Nonlinear global Fréchet regression for random objects via weak conditional expectation

Satarupa Bhattacharjee, Bing Li, and Lingzhou Xue

Department of Statistics, The Pennsylvania State University
University Park, PA 16802, U.S.A.

Abstract

We introduce a nonlinear global regression model for object-valued predictor and response tuples. Random object data are complex non-Euclidean data taking value in general metric space, possibly devoid of any underlying vector space structure. Such data are getting increasingly abundant with the rapid advancement in technology. Examples include probability distributions, positive semi-definite matrices, and data on Riemannian manifolds. We propose the notion of a weak conditional Fréchet mean to aid the object regression framework. One of the main contributions is to establish a connection between the conditional Fréchet mean and the weak conditional Fréchet mean, the latter can being a generalization of the former. The motivation is based on Carleman operators and their inducing functions in the particular case of the classical Euclidean data. The state-of-the-art global Fréchet regression approach by Petersen and Müller (2019) emerges as a special case of the proposed model. We require that the metric space where the predictors reside admits a reproducing kernel Hilbert space embedding that is rich enough to characterize the joint probability distribution of the responses and the predictors, while the intrinsic geometry of the metric space where the responses lie is utilized to study the asymptotic convergence of the proposed estimates. Numerical studies, including both simulations and a data application, are conducted to investigate the performance of our estimator in a finite sample.

1 Introduction

Encountering complex non-Euclidean data-taking values in a general metric space that may defy any inherent linear structure has become increasingly common in the areas such as biological or social sciences with the rapid advancement of technology. Examples of such “*random object*” data, recorded in the form of images, shapes, networks, or life tables include distributional data in Wasserstein space (Delicado and Vieu, 2017; Le Gouic and Loubes, 2017), symmetric positive definite matrix objects (Dryden et al., 2009), data on the surface of the sphere (Di Marzio et al., 2014), phylogenetic trees (Billera et al., 2001) among others. Since the data are metric space valued, many classical notions of statistics, such as the definition of sample or population mean as an average or expected value, do not apply anymore and need to be replaced by barycenters or Fréchet means (Fréchet, 1948). In the regression context, the conditional Fréchet mean for random object responses Y, residing in a metric space (Ω_Y, d_Y), and Euclidean predictors $X \in \mathbb{R}^p$, as (Hein, 2009; Petersen and Müller, 2019)

$$E_\oplus(Y|X=x) = m_\oplus(x) := \text{argmin}_{y\in\Omega_Y} E[d_Y^2(Y,y)|X=x]. \tag{1}$$

The Fréchet regression proposed by Petersen and Müller (2019) generalizes the globally linear least squares method and the nonparametric local linear regression to fit the conditional Fréchet mean. The globally linear approach, in particular, targets an alternative formulation than (1) given by

$$\tilde{m}_\oplus(x) = \text{argmin}_{y\in\Omega_Y} E[s(X,x) d_Y^2(Y,y)], \tag{2}$$

where the weight function $s(X,x) = 1 + (x-\mu_X)^\intercal \Sigma_X^{-1}(X-\mu_X)$ varies globally and linearly with the output points $x \in \mathbb{R}^p$, hence the nomenclature; μ_X and Σ_X being the expectation and covariance matrix for the predictors X.

Model (2) coincides with model (1) in the special case of multiple linear regression with Euclidean responses and predictors. However, for a general metric space-valued response $Y \in \Omega_Y$, the above two targets are different, thus making the regression relationship for general metric-valued data quite restrictive. Although the local regression, which indeed targets (1) with an asymptotically negligible bias, is more flexible, it is effective only when the dimension of the predictor is relatively low. As this dimension gets higher, its accuracy drops significantly- a phenomenon known as the curse of dimensionality. Recently Bhattacharjee and Müller (2021) developed a single index Fréchet regression that projects the multivariate predictors onto a desired direction parameter vector to form a single index, thus facilitating inference for Fréchet regression. However, the model assumptions are still somewhat restrictive, and in general, the Fréchet regression framework only can accommodate Euclidean predictors.

In this work, we propose a non-linear global object regression framework that can accommodate both responses and predictors residing in arbitrary metric spaces. Our main two contributions are listed as follows.

Firstly, as discussed before, the conditional Fréchet mean in (1) might be a significantly different target from the global Fréchet mean (2) proposed by Petersen and Müller (2019), owing to the lack of linearity in a general abstract metric space. Hence the interpretation or validity of such a "globally linear" model can be brought into question. We propose a significant step up in bridging the discrepancy between two targets and extending the global linear regression to a more general globally non-linear object regression.

In order to answer this question, one first needs to ponder what a polynomial regression model even looks like in a metric space. A convenient vehicle to link random object data analysis to non-linear global RKHS (Reproducing Kernel Hilbert Space) regression models, beyond linear or polynomial regression to an arbitrary non-linear function, is achieved through weak conditional moments on $d_Y^2(Y, y)$. Li and Song (2022) first introduced this new statistical construct as a generalization of conditional expectation based on Carleman operators and their induced functions. The key idea of this approach for classical Hilbertian data is that it replaces the L^2 space for the projection that characterizes the conditional expectation by an arbitrary Hilbert space, while still maintaining the unbiasedness of the regression estimate. This concept of the weak conditional mean is further developed and some key properties of the construct proven to facilitate a better understanding. Furthermore, the weak conditional mean for Euclidean data is extended to *weak conditional Fréchet means* to define a random object regression model via some kernelized version of the predictors. The global linear regression model by Petersen and Müller (2019) emerges as a special case of a linear kernel. In fact, we discuss four types of conditional means- the conditional expectation and its weak version for both Euclidean and object-valued data in explicit detail and establish connection among the four notions (see Figure 1).

Secondly, beyond scalar-or-vector-valued predictors, studying the relation between two arbitrary random objects is also increasingly important. Unfortunately, not much exists in the state-of-the-art literature in this context, barring special cases of distribution-on-distribution regression (Chen et al., 2019, 2021; Ghodrati and Panaretos, 2022). Our proposed method accommodates more general predictors such as random vectors, functions, or even object-valued predictors, as long as the predictor space is rich enough to admit an RKHS embedding. We discuss the details of constructing appropriate kernels to generate such RKHSs and study the relevant operators generated to achieve this goal.

The rest of the paper is organized as follows. Section 2 defines the preliminary setup of the problem and focuses on the construction of the weak conditional mean for the classical/ Euclidean paradigm in detail. It is important to note that Section 2 by itself is a key contribution in the state-of-the-art literature for the Hilbert space valued functional data. In Section 3 we define the weak condition moments for object responses and predictors, establish the global non-linear object regression model and study its connections to the global linear object regression framework. In Section 4, we propose a suitable estimator for the weak conditional Fréchet mean from the observed data. In this vein, the construction of the underlying RKHS is discussed and an M-estimation setting is devised. Section 5 establishes the asymptotic convergence rates of the proposed methods. Simulation results are presented in Section 6 to show the numerical performances of the proposed methods. In Section 7, we analyze a real application of the proposed method for the mortality-vs-fertility distributions. All proofs are presented in the Supplementary Material.

2 Weak conditional mean and further development

In this section we first introduce the notations with a focus on the construction of a reproducing kernel Hilbert space on the space where the predictor objects lie. Next, we outline the basic idea underlying the construction of the weak conditional expectation in Li and Song (2022). We will also derive some new properties of weak conditional expectation, and give a more general the theory about the weak conditional expectation that is needed in later development.

2.1 Random objects and reproducing kernels

Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let (Ω_X, d_X) and (Ω_Y, d_Y) be metric spaces, where Ω_X and Ω_Y are sets and d_X and d_Y are the metrics. Let $\mathcal{F}_X$ and $\mathcal{F}_Y$ be the Borel σ-fields in Ω_X and Ω_Y corresponding to the open sets determined by d_X and d_Y. Let $X : \Omega \to \Omega_X$ and $Y : \Omega \to \Omega_Y$ be random elements that are measurable, respectively, with respect to $\mathcal{F}/\mathcal{F}_X$ and $\mathcal{F}/\mathcal{F}_Y$. Such random elements are called *statistical objects.* Let

$$P_{XY} = P \circ (X, Y)^{-1}, \quad P_X = P \circ X^{-1}, \quad P_Y = P \circ Y^{-1}$$

be the distributions of (X, Y), X, Y.

We will assume that there exists a positive definite kernel $\kappa_X : \Omega_X \times \Omega_X \to \mathbb{R}$. While there are sufficient conditions for a metric space to possess such kernels, we make this requirement our general assumption.

Assumption 1 *There is a positive definite kernel $\kappa_X : \Omega_X \times \Omega_X \to \mathbb{R}$.*

For example, if Ω_X is of negative type, then the metric-induced kernel is positive definite (Sejdinovic, Sriperumbudur, Gretton and Fukumizu, 2013). Furthermore, Zhang, Xue, and Li (2021) showed that, if Ω_X is complete and separable, and there is a continuous injection from $\rho : \Omega_X \to \mathcal{H}$ for some separable Hilbert space $\mathcal{H}$, then, for any analytic function $F(t) = \sum_{i=1}^{\infty} a_i t^i$ with $a_i > 0$, the function $\kappa : \Omega_X \times \Omega_X \to \mathbb{R}$ of the form $F(\langle \rho(x_1), \rho(x_2) \rangle_{\mathcal{H}})$ is a cc-universal kernel (Michelli et al 2006). Let $\kappa_G(x, x') = \exp(-\gamma_X d_X^2(x, x'))$ and $\kappa_L(x, x') = \exp(-\gamma_X d_X^2(x, x'))$ denote the Gaussian and Laplacian kernels, respectively. Zhang et al. (2021) showed that both κ_G and κ_L on a complete and separable metric space Ω_X are positive definite and universal, and the RKHS $\mathcal{H}_X$ generated by such kernels is dense in $L^2(P_X)$.

Note that we do not impose the above assumption on Ω_Y.

2.2 Weak conditional mean via uncentered regression operator

We first define the extended Carlman operator, which is a slight extension of the definition in Weidmann (2012).

Definition 1 (Carleman operator) *Let $\mathcal{G}$ be a set, $\mathcal{M}$ a Hilbert space of real-valued functions on $\mathcal{G}$, $\mathcal{H}$ another Hilbert space, and $A : \mathcal{H} \to \mathcal{M}$ a linear operator. If, for each $x \in \mathcal{G}$, the linear functional*

$$A_x : \mathcal{H} \to \mathbb{R}, \ f \mapsto (Af)(x)$$

is bounded, then we call A an extended Carleman operator. The Riesz representation $\lambda_A(x)$ of A_x is called the inducing function of A.

In the rest of the paper, $\mathcal{G}$ is the metric space Ω_X, $\mathcal{M}_X$ is the RKHS generated by κ_X, $\mathcal{H}$ is the real line $\mathbb{R}$, and $A : \mathbb{R} \to \mathcal{M}_X$ is the regression operator.

We next introduce the regression operator. Let $\mathcal{H}_U$ be a generic Hilbert space, and let $U : \Omega \to \mathcal{H}_U$ be a random element. We make the following assumption.

Assumption 2 *$\mathcal{M}_X$ and $\mathcal{H}_U$ are separable.*

These conditions are mild: for example, by Hsing and Eubank (2015), Theorem 2.7.5, if Ω_X is separable and κ_X is continuous, then $\mathcal{M}_X$ is separable. Since $\mathcal{H}_U$ will be taken to be $\mathbb{R}$ for the rest of the paper, it is separable. Consider the tensor products

$$\kappa_X(\cdot, X) \otimes \kappa_X(\cdot, X), \quad \kappa_X(\cdot, X) \otimes U.$$

The above quantities are members of the tensor product spaces $\mathcal{M}_X \otimes \mathcal{M}_X$ and $\mathcal{M}_X \otimes \mathcal{H}_U$, respectively. By simple calculation,

$$\begin{aligned} \|\kappa_X(\cdot, X) \otimes \kappa_X(\cdot, X)\|_{\mathcal{M}_X \otimes \mathcal{M}_X} &= \kappa_X(X, X), \\ \|\kappa_X(\cdot, X) \otimes U\|_{\mathcal{M}_X \otimes \mathcal{H}_U} &= \sqrt{\kappa_X(X, X)}\|U\|. \end{aligned} \tag{3}$$

We make the following assumption.

Assumption 3 *$E\kappa_X(X, X) < \infty$, $E(\sqrt{\kappa_X(X, X)}\|U\|) < \infty$.*

Since $\mathcal{M}_X$ and $\mathcal{H}_U$ are separable, $\mathcal{M}_X \otimes \mathcal{M}_X$ and $\mathcal{M}_X \otimes \mathcal{H}_U$ are separable. Furthermore, by Assumption 3 and relations in (3), we have

$$E(\|\kappa_X(\cdot, X) \otimes \kappa_X(\cdot, X)\|_{\mathcal{M}_X \otimes \mathcal{M}_X}) < \infty, \quad E(\|\kappa_X(\cdot, X) \otimes U\|_{\mathcal{M}_X \otimes \mathcal{H}_U}) < \infty.$$

By Theorem 2.6.5 of Hsing and Eubank (2015), the following Bochner integrals

$$\int_\Omega \kappa_X(\cdot, X) \otimes \kappa_X(\cdot, X) dP, \quad \int_\Omega \kappa_X(\cdot, X) \otimes U dP$$

are defined. They will be denoted by M_{XX} and M_{XU}, respectively, and will be called the covariance operator of X and the cross-covariance operator from $\mathcal{H}_U$ to $\mathcal{M}_X$. It can be shown that, for any $f, g \in \mathcal{M}_X$ and $h \in \mathcal{H}_U$, we have

$$\langle f, M_{XX} \rangle_{\mathcal{M}_X} = E[f(X)g(X)], \quad \langle f, M_{XU} h \rangle_{\mathcal{M}_X} = E[f(X)\langle U, h \rangle_{\mathcal{H}_U}]. \tag{4}$$

Henceforth, for a linear operator $A : \mathcal{H} \to \mathcal{M}$, we let $\mathrm{ran}(A)$ denote the range of A and $\ker(A)$ denote the kernel of A; that is, $\mathrm{ran}(A) = \{Af : f \in \mathcal{H}\}$ and $\ker(A) = \{f \in \mathcal{H}, Af = 0\}$. Furthermore, let $\overline{\mathrm{ran}}(A)$ denote the closure of $\mathrm{ran}(A)$. We make the following assumption.

Assumption 4 $\ker(M_{XX}) = \{0\}$ *and* $\operatorname{ran}(M_{XU}) \subseteq \operatorname{ran}(M_{XX})$.

This assumption is very mild. By (4), $M_{XX}f = 0$ implies $E[f^2(X)] = 0$, which implies that $f(X) = 0$ almost surely. If κ_X is continuous, then $f(X) = 0$ everywhere. Hence, if κ_X is continuous, then $\ker(M_{XX}) = \{0\}$. As argued in Li (2018), the assumption $\operatorname{ran}(M_{XU}) \subseteq \operatorname{ran}(M_{XX})$ is a smoothness assumption about the relation between U and X. Under $\ker(M_{XX}) = \{0\}$, $M_{XX} : \mathcal{M}_X \to \operatorname{ran}(M_{XX})$ is an injective function. Thus the inverse function $M_{XX}^{-1} : \operatorname{ran}(M_{XX}) \to \mathcal{M}_X$ is defined. By $\operatorname{ran}(M_{XU}) \subseteq \operatorname{ran}(M_{XX})$, the operator

$$R_{XU} = M_{XX}^{-1} M_{XU}$$

is well defined, and is called the regression operator (Lee, Li, and Zhao, 2016). Note, however, since M_{XX} is a trace class operator, M_{XX}^{-1} is an unbounded operator. Nevertheless, as argued in Li (2018), it is entirely reasonable to assume R_{XU} to be a bounded or even compact operator, which imposes again a type of smoothness on the relation between U and X.

Assumption 5 $R_{XU} : \mathcal{H}_U \to \mathcal{M}_X$ *is a bounded operator.*

As we show below, this assumption actually implies that R_{XU} is an extended Carleman operator.

Proposition 1 *If R_{XU} is a bounded operator, then it is an extended Carleman operator.*

The next theorem is the key property of the regression operator. Since it is more general that those given in Lee, Li, and Zhao (2016) and Li and Song (2022), we provide a proof here.

Theorem 1 *If Assumptions 1 through 5 are satisfied and, for any $\alpha \in \mathcal{H}_U$, $E(\langle \alpha, Y \rangle_{\mathcal{H}_U} | X)$ is in the $L_2(P_X)$-closure of $\mathcal{M}_X$, then*

1. $E(\langle \alpha, Y \rangle_{\mathcal{H}_U} | X) \in \operatorname{ran}(R_{XU})$ *almost surely;*
2. *for any $\alpha \in \mathcal{H}_U$, $R_{XU}(\alpha)(X) = E[\langle \alpha, U \rangle_{\mathcal{H}_U} | X]$ almost surely.*

As a special case, when $\mathcal{M}_X$ is dense in $L_2(P_X)$, the conclusion of the theorem holds, because in that case $E[\langle \alpha, U \rangle_{\mathcal{H}_U} | X]$ is always in the $L_2(P_X)$-closure of $\mathcal{M}_X$. This was the result proved in Li and Song (2022). The weak conditional mean is defined as the inducing function of the linear operator R_{XU}.

Definition 2 *If Assumptions 1 through 5 are satisfied, then the random element*

$$\omega \mapsto \lambda_{R_{XU}}(X(\omega)), \quad \Omega \to \mathcal{H}_U$$

is the weak conditional expectation of U given X; that is $\lambda_{R_{XU}}(X) = E(U \,\vdots\, X)$.

It follows easily from Theorem 1 that the weak conditional expectation reduces to the true conditional expectation under assumptions therein.

Corollary 1 *Under the assumptions in Theorem 1, we have*

$$E(U \,\vdots\, X) = E(U|X).$$

2.3 Weak conditional mean via centered regression operator

An alternative definition of the regression operator, as given in Lee et al. (2016), is the centered version of R_{XU}. Let

$$\Sigma_{XX} = E[(\kappa_X(\cdot, x) - \mu_X) \otimes (\kappa_X(\cdot, x) - \mu_X)], \quad \Sigma_{XU} = E[(\kappa_X(\cdot, x) - \mu_X) \otimes (U - \mu_U)].$$

These operators are defined under Assumption 3. We make similar range assumption as Assumption 4.

Assumption 6 $\operatorname{ran}(\Sigma_{XU}) \subseteq \operatorname{ran}(\Sigma_{XX})$.

In general, $\ker(\Sigma_{XX}) \neq \{0\}$, and so function $\Sigma_{XX} : \mathcal{M}_X \to \mathcal{M}_X$ is not invertible. However, the restricted operator $\Sigma_{XX}|_{\overline{\operatorname{ran}}(\Sigma_{XX})}$ is an invertible function. We call its inverse $[\Sigma_{XX}|_{\overline{\operatorname{ran}}(\Sigma_{XX})}]^{-1}$ the Moore-Penrose inverse, and denote it by $\Sigma_{XX}^{\dagger}$. Note that this is a mapping from $\operatorname{ran}(\Sigma_{XX})$ to $\overline{\operatorname{ran}}(\Sigma_{XX})$. Under Assumption 6, the operator

$$R_{XU}^{(c)} := \Sigma_{XX}^{\dagger}\Sigma_{XU}$$

is well defined, and, to distinguish it from R_{XU} above, we denote it by $R_{XU}^{(c)}$ and call it the centered regression operator.

Assumption 7 $R_{XU}^{(c)}$ *is a bounded operator.*

Using $R_{XU}^{(c)}$, we now give the alternative definition of the weak conditional expectation. It turns out that this alternative definition deal with the constant function better $f(x) = \text{constant}$ better than the uncentered version.

Definition 3 *Suppose $R^{(c)}_{XU}$ is defined and is a Carleman operator. Then the following random element*

$$E(U) + \lambda_{R^{(c)}_{XU}}(X) - E[\lambda_{R^{(c)}_{XU}}(X)]$$

is called the weak conditional expectation of U given X with respect to $\mathcal{M}_X$.

The next proposition is a parallel result of Theorem 1 for the centered regression operator. We will say that a function f belongs to a subset of $L_2(P_X)$ modulo constant if there is a constant c such that $f + c$ belongs to that subset.

Proposition 2 *If Assumptions 1, 2, 3, 6, and 7 are satisfied and, for any $\alpha \in \mathcal{H}_U$, $E(\langle \alpha, Y\rangle_{\mathcal{H}_U}|X)$ belongs to the $L_2(P_X)$-closure of $\mathcal{M}_X$ modulo constant, then*

1. *$E(\langle \alpha, Y\rangle_{\mathcal{H}_U}|X) \in \mathrm{ran}(R_{XU})$ modulo constant almost surely;*

2. *for any $\alpha \in \mathcal{H}_U$,*

$$E[\langle \alpha, U\rangle_{\mathcal{H}_U}|X] = \langle \alpha, E(U)\rangle_{\mathcal{H}_U} + R^{(c)}_{XU}(\alpha)(X) - E[R^{(c)}_{XU}(\alpha)(X)]. \qquad (5)$$

The proof is similar to that of Theorem 1 and is omitted. The advantage of Definition 3 over Definition 2 is that the former does not require the function $x \mapsto 1$ to be a member of $\mathcal{M}_X$, while the latter usually does, as shown in the next corollary. In the following, $\mathbb{1}_X : \Omega_X \to \mathbb{R}$ stands for the function $x \mapsto 1$.

Corollary 2 *Suppose*

1. *both R_{XU} and $R^{(c)}_{XU}$ are defined and bounded;*

2. *for any $\alpha \in \mathcal{H}_U$, $E(\langle U, \alpha\rangle_{\mathcal{H}_U}|X)$ is in the $L_2(P_X)$-closure of $\mathcal{M}_X$;*

3. *$E(U) - E[\lambda_{R^{(c)}_{XU}}(X)] \neq 0$.*

Then $\mathbb{1}_X$ belongs $\mathcal{M}_X$ almost surely.

The next simple example illustrates the advantage of $\mu_U + \lambda_{R^{(c)}_{XU}}(X) - E[\lambda_{R^{(c)}_{XU}}(X)]$ over $\lambda_{R_{XU}}(X)$ as the definition of weak conditional expectation.

Example 1 Suppose U and X are random vectors in $\mathbb{R}^q$ and $\mathbb{R}^p$, respectively. Assume that

$$E(U|X) = a + B^\intercal X.$$

where a is a nonzero vector in $\mathbb{R}^p$, and B is a matrix in $\mathbb{R}^{p\times q}$. Under this model, it can be easily shown that

$$E(U|X) = E(U) + [\mathrm{cov}(U,X)][\mathrm{var}(X)]^{-1}(X - E(X)). \tag{6}$$

Let $\mathcal{H}_U$ be the Euclidean space $\mathbb{R}^q$ and $\mathcal{M}_X$ is the Hilbert space consisting of functions of the form $\{a^\intercal x : a \in \mathbb{R}^p\}$ with inner product defined by

$$\langle a_1^\intercal(\cdot), a_2^\intercal(\cdot)\rangle_{\mathcal{M}_X} = a_1^\intercal a_2.$$

The space $\mathcal{M}_X$ can be viewed as an RKHS with kernel $\kappa_X(a_1^\intercal(\cdot), a_2^\intercal(\cdot)) = a_1^\intercal a_2$. In this case

$$M_{XX} = E[((\cdot)^\intercal X) \otimes ((\cdot)^\intercal X)], \quad M_{XU} = E[((\cdot)^\intercal X) \otimes U].$$

The space $\mathcal{M}_X$ is isomorphic to $\mathbb{R}^p$ with the isomorphism $T : \mathcal{M}_X \to \mathbb{R}^p,\ a^\intercal(\cdot) \mapsto a$. Furthermore, it can be easily shown that $TM_{XX}T^* = E(XX^\intercal)$ and $TM_{XU} = E(XU^\intercal)$. Hence

$$\begin{aligned}
R_{XU}(\alpha)(X) &= \langle R_{XU}(\alpha), (\cdot)^\intercal X\rangle_{\mathcal{M}_X} \\
&= (TR_{XU}(\alpha))^\intercal (T((\cdot)^\intercal X)) \\
&= (TR_{XU}(\alpha))^\intercal X \\
&= (TM_{XX}^{-1}T^*TM_{XU}\alpha)^\intercal X \\
&= \alpha^\intercal[(E(XX^\intercal))^{-1}E(XU^\intercal)]^\intercal X,
\end{aligned}$$

which implies $\lambda_{R_{XU}} = [(E(XX^\intercal))^{-1}E(XU^\intercal)]^\intercal X$. Clearly, this is not the same as the right-hand side of (6).

Next, let's consider the centered version. Similar to the above argument, we can show that

$$R_{XU}^{(c)}(\alpha)(X) = \alpha^\intercal[(\mathrm{var}(X))^{-1}\mathrm{cov}(X,U)]^\intercal X,$$

implying $\lambda_{R_{XU}^{(c)}}(X) = [(\mathrm{var}(X))^{-1}\mathrm{cov}(X,U)]^\intercal X$. Hence

$$E(U) + \lambda_{R_{XU}^{(c)}}(X) - E[\lambda_{R_{XU}^{(c)}}(X)] = E(U) + [(\mathrm{var}(X))^{-1}\mathrm{cov}(X,U)]^\intercal(X - EX),$$

which is exactly the right-hand side of (6). □

This example shows that, when $\mathcal{M}_X$ does not contain $\mathbb{1}_X$, $\lambda_{R_{XY}}(X)$ is not the right generalization of $E(U|X)$. In comparison, $E(U) + \lambda_{R^{(c)}_{XU}}(X) - E[\lambda_{R^{(c)}_{XU}}(X)]$ gives the right generalization without requiring $\mathcal{M}_X$ to contain $\mathbb{1}_X$. The next theorem shows that, when $\mathcal{M}_X$ does contain the $\mathbb{1}_X$, the two definitions are equivalent.

Theorem 2 *If R_{XU} and $R^{(c)}_{XU}$ are defined and bounded, and $\mathcal{M}_X$ contains $\mathbb{1}_X$, then*

$$\lambda_{R_{XU}}(X) = E(U) + \lambda_{R^{(c)}_{XU}}(X) - E[\lambda_{R^{(c)}_{XU}}(X)]$$

almost surely.

Throughout the rest of the paper, we will adopt Definition 3 as our definition of the weak conditional expectation, and will denote it by $E(U \,¦\, X)$.

3 Weak conditional Fréchet mean

3.1 Weak conditional Fréchet mean and its properties

Having defined the weak conditional expectation of $E(U \,¦\, X)$, we now define the weak conditional Fréchet mean of a random object Y in the metric space (Ω_Y, d_Y). For any fixed $y \in \Omega_Y$, let $U = d^2(y, Y)$ and $\mathcal{H}_U = \mathbb{R}$. Assuming (X, U) satisfies Assumptions Assumptions 1, 2, 3, 6, and 7, the weak conditional mean $E[d^2(y, Y) \,¦\, X]$ is well defined.

Definition 4 *Suppose X and $U = d^2(y, Y)$ satisfy Assumptions 1, 2, 3, 6, and 7. The weak conditional Fréchet mean of Y given X, denoted by $E_\oplus(Y \,¦\, X = x)$, is the minimizer of $E[d^2(y, Y) \,¦\, X = x]$. That is,*

$$E_\oplus(Y \,¦\, X = x) = \mathrm{argmin}_{y \in \Omega_Y} E[d_Y^2(Y, y) \,¦\, X = x].$$

We use $E_\oplus(Y \,¦\, X)$ to denote the function $x \mapsto E_\oplus(Y \,¦\, X = x)$.

In plain language, the weak conditional Fréchet mean is any minimizer (over $y \in \Omega_Y$) of the weak conditional mean of $d^2(y, Y)$ given X. The next proposition gives an explicit expression of $E(U \,¦\, X)$ when U when U is a random scalar.

Corollary 3 *Suppose $\mathcal{H}_U = \mathbb{R}$ and (X, U) satisfies Assumptions 1, 2, 3, 6, and 7. Then*

$$E(U \,¦\, X) = E(U) + \langle \kappa_X(\cdot, X) - \mu_X, \Sigma^\dagger_{XX} E[(\kappa_X(\cdot, X) - \mu_X)U] \rangle_{\mathcal{M}_X}. \qquad (7)$$

where $(\kappa_X(\cdot, x) - \mu_X)U$ denotes the function $x \mapsto (\kappa_X(\cdot, x) - \mu_X)U$.

By this corollary, the weak condition Fréchet mean can be written more explicitly as

$$\begin{aligned} f_\oplus(x) := E_\oplus(Y \mathrel{\vdots} X = x) = \operatorname{argmin}_{y\in\Omega_Y} \big[E(d^2(Y,y)) + \\ \langle \kappa_X(\cdot, X) - \mu_X, \Sigma_{XX}^\dagger E[(\kappa_X(\cdot, x) - \mu_X) d^2(Y,y)] \rangle_{\mathcal{M}_X} \big]. \qquad (8) \end{aligned}$$

Denoting $d_Y^2(Y,y)$ as $U(y)$, and the operator $E[(\kappa_X(\cdot, X) - \mu_X)d^2(Y,y)]$ as $\Sigma_{XU(y)}$ one can rewrite (8) as

$$f_\oplus(x) = E_\oplus(Y \mathrel{\vdots} X) = \operatorname{argmin}_{y\in\Omega_Y} \left[E(U(y)) + \langle \kappa_X(\cdot, X) - \mu_X, \Sigma_{XX}^\dagger \Sigma_{XU(y)} \rangle_{\mathcal{M}_X} \right]. \qquad (9)$$

We take $E_\oplus(Y \mathrel{\vdots} X)$ as our population target for estimation in nonlinear global Fréchet regression, which offers great flexibility. First, when we employ a universal kernel such as the Gaussian kernel of the Laplacian kernel, we are guaranteed to recover the conditional Fréchet mean. Indeed, by Proposition 2, we have the following corollary.

Corollary 4 *Suppose X and $U = d_Y(Y,y)^2$ satisfy Assumptions 1, 2, 3, 6, and 7. If $\mathcal{M}_X$ is dense in $L_2(P_X)$ modulo constant, then*

$$E_\oplus(Y|X) = E_\oplus(Y \mathrel{\vdots} X).$$

Secondly, even when $\mathcal{M}_X$ is not dense in $L_2(P_X)$ modulo constant, it still makes sense to use $E_\oplus(Y \mathrel{\vdots} X)$, because it has the following optimality property. Let $\mathcal{N}_X$ denote the $L_2(P_X)$-closure of $\mathcal{M}_X + \operatorname{span}(\mathbb{1}_X)$. That is, a member of $\mathcal{N}_X$ can be written as the limit of functions of the form $f_n + c_n$, where $f_n \in \mathcal{M}_X$ and c_n is a constant.

Theorem 3 *If $R_{XU}^{(c)}$ is defined and bounded, then, for any $f \in \mathcal{N}_X$,*

$$E\{[E(U|X) - E(U \mathrel{\vdots} X)]^2\} \le E\{[E(U|X) - f(X)]^2\}.$$

This theorem shows that, even when $E_\oplus(Y \mathrel{\vdots} X)$ is different from $E_\oplus(Y|X)$, the former is closest to the latter in the sense that the objective function by which we obtain the former is closer to the objective by which we obtain the latter than any other function in the $L_2(P_X)$-closure of $\mathcal{M}_X + \operatorname{span}(\mathbb{1}_X)$.

When Ω_Y is a Hilbert space, say $\mathcal{H}_Y$, the weak Fréchet conditional mean is defined as the minimizer of the weak conditional mean of the squared norm of the difference between $\|Y - y\|^2_{\mathcal{H}_Y}$. By making analogy with the fact that, in terms of the true conditional mean, $E(Y|X)$ is indeed the minimizer of $E(\|Y = y\|^2|X)$, it seems plausible to expect that $E(Y\,¦\,X)$ is the minimizer of $E(\|Y - y\|^2\,¦\,X)$ over $\mathcal{H}_Y$. This is indeed the case, as shown in the next theorem.

Theorem 4 *If Ω_Y is a Hilbert space, R_{UX} is defined and bounded, then*

$$E_\oplus(Y\,¦\,X) = E(Y\,¦\,X).$$

So far we have considered four types of conditional means: the conditional mean $E(Y|X)$, the Fréchet conditional mean $E_\oplus(Y|X)$, the weak conditional mean $E(Y\,¦\,X)$, and the weak Fréchet conditional mean $E_\oplus(Y\,¦\,X)$. The conditional expectation $E(Y|X)$ can be seen as the orthogonal projection onto the closed subspace $L^2(P_X)$ that minimizes the expected squared difference $E(Y - X)^2$ among all random variables X, so in a sense it is the best predictor of Y based on the information in the σ-algebra generated by a random variable X. Thus, more informally, $E(Y|X) = \Pi_{L_2(P_X)}(Y)$. For random functions X and Y taking values in general Hilbert-spaces $\mathcal{H}_1$ and $\mathcal{H}_2$, respectively, weak conditional mean is given by the projection $E(Y|X) = \Pi_{\mathcal{H}_1}(Y)$. Both the concepts have now been generalized for metric space valued data and the next corollary summarizes their relations (also see Figure 1).

Corollary 5 *Suppose R_{UX} is defined and bounded. Then*

1. *If Ω_Y is a Hilbert space, then*

$$E_\oplus(Y|X) = E(Y|X), \quad E_\oplus(Y\,¦\,X) = E(Y\,¦\,X)$$

2. *If $\mathcal{M}_X$ is dense in $L_2(P_X)$ modulo constant, then*

$$E(Y|X) = E(Y\,¦\,X), \quad E_\oplus(Y|X) = E_\oplus(Y\,¦\,X).$$

3.2 Relation with global linear Fréchet regression

Interestingly, as the next theorem shows, the weak conditional Fréchet mean reduces to the objective function of the global linear Fréchet regression introduced by Petersen and Müller (2019) in a special case, where κ_X is the linear kernel $c + x_1^\intercal x_2$. Let $\Sigma_X = \text{var}(X)$ be the covariance matrix of he random vector X.

$$\begin{tikzcd}
E(Y|X) \arrow[r, leftrightarrow, "\mathcal{M}_X \text{ is dense in } L_2(P_X)"] \arrow[d, leftrightarrow, "\Omega_Y \text{ is a Hilbert space}"'] & E\,(Y¦X) \arrow[d, leftrightarrow, "\Omega_Y \text{ is a Hilbert space}"] \\
E_\oplus(Y|X) \arrow[r, leftrightarrow, "\mathcal{M}_X \text{ is dense in } L_2(P_X)"] & E_\oplus(Y¦X)
\end{tikzcd}$$

Figure 1: Diagram describing the inter-relation between different types of conditional means.

Theorem 5 *If Σ_X is invertible, κ_X is the linear kernel $c + x_1^\intercal x_2$. Then*

$$E[d_Y^2(Y,y)¦X=x] = E\left\{[1 + (x - EX)^\intercal \Sigma_X^{-1}(X - EX)]d_Y^2(Y,y)\right\}.$$

When κ_X is any arbitrary kernel such as a linear kernel and is not necessarily a universal kernel, the weak conditional Fréchet mean $E_\oplus(Y¦X)$ is not same as the conditional Fréchet mean $E_\oplus(Y¦X)$. For example, as shown above, the target for the global Fréchet regression, which emerges as a special case of the weak conditional Fréchet means corresponding to a linear kernel, is different from the conditional Fréchet regression function $E_\oplus(Y|X)$. However, the regression relationship between two random objects $(X,Y) \in \Omega_X \times \Omega_Y$ expressed through the weak Fréchet conditional mean is interesting and worth investigating in its own right. This alternative formulation is described through an RKHS embedding in the predictor space, thus accommodating random objects lying in the general metric space as a predictor. The characterization of the dependence between Y and X is global and nonlinear, and no bandwidth parameter is required to fine-tune the regression function.

3.3 Existence and uniqueness of $E_\oplus(Y¦X)$

We now turn to the existence and uniqueness of the weak Fréchet conditional mean. Because the objective function $E_\oplus(d^2(Y,y)|X)$ cannot in general be expressed as an integral with respect to a probability measure, the existing methods (Satarupa, please provide references) used for proving the existence and uniqueness for the Fréchet conditional mean cannot be used. Nevertheless, reasonably general statements about existence and uniqueness can be made under some conditions.

For existence, by the extreme value theorem, if the function $y \mapsto E(d_Y^2(Y, y) ¦ X = x)$ and Ω_Y is compact, then there is a y_0 in Ω_Y that minimizes $E(d_Y^2(Y, y) ¦ X = x)$, which then is a weak Fréchet conditional mean.

For the uniqueness, we do not yet have a general result except for the following special case. The investigation for general sufficient conditions for uniqueness is an interesting open problem, which we leave to future research. Here, we only prove uniqueness in two special cases. Again, let $U = d(Y, y)^2$.

Proposition 3 *Suppose*

1. $R_{UY}^{(c)}$ *is defined and bounded;*

2. $\mathcal{M}_X$ *is dense in* $L_2(P_X)$ *modulo constants;*

3. Ω_Y *is a global nonpositive curvature metric space.*

Then $E_\oplus(Y ¦ X)$ *exists and is unique.*

For the definition and the related theories for the a global nonpositive curvature metric space, see Sturm (2003). The second special case is when Ω_Y is a negative-type metric space.

Definition 5 (Negative type metric space) *The space* (M, ρ) *with a semi-metric* ρ *is of negative type if for all* $n \geq 2$, $z_1, z_2, \ldots, z_n \in M$ *and* $\alpha_1, \alpha_2, \ldots, \alpha_n \in \mathbb{R}$, *with* $\sum_{i=1}^n \alpha_i = 0$, *one has* $\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \rho(z_i, z_j) \leq 0$.

Proposition 4 *Suppose* Ω_Y *is a negative type metric space,* $\mathcal{H}$ *is a Hilbert space,* $\rho : \Omega_Y \to \mathcal{H}$ *is a continuous injection, and* $\rho : \Omega_Y \to \rho(\Omega_Y)$ *is an isometry. The minimizer* $E_\oplus[Y ¦ X] = \mathrm{argmin}_{z \in \rho(\Omega_Y)} E[\|\rho(Y) - z\|_{\mathcal{H}}^2 ¦ X]$ *exists, is unique if either one of the following holds:*

1. $\mathcal{M}_X$ *is dense in* $L_2(P_X)$ *modulo constants*

2. $\mathcal{M}_X$ *is a finite RKHS and* $\rho(\Omega_Y)$ *is convex and closed in* $\mathcal{H}$.

In the latter case, $E_\oplus(Y ¦ X)$ *corresponds to the* $\mathcal{H}$*-orthogonal projection of* $E[\rho(Y) ¦ X]$ *on* $\rho(\Omega_Y)$.

The existence of such a isometric continuous map not a strong requirement. For example, if Ω_Y is a separable metric space of negative type, one can always define the distance-induced kernel $\kappa : \Omega_Y \times \Omega_Y \to \mathbb{R}$ as

$$\kappa(y, y') = \frac{1}{2}[d_Y(y, y_0) + d_Y(y', y_0) - d_Y(y, y')],$$

for any fixed element $y_0 \in \Omega_Y$. Then there us a unique RKHS $\mathcal{H}$ generated by this κ and the map $\rho : \Omega_Y \to \mathcal{H}$ defined by $\rho(y) = \kappa(\cdot, y)$ satisfies all the requirements of the above proposition. Further, for many commonly observed object-valued data, the image set $\rho(\Omega_Y)$ is closed and convex in the underlying Hilbert space $\mathcal{H}$. Some examples are discussed in the following.

Example 1: The space of univariate probability distributions G on $\mathbb{R}$ such that $\int_{\mathbb{R}} x^2 G(x) < \infty$, equipped with the Wasserstein-2 metric. For two such distributions G_1 and G_2, the Wasserstein-2 metric between G_1 and G_2 is given by

$$d_W^2(G_1, G_2) = \int_0^1 (G_1^{-1}(t) - G_2^{-1}(t))^2 dt, \tag{10}$$

where G_1^{-1} and G_2^{-1} are the quantile functions corresponding to G_1 and G_2, respectively. The weak conditional Fréchet mean for distributional objects endowed with the Wasserstein-2 metric d_W as defined above is given by the distributional object whose corresponding quantile function is equal to the $L^2([0,1])$-orthogonal projection of $E[Q_Y \,¦\, X]$ on $Q(\Omega_Y)$, where $Q(\Omega_Y)$ denotes the space of distributions represented as quantile functions and

$$E[Q_Y \,¦\, X] = E(Q_Y) + \langle \kappa_X(\cdot, x) - \mu_X,\ \Sigma_{XX}^{\dagger}\ E\left((\kappa_X(\cdot, X) - \mu_X)Q_Y\right)\rangle_{\mathcal{M}_X}.$$

Example 2: The space of symmetric positive semi-definite matrices with unit diagonal, Ω_Y, endowed with the Frobenius metric d_F. For any two elements $A, B \in (\Omega_Y, d_F)$, their Frobenius distance is given by

$$d_F^2(A, B) = \sqrt{\text{trace } ((A - B)(A - B)^T)}. \tag{11}$$

The weak conditional Fréchet mean for spd matrix objects equipped with the Frobenius metric d_F is given by the orthogonal projection of $B(x)$ onto the space of correlation matrices, where $B(x)$ has the (j, k)-th entry as

$$B_{jk}(x) = E(Y_{jk}) + \langle \kappa_X(\cdot, x) - \mu_X,\ \Sigma_{XX}^{\dagger}\ E\left((\kappa_X(\cdot, X) - \mu_X)Y_{jk}\right)\rangle_{\mathcal{M}_X}.$$

Here Y_{jk} is the (j,k)-th entry of $Y \in (\Omega_Y, d_F)$. The existence, uniqueness, and explicit form of the weak conditional Fréchet mean can also be derived for other Euclidean and pseudo-Euclidean metrics such as power metric, log-affine metric, Cholesky metric and so on (Dryden et al., 2010; Lin, 2019).

4 Estimation

In the last section, we have described the solution to the nonlinear object regression framework at the population level. In the following, we implement the regression at the sample level. The key steps involve the construction of the sample estimate for the regression function as an M-estimator based on i.i.d. paired observations $(X_i, Y_i)_{i=1}^n$. In order to quantify the sample objective function minimized by the regression estimator, we need to express the underlying RKHS $\mathcal{H}_X$ and the relevant auto covariance and pseudo-cross covariance operators with a coordinate representation system (see, e.g., Horn and Johnson (2012); Li (2018)).

4.1 Coordinate representation

Suppose that $\mathcal{L}_1$ is a finite dimensional linear space with basis $\mathcal{B} = \{\xi_1, \xi_2, \ldots, \xi_p\}$. Then for any $\xi \in \mathcal{L}_1$, there is a unique vector $(a_1, a_2, \ldots, a_p)^\intercal \in \mathbb{R}^p$ such that $\xi = \sum_{i=1}^p a_i \xi_i$. The vector $(a_1, a_2, \ldots, a_p)^\intercal$ is called the coordinate of ξ with respect to $\mathcal{B}$, and denoted by $[\xi]_{\mathcal{B}}$. Throughout this section, we will use this notation to describe coordinate representation. Next, we introduce the coordinate representation of a linear operator between two (finite-dimensional) linear spaces. Suppose $\mathcal{L}_2$ is another linear space with basis $\mathcal{C} = \{\eta_1, \eta_2, \ldots, \eta_q\}$ and A is a linear operator from $\mathcal{L}_1$ $\mathcal{L}_2$. Then for any $\eta \in \mathcal{L}_1$, we have

$$\begin{aligned} A\xi = A\left(\sum_{i=1}^p \left([\xi]_{\mathcal{B}}\right)_i \xi_i\right) &= \sum_{i=1}^p \left([\xi]_{\mathcal{B}}\right)_i (A\xi_i) \\ = \sum_{i=1}^p \left([\xi]_{\mathcal{B}}\right)_i \sum_{j=1}^q \left([A\xi_i]_{\mathcal{C}}\right)_j \eta_j &= \sum_{j=1}^q \left\{\left({}_{\mathcal{C}}[A]_{\mathcal{B}}\right)\left([\xi]_{\mathcal{B}}\right)\right\}_j \eta_j, \end{aligned}$$

where ${}_{\mathcal{C}}[A]_{\mathcal{B}}$ is the $q \times p$ matrix with (i,j)th entry $([A\xi_j]_{\mathcal{C}})_i$. The above equation implies that $[A\xi]_{\mathcal{C}} = ({}_{\mathcal{C}}[A]_{\mathcal{B}})([\xi]_{\mathcal{B}})$. Therefore we call the matrix ${}_{\mathcal{C}}[A]_{\mathcal{B}}$ the coordinate representation of the linear operator A with respect to the bases $\mathcal{B}$ and $\mathcal{C}$. Similarly, for two Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$, with spanning systems $\mathcal{B}_1$ and $\mathcal{B}_2$, and a linear operator $A : \mathcal{H}_1 \to \mathcal{H}_2$, we use the notation ${}_{\mathcal{B}_1}[A]_{\mathcal{B}_2}$ to represent the coordinate representation of A relative to spanning systems $\mathcal{B}_1$ and $\mathcal{B}_2$.

4.2 Construction of the RKHS $\mathcal{H}_X$ and Model Fitting

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. observations of $(X, Y) \in \Omega_X \times \Omega_Y$. The RKHS $\mathcal{H}_X$ is spanned by $\{\kappa_X(\cdot, X_i) : i = 1, \ldots .n\}$ equipped with the inner product

$$\langle f, g \rangle_{\mathcal{M}_X} = [f]^{\intercal} K_X [g],$$

for any $f, g \in \mathcal{H}_X$, where K_X is the $n \times n$ Gram matrix whose (i,j)th entry is $\kappa_X(X_i, X_j)$, $i, j = 1, \ldots, n$. Further, since the evaluation functional of the objective functions the weak conditional Fréchet mean minimizes depend on $y \in \Omega_Y$, we denote $U = U(y) = d_Y^2(Y, y)$. Similarly define $V(y) = d_Y(Y, y)$, and the sample observations as $U_i(y) = d_Y^2(Y_i, y)$ and$V_i(y) = d_Y(Y_i, y)$, respectively.

At the sample level, we estimate Σ_{XX} $\Sigma_{XU(y)}$, and $\Sigma_{XV(y)}$ by replacing the expectations $E(\cdot)$ with the sample moments $E_n(\cdot)$ with respect to the empirical measure whenever possible. For example, we estimate Σ_{XX} by $\hat{\Sigma}_{XX} = \frac{1}{n}\sum_{i=1}^n (\kappa_X(\cdot, X_i) - \hat{\mu}_X) \otimes (\kappa_X(\cdot, X_i) - \hat{\mu}_X)$, where $\hat{\mu}_X = \frac{1}{n}\sum_{i=1}^n \kappa_X(\cdot, X_i)$. The sample estimate for $\Sigma_{XU(y)}$ and $\Sigma_{XV(y)}$ for any given $y \in \Omega_Y$ is similarly defined as $\hat{\Sigma}_{XU(y)} = \frac{1}{n}\sum_{i=1}^n (\kappa_X(\cdot, X_i) - \hat{\mu}_X) U_i(y)$, and $\hat{\Sigma}_{XV(y)} = \frac{1}{n}\sum_{i=1}^n (\kappa_X(\cdot, X_i) - \hat{\mu}_X) V_i(y)$, respectively. Suppose, the subspace $\overline{\text{ran}}(\hat{\Sigma}_{XX})$ is spanned by the set $\mathcal{B}_X = \{\kappa_X(\cdot, X_i) - E_n(\kappa_X(\cdot, X_i)) : i = 1, \ldots, n\}$. We then have the following coordinate representations of auto covariance and pseudo-cross covariance operators, for any $y \in \Omega_Y$,

$${}_{\mathcal{B}_X}[\hat{\Sigma}_{XX}]_{\mathcal{B}_X} = n^{-1} G_X, \ [\hat{\Sigma}_{XU(y)}]_{\mathcal{B}_X} = [\hat{\Sigma}_{XV(y)}]_{\mathcal{B}_X} = n^{-1} G_X, \ {}_{\mathcal{B}_X}[\hat{\Sigma}_{XX}^{\dagger}]_{\mathcal{B}_X} = n^{-1} G_X^{\dagger},$$

where $G_X = Q K_X Q$ and $G_X^{\dagger}$ is the Moore-Penrose inverse of G_X via the Tikhonov-regularized inverse $(G_X + \epsilon_X I_n)^{-1}$ to prevent overfitting, where $\epsilon_X > 0$ is a tuning constant. Here Q denotes the projection matrix $I_n - \frac{1}{n} 1_n 1_n^T$ with $Q^2 = Q$. For a detailed discussion see e.g. Section 12.4 of Li (2018).

Recalling the definition of the population level weak conditional Fréchet mean $E[Y \,\vdots\, X = x]$ from (8) given by

$$f_{\oplus}(x) = \underset{y \in \Omega_Y}{\operatorname{argmin}}\ J(y), \text{ where } J(y) = E[U(y)] + \langle \kappa_X(\cdot, x) - \mu_X, \Sigma_{XX}^{\dagger} \Sigma_{XU(y)} \rangle_{\mathcal{M}_X}, \quad (12)$$

we define the following estimator

$$\hat{f}_\oplus(x) = \underset{y \in \Omega_Y}{\operatorname{argmin}}\ J_n(y), \text{ where } J_n(y) = \frac{1}{n}\sum_{i=1}^n U_i(y) + \langle \kappa_X(\cdot, x) - \hat{\mu}_X, \hat{\Sigma}_{XX}^\dagger \hat{\Sigma}_{XU(y)} \rangle_{\mathcal{M}_X}. \tag{13}$$

To obtain a more explicit computable form of the above, it remains to identify the coordinate of $\kappa_X(\cdot, x) - \hat{\mu}_X$ with respect to the spanning system $\{\kappa_X(\cdot, X_i) - \hat{\mu}_X : i = 1, \ldots, n\}$. Suppose that $[\kappa_X(\cdot, x) - \hat{\mu}_X] = c_x$ for some $c_x \in \mathbb{R}^n$. Then

$$\langle \kappa_X(\cdot, x) - \hat{\mu}_X, \kappa_X(\cdot, X_i) - \hat{\mu}_X \rangle_{\mathcal{M}_X} = e_i^\intercal K_X c_x - \frac{1}{n}(e_i^\intercal K_X 1_n)(1_n^\intercal c_x) = e_i^\intercal K_X Q c_x,$$

where e_i denotes the vector whose ith component is 1 and all others are 0. Taking $i = 1, \ldots, n$, we have $d_x = K_X Q c_x$, where d_x is the vector of length n with ith component $\kappa_X(X_i, x) - E_n(\kappa_X(X_i, x))$. With the Tikhonov regularization, we obtain the solution $c_x = Q(K_X + \epsilon_X I_n)^{-1} d_x$. Thus, the empirical objective function in (13) becomes

$$J_n(y) = \frac{1}{n} h_Y^\intercal 1_n + h_Y^\intercal G_X (G_X + \epsilon_X I_n)^{-1} c_x,$$

where h_Y is the vector with the i-th component $U_i(y)$, $i = 1, \ldots, n$, and $1_n = (1, 1, \ldots, 1)^\intercal$.

4.3 Tuning parameter selection

We use the general cross-validation criterion (Golub et al., 1979) to determine the tuning constant ϵ_X involved in the Tikhonov-regularization of the inverse auto-covariance operator $\Sigma_{XX}^\dagger$.

$$\mathrm{GCV}(\epsilon_X) = \frac{1}{n}\sum_{i=1}^n \frac{d_Y^2(Y_i, \hat{Y}_i)}{\left(1 - \mathrm{tr}[Q G_X (G_X + \epsilon_X I_n)^{-1} + 1_n 1_n^\intercal / n]/n\right)^2}, \tag{14}$$

where Y_i and $\hat{Y}_i$ are respectively the observed and predicted responses for the i-th subject, $i = 1, \ldots, n$. The numerator of this criterion quantifies the prediction error while the denominator controls the degree of overfitting. We minimize the criterion over a grid $\{10^{-6}, \ldots, 10^{-1}\}$ to find the optimal tuning constants.

5 Convergence results

In this section, we develop the asymptotic convergence results for the proposed object regression method. In particular, the convergence of the auto-covariance and pseudo-cross-covariance operators with a suitable rate is established, which is used in turn to show the convergence of the regression estimate using the M-estimation theory.

5.1 Convergence of regression operators

The asymptotic properties of the empirical estimates of the mean and auto covariance operator defined on the RKHS $\mathcal{H}_X$ have been well-studied in the literature (see e.g., Sang and Li (2022); Fukumizu et al. (2007); Lee et al. (2013). For completion, we list the properties here

Lemma 1 *Under Assumptions 1- 3, and 6- 7,*

(1) $\|\hat{\mu}_X - \mu_X\|_{\mathcal{M}_X} = O_P(n^{-1/2})$.

(2) $\|\hat{\Sigma}_{XX} - \Sigma_{XX}\|_{OP} = O_P(n^{-1/2})$.

Suppose the eigenvalue and eigenfunction sequence of Σ_{XX} is given by $\{(\lambda_j, \phi_j) : j = 1, 2, \dots\}$. By Mercer's theorem, the spectral decomposition of the variance operator Σ_{XX} is given by

$$\Sigma_{XX} = \sum_{j=1}^{\infty} \lambda_j \phi_j \otimes \phi_j. \tag{15}$$

Typically, for a positive definite κ_X, Σ_{XX} is a compact operator whose eigenvalues decay to 0, hence $\Sigma_{XX}^{\dagger}$ is unbounded. However, it is reasonable to assume that the regression operators $R_{XV(y)} := \Sigma_{XX}^{\dagger}\Sigma_{XV(y)}$ and $R_{XU(y)} := \Sigma_{XX}^{\dagger}\Sigma_{XU(y)}$ to be bounded uniformly for all $y \in \Omega_Y$. We assume a degree of smoothness smoothness on the joint distribution of (X, Y), requiring that the output functions for the regression operator must be sufficiently concentrated on the low-frequency components of Σ_{XX}. The following assumption is a stronger version of Assumption 7.

Assumption 8 $\sup_{y \in \Omega_Y} E\left(|\phi_j(X) - E(\phi_j(X))|\ d_Y^k(Y, y)\right) \leq \lambda_j^2$, $k = 1, 2$.

i.e., $R_{XU(y)} := \Sigma_{XX}^{\dagger}\Sigma_{XU(y)}$ and $R_{XV(y)} := \Sigma_{XX}^{\dagger}\Sigma_{XV(y)}$; are bounded operators uniformly for all $y \in (\Omega_Y, d_Y)$, in other words $\text{ran}(\Sigma_{XU(y)})$, which can possibly depend on y, is entirely contained in the $\mathrm{ran}(\Sigma_{XX})$ uniformly across all possible $y \in \Omega_Y$, similarly for $\Sigma_{XV(y)}$. This is a generalization of Assumptions 4 and 5 for the psedu-cross covariance operators indexed by $y \in \Omega_Y$. Condition 8 guarantees that that the composite operators $\Sigma_{XX}^{\dagger}\Sigma_{XU(y)}$ is well-defined, bounded and compact, uniformly for all $y \in \Omega_Y$, for $k = 1, 2$. This implies that $\Sigma_{XX}^{\dagger}\Sigma_{XU(y)}$ must send all incoming functions into the low-frequency range of the eigenspaces of Σ_{XX} with relatively large eigenvalues uniformly for all $y \in \Omega_Y$, for $k = 1, 2$. That is, $\Sigma_{XU(y)}$ is smooth uniformly for all $y \in \Omega_Y$ in the sense that its outputs are low-frequency components of Σ_{XX},, similarly for $\Sigma_{XV(y)}$.

The consistent estimation for the pseudo-cross covariance operators is derived uniformly over all elements $y \in \Omega_Y$, under the following assumption on the intrinsic geometry and complexity of the response space (Ω_Y, d_Y), which can be quantified by a bound on the entropy integral of Ω_Y.

Assumption 9 *The entropy integral of Ω_Y is finite, i.e.,*

$$J := \int_0^1 \sqrt{1 + \log N(\epsilon, \Omega_Y, d)}\, d\epsilon < \infty,$$

where $N(\epsilon, \Omega_Y, d)$ is the covering number for the space Ω_Y using balls of radius ϵ.

This assumption is satisfied by most of the commonly observed random objects such as the space of univariate distributions with Wasserstein metric, space of positive semi-definite matrices with a suitable choice of metric, data on the surface of an $n-$ sphere with the intrinsic geodesic metric, and so on (see e.g. Dubey and Müller (2019) and the references therein).

Lemma 2 *Under Assumptions 1- 3, and 6- 9,*

$$\sup_{y \in \Omega_Y} \|\hat{\Sigma}_{XU(y)} - \Sigma_{XU(y)}\|_{OP} = O_P(n^{-1/2}); \quad \sup_{y \in \Omega_Y} \|\hat{\Sigma}_{XV(y)} - \Sigma_{XV(y)}\|_{OP} = O_P(n^{-1/2}).$$

The consistent estimation for the regression operators is described in the following lemma, under further smoothness conditions on the regression relationship between X and Y.

Assumption 10 *For all $j \in \mathbb{N}$, there is a $0 < \beta \leq 1$ such that $\sup_{y \in \Omega_Y} E\left(|\phi_j(X) - E(\phi_j(X))|\, d_Y^{(k)}(Y, y)\right) \leq \lambda_j^{2+\beta}$, for $k = 1, 2$, i.e. there is a bounded linear operator $S_{XY} : \mathcal{H}_X \to \mathcal{H}_X$ such that $\sup_{y \in \Omega_Y} \Sigma_{XX}^{(1+\beta)^{\dagger}} \Sigma_{XY}^{(k)}(y)$ is a bounded linear operator uniformly over all $y \in \Omega_Y$ for $k = 1, 2$.*

Suppose $n^{-1/2} \prec \epsilon_n \prec 0$. For any β as defined in Assumption 10, define

$$\alpha_n = \epsilon_n^{\beta} + \epsilon_n^{-1} n^{-1/2}. \tag{16}$$

Proposition 5 *Under Assumptions 1- 3, and 6- 10,*

$$\sup_{y \in \Omega_Y} \|\hat{\Sigma}_{XX}^{\dagger} \hat{\Sigma}_{XU(y)} - \Sigma_{XX}^{\dagger} \Sigma_{XU(y)}\|_{OP} = O_P(\alpha_n); \ \sup_{y \in \Omega_Y} \|\hat{\Sigma}_{XX}^{\dagger} \hat{\Sigma}_{XV(y)} - \Sigma_{XX}^{\dagger} \Sigma_{XV(y)}\|_{OP} = O_P(\alpha_n),$$

where α_n is as given in (16).

5.2 Estimation of weak conditional Fréchet mean

Having established the convergence of the pseudo-regression operators, we proceed to derive the convergence results for the weak Fréchet conditional mean in (13). We require the following assumptions regarding the intrinsic geometry of the response space, which are the key to establishing the rate of convergence of any M-estimator, namely, the assumption of well-separateness of the minimizer, an upper bound on the entropy integral of the underlying metric space, and a local lower bound on the curvature of the objective functions listed in the Appendix.

Theorem 6 *Under Assumptions 1- 3, 6- 10, and the technical assumptions 11- 12 in the Appendix, for any $x \in (\Omega_X, d_X)$,*

$$d_Y(\hat{f}_{\oplus}(x), f_{\oplus}(x)) = o_P(1).$$

Theorem 7 *Under Assumptions 1- 3, 6- 10, and the technical assumptions 11- 13 in the Appendix, for any $x \in (\Omega_X, d_X)$,*

$$d_Y(\hat{f}_{\oplus}(x), f_{\oplus}(x)) = O_P(\alpha_n),$$

where α_n is as given in (16).

For most commonly observed random objects β in Assumption 13 is 2, yielding an asymptotic rate of convergence for the M-estimator as $O_P(\alpha_n^{-1})$. With a suitable rate carried from the RKHS regression literature, one can derive the rate of convergence as a function of the sample size n. For example, in Li and Song (2017), $\alpha_n \approx n^{-1/4}$, which is improved upon by Sang and Li (2022) as $\alpha_n \approx n^{-1/3}$. This improved rate can be incorporated in the rate of convergence for the weak conditional Fréchet mean to yield an optimal rate of $O_P(n^{-1/3})$.

6 Simulation Studies

In this section, we evaluate the numerical performances of the proposed object-on-object regression method under different simulation settings for commonly observed random objects.

In all of the following simulation scenarios, we consider Gaussian radial basis kernel $\kappa_G(y, y') = \exp(-\gamma_X d^2(y, y'))$ as a candidate to construct the underlying RKHS $\mathcal{H}_X$ in the predictor space. We choose the parameters γ_X as the fixed quantity

$$\gamma_X = \frac{\rho_Y}{2\sigma_G^2},\ \sigma_G^2 = \binom{n}{2}^{-1} \sum_{i<j} d^2(X_i, X_j),\ \rho_Y = 1.$$

The same choices of tuning parameters were used in Lee et al. (2013); Li and Song (2017); Zhang et al. (2022). The metrics d_X and d_Y are chosen appropriately to enhance the interpretability of the results in each of the following scenarios considered.

6.1 Scenario 1: Univariate distribution-on-object regression

In the first scenario considered, we have univariate distributional objects as responses coupled with various types of random objects as predictors. Let (Ω_Y, d_Y) be the metric space of univariate distributions endowed with Wasserstein metric $d_Y = d_W$, as described in (10) Section 3.3. A sample of distributional object response, $Y_1, \ldots, Y_n$ is observed in equivalent forms of CDF, quantile functions, or densities. However, the distributions $Y_1, \ldots, Y_n$ are usually not fully observed in practice and the latent curves need to be recovered from the discrete observations $\{Y_{ij}\}_{j=1}^m$, $i = 1, \ldots, n$, one encounters in reality. For this, we employ nonparametric smoothing with a suitable bandwidth choice implemented by the *CreateDensity()* function in the *frechet* R package (Chen et al., 2020). While considering distributional predictors, the trajectories X_i are recovered from the discrete observations $\{X_{ij}\}_{j=1}^m$; $i = 1 \ldots, n$ in a similar manner.

The random distributional response Y is generated conditional on X by adding noise to the quantile functions, which are demonstrated in the following simulation settings for various types of predictor objects. Generally, we let $Y = N(\zeta(x), \eta^2(x))$, where the mean and variance of the response distribution are dependent on X. To this end, the auxiliary distribution parameters μ_Y and σ_Y, given X, are independently sampled such that $E(\mu_Y|X = x) = \zeta(x)$ and $E(\sigma_Y^2|X = x) = \eta^2(x)$, and the corresponding distributional response in its qunatile representation is constructed as $Q_Y(\cdot) = \mu_Y + \sigma_Y \Phi^{-1}(\cdot)$.

The minimization problem in (13) is solved by considering quantile function representation of the distributional responses. If Q_{Y_i} is the quantile function corresponding to Y_i, $i = 1, \dots, n$; and $\hat{Q}_\oplus(\cdot; x)$ is the quantile function corresponding to the distribution $\hat{f}_\oplus(x)$ in (13), using similar logic as the proof of Proposition 4,

$$\hat{Q}_\oplus(\cdot; x) = \operatorname{argmin}_{q \in Q(\Omega_Y)} \|q - \frac{1}{n}\sum_{i=1}^{n} w_{in}(x) Q_{Y_i}\|_{L^2[0,1]}$$

Existence and uniqueness of the solution of the above and therefore of 13 is guaranteed $\hat{Q}_\oplus(\cdot; x)$ corresponds to the orthogonal projection of $g_x := \frac{1}{n}\sum_{i=1}^{n} w_{in}(x) Q_{Y_i}$ as an element of the Hilbert space $L^2([0,1])$ on the closed and convex set $Q(\Omega_Y)$, where $Q(\Omega_Y)$ is the space of quantile functions corresponding to distributions in (Ω_Y, d_W), as shown in Proposition 4.

Taking an equidistant grid $\{u_j\}_{j=1}^M$ on $[0,1]$ and evaluating $g_j := g_x(u_j)$, a discretized version, $\hat{Q}^*$, of the approximation of $\hat{Q}_\oplus(\cdot; x)$ is computed by solving the constrained quadratic program problem $\hat{Q}^* = \operatorname{argmin}_{q \in \mathbb{R}^M} \|g - q\|_E$ such that $q_1 \leq q_2 \cdots \leq q_M$. We employ an OSQP solver to implement this in practice.

We set $n = 200, 400$, $m = 50, 100$ and generate n samples $(X_i, \{Y_{ij}\}_{j=1}^m)_{i=1}^n$. We use half of them to train the predictors via the proposed object regression method and then evaluate the discrepancy between the estimated and true responses using the rest of the data set by computing the Wasserstein distance metric (10) between the two distributions. The tuning parameter for computing $\Sigma_{XX}^\dagger$ is determined by the method described in Section 4.3. The experiment is repeated $B = 100$ times, and averages and standard errors (in parentheses) of the prediction error are computed as

$$\text{RMPE} := \frac{1}{B}\sum_{b=1}^{B} d_W(Y_b^{\text{test}}, \hat{Y}_b^{\text{test}}), \tag{17}$$

where Y_b^{test} and $\hat{Y}_b^{\text{test}}$ are the observed and predicted responses in the test set, respectively, for the b-th replicate, $b = 1 \dots, B$.

Model-I.1 (Euclidean predictors): $\mu_Y | X \sim N((\beta^\intercal X)^2, \nu_1^2)$ and $\sigma_Y | X \sim Gamma((\gamma^\intercal X)^2/\nu_2, \nu_2/(\gamma^\intercal X))$.

Model-I.2 (Euclidean predictors): After sampling the distribution parameters as in the previous setting, the resulting distribution is then "transported" in Wasserstein space via a random transport map T, that is uniformly sampled from a family of perturbation/ distortion functions $\{T_k : k \in \pm 1, \pm 2, \}$, where $T_k(x) = x - \sin(kx)/|k|$. The transported distribution is given by $T\#(\mu_Y + \sigma_Y \Phi^{-1})$, where $T\#p$ is a push-forward measure such that $T\#p(A) = p\left(\{x : T(x) \in A\}\right)$, for any measurable function $T : \mathbb{R} \to \mathbb{R}$, distribution $p \in (\Omega_Y, d_W)$, and set $A \subset \mathbb{R}$. We sample the random transport map T uniformly from the collection of maps described above; p denotes a Gaussian distribution with parameters $\zeta(x) = (\beta^\intercal X)^2$ and $\eta^2(x) = (\gamma^\intercal X)^2$. The distributions thus generated are not Gaussian anymore due to transportation. The conditional Fréchet mean can be shown to remain at $\mu_Y + \sigma_Y \Phi^{-1}$ as before.

For Models I.1 and I.2, the Euclidean vector predictor $X \in \mathbb{R}^p$ is generated as follows: (i) we first generate $U_1, \ldots, U_p$ from the AR(1) model with mean 0 and covariance matrix $\Sigma = (0.5^{|i-j|})_{i,j}$, and then (ii) generate $X_j = 2\Phi(U_j) - 1$, $j = 1, \ldots, p$, where Φ is the c.d.f. of $N(0, 1)$. We select $\nu_1^2 = 0.1$, $\nu_2 = 0.25$, $\beta = (1, -2, 0, 1)$, and $\gamma = c(0.1, 0.2, 1, 0.3)$ in the above models.

The performance of our method, denoted by global nonlinear Fréchet regression (GNLFR), is compared with the globally linear Fréchet regression (GLFR) method by Petersen and Müller (2019) for varying levels of the predictor dimension, sample size, and number of discrete observations for each sample of distributions, namely p, n, and m, respectively. Table 1 summarizes the results. The prediction error decreases generally corresponding to a lower dimension p of the predictor, a higher sample size n, and a denser design (higher m) over which the response is sampled. Across the board, our method outperforms the GLFR method in terms of prediction accuracy. Especially, for Setting I.2 the GNLFR method proves significantly better, which is not unexpected given the highly non-linear data-generating mechanism for this setting.

For models I.3-I.5 below, we consider univariate distribution-on-distribution regression.

Model-I.3 (Univariate distributions as predictors): $\mu_Y | X \sim N(\exp(W_2^2(X, \mu_1)) + \exp(W_2^2(X, \mu_2)), \nu_1^2)$ and $\sigma_Y | X = 0.1$.

Model-I.4 (Univariate distributions as predictors): $\mu_Y | X \sim N(\exp(W_2^2(X, \mu_1))$

Table 1: Table showing the Monte Carlo mean (standard error) estimation errors as per (17) for the proposed global nonlinear Fréchet regression (GNLFR) and the global linear Fréchet regression by Petersen and Müller (2019) (GLFR), for Euclidean predictors and univariate distributional responses in Scenario I.1-I.2. The lowest number in a row corresponding to each data generating mechanism is highlighted.

	I.1 (GNLFR)		I.1 (GLFR)		I.2 (GNLFR)		I.2 (GLFR)	
(p,n)\m	50	100	50	100	50	100	50	100
(4,200)	0.037 (0.012)	**0.024** (0.016)	0.053 (0.021)	0.038 (0.014)	0.110 (0.081)	**0.087** (0.070)	0.230 (0.012)	0.181 (0.011)
(10,200)	0.051 (0.019)	**0.042** (0.015)	0.060 (0.017)	0.049 (0.020)	0.187 (0.031)	**0.112** (0.023)	0.334 (0.045)	0.278 (0.031)
(20,200)	0.058 (0.018)	**0.051** (0.018)	0.071 (0.020)	0.065 (0.019)	0.210 (0.029)	**0.153** (0.028)	0.431 (0.025)	0.391 (0.022)
(4,400)	0.021 (0.009)	**0.013** (0.009)	0.034 (0.010)	0.021 (0.011)	0.089 (0.021)	**0.047** (0.022)	0.134 (0.020)	0.086 (0.021)
(10,400)	0.029 (0.010)	**0.023** (0.011)	0.037 (0.009)	0.024 (0.008)	0.174 (0.019)	**0.133** (0.020)	0.356 (0.012)	0.239 (0.014)
(20,400)	0.041 (0.013)	**0.033** (0.011)	0.081 (0.015)	0.043 (0.015)	0.189 (0.016)	**0.122** (0.016)	0.451 (0.013)	0.378 (0.015)

$, \nu_1^2)$ and $\sigma_Y|X = Gamma(W_2^2(X, \mu_2), W_2(X, \mu_2))$.

Model-I.5 (Univariate distributions as predictors): $\mu_Y|X \sim N(\exp(H(X, \mu_1)), 0.2^2); \sigma_Y|X = \exp(H(X, \mu_2))$.

In the above we let $\nu_1^2 = 0.1$, $\mu_1 = Beta(2, 1)$ and $\mu_2 = Beta(2, 3)$ and generate discrete observations from distributional predictors by $\{X_{ij}\}_{j=1}^m \overset{i.i.d.}{\sim} Beta(a_i, b_i)$, where $a_i \overset{i.i.d.}{\sim} Gamma(2, \text{rate} = 1)$ and $b_i \overset{i.i.d.}{\sim} Gamma(2, \text{rate} = 3)$. $W_2(\cdot, \cdot)$ and $H(\cdot, \cdot)$ denote, respectively, the Wasserstein-2 distance and the Hellinger distance between two univariate distributional objects. The Hellinger distance between two Beta distributions $\mu = Beta(a_1, b_1)$ and $\nu = Beta(a_2, b_2)$ can be represented explicitly as

$$H(\mu, \nu) = 1 - \int \sqrt{f_\mu(t) f_\nu(t)} dt = 1 - \frac{B((a_1 + a_2)/2, (b_1 + b_2)/2)}{\sqrt{B(a_1, b_1) B(a_2, b_2)}},$$

where $B(\alpha, \beta)$ is the *Beta* function.

Note that by virtue of the Gram matrix of the underlying RKHS kernel κ_X, the predictor space is now embedded into a Hilbert space, hence finding the weak conditional Fréchet mean reduces to solving a constrained quasi-quadratic optimization problem and projecting back into the solution space.

The performance of our method, denoted by global nonlinear Fréchet regression (GNLFR), is compared with the distribution-on-distribution Wasserstein regression (WR) proposed by Chen et al. (2021) for varying choices of the sample size and predictor dimension (n, m) (see Table 2). We observed a decrease in the RMPE as per (17) for all the settings as the sample size n was increased, favorably for the denser design with a higher m. For Setting I.3, the our method fairs comparably well with the WR method, but for more non-linear data generation mechanisms as in settings I.4 and I.5, our method outperforms the WR method. Further, our method uses the intrinsic geometry of the space, as compared to the WR method, which utilizes the psuedo-Riemannian structure of the Wasserstein space, thus making our estimation more reliable and robust.

Table 2: Table showing the Monte Carlo mean (standard error) estimation errors as per (17) for univariate distribution-on-distribution regression in Scenario I according to models I.3- I.5, corresponding to the proposed global nonlinear Fréchet regression (GNLFR) and the Wasserstein Regression (WR) method by Chen et al. (2021). The lowest number in a row corresponding to each data generating mechanism is highlighted.

(n, m)	I.3 (GNLFR)	I.3 (WR)	I. 4 (GNLFR)	I.4 (WR)	I.5 (GNLFR)	I.5 (WR)
(200, 50)	0.314 (0.121)	**0.298** (0.191)	**0.461** (0.110)	0.514 (0.093)	**0.491** (0.110)	0.820 (0.217)
(200, 100)	**0.268** (0.091)	0.272 (0.110)	**0.381** (0.125)	0.443 (0.112)	**0.407** (0.099)	0.788 (0.098)
(400, 50)	0.159 (0.092)	**0.155** (0.082)	**0.218** (0.160)	0.310 (0.188)	**0.251** (0.181)	0.549 (0.167)
(400, 100)	**0.134** (0.086)	0.141 (0.079)	**0.172** (0.155)	0.256 (0.167)	**0.177** (0.120)	0.422 (0.115)

We next consider the scenario where X is a two-dimensional random Gaussian distribution in Models I.6-I.7. Similar data generation mechanism was followed in Zhang et al. (2022).

Model-I.6 (Multivariate distributions as predictors): $\mu_Y|X \sim N(\exp(W_2(X, \mu_1)), \nu_1^2)$ and $\sigma_Y|X = 0.1$, with $\mu_1 \sim N((-1, 0)^\intercal, \mathrm{diag}(1, 0.5))$.

Model-I.7 (Multivariate distributions as predictors): $\mu_Y|X \sim N(\exp(W_2(X, \mu_1)), \nu_1^2)$ and $\sigma_Y|X = \tau_1^\intercal \Lambda \tau_2$, with $\mu_1 \sim N((-1, 0)^\intercal, \mathrm{diag}(1, 0.5))$; $\tau_1 = (1/\sqrt{2}, 1/\sqrt{2})^\intercal$, $\tau_2 = (1/\sqrt{2}, -1/\sqrt{2})^\intercal$, $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2)$, where $(\lambda_1, \lambda_2)|X \sim N(W_2(X, \mu_2)(1, 1)^\intercal, 0.25 I_2)$, $\mu_2 \sim N((0, 1)^\intercal, \mathrm{diag}(0.5, 1))$.

When computing $W_2(X, \mu_1)$ and $W_2(X, \mu_2)$, we use the following explicit representations of the Wasserstein distance between two Gaussian distributions:

$$W_2^2(N(m_1, \Sigma_1), N(m_2, \Sigma_2)) = ||m_1 - m_2||^2 + ||\Sigma_1 - \Sigma_2||_F, \tag{18}$$

Table 3 shows a lower RMPE for the less complex setting I.6, while the performance of the method improves for higher n, m as before.

Table 3: Table showing the mean and s.e. (in parenthesis) of the prediction errors as per (17), for multivariate distributions as predictors coupled with univariate distributions as responses, as described in Models I.6-I.7 in Scenario I. The lowest number in a row corresponding to each data-generating mechanism is highlighted.

	I.6		I.7	
n\m	50	100	50	100
200	0.619 (0.110)	**0.534** (0.100)	0.719 (0.142)	**0.578** (0.131)
400	0.467 (0.091)	**0.388** (0.092)	0.635 (0.110)	**0.541** (0.112)

In Model I.8, Hilbertian random functions are taken as predictor objects coupled with univariate distribution responses, where the distribution of the response varies conditional on the predictor values as before.

Model-I.8 (Random functions as predictors): The predictor trajectories X and associated noisy measurements were generated as follows. Suppose that the simulated process X has the mean function $\mu_X(s) = s + \sin(s)$, with covariance function constructed from two eigenfunctions, $\phi_1(s) = \sqrt{2}\sin(2\pi ks)$ and $\phi_2(s) = \sqrt{2}\cos(2\pi ks)$, $0 \leq s \leq 1$. We chose $\lambda_1 = 1, \lambda_2 = 0.7$ and $\lambda_k = 0$ for $k \geq 3$, as eigenvalues, and the FPC scores ξ_k; $(k = 1, 2)$ were generated from $N(0, \lambda_k)$. Using the Kerhunen-Loéve expansion the predictor process is then given by $X(s) = \mu_X(s) + \sum_{k=1}^{\infty} \xi_k \phi_k(s)$. To adequately reflect both a dense design and an irregular/sparse measurement paradigm, we assume that there is a random number N_i of random measurement times for X_i for the i-th subject, which are denoted as $S_{i1}, \ldots, S_{iN_i}$ and contaminated with measurement errors ϵ_{ij}, $1 \leq j \leq N_i$, $1 \leq i \leq n$. The errors are assumed to be i.i.d. with $E(\epsilon_{ij}) = 0$ $E[\epsilon_{ij}^2] = \sigma_X^2 = 0.1$, and independent of functional principal component scores ξ_{ik} that satisfy $E[\xi_{ik}] = 0$, $E[\xi_{ik}\xi_{ik'}] = 0$ for $k \neq k'$, and $E[\xi_{ik}^2] = \lambda_k$. Thus, for the i-th sample, the predictor measurement with noise is represented as $U_{ij} = \mu_X(S_{ij}) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k(S_{ij}) + \epsilon_{ij}$, $i = 1, \ldots, n,\ j = 1, \ldots, N_i$. The data generation mechanism above is similar to Yao et al. (2005) and both a sparse and a dense grid of observation are considered with $N_i = 50$ and $N_i \in \{3, \ldots, 5\}$, respectively. Finally, the response as a univariate distribution is constructed as $Y \sim N(\mu_Y, \sigma_Y)$, and the auxiliary parameters conditional on $X(\cdot)$ are generated independently as $\mu_Y | X \sim N((\xi_1, \xi_2)^\intercal \mathrm{diag}(\lambda_1, \lambda_2)(1, -1), \nu_1^2)$ and $\sigma_Y | X = 0.1$.

Again, it is evident from Table 4, that the method yields better prediction error when the sample size and number of discrete observations per sample in the response is high, favorable for the dense design paradigm for the predictor functions.

Table 4: Table showing the average prediction error as per (17) along with the standard error for Hilbertian objects as predictors and univariate distributions as responses, as described in Models I.8 under sparse and dense predictor design. The lowest number in a row is highlighted.

	I.8 (dense design)		I.8 (sparse design)	
n\m	50	100	50	100
200	0.334(0.051)	**0.270** (0.049)	0.483 (0.130)	**0.379** (0.124)
400	0.211 (0.031)	**0.176** (0.032)	0.410 (0.022)	**0.347** (0.022)

6.2 Scenario 2: Multivariate distribution-on-object regression

We now consider the scenario where both X and Y are two-dimensional random Gaussian distributions. The construction of the kernel κ_X is done using the sliced 2-Wasserstein distance, which is obtained by computing the average Wasserstein distance of the projected univariate distributions along randomly picked directions. To define formally,

Definition 6 (Sliced Wasserstein metric) *let μ_1 and μ_2 be two measures in $\mathcal{P}_p(M)$, the set of Borel probability measures on $(M, \mathcal{B}(M))$ that have finite $p-$th moment and is dominated by the Lebesgue measure on $\mathbb{R}^d$, with $M \subset of \mathbb{R}^d$, $d > 1$. Let S^{d-1} be the unit sphere in $\mathbb{R}^d$. For $\theta \in S^{d-1}$, let $T_\theta : \mathbb{R}^d \to \mathbb{R}$ be the linear transformation $x \mapsto \langle \theta, x \rangle$. Further, let $\mu_1 \circ T_\theta^{-1}$ and $\mu_2 \circ T_\theta^{-1}$ be the push-forward measures by the mapping T_θ. The sliced $p-$ Wasserstein distance between μ_1 and μ_2 is then defined by*

$$SW_p(\mu_1, \mu_2) = \left(\int_{S^{d-1}} W_p^p(\mu_1 \circ T_\theta^{-1}, \mu_2 \circ T_\theta^{-1}) d\theta \right)^{\frac{1}{p}}. \tag{19}$$

For $p = 2$, Kolouri et al. (2016) show that the square of sliced Wasserstein distance is conditionally negative definite and hence that the Gaussian RBF kernel defined as $\kappa_X(x, x') = \exp(-\gamma_X SW_2^2(x, x'))$ is a positive definite kernel.

We generate discrete observations for the predictor distributions $X_i, i = 1, \ldots, n$ given by $\{X_{ij}\}_{j=1}^m \overset{i.i.d.}{\sim} N(a_i(1,1)^T, b_i I_2)$, where $a_i \overset{i.i.d.}{\sim} N(0.5, 0.5^2)$ and $b_i \overset{i.i.d.}{\sim} Beta(2,3)$. For computing the Gram matrix associated with the multivariate predictor distribution supported on $M \subset \mathbb{R}^d$, $d > 1$ the sliced Wasserstein distance is estimated using a Monte Carlo method as

$$SW_2(\mu_{X_i}, \mu_{X_k}) \approx \left(\frac{1}{L} \sum_{l=1}^{L} W_2^2(\mu_{X_i} \circ T_\theta^{-1}, \mu_{X_k} \circ T_\theta^{-1}) \right)^{\frac{1}{2}},$$

where $\mu_{X_i} = \frac{1}{m} \sum_{j=1}^m \delta_{X_{ij}}$ is the empirical measure for the $i-$th sample, $i = 1, \ldots, n$, $\{\theta_l\}_{l=1}^L$ are i.i.d. samples drawn from the uniform distribution on $S^{d-1} \subset \mathbb{R}^d$. The approximation error depends on the number of Monte Carlo samples L. In our simulation settings, we set $L = 50$.

The random responses $Y = N(\mu_Y, \Sigma_Y)$, where $\mu_Y \in \mathbb{R}^2$ and $\Sigma_Y \in \mathbb{R}^{2\times 2}$ are then generated according to the following models.

Model-II.1 (Multivariate distributions as predictors): $\mu_Y|X \sim N(W_2(X, \mu_1)(1, 1)^\intercal, I_2)$ and $\Sigma_Y|X = \text{diag}(1, 1)$.

Model-II.2 (Multivariate distributions as predictors): $\mu_Y|X \sim N(W_2(X, \mu_1)(1, 1)^\intercal, I_2)$ and $\Sigma_Y|X = \Gamma\Lambda\Gamma^\intercal$, where $\Gamma = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}$, $\Lambda = \text{diag}(\lambda_1, \lambda_2)$ with $(\lambda_1, \lambda_2)|X \overset{i.i.d.}{\sim} tGamma(W_2^2(X, \mu_2), W_2(X, \mu_2), (0.2, 2))$, where μ_1 and μ_2 are two fixed measures defined by $\mu_1 = N((-1, 0)^\intercal, \text{diag}(1, 0.5))$ and $\mu_2 = N((0, 1)^\intercal, \text{diag}(0.5, 1))$, and $tGamma(\alpha, \beta, (r_1, r_2))$ is the truncated gamma distribution on range (r_1, r_2) with shape parameter α and rate parameter β. The Wasserstein distance between the bivariate Gaussian distributions is computed as per (18).

Since the dimension d of the random probability measures that we study here is more than 1, one does not have an analytic form for the barycenter and the optimization algorithms to obtain it are complex, in contrast to the case $d = 1$, where the quantile representation of Wasserstein distance leads to an explicit solution via the L^2 mean of the quantile functions. The computation of Wasserstein barycenters in multidimensional Euclidean space has been intensively studied (e.g., Rabin et al. (2012); Álvarez-Esteban et al. (2016); Dvurechenskii et al. (2018); Peyré et al. (2019), and one of the most popular methods utilize the Sinkhorn divergence (Cuturi, 2013), which is an entropy-regularized version of the Wasserstein distance that allows for computationally efficient solutions of the barycenter problem, however at the cost of introducing a bias, as is common for regularized estimation. Due to the gain in efficiency, we adopt this approach in our implementations using the R package *WSGeometry* (Heinemann and Bonneel, 2021).

Using the same choices for n, m, and the tuning parameters, we again split the data into a training and a test set, and use the training set to implement the proposed object regression method at the output predictor points to predict the response in the testing set. The whole process is repeated $B = 100$ times and the prediction error computed between the observed and predicted bi-variate distributional responses in the test set using the average Sliced Wasserstein distance between them, as per (19). The mean and standard error of this root mean prediction error is shown in Table 5, where a similar pattern of decreased RMPE for a combination of higher sample size and denser observation grid for the paired sample of distribution is noted.

Table 5: Table showing the Monte Carlo mean (standard error) prediction errors for Scenario II. The lowest number in a row is highlighted across different model settings.

	II.1		II.2	
n\m	50	100	50	100
200	0.620 (0.134)	**0.442** (0.130)	0.811 (0.200)	**0.693** (0.177)
400	0.319 (0.094)	**0.178** (0.092)	0.543 (0.160)	**0.329** (0.152)

6.3 Scenario 3: SPD matrix object-on-object regression

A common type of random object encountered in brain imaging studies is functional connectivity correlation matrices, which are positive semi-definite symmetric matrices. Let (Ω_Y, d_F) be the space of $r \times r$ symmetric positive definite (SPD) matrices endowed with Frobenius distance $d_F(Y_1, Y_2) = ||Y_1 - Y_2||_F$ as defined in (11) in Section 3.3. Two simulation scenarios are considered as follows.

Model-III.1 (Euclidean predictors): The real-valued predictors X_i are independently sampled from a $Beta(1/2, 2)$, while the SPD matrix responses Y_i conditional on X_i are generated according to the model $Y_i = \tilde{Y}_i\tilde{Y}_i^T$, with $\tilde{Y}_i|X_i = \mu(X_i) + [\Sigma(X_i)]^{-1/2}Z_i$, where for a fixed dimension r, the mean vector $\mu(x)$ has components $\mu_j(x) = b_j - 2(x - c_j)^2$, $j = 1, \ldots, r$. Here $b_j \sim U(2, 4)$ and $c_j \sim U(0, 1)$, and Z_i are sampled independently of X_i as a standard $r-$dimensional Gaussian random vector. the covariance $\Sigma(x)$ is formed by generating a $r \times r$ matrix A with independent $N(0, 0.5)$ random variables in each entry, then computing $S = 0.5(A + A^T)$. A second $r \times r$ matrix V is generated with elements drawn independently as $U(0, 0.5)$, from which $\theta = 0.5(V + V^T)$ is computed. Finally, with Exp denoting matrix exponentiation and $\odot$ the Hadamard product, we form $\Sigma(x) = (x + 2x^3)Exp[S \odot \sin(2\pi\theta(x + 0.1))]$.

Model-III.2 (SPD matrix objects as predictors): The predictors are now themselves SPD matrices. This is generated as the covariance matrix computed from a p-variate Gaussian random vector with independent components each with mean 0 and variance 1 for each sample. The predictors are projected down on a desired direction vector β whose each component $\beta_j \sim U(0, 1)$, $j = 1, \ldots, p$ to compute $\tilde{X}_i = X_i\beta$. Here, we choose $p = 5$. Now the response matrices are generated as before in Model

III.2 conditional on $\tilde{X}_i$.

In order to apply the proposed method, again the Gaussian RBF kernel given by $\kappa_X(x, x') = \exp(-\gamma_X d_F^2(x, x'))$ is taken to compute the Gram matrix in the predictor space, with the tuning parameter chosen as before. From a sample $(X_i, Y_i)_{i=1}^n$ the minimization in (13) can be reformulated by setting $\hat{f}_\oplus(x) = \frac{1}{n}\sum_{i=1}^n w_{in}(x)Y_i$ and computing the correlation matrix which is nearest to the matrix $\hat{f}_\oplus(x)$, which is implemented by the alternating projections algorithm via the *nearPD()* function in the *Matrix* R package.

We compare performances of the proposed method for a combination of sample size and the dimension of the response matrices given by n and r, respectively, by computing the Frobenius distance between the true and the predicted SPD matrix responses in the test set, using the model fit on the training set, as described before. The first two columns of Table 6 display the average prediction error across 100 replications of the above process. Our method fares better for increased sample size, while the dimension of the response SPD matrices is lower in both simulation scenarios.

Table 6: Table showing the Monte Carlo mean (standard error) estimation errors for Scenarios III and IV. The lowest number in a row is highlighted across different model settings.

	III.1		III.2		IV.1	
n\r	5	20	5	20	5	20
200	**0.119** (0.041)	0.275 (0.040)	**0.226** (0.130)	0.786 (0.110)	**0.161** (0.011)	0.235 (0.031)
400	**0.048** (0.037)	136 (0.035)	**0.127** (0.110)	0.502 (0.097)	**0.079** (0.012)	0.145 (0.029)

6.4 Scenario 4: Network object-on-object regression

Model-IV.1 (Euclidean predictors): Let $G = (V, E)$ be a simple (with no self-loops), weighted, undirected network with a set of nodes $V = \{v_1, \dots, v_r\}$ and a set of edge weights $E = \{w_{ij} : w_{ij} \geq 0,\ i, j = 1, \dots, r\}$, where $w_{ij} = 0$ indicates v_i and v_j are not connected and $w_{ij} > 0$ otherwise, with $w_{ij} < M$ for some $M > 0$. A network can be uniquely represented by its graph Laplacian $L = (l_{ij})$, where $l_{ij} = -w_{ij}$ if $i \neq j$ and $l_{ij} = \sum_{k \neq i} w_{ik}$ if $i = j$, for $i, j = 1, \ldots, r$. The space of graph Laplacians is given by $\mathcal{L}_r = \{L = (l_{ij}) : L = L^\intercal,\ L1_r = 0_r,\ -W \leq l_{ij} \leq 0 \ \text{ for some W } \geq 0 \text{ and } i \neq j\}$, where 1_r and 0_r are the r-vectors of ones and zeroes, respectively. Note that $\mathcal{L}_r$ is not a linear space, but a bounded, closed, and convex subset in $\mathbb{R}^{r^2}$ of dimension $r(r-1)/2$. Owing to the fact that $x^\intercal Lx \geq 0$ for all $x \in \mathbb{R}^r$ and $L \in \mathcal{L}_r$, it can be seen as a metric space of positive-semidefinite matrix objects, equipped with a suitable choice of metric such as the Frobenius or power metric.

To assess the performance of our proposed methods, we consider the space $(\mathcal{L}_r, d_F)$, where d_F is the Frobenius metric as per (11). The data generation mechanism is as follows. Denote the half vectorization excluding the diagonal of a symmetric and centered matrix by $vech$, with inverse operation $vech^{-1}$. By the symmetry and centrality, every graph Laplacian L is fully known by its upper (or lower) triangular part, which can then be vectorized into $vech(L)$, a vector of length $d = r(r-1)/2$. We construct the conditional distributions $F_{L|X}$ by assigning an independent beta distribution to each element of $vech(L)$. Specifically, a random sample $(\beta_1, \ldots, \beta_d)^\intercal$ is generated using beta distributions whose parameters depend on the scalar predictor X and vary under different simulation scenarios. The random response L is then generated conditional on X through an inverse half vectorization $vech^{-1}$ applied to $(\beta_1, \ldots, \beta_d)^\intercal$. The the true regression function $m(x)$ is defined as $m(x) = vech^{-1}(-x, \ldots, -x)$, $L = vech^{-1}(\beta_1, \ldots, \beta_d)^\intercal$, where $\beta_j \overset{i.i.d.}{\sim} Beta(X, 1-X)$. To ensure that the random response L generated in simulations resides in $\mathcal{L}_r$, the off-diagonal entries $-\beta_j$ $j = 1, \ldots, d$, need to be nonpositive and bounded below. Thus we choose $\beta_j \overset{i.i.d.}{\sim} Beta(X, 1-X)$. The scalar predictor X_i are randomly sampled from a $Unif(0,1)$ distribution to obtain the samples of pairs (X_i, L_i), $i = 1, \ldots, n$, setting $r = 5, 20$, and following the above procedure. The prediction error w.r.t the Frobenius metric is shown in the rightmost column of Table 6. The method performs better for higher n and lower r.

7 Data Analysis

In this application, we explore the relationship between the distribution of age-at-death and that of the mother's age at birth at a country level. Going beyond summary statistics such as mortality or fertility rate, viewing the entire distributions as samples of data is more informative and insightful to understanding the nature of human longevity and its dependence on relevant predictors. The data is obtained from the UN World Population Prospects 2019 Databases (`https://population.un.org`). For this analysis, we focus on $n = 194$ countries over the period of time $2015 - 2020$. The mortality data is available in the form of life tables over the age interval $[0, 110]$ (all in years) while the number of births is categorized by the mother's age every five years over the age bracket $[15, 50]$. We used bin widths equal to 5 years to construct the histograms for the mortality and fertility distributions, respectively, and proceeded to obtain the smooth densities by applying local linear regression using the *frechet* package at the country level. The domains of the age-at-death and mother's age-at-birth densities are $[0, 110]$ and $[15, 50]$ years, respectively. The densities are assumed to lie in the space of univariate distributions equipped with the Wasserstein metric (Ω_Y, d_W) in (10). Figure 2 shows the sample of densities observed.

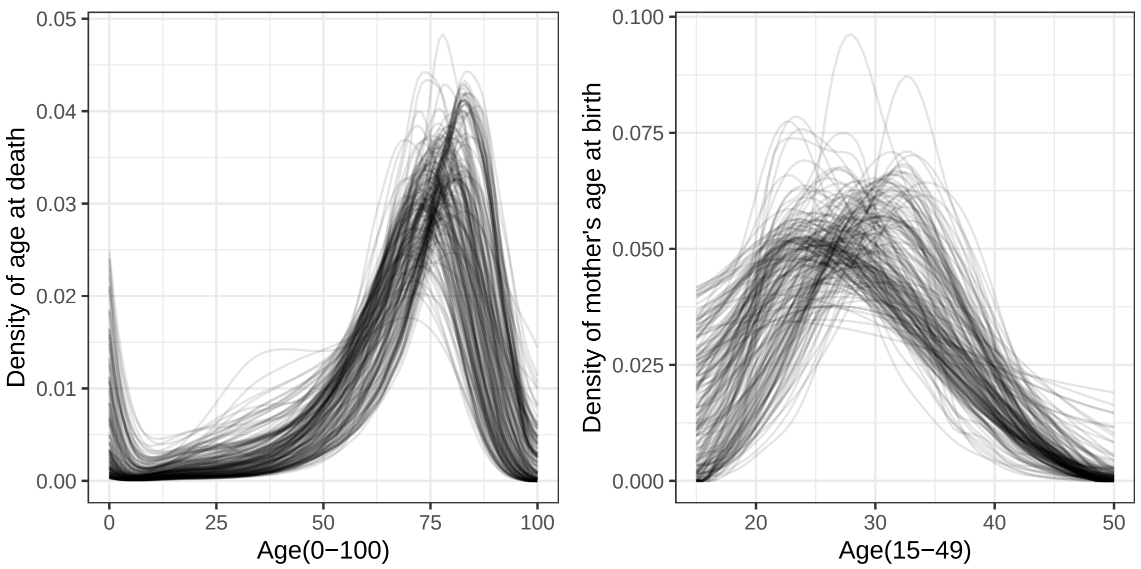


Figure 2: Visualization of distributional objects represented as densities of age at death and mother's age at birth for a sample of 194 countries.

We applied the proposed object-on-object regression method with age-at-death densities as responses and mother's age-at-birth densities as predictors to compare the evolution of mortality distributions among different countries. We show the leave-one-out prediction results together with the observed distributional predictors and responses in Figure 3 for a select few countries, which showcases different patterns of mortality change over changes in the predictor distribution. The Wasserstein distance between the observed and predicted distributions is also shown. Specifically, we selected the countries Bangladesh, Argentina, the USA, Japan, the UK, and Norway, ordered by the lowest to the highest value of the mode of the mother's age-at-death densities. Both the observed and predicted age-at-death densities across the panels from left to right are seen to be more right-shifted, indicating increased longevity corresponding to a higher age at birth for the mother. Further, for Japan, Norway, and the USA, the rightward mortality shift is seen to be more expressed than suggested by the prediction, indicating that longevity extension is more than anticipated, while the mortality distribution for the UK seems to shift to the right at a slower pace than predicted, leading to a relatively larger WD with a value of 0.8 between the observed and predicted response. In contrast, the regression fit for Argentina and Bangladesh are quite accurate.

The effect of the mother's age-at-birth is elicited in Figure 4a, where the model is fitted for varying levels of the mode of the predictor distribution. The fitted densities are color coded such that blue to red indicates smaller to larger values of the mode of the age-at-birth densities. We find that lower age-at-birth of the mother is associated with left-shifted age-at-death distributions in general, while modes at higher age-at-birth correspond to a shift of the mode of the age-at-death toward the right. Child mortality has an association with both low and high values of age-at-birth for the mother, which concurs with the observations made earlier.

The fit of the model is further demonstrated by computing the estimation error by virtue of the residual map for the i-th subject, $T_i : \Omega_Y \to \Omega_Y$, defined as the optimal transport map $T_i = \nu_i \# \hat{\nu}_i$, that pushes forward the observed response ν_i to the fitted value $\hat{\nu}_i$. Using the theory of optimal transport for univariate distributions (Villani et al., 2009), this map can be explicitly computed as $T_i = Q_{\hat{\nu}_i} \circ F_{\nu_i}$, where $Q_{\hat{\nu}_i}$ and F_{ν_i} are, respectively, the quantile function and the CDF of the distributions $\hat{\nu}_i$ and ν_i. Using these residual maps one can obtain an analog of the "residual plot" in the classical regression case, compared to the identity map. Looking at the deviation from the identity map one can see in which parts the support of the distributions, the model provides a good fit, and where less so and the departure from the identity can serve as a diagnostic tool for the validity of the model. Note that, contrary to classical regression, where the residuals add up to zero by construction, the residual maps are not constrained to have a mean equal to the identity.

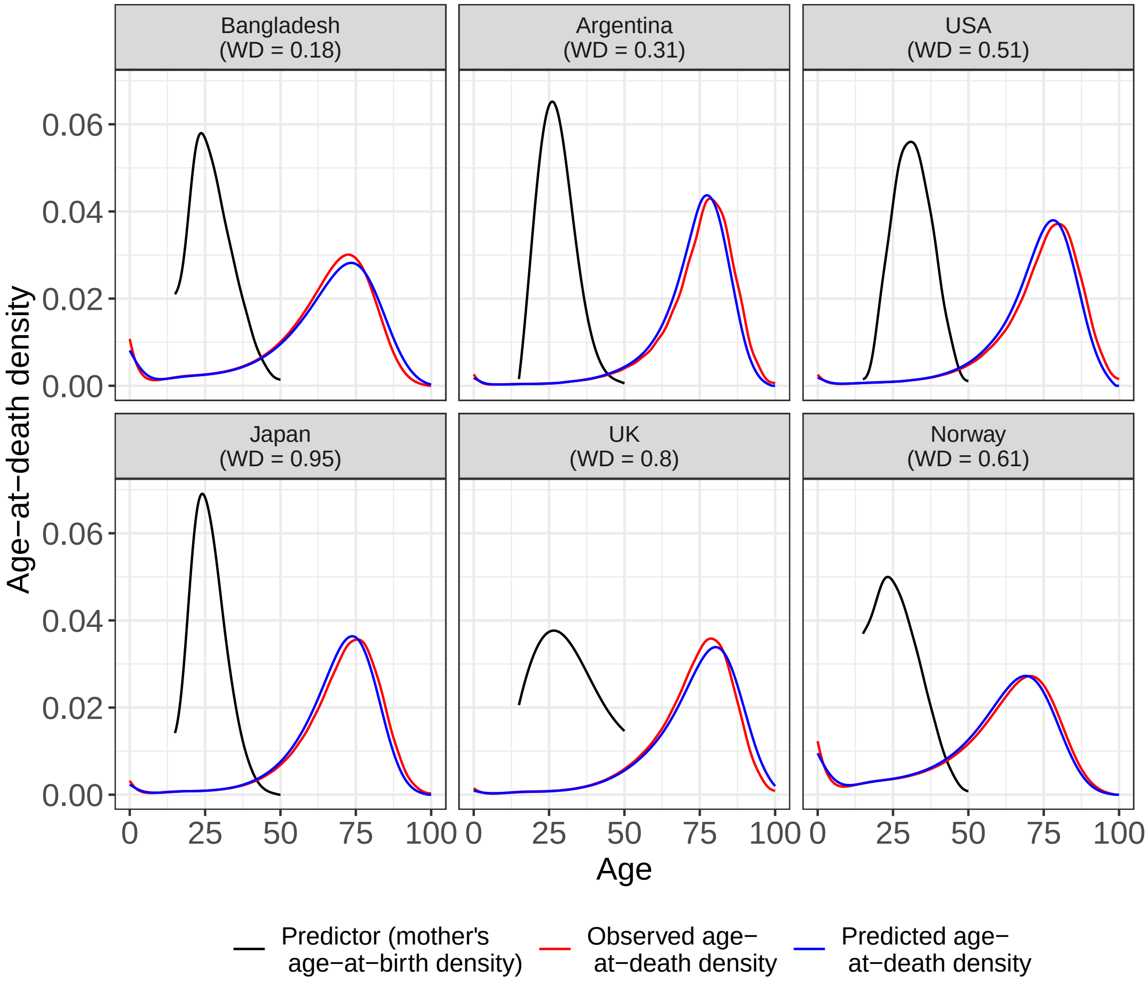
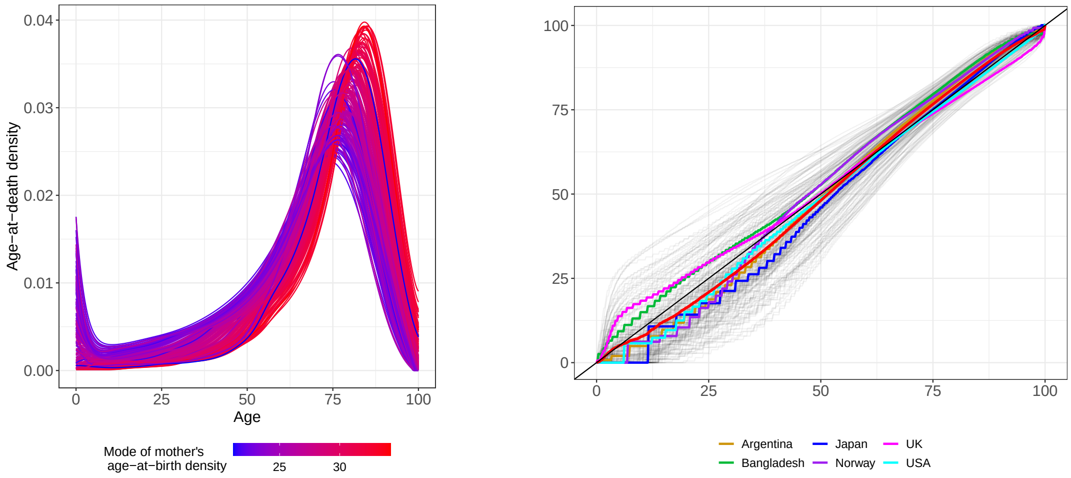


Figure 3: Visualization of distributional objects represented as densities of age at death and mother's age at birth for a sample of 194 countries.

The residual maps computed for each of the 194 countries are plotted in Figure 4b. One can see that the pointwise variability is much more prominent for younger ages and decreases for progressively older ages, indicating many other plausible factors affecting mortality at younger ages. The identity map is overlaid in black. The mean of residuals plotted in red lies very close to the identity map, which provides evidence in support of the validity of our model. The residual maps of the specific countries considered in Figure 3 are highlighted. Similar patterns of right-shifted distributions, especially near the age-at-death $[15, 40]$ years are observed for the highlighted countries. For example, while the evolution of the mortality distributions for Japan and



(a) The changes in the density of the age-at-death distribution as the mode of the distribution of the mother's age-at-birth ranges from low (blue) to high (red) are displayed.

(b) Residual maps corresponding to $n = 197$ countries are plotted in gray, with specific countries highlighted. The identity map and the average residual map are overlaid in black and red, respectively.

the USA can be viewed as mainly a rightward shift over calendar years, this is not the case for the UK, where compared with the fitted response, the actual rightward shift of the mortality distribution seems to be accelerated for those above age 65, and decelerated for those below age 65.

To evaluate the out-of-sample prediction performance of the method, we randomly split the dataset into a training set and a test set, and use the fits obtained from the training set to predict the responses to the test set using only the predictors present in the test set. As a measure of the efficacy of the fitted model, we compute the root mean squared prediction error (RMPE) as the Wasserstein discrepancy between the observed and the predicted distributions in the test set. We repeat the process 100 times to obtain the average RMPE, which comes out low (0.693 with a standard error of 0.151), supporting the efficacy of the model.

8 Discussion

In this paper, we have proposed a nonlinear global object-on-object regression method based on the intrinsic geometry of the metric space where the responses reside coupled with suitable linear operators defined via the reproducing kernel Hilbert space on the predictor space. This contribution is one of the first to model the regression relationship between metric-valued objects, beyond scalar-or-vector-valued predictors. Further, the lack of linearity in an abstract metric space can result in a significant difference between conditional and globally linear Fréchet means proposed by Petersen and Müller (2019), leading to questions about the validity of such globally linear models. To address this, we introduce a novel method extending global linear regression to a general global non-linear object regression. We employ generalized weak conditional Fréchet moments as a way to link random object data analysis to non-linear global RKHS regression models, allowing for arbitrary non-linear functions beyond linear or polynomial regression.

The concept of weak Fréchet moments can be easily extended to Fréchet median or as a minimizer of Huber loss, by substituting $E[d^2_Y(Y,\cdot)\,\vdots\,X]$ by $E[\rho_Y(Y,\cdot)\,\vdots\,X]$, for any appropriate convex loss function ρ_Y in the metric space (Ω_Y, d_Y), depending on the context and interpretation of the problem. This calls for future research. The selection of a suitable metric is also an open problem.

Further, the rate of convergence of the proposed estimator is derived as $\approx n^{-1/4}$, which entails from the work of Li and Song (2017). This rate can be further improved using a suitable rate carried out from the RKHS regression literature.

A Technical assumptions for M-estimators

Assumption 11 *The weak conditional Fréchet means $f_\oplus(x)$ and $\hat{f}_\oplus(x)$ exist and are unique, the latter almost surely. Further, the minimizer at the population level is well separated. i.e., for any $\epsilon > 0$,*

$$\inf_{d_Y(y,f_\oplus(x))>\epsilon} J(y,x) - J(f_\oplus(x),x) > 0.$$

Assumption 12 *Let $B_\delta(f_\oplus(x)) \subset \Omega_Y$ be the ball of radius δ, centered at $f_\oplus(x)$ and $N(\epsilon, B_\delta(f_\oplus(x)), d_Y)$ be its covering number using balls of radius ϵ. Then the entropy integral is computed from the covering number given by*

$$J = J(\delta) := \int_0^1 \sqrt{1 + \log N(\delta\epsilon, B_\delta(f_\oplus(x)), d_Y)}d\epsilon = O(1) \text{ as } \delta \to 0.$$

Assumption 13 *There exist constants $\eta > 0, C > 0$, and $\beta > 1$, possibly depending on $x \in (\Omega_X, d_X)$, such that*

$$J(y,x) - J(f_\oplus(x),x) \geq Cd^\beta(y, f_\oplus(x)),$$

for any small neighborhood $d_Y(y, f_\oplus(x)) < \eta$.

Assumption 11 is commonly used to establish the consistency of an M-estimator; see Chapter 3.2 in Van der Vaart and Wellner (2000). In particular, it ensures that weak convergence of the empirical process $\tilde{J}_n$ to the population process J, which in turn implies convergence of their minimizers. The conditions on the covering number in Assumption 12 and curvature in Assumption 13 arise from empirical process theory and control the behavior of $\tilde{J}_n - J$ near the minimum, which is necessary to obtain rates of convergence. These assumptions are again satisfied for many random objects of interest the common examples of random objects such as distributions, covariance matrices, networks, and so on (see Propositions 1-3 of Petersen and Müller (2019)).

References

Álvarez-Esteban, P. C., Del Barrio, E., Cuesta-Albertos, J., and Matrán, C. (2016). A fixed-point approach to barycenters in wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762.

Bhattacharjee, S. and Müller, H.-G. (2021). Single index fr\'echet regression. *arXiv preprint arXiv:2108.05437*.

Billera, L. J., Holmes, S. P., and Vogtmann, K. (2001). Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics*, 27(4):733–767.

Chen, Y., Gajardo, A., Fan, J., Zhong, Q., Dubey, P., Han, K., Bhattacharjee, S., and Müller, H. (2020). frechet: statistical analysis for random objects and non-euclidean data. *R package version 0.2. 0*.

Chen, Y., Lin, Z., and Müller, H.-G. (2021). Wasserstein regression. *Journal of the American Statistical Association*, pages 1–14.

Chen, Z., Bao, Y., Li, H., and Spencer Jr, B. F. (2019). Lqd-rkhs-based distribution-to-distribution regression methodology for restoring the probability distributions of missing shm data. *Mechanical Systems and Signal Processing*, 121:655–674.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

Delicado, P. and Vieu, P. (2017). Choosing the most relevant level sets for depicting a sample of densities. *Computational Statistics*, 32(3):1083–1113.

Di Marzio, M., Panzera, A., and Taylor, C. C. (2014). Nonparametric regression for spherical data. *Journal of the American Statistical Association*, 109(506):748–763.

Dryden, I. L., Koloydenko, A., and Zhou, D. (2009). Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *Annals of Applied Statistics*, 3:1102–1123.

Dryden, I. L., Pennec, X., and Peyrat, J.-M. (2010). Power euclidean metrics for covariance matrices with application to diffusion tensor imaging. *arXiv preprint arXiv:1009.3045*.

Dubey, P. and Müller, H.-G. (2019). Fréchet analysis of variance for random objects. *Biometrika*, 106(4):803–821.

Dvurechenskii, P., Dvinskikh, D., Gasnikov, A., Uribe, C., and Nedich, A. (2018). Decentralize and randomize: Faster algorithm for wasserstein barycenters. *Advances in Neural Information Processing Systems*, 31.

Fréchet, M. R. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. *Annales de l'institut Henri Poincaré*, 10(4):215–310.

Fukumizu, K., Bach, F. R., and Gretton, A. (2007). Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8(2).

Ghodrati, L. and Panaretos, V. M. (2022). Distribution-on-distribution regression via optimal transport maps. *Biometrika*, 109(4):957–974.

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223.

Hein, M. (2009). Robust nonparametric regression with metric-space valued output. *Advances in neural information processing systems*, 22.

Heinemann, F. and Bonneel, N. (2021). Wsgeometry: compute wasserstein barycenters, geodesics, pca and distances. *R package version 0.1. 0*.

Horn, R. A. and Johnson, C. R. (2012). *Matrix analysis*. Cambridge university press.

Kolouri, S., Zou, Y., and Rohde, G. K. (2016). Sliced wasserstein kernels for probability distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5258–5267.

Le Gouic, T. and Loubes, J.-M. (2017). Existence and consistency of wasserstein barycenters. *Probability Theory and Related Fields*, 168(3):901–917.

Lee, K.-Y., Li, B., and Chiaromonte, F. (2013). A general theory for nonlinear sufficient dimension reduction: Formulation and estimation.

Lee, K.-Y., Li, B., and Zhao, H. (2016). Variable selection via additive conditional independence. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 1037–1055.

Li, B. (2018). *Sufficient dimension reduction: Methods and applications with R*. CRC Press.

Li, B. and Song, J. (2017). Nonlinear sufficient dimension reduction for functional data.

Li, B. and Song, J. (2022). Dimension reduction for functional data based on weak conditional moments. *The Annals of Statistics*, 50(1):107–128.

Lin, Z. (2019). Riemannian geometry of symmetric positive definite matrices via cholesky decomposition. *SIAM Journal on Matrix Analysis and Applications*, 40(4):1353–1370.

Petersen, A. and Müller, H.-G. (2019). Fréchet regression for random objects with Euclidean predictors. *The Annals of Statistics*, 47(2):691–719.

Peyré, G., Cuturi, M., et al. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Rabin, J., Peyré, G., Delon, J., and Bernot, M. (2012). Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision: Third International Conference, SSVM 2011, Ein-Gedi, Israel, May 29–June 2, 2011, Revised Selected Papers 3*, pages 435–446. Springer.

Sang, P. and Li, B. (2022). Nonlinear function-on-function regression by rkhs. *arXiv preprint arXiv:2207.08211*.

Sturm, K.-T. (2003). Probability measures on metric spaces of nonpositive. *Heat Kernels and Analysis on Manifolds, Graphs, and Metric Spaces: Lecture Notes from a Quarter Program on Heat Kernels, Random Walks, and Analysis on Manifolds and Graphs: April 16-July 13, 2002, Emile Borel Centre of the Henri Poincaré Institute, Paris, France*, 338:357.

Van der Vaart, A. and Wellner, J. (2000). *Weak Convergence and Empirical Processes: with Applications to Statistics (Springer Series in Statistics)*. Springer, corrected edition.

Villani, C. et al. (2009). *Optimal transport: old and new*, volume 338. Springer.

Weidmann, J. (2012). *Linear operators in Hilbert spaces*, volume 68. Springer Science & Business Media.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005). Functional linear regression analysis for longitudinal data.

Zhang, Q., Li, B., and Xue, L. (2022). Nonlinear sufficient dimension reduction for distribution-on-distribution regression. *arXiv preprint arXiv:2207.04613*.

Zhang, Q., Xue, L., and Li, B. (2021). Dimension reduction and data visualization for fréchet regression. *arXiv preprint arXiv:2110.00467*.