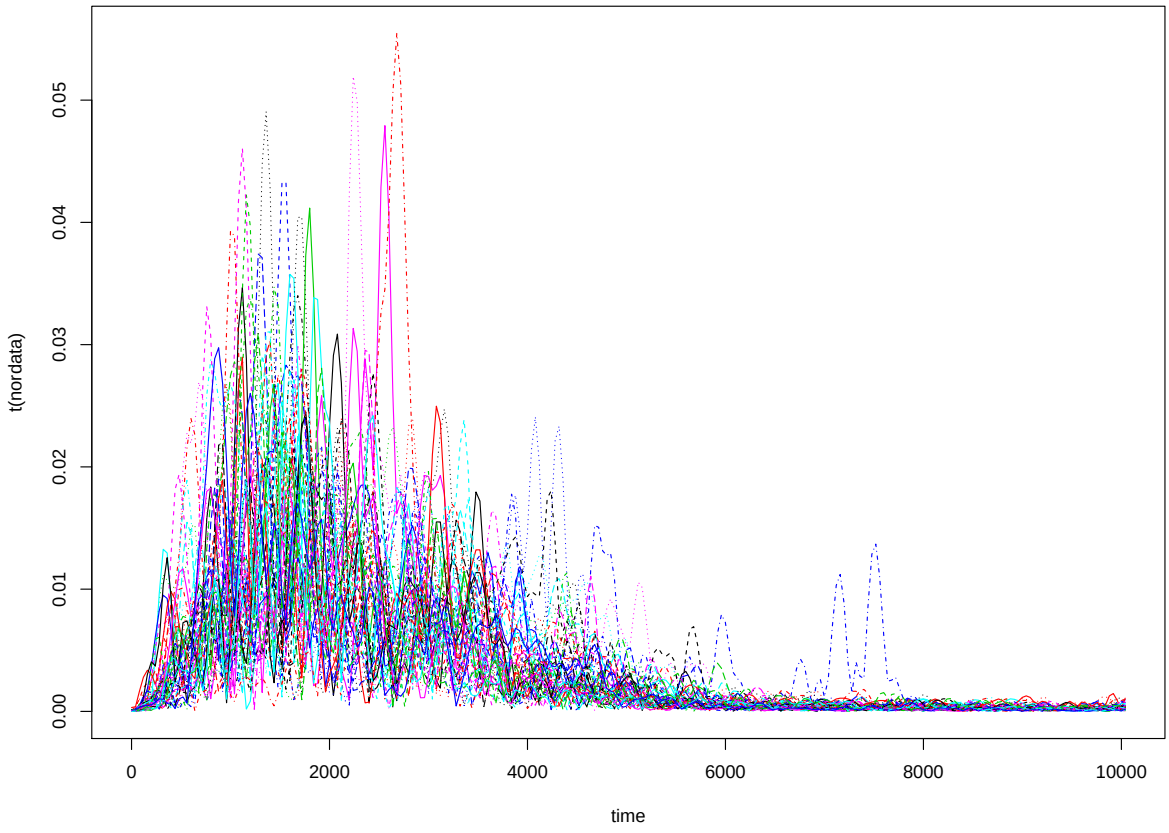

Figure 6: Oto-emission curves. 48 TEOAE curves following to frequencies ranging from 0Hz to 10kHz.

a regression model to link between the age and these distributions representing the response of the cochlea at frequencies ranging from 0Hz to 10kHz. More precisely, we estimate the age for each level of response normalized and treated as a distribution μ by using our proposed function as follows

$$\hat{f}(\mu_i) = \gamma^2 \sum_{j=1}^{n} \hat{\alpha}_j \exp\left(-\frac{\int_0^1 (F_{\mu_i}^{-1}(t) - F_{\mu_j}^{-1}(t))^2 dt}{l}\right), \tag{21}$$

where $F_{\mu}^{-1}(U)$ defined as (5) and the value of $\hat{\alpha}_j$ is chosen by optimal parameter λ in (16). We estimate the integral in (21) by following formula

$$\int_0^1 (F_{\mu_i}^{-1}(t) - F_{\mu_j}^{-1}(t))^2 dt = \sum_{m=1}^{M} \left[F_{\mu_i}^{-1}\left(\frac{m}{M}\right) - F_{\mu_j}^{-1}\left(\frac{m}{M}\right)\right]^2, \tag{22}$$

where we can understand each F_{μ_i} is an experimental distribution function of μ_i and M is the number of discretized frequencies. As far as we know, each individual is associated with a curve, which after normalization without lost relationship among original data, it is considered as a distribution μ_i. To calculate $F_{\mu_i}^{-1}\left(\frac{m}{M}\right)$, we arrange each curve in ascending order, for instance we denote $X_\mu(1) \leq X_\mu(2) \leq \cdots \leq X_\mu(M)$ following to distribution μ and $\sum_{m=1}^{M} X_\mu(m) = 1$, so $F_{\mu_i}^{-1}\left(\frac{m}{M}\right) = X_{\mu_i}(m)$. Hence, we write again the formula (22)

$$\int_0^1 (F_{\mu_i}^{-1}(t) - F_{\mu_j}^{-1}(t))^2 dt = \sum_{m=1}^{M} \left(X_{\mu_i}(m) - X_{\mu_j}(m)\right)^2, \tag{23}$$

where X_{μ_i} is a curve X following to distribution μ_i.
In our simulation, we choose $\gamma = 1$, the value of $l = 10$ and the value of $\lambda > 0$. We aim to study the age in relation with its TEOAE curve of 48 subjects, recorded on human population in South Africa, with the range of frequency from 0Hz to 10kHz. See the Figure 7 to show the differences between the age of 15 to 50 years old. Following the estimated function in (21), we take 47 distributions $\{\mu_j\}_{j=1}^{47}$ for training set to calculate estimation value of $\hat{\alpha}_j$ and try to estimate real age of a remaining individual μ_i with $i \neq j$. And the results are showed clearly in the Figure 8 and Figure 9 about the exact age and predicted age.

Hence in figure 7 and Figure 9, we applied effectively our proposed estimation function in predicting age from its TEOAE data. By choosing the optimal parameters γ, l and λ we could predict very well the exact ages belonging to the age class $[20, 30]$ and negligible errors in other age classes. This is quite reasonable when seeing in the Figure 7 that the age distributed diversity almost from 20 to 30 years old, so our proposed estimation function learnt very well to predict age in this age class. Thus by using the distribution regression model, we investigated the relationship between the evoked responses from clicks covering the frequencies range up to 10kHz and its evolutionary ages.

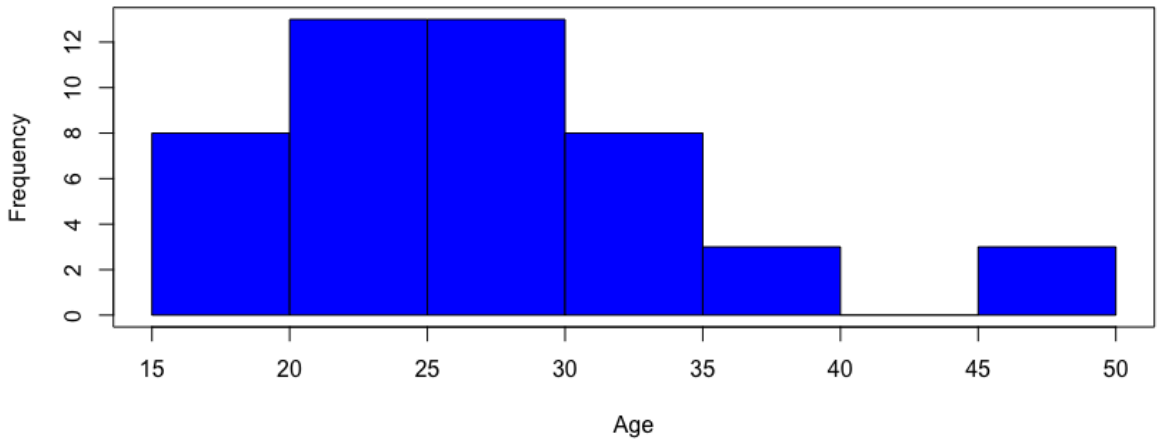

Figure 7: Histogram of real age in a human population. The age distribute diversity from 15 to 30, however, there is a few of individual of age from 35 to 40 and 45 to 50. And there exists no individual have age from 40 to 45.

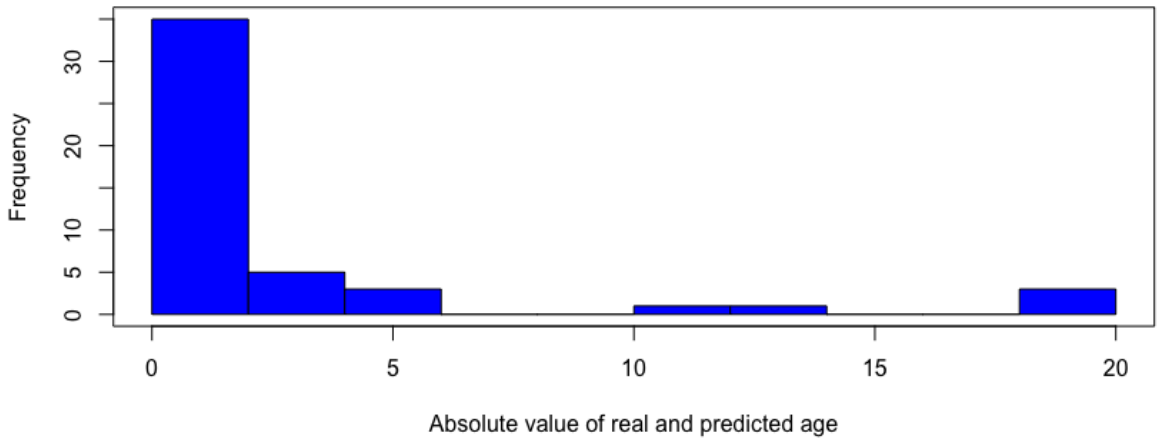

Figure 8: Histogram of difference between real and predicted age for OAE. In the first column, in which the difference between real and estimate age is very small closing to zero, this means more accuracy between real and predicted age. Almost the ages from 20 to 35 lie in this column.

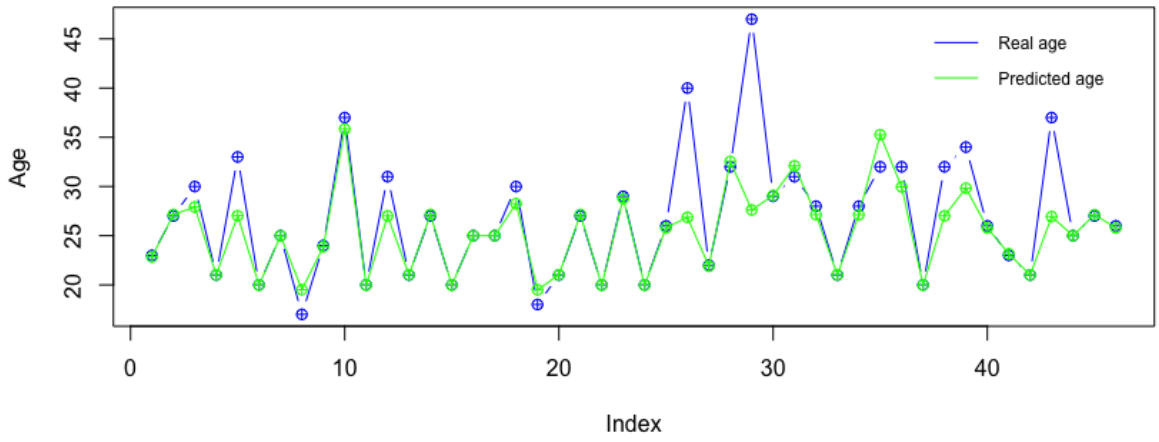

Figure 9: Real and Predicted Age. By using the optimized parameters of $\gamma = 1$, $l = 10$ and $\lambda > 0$ depending on age class, we obtain almost the exacted ages belonging to the age class $[20, 30]$ with λ around 15. For instance, we can predict very well the exact ages 20, 21, 23, 24, 25, 27, 29 corresponding to the predicted ages 19.50, 21.02, 22.83, 23.87, 24.89, 27.15, 28.76.

5 Discussion

In this paper, we have introduced a new estimated function for regression model with distribution inputs. More precisely, we effectively used class of positive definite kernel produced by Wasserstein distance, built in [12] by proving that it is a kind of universal kernel. Researching the universal kernel theories, we detected a very good property of our universal kernel to build a RKHS. Then we obtained a particular estimation from Representer theorem for our distribution regression problem, these works showed that the relation between the random distribution and the real number response can be learnt by using directly the regularized empirical risk over RKHS. Our proposed estimation is clearly better than state-of-the-art-ones in simulated data. More interestingly, we researched successfully TEOAE curve of each individual in human population as a distribution which after normalization. We then investigated the relationship between age and its TEOAE that the response involves with age and the effect of age in hearing issues is deeply related to the change of cochlear. This is a new interesting approach in the field of Biostatistics, in which we indicated the evolution of hearing capacity under statistical domain - distribution regression model. We believe that our paper tackles an important issue for data science experts willing to predict problems in regression with probability distributions as input. The extension of this work on distributions for general dimensions should be addressed in a further work, using for instance as a kernel the one buit in [25].

References

[1] J.-M. Azaïs, "Le modèle linéaire par l'exemple," 2006.

[2] M. H. Kutner, C. Nachtsheim, and J. Neter, *Applied linear regression models*. McGraw-Hill/Irwin, 2004.

[3] J. Neter, M. H. Kutner, C. J. Nachtsheim, and W. Wasserman, *Applied linear statistical models*, vol. 4. Irwin Chicago, 1996.

[4] J. O. Ramsay and B. W. Silverman, *Applied functional data analysis: methods and case studies*. Springer, 2007.

[5] C. Preda, "Regression models for functional data by reproducing kernel hilbert spaces methods," *Journal of statistical planning and inference*, vol. 137, no. 3, pp. 829–840, 2007.

[6] H. Kadri, E. Duflos, P. Preux, S. Canu, and M. Davy, "Nonlinear functional regression: a functional rkhs approach," in *Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS'10)*, vol. 9, pp. 374–380, 2010.

[7] A. Berlinet and C. Thomas-Agnan, *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

[8] A. Smola, A. Gretton, L. Song, and B. Schölkopf, "A hilbert space embedding for distributions," in *International Conference on Algorithmic Learning Theory*, pp. 13–31, Springer, 2007.

[9] C. Villani, *Optimal transport: old and new*, vol. 338. Springer Science & Business Media, 2008.

[10] S. Kolouri, Y. Zou, and G. K. Rohde, "Sliced Wasserstein kernels for probability distributions," *CoRR*, vol. abs/1511.03198, 2015.

[11] G. Peyré, M. Cuturi, and J. Solomon, "Gromov-Wasserstein averaging of kernel and distance matrices," in *ICML 2016*, 2016.

[12] F. Bachoc, F. Gamboa, J.-M. Loubes, and N. Venet, "A gaussian process regression model for distribution inputs," *IEEE Transactions on Information Theory*, 2017.

[13] A. Christmann and I. Steinwart, "Universal kernels on non-standard input spaces," in *Advances in neural information processing systems*, pp. 406–414, 2010.

[14] B. K. Sriperumbudur, K. Fukumizu, and G. R. Lanckriet, "Universality, characteristic kernels and rkhs embedding of measures," *Journal of Machine Learning Research*, vol. 12, no. Jul, pp. 2389–2410, 2011.

[15] C. A. Micchelli, Y. Xu, and H. Zhang, "Universal kernels," *Journal of Machine Learning Research*, vol. 7, no. Dec, pp. 2651–2667, 2006.

[16] W. Whitt, "Bivariate distributions with given marginals," *The Annals of statistics*, pp. 1280–1289, 1976.

[17] M. G. Cowling, "Harmonic analysis on semigroups," *Annals of Mathematics*, pp. 267–283, 1983.

[18] P. Embrechts and M. Hofert, "A note on generalized inverses," *Mathematical Methods of Operations Research*, vol. 77, no. 3, pp. 423–432, 2013.

[19] G. Kimeldorf and G. Wahba, "Some results on tchebycheffian spline functions," *Journal of mathematical analysis and applications*, vol. 33, no. 1, pp. 82–95, 1971.

[20] A. J. Smola and B. Schölkopf, *Learning with kernels*, vol. 4. Citeseer, 1998.

[21] J. J. Eggermont, *Hearing Loss: Causes, Prevention, and Treatment*. Academic Press, 2017.

[22] P. X. Joris, C. Bergevin, R. Kalluri, M. Mc Laughlin, P. Michelet, M. van der Heijden, and C. A. Shera, "Frequency selectivity in old-world monkeys corroborates sharp cochlear tuning in humans," *Proceedings of the National Academy of Sciences*, vol. 108, no. 42, pp. 17516–17520, 2011.

[23] T. O-Uchi, J. Kanzaki, Y. Satoh, S. Yoshihara, A. Ogata, Y. Inoue, and H. Mashino, "Age-related changes in evoked otoacoustic emission in normal-hearing ears," *Acta Oto-Laryngologica*, vol. 114, no. sup514, pp. 89–94, 1994.

[24] L. Collet, A. Moulin, M. Gartner, and A. Morgon, "Age-related changes in evoked otoacoustic emissions," *Annals of Otology, Rhinology & Laryngology*, vol. 99, no. 12, pp. 993–997, 1990.

[25] F. Bachoc, A. Suvorikova, J.-M. Loubes, and V. Spokoiny, "Gaussian process forecast with multidimensional distributional entries," *arXiv preprint arXiv:1805.00753*, 2018.