RANDOM FORESTS WEIGHTED LOCAL FRÉCHET REGRESSION WITH THEORETICAL GUARANTEE

BY RUI QIU¹, ZHOU YU², RUOQING ZHU³

¹School of Statistics, East China Normal University, 52204404010@stu.ecnu.edu.cn
²School of Statistics, East China Normal University, zyu@stat.ecnu.edu.cn
³Department of Statistics, University of Illinois at Urbana-Champaign, rqzhu@illinois.edu

Statistical analysis is increasingly confronted with complex data from general metric spaces, such as symmetric positive definite matrix-valued data and probability distribution functions. [47] and [17] establish a general paradigm of Fréchet regression with complex metric space valued responses and Euclidean predictors. However, their proposed local Fréchet regression approach involves nonparametric kernel smoothing and suffers from the curse of dimensionality. To address this issue, we in this paper propose a novel random forests weighted local Fréchet regression paradigm. The main mechanism of our approach relies on the adaptive kernels generated by random forests. Our first method utilizes these weights as the local average to solve the Fréchet mean, while the second method performs local linear Fréchet regression, making both methods locally adaptive. Our proposals significantly improve existing Fréchet regression methods. Based on the theory of infinite order U-processes and infinite order $M_m$-estimator, we establish the consistency, rate of convergence, and asymptotic normality for our proposed random forests weighted Fréchet regression estimator, which covers the current large sample theory of random forests with Euclidean responses as a special case. Numerical studies show the superiority of our proposed two methods for Fréchet regression with several commonly encountered types of responses such as probability distribution functions, symmetric positive definite matrices, and sphere data. The practical merits of our proposals are also demonstrated through the application to the human mortality distribution data.

1. Introduction. In recent years, non-Euclidean statistical analysis has received increasing attention due to the demands from modern applications, such as the covariance or correlation matrices for the functional brain connectivity in neuroscience and probability distributions in CT hematoma density data. To this end, [29] proposed nonparametric Nadaraya-Watson estimators for response variables being random objects, which are random elements in general metric spaces that by default do not have a vector space structure. [47] further introduced the general framework of Fréchet regression and established the methodology and theory for both global and local Fréchet regression analysis of complex random objects. [17] continued to derive the uniform convergence rate of local Fréchet regression. [39, 58] considered nonparametric modeling with responses being symmetric positive-definite matrices, which is a specific type of random objects. [48] provided a summary and theoretical analysis of nonparametric Fréchet regression methods. These methods certainly build a concrete foundation of statistical modeling with non-Euclidean responses. However, methods mentioned above rely on nonparametric kernel smoothing and thus can be problematic when the dimension of predictor $X$ is relatively high, limiting the scope of Fréchet regression in real applications [4, 57, 59].

Keywords and phrases: Global Fréchet regression, Local Fréchet regression, Mortality distributions, Metric space, Random forests, Random Objects.
Random forests, as pioneered by Leo Breiman [12], are a popular and promising tool for high dimensional statistical learning for Euclidean data. Random Forests are an ensemble model that combines the strength of multiple randomized trees. The tree is itself a nonparametric algorithm based on recursive partitions of the input space, which allows for flexible modeling of interactions in high dimensions. The ensemble learning procedure of random forests further reduces the tree’s variance by averaging multiple trees, thereby significantly improving the performance. Random forests demonstrate substantial gains in regression and classification compared to classical statistical methods and enjoy a wide range of applications. Moreover, the trees can be generated parallelly, making random forests more attractive computationally.

Research into random forests has gained considerable momentum in recent years due to their tremendous popularity. Many variant algorithms of random forests have been proposed, including weighted random forests [56], quantile random forests [40, 49], online random forests [19, 53], Mondrian forests [35, 43, 44], reinforcement learning forests [60], survival forests [30], generalized random forests [3] and many others. In the meantime, a series of theoretical results of the random forest has been established. [9] first proved the consistency of purely random forests for classification. For regression problems, [2, 26] further made a complete analysis of the variance and bias of purely random forests. [8, 25] established the consistency and convergence rate of the centered random forests for regression and classification, respectively. [21] provided the convergence rate of $q$-quantile random forests. In particular, [34] improved the rate of the median random forests. [51] proved the $L^2$ consistency of Breiman’s original random forests for the first time under the assumption of additive model structure. [41] formulated random forests as infinite order incomplete U-statistics and studied their asymptotic normality. [55] further established the central limit theorem for random forests based on honest tree construction.

However, most methodology developments, theoretical investigations, and real applications of random forests focus on classical Euclidean responses $Y$ and predictors $X$. It is then of great interest to generalize the random forests with metric space-valued responses, which is expected to work better than existing Fréchet regression methods when the predictor dimension is moderately large. To this end, [14] proposed Fréchet trees and Fréchet random forests based on CART trees and Breiman’s random forests. However, they only provided the consistency result for a single Fréchet tree but not for Fréchet random forests. On the other hand, recent developments [10, 24, 36, 40, 50] reveal the fact that random forests implicitly construct a kernel-type weighting function. This proliferation of work points toward a general synthesis between the core of nonparametric kernel smoothing and the ability to encompass data-adaptive weighting by random forests. Taking a step forward, we in this paper propose a novel random forest weighted local Fréchet regression paradigm with metric-space valued response variables and scalar predictor variables.

Our major contributions are summarized from the following three perspectives. First, to the best of our knowledge, this is the first attempt to adopt random forests as an adaptive kernel for Fréchet regression. Our approach is designed for metric space-valued responses and includes classical random forests with Euclidean responses as a special case. Different from [14], our proposal articulates a new formulation of Fréchet regression based on random forests that has an intrinsic relationship with classical nonparametric kernel regression. Second, compared to [14], our method is more concise, which allows us to take a substantial step towards the asymptotic theory of local Fréchet regression based on random forests rigorously. We systematically investigate the theoretical properties, including consistency, rate of convergence, and asymptotic normality. The consistency and rate of convergence are derived based on the theory of infinite order $U$-statistic and $U$-process. To study the asymptotic normality, we extend the current theory of finite order $M_m$-estimator to infinite order $M_{m_o}$-estimator. The new technical tools developed to establish the central limit theorem of infinite
order $M_{m_n}$-estimator can be of independent interest. And our asymptotic normality result also covers that of random forests with Euclidean responses [55] as a special case. Last but not least, the perspective from which we view the random forests facilitates the generalization of our method to the random forests weighted local linear Fréchet regression. The random forests weighted Fréchet regression and local linear Fréchet regression collectively make up a coherent system and new framework for Fréchet regression.

The rest of the paper is organized as follows. In Section 2, we give an overview of Fréchet regression and introduce the random forests weighted Fréchet regression (RFWFR) method. In Section 3, we establish the consistency and develop the asymptotic convergence rate and asymptotic normality of RFWFR. In Section 4, we present the random forests weighted local linear Fréchet regression (RFWLLFR) approach as the generalization of RFWFR and confirm its consistency in estimation. In Section 5, we present the random forests weighted local linear Fréchet regression (RFWLLFR) approach as the generalization of RFWFR and confirm its consistency in estimation. In Section 6, we apply our methods to the Human Mortality Data, where the responses are age-at-death distributions. Section 7 concludes the paper with some discussions. All proofs are relegated to the supplement material.

2. Random Forests Weighted Fréchet Regression.

2.1. Preliminaries. As our main focus is Fréchet regression with metric space valued responses, we first briefly review Fréchet regression. Then we introduce how to construct a Fréchet tree [14], which generalizes the classical regression tree to the case with metric space valued responses.

2.1.1. Fréchet Regression. Let $(\Omega, d)$ be a metric space equipped with a specific metric $d$. Let $\mathcal{R}^p$ be the $p$-dimensional Euclidean space. Fréchet regression aims at capturing the relationship between $Y \in \Omega$ and $X \in \mathcal{R}^p$. For general metric space, the conventional concepts of mean and variance can be generalized to the Fréchet version [23] as

$$
\mu_Y = \arg\min_{y \in \Omega} E \left[ d^2(Y, y) \right], \quad V_Y = E \left[ d^2(Y, \mu_Y) \right].
$$

Note that $\mu_Y$ and $V_Y$ coincide with the classical mean and variance when $\Omega = \mathcal{R}$. Moreover, when $\Omega = \mathcal{R}$, the central role of classical regression is to estimate the conditional expectation

$$
m(x) = E[Y \mid X = x] = \arg\min_{y \in \mathcal{R}} E \left[ (Y - y)^2 \mid X = x \right].
$$

Replacing the Euclidean distance with the intrinsic metric $d$ of $\Omega$, Fréchet conditional mean [47] can then be defined as

$$
m_{\oplus}(x) = \arg\min_{y \in \Omega} M_{\oplus}(x, y) = \arg\min_{y \in \Omega} E \left[ d^2(Y, y) \mid X = x \right].
$$

Let $F$ be the joint distribution of $(X, Y)$ defined on $\mathcal{R}^p \times \Omega$. Given an i.i.d training sample $D_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$ with $(X_i, Y_i) \sim F$, the goal of Fréchet regression is to estimate $m_{\oplus}(x)$ in the sample level. For this purpose, [29] generalized the Nadaraya-Watson regression to the Fréchet version as

$$
\hat{m}^{NW}_{\oplus}(x) = \arg\min_{y \in \Omega} \frac{1}{n} \sum_{i=1}^n K_h \left( X_i - x \right) d^2(Y_i, y).
$$

where $K$ is a smoothing kernel such as the Epanechnikov kernel or Gaussian Kernel and $h$ is a bandwidth, with $K_h(\cdot) = h^{-1} K(\cdot/h)$. When $\Omega = \mathcal{R}$, we can verify that

$$
\hat{m}^{NW}_{\oplus}(x) = \frac{\sum_{i=1}^n K_h \left( X_i - x \right) Y_i}{\sum_{i=1}^n K_h \left( X_i - x \right)}.
$$
which is exactly the classical Nadaraya-Watson estimator. [47] recharacterized the standard multiple linear regression and local linear regression as a function of weighted Fréchet means, and proposed global Fréchet regression and local Fréchet regression as follows

\[
\hat{m}_\Omega(x) = \arg \min_{y \in \Omega} \frac{1}{n} \sum_{i=1}^{n} s_{in}(x) d^2(Y_i, y).
\]

For global Fréchet regression,

\[
s_{in}(x) = 1 + (X_i - \bar{X})^T \Sigma^{-1}(x - \bar{X}),
\]

where \(\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i\) and \(\Sigma = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^T\). For Euclidean responses with \(\Omega = \mathcal{R}\), \(\hat{m}_\Omega(x)\) is just the classical ordinary least square estimator.

For local Fréchet regression with \(p = 1\),

\[
s_{in}(x) = \frac{1}{\sigma_0^2} K_h(X_i - x) [\hat{\mu}_2 - \hat{\mu}_1 (X_i - x)],
\]

where \(\hat{\mu}_j = \frac{1}{n} \sum_{i=1}^{n} K_h(X_i - x)(X_i - x)^j\), \(\sigma_0^2 = \hat{\sigma}_0^2 = \hat{\mu}_0^2 - \hat{\mu}_1^2\). For Euclidean responses with \(\Omega = \mathcal{R}\), local Fréchet regression coincides with the classical local linear nonparametric regression.

We see that Nadaraya-Watson Fréchet regression and local Fréchet regression involve kernel weighting function in the estimation procedure, which limit their applications when \(p \geq 3\). To address this issue, we can borrow the strength of random forests to generate a more powerful weighting function for moderately large \(p\).

2.1.2. Fréchet Trees. A regression tree \(T\) splits the input space recursively from the root node (the entire input space). At each split, the parent node is divided into two child nodes along a certain feature direction and a certain cutoff point, which are decided by a specific splitting criterion. After many splits, the child node becomes small enough to form a leaf node, and the sample data points within the leaf node are used to estimate the conditional mean. In essence, the tree is also a localized method for nonparametric regression. It divides the entire sample space into small rectangles of different sizes. The depth of the tree (or the terminal node size) is a crucial factor that affects the performance. The deeper the tree grows, the smaller the bias of the model is, but the larger the variance is.

We can adopt the CART-split criterion, which is the most commonly used criterion in application. The difference is that the impurity of data from a general metric space is no longer measured by the variance under the Euclidean distance. Instead, we use the Fréchet variance [14]. A split on an internal node \(A\) can be represented by a pair \((j, c), j \in \{1, \ldots, p\}\), indicating that \(A\) is split at position \(c\) along the direction of feature \(j\). We select the optimal \((j^*_n, c^*_n)\) to decrease the sample Fréchet variance as much as possible, so that the sample points in the same child node have great similarity. The split criterion is

\[
\mathcal{L}_n(j, c) = \frac{1}{N_n(A)} \left( \sum_{i : X_i \in A} d(X_i, \bar{Y}_A) - \sum_{i : X_i \in A_{j,l}} d(X_i, \bar{Y}_{A_{j,l}}) - \sum_{i : X_i \in A_{j,r}} d(X_i, \bar{Y}_{A_{j,r}}) \right)
\]

where \(A_{j,l} = \{x \in A : x^{(j)} < c\}, A_{j,r} = \{x \in A : x^{(j)} \geq c\}\), \(N_n(A)\) is the number of observations falling into the node \(A\), and \(\bar{Y}_A = \arg \min_{y \in \Omega} \sum_{i : X_i \in A} d^2(Y_i, y)\), i.e., the sample Fréchet mean of \(Y_i\)’s associated to the observations belonging to the node \(A\). \(\bar{Y}_{A_{j,l}}\) and \(\bar{Y}_{A_{j,r}}\) are defined similarly. Then the optimal split pair is decided by

\[
(j^*_n, c^*_n) = \arg \max_{j,c} \mathcal{L}_n(j, c).
\]
When the construction of the tree $T$ is completed, let $L_T(x)$ be the leaf node of $T$ containing $X = x$. Then $m_{\oplus}(x)$ can be possibly estimated by the sample Fréchet mean of $Y_i$'s corresponding to the observations belonging to $L_T(x)$, that is

$$\hat{m}_{\oplus}(x) = \arg\min_{y \in \Omega} \sum_{i : X_i \in L_T(x)} d^2(Y_i, y).$$

And $\hat{m}_{\oplus}(x)$ is called the Fréchet tree estimator.

2.2. Random Forests Weighted Fréchet Regression. A single tree model may suffer from large bias or large variance depending on the tuning. To improve the predictive performance, we can aggregate individual trees to form a random forest. The prediction error of the random forest is related to the number of trees and the correlation among different trees. We usually introduce more random effects in a tree-building process. For example, a subset of features is randomly selected before each split, and the split direction is designed based on these randomly selected features only. In this way, we reduce the correlation between trees and further improve the prediction accuracy of random forests. Here, we denote $\xi$ as the collection of these random effects.

We first consider the classical random forests with Euclidean responses ($\Omega = \mathbb{R}$). And each tree is trained on a subsample $D_n^b = \left\{(X_{i_{b,1}}, Y_{i_{b,1}}), (X_{i_{b,2}}, Y_{i_{b,2}}), \ldots, (X_{i_{b,s_n}}, Y_{i_{b,s_n}})\right\}$ of size $s_n$ of the training data set $D_n$, with $1 \leq i_{b,1} < i_{b,2} < \ldots < i_{b,s_n} \leq n$. We may assume that $s_n \to +\infty$ and $s_n/n \to 0$ as the sample size $n$ tends to infinity. Resampling is done here without replacement for the convenience of theoretical analysis. The $b$-th tree $T_b$ gives an estimator of $m(x)$

$$T_b(x; D_n^b, \xi_b) = \frac{1}{N(L_b(x))} \sum_{i : X_i \in L_b(x)} Y_i,$$

where $N(L_b(x))$ is the number of samples in $L_b(x)$, which is the leaf node contains $x$ of the $b$-th tree.

For the random forest constructed based on a total number of $B$ randomized regression trees, the estimator of $m(x)$ becomes

$$RF(x) = \frac{1}{B} \sum_{b=1}^B T_b(x; D_n^b, \xi_b) = \frac{1}{B} \sum_{b=1}^B \frac{1}{N(L_b(x))} \sum_{i : X_i \in L_b(x)} Y_i. \quad (2)$$

In fact, we can view the random forests from another perspective and regard it as a weighted average of the training responses,

$$RF(x) = \frac{1}{B} \sum_{b=1}^B \sum_{i=1}^n \frac{1}{N(L_b(x))} \left\{X_i \in L_b(x)\right\} Y_i = \sum_{i=1}^n \left(\frac{1}{B} \sum_{b=1}^B \frac{1}{N(L_b(x))} \left\{X_i \in L_b(x)\right\}\right) Y_i = \sum_{i=1}^n \alpha_i(x) Y_i, \quad (3)$$

where $\alpha_i(x) = \frac{1}{B} \sum_{b=1}^B \frac{1\{X_i \in L_b(x)\}}{N(L_b(x))}$ is defined as the random forests weight.

We now generalize the Euclidean random forests to the Fréchet random forests when $\Omega$ is a general metric space. We first rewrite the explicit expression of the random forests estimator.
as the implicit minimizer of certain objective functions. There are two generalization methods
corresponding to the two views (2) and (3) of random forests above. We can rewrite (2) as
\[
RF(x) = \arg\min_{y \in \mathcal{R}} \frac{1}{B} \sum_{b=1}^{B} \left[ \arg\min_{y \in \mathcal{R}} \left( \frac{1}{N(L_b(x))} \sum_{i : X_i \in L_b(x)} (Y_i - y)^2 \right) - y \right]^{2}.
\]

Then the first generalization to metric space valued response is simply replacing the Eu-
clidean distance by the metric \(d\) of \(\Omega\), that is,
\[
RF^{(1)}(x) = \arg\min_{y \in \Omega} \frac{1}{B} \sum_{b=1}^{B} d\left( \arg\min_{y \in \Omega} \frac{1}{N(L_b(x))} \sum_{i : X_i \in L_b(x)} d^2(Y_i, y), y \right).
\]

Alternatively, we can start from (3) and rewrite the random forests estimator as
\[
RF(x) = \arg\min_{y \in \mathcal{R}} \sum_{i=1}^{n} \alpha_i(x)(Y_i - y)^2.
\]

We propose the second generalization of random forests for metric space valued responses as
\[
RF^{(2)}(x) = \arg\min_{y \in \Omega} \sum_{i=1}^{n} \alpha_i(x)d^2(Y_i, y).
\]

In our second generalization (5), random forests produce the weighting function for Fréchet
regression but do not participate in the prediction. To this end, (5) is essentially a local con-
tant estimator based on Fréchet random forests kernel. We call it random forests weighted
Fréchet regression (RFWFR). It is a nonparametric statistical tool to analyze metric space
valued or more general non-Euclidean data. Our proposed RFWFR is expected to outper-
form Nadaraya-Watson Fréchet regression estimator [29] and the local Fréchet regression
estimator [47] for the following two reasons. Firstly, the random forests kernel can handle
large \(p\). Secondly, the random forests kernel is adaptive in the sense that it incorporates both
the information of \(X\) and \(Y\), while the standard smoothing kernel only uses the information
of \(X\).

The first generalization (4) is actually the Fréchet random forests proposed by [14]. The
idea behind this is to average the results of each tree. However, our proposed second general-
ization RFWFR (5) looks more concise because it involves only one “argmin”. To acquire
the random forests weighting function \(\alpha_i(x)\), we still need to construct all trees. But we do
not need to calculate the sample Fréchet mean of the leaf nodes of the trees, as the prediction
of each tree is not needed. It is worth noting that when \(\Omega = \mathcal{R}\), (4) and (5) are equivalent.
However, for general metric space \((\Omega, d)\), (4) and (5) may not be the same. Based on the
second type of generalization, we can rigorously establish the asymptotic properties of our
proposed RFWFR. And these results fill the theoretical gap left in [14].

The classical Nadaraya-Watson regression, random forests, Nadaraya-Watson Fréchet re-
gression [29] and our proposed RFWFR are essentially all local constant estimators, of which
the first two have explicit formulas and the latter two are obtained through implicit opti-
mization. The relationships among the four type estimators are shown in Figure 1.
Fig 1: The relationships among the four local constant estimators.

3. Theory. This section is devoted to the asymptotic analysis of RFWFR. To facilitate further theoretical investigations, we follow [55] and assume that $B \to \infty$. Let $\bar{\alpha}_i(x) = \lim_{B \to \infty} \alpha_i(x) = \lim_{B \to \infty} \frac{1}{B} \sum_{b=1}^{B} \frac{1\{X_i \in L_b(x)\}}{N(L_b(x))}$. Instead of studying the $RF^{(2)}(x)$ defined in (5), we consider

$$\hat{r}_\oplus(x) = \arg\min_{y \in \Omega} \tilde{R}_n(x,y) = \arg\min_{y \in \Omega} \sum_{i=1}^{n} \bar{\alpha}_i(x) d^2(Y_i,y)$$

and develop the corresponding large sample theory. And we see that

$$\hat{r}_\oplus(x) = \arg\min_{y \in \Omega} \lim_{B \to \infty} \frac{1}{B} \sum_{b=1}^{B} \frac{1\{X_i \in L_b(x)\}}{N(L_b(x))} d^2(Y_i,y)$$

$$= \arg\min_{y \in \Omega} \lim_{B \to \infty} \frac{1}{B} \sum_{b=1}^{B} \frac{1}{N(L_b(x))} \sum_{i:X_i \in L_b(x)} d^2(Y_i,y)$$

$$= \arg\min_{y \in \Omega} \left( \frac{n}{s_n} \right)^{-1} \sum_{k} \mathbb{E}_{\xi \sim \Xi} \left( \frac{1}{N(L(x;\mathcal{D}^k_n,\xi))} \sum_{i:X_i \in L(x;\mathcal{D}^k_n,\xi)} d^2(Y_i,y) \right),$$

where the summation about $k$ is taken over all $\binom{n}{s_n}$ subsamples of size $s_n$. $\mathcal{D}^k_n$ is the $k$-th subsample of $\mathcal{D}_n$, $L(x;\mathcal{D}^k_n,\xi)$ is the leaf node containing $x$ of the tree constructed by the subsample $\mathcal{D}^k_n$ and the expectation is taken about the random effect $\xi$. As $s_n \to \infty$ when $n \to \infty$, the objective function of $\hat{r}_\oplus(x)$ is actually an infinite order U-statistic with rank $s_n$ for any fixed $y \in \Omega$. Then the infinite order U-process theory will play a key role in our theoretical studies.

Based on (6) and (7), we can define two population level versions of $\hat{r}_\oplus(x)$ as follows.

$$\tilde{r}_\oplus(x) = \arg\min_{y \in \Omega} \tilde{R}_n(x,y)$$

$$= \arg\min_{y \in \Omega} nE \left[ \bar{\alpha}_i(x) d^2(Y_i,y) \right]$$

$$= \arg\min_{y \in \Omega} \mathbb{E} \left( \frac{1}{N(L(x;\mathcal{D}^k_n,\xi))} \sum_{i:X_i \in L(x;\mathcal{D}^k_n,\xi)} d^2(Y_i,y) \right),$$

where $\mathbb{E}$ is the expectation.
where the expectation is taken about all randomness. Recall the goal of the Fréchet regression in (1) is
\[
m_m(x) = \arg\min_{y \in \Omega} M_m(x, y) = \arg\min_{y \in \Omega} E \left[ d^2(Y, y) \mid X = x \right].
\]
We mainly conduct our asymptotic analysis by separating \(d(\hat{r}_m(x), m_m(x))\) into two parts. One is the variance term \(d(\hat{r}_m(x), \tilde{r}_m(x))\), and the other is bias term \(d(\tilde{r}_m(x), m_m(x))\). As \(\hat{r}_m(x)\) have two different expressions, choosing a suitable form will bring great convenience for our theoretical developments. When we adopt (7) and (9), the theory of infinite order U-statistics and U-processes can be utilized. And when we adopt (6) and (8), the perspective of weighted average is helpful.

3.1. Consistency. We now consider the point-wise consistency of RFWFR. To study the variance term \(d(\hat{r}_m(x), \tilde{r}_m(x))\) and the bias term \(d(\tilde{r}_m(x), m_m(x))\), we need the following regularity assumptions.

(A1) \((\Omega, d)\) is a bounded metric space, i.e., \(\text{diam}(\Omega) = \sup_{y_1, y_2 \in \Omega} d(y_1, y_2) < \infty\).

(A2) \((X, Y)\) has uniformly continuous and bounded density \(\rho\). And the marginal \(\rho_X\) is bounded away from zero such that \(0 < \rho_{\text{min}} \leq \rho_X\).

(A3) \(\text{diam}(L(x)) \to 0\) in probability.

(A4) The object \(m_m(x)\) exists and is unique. For all \(n\), \(\tilde{r}_m(x)\) and \(\hat{r}_m(x)\) exist and are unique, the latter almost surely. Additionally, for any \(\varepsilon > 0\),
\[
\inf_{d(y, m_m(x)) > \varepsilon} \{M_m(x, y) - M_m(x, m_m(x))\} > 0,
\]
\[
\liminf_n \inf_{d(y, \tilde{r}_m(x)) > \varepsilon} \left\{ \tilde{R}_n(x, y) - \tilde{R}_n(x, \tilde{r}_m(x)) \right\} > 0.
\]

Assumptions (A1)-(A2) are commonly used conditions to study the Fréchet tree, see [14]. Suppose we require that the termination condition for the growth of each tree is that the number of samples in the leaf nodes does not exceed a certain constant. In that case, each tree will grow deeper and deeper as \(n \to +\infty\), and thus, the assumption (A3) will hold. The assumption (A4) is also a regular condition to guarantee the consistency of M-estimator in the literature of empirical process. Please refer to Corollary 3.2.3 in [54] for more details. [47] imposed similar assumptions to analyze global or local Fréchet regression.

**Lemma 3.1.** Suppose that, for a fixed \(x \in \mathbb{R}^p\), (A1) and (A4) hold. Then,
\[
d(\hat{r}_m(x), \tilde{r}_m(x)) = o_p(1).
\]

We then deal with the bias term. In addition to assumptions (A1)–(A4), we further require that trees are fitted with honesty as defined in [55]. We now give a brief introduction of honest random forests with Euclidean responses.

A tree is honest if the Euclidean response \(Y_i\) from each training example is only used to estimate \(\mu(x) = E[Y \mid X = x]\) or to decide where to place the splits, but not both. \(Y_i\)'s used for prediction are called prediction points, and \(Y_i\)'s used to construct the tree are called structure points. A random forest is honest if it is composed of honest trees. The advantage of such random forests is that the model construction process and the prediction process are independent. This brings great convenience to the analysis of the theoretical property of random forests. [55] achieve both consistency and the central limit theorem of honest random forests provided that the sub-sample size \(s_n\) scales with an appropriate rate of \(n\). Moreover, their theoretical results apply to a wide range of random forest algorithms, including the classical approach based on the CART criteria.
The simplest way to achieve honesty is that the splitting rule of trees only depends on predictors \( X \) like the purely random forests [2, 26]. If the information of \( Y \) is also considered, then the double-sample tree [55] is a common approach to generate an honest tree. It divides the entire sample into two parts: the construction set \( \mathcal{J} \) and the prediction set \( \mathcal{I} \) such that \(|\mathcal{J}| = \lfloor s_n/2 \rfloor \) and \(|\mathcal{I}| = \lceil s_n/2 \rceil \). During the tree growing process, the splits are chosen using any data from the \( \mathcal{J} \) sample and \( X \)-observations from the \( \mathcal{I} \) sample, but without using \( Y \)-observations from the \( \mathcal{I} \) sample. And the estimation of leaf-wise responses only adopts the \( \mathcal{I} \) sample observations.

The honesty assumption can also be adapted to Fréchet trees, and we then have

\[
E \left( \frac{1}{N(L(x; D^k_n, \xi))} \sum_{i: X_i \in L(x; D^k_n, \xi)} d^2(Y_i, y) \right) = E \left[ E \left[ d^2(Y, y) | X \in L(x) \right] \right].
\]

We emphasize here that \( N(L(x; D^k_n, \xi)) \) and \( X_i \in L(x; D^k_n, \xi) \) only involve prediction points. Then (9) can be further rewritten as

\[
\hat{r}_{\oplus}(x) = \arg\min_{y \in \Omega} \hat{R}_{\oplus}(x, y) = \arg\min_{y \in \Omega} E \left[ E \left[ d^2(Y, y) | X \in L(x) \right] \right].
\]

Based on this formulae (10) of \( \hat{r}_{\oplus}(x) \), the bias term is asymptotically negligible.

**Lemma 3.2.** Suppose that for a fixed \( x \in \mathcal{R}^p \), (A1)–(A4) hold and the Fréchet trees are honest. Then,

\[
d(\hat{r}_{\oplus}(x), m_{\oplus}(x)) = o(1).
\]

Combining the two lemmas above, we are now ready to establish the consistency of \( \hat{r}_{\oplus}(x) \). And we further require that the Fréchet trees are symmetric under the condition of honesty. The Fréchet tree is symmetric if the (randomized) output of the tree does not depend on the order \((i = 1, 2, \ldots, s_n)\) in which the training examples are indexed.

**Theorem 3.3.** Suppose that for a fixed \( x \in \mathcal{R}^p \), (A1)–(A4) hold and the Fréchet trees are honest and symmetric. Then the RFWFR is point-wise consistent, that is,

\[
d(\hat{r}_{\oplus}(x), m_{\oplus}(x)) = o_p(1).
\]

Our consistency result above is different from [14] in two aspects. Firstly, consistency result in [14] is only for a single Fréchet tree, while the result here holds for the estimator based on Fréchet random forests with multiple trees. Secondly, [14] focused on the excess risk of Fréchet regression, while our consistency result is established directly based on the natural distance measure between the RFWFR \( \hat{r}_{\oplus}(x) \) and the target regression function \( m_{\oplus}(x) \).

If the previous last two assumptions are suitably strengthened, we can further obtain the uniform convergence results for \( \hat{r}_{\oplus}(x) \). Let \( \| \cdot \| \) be the Euclidean norm on \( \mathcal{R}^p \) and \( B > 0 \).

(U0) For any \( \|x\| \leq B \), \( M_{\oplus}(x, y) \) is equicontinuous, i.e.,

\[
\lim_{\hat{x} \to x} \sup_{y \in \Omega} |M_{\oplus}(\hat{x}, y) - M_{\oplus}(x, y)| = 0.
\]

(U3) \( \sup_{\|x\| \leq B} \text{diam}(L(x)) \to 0 \) in probability.

(U4) For all \( \|x\| \leq B \), \( m_{\oplus}(x) \), \( \hat{r}_{\oplus}(x) \) and \( \hat{r}_{\oplus}(x) \) exist and are unique, the latter almost surely. Additionally, for any \( \varepsilon > 0 \),

\[
\inf_{\|x\| \leq B} \inf_{d(y, m_{\oplus}(x)) > \varepsilon} \{M_{\oplus}(x, y) - M_{\oplus}(x, m_{\oplus}(x))\} > 0.
\]
convergence rate results. The convergence rate of the variance term. A little setup is needed before we can describe the sign-symmetrization technique. Our theoretical studies will adopt this inequality to establish the uniform entropy integrals. [28] discovered a stronger maximal inequality by the partial

\[ \sup_{\|x\| \leq B} d(\tilde{r}_{\oplus}(x), m_{\oplus}(x)) = o_p(1). \]

3.2. Rate of Convergence. Following the path of the previous section, we consider the convergence rate of variance term and bias term respectively. We focus on the variance term in the first place. From (7) and (9), we recall that

\[
\hat{R}_n(x, y) = \left( \frac{n}{s_n} \right)^{-1} \sum_{k} E_{\xi \sim \Xi} \left( \frac{1}{N(L(x; D_{n,k}^\xi, \xi))} \sum_{i: X_i \in L(x; D_{n,k}^\xi, \xi)} d^2(Y_i, y) \right),
\]

\[
\bar{R}_n(x, y) = E \left( \frac{1}{N(L(x; D_{n,k}^\xi, \xi))} \sum_{i: X_i \in L(x; D_{n,k}^\xi, \xi)} d^2(Y_i, y) \right).
\]

By Theorem 3.2.5 of [54], the key to determine the convergence rate is to establish the upper bound of the maximum value of the centered and scaled process \( \sqrt{n} \left( \hat{R}_n - \bar{R}_n \right) \).

Recall that \( \hat{R}_n \) is an infinite order U-statistic for any fixed \( y \in \Omega \). When we consider the corresponding process, the maximal inequality of the infinite order U-process is the most important tool. Some related papers have studied the maximal inequality of U-processes. [52] established the maximal inequality for degenerate U-processes of arbitrary order based on moment inequalities. [1] obtained a similar maximal inequality by exponential inequalities, which is stronger than that of [52]. However, both the results are limited to U-processes of fixed order. [27] first extended the maximal inequality of [52] to infinite order U-processes by the complete sign-symmetrization technique. Unfortunately, they are unable to extend the maximal inequality of [1] to the infinite order case because such results rely on a symmetrization inequality of [18] which incurs upper bounds that grow with the order much too quickly. [15] gave a local maximal inequality for degenerate infinite order U-processes in the form of the uniform entropy integrals. [28] discovered a stronger maximal inequality by the partial sign-symmetrization technique. Our theoretical studies will adopt this inequality to establish the convergence rate of the variance term. A little setup is needed before we can describe the convergence rate results.

Let \( Z_1 = (X_1, Y_1), \ldots, Z_n = (X_n, Y_n) \), and define

\[
H_n(Z_{i_1}, \ldots, Z_{i_{s_n}}, y) = E_{\xi \sim \Xi} \left( \frac{1}{N(L(x; D_{n,k}^\xi, \xi))} \sum_{i: X_i \in L(x; D_{n,k}^\xi, \xi)} \left( d^2(Y_i, y) - d^2(Y_i, \tilde{r}_{\oplus}(x)) \right) \right).
\]

Consider the function class

\[
\mathcal{H}_\delta := \{ H_n(z_1, \ldots, z_{s_n}, y) : d(y, \tilde{r}_{\oplus}(x)) < \delta \}.
\]
Let $Z_i^0 = (X_i, Y_i)$, for $i = 1, \ldots, n$, and let $\{Z_i^1\}_{i=1}^n$ be i.i.d., independent of $\{Z_i^0\}_{i=1}^n$, with the same distribution. For $\forall H_n(y_1), H_n(y_2) \in \mathcal{H}_\delta$, define the following random pseudometric

$$d_j(H_n(y_1), H_n(y_2)) = \frac{\sum_{k=1}^{n} \left| \sum_{a \in (n), a_{1:k} = k} H_n \left( Z_{a_1}^{0,1}; y_1 \right) - H_n \left( Z_{a_1}^{0,1}; y_2 \right) \right|}{\sum_{a \in (n), a_{1:k}} G_\delta \left( Z_{a_1}^{0,1} \right)}.$$ 

where $Z_{a_1}^{0,1} = \left( Z_{a_1}^0, \ldots, Z_{a_1}^0, Z_{a_1+1}^1, \ldots, Z_{a_n}^1 \right), (n)_{s_n}$ represents all the permutations of taking $s_n$ distinct elements from the set $\{1, 2, \ldots, n\}$, and $G_\delta$ is the envelope function for $\mathcal{H}_\delta$ such that $|H_n| \leq G_\delta$ for every $H_n \in \mathcal{H}_\delta$.

We assume the following assumptions to analyze the convergence rate of the variance term.

(A5) There exist constants $A$ and $V$ such that

$$\max_{j \leq s_n} N(\varepsilon, d_j, \mathcal{H}_\delta) \leq A e^{-V}$$

as $\delta \to 0$ for any $\varepsilon \in (0, 1]$, where $N(\varepsilon, d_j, \mathcal{H}_\delta)$ is the $\varepsilon$-covering number of the function class $\mathcal{H}_\delta$ based on the pseudometric $d_j$ we introduce.

(A6) There exist $\delta_1 > 0, C_1 > 0$ and $\beta_1 > 1$, possibly depending on $x$, such that, whenever $d(y, \tilde{r}_\oplus(x)) < \delta_1$, we have

$$\liminf_n \left[ \tilde{R}_n(x, y) - \tilde{R}_n(x, \tilde{r}_\oplus(x)) - C_1 d(y, \tilde{r}_\oplus(x))^{\beta_1} \right] \geq 0.$$ 

The assumption (A5) along with the pseudometric $d_j$ were proposed by [27, 28] to establish the maximal inequality of infinite order U-processes. From the perspective of empirical process, (A5) regulates $\mathcal{H}_\delta$ a Euclidean class. Knowing that a class of functions is Euclidean aids immensely in establishing the convergence rate. The assumption (A6) comes from [47]. It is also an extension of the condition that control the rate of convergence of $M$-estimator. Please refer to Theorem 3.2.5 of [54] for more details.

**Theorem 3.5.** Suppose that for a fixed $x \in \mathcal{R}^p$, (A1), (A4), (A5) and (A6) hold. Then we have

$$d(\tilde{r}_\oplus(x), \tilde{r}_\oplus(x)) = O_p \left[ \left( \frac{s_n^2 \log s_n}{n} \right)^{\frac{1}{\pi_{\oplus}(x)}} \right].$$

**Remark 3.6.** Consider the special case when $\Omega$ is a bounded subset of $\mathcal{R}$. If the trees are honest, then

$$\tilde{R}_n(x, y) = E \left[ E \left[ (Y - y)^2 \mid X \in L(x) \right] \right] \quad \text{and} \quad \tilde{r}_\oplus(x) = E \left[ E(Y \mid X \in L(x)) \right].$$

And we can further get that

$$\tilde{R}_n(x, y) - \tilde{R}_n(x, \tilde{r}_\oplus(x))$$

$$= E \left[ E \left[ (Y - y)^2 \mid X \in L(x) \right] \right] - E \left[ E \left[ (Y - \tilde{r}_\oplus(x))^2 \mid X \in L(x) \right] \right]$$

$$= y^2 - 2y E \left[ E(Y \mid X \in L(x)) \right] + \left[ E \left[ E(Y \mid X \in L(x)) \right] \right]^2$$

$$= (y - \tilde{r}_\oplus(x))^2,$$

which indicates that $\beta_1 = 2$ in the assumption (A6) for Euclidean responses. By the results of Theorem 3.5, $d(\tilde{r}_\oplus(x), \tilde{r}_\oplus(x)) = O_p \left[ \left( \frac{s_n^2 \log s_n}{n} \right)^{\frac{1}{2}} \right]$. The rate derived here is slower than
as described in our asymptotic normality results in the next section. The reason is that existing maximum inequalities of infinite order U-processes are not strong enough. More sophisticated tools of infinite order U-processes are expected to be developed to further improve this convergence rate as \( O_p \left( \frac{s_n \log s_n}{n} \right)^{\frac{1}{2(\beta - 1)}} \).

Now we turn to the convergence rate of the bias term, which is closely related to the construction of Fréchet trees. To bound the bias of random forests with Euclidean responses and to further eliminate the bias when deriving the asymptotic normality, [55] requires some more conditions other than honesty for the tree. Here we place the same requirements on the construction of Fréchet trees.

(i) (Random-split) At each node split, marginalizing over \( \xi \), the probability that \( X_j \) is selected as the split variable is bounded below by \( \frac{\pi}{p} \) for some \( 0 < \pi \leq 1 \), for all \( j = 1, \ldots, p \).

(ii) (\( \alpha \)-regular) After each split is completed, each child node contains at least a fraction \( \alpha > 0 \) of the available training sample. Moreover, the tree stops to grow if every leaf node contains only between \( k \) and \( 2k - 1 \) observations, where \( k \in \mathbb{N} \) is some fixed tuning parameter.

Some more assumptions are needed to get the convergence rate of the bias term.

(A7) For each \( y \), \( M_{\oplus}(x, y) \) is Lipschitz-continuous about \( x \), and the Lipschitz constant has a common upper bound \( K \).

(A8) There exist \( \delta_2 > 0, C_2 > 0 \) and \( \beta_2 > 1 \), possibly depending on \( x \), such that, whenever \( d(y, m_{\oplus}(x)) < \delta_2 \), we have \( M_{\oplus}(x, y) - M_{\oplus}(x, m_{\oplus}(x)) \geq C_2 d(y, m_{\oplus}(x))^\beta_2 \).

The Lipschitz-continuity in the assumption (A7) allows us to control the bias term by restricting the diameter of the sample space represented by the leaf node. The assumption (A8) is similar to the assumption (A6). Both the two assumptions are adaptions of [55] to Fréchet regression with metric space valued responses.

**Theorem 3.7.** Suppose \( X_1, \ldots, X_n \in [0, 1]^p \). Assume that for a fixed \( x \in [0, 1]^p \), (A1), (A2), (A4), (A7) and (A8) hold, and the Fréchet trees are \( \alpha \)-regular, random-split and honest. Then, provided that \( \alpha \leq 0.2 \), we have

\[
d(\tilde{r}_{\oplus}(x), m_{\oplus}(x)) = O \left( \frac{s_n \log s_n}{n} \right)^{\frac{1}{2(\beta - 1)}}.
\]

as \( n \to \infty \).

**Remark 3.8.** Consider again the special case when \( \Omega \) is a bounded subset of \( \mathbb{R} \). We then have

\[
M_{\oplus}(x, y) - M_{\oplus}(x, m_{\oplus}(x)) \\
= E \left[ (Y - y)^2 \mid X = x \right] - E \left[ (Y - E(Y|X = x))^2 \mid X = x \right] \\
= (y - E(Y|X = x))^2 \\
= (y - m_{\oplus}(x))^2.
\]

That is to say, the assumption (A8) holds when \( \beta_2 = 2 \). By the above Theorem 3.7,

\[
d(\tilde{r}_{\oplus}(x), m_{\oplus}(x)) = O \left( \frac{s_n \log s_n}{n} \right)^{\frac{1}{2(\beta_2 - 1)}}.
\]
as \( n \to \infty \). This rate coincides with Theorem 3 of [55], which indicates that our asymptotic bias result generalizes that of Euclidean random forests to Fréchet random forests.

Combining Theorem 3.5 and Theorem 3.7, we get the convergence rate for \( \hat{r} \oplus (x) \).

**Corollary 3.9.** Suppose \( X_1, \ldots, X_n \in [0, 1]^p \). Assume that for a fixed \( x \in [0, 1]^p \), (A1), (A2), (A4)–(A8) hold, and the Fréchet trees are \( \alpha \)-regular, random-split, honest and symmetric. Then, provided that \( \alpha \leq 0.2 \), we have

\[
d(\hat{r} \oplus (x), m \oplus (x)) = O_p \left( s_n^{-2} \left( \frac{\log s_n}{n} \right)^{2(\beta_1-1)} + s_n^{-2} \frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})} \frac{1}{p \beta_2 - 1} \right).
\]

### 3.3. Asymptotic Normality

There are two major difficulties in deriving the asymptotic normality of our proposed RFWFR. On the one hand, it is not the classical \( M \)-estimator, but the \( M_m \)-estimator [11] and even the \( M_{m_n} \)-estimator with infinite order U-processes. So we have to deal with the most general \( M_{m_n} \)-estimator, where \( m_n \) diverges to infinity. On the other hand, the \( M_{m_n} \)-estimator here takes value not in Euclidean space but in general metric space, which will also bring difficulties to the study of asymptotic limiting distribution. To address the first difficulty, we will generalize the result in section 2.5 of [11] to acquire the probability representation and the asymptotic normality of the \( M_{m_n} \)-estimator. As for the second issue, the seminal work of [5, 6, 7] concluded that the map of the sample Fréchet mean is asymptotically normally distributed around the map of the Fréchet mean under certain assumptions. We follow their developments in combination with our developed asymptotic tool for \( M_{m_n} \)-estimator to establish the asymptotic normality of RFWFR finally.

First, we generalize the theory of \( M_m \)-estimator in [11] to the case that \( m \) tends to infinity along with \( n \).

**Definition 3.10.** Let \( Z_1, Z_2, \ldots, Z_{m_n} \) be i.i.d. \( Z \)-valued random variables and \( \theta \in \mathbb{R}^d \). Let \( f_n(z_1, z_2, \ldots, z_{m_n}, \theta) \) be a real valued measurable function which is symmetric in the arguments \( z_1, z_2, \ldots, z_{m_n} \) for each \( n \). Define

\[
Q_n(\theta) = E f_n(Z_1, Z_2, \ldots, Z_{m_n}, \theta)
\]

and

\[
\hat{\theta}_n = \arg\min_{\theta \in \mathbb{R}^d} Q_n(\theta).
\]

\( \hat{\theta}_n \) is called the \( M_{m_n} \)-parameter.

**Definition 3.11.** Let \( Z_1, Z_2, \cdots, Z_n \) be a sequence of i.i.d. observations. Define

\[
\hat{Q}_n(\theta) = \left( \frac{n}{m_n} \right)^{-1} \sum_{1 \leq i_1 < i_2 < \cdots < i_{m_n} \leq n} f_n(Z_{i_1}, Z_{i_2}, \ldots, Z_{i_{m_n}}, \theta)
\]

and

\[
\hat{\theta}_n = \arg\min_{\theta \in \mathbb{R}^d} \hat{Q}_n(\theta).
\]

\( \hat{\theta}_n \) is called the \( M_{m_n} \)-estimator of \( \theta_n \).
In the above definition, a hidden assumption is $m_n/n \to 0$. Actually, $\hat{Q}_n(\theta)$ is an infinite order U-process about $\theta$. Since $\hat{Q}_n(\theta)$ is the sample analogue of $Q_n(\theta)$, $\hat{\theta}_n$ is a reasonable estimator of $\theta_n$. When $m_n = 1$, $\hat{\theta}_n$ is the classical M-estimator. When $m_n = m$ is a fixed positive integer, $\hat{\theta}_n$ is the $M_m$-estimator studied in [11]. By an appropriate selection theorem, it is often possible to choose a measurable version of $\hat{\theta}_n$. We always work with such a version. Please refer to section 2.3 of [11] for more details.

Let $g_n$ be a measurable sub-gradient of $f_n(z_1, z_2, \ldots, z_{m_n}, \theta)$ about $\theta$. Define

$$K_n = \text{Var} \left[ E \left( g_n \left( Z_1, Z_2, \ldots, Z_{m_n}, \theta_n \right) | Z_1 \right) \right],$$

$$U_n = \left( \frac{n}{m_n} \right)^{-1} \sum_{1 \leq i_1 < i_2 < \ldots < i_m \leq n} g_n \left( Z_{i_1}, Z_{i_2}, \ldots, Z_{i_m}, \theta_n \right).$$

In order to achieve the asymptotic normality of the $M_{m_n}$-estimator, we need the following assumptions.

(i) $f_n(z_1, z_2, \ldots, z_{m_n}, \theta)$ is measurable in $(z_1, z_2, \ldots, z_{m_n})$ and convex in $\theta$ for each $n$.

(ii) $Q_n(\theta)$ is finite for each $\theta$ and $n$.

(iii) $\theta_n$ exists and is unique for each $n$, and $f_n(z_1, z_2, \ldots, z_{m_n}, \theta)$ is twice differentiable on an appropriate neighborhood of $\theta_n$.

(iv) $E \left| g_n \left( Z_1, Z_2, \ldots, Z_{m_n}, \theta_n \right) \right|^2 < \infty$ for each $n$, and $m_n \lambda_{\min}(K_n) \to 0$, where $\lambda_{\min}(K_n)$ denotes the smallest eigenvalue of $K_n$.

(v) $H_n = \nabla^2 Q(\theta_n)$ exists and is positive definite for each $n$ and $\lambda_{\min}(H_n) \to 0$.

Through the definition, we know that the $M_{m_n}$-estimator is an implicit solution to the infinite order U-process. Under the above assumptions, we can derive $\hat{\theta}_n$ a weak representation through the linearization of the infinite order U-statistic $U_n$. Therefore, the asymptotic normality of infinite order U-statistics determines the asymptotic normality of the $M_{m_n}$-estimator. [41] gives sufficient conditions for asymptotic normality of such U-statistics. However, these conditions can not hold simultaneously. [20] then developed conditions which can be verified on the basis of [41]. But both of their results require that the order of the infinite order U-statistics is $o(\sqrt{n})$. [45] further improved this result when the order of the infinite order U-statistics is $o(n)$. [55] also gave the same rate (ignoring the log-factors) when focusing on random forests with some additional requirements on the construction of trees. Our assumption (iv) here is to ensure the asymptotic normality of $U_n$ by the result of [45]. And assumptions (i), (ii), (iii) and (v) are adaptions of that for studying $M_m$ estimator in [11].

**Theorem 3.12.** Suppose that assumptions (i)-(v) hold, then for any sequence of measurable minimizers $\{\theta_n, n \geq 1\}$,

(a) $\hat{\theta}_n - \theta_n = -H_n^{-1} U_n + o_p \left( \frac{m_n}{\sqrt{n}} \right)$.

(b) $\sqrt{n} \Lambda_n^{-\frac{1}{2}} \left( \hat{\theta}_n - \theta_n \right) \overset{d}{\to} N(0, I)$, where

$$\Lambda_n = m_n^2 H_n^{-1} K_n H_n^{-1}.$$

Theorem 3.12 establishes the asymptotic normality of $M_{m_n}$-estimator, which is a generalization of the central limit theorem of the $M_m$-estimator given in [11]. Next we derive the asymptotic normality of RFWFR by applying this result. Here we continue to adopt the expressions (7) and (9). For the consistency of notations, let

$$h_n \left( Z_{i_1}, Z_{i_2}, \ldots, Z_{i_{m_n}}, y \right) = E_{\xi \sim \Xi} \left( \frac{1}{N(L(x; D_{1n}^k, \xi))} \sum_{i : X_i \in L(x; D_{1n}^k, \xi)} d^2(Y_i, y) \right)$$
where \( Z_i = (X_i, Y_i) \) and \( \mathcal{D}_n^k = \{ Z_{i_1}, Z_{i_2}, \ldots, Z_{i_n} \} \). Then

\[
\tilde{r}_\oplus(x) = \arg\min_{y \in \Omega} \tilde{R}_n(x, y) = \arg\min_{y \in \Omega} \left( \frac{n}{s_n} \right)^{-1} \sum_{1 \leq i_1 < i_2 < \cdots < i_n \leq n} h_n \left( Z_{i_1}, Z_{i_2}, \ldots, Z_{i_n}, y \right),
\]

\[
\bar{r}_\oplus(x) = \arg\min_{y \in \Omega} \bar{R}_n(x, y) = \arg\min_{y \in \Omega} \bar{h}_n \left( Z_1, Z_2, \ldots, Z_s, y \right).
\]

Unfortunately, since \( y \) is not in Euclidean space, we cannot compute the derivative about \( y \) and \( \tilde{r}_\oplus(x) - \bar{r}_\oplus(x) \) has no sense. In order to apply Theorem 3.12, we consider mapping \( y \) to the Euclidean space locally and establish asymptotic normality, which is the standard procedure for the asymptotic analysis of sample Fréchet mean as introduced in [5, 6, 7]. We assume the following conditions to establish the central limit theorem of the proposed RFWFR.

\begin{enumerate}
    \item \((A9)\) \( h_n(z_1, z_2, \ldots, z_s, y) \) is measurable in \((z_1, z_2, \ldots, z_s)\) and \( \bar{R}_n(x, y) < \infty \) for each \( y \) and \( n \).
    \item \((A10)\) \( \tilde{r}_\oplus(x) \) exists and is unique for each \( n \).
    \item \((A11)\) \( \tilde{r}_\oplus(x) \in G \) for large \( n \), where \( G \) is a measurable subset of \( \Omega \). And there is a homeomorphism \( \phi : G \to U \), where \( U \) is an open subset of \( \mathbb{R}^d \) for some \( d \geq 1 \), and \( G \) is given its relative topology on \( \Omega \). Also \( u \mapsto f_n(z_1, z_2, \ldots, z_s, u) = h_n \left( z_1, z_2, \ldots, z_s, \phi^{-1}(u) \right) \) is twice differentiable on an appropriate neighborhood of \( \phi(\tilde{r}_\oplus(x)) \).
    \item \((A12)\) \( f_n(z_1, z_2, \ldots, z_s, u) \) is convex in \( u \) for each \( n \).
    \item \((A13)\) Let \( g_n \) be a measurable sub-gradient of \( f_n \) about \( u \). Define
    \[
    g_n = E \left( g_n \left( Z_1, Z_2, \ldots Z_s, \phi(\tilde{r}_\oplus(x)) \right) | Z_1 \right),
    \]
    \[
    K_n = \text{Var} \left( g_n \right).
    \]

    Then \( E|g_n(Z_1, Z_2, \ldots, Z_s, \phi(\tilde{r}_\oplus(x))|^2 < \infty \) for each \( n \), and \( s_n \lambda_{\min}(K_n) \to 0 \).
    \item \((A14)\) \( H_n = \nabla^2 E f_n(Z_1, Z_2, \ldots, Z_s, \phi(\tilde{r}_\oplus(x))) \) exists and is positive definite for each \( n \), and \( \lambda_{\min}(H_n) \to 0 \).
\end{enumerate}

The assumption \((A11)\) is crucial just like the assumption for establishing the asymptotic normality of sample Fréchet mean as suggested in [5]. Other assumptions are adaptations of that in Theorem 3.12.

**Theorem 3.13.** Suppose that for a fixed \( x \in \mathcal{R}^p \), \((A1), (A4)\) and \((A9)\)–\((A14)\) hold. Then

\[
\sqrt{n} \Lambda_n^{-\frac{1}{2}} \left( \phi(\tilde{r}_\oplus(x)) - \phi(\bar{r}_\oplus(x)) \right) \overset{d}{\to} \mathcal{N}(0, I)
\]

where

\[
\Lambda_n = s_n^2 H_n^{-1} K_n H_n^{-1}.
\]

**Remark 3.14.** This result is established for random forests with responses in the most general metric space, which can be applied to both intrinsic and extrinsic analysis. The metric can be the geodesic distance on a Riemannian manifold with some metric tensor for the intrinsic analysis. For the extrinsic analysis, the metric can be the distance from embedding the output space into the Euclidean space. However, we only utilize the final metric in either case and do not care about the Riemannian structure or other information behind it. Adding more structure information of the metric space to the nonparametric Fréchet regression may result in very similar results to the direct use of metric. See [47] for more discussions.
REMARK 3.15. Consider the special case when $\Omega$ is a bounded subset of $\mathcal{R}$. For responses that are Euclidean, we naturally choose $\phi$ as the identity mapping. Then we have

$$f_n(Z_1, \ldots, Z_{s_n}, u) = h_n(Z_1, \ldots, Z_{s_n}, u) = E_{\xi \sim \Xi} \left( \frac{1}{N(L(x; D_n^k, \xi))} \sum_{i: X_i \in L(x; D_n^k, \xi)} (Y_i - u)^2 \right).$$

And we further get

$$Q_n(u) = E \left( \frac{1}{N(L(x; D_n^k, \xi))} \sum_{i: X_i \in L(x; D_n^k, \xi)} (Y_i - u)^2 \right),$$

$$\hat{Q}_n(u) = \left( \frac{n}{s_n} \right)^{-1} \sum_k E_{\xi \sim \Xi} \left( \frac{1}{N(L(x; D_n^k, \xi))} \sum_{i: X_i \in L(x; D_n^k, \xi)} (Y_i - u)^2 \right).$$

In addition,

$$u_n = \arg\min_{u \in U} Q_n(u) = E \left( \frac{1}{N(L(x; D_n^k, \xi))} \sum_{i: X_i \in L(x; D_n^k, \xi)} Y_i \right),$$

$$\hat{u}_n = \arg\min_{u \in U} \hat{Q}_n(u) = \left( \frac{n}{s_n} \right)^{-1} \sum_k E_{\xi \sim \Xi} \left( \frac{1}{N(L(x; D_n^k, \xi))} \sum_{i: X_i \in L(x; D_n^k, \xi)} Y_i \right).$$

Since $g_n$ is the sub-gradient of $f_n$ about $u$, we have

$$g_n(Z_1, \ldots, Z_{s_n}, u) = -2 \left[ E_{\xi \sim \Xi} \left( \frac{1}{N(L(x; D_n^k, \xi))} \sum_{i: X_i \in L(x; D_n^k, \xi)} Y_i \right) - u \right].$$

Let $\zeta_1, n = \text{Var} \left( E \left( E_{\xi \sim \Xi} \left( \frac{1}{N(L(x; D_n^k, \xi))} \sum_{i: X_i \in L(x; D_n^k, \xi)} Y_i \right) \mid Z_1 \right) \right)$, then $K_n = 4\zeta_1, n$ and $\Lambda_n = s_n^2 H_n^{-1} K_n H_n^{-1} = s_n^2 \zeta_1, n$. If $s_n \zeta_1, n \rightarrow 0$, the assumption (A13) holds. Then we see that

$$\frac{\sqrt{n} (\hat{u}_n - u_n)}{s_n \zeta_1, n} \xrightarrow{d} N(0, 1),$$

i.e.,

$$\frac{\sqrt{n} (\hat{\tau}(x) - \tilde{\tau}(x))}{\sqrt{s_n^2 \zeta_1, n}} \xrightarrow{d} N(0, 1).$$

Note that $\hat{\tau}(x)$ is exactly the prediction of the Euclidean random forests at $x$, and (11) becomes the standard asymptotic normality results of Euclidean random forests [41, 45, 55].

If the homeomorphism $\phi$ in Theorem 3.13 is Lipschitz-continuous, we can further discuss the asymptotic normality of $\phi(\hat{\tau}(x))$ about $\phi(m_{\oplus}(x))$. Denote the Lipschitz constant of the Lipschitz homeomorphism $\phi$ as $L$, then by Theorem 3.7

$$\|\phi(\hat{\tau}(x)) - \phi(m_{\oplus}(x))\| \leq Ld(\hat{\tau}(x), m_{\oplus}(x)) = O \left( \frac{s_n^{- \frac{1}{2} \log((1 - \alpha)^{-1})}}{\log(\alpha^{-1})} \frac{\pi^{-1}}{\beta^{1/2}} \right).$$
Let $s_n = n^\beta$, then

$$\frac{\|\phi(\hat{r}_\otimes(x)) - \phi(\hat{m}_\otimes(x))\|}{\|\Lambda_n^{\frac{1}{2}}/\sqrt{n}\|} = O\left(\frac{n^{\frac{1}{2}}} n^{1-\beta} \left(1 + \frac{\log(1-\alpha)^{-1}}{\pi p \log(\alpha^{-1}) \beta^2 - 1}\right)\right).$$

The right-hand-side converges to 0 provided that

$$\beta > \left(1 + \frac{\log(1-\alpha)^{-1}}{\pi^{-1} p \log(\alpha^{-1}) \beta^2 - 1}\right)^{-1} = 1 - \left(1 + \frac{p \log(\alpha^{-1}) (\beta^2 - 1)}{\pi \log((1-\alpha)^{-1})}\right)^{-1}.$$

Then we have the following corollary by Theorem 3.13 and the Slutsky’s theorem.

**Corollary 3.16.** Suppose $X_1, \ldots, X_n \in [0, 1]^p$. Assume that for a fixed $x \in [0, 1]^p$, (A1), (A2), (A4) and (A7)–(A14) hold. And assume that the homeomorphism $\phi$ is Lipschitz-continuous. Moreover, the Fréchet trees are $\alpha$-regular, random-split, honest and symmetric. Then if $\alpha \leq 0.2$ and the subsample size $s_n$ satisfies

$$s_n \asymp n^\beta$$

for some $\beta_{\text{min}} := 1 - \left(1 + \frac{p \log(\alpha^{-1}) (\beta^2 - 1)}{\pi \log((1-\alpha)^{-1})}\right)^{-1} < \beta < 1,$

we have

$$\sqrt{n} \Lambda_n^{-\frac{1}{2}} (\phi(\hat{r}_\otimes(x)) - \phi(\hat{m}_\otimes(x))) \xrightarrow{d} \mathcal{N}(0, I)$$

where

$$\Lambda_n = s_n^2 H_n^{-1} K_n H_n^{-1}.$$

**Remark 3.17.** Consider the special case when $\Omega$ is a bounded subset of $\mathcal{R}$. And we take $\phi$ as the identity mapping and $\beta_2 = 2$. This mapping is naturally Lipschitz-continuous. By Corollary 3.16, we have

$$\frac{\sqrt{n} (\hat{r}_\otimes(x) - \hat{m}_\otimes(x))}{\sqrt{s_n^2 \zeta_{1,n}}} \xrightarrow{d} \mathcal{N}(0, 1)$$

when

$$s_n \asymp n^\beta$$

for some $\beta_{\text{min}} := 1 - \left(1 + \frac{p \log(\alpha^{-1}) (\beta^2 - 1)}{\pi \log((1-\alpha)^{-1})}\right)^{-1} < \beta < 1.$

This result coincides with Theorem 1 of [55] of random forests with Euclidean responses. Therefore, our asymptotic normality established for Fréchet regression with general metric space valued responses includes the most updated results of Euclidean random forests as a special case.

4. Random Forests Weighted Local Linear Fréchet Regression. We have proposed random forests weighted Fréchet regression in section 2, which is a Nadaraya-Watson type estimator using random forests weighting function. A very natural extension is to carry out local linear Fréchet regression further. As we know, the local linear estimator is based on the first-order Taylor expansion of the regression function, while the Nadaraya-Watson estimator is a local constant type method. The local linear estimator is more flexible and is thus expected to be more precise, especially near the boundaries. Local Fréchet regression proposed by [47] is just a local linear estimator adapted to the cases with metric space valued responses.
Similar to the classical local linear estimator, their method still suffers from the curse of dimensionality. In particular, if the predictors contain some irrelevant variables, using classical kernel functions often has worse performance. On the contrary, the mechanism of random forests can help combat the curse of dimensionality. For sparse and high-dimensional predictors, it can select more relevant variables to divide the sample space \[8, 34\]. Due to these superior characteristics of the random forests, \[10\] and \[24\] use random forests kernel instead of those classical smoothing kernel functions. But their research is limited to classical Euclidean regression problems. This section proposes the second method called random forests weighted local linear Fréchet regression (RFWLLFR), which adopts the random forests kernel to local Fréchet regression.

Recall that \( \Omega \) in our setting is a general metric space not limited to Riemannian manifolds. Then there is not a Taylor expansion similar to the Euclidean case. To extend our proposed RFWFR, we first review the classical local linear estimator with \( p = 1 \) and \( \Omega = \mathcal{R} \). Consider the following optimization problem,

\[
\left( \hat{\beta}_0, \hat{\beta}_1 \right) = \arg\min_{\beta_0, \beta_1} \frac{1}{n} \sum_{i=1}^{n} K_h (X_i - x) (Y_i - \beta_0 - \beta_1 (X_i - x))^2.
\]

Then the local linear estimator is

\[
\hat{l}(x) = \hat{\beta}_0 = e_1^T (X^T W X)^{-1} \sum_{i=1}^{n} \left( \frac{1}{X_i - x} \right) K_h (X_i - x) Y_i,
\]

where

\[
X := \begin{pmatrix} 1 & X_1 - x \\ 1 & X_2 - x \\ \vdots \\ 1 & X_n - x \end{pmatrix}_{n \times 2}, \quad W := \text{diag}(K_h (X_1 - x), \ldots, K_h (X_n - x)), \quad e_1 := \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.
\]

Keep in mind that random forests show a good behavior for applications in high-dimensional setting. If we view the random forests as a weight generating function, it is easy to obtain the local linear regression based on random forests kernel. For \( p \geq 1 \) and \( \Omega = \mathcal{R} \), \[10\] and \[24\] considered

\[
\left( \hat{\beta}_0, \hat{\beta}_1 \right) = \arg\min_{\beta_0, \beta_1} \frac{1}{n} \sum_{i=1}^{n} \alpha_i (x) (Y_i - \beta_0 - \beta_1^T (X_i - x))^2,
\]

where \( \alpha_i (x) = \frac{1}{B} \sum_{b=1}^{B} \frac{1}{1 \{X_i \in L_b(x)\}} \). Then define the random forests weighted local linear estimator as

\[
\hat{l}(x) = \hat{\beta}_0 = e_1^T (X^T A X)^{-1} \sum_{i=1}^{n} \left( \frac{1}{X_i - x} \right) \alpha_i (x) Y_i
\]

where

\[
X := \begin{pmatrix} 1 & X_{11} - x_1 & \cdots & X_{1p} - x_p \\ 1 & X_{21} - x_1 & \cdots & X_{2p} - x_p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} - x_1 & \cdots & X_{np} - x_p \end{pmatrix}_{n \times (p+1)}, \quad A := \text{diag}(\alpha_1(x), \ldots, \alpha_n(x)).
\]

In order to generalize (13) to metric space valued responses, we need to rewrite it as an implicit form

\[
\hat{l}(x) = \arg\min_{y \in \mathcal{R}} e_1^T (X^T A X)^{-1} \sum_{i=1}^{n} \left( \frac{1}{X_i - x} \right) \alpha_i (x) (Y_i - y)^2.
\]
Replacing the Euclidean distance with the general metric $d$, the random forests weighted local linear Fréchet regression for $p \geq 1$ and general metric space $(\Omega, d)$ is proposed as

\begin{equation}
\hat{l}_\oplus(x) = \arg\min_{y \in \Omega} e_1^T (X^T A X)^{-1} \sum_{i=1}^n \left( \frac{1}{X_i - x} \right) \alpha_i(x) d^2(Y_i, y).
\end{equation}

In fact, we can also get the random forests weighted Fréchet regression again through the above generalization process. What we need to do is to only replace the first-order Taylor expansion in (12) with zero-order.

Consider the special case with $p = 1$, (14) have the following equivalent expressions.

\begin{align*}
\hat{l}_\oplus(x) &= \arg\min_{y \in \Omega} \sum_{i=1}^n e_1^T (X^T A X)^{-1} \left( \frac{1}{X_i - x} \right) \alpha_i(x) d^2(Y_i, y) \\
&= \arg\min_{y \in \Omega} \sum_{i=1}^n e_1^T \left( \sum_{i=1}^n \alpha_i(x) (X_i - x) \sum_{i=1}^n \alpha_i(x) (X_i - x)^2 \right)^{-1} \left( \frac{1}{X_i - x} \right) \alpha_i(x) d^2(Y_i, y) \\
&= \arg\min_{y \in \Omega} \frac{1}{n} \sum_{i=1}^n e_1^T \left( \frac{\hat{\mu}_2 - \hat{\mu}_1}{\hat{\mu}_0 - \hat{\mu}_1} \right) \left( \frac{1}{X_i - x} \right) \alpha_i(x) d^2(Y_i, y) \\
&= \arg\min_{y \in \Omega} \frac{1}{n} \sum_{i=1}^n t_{in}(x) d^2(Y_i, y)
\end{align*}

where $t_{in}(x) = \frac{1}{\hat{\mu}_0 \hat{\mu}_2 - \hat{\mu}_1^2} \alpha_i(x) [\hat{\mu}_2 - \hat{\mu}_1 (X_i - x)]$, $\hat{\mu}_j = \frac{1}{n} \sum_{i=1}^n \alpha_i(x) (X_i - x)^j$.

Apart from the weight generating function, this form coincides with the local Fréchet regression proposed by [47]. Moreover, acquiring the local Fréchet regression estimators by directly from the corresponding explicit Euclidean forms is more straightforward. Our proposed RFWFR and RFWLLFR are both weighted Fréchet means, but the weights of RFWLLFR may be negative.

Classical local linear regression, random forests weighted local linear regression [10, 24], local Fréchet regression [47] and our method RFWLLFR are essentially all local linear estimators. The first two have explicit forms, and the latter two are obtained through implicit minimization. The relationships among the four type estimators are shown in Figure 2.

In order to study the consistency of RFWLLFR, we assume the follow conditions.

\begin{align*}
&\text{(A15)} \quad X \sim \text{Uniform} \left([0,1]^p\right). \\
&\text{(A16)} \quad N(L(x)) \to \infty. \\
&\text{(A17)} \quad \text{The Fréchet trees are trained in such a way that for each } y \in \Omega, \\
&\quad \max_{i,b} \left| \left\{ X_i \in L_b(x) \right\} \mid M_{\oplus}(X_i, y) - M_{\oplus}(x, y) \right| \xrightarrow{P} 0 \\
&\quad \text{That is, the leaf node containing } x \text{ shrink such that the maximal variation of the function } M_{\oplus} \text{ within a cell shrinks to 0 in probability for each } y \in \Omega. \\
&\text{(A18)} \quad \text{The objects } m_{\oplus}(x) \text{ and } \hat{l}_{\oplus}(x) \text{ exist and are unique, the latter almost surely. And for any } \varepsilon > 0, \\
&\quad \inf_{d(y, m_{\oplus}(x)) > \varepsilon} \left\{ M_{\oplus}(x, y) - M_{\oplus}(x, m_{\oplus}(x)) \right\} > 0.
\end{align*}
Assumptions (A15)-(A17) are the conditions used in [10] to establish the consistency of nonparametric regression estimators with random forests as adaptive nearest neighbors. They are all about the sample distribution and the random forests construction, which of course also apply to our proposed method. The assumption (A17) is a general condition, and (A17) can be deduced from assumptions (A2) and (A3). It is important to note here that the assumption (A16) is not required for the consistency of RFWFR. But it provides a guarantee that the law of large numbers can be used for the samples in the leaf node $L(x)$, which is critical to the proof of the consistency of RFWLLFR. Another reasonable interpretation is that the Fréchet random forests provide weights for the final local linear regression and should not use the forests to model strong, smooth signals to prevent overfitting phenomenon. In other words, the trees that form the forest here should not grow too deep. These can also be observed in our later simulations, where moderately grown trees can improve the performance of RFWLLFR. But RFWFR often requires deeper trees. In addition to the assumptions above, the honesty condition is still necessary to prove consistency of RFWLLFR.

**Theorem 4.1.** Suppose that for a fixed $x \in [0, 1]^p$, (A1), (A15)–(A18) hold and the Fréchet trees are honest. Then the RFWLLFR is point-wise consistent, that is,

$$d(\hat{l}_\Theta(x), m_\Theta(x)) = o_p(1).$$

5. Simulations. In this section, we consider three Fréchet regression scenarios to evaluate the performance of the two methods proposed above. The three scenarios adopt probability distributions, symmetric positive definite matrices, and spherical data as responses. We include the global Fréchet regression (GFR) and local Fréchet regression (LFR) methods [47], and the Fréchet random forests (FRF) [14] for comparisons. Throughout this section, GFR and LFR can be implemented by R-package "frechet" [16]. The parameter window width is chosen via cross-validation. And FRF can be implemented by the R-package "FrechForest" [13] with a slight modification through adding the three new types of responses and their corresponding metrics considered in this section into their package. RFWFR and RFWLLFR are also implemented in R. For simplicity, the Fréchet trees in our simulations are not necessarily honest. We use 100 trees for the construction of random forests. There are two hyper-parameters for each Fréchet tree: the depth of trees and the number of features randomly selected at each internal node. For a fair comparison, each method also choose the hyper-parameters by cross-validation. We consider $3 \sim \lceil log_2 n \rceil$ for the range of tuning about the depth of trees, where $n$ is the number of training samples.

For responses being symmetric positive definite matrices, the intrinsic local polynomial regression (ILPR) [58] and the manifold additive model (MAM) [39] are two promising tools which take advantage of the geometric structure of the Riemannian manifold. We also
included the two methods for comparisons for Fréchet regression with symmetric positive
definite matrices.

In the following simulations, each setting is repeated 100 times. For the $r$-th Monte Carlo
test of a particular sample size, $\hat{m}_{\oplus}$ denotes the fitted function based on the method $\hat{m}_{\oplus}$. The
quality of the estimation is measured quantitatively by the integrated squared errors (ISE)
$$ISE_r(\hat{m}_{\oplus}) = \int_{[0,1]^p} d^2(\hat{m}_{\oplus}(x), m_{\oplus}(x)) \, dx.$$ 
$ISE_r$ can be further approximated based on $N$ new random points
$$\hat{ISE}_r(\hat{m}_{\oplus}) = \frac{1}{N} \sum_{i=1}^N d^2(\hat{m}_{\oplus}(X_i), m_{\oplus}(X_i)).$$
We evaluate the performance of the method $\hat{m}_{\oplus}$ in terms of mean integrated squared error
(MISE), approximated by the average of 100 simulation runs
$$\text{(15)} \quad \text{MISE}(\hat{m}_{\oplus}) \approx \frac{1}{100} \sum_{r=1}^{100} \hat{ISE}_r(\hat{m}_{\oplus}).$$

5.1. Fréchet regression for distributions. Let $(\Omega, d)$ be the metric space of probability
distributions on $\mathcal{R}$ with finite second order moments and the quadratic Wasserstein metric $d_w$. For two such distributions $Y_1$ and $Y_2$, the squared Wasserstein distance is defined by
$$\text{(16)} \quad d^2_W(Y_1, Y_2) = \int_0^1 (Y_1^{-1}(t) - Y_2^{-1}(t))^2 \, dt,$$
where $Y_1^{-1}$ and $Y_2^{-1}$ are the quantile functions corresponding to $Y_1$ and $Y_2$, respectively.

We generate i.i.d $X_1, \ldots, X_n$ from a $p$-dimensional random vector whose components are
independent $U[0,1]$ random variables. We generate random normal distribution $Y$ with mean
and variance parameters as random variables dependent on $X$ such that the Fréchet regression
function is
$$m_{\oplus}(X) = E(Y(\cdot)|X) = \mu_X + \sigma_X \Phi(\cdot),$$
where $\Phi$ is the standard normal distribution function. In addition, $\mu_X$ and $\sigma_X$ are specified
in the following.

**Setting I-1:**
$$\mu_X \sim N\left(10(e_1^TX)^2 \left(2\beta^TX - 1\right), \sigma^2\right) \quad \text{and} \quad \sigma_X = 1,$$
For $p = 2$:
$$\beta = (0.75, 0.25).$$
For $p = 5, 10$:
$$\beta = (0.1, 0.2, 0.3, 0.4, 0, \ldots, 0)$$
For $p = 20$:
$$\beta = (0.1, 0.2, 0.3, 0.4, 0, \ldots, 0, 0.1, 0.2, 0.3, 0.4) / 2.$$

**Setting I-2:**
$$\mu_X \sim N\left(\sin\left(4\pi\beta_1^TX\right) \left(2\beta_2^TX - 1\right), \sigma^2\right) \quad \text{and} \quad \sigma_X = 2|X_1 - X_2|.$$
We also consider four situations of the dimension of $X : p = 2, 5, 10, 20$.  
For $p = 2$:

$$\beta_1 = (0.75, 0.25), \beta_2 = (0.25, 0.75).$$

For $p = 5, 10, 20$:

$$\beta_1 = (0.1, 0.2, 0.3, 0.4, 0, \ldots, 0), \beta_2 = (0, \ldots, 0, 0.1, 0.2, 0.3, 0.4).$$

In setting I-1, only the mean $\mu_X$ is related to $X$. But in setting I-2, the mean $\mu_X$ and standard deviation $\sigma_X$ are both related to $X$. And we set $n = 100, 200$ for $p = 2$; $n = 200, 500$ for $p = 5$; $n = 500, 1000$ for $p = 10$ and $n = 1000, 2000$ for $p = 20$. In particular when $p = 20$, the design of $\beta$ corresponds to the sparse high dimension Fréchet regression. For computation simplicity, the quantile function of each $Y_i$ is discretized as the 21 quantile points corresponding to the equispaced grids on $[0, 1]$. It can then be verified from (16) that the Wasserstein distance between the two distributions is actually the Euclidean distance between their quantile points. Therefore, the method FRF and our RFWFR should have the same output.

The simulation results are shown in Table 1. Since GFR is actually a global linear regression method, it is also suitable for the case when $X$ is high-dimensional like the other three methods based on Fréchet trees. However, as the mean function of $\mu_X$ is nonlinear, its performance is the worst and can be used as a benchmark for comparing other methods. From the results in this table we can see that the performance of GFR can not be improved significantly by simply increasing the number of training samples, but it performs better when the number of effective variables increases for setting I-1. LFR is a local method relying heavily on kernel smoothing. And the "frechet" package can only handle the case where the dimension of $X$ is less than 2. It can be found that LFR is the most unstable among all local methods. For setting I-1, the best performance is concentrated in RFWLLFR. As the dimension of $X$ increases, RFWLLFR is more stable than RFWFR, and its error does not increase much, indicating that it is easier to capture nonlinear signals. However, for the more complex setting I-2, the performance of RFWFR and RFWLLFR is close. Especially in the high-dimensional case, RFWFR begins to outperform RFWLLFR. We use boxplots (Figure 3) to show the results more clearly, under $p = 2, 10$. We also select several combinations of $(n, p)$ from setting I-2 to study the effect of noise $\sigma$ on the performance of each method. See the results summarized in Table 2. It can be seen that GFR is almost unaffected by noise. LFR performs poorly when the noise level is high. And our proposed RFWFR and RFWLLFR are still better than GFR and LFR in general.

Since bias tends to be the main source of error in random forests with enough number of trees, these trees tend to grow relatively deep. RFWFR is similar in nature to random forests, so it also prefers to use deeper trees. As for RFWLLFR, knowing that a more powerful local linear regression will be used for the final model fitting, we do not want to capture too much signal in the data during the construction of the random forest. So shallower trees are often used to avoid overfitting for RFWLLFR. For example, Figure 4 shows that the effect of the depth of Fréchet trees on the performance of our two methods based on setting I-1 with $p = 10$ and $n = 500$. RFWLLFR achieves the optimal performance when the depth of Fréchet trees is six, while RFWFR prefer deeper Fréchet trees.

5.2. Fréchet regression for symmetric positive-definite matrices. Let $(\Omega, d)$ be the metric space $S^n_{dd}$ of $m \times m$ symmetric positive-definite matrices endowed with certain metric $d$. There are many options for metrics, this section only focuses on the Log-Cholesky metric ([37]) and the affine-invariant metric ([22, 42, 46]). For a matrix $Y$, let $|Y|$ denotes the strictly lower triangular matrix of $Y$, $\mathbb{D}(Y)$ denotes the diagonal part of $Y$ and $\|Y\|_F$ denotes the Frobenius norm. It is well known that if $Y$ is a symmetric positive-definite matrix, there is
### Table 1

**MISE (standard deviation) of different methods for setting $I$ with $\sigma = 0.2$ for 100 simulation runs. Bold-faced number indicates the best performer.**

| $(p, n)$ | $\sigma = 0.1$ | $\sigma = 0.2$ | $\sigma = 0.5$ |
|----------|-----------------|-----------------|-----------------|
| $(2, 100)$ | 0.3026 (0.0283) | 0.3023 (0.0278) | 0.3032 (0.0274) |
| $(2, 200)$ | 0.2331 (0.0240) | 0.2335 (0.0241) | 0.2363 (0.0245) |
| $(5, 500)$ | 0.2434 (0.0295) | 0.2438 (0.0297) | 0.2462 (0.0302) |
| $(10, 1000)$ | 0.2452 (0.0285) | 0.2456 (0.0286) | 0.2479 (0.0288) |

### Table 2

**MISE (standard deviation) of different methods for $(2, 200), (5, 500), (10, 1000), (20, 2000)$ from setting I-2 with different $\sigma$ for 100 simulation runs. Bold-faced number indicates the best performer.**

| $(p, n)$ | $\sigma = 0.1$ | $\sigma = 0.2$ | $\sigma = 0.5$ |
|----------|-----------------|-----------------|-----------------|
| $(2, 100)$ | 0.3026 (0.0283) | 0.3023 (0.0278) | 0.3032 (0.0274) |
| $(2, 200)$ | 0.2331 (0.0240) | 0.2335 (0.0241) | 0.2363 (0.0245) |
| $(5, 500)$ | 0.2434 (0.0295) | 0.2438 (0.0297) | 0.2462 (0.0302) |
| $(10, 1000)$ | 0.2452 (0.0285) | 0.2456 (0.0286) | 0.2479 (0.0288) |

A lower triangular matrices $P$ whose diagonal elements are all positive such that $PP^T = Y$. This $P$ is called the Cholesky factor of $Y$, denoted as $\mathcal{L}(Y)$.

For an $m \times m$ symmetric matrix $A$, $\exp(A) = I_m + \sum_{j=1}^{\infty} \frac{1}{j!} A^j$ is a symmetric positive-definite matrix. Conversely, for a symmetric positive-definite matrix $Y$, the matrix logarithmic map is $\log(Y) = A$ such that $\exp(A) = Y$.

For two symmetric positive-definite matrices $Y_1$ and $Y_2$, the Log-Cholesky distance is defined by

$$d_L(Y_1, Y_2) = d_L(\mathcal{L}(Y_1), \mathcal{L}(Y_2))$$
where  
\[ d_C(P_1, P_2) = \left\{ \| P_1 - P_2 \|_F^2 + \| \log \mathbb{D}(P_1) - \log \mathbb{D}(P_2) \|_F^2 \right\}^{1/2} \]

And the affine-invariant distance is defined by
\[ d_A(Y_1, Y_2) = \left\| \log \left( Y_1^{-1/2} Y_2 Y_1^{-1/2} \right) \right\|_F. \]

We generate i.i.d \( X_1, \ldots, X_n \) from a \( p \)-dimensional random vector whose components are independent \( U[0, 1] \) random variables. And the response \( Y \) is generated via symmetric matrix variate Normal distribution. Consider the simplest case, we say an \( m \times m \) symmetric matrix \( A \sim N_{mm}(M; \sigma^2) \) if \( A = \sigma^2 Z + M \) where \( M \) is an \( m \times m \) symmetric matrix and \( Z \) is an \( m \times m \) symmetric random matrix with independent \( N(0, 1) \) diagonal elements and \( N(0, 1/2) \) off-diagonal elements. For more details on symmetric matrix variate Normal distribution, please refer to [59]. We first consider the following settings with \( Y \) being symmetric positive-definite matrices.
Setting II-1:

$$\log(Y) \sim N_{mn}(\log\{D(X)\}, \sigma^2),$$

where $D(X) = \exp\left(\begin{pmatrix} 1 & \rho(X) \\ \rho(X) & 1 \end{pmatrix}\right)$ and $\rho(X) = \cos (4\pi(\beta^T X))$. The choice of $\beta$ corresponds to $p = 2, 5, 10, 20$ is the same as setting I-1.

Setting II-2:

$$\log(Y) \sim N_{mn}(\log\{D(X)\}, \sigma^2),$$

where $D(X) = \begin{pmatrix} 1 & \rho_1(X) \rho_2(X) \\ \rho_1(X) & 1 \end{pmatrix}$ and $\rho_1(X) = 0.8 \cos(4\pi(\beta_1^T X))$, $\rho_2(X) = 0.4 \cos(4\pi(\beta_2^T X))$. The choice of $(\beta_1, \beta_2)$ corresponds to $p = 2, 5, 10, 20$ is the same as setting I-2.

We again compare our methods with GFR, LFR and FRF. From (17), it is easy to see that the Log-Cholesky distance between two symmetric positive-definite matrices is essentially the Frobenius distance between the matrices after some transformations. Therefore, similar to the Fréchet regression for distributions, the FRF and RFWFR would have the same output. However, for the affine-invariant distance, they are different. Results about setting II with
Log-Cholesky metric are showed in Table 3. When $p = 2$, LFR has the best performance. When $p > 2$, RFWLLFR performs the best in most cases. And the boxplots for $p = 2, 10$ shown in Figure 5 give a clearer presentation. The results with the affine-invariant metric summarized in Table 4 also advocates the local linear method RFWLLFR. Moreover, for this non-Euclidean type metric, we observe slight differences between FRF and RFWFR.

| Model | $(p, n)$ | GFR | LFR | FRF/RFWFR | RFWLLFR |
|-------|----------|-----|-----|-----------|---------|
| II-1  | (2, 100) | 1.264 (0.116) | 0.088 (0.039) | 0.177 (0.061) | 0.128 (0.072) |
|       | (2, 200) | 1.209 (0.089) | 0.038 (0.016) | 0.082 (0.019) | 0.054 (0.019) |
|       | (5, 200) | 1.281 (0.115) | NA | 0.507 (0.082) | 0.397 (0.096) |
|       | (5, 500) | 1.267 (0.104) | NA | 0.299 (0.046) | 0.190 (0.040) |
|       | (10, 500) | 1.279 (0.104) | NA | 0.586 (0.101) | 0.575 (0.101) |
|       | (10, 1000) | 1.253 (0.096) | NA | 0.420 (0.084) | 0.407 (0.082) |
|       | (20, 1000) | 0.973 (0.114) | NA | 0.623 (0.088) | 0.485 (0.075) |
|       | (20, 2000) | 0.956 (0.122) | NA | 0.522 (0.076) | 0.379 (0.058) |
| II-2  | (2, 100) | 1.932 (0.135) | 0.240 (0.098) | 0.367 (0.068) | 0.283 (0.079) |
|       | (2, 200) | 1.898 (0.152) | 0.109 (0.020) | 0.188 (0.035) | 0.143 (0.032) |
|       | (5, 200) | 1.980 (0.136) | NA | 0.855 (0.114) | 0.674 (0.116) |
|       | (5, 500) | 1.935 (0.143) | NA | 0.543 (0.065) | 0.369 (0.048) |
|       | (10, 500) | 1.971 (0.136) | NA | 1.057 (0.155) | 1.056 (0.134) |
|       | (10, 1000) | 1.949 (0.140) | NA | 0.845 (0.134) | 0.839 (0.119) |
|       | (20, 1000) | 1.970 (0.116) | NA | 1.251 (0.200) | 1.362 (0.171) |
|       | (20, 2000) | 1.962 (0.140) | NA | 1.071 (0.178) | 1.191 (0.158) |

Table 3
MISE (standard deviation) of different methods for setting II with $\sigma^2 = 0.2$ and Log-Cholesky metric for 100 simulation runs. Bold-faced number indicates the best performer.

The abelian group structure inherited from either the Log-Cholesky metric or the Log-Euclidean metric framework can turn the space of symmetric positive-definite matrices into a Riemannian manifold and further a bi-invariant Lie group [39]. [39] further proposed an additive model for the regression of symmetric positive-definite matrices valued response and an effective method called the manifold additive model (MAM). Their numerical studies show that the proposed method enjoys superior numerical performance compared with the intrinsic local polynomial regression (ILPR, [58]), especially when the underlying model is fully additive. However, [39] only consider $p = 3, 4$ in their simulation studies. In the next, we adopt the settings in [39] to make a comprehensive comparisons among MAM, ILPR, GFR, FRF, RFWFR, and RFWLLFR. MAM can be implemented with the R-package "matrix-manifold" [38].

Setting III: Let $X$ be a $p$-dimensional random vector whose components are independent $U[0, 1]$ random variables. The response $Y$ is generated via

$$Y = \mu \oplus w(X) \oplus \zeta,$$

where $\mu$ is the $3 \times 3$ identity matrix, $w(X) = \exp_{\tau_{\mu,e}} f(X)$, $e$ is the identity element of the group, $\exp_{\tau_{\mu,e}}$ denotes the parallel transport from $\mu$ to $e$, $\exp(\cdot)$ denotes the Lie exponential map, $\oplus$ denotes the group operation, and $\zeta$ is the random noise. The noise $\zeta$ is generated according to $\log \zeta = \sum_{j=1}^{6} Z_j v_j$, where $\log(\cdot)$ denotes the Lie log map, $Z_1, \ldots, Z_6$ are independently sampled from $N(0, \sigma^2)$, and $v_1, \ldots, v_6$ are an orthonormal basis of the tangent space $T_e S^+_3$. 

The signal-to-ratio (SNR) is measured by $\text{SNR} = E \| \log w (X) \|^2 / E \| \log \zeta \|^2$. Take the value of the parameter $\sigma^2$ to cover two choices for the SNR, namely, $\text{SNR} = 2$ and $\text{SNR} = 4$. Refer to [39] for the details of the notations and concepts here. We consider the following setting about $f(X)$.

**III-1:** $f (X) = \sum_{k=1}^q f_k (x_k)$ with $f_k (x_k)$ being an $3 \times 3$ matrix whose $(j, l)$-entry is $g(x_k; j, l, q) = \exp(-|j - l|/q) \sin(2q\pi (x_k - (j + l)/q))$.

**III-2:** $f (X) = f_{12} (x_1, x_2) \prod_{k=3}^q f_k (x_k)$, where $f_{12} (x_1, x_2)$ is an $3 \times 3$ matrix whose $(j, l)$-entry is $\exp\{-(j + l) (x_1 + x_2)\}$, and $f_k (x_k)$ is an $3 \times 3$ matrix whose $(j, l)$-entry is $\sin(2\pi x_k)$.

To maintain the consistency of the simulations, we use the same way as [39] to measure the quality of the estimation, which is slightly different from (15). For settings III-1 and III-2, we consider $p = 3, 4, 10, 20$ and two choices of $n$ for each $p$. For $p = 3, 4$, the setting is the same as [39]. For $p = 10, 20$, we increase the dimension of $X$, but $Y$ is still only related to the
first four components of $X$, i.e., $q = 4$. Table 5 shows the results. For the setting III-1 where the underlying model is additive, MAM shows clear advantages when $p$ is relatively small. Although the three methods based on Fréchet trees are not optimal, they are significantly better compared to ILPR. For non-additive setting III-2, FRF/RFWFR tends to perform the best. This setting also indicates that there are indeed cases where RFWLLFR will perform worse than RFWFR. It may be more efficient for complex settings to use RFWFR, whose mechanism is similar to that of random forests. Since both MAM and GFR make specific model assumptions, setting III-2 is not suitable for these two methods, although MAM performs slightly better than GFR. In particular, when $p$ is greater than 10, the implementations of MAM and ILPR often fail to work, and they are not feasible when the dimension of $X$ is too large. Through this experiment, we again demonstrate the outstanding performance of our method in the field of high-dimensional non-parametric Fréchet regression. In real applications, we often lack prior knowledge of the structure of the underlying model, so it is important to develop regression methods that suit all situations with excellent performance.

5.3. Fréchet regression for spherical data. Let $(\Omega, d)$ be the metric space $S^2$ of sphere data endowed with the geodesic distance $d_g$. For two points $Y_1, Y_2 \in S^2$, the geodesic distance is defined by

$$d_g(Y_1, Y_2) = \arccos(Y_1^T Y_2).$$

We generate i.i.d $X_1, \ldots, X_n$ from a $p$-dimensional random vector whose components are independent $U[0, 1]$ random variables. And $Y_i$ are generated by the following two settings.

**Setting IV-1**: Let the Fréchet regression function be

$$m_\oplus(X) = ((1 - (\beta_1^T X)^2)^{1/2}) \cos(\pi(\beta_1^T X)), (1 - (\beta_1^T X)^2)^{1/2}) \sin(\pi(\beta_2^T X)), \beta_1^T X)^T.$$

We generate binary Normal noise $\varepsilon_i$ on the tangent space $T_{m_\oplus(X_i)} S^2$, then map $\varepsilon_i$ back to $S^2$ by Riemannian exponential map to get $Y_i$. Specifically, we first independently generate

| Model | $(p, n)$ | FRF | RFWFR | RFWLLFR |
|-------|----------|-----|-------|---------|
| II-1  | (2, 100) | 0.161433 (0.041988) | 0.161432 (0.041991) | **0.124380 (0.046016)** |
|       | (2, 200) | 0.080703 (0.015133) | 0.080707 (0.015134) | **0.063739 (0.016001)** |
|       | (5, 200) | 0.408030 (0.063235) | 0.408010 (0.063239) | **0.332478 (0.073828)** |
|       | (5, 500) | 0.243646 (0.035725) | 0.243662 (0.035725) | **0.166254 (0.029714)** |
|       | (10, 200)| 0.080703 (0.015133) | 0.080707 (0.015134) | **0.063739 (0.016001)** |
|       | (10, 500)| 0.366012 (0.046278) | 0.366009 (0.046281) | **0.280171 (0.035996)** |
|       | (10, 1000)| 0.243646 (0.035725) | 0.243662 (0.035725) | **0.166254 (0.029714)** |

**MISE (standard deviation) of different methods for setting II with $\sigma^2 = 0.2$ and affine-invariant metric for 100 simulation runs. Bold-faced number indicates the best performer.**

**Table 4**
\[ \delta_1, \delta_2 \sim N (0, 0.2^2) \], then let \( \varepsilon_i = \delta_1 v_1 + \delta_2 v_2 \), where \( \{v_1, v_2\} \) forms an orthogonal basis of tangent space \( T_{m \oplus (X_i)} S^2 \). Then \( Y_i \) can be generated by

\[
Y_i = \text{Exp}_{m \oplus (X_i)} (\varepsilon_i) = \cos (\|\varepsilon_i\|) m_{\oplus} (X_i) + \sin (\|\varepsilon_i\|) \frac{\varepsilon_i}{\|\varepsilon_i\|}, \quad i = 1, \ldots, n
\]

where \( \| \cdot \| \) is the Euclidean norm.

We consider the following four situations of the dimension of \( X : p = 2, 5, 10, 20 \).

For \( p = 2 \):

\[ \beta_1 = (1, 0), \beta_2 = (0, 1). \]

For \( p = 5, 10, 20 \):

\[ \beta_1 = (0.1, 0.2, 0.3, 0.4, 0, \ldots, 0), \beta_2 = (0, \ldots, 0, 0.1, 0.2, 0.3, 0.4). \]

**Setting IV-2:** Consider the following model

\[
Y_i = (\sin(\beta_1^T X_i + \varepsilon_{i1}) \sin(\beta_2^T X_i + \varepsilon_{i2}), \sin(\beta_1^T X_i + \varepsilon_{i1}) \cos(\beta_2^T X_i + \varepsilon_{i2}), \cos(\beta_1^T X_i + \varepsilon_{i1}))^T,
\]

**Table 5**

| Model | \((p, n)\) | MAM | ILPR | GFR | FRF | RFWFR | RFWLLFR |
|-------|-------------|-----|------|-----|-----|-------|----------|
|       |             |     |      |     |     |       |          |
| III-1 | (3, 100)    | 0.415 (0.023) | 0.922 (0.126) | 0.970 (0.012) | 0.714 (0.018) | 0.794 (0.023) |
| (SNR = 2) | (3, 200)    | 0.299 (0.017) | 0.976 (0.064) | 0.956 (0.008) | 0.613 (0.013) | 0.655 (0.017) |
| | (4, 100)    | 0.527 (0.028) | 0.965 (0.033) | 0.986 (0.013) | 0.805 (0.018) | 0.923 (0.029) |
| | (4, 200)    | 0.357 (0.019) | 0.916 (0.021) | 0.970 (0.010) | 0.748 (0.014) | 0.832 (0.018) |
| III-2 | (SNR = 2)   | (10, 500) | NA | NA | 0.967 (0.008) | 0.774 (0.014) | 0.929 (0.015) |
| (SNR = 2) | (10, 1000)  | NA | NA | 0.959 (0.009) | 0.742 (0.012) | 0.879 (0.011) |
| | (20, 1000)  | NA | NA | 0.966 (0.009) | 0.778 (0.011) | 0.956 (0.014) |
| | (20, 2000)  | NA | NA | 0.958 (0.008) | 0.752 (0.010) | 0.917 (0.012) |
| III-1 | (SNR = 4)   | (10, 500) | NA | NA | 0.838 (0.061) | 0.681 (0.076) | 0.773 (0.068) |
| (SNR = 4) | (10, 1000)  | NA | NA | 0.829 (0.061) | 0.643 (0.077) | 0.718 (0.070) |
| | (20, 1000)  | NA | NA | 0.836 (0.061) | 0.736 (0.079) | 0.838 (0.074) |
| | (20, 2000)  | NA | NA | 0.830 (0.060) | 0.698 (0.078) | 0.783 (0.072) |

**MISE** (standard deviation) of different methods for setting III with SNR= 2, 4 and Log-Cholesky metric for 100 simulation runs. Bold-faced number indicates the best performer.
where the random noise $\varepsilon_{1i}, \varepsilon_{2i} \sim N(0, 0.2^2)$ are generated independently. The four situations correspond to $p = 2, 5, 10, 20$ are the same as setting IV-1.

Setting IV-1 is similar to [47, 59], and setting IV-2 is similar to [57]. For Fréchet regression with sphere data, we focus on the comparison of FRF, RFWFR, and RFWLLFR. For the geodesic distance $d_g$, the FRF and RFWFR will have different estimations. We summarize the estimation error in Table 6. RFWLLFR performs best in all cases. To vividly demonstrate the performances of RFWFR and RFWLLFR, we in Figure 6 exhibit the prediction of nine points with $p = 2$ and $n = 200$, which also verifies the advantage of RFWLLFR.

| Model | $(p, n)$ | FRF | RFWFR | RFWLLFR |
|-------|----------|-----|-------|----------|
| IV-1  | (2, 100) | 0.32329 (0.007413) | 0.32327 (0.007409) | **0.019839** (0.005842) |
|       | (2, 200) | 0.32078 (0.003967) | 0.32079 (0.003967) | **0.010781** (0.002704) |
|       | (5, 200) | 0.30237 (0.004693) | 0.30237 (0.004693) | **0.017935** (0.003892) |
|       | (5, 500) | 0.21684 (0.002725) | 0.21683 (0.002724) | **0.012164** (0.001998) |
|       | (10, 500)| 0.35012 (0.004252) | 0.35018 (0.004253) | **0.013339** (0.001915) |
|       | (10, 1000)| 0.27184 (0.003063) | 0.27188 (0.003065) | **0.010851** (0.001389) |
|       | (20, 1000)| 0.34698 (0.003879) | 0.34701 (0.003879) | **0.033069** (0.004379) |
|       | (20, 2000)| 0.29362 (0.003503) | 0.29362 (0.003502) | **0.028837** (0.003453) |
| IV-2  | (2, 100) | 0.010498 (0.002954) | 0.010501 (0.002954) | **0.004226** (0.001735) |
|       | (2, 200) | 0.007982 (0.001836) | 0.007984 (0.001837) | **0.003150** (0.000956) |
|       | (5, 200) | 0.008893 (0.001612) | 0.008894 (0.001612) | **0.005192** (0.001418) |
|       | (5, 500) | 0.006562 (0.001216) | 0.006563 (0.001216) | **0.002946** (0.000637) |
|       | (10, 500)| 0.008936 (0.001378) | 0.008938 (0.001378) | **0.005266** (0.000962) |
|       | (10, 1000)| 0.007110 (0.000951) | 0.007111 (0.000952) | **0.004190** (0.000626) |
|       | (20, 1000)| 0.009090 (0.001258) | 0.009096 (0.001259) | **0.006311** (0.000975) |
|       | (20, 2000)| 0.007427 (0.000938) | 0.007429 (0.000938) | **0.005117** (0.000692) |

Table 6
MISE (standard deviation) of different methods for setting IV with $\sigma = 0.2$ for 100 simulation runs. Bold-faced number indicates the best performer.
Fig 6: The plots of predictions of $m_{\geq}(x)$ given by RFWFR (left) and RFWLLFR (right) in a simulation run of $(p = 2, n = 200)$. The upper panels show results of setting IV-1, while the bottom show results of setting IV-2. The red points represent the real points, and the black points represent the predicted points.

6. Real Application. In this part, we apply the two methods proposed in this paper to a practical example with distribution function as the responses. Here we adopt the human mortality data, which is analyzed by [59] for Fréchet sufficient dimension reduction. The data are collected from United Nation Databases (http://data.un.org/) and UN World Population Prospects 2019 Databases (https://population.un.org/wpp/Download). The data is about the average number of deaths from 2010 to 2015 across 162 countries. The average number of deaths is recorded according to the different ranges of age. Specifically, the data is treated as histograms of death at age, with bin widths equal to 5 years. We can get the corresponding density function as our response value through these histograms of deaths versus age. Taking distribution function as the outcome of interest allows us to contain more information than summary statistics. As for the predictor variables, [59] consider the following 9 factors: (1) Population Density: population per square Kilometer; (2) Sex Ratio: number of males per 100 females in the population; (3) Mean Childbearing Age: the average age of mothers at the birth of their children; (4) Gross Domestic Product (GDP) per Capita; (5) Gross Value Added (GVA) by Agriculture: the percentage of agriculture, hunting, forestry, and fishing activities of gross value added; (6) Consumer price index: treat 2010 as the base year; (7) Unemployment Rate; (8) Expenditure on Health (percentage of GDP); (9) Arable Land (percentage of total land area). These factors involve population, economy, health, and geography, which are closely related to the human life span. We use the "frechet" package
to transform the histograms to smoothed probability density functions. For comparison, we also implement GFR, LFR and FRF. Since the R-package of LRF is only applicable when $X$ has no more than 2-dimensions, we then adopt the sufficient dimension reduced predictors given by [59]. For the the human mortality data, [59] verified that the structure dimension of the central space is 2 and obtained the following two directions for further modelling, which are

$$\hat{\beta}_1 = (0.080, -0.302, 0.359, 0.856, 0.159, -0.005, 0.025, 0.114, -0.028)^T,$$

$$\hat{\beta}_2 = (-0.009, -0.265, -0.610, 0.284, -0.661, 0.016, -0.194, 0.054, -0.001)^T.$$ We use these two directions to convert the 9-dimensional $X$ into 2-dimensional vectors $(\hat{\beta}_1^T X, \hat{\beta}_2^T X)$ as the input of LRF.

We then perform 9-fold cross-validation to evaluate the performance. Therefore, each training data contains 144 countries, and the testing data contains 18 countries. We use the best hyper-parameters based on the average testing error for each method. The final average testing error is 62.61 for GFR, 54.77 for LFR, 37.85 for FRF/RFWFR, and 43.16 for RFWLLFR. FRF/RFWFR has the best performance. The sample size (162) is not large enough for the 9-dimensional local regression problem. And there are significant differences between these 162 countries, which further increase the difficulty of regression.

To reflect the performance of each method more clearly, we give the plots of mortality density predictions versus the first estimated sufficient predictors $\beta_1^T X$ (see Figure 7). For reference, the first plot of Figure 7 is the smooth real density function fitted according to the human mortality data versus the first estimated sufficient predictors. The value of the first sufficient predictor largely reflects the degree of development of a country. The country with a small value of the first sufficient predictor has a backward medical level, which greatly increases the infant mortality rate. Please refer to [59] for more discussions. We also give the plots of the absolute differences between mortality densities prediction and the real density functions in Figure 8 for further comparisons. We can see that the areas where the predictions have a large error are the steep areas of the true density functions. Such areas include the elderly age group (80 ~ 100 years old) and the infant age group (near 0 years old). GFR predicts poorly in these two regions. It underestimates the infant mortality rate, and the prediction of the elderly age group shows a trend that the maximum value of the density function increases as the first sufficient predictor increases. Compared with GFR, LFR based on the two sufficient dimension reduced predictors greatly improved the prediction of the elderly age group but worsened the prediction of the infant age group. FRF/RFWFR optimize the prediction of the infant age group compared with LFR, and do a better job in more regions. RFWLLFR has the best performance in the infant age group, which shows that it has a greater advantage at boundary points. However, RFWLLFR performs poorly in the elderly age group. Overall speaking, RFWFR and RFWLLFR have complementary advantages for real data applications.
Fig 7: The top plot is the real mortality distributions versus the first sufficient predictors, and the remaining plots are the distributions predicted by each method versus the first sufficient predictors.
Fig 8: Absolute differences between the distributions predicted by each method and the real mortality distributions.

7. Conclusion. We provide two flexible and complementary Fréchet regression estimators based on random forests, which mitigate the curse of dimensionality for Fréchet regression. And the two methods certainly extend random forests to the case with metric space valued responses. We also establish the consistency, asymptotic convergence rate, and asymptotic normality of the proposed random forests weighted Fréchet regression estimator. Moreover, our theoretical findings include the most updated theoretical result of random forests with Euclidean responses as a special case. The theory developed for $M_{m_m}$-estimator based on infinite order U-processes and U-Statistics is of independent interest. Numerical performance demonstrated in both simulation studies and real applications advocates our proposals.

In the present work, we focus on the theory of random forests weighted Fréchet regression estimator. Theoretical investigation into the random forests weighted local linear Fréchet regression estimator including convergence rate and asymptotic normality is challenging for future research. Considering suitable ridge penalties as suggested by [24] in combination with our local linear method should be helpful for further achieving asymptotic normality. Moreover, uniform convergence rates of our proposals deserve further delicate analysis.
For Fréchet regression, our two methods only use the most basic information of a metric space. So our method has a wide range of applicability. When the metric space is a specific Riemannian manifold, more information can be considered in the construction of the model. Taylor expansions can be implemented on the tangent plane of the Riemannian manifold based on some specific geometric structure. And some advanced statistical tools designed for responses lying on the Riemannian manifold were developed, like the ILPR [58] and MAM [39]. For metric space being a specific Hilbert space, vector operations and an inner product structure are available, which inspires several promising nonparametric Hilbertian regression such as [31, 32, 33]. For the above two types of responses, we can consider nonparametric regression framework based on random forest kernels like our proposed RFWFR and RFWLLFR in future research. When more information of the output space is considered, models are expected to be more specific and targeted. It is of great interest to incorporate random forests for regression with specific structure information of non-Euclidean responses. Another interesting direction is to follow [14, 33] and study random forests weighted Fréchet regression when the predictors are also non-Euclidean. We leave these to future research.

REFERENCES

[1] Arcones, M. A. and Giné, E. (1993). Limit theorems for U-processes. *The Annals of Probability* 1494–1542.
[2] Arlot, S. and Genauer, R. (2014). Analysis of purely random forests bias. *arXiv preprint arXiv:1407.3939.*
[3] Athey, S., Tibshirani, J. and Wager, S. (2019). Generalized random forests. *The Annals of Statistics* 47:1148–1178.
[4] Bhattacharjee, S. and Müller, H.-G. (2021). Single Index Fréchet Regression. *arXiv preprint arXiv:2108.05437.*
[5] Bhattacharya, R. and Lin, L. (2017). Omnibus CLTs for Fréchet means and nonparametric inference on non-Euclidean spaces. *Proceedings of the American Mathematical Society* 145:413–428.
[6] Bhattacharya, R. and Patrangenaru, V. (2003). Large sample theory of intrinsic and extrinsic sample means on manifolds. *The Annals of Statistics* 31:1–29.
[7] Bhattacharya, R. and Patrangenaru, V. (2005). Large sample theory of intrinsic and extrinsic sample means on manifolds: II. *Annals of statistics* 1225–1259.
[8] Biau, G. (2012). Analysis of a random forests model. *The Journal of Machine Learning Research* 13:1063–1095.
[9] Biau, G., Devroye, L. and Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research* 9.
[10] Blomqvist, A., Talwalkar, A., Yu, B. and Wu, C. (2016). Supervised neighborhoods for distributed nonparametric regression. In *Artificial Intelligence and Statistics* 1450–1459. PMLR.
[11] Bose, A. and Chatterjee, S. (2018). U-statistics, Mm-estimators and Resampling. Springer.
[12] Breiman, L. (2001). Random forests. *Machine learning* 45:5–32.
[13] Capitaine, L. (2021). FrechForest: Frechet Random Forests. *R package version 0.95.*
[14] Capitaine, L., Bigot, J., Thibaut, R. and Genauer, R. (2019). Fréchet random forests for metric space valued regression with non euclidean predictors. *arXiv preprint arXiv:1906.01741.*
[15] Chen, X. and Kato, K. (2020). Jackknife multiplier bootstrap: finite sample approximations to the U-process supremum with applications. *Probability Theory and Related Fields* 176:1097–1163.
[16] Chen, Y., Gajardo, A., Fan, J., Zhong, Q., Dubey, P., Han, K., Bhattacharjee, S. and Müller, H.-G. (2020). frechet: Statistical Analysis for Random Objects and Non-Euclidean Data. *R package version 0.2.0.*
[17] Chen, Y. and Müller, H.-G. (2020). Uniform convergence of local Fréchet regression, with applications to locating extrema and time warping for metric-space valued trajectories. *arXiv preprint arXiv:2006.13548.*
[18] De la Peña, V. H. (1992). Decoupling and Khintchine's inequalities for U-statistics. *The Annals of Probability* 1877–1892.
[19] Denil, M., Matheson, D. and Freitas, N. (2013). Consistency of online random forests. In *International conference on machine learning* 1256–1264. PMLR.
[20] DiCiccio, C. and Romano, J. P. (2020). CLT for U-statistics with growing dimension.
[21] Durox, R. and Scornet, E. (2018). Impact of subsampling and tree depth on random forests. ESAIM: Probability and Statistics 22 96–128.
[22] Fletcher, P. T. and Joshi, S. (2007). Riemannian geometry for the statistical analysis of diffusion tensor data. Signal Processing 87 250–262.
[23] Fréchet, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. In Annales de l’institut Henri Poincaré 10 215–310.
[24] Friedberg, R., Tibshirani, J., Athey, S. and Wager, S. (2020). Local linear forests. Journal of Computational and Graphical Statistics 1–15.
[25] Gao, W. and Zhou, Z.-H. (2020). Towards Convergence Rate Analysis of Random Forests for Classification. Advances in Neural Information Processing Systems 33.
[26] Genuer, R. (2012). Variance reduction in purely random forests. Journal of Nonparametric Statistics 24 543–562.
[27] Heilig, C. and Nolan, D. (2001). Limit theorems for the infinite-degree U-process. Statistica Sinica 289–302.
[28] Heilig, C. M. (1997). An empirical process approach to U-processes of increasing degree. University of California, Berkeley.
[29] Hein, M. (2009). Robust nonparametric regression with metric-space valued output. In Advances in Neural Information Processing Systems 718–726. Citeseer.
[30] Ishwaran, H., Kogalur, U. B., Blackstone, E. H. and Lauer, M. S. (2008). Random survival forests. The annals of applied statistics 2 841–860.
[31] Jeon, J. M., Lee, Y. K., Mammen, E. and Park, B. U. (2021+). Locally polynomial Hilbertian additive regression. To appear in Bernoulli.
[32] Jeon, J. M. and Park, B. U. (2020). Additive regression with Hilbertian responses. The Annals of Statistics 48 2671–2697.
[33] Jeon, J. M., Park, B. U. and Van Keilegom, I. (2021). Additive regression for non-Euclidean responses and predictors. The Annals of Statistics 49 2611–2641.
[34] Klusowski, J. (2021). Sharp analysis of a simple model for random forests. In International Conference on Artificial Intelligence and Statistics 757–765. PMLR.
[35] Lakshminarayanan, B., Roy, D. M. and Teh, Y. W. (2014). Mondrian forests: Efficient online random forests. Advances in neural information processing systems 27 3140–3148.
[36] Lin, Y. and Jeon, Y. (2006). Random forests and adaptive nearest neighbors. Journal of the American Statistical Association 101 578–590.
[37] Lin, Z. (2019). Riemannian geometry of symmetric positive definite matrices via cholesky decomposition. SIAM Journal on Matrix Analysis and Applications 40 1353–1370.
[38] Lin, Z. (2020). Matrix-manifold: Basic Operations and Functions on Matrix Manifolds. R package version 0.1.0.
[39] Lin, Z., Müller, H.-G. and Park, B. U. (2020). Additive Models for Symmetric Positive-Definite Matrices, Riemannian Manifolds and Lie groups. arXiv preprint arXiv:2009.08789.
[40] Meinshausen, N. and Ridgeway, G. (2006). Quantile regression forests. Journal of Machine Learning Research 7.
[41] Mentch, L. and Hooker, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. The Journal of Machine Learning Research 17 841–881.
[42] Moakher, M. (2005). A differential geometric approach to the geometric mean of symmetric positive-definite matrices. SIAM Journal on Matrix Analysis and Applications 26 735–747.
[43] Mourtada, J., Gaiffas, S. and Scornet, E. (2017). Universal consistency and minimax rates for online mondrian forests. In Advances in neural information processing systems 3758–3767.
[44] Mourtada, J., Gaiffas, S. and Scornet, E. (2020). Minimax optimal rates for Mondrian trees and forests. The Annals of Statistics 48 2253–2276.
[45] Peng, W., Coleman, T. and Mentch, L. (2019). Asymptotic distributions and rates of convergence for random forests via generalized u-statistics. arXiv preprint arXiv:1905.10651.
[46] Pennec, X., Fillard, P. and Ayache, N. (2006). A Riemannian framework for tensor computing. International Journal of computer vision 66 41–66.
[47] Petersen, A. and Müller, H.-G. (2019). Fréchet regression for random objects with Euclidean predictors. The Annals of Statistics 47 691–719.
[48] Schotz, C. (2021). The Fréchet Mean and Statistics in Non-Euclidean Spaces, PhD thesis.
[49] Scornet, E. (2016). On the asymptotics of random forests. Journal of Multivariate Analysis 146 72–83.
[50] Scornet, E. (2016). Random forests and kernel methods. IEEE Transactions on Information Theory 62 1485–1500.
[51] Scornet, E., Biau, G. and Vert, J.-P. (2015). Consistency of random forests. The Annals of Statistics 43 1716–1741.
[52] SHERMAN, R. P. (1994). Maximal inequalities for degenerate U-processes with applications to optimization estimators. *The Annals of Statistics* 439–459.

[53] TADDY, M. A., GRAMACY, R. B. and POLSON, N. G. (2011). Dynamic trees for learning and design. *Journal of the American Statistical Association* 106 109–123.

[54] VAN DER VAART, A. and WELLNER, J. A. (1996). *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media.

[55] WAGER, S. and ATHEY, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* 113 1228–1242.

[56] WINHAM, S. J., FREIMUMTH, R. R. and BIERNACKA, J. M. (2013). A weighted random forests approach to improve predictive performance. *Statistical Analysis and Data Mining: The ASA Data Science Journal* 6 496–505.

[57] YING, C. and YU, Z. (2022). Fréchet Sufficient Dimension Reduction for Random Objects. *To appear in Biometrika*.

[58] YUAN, Y., ZHU, H., LIN, W. and MARRON, J. S. (2012). Local polynomial regression for symmetric positive definite matrices. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74 697–719.

[59] ZHANG, Q., XUE, L. and LI, B. (2021). Dimension Reduction and Data Visualization for Fréchet Regression. *arXiv preprint arXiv:2110.00467*.

[60] ZHU, R., ZENG, D. and KOSOROK, M. R. (2015). Reinforcement learning trees. *Journal of the American Statistical Association* 110 1770–1784.