Journal of Machine Learning Research 25 (2024) 1-69 Submitted 6/23; Revised 11/23; Published 3/24

Random Forest Weighted Local Fréchet Regression with Random Objects

Rui Qiu RQIU_STAT@OUTLOOK.COM
School of Statistics, KLATASDS-MOE
East China Normal University
Shanghai 200062, China

Zhou Yu* ZYU@STAT.ECNU.EDU.CN
School of Statistics, KLATASDS-MOE
East China Normal University
Shanghai 200062, China

Ruoqing Zhu RQZHU@ILLINOIS.EDU
Department of Statistics
University of Illinois at Urbana-Champaign
Champaign, IL 61820, USA

Editor: Genevera Allen

Abstract

Statistical analysis is increasingly confronted with complex data from metric spaces. Petersen and Müller (2019) established a general paradigm of Fréchet regression with complex metric space valued responses and Euclidean predictors. However, the local approach therein involves nonparametric kernel smoothing and suffers from the curse of dimensionality. To address this issue, we in this paper propose a novel random forest weighted local Fréchet regression paradigm. The main mechanism of our approach relies on a locally adaptive kernel generated by random forests. Our first method uses these weights as the local average to solve the conditional Fréchet mean, while the second method performs local linear Fréchet regression, both significantly improving existing Fréchet regression methods. Based on the theory of infinite order U-processes and infinite order M_{m_n}-estimator, we establish the consistency, rate of convergence, and asymptotic normality for our local constant 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 methods with several commonly encountered types of responses such as distribution functions, symmetric positive-definite matrices, and sphere data. The practical merits of our proposals are also demonstrated through the application to New York taxi data and human mortality data.

Keywords: metric space, Fréchet regression, random forest, nonparametric regression, infinite order U-process

1. Introduction

In recent years, non-Euclidean statistical analysis has received increasing attention due to demands from modern applications, such as the covariance or correlation matrices for functional brain connectivity in neuroscience and probability distributions in CT hematoma density data. To this end, Hein (2009) proposed nonparametric Nadaraya-Watson estimators for response variables be-

*. Corresponding author.

ing random objects, which are random elements in general metric spaces that by default do not have a vector space structure. Petersen and Müller (2019) 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. Chen and Müller (2022) continued to derive the uniform convergence rate of local Fréchet regression. Yuan et al. (2012) and Lin et al. (2022) considered nonparametric modeling with responses being symmetric positive-definite matrices, which are a specific type of random object. 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 (Zhang et al., 2023; Ying and Yu, 2022). Recently, Bhattacharjee and Müller (2023) and Ghosal et al. (2023) proposed single index Fréchet regression to resolve this dilemma but requires strong model assumptions.

Random forest, as pioneered by Leo Breiman (Breiman, 2001), is a popular and promising tool for relatively high-dimensional statistical learning for Euclidean data. It is an ensemble model that combines the strength of multiple randomized trees. Moreover, trees can be generated parallelly, making random forests more attractive computationally. Random forests demonstrate substantial gains in various learning tasks compared to classical statistical methods, such as survival analysis (e.g., Ishwaran et al., 2008; Steingrimsson et al., 2019). Theoretical research into random forests has gained considerable momentum in recent years due to their tremendous popularity. Biau et al. (2008) first proved the consistency of purely random forests for classification. For regression problems, Genuer (2012) and Arlot and Genuer (2014) further made a complete analysis of the variance and bias of purely random forests. Biau (2012) and Gao and Zhou (2020) established the consistency and convergence rate of the centered random forests for regression and classification, respectively. Duroux and Scornet (2018) provided the convergence rate of q-quantile random forests. In particular, Klusowski (2021) improved the rate of the median random forests. Scornet et al. (2015) proved the L^2 consistency of Breiman's original random forests for the first time under the assumption of additive model structure. Mentch and Hooker (2016) formulated random forests as infinite order incomplete U-statistics and studied their asymptotic normality. Wager and Athey (2018) 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 and predictors. In a recent study by Yao et al. (2022), a general framework based on forests was introduced for estimating a survival function considering time-varying covariates. Additionally, there is a significant interest in generalizing 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, Capitaine et al. (2019) proposed Fréchet trees and Fréchet random forests based on regression trees and Breiman's random forests. On the other hand, recent developments (e.g., Lin and Jeon, 2006; Meinshausen and Ridgeway, 2006; Bloniarz et al., 2016; Athey et al., 2019; Friedberg et al., 2020) 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 locally 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 superior performance and desirable statistical properties.

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 a kernel for Fréchet

regression. Our proposal called random forest weighted local constant Fréchet regression articulates a new formulation of Fréchet regression based on random forests that has an intrinsic relationship with classical nonparametric kernel regression. Second, compared to Capitaine et al. (2019), our method is more concise in terms of formulation, which allows us to take a substantial step towards the asymptotic theory of local Fréchet regression based on random forests rigorously. Following the line of research introduced by Wager and Athey (2018) based on trees with honesty and other properties, the consistency and rate of convergence are derived based on the theory of infinite order U-statistics and U-processes. To study the asymptotic normality, we extend the current theory of finite order M_m-estimator to infinite order M_{m_n}-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 (Wager and Athey, 2018) 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 local linear version (Bloniarz et al., 2016; Friedberg et al., 2020). This extension achieves better smoothness of the resulting estimator. The random forest weighted local constant Fréchet regression and local linear Fréchet regression collectively make up a coherent system and a 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 forest weighted local constant Fréchet regression (RFWLCFR) method. In Section 3, we establish the consistency and develop other asymptotic theories of RFWLCFR. In Section 4, we present the random forest weighted local linear Fréchet regression (RFWLLFR) approach as the generalization of RFWLCFR and confirm its consistency in estimation. In Section 5, a novel measure of variable importance is introduced. In Section 6, we conduct comprehensive simulation studies to examine our proposals in different settings, including probability distributions, symmetric positive-definite matrices, and spherical data. In Section 7, we apply our methods to the New York taxi data, where the response is a taxi ride network, and human mortality data, where the response is age-at-death distribution. Section 8 concludes the paper with some discussions. All proofs and additional material are presented in the appendices.

2. Proposed Method

Before formally presenting our first method in Section 2.2, we undertake some preliminary preparations. This encompasses furnishing a background introduction to Fréchet regression and explaining the process of constructing Fréchet trees, which is a fundamental constituent of our method.

2.1 Preliminaries

2.1.1 FRÉCHET REGRESSION

Let (Ω, d) be a metric space equipped with a specific metric d. Let $\mathcal{R}^p$ be the p-dimensional Euclidean space. We consider a random pair $(X, Y) \sim F$, where $X \in \mathcal{R}^p$, $Y \in \Omega$ and F is the joint distribution of (X, Y). We denote the marginal distributions of X and Y as F_X and F_Y, respectively. The conditional distributions $F_{X|Y}$ and $F_{Y|X}$ are also assumed to exist. When $\Omega \subseteq \mathcal{R}$, the target of classical regression is to estimate the conditional mean

$$m(x) = E\big(Y \mid X = x\big) = \underset{y \in \mathcal{R}}{\operatorname{argmin}} E\Big\{\big(Y - y\big)^2 \mid X = x\Big\}.$$

By replacing the Euclidean distance with the intrinsic metric d of Ω, conditional Fréchet mean (Petersen and Müller, 2019) can then be defined as

$$m_{\oplus}(x) = \underset{y\in\Omega}{\operatorname{argmin}} M_{\oplus}(x,y) = \underset{y\in\Omega}{\operatorname{argmin}} E\big\{d^2(Y,y) \mid X = x\big\}.$$

Given an i.i.d training sample $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$ 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, Hein (2009) generalized the Nadaraya-Watson regression to the Fréchet version as

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

where K is a smoothing kernel such as the Epanechnikov kernel or Gaussian Kernel and h is a bandwidth, with $K_h(\cdot) = h^{-1}K(\cdot/h)$. Petersen and Müller (2019) 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

$$\hat{m}_{\oplus}(x) = \underset{y\in\Omega}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} s_{in}(x) d^2\left(Y_i, y\right),$$

where $s_{in}(x)$ has different expressions for global and local Fréchet regression.

The Nadaraya-Watson Fréchet regression and local Fréchet regression both involve kernel weighting function K in the estimation procedure, which limits their applications when $p \geq 3$. To address this issue, we aim to borrow the strength of random forests to generate a more powerful weighting function for moderately large p. Figure 1 depicts flow statistics and predictive outcomes of yellow taxi traffic in Manhattan, New York, while further details are provided in Section 7. This problem can be formulated as a Fréchet regression problem where the response variable is a network (matrix) and 14 predictor variables are considered. Notably, when $p = 14$, both the Nadaraya-Watson Fréchet regression and local Fréchet regression techniques exhibit significant limitations. Here we employ Fréchet sufficient dimension reduction (Ying and Yu, 2022) to help realize the local Fréchet regression. The global Fréchet regression, although not restricted by dimensionality, relies on the assumption of linearity for satisfactory performance. Instead, it is evident from Figure 1 that the two methods we will propose in this paper have a higher prediction accuracy than global Fréchet regression.

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 (Fréchet) mean.

In this paper, we use Fréchet trees to refer to regression trees that handle metric space valued responses, regardless of their splitting criterion. Here we introduce an adaptive criterion—variance reduction splitting criterion, which uses information from both the predictor $X = (X^{(1)}, \ldots, X^{(p)})$ and response Y in the node splitting decision. The impurity of Y from a general metric space is no

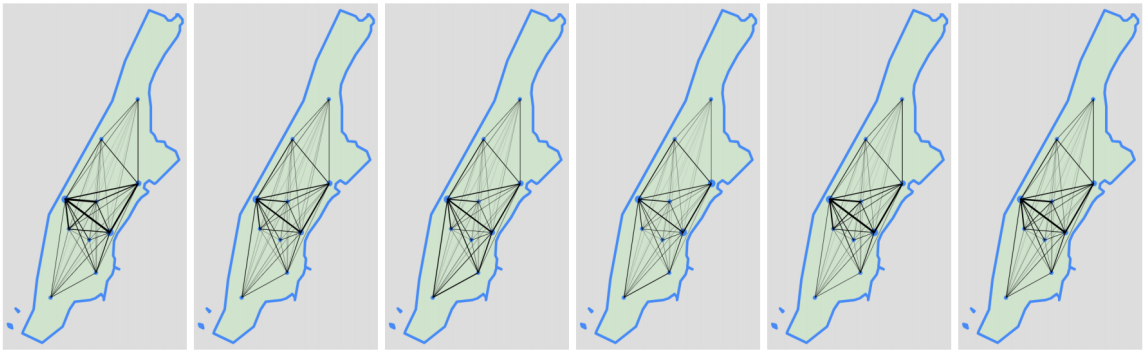


Figure 1: The first plot illustrates the flow statistics of yellow taxis in ten distinct zones of Manhattan, New York, during a certain time period. The thickness of the edges connecting vertices corresponds to the level of inter-zone traffic, while the size of vertices represents the total traffic volume within each zone. The remaining five plots from left to right are the predictions given by the global Fréchet regression, local Fréchet regression after dimension reduction, single index Fréchet regression, RFWLCFR and RFWLLFR.

longer measured by the variance under the Euclidean distance. Instead, we use the Fréchet variance. 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 $X^{(j)}$. We select the optimal (j_n^*, c_n^*) to decrease the sample Fréchet variance as much as possible so that the sample points grouped in the same child node exhibit a high degree of similarity. Specifically, the splitting criterion is

$$\mathcal{L}_n(j,c) = \frac{1}{N_n(A)} \left\{ \sum_{i:X_i \in A} d^2\left(Y_i, \bar{Y}_A\right) - \sum_{i:X_i \in A_{j,l}} d^2\left(Y_i, \bar{Y}_{A_{j,l}}\right) - \sum_{i:X_i \in A_{j,r}} d^2\left(Y_i, \bar{Y}_{A_{j,r}}\right) \right\},$$

where $A_{j,l} = \left\{x \in A : x^{(j)} < c\right\}, A_{j,r} = \left\{x \in A : x^{(j)} \geq c\right\}$, $N_n(A)$ is the number of samples falling into the node A, and $\bar{Y}_A = \operatorname{argmin}_{y \in \Omega} \sum_{i:X_i \in A} d^2\left(Y_i, y\right)$, *i.e.*, the sample Fréchet mean of Y_i's associated to the samples 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

$$(j_n^*, c_n^*) = \operatorname*{argmax}_{j,c} \mathcal{L}_n(j,c).$$

2.2 Local Constant Method

A single tree model may suffer from large bias or large variance depending on the tuning. To improve the predictive accuracy, we can aggregate multiple trees to form a random forest. The prediction error of random forests is closely related to the correlation among different trees. In addition to resampling the training data set for the growing of individual trees, auxiliary randomness is often introduced to further reduce the correlation between trees and thus improve the performance of random forests. For example, a subset of features is randomly selected before each split, and the split direction is designed based on the subset only. Here, we denote $\xi \sim \Xi$ as a source of auxiliary randomness.

We first consider the classical random forests with Euclidean responses. And each tree is trained on a subsample $\mathcal{D}_n^b = \left\{\left(X_{i_{b,1}}, Y_{i_{b,1}}\right), \left(X_{i_{b,2}}, Y_{i_{b,2}}\right), \ldots, \left(X_{i_{b,s_n}}, Y_{i_{b,s_n}}\right)\right\}$ of the training data set $\mathcal{D}_n$, with $1 \leq i_{b,1} < i_{b,2} < \ldots < i_{b,s_n} \leq n$. Throughout the paper, we assume that the subsample size $s_n \to +\infty$ and $s_n/n \to 0$ as n tends to infinity. Data resampling is done here without replacement (see Scornet et al. (2015); Mentch and Hooker (2016); Wager and Athey (2018)). The bth tree T_b constructed by $\mathcal{D}_n^b$ and a random draw $\xi_b \sim \Xi$ gives an estimator of $m(x)$

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

where $N(L_b(x; \mathcal{D}_n^b, \xi_b))$ is the number of samples in $L_b(x; \mathcal{D}_n^b, \xi_b)$, the leaf node containing x of T_b. For the random forest constructed by B randomized trees, $m(x)$ can be estimated by

$$\hat{r}(x) = \frac{1}{B} \sum_{b=1}^{B} T_b(x; \mathcal{D}_n^b, \xi_b) = \frac{1}{B} \sum_{b=1}^{B} \frac{1}{N(L_b(x; \mathcal{D}_n^b, \xi_b))} \sum_{i: X_i \in L_b(x; \mathcal{D}_n^b, \xi_b)} Y_i. \tag{1}$$

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

$$\hat{r}(x) = \sum_{i=1}^{n} \alpha_i(x) Y_i, \tag{2}$$

where $\alpha_i(x) = \frac{1}{B} \sum_{b=1}^{B} \frac{1\left\{X_i \in L_b(x; \mathcal{D}_n^b, \xi_b)\right\}}{N(L_b(x; \mathcal{D}_n^b, \xi_b))}$ is defined as the random forest kernel.

We now generalize the Euclidean random forests to the Fréchet version when Ω is a general metric space. The generalization process involves two distinct approaches, each aligning with one of the two perspectives presented in (1) and (2). We first rewrite the explicit expression of the random forest estimator (1) as the implicit minimizer of some objective function

$$\hat{r}(x) = \operatorname*{argmin}_{y \in \mathcal{R}} \frac{1}{B} \sum_{b=1}^{B} \left[\operatorname*{argmin}_{y' \in \mathcal{R}} \left\{ \frac{1}{N(L_b(x; \mathcal{D}_n^b, \xi_b))} \sum_{i: X_i \in L_b(x; \mathcal{D}_n^b, \xi_b)} \left(Y_i - y'\right)^2 \right\} - y \right]^2.$$

Then the first generalization for metric space valued responses is simply replacing the Euclidean distance by the metric d of Ω, that is,

$$\hat{r}_{\oplus}^{(1)}(x) = \operatorname*{argmin}_{y \in \Omega} \frac{1}{B} \sum_{b=1}^{B} d^2 \left(\operatorname*{argmin}_{y' \in \Omega} \frac{1}{N(L_b(x; \mathcal{D}_n^b, \xi_b))} \sum_{i: X_i \in L_b(x; \mathcal{D}_n^b, \xi_b)} d^2\left(Y_i, y'\right), y \right). \tag{3}$$

Alternatively, we can start from (2) and rewrite the random forest estimator as

$$\hat{r}(x) = \operatorname*{argmin}_{y \in \mathcal{R}} \sum_{i=1}^{n} \alpha_i(x) \left(Y_i - y\right)^2.$$

Then we propose the second generalization for metric space valued responses as

$$\hat{r}_{\oplus}^{(2)}(x) = \operatorname*{argmin}_{y \in \Omega} \sum_{i=1}^{n} \alpha_i(x) d^2\left(Y_i, y\right). \tag{4}$$

The first generalization (3) is actually the Fréchet random forest proposed by Capitaine et al. (2019). The idea behind it is to average the results of each Fréchet tree. However, our proposed second generalization (4) looks more concise because it involves only one "argmin", which brings convenience to our theoretical derivation. To acquire the random forest kernel $\alpha_i(x)$, we still need to construct all Fréchet trees. It is worth noting that when $\Omega \subseteq \mathcal{R}$, (3) and (4) are equivalent. However, for a general metric space Ω, (3) and (4) may not be the same. Therefore these are two different methods. Building upon the framework of the weighted Fréchet mean outlined in (4), we are able to systematically establish the asymptotic theory. These results fill the theoretical gap left in Capitaine et al. (2019) and encompass the theory for classical Euclidean random forests as a specific case. Additionally, this framework offers the potential for generalizing our method to a local linear version (see Section 4).

In the generalization (4), random forests produce the weighting function $\alpha_i(x)$ for Fréchet regression but do not participate in the prediction. Since (4) is essentially a local constant estimator based on the random forest kernel, we call it random forest weighted local constant Fréchet regression (RFWLCFR). Our proposal is expected to outperform Nadaraya-Watson Fréchet regression estimator (Hein, 2009) and the local Fréchet regression estimator (Petersen and Müller, 2019) for the following two reasons. Firstly, the random forest kernel can handle moderately large p. Secondly, if the split criterion of Fréchet trees depends on the response Y, the random forest kernel will be adaptive in the sense that it incorporates the information of both X and Y. Here we use a simulation example to illustrate the adaptiveness of a random forest kernel based on 100 Fréchet trees with the variance reduction splitting criterion introduced before.

Example 1 *Consider a Fréchet regression problem for a spherical response Y equipped with the geodesic distance and a predictor $X = (X^{(1)}, X^{(2)}) \sim \mathcal{U}\left([0,1]^2\right)$ as*

$$Y = \left(\sin(X^{(1)}+\varepsilon)\sin(X^{(1)}+\varepsilon), \sin(X^{(1)}+\varepsilon)\cos(X^{(1)}+\varepsilon), \cos(X^{(1)}+\varepsilon)\right)^{\mathrm{T}}, \tag{5}$$

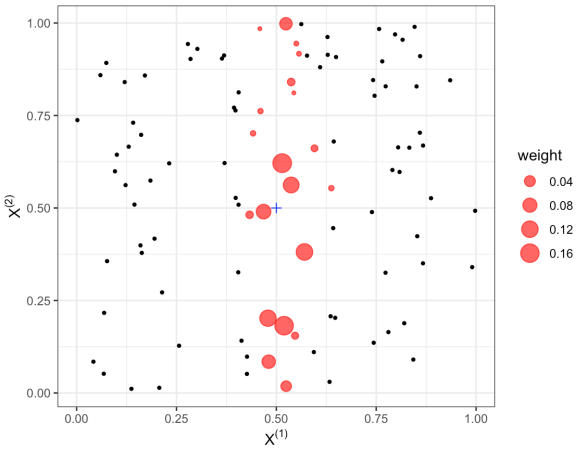


Figure 2: Weights given by the random forest kernel. Each point represents a training sample. The red points represent samples whose weights to $(0.5, 0.5)$ are greater than 0 and the diameter of these points indicates the size of the weights.

where $\epsilon \sim \mathcal{N}\left(0, 0.2^2\right)$. *The values of the random forest kernel at* 100 *training samples when making a prediction at the center* $(0.5, 0.5)$ *are displayed in Fig. 2. It can be observed that the weights decay much more quickly along the* $X^{(1)}$ *direction and are less influenced by the value of* $X^{(2)}$. *Under the construction mechanism of the random forest, samples that are close to the target point in the* $X^{(1)}$ *direction are considered important for prediction, which is consistent with the fact that* Y *is only relevant to* $X^{(1)}$ *in* (5). *Unlike the random forest kernel, the Euclidean distance based kernel does not have such an adaptive nature, and the local neighborhoods determined by it for the target point will not spread out in irrelevant directions.*

3. Theoretical Properties

In this section, we first introduce the population objective form of our method RFWLCFR, and then establish its theoretical properties from three perspectives: consistency, convergence rate, and asymptotic distribution.

3.1 Population Target

Here we consider a random pair $(X, Y) \sim F$, where X and Y take values in $[0,1]^p$ and Ω. To facilitate further theoretical investigations with the infinite order U-statistic and U-process tools, we follow Wager and Athey (2018) and assume that $B \to \infty$ given infinite computing power. Let

$$\bar{\alpha}_i(x) = \binom{n}{s_n}^{-1} \sum_k E_{\xi\sim\Xi} \frac{1\left\{X_i \in L(x; \mathcal{D}_n^k, \xi)\right\}}{N(L(x; \mathcal{D}_n^k, \xi))}, \tag{6}$$

where the summation about k is taken over all $\binom{n}{s_n}$ subsamples of size s_n, $\mathcal{D}_n^k$ is the kth subsample of $\mathcal{D}_n$, and the expectation is taken about the random effect ξ. $B \to \infty$ is equivalent to taking into account all Fréchet trees constructed by each subsample of $\mathcal{D}_n$ and all ξ conditioned on each subsample, which leads to the random forest kernel (6). Now we consider the infinite forest version

$$\hat{r}_{\oplus}(x) = \underset{y\in\Omega}{\operatorname{argmin}} \hat{R}_n(x, y) = \underset{y\in\Omega}{\operatorname{argmin}} \sum_{i=1}^{n} \bar{\alpha}_i(x)\, d^2(Y_i, y), \tag{7}$$

and develop the corresponding large sample theory. It can be observed that

$$\begin{aligned}
\hat{r}_{\oplus}(x) =& \underset{y\in\Omega}{\operatorname{argmin}} \sum_{i=1}^{n} \left[\binom{n}{s_n}^{-1} \sum_k E_{\xi\sim\Xi} \frac{1\left\{X_i \in L(x; \mathcal{D}_n^k, \xi)\right\}}{N(L(x; \mathcal{D}_n^k, \xi))} \right] d^2(Y_i, y) \\
=& \underset{y\in\Omega}{\operatorname{argmin}} \binom{n}{s_n}^{-1} \sum_k E_{\xi\sim\Xi} \left\{ \frac{1}{N(L(x; \mathcal{D}_n^k, \xi))} \sum_{i: X_i \in L(x; \mathcal{D}_n^k, \xi)} d^2(Y_i, y) \right\}.
\end{aligned} \tag{8}$$

As $s_n \to \infty$ when $n \to \infty$, the objective function $\hat{R}_n(x, y)$ of $\hat{r}_{\oplus}(x)$ is an infinite order U-statistic with rank s_n for any fixed $y \in \Omega$. The assumption that B is large enough for Monte Carlo effects not to matter is inspired by Scornet et al. (2015); Wager and Athey (2018); Cevid et al. (2022) and many other works. In practice, we can choose B as large as possible. In Remark 2, we provide a discussion regarding the issue of finite B.

Based on (7) and (8), we define two population level versions of $\hat{r}_{\oplus}(x)$ as follows.

$$\begin{aligned}\tilde{r}_{\oplus}(x) &= \underset{y\in\Omega}{\operatorname{argmin}}\tilde{R}_n(x,y) = \underset{y\in\Omega}{\operatorname{argmin}}\, nE\left\{\bar{\alpha}_i\left(x\right)d^2\left(Y_i,y\right)\right\} \\ &= \underset{y\in\Omega}{\operatorname{argmin}}E\left\{\frac{1}{N(L(x;\mathcal{D}_n^k,\xi))}\sum_{i:X_i\in L(x;\mathcal{D}_n^k,\xi)}d^2\left(Y_i,y\right)\right\},\end{aligned} \tag{9}$$

where the expectation is taken about all randomness. We can separate $d(\hat{r}_{\oplus}(x), m_{\oplus}(x))$ into the bias term $d(\tilde{r}_{\oplus}(x), m_{\oplus}(x))$ and the variance term $d(\hat{r}_{\oplus}(x), \tilde{r}_{\oplus}(x))$ for asymptotic analysis.

3.2 Consistency

To study the pointwise consistency of $\hat{r}_{\oplus}(x)$, we assume the following regularity conditions.

(A1) (Ω, d) is a bounded metric space, *i.e.*, $\operatorname{diam}(\Omega) = \sup_{y_1,y_2\in\Omega} d\left(y_1,y_2\right) < \infty$.

(A2) The marginal density f of X, as well as the conditional densities g_y of $X \mid Y = y$, exist and are bounded and continuous, the latter for all $y \in \Omega$. And f is also bounded away from zero such that $0 < f_{\min} \leq f$. Additionally, for any open $V \subseteq \Omega$, $\int_V \mathrm{d}F_{Y|X}(x,y)$ is continuous as a function of x.

(A3) $\operatorname{diam}\left(L(x)\right) \to 0$ in probability, where $L(x)$ is the leaf node containing x of any Fréchet tree in the random forest.

(A4) The object $m_{\oplus}(x)$ exists and is unique. For all n, $\tilde{r}_{\oplus}(x)$ and $\hat{r}_{\oplus}(x)$ exist and are unique, the latter almost surely. Additionally, for any $\varepsilon > 0$,

$$\begin{aligned}&\inf_{d(y,m_{\oplus}(x))>\varepsilon}\left\{M_{\oplus}(x,y) - M_{\oplus}\left(x,m_{\oplus}(x)\right)\right\} > 0, \\ &\liminf_n \inf_{d(y,\tilde{r}_{\oplus}(x))>\varepsilon}\left\{\tilde{R}_n(x,y) - \tilde{R}_n\left(x,\tilde{r}_{\oplus}(x)\right)\right\} > 0.\end{aligned}$$

Assumptions (A1), (A2) and (A4) are commonly used conditions to study the Fréchet regression, see Petersen and Müller (2019). If the termination condition for the growth of each Fréchet tree is that the number of samples in the leaf nodes does not exceed a certain constant, for example, the Fréchet tree is α-regular, which will be mentioned in Section 3.3, the assumption (A3) will hold (see Lemma 2 of Wager and Athey (2018)). Similar conditions can also be found in Denil et al. (2013). The assumption (A4) is also a regular condition to guarantee the consistency of M-estimators (see Corollary 3.2.3 of van der Vaart and Wellner (1996)). The simulation in Section 6 will consider three kinds of metric spaces: probability distributions equipped with the Wasserstein metric, symmetric positive definite matrices equipped with Log-Cholesky metric or the affine-invariant metric, and unit sphere equipped with geodesic distance. The first two can satisfy the assumption (A4) naturally or under very weak conditions. For the last one, the uniqueness of Fréchet means is generally not guaranteed but can be satisfied under certain circumstances, for example restricting the support of the underlying distribution.

In addition to assumptions (A1)–(A4), we further require that Fréchet trees are constructed with honesty and symmetry as defined in Wager and Athey (2018). More instructions about honesty are given in Appendix A and can easily be adapted to Fréchet trees.

(a) (*Honest*) The Fréchet tree is honest if the training examples whose responses have been used to decide where to place the splits can not be involved in the calculation of the random forest kernel.

(b) (*Symmetric*) The Fréchet tree is symmetric if the output of the tree does not depend on the order $(i = 1, 2, \dots)$ in which the training examples are indexed.

Theorem 1 *Suppose that for a fix $x \in [0,1]^p$, (A1)–(A4) hold and Fréchet trees are honest and symmetric. Then $\hat{r}_{\oplus}(x)$ is pointwise consistent, that is,*

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

Remark 2 *Building infinite Fréchet trees is actually computationally unfeasible. This issue also represents a perennial challenge associated with U-statistics, particularly when dealing with large n. In fact, the above results also hold true with $B = B_n < \binom{n}{s_n}$ where $B_n \to \infty$ as $n \to \infty$. Consider*

$$\hat{r}_{\oplus}(x) = \underset{y\in\Omega}{\operatorname{argmin}} \hat{R}_n(x,y) = \underset{y\in\Omega}{\operatorname{argmin}} \frac{1}{B_n} \sum_{b=1}^{B_n} \left\{ \frac{1}{N(L_b(x;\mathcal{D}_n^b,\xi_b))} \sum_{i:X_i\in L_b(x;\mathcal{D}_n^b,\xi_b)} d^2\left(Y_i, y\right) \right\}.$$

In this case, $\hat{R}_n(x,y)$ is called an incomplete infinite order U-statistic with a random kernel (about ξ_b) for each fixed $y \in \Omega$, which in general does not fit within the framework of infinite order U-statistics. Since the randomization parameter $\xi \sim \Xi$ here is independent of the training sample $\mathcal{D}_n$, the consistency of $\hat{R}_n(x,y) - \tilde{R}_n(x,y) = o_p(1)$ for each $y \in \Omega$ still holds. Then by the same proof as Theorem 1, $\hat{r}_{\oplus}(x)$ is still pointwise consistent with finite Fréchet trees. But the other tool of incomplete infinite order U-processes with a random kernel is not clear so far. So the convergence rate and asymptotic normality in the following content will still be developed under the constraint $B \to \infty$ due to the vast challenges of theoretical techniques.

If the previous assumptions are suitably strengthened (see the assumption (U1–U4) before the proof of the following theorem in the Appendix E), we can further obtain the uniform convergence results for $\hat{r}_{\oplus}(x)$. Let $\|\cdot\|$ be the Euclidean norm on $\mathcal{R}^p$ and $J > 0$.

Theorem 3 *Suppose that (A1), (U1)–(U4) hold and Fréchet trees are honest and symmetric. Then*

$$\sup_{\|x\|\le J} d(\hat{r}_{\oplus}(x), m_{\oplus}(x)) = o_p(1).$$

3.3 Rate of Convergence

We proceed to analyze the convergence rates of the bias term and the variance term separately. The convergence rate of the bias term is closely related to the construction of Fréchet trees. Here we follow Wager and Athey (2018) to place the following additional requirements on the construction of Fréchet trees.

(c) (*Random-split*) At each node split, marginalizing over ξ, the probability that $X^{(j)} (1 \le j \le p)$ is selected as the split variable is bounded below by π/p for some $0 < \pi \le 1$.

(d) (α-*regular*) After each splitting, each child node contains at least a fraction $\alpha > 0$ of the available training examples which will be used to calculate the random forest kernel. Moreover, the tree stops growing if every leaf node contains only between k and $2k-1$ observations, where k is some fixed integer.

To derive the convergence rates, some additional assumptions are required. We start with some notations. Let $Z_i = (X_i, Y_i)$ and $\mathcal{D}_n^k = \left(Z_{i_{k,1}}, Z_{i_{k,2}}, \ldots, Z_{i_{k,s_n}}\right)$, and define

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

Consider the function class

$$\mathcal{H}_\delta := \{H_n(z_1, \ldots, z_{s_n}, y) : d\left(y, \tilde{r}_\oplus(x)\right) < \delta\}.$$

Let $Z_i^0 = (X_i, Y_i)$, for $i = 1, \ldots, n$, and let $\left\{Z_i^1\right\}_{i=1}^n$ be i.i.d., independent of $\left\{Z_i^0\right\}_{i=1}^n$ with the same distribution. For $\forall H_n(y_1), H_n(y_2) \in \mathcal{H}_\delta$, define the following random pseudometric

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

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

With the above preparation, the assumptions are specified as follows.

(A5) For each y, $M_\oplus(x, y)$ is Lipschitz-continuous about x, and the Lipschitz constant has a common upper bound K.

(A6) There exist $\delta_1 > 0, C_1 > 0$ and $\beta_1 > 1$, possibly depending on x, such that, whenever $d\left(y, m_\oplus(x)\right) < \delta_1$, we have $M_\oplus(x, y) - M_\oplus\left(x, m_\oplus(x)\right) \geq C_1 d\left(y, m_\oplus(x)\right)^{\beta_1}$.

(A7) There exist constants A and V such that

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

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

(A8) There exist $\delta_2 > 0, C_2 > 0$ and $\beta_2 > 1$, possibly depending on x, such that, whenever $d\left(y, \tilde{r}_\oplus(x)\right) < \delta_2$, we have

$$\liminf_n \left\{\tilde{R}_n(x, y) - \tilde{R}_n\left(x, \tilde{r}_\oplus(x)\right) - C_2 d\left(y, \tilde{r}_\oplus(x)\right)^{\beta_2}\right\} \geq 0.$$

The Lipschitz continuity in the assumption (A5) allows us to control the bias term by restricting the diameter of the sample space represented by the leaf node. The assumption (A7) along with the pseudometric d_j were proposed by Heilig (1997) and Heilig and Nolan (2001) to establish the maximal inequality of infinite order U-processes. From the perspective of empirical process, (A7) regulates $\mathcal{H}_\delta$ a Euclidean class. Knowing that a class of functions is Euclidean aids immensely in establishing the convergence rate of the variance term. The assumptions (A6) and (A8) comes from Petersen and Müller (2019). (A8) is also an extension of the condition that controls the convergence rate of M-estimators. Please refer to Theorem 3.2.5 of van der Vaart and Wellner (1996) for more details. First, we establish the rate for the bias term as follows.

Lemma 4 *Suppose that for a fixed $x \in [0, 1]^p$, (A1), (A2), (A4), (A5) and (A6) hold, and the Fréchet trees are α-regular, random-split and honest. Then, as $n \to \infty$, provided that $\alpha \leq 0.2$, we have*

$$d(\tilde{r}_\oplus(x), m_\oplus(x)) = O\left(s_n^{-\frac{1}{2}\frac{\log(1-\alpha)}{\log(\alpha)}\frac{\pi}{p}\frac{1}{\beta_1 - 1}}\right).$$

Remark 5 *Consider the special case when $\Omega \subseteq \mathcal{R}$. Then*

$$M_{\oplus}(x,y) - M_{\oplus}\left(x, m_{\oplus}(x)\right) = \left\{y - m_{\oplus}(x)\right\}^2.$$

That is to say, the assumption (A6) holds when $\beta_1 = 2$. By the above lemma 4,

$$d(\tilde{r}_{\oplus}(x), m_{\oplus}(x)) = O\left(s_n^{-\frac{1}{2}\frac{\log(1-\alpha)}{\log(\alpha)}\frac{\pi}{p}}\right)$$

as $n \to \infty$. This rate coincides with Theorem 3 of Wager and Athey (2018), which indicates that our asymptotic bias result generalizes that of Euclidean random forests to random forests with metric space valued responses.

Then we turn to the convergence rate of the variance term.

Lemma 6 *Suppose that for a fixed $x \in [0,1]^p$, (A1), (A4), (A7) and (A8) hold, and the Fréchet trees are symmetric. Then we have*

$$d(\hat{r}_{\oplus}(x), \tilde{r}_{\oplus}(x)) = O_p\left(\left(\frac{s_n^2 \log s_n}{n}\right)^{\frac{1}{2(\beta_2-1)}}\right).$$

Remark 7 *When $\Omega \subseteq \mathcal{R}$, and if the trees are honest, then*

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

And we can further get

$$\begin{aligned}
&\tilde{R}_n(x,y) - \tilde{R}_n\left(x, \tilde{r}_{\oplus}(x)\right) \\
=&E\left[E\left\{(Y-y)^2 \mid X \in L(x)\right\}\right] - E\left[E\left\{(Y-\tilde{r}_{\oplus}(x))^2 \mid X \in L(x)\right\}\right] \\
=&y^2 - 2yE\left[E\left\{Y \mid X \in L(x)\right\}\right] + \left(E\left[E\left\{Y \mid X \in L(x)\right\}\right]\right)^2 \\
=&\left\{y - \tilde{r}_{\oplus}(x)\right\}^2,
\end{aligned}$$

which indicates that $\beta_2 = 2$ in the assumption (A8) for Euclidean responses. By the result of Lemma 6, $d(\hat{r}_{\oplus}(x), \tilde{r}_{\oplus}(x)) = O_p\big((s_n^2 \log s_n/n)^{1/2}\big)$. The rate derived here is slower than $(s_n/n)^{1/2}$ as described in our asymptotic normality result (Remark 15 in Appendix B). The reason is that the existing maximal inequalities of infinite order U-processes are not strong enough. More refined tools of infinite order U-processes are expected to be developed to further improve this convergence rate.

Combining Lemma 4 and Lemma 6, we get the convergence rate for $\hat{r}_{\oplus}(x)$.

Theorem 8 *Suppose that for a fixed $x \in [0,1]^p$, (A1), (A2), (A4)–(A8) hold, and Fréchet trees are α-regular, random-split, honest and symmetric. If $\alpha \leq 0.2$, then*

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

3.4 Asymptotic Normality

There are two major challenges in deriving the asymptotic normality of $\hat{r}_{\oplus}(x)$. On the one hand, $\hat{r}_{\oplus}(x)$ has no explicit expression like Euclidean random forests and it is not the classical M-estimator, but M_m-estimator (Bose and Chatterjee, 2018) and even 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 Bose and Chatterjee (2018) to acquire the probability representation and asymptotic normality of the M_{m_n}-estimator. As for the second issue, the seminal work of Bhattacharya and Lin (2017) and Bhattacharya and Patrangenaru (2003, 2005) 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 $\hat{r}_{\oplus}(x)$ finally. While these difficulties have been reasonably addressed, the conditions under which asymptotic normality holds are technical and not easily verifiable in practice. Consequently, the asymptotic normality result is primarily intended for theoretical completeness and may not be convenient for practical statistical inference. Therefore we have placed this section in the Appendix B for interested readers. It is noteworthy that our result encompasses the asymptotic normality of the Euclidean random forest as a special case. In addition, the theory developed for M_{m_n}-estimator based on infinite order U-processes and U-Statistics is of independent interest. Overall, the development of a more practically significant asymptotic distribution theory remains a challenging endeavor for our problem. When Y comes from specialized metric spaces such as Riemannian manifolds or Hilbert spaces, the incorporation of additional space properties may lead to further breakthroughs. However, it is beyond the scope of this paper.

4. Local Linear Smoothing

In Section 2, we have proposed RFWLCFR, which is a Nadaraya-Watson type estimator using the random forest kernel. A very natural extension is to carry out local linear Fréchet regression further. The local linear estimator is more flexible and accurate in capturing smooth signals. Local Fréchet regression proposed by Petersen and Müller (2019) is a novel local linear estimator adapted to cases with metric space valued responses. Similar to the classical local linear estimator, their method still suffers from the curse of dimensionality. This section proposes the second method called random forest weighted local linear Fréchet regression (RFWLLFR), which adopts the random forest kernel to local linear Fréchet regression.

Bloniarz et al. (2016) and Friedberg et al. (2020) considered a local linear regression with Euclidean responses based on the random forest kernel

$$\left(\hat{\beta}_0, \hat{\beta}_1\right) = \underset{\beta_0,\beta_1}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \alpha_i(x) \left\{Y_i - \beta_0 - \beta_1^{\mathrm{T}} \left(X_i - x\right)\right\}^2, \tag{10}$$

where $\alpha_i\left(x\right) = \frac{1}{B}\sum_{b=1}^{B} \frac{1\left\{X_i \in L_b(x;\mathcal{D}_n^b,\xi_b)\right\}}{N(L_b(x;\mathcal{D}_n^b,\xi_b))}$. Then define the random forest weighted local linear estimator as

$$\hat{l}(x) = \hat{\beta}_0 = e_1^{\mathrm{T}}(\tilde{X}^{\mathrm{T}} A \tilde{X})^{-1} \sum_{i=1}^{n} \begin{pmatrix} 1 \\ X_i - x \end{pmatrix} \alpha_i(x) Y_i \tag{11}$$

where

$$\tilde{X} := \begin{pmatrix} 1 & (X_1 - x)^{\mathrm{T}} \\ 1 & (X_2 - x)^{\mathrm{T}} \\ \vdots & \vdots \\ 1 & (X_n - x)^{\mathrm{T}} \end{pmatrix}, A := \operatorname{diag}\left(\alpha_1(x), \ldots, \alpha_n(x)\right), e_1 := (1, 0, \cdots, 0)^{\mathrm{T}}.$$

To generalize (11) with metric space valued responses, we rewrite it as an implicit form

$$\hat{l}(x) = \underset{y \in \mathcal{R}}{\operatorname{argmin}}\, e_1^{\mathrm{T}} (\tilde{X}^{\mathrm{T}} A \tilde{X})^{-1} \sum_{i=1}^{n} \begin{pmatrix} 1 \\ X_i - x \end{pmatrix} \alpha_i(x)(Y_i - y)^2.$$

Replacing the Euclidean distance with a metric d, the RFWLLFR for a general metric space (Ω, d) is proposed as

$$\hat{l}_{\oplus}(x) = \underset{y \in \Omega}{\operatorname{argmin}}\, e_1^{\mathrm{T}} (\tilde{X}^{\mathrm{T}} A \tilde{X})^{-1} \sum_{i=1}^{n} \begin{pmatrix} 1 \\ X_i - x \end{pmatrix} \alpha_i(x) d^2(Y_i, y). \tag{12}$$

Our proposed local constant and local linear methods are both to calculate weighted Fréchet means, but the weights of the local linear method may be negative. Due to the complexity of RFWLLFR, we only study the consistency of $\hat{l}_{\oplus}(x)$ here. To this end, we assume the following conditions.

(A9) $X \sim \mathcal{U}\left([0,1]^p\right)$, the uniform distribution on $[0,1]^p$.

(A10) $N(L_b(x; \mathcal{D}_n^b, \xi_b)) \to \infty$ for $b = 1, \ldots, B$.

(A11) The Fréchet trees are trained in such a way that for each $y \in \Omega$

$$\max_{1 \le i \le n, 1 \le b \le B} \left[1\{X_i \in L_b(x; \mathcal{D}_n^b, \xi_b)\} \left| M_{\oplus}(X_i, y) - M_{\oplus}(x, y) \right| \right] \xrightarrow{p} 0.$$

That is, the leaf node containing x shrinks such that the maximal variation of the function $M_{\oplus}(\cdot, y)$ within a cell shrinks to 0 in probability for each $y \in \Omega$.

(A12) $m_{\oplus}(x)$ and $\hat{l}_{\oplus}(x)$ exist and are unique, the latter almost surely. For any $\varepsilon > 0$,

$$\inf_{d(y, m_{\oplus}(x)) > \varepsilon} \left\{ M_{\oplus}(x, y) - M_{\oplus}\left(x, m_{\oplus}(x)\right) \right\} > 0.$$

Assumptions (A9)-(A11) are similar to the conditions used in Bloniarz et al. (2016) to establish the consistency of a nonparametric regression estimator using random forests as adaptive nearest neighbor generators. The assumption (A11) is a general condition, which can be deduced from assumptions (A2) and (A3). It is important to note that assumption (A10) is not required for the consistency of our local constant method. But it provides a guarantee that the law of large numbers can be used for the samples in the leaf node, which derives the consistency of $\hat{l}_{\oplus}(x)$ even when B in the random forest kernel is a fixed constant. Another reasonable interpretation is that the random forest provides weights for the final local linear regression and should not be used to model strong, smooth signals to prevent overfitting phenomena. In other words, the Fréchet trees that form the forest here should not grow too deep. These can also be observed in Figure 12 from Appendix C.2, where moderately grown trees can improve the performance of $\hat{l}_{\oplus}(x)$. But our local constant method often requires deeper trees. In addition to the assumptions above, the honesty condition is still necessary to prove the consistency of $\hat{l}_{\oplus}(x)$.

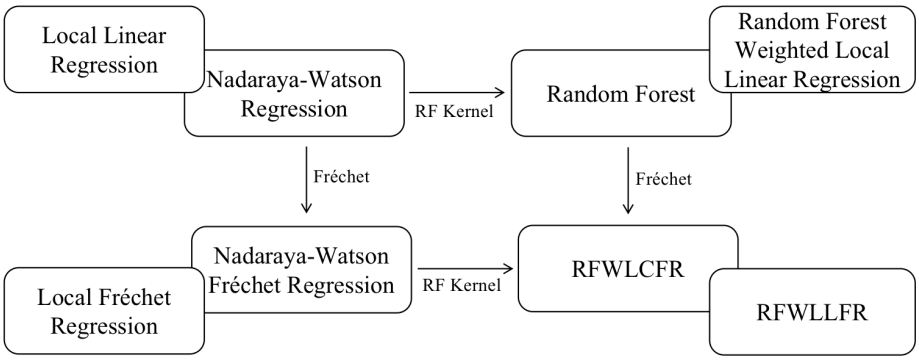


Figure 3: The relationships among the eight local estimators.

Theorem 9 *Suppose that for a fixed $x \in [0,1]^p$, (A1), (A9)–(A12) hold and Fréchet trees are honest. Then $\hat{l}_{\oplus}(x)$ is pointwise consistent, that is,*

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

Remark 10 *As can be seen from Remark 2, we have weakened the requirement of our local constant method for the number B of trees, which eliminates the gap between theoretical investigations and practical applications. Certainly, if we reimpose assumptions of the local linear method here on it, a parallel proof guarantees that B can be further weakened to a fixed constant.*

After introducing our two methods, we briefly summarize all relevant methods. The classical Nadaraya-Watson regression, random forest, Nadaraya-Watson Fréchet regression (Hein, 2009) and RFWLCFR are essentially all local constant estimators. The classical local linear regression, random forest weighted local linear regression (Bloniarz et al., 2016; Friedberg et al., 2020), local Fréchet regression (Petersen and Müller, 2019) and RFWLLFR are essentially all local linear estimators. The relationship among the eight types of estimators is shown in Fig. 3.

5. Forest Kernel-based Permutation Variable Importance

As a byproduct of the classical random forest algorithm, one can effectively assess the importance of each variable using the out-of-bag (oob) samples (Breiman, 2001), which refer to those observations not used in constructing the tree. Specifically, when the bth tree is grown, the oob samples are passed down the tree, and the prediction accuracy is recorded. Then, the jth variable values are randomly permuted within the oob samples, and accuracy is re-evaluated. The average decrease in accuracy, resulting from this permutation, serves as a measure of the importance of variable j in the random forest. The above process can be readily extended to non-Euclidean scenarios by replacing the Euclidean distances with general distances. A notable disadvantage of this type of permutation-based measure is that it is formulated at the tree level, and averaged over the forest. Hence this method does not correctly reflect the reduced accuracy on the forest kernel due to the permutation. Here, we introduce a novel method that uses the forest kernel for oob permutation variable importance.

For each sample (X_i, Y_i) from the entire training data $\mathcal{D}_n$, we can collect the Fréchet trees whose construction (X_i, Y_i) did not participate in. Consequently, (X_i, Y_i) serves as a natural test

point for a small forest composed of these trees. We predict the response of X_i by RFWLCFR based on the kernel generated by this small forest and record the prediction error. Note that the oob prediction mechanic here is also similar to the ones used in the jackknife confidence interval (Wager et al., 2014). These errors are then averaged over all training samples as a baseline. To quantify the importance of the jth variable, we randomly permute the value for the variable j within $\mathcal{D}_n$. The prediction error is recalculated by treating each permuted observation as a testing point. The decrease in accuracy on the permuted data compared to the baseline is a variable importance measure for variable j. The complete algorithm is summarized in Algorithm 1. Although Algorithm 1 incurs more computational overhead compared to the classical tree-level measure, as it involves identifying unrelated trees and making predictions for each observation, the extra computational cost is not significant. The advantage of our method lies in its higher precision, given that the prediction process is executed based on a forest-level kernel. In addition, repeating the algorithm multiple times and averaging the variable importance can enhance result stability. In cases where there is an excessive number of features, the variable importance ranking obtained from Algorithm 1 can guide us in discarding less important variables, thereby mitigating the challenges arising from the curse of dimensionality.

Algorithm 1 : Variable importance calculation

Inputs: A training set $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$, number of Fréchet trees B.

Step 1. Construct a random forest consisting of B Fréchet trees $\{T_b(x; \mathcal{D}_n^b, \xi_b)\}_{b=1}^B$ based on $\mathcal{D}_n$, which generate the random forest kernel for the achievement of RFWLCFR.

Step 2.

for $i = 1$ to n **do**

Identify the collection $\mathcal{T}_i$ of Fréchet trees whose growth (X_i, Y_i) did not participate in: $\mathcal{T}_i = \{T_b(x; \mathcal{D}_n^b, \xi_b) : 1 \le b \le B, (X_i, Y_i) \notin \mathcal{D}_n^b\}$.

Predict the response of X_i with RFWLCFR, denoted by $\hat{r}_{\oplus}^{oob}(X_i)$, based on the random forest kernel provided by $\mathcal{T}_i$.

end for

Record the mean square error: $R_0 = \frac{1}{n}\sum_{i=1}^n d^2(\hat{r}_{\oplus}^{oob}(X_i), Y_i)$.

Step 3.

for $j = 1$ to p **do**

Permute the values for the jth variable randomly in $\{X_i\}_{i=1}^n$ and repeat Step 2 with the permuted data and the same $\mathcal{T}_i, 1 \le i \le n$, acquired in Step 2; Record the corresponding mean square error R_j.

end for

Step 4. Calculate the variable importance for the jth variable: $\mathrm{VI}(X^{(j)}) = R_j - R_0, 1 \le j \le p$.

6. Simulations

In this section, we consider three Fréchet regression scenarios including probability distributions, symmetric positive definite matrices, and spherical data to evaluate the performance of the two methods proposed in this paper. We include the global Fréchet regression (GFR) and local Fréchet regression (LFR) (Petersen and Müller, 2019), and the Fréchet random forest (FRF) (Capitaine et al., 2019) for comparisons. Additionally, the single index Fréchet regression (IFR) (Bhat-

tacharjee and Müller, 2023) is also considered for a setting of the single index model with symmetric positive-definite matrix responses. Throughout this section, GFR and LFR can be implemented by R-package "frechet" (Chen et al., 2020). FRF can be implemented by the R-package "FrechForest" (Capitaine, 2021) with a slight modification by adding the three new types of responses and their corresponding metrics into the package. And Julia code for the implementation of IFR can be found in the GitHub platform. Our RFWLCFR and RFWLLFR are also implemented in R. For simplicity, the Fréchet trees in our simulations are not necessarily honest. All random forests are constructed by 100 Fréchet trees. There are three hyperparameters for each Fréchet tree: the size s_n of each subsample, the depth of Fréchet trees and the number of features randomly selected at each internal node. The choice of s_n is very tedious and time-consuming. Here we instead acquire all subsamples by sampling from the training data set $\mathcal{D}_n$ with replacement, which is commonly used in random forest codes. When the size n of $\mathcal{D}_n$ is large enough, each subsample is expected to have the fraction $(1-1/e) \approx 63.2\%$ of the unique examples of $\mathcal{D}_n$. We consider $3 \sim \lceil \log_2 n \rceil$ for the range of tuning about the depth of Fréchet trees, where n is the number of training samples. For a fair comparison, each method chooses the hyperparameters by cross-validation.

In the following simulations, each setting is repeated 100 times. For the rth Monte Carlo test, $\hat{m}_\oplus^r$ denotes the fitted function based on the method $\hat{m}_\oplus$ and the quality of the estimation is measured quantitatively by the mean squared error

$$\mathrm{MSE}_r(\hat{m}_\oplus) = \frac{1}{100} \sum_{i=1}^{N} d^2(\hat{m}_\oplus^r(X_i), m_\oplus(X_i))$$

based on 100 new testing points.

6.1 Fréchet Regression for Distributions

Let (Ω, 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

$$d_W^2(Y_1, Y_2) = \int_0^1 \left\{ Y_1^{-1}(t) - Y_2^{-1}(t) \right\}^2 dt, \tag{13}$$

where Y_1^{-1} and Y_2^{-1} are quantile functions corresponding to Y_1 and Y_2, respectively.

Let $X_1, \ldots, X_n \sim \mathcal{U}([0,1]^p)$, and we generate random normal distribution Y by

$$Y = \mathcal{N}\left(\mu_Y, \sigma_Y^2\right),$$

where μ_Y and σ_Y are random variables dependent on X as described in the following.

Setting I-1:

$$\mu_Y \sim \mathcal{N}\left(5\beta^{\mathrm{T}} X - 2.5, \sigma^2\right), \quad \sigma_Y = 1.$$

We consider four situations of the dimension of X : $p = 2, 5, 10, 20$.
(i) $p = 2$: $\beta = (0.75, 0.25)$;
(ii) $p = 5, 10$: $\beta = (0.1, 0.2, 0.3, 0.4, 0, \ldots, 0)$;
(iii) $p = 20$: $\beta = (0.1, 0.2, 0.3, 0.4, 0, \ldots, 0, 0.1, 0.2, 0.3, 0.4)/2$.

Setting I-2:

$$\mu_Y \sim \mathcal{N}\left(\sin\left(4\pi\beta_1^{\mathrm{T}}X\right)\left(2\beta_2^{\mathrm{T}}X-1\right),\sigma^2\right), \quad \sigma_Y = 2|e_1^{\mathrm{T}}X - e_2^{\mathrm{T}}X|,$$

where e_i is a vector of zeros with 1 in the ith element. We also consider four situations.
(i) For $p = 2$: $\beta_1 = (0.75, 0.25), \beta_2 = (0.25, 0.75)$.
(ii) For $p = 5, 10, 20$: $\beta_1 = (0.1, 0.2, 0.3, 0.4, 0, \ldots, 0)$, $\beta_2 = (0, \ldots, 0, 0.1, 0.2, 0.3, 0.4)$.

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 $p = 20$. 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 (13) that the Wasserstein distance between the two distributions is actually the Euclidean distance between their quantile points. Therefore, our RFWLCFR and the method FRF will have the same output.

For the fairness of comparison, we begin with the setting I-1 where data are generated in a linear way. It is compliant with the initial assumptions of GFR. The results are recorded in Table 1. Not surprisingly, GFR is the best performer, followed by RFWLLFR. It is worth noting that in all cases, RFWLLFR performs best when the depth of Fréchet trees is 3 (under the constraint $3 \sim \lceil \log_2 n \rceil$) and only one feature is randomly selected at each internal node. In fact, if the input space is not partitioned, *i.e.*, the Fréchet trees are of depth 1, RFWLLFR will do the same thing as GFR except RFWLLFR has a subsampling process from the training data. For setting I-2, as the Fréchet regression function is nonlinear, the performance of GFR is the worst. And we observe that the performance of GFR can not be improved significantly by simply increasing the number of training samples. For the low-dimensional case, the best performance is concentrated in RFWLLFR, indicating that it is easier to capture nonlinear signals. But as the dimension of X increases, the requirement of data size increases rapidly for local methods and the fitting becomes more challenging. Instead, a more straightforward method RFWLCFR begins to outperform RFWLLFR. LFR is a

Model	(p, n)	GFR	LFR	RFWLCFR/FRF	RFWLLFR
I-1	$(2, 100)$	**0.0014** (0.0012)	0.0688 (0.0707)	0.0269 (0.0071)	0.0031 (0.0015)
	$(2, 200)$	**0.0006** (0.0006)	0.0452 (0.0363)	0.0179 (0.0037)	0.0014 (0.0008)
	$(5, 200)$	**0.0011** (0.0006)	NA	0.0468 (0.0085)	0.0028 (0.0010)
	$(5, 500)$	**0.0005** (0.0003)	NA	0.0299 (0.0049)	0.0011 (0.0004)
	$(10, 500)$	**0.0009** (0.0004)	NA	0.0401 (0.0059)	0.0019 (0.0005)
	$(10, 1000)$	**0.0004** (0.0002)	NA	0.0284 (0.0039)	0.0009 (0.0002)
	$(20, 1000)$	**0.0009** (0.0003)	NA	0.0542 (0.0088)	0.0015 (0.0004)
	$(20, 2000)$	**0.0004** (0.0001)	NA	0.0444 (0.0059)	0.0007 (0.0002)
I-2	$(2, 100)$	0.3045 (0.0306)	0.0944 (0.0883)	0.0410 (0.0090)	**0.0317** (0.0141)
	$(2, 200)$	0.3023 (0.0278)	0.0745 (0.1550)	0.0254 (0.0050)	**0.0186** (0.0073)
	$(5, 200)$	0.2482 (0.0206)	NA	0.0772 (0.0136)	**0.0719** (0.0121)
	$(5, 500)$	0.2335 (0.0241)	NA	0.0557 (0.0087)	**0.0502** (0.0083)
	$(10, 500)$	0.2416 (0.0261)	NA	**0.1053** (0.0195)	0.1134 (0.0171)
	$(10, 1000)$	0.2438 (0.0297)	NA	**0.0901** (0.0174)	0.0927 (0.0148)
	$(20, 1000)$	0.2442 (0.0245)	NA	**0.1399** (0.0260)	0.1554 (0.0197)
	$(20, 2000)$	0.2456 (0.0286)	NA	**0.1257** (0.0238)	0.1337 (0.0194)

Table 1: Average MSE (standard deviation) of different methods for setting I-1,2 with $\sigma = 0.2$ for 100 simulation runs. Bold-faced numbers indicate the best performers.

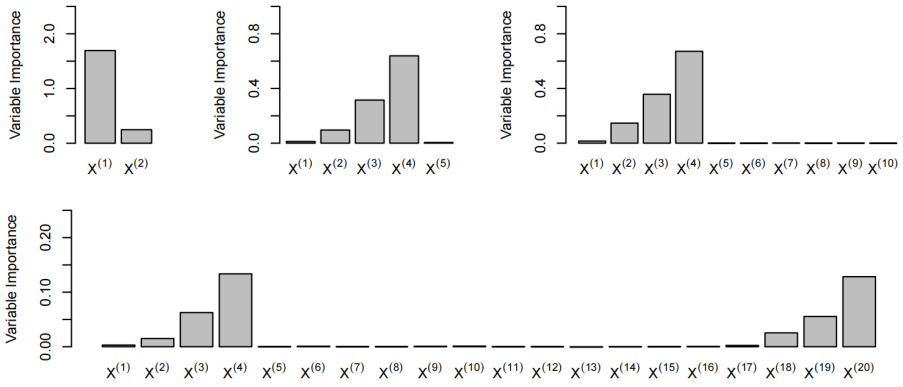


Figure 4: The variable importance for all variables of setting I-1.

local method relying heavily on kernel smoothing. The corresponding function in "frechet" package can only handle the case where the dimension of X is less than 3. When $p = 2$, the method LFR does not perform as well as RFWLCFR and RFWLLFR. For the effect of noise size σ and the case that the components of X are correlated, please refer to Appendix C.2.

As variable importance is readily apparent in setting I-1, we use this setting to examine the variable importance measure introduced in Section 5. Our analysis concentrates on four cases of (n, p): $(2, 100), (5, 200), (10, 500)$ and $(20, 1000)$. Following the Algorithm 1, we calculate the importance for all variables in one simulation run, as presented in Figure 4. It is evident that the importance rankings closely align with the truth. This indicates the ability of our method to accurately discern the order of importance among the feature variables for predicting the response values.

6.2 Fréchet Regression for Symmetric Positive-definite Matrices

Let (Ω, d) be the metric space $\mathcal{S}_m^+$ of $m \times m$ symmetric positive-definite (SPD) matrices endowed with metric d. There are many options for metrics, this section focuses on the Log-Cholesky metric and the affine-invariant metric. For a matrix Y, let $\lfloor Y \rfloor$ denote the strictly lower triangular matrix of Y, $\mathbb{D}(Y)$ denote the diagonal part of Y and $\|Y\|_F$ denote the Frobenius norm. It is well known that if Y is a symmetric positive-definite matrix, there is a lower triangular matrix P whose diagonal elements are all positive such that $PP^{\mathrm{T}} = Y$. This P is called the Cholesky factor of Y, devoted as $\mathscr{L}(Y)$.

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 metric (Lin, 2019) is defined by

$$d_L(Y_1, Y_2) = d_{\mathcal{L}}(\mathscr{L}(Y_1), \mathscr{L}(Y_2)),$$

where $d_{\mathcal{L}}(P_1, P_2) = \{\|\lfloor P_1 \rfloor - \lfloor P_2 \rfloor\|_F^2 + \|\log \mathbb{D}(P_1) - \log \mathbb{D}(P_2)\|_F^2\}^{1/2}$. And the affine-invariant metric (Moakher, 2005; Pennec et al., 2006) is defined by

$$d_A(Y_1, Y_2) = \left\| \log \left(Y_1^{-1/2} Y_2 Y_1^{-1/2} \right) \right\|_F.$$

Model	(p, n)	GFR	LFR	IFR	RFWLCFR/FRF	RFWLLFR
II-1	$(2, 100)$	1.264 (0.116)	**0.088** (0.039)	0.981 (0.280)	0.177 (0.061)	0.128 (0.072)
	$(2, 200)$	1.209 (0.089)	**0.038** (0.016)	0.863 (0.283)	0.082 (0.019)	0.054 (0.019)
	$(5, 200)$	1.281 (0.115)	NA	1.202 (0.180)	0.507 (0.082)	**0.397** (0.096)
	$(5, 500)$	1.267 (0.104)	NA	1.167 (0.216)	0.299 (0.046)	**0.190** (0.040)
	$(10, 500)$	1.279 (0.104)	NA	1.252 (0.115)	0.586 (0.101)	**0.575** (0.101)
	$(10, 1000)$	1.253 (0.096)	NA	1.242 (0.099)	0.420 (0.084)	**0.407** (0.082)
	$(20, 1000)$	0.973 (0.114)	NA	0.971(0.110)	0.623 (0.088)	**0.485** (0.075)
	$(20, 2000)$	0.956 (0.122)	NA	0.591(0.073)	0.522 (0.076)	**0.379** (0.058)
II-2	$(2, 100)$	1.932 (0.135)	**0.240** (0.098)	NA	0.367 (0.068)	0.283 (0.079)
	$(2, 200)$	1.898 (0.152)	**0.109** (0.020)	NA	0.188 (0.035)	0.143 (0.032)
	$(5, 200)$	1.980 (0.136)	NA	NA	0.855 (0.114)	**0.674** (0.116)
	$(5, 500)$	1.935 (0.143)	NA	NA	0.543 (0.065)	**0.369** (0.048)
	$(10, 500)$	1.971 (0.136)	NA	NA	1.057 (0.155)	**1.056** (0.134)
	$(10, 1000)$	1.949 (0.140)	NA	NA	0.845 (0.134)	**0.839** (0.119)
	$(20, 1000)$	1.970 (0.116)	NA	NA	**1.251** (0.200)	1.362 (0.171)
	$(20, 2000)$	1.962 (0.140)	NA	NA	**1.071** (0.178)	1.191 (0.158)

Table 2: Average MSE (standard deviation) of different methods for setting II-1,2 with $\sigma = 0.2$ and Log-Cholesky metric over 100 simulation runs. Bold-faced numbers indicate the best performers.

In practical applications, we need to choose the appropriate metric according to the need. The Log-Cholesky metric is faster in the calculation, while the affine-invariant metric has the congruence invariance property. More discussion refers to Lin (2019).

We generate $X_1, \ldots, X_n$ from the uniform distribution $\mathcal{U}([0,1]^p)$. And the response Y is generated via symmetric matrix variate normal distribution (Zhang et al., 2023). Consider the simplest case, we say an $m \times m$ symmetric matrix $A \sim \mathcal{N}_{mm}(M; \sigma^2)$ if $A = \sigma Z + M$ where M is an $m \times m$ symmetric matrix and Z is an $m \times m$ symmetric random matrix with independent $\mathcal{N}(0,1)$ diagonal elements and $\mathcal{N}(0, 1/2)$ off-diagonal elements. We consider the following settings with Y being SPD matrices.

Setting II-1:

$$\log(Y) \sim \mathcal{N}_{mm}\left(D(X), \sigma^2\right)$$

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

Setting II-2:

$$\log(Y) \sim \mathcal{N}_{mm}\left(D(X), \sigma^2\right)$$

with $D(X) = \begin{pmatrix} 1 & \rho_1(X) & \rho_2(X) \\ \rho_1(X) & 1 & \rho_1(X) \\ \rho_2(X) & \rho_1(X) & 1 \end{pmatrix}, \rho_1(X) = 0.8\cos\left(4\pi(\beta_1^T X)\right)$ and $\rho_2(X) = 0.4\cos\left(4\pi(\beta_2^T X)\right)$. The choice of (β_1, β_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. Additionally, since the setting II-1 is a single index model, we include IFR as another competitor. Since 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 regression for distributions, RFWL-

Model	(p, n)	FRF	RFWLCFR	RFWLLFR
II-1	$(2, 100)$	0.164133 (0.041988)	0.164132 (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)
	$(5, 500)$	0.243664 (0.035725)	0.243662 (0.035725)	**0.166254** (0.029714)
	$(10, 500)$	0.501018 (0.079189)	0.500904 (0.079182)	**0.464811** (0.079191)
	$(10, 1000)$	0.366349 (0.069003)	0.366257 (0.068997)	**0.331014** (0.067757)
	$(20, 1000)$	0.502299 (0.073056)	0.502238 (0.073048)	**0.390836** (0.057455)
	$(20, 2000)$	0.425098 (0.062546)	0.425046 (0.062547)	**0.304537** (0.044358)
II-2	$(2, 100)$	0.286285 (0.059663)	0.286285 (0.059655)	**0.238548** (0.062099)
	$(2, 200)$	0.154080 (0.025816)	0.154085 (0.025813)	**0.128763** (0.023595)
	$(5, 200)$	0.627837 (0.088528)	0.627769 (0.088545)	**0.499721** (0.087616)
	$(5, 500)$	0.396012 (0.046278)	0.395999 (0.046281)	**0.280171** (0.035996)
	$(10, 500)$	0.794744 (0.128331)	0.794220 (0.128334)	**0.762684** (0.100195)
	$(10, 1000)$	0.638420 (0.110770)	0.637765 (0.110686)	**0.605356** (0.089589)
	$(20, 1000)$	0.927755 (0.147346)	**0.927363** (0.147369)	0.991135 (0.111894)
	$(20, 2000)$	0.795528 (0.141159)	**0.795009** (0.141150)	0.878269 (0.118397)

Table 3: Average MSE (standard deviation) of different methods for setting II-1,2 with $\sigma = 0.2$ and affine-invariant metric over 100 simulation runs. Bold-faced numbers indicate the best performers.

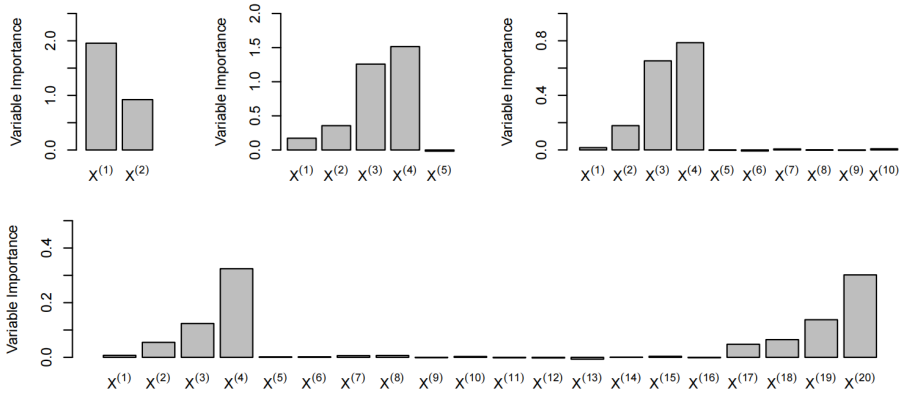


Figure 5: The variable importance for all variables of setting II-1.

CFR and FRF would have the same output. Results about setting II-1,2 with Log-Cholesky metric are shown in Table 2. In setting II-1, IFR performs better than GFR, but still falls short of our two methods. IFR learns coefficients of a single index model within 500 randomly generated direction vectors. This finite traversal optimization approach incurs a loss in precision. Taken together, LFR has the best performance when $p = 2$. But when $p > 2$, RFWLLFR performs the best in most cases. The results with the affine-invariant metric summarized in Table 3 also advocate RFWLLFR except for the high dimensional case of setting II-2. Moreover, for the affine-invariant metric, we observe slight differences between RFWLCFR and FRF.

Similar to Section 6.1, we further evaluate the validity of the proposed variable importance method in the scenario of matrix responses. We consider the setting II-1 with the Log-Cholesky metric and still pick (n, p) for $(2, 100)$, $(5, 200)$, $(10, 500)$ and $(20, 1000)$. The results are illustrated in Figure 5. Once again, it can be observed that our method produces nearly accurate importance rankings for the variables.

6.3 Fréchet Regression for Spherical Data

Let (Ω, d) be the metric space $\mathbb{S}^2$ of sphere data endowed with the geodesic distance d_g. For two points $Y_1, Y_2 \in \mathbb{S}^2$, the geodesic distance is defined by

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

We generate i.i.d $X_1, \ldots, X_n \sim \mathcal{U}([0,1]^p)$. And Y_i are generated by the following two settings.

Setting III-1: Let the Fréchet regression function be

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

We generate binary Normal noise ε_i on the tangent space $T_{m_{\oplus}(X_i)}\mathbb{S}^2$, then map ε_i back to $\mathbb{S}^2$ by Riemannian exponential map to get Y_i. Specifically, we first independently generate $\delta_{i1}, \delta_{i2} \overset{iid}{\sim} \mathcal{N}\left(0, 0.2^2\right)$, then let $\varepsilon_i = \delta_{i1}v_1 + \delta_{i2}v_2$, where $\{v_1, v_2\}$ forms an orthogonal basis of tangent space $T_{m_{\oplus}(X_i)}\mathbb{S}^2$. Then Y_i can be generated by

$$Y_i = \mathrm{Exp}_{m_{\oplus}(X_i)}(\varepsilon_i) = \cos(\|\varepsilon_i\|)\, m_{\oplus}(X_i) + \sin(\|\varepsilon_i\|)\frac{\varepsilon_i}{\|\varepsilon_i\|},$$

Model	(p, n)	FRF	RFWLCFR	RFWLLFR
III-1	$(2, 100)$	0.032329 (0.007413)	0.032327 (0.007409)	**0.019839** (0.005842)
	$(2, 200)$	0.023078 (0.003967)	0.023079 (0.003967)	**0.010781** (0.002704)
	$(5, 200)$	0.030237 (0.004693)	0.030237 (0.004693)	**0.017935** (0.003892)
	$(5, 500)$	0.021684 (0.002725)	0.021683 (0.002724)	**0.012164** (0.001998)
	$(10, 500)$	0.035012 (0.004252)	0.035018 (0.004253)	**0.013339** (0.001915)
	$(10, 1000)$	0.027184 (0.003063)	0.027188 (0.003065)	**0.010851** (0.001389)
	$(20, 1000)$	0.034698 (0.003879)	0.034701 (0.003879)	**0.033069** (0.004379)
	$(20, 2000)$	0.029362 (0.003503)	0.029362 (0.003502)	**0.028837** (0.003453)
III-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)
	$(20, 1000)$	0.009005 (0.001258)	0.009006 (0.001259)	**0.006311** (0.000975)
	$(20, 2000)$	0.007427 (0.000938)	0.007429 (0.000938)	**0.005117** (0.000692)

Table 4: Average MSE (standard deviation) of different methods for setting III-1,2 over 100 simulation runs. Bold-faced numbers indicate the best performers.

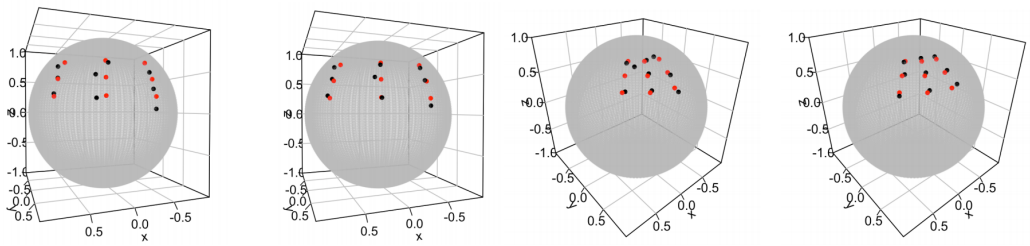


Figure 6: The plots of predictions of $m_{\oplus}(x)$ given by RFWLCFR (the 1st and 3rd panels) and RFWLLFR (the 2nd and 4th panels) in a simulation run of $p = 2, n = 200$. The left two panels show the results of setting III-1, while the right two show the results of setting III-2. The red points represent real points, and the black points represent predicted points.

where $\|\cdot\|$ is the Euclidean norm. Consider the following four kinds of dimensions
(i) $p = 2$: $\beta_1 = (1, 0), \beta_2 = (0, 1)$;
(ii) $p = 5, 10, 20$: $\beta_1 = (0.1, 0.2, 0.3, 0.4, 0, \ldots, 0)$, $\beta_2 = (0, \ldots, 0, 0.1, 0.2, 0.3, 0.4)$.

Setting III-2: Consider the following model

$$Y_i = \left(\sin(\beta_1^{\mathrm{T}} X_i + \varepsilon_{i1}) \sin(\beta_2^{\mathrm{T}} X_i + \varepsilon_{i2}), \sin(\beta_1^{\mathrm{T}} X_i + \varepsilon_{i1}) \cos(\beta_2^{\mathrm{T}} X_i + \varepsilon_{i2}), \cos(\beta_1^{\mathrm{T}} X_i + \varepsilon_{i1})\right)^{\mathrm{T}},$$

where the random noise $\varepsilon_{i1}, \varepsilon_{i2} \overset{iid}{\sim} \mathcal{N}\left(0, 0.2^2\right)$ are generated independently. The four situations corresponding to $p = 2, 5, 10, 20$ are the same as setting III-1.

Setting III-1 is similar to Petersen and Müller (2019) and Zhang et al. (2023), and setting III-2 is similar to Ying and Yu (2022). For Fréchet regression with sphere data, we focus on the comparison of FRF, RFWLCFR and RFWLLFR. RFWLCFR and FRF will have different outputs under the geodesic distance d_g. We summarize the results in Table 4. RFWLLFR performs best in all cases. To vividly describe the performance of RFWLCFR and RFWLLFR, Figure 6 exhibits the prediction of nine given testing points with $p = 2$ and $n = 200$ for setting III-1,2, which verifies the advantage of RFWLLFR.

7. Real Application

In this section, we use New York taxi data and mortality data to validate the advanced performance of our methods in practical applications.

7.1 New York Taxi Data

The New York City Taxi and Limousine Commission provides detailed records on yellow taxi rides, including pick-up and drop-off dates and times, pick-up and drop-off locations, trip distances, payment types, and other information. The data can be downloaded from `https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page`. In line with Dubey and Müller (2020), we transform the raw data into network data (adjacency matrices), where nodes represent zones and edges are weighted by the number of taxi rides that picked up in one zone and dropped off in another within a single hour. Specifically, we take the following steps to gather adjacency matrices:

1. Due to resource constraints, we only use data from January and February 2019 (59 days).

2. We further filter the observations to only include pick-ups and drop-offs that occurred within Manhattan (excluding islands).

3. We divide Manhattan into 10 zones and labeled them according to Dubey and Müller (2020). Details can be found in Appendix D. Then each network has 10 nodes, and the corresponding adjacency matrix has dimensions 10×10.

4. For each hour, we collect the number of pairwise connections between nodes based on pick-ups and drop-offs, which corresponds to the weights between nodes. We normalize the weights by the maximum edge weight in each hour, scaling them to the range $[0, 1]$.

We acquire a total of 1416 adjacency matrices of 10×10, which describe the taxi movements between zones in Manhattan. To facilitate the Fréchet regression analysis for network responses, we transform these matrices into SPD matrices by applying the matrix exponential mapping $\exp(\cdot)$ to them. Additionally, from the taxi data, we collect nine potential features with values averaged over each hour:

- *Ave. Distance*: mean distance traveled, standardized
- *Ave. Fare*: mean fare, standardized
- *Ave. Passengers*: mean number of passengers, standardized
- *Ave. Tip*: mean tip, standardized
- *Cash*: sum of cash indicators for type of payment, standardized
- *Credit*: sum of credit indicators for type of payment, standardized
- *Dispute*: sum of dispute indicators for type of payment, standardized
- *Free*: sum of free indicators for type of payment, standardized
- *Late Hour*: indicator for the hour being between 11pm and 5am, standardized

We also gather weather data for January and February 2019 from `https://www.wunderground.com/history/daily/us/ny/new-york-city/KLGA/date`, which yields 5 weather variables as potential features:

- *Day's Ave. Temp*: daily mean temperature, standardized
- *Day's Ave. Humid*: daily mean humidity, standardized
- *Day's Ave. Wind*: daily mean wind speed, standardized
- *Day's Ave. Press*: daily mean barometric pressure, standardized
- *Day's Total Precip*: daily total precipitation, standardized

In total, we have 14 potential features. The data set consisting of 1416 samples is partitioned randomly into three parts for Fréchet regression: a training set of size 850, a validation set of size 283, and a testing set of size 283, following a ratio of $6 : 2 : 2$. We train GFR, LFR, IFR, RFWLCFR,

and RFWLLFR on the training set using the Log-Cholesky metric and fine-tune their hyperparameters on the validation set. Subsequently, we retrain these methods on the combined training and validation sets of size 1133 using the selected hyperparameters to obtain the final models. Their performance is evaluated on the testing set by computing the mean squared errors based on the Log-Cholesky metric. However, the R-package for LFR is only applicable when the dimension of the feature space does not exceed 2. This limitation restricts the use of LFR in the current regression task. An effective solution is to consider the Fréchet sufficient dimension reduction method called the weighted inverse regression ensemble (WIRE), as proposed by Ying and Yu (2022). Unfortunately, the structural dimension of the central space is estimated to be 3. To ensure the availability of LFR, we are compelled to discard the third sufficient dimension reduction direction and project the original 14-dimensional feature vector onto a 2-dimensional subspace. This inevitably results in some loss of information. Specifically, the first two sufficient dimension reduction directions obtained through WIRE are as follows:

$$\begin{aligned}\hat{\beta}_1 = (&-0.911, 0.049, -0.152, -0.095, 0.303, 0.013, 0.035, -0.079,\\ &0.157, 0.076, 0.007, -0.021, 0.058, 0.033)^{\mathrm{T}};\\ \hat{\beta}_2 = (&0.125, -0.057, 0.178, 0.354, 0.740, 0.258, -0.091, -0.220,\\ &-0.390, -0.014, 0.020, -0.005, -0.027, 0.001)^{\mathrm{T}}.\end{aligned} \tag{14}$$

The two directions transform X into a 2-dimensional vector $(\hat{\beta}_1^{\mathrm{T}}X, \hat{\beta}_2^{\mathrm{T}}X)$ as the input of LRF. Additionally, the above result of the Fréchet sufficient dimension reduction shows that the underlying model can not be a single index model. However, despite this, we still intend to employ the IFR method for the sake of comparative analysis.

With the above preparations, we obtain the following testing errors: 1.377 for GFR, 0.944 for LFR, 2.620 for IFR, 0.568 for RFWLCFR, and 0.576 for RFWLLFR. In light of the results, RFWLCFR exhibits the highest prediction accuracy, followed closely by RFWLLFR. The superior performance of LFR compared to GFR strongly suggests the presence of a nonlinear regression relationship. The poor performance of IFR can be easily understood, as the central space discussed before has a structural dimension of 3, rendering the single index method ineffective. As expected, IFR performs worse than LFR, which benefits from two sufficient dimension reduction directions. We can further obtain the predicted taxi ride networks by applying the inverse mapping (matrix logarithmic map) $\log(\cdot)$ to the predicted matrices given by the methods described above. To visually illustrate the disparities in the outcomes of these methods, we randomly select three samples from the testing set and plot their corresponding true and predicted networks, as depicted in Figure 7. Please note that the plots for the first sample have been previously shown in Figure 1 and are therefore excluded here. Obviously, all the regression methods effectively capture the structural characteristics and weight information of true networks. But our methods stand out in their capacity to handle intricate details, yielding predicted outcomes that closely approximate the true networks. Corresponding heatmaps (Figure 8) have also been generated to complement the visualization. These results numerically and visually verify the superiority of our methods and demonstrate their potential in complex network learning.

We now measure the importance of each feature using Algorithm 1. The result is shown in Figure 9. This ranking of importance appears to be reasonable. Firstly, five weather features ($X^{(10)}$ to $X^{(14)}$) are deemed unimportant. One main reason is that daily averages of these weather features may not accurately capture their hourly impact on taxi traffic. And our data is sourced from the

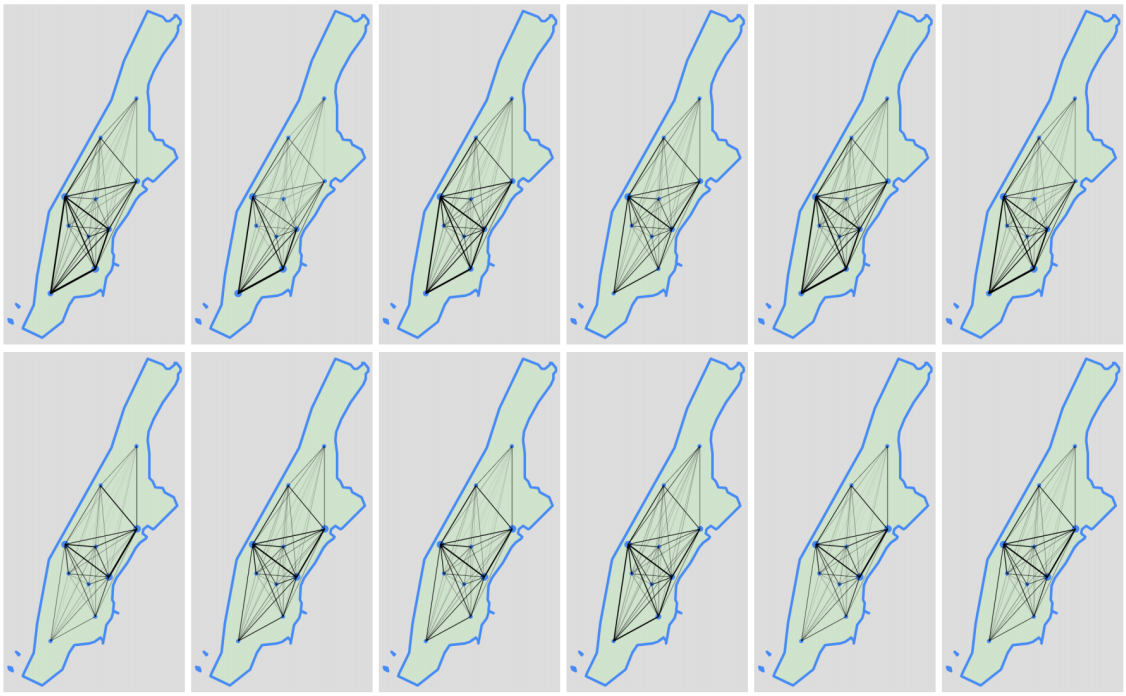


Figure 7: True (left) and fitted (right, in order of GFR, LFR after dimension reduction, IFR, RFWLCFR, RFWLLFR) networks with 10 zones for the remaining two test samples. The thickness of the edges connecting vertices corresponds to their weights, while the size of vertices represents the total traffic volume within each zone.

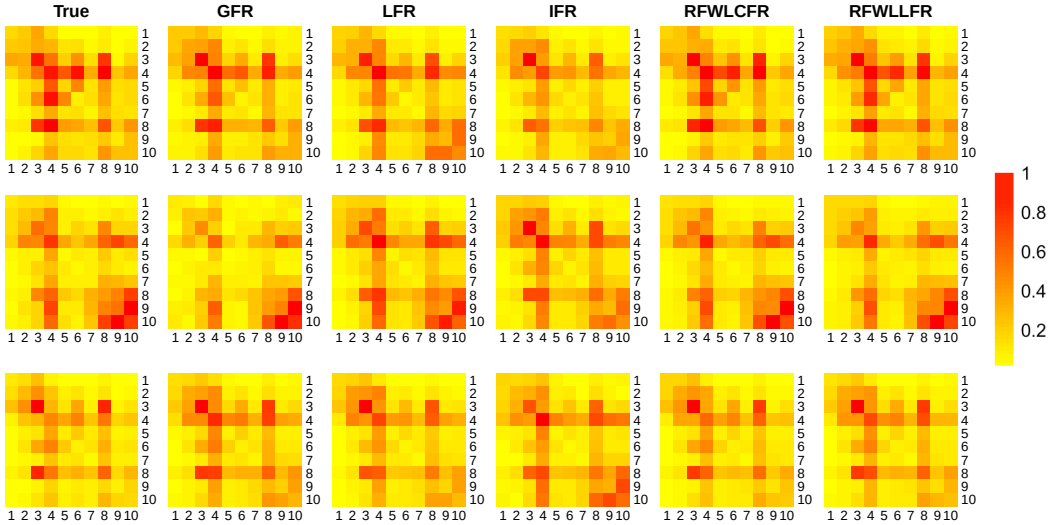


Figure 8: Networks with 10 zones represented as heatmaps for the three test samples.

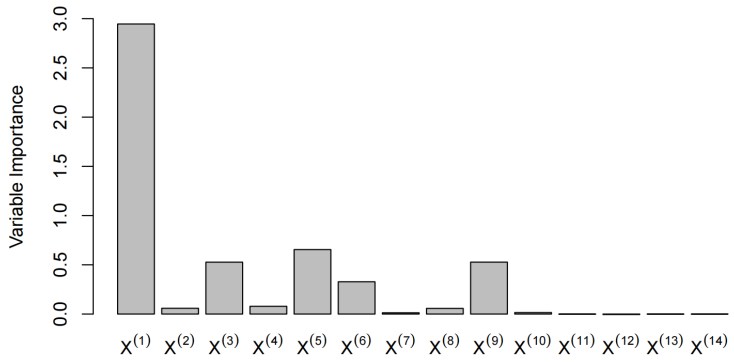


Figure 9: The variable importance for 14 features of New York taxi data.

consistent winter season in which daily weather conditions are relatively stable. The limited variability in weather conditions may account for the constrained explanatory power of these weather features in capturing fluctuations in taxi traffic. Another evidence support is the observation that the last five coefficients of the first two sufficient dimension reduction directions, β_1 and β_2 in (14), are all notably small. Among these five features, the two most significant ones are Day's Ave. Temp ($X^{(10)}$) and Day's Total Precip ($X^{(14)}$), as adverse weather conditions such as low temperatures or rain are more likely to hinder travel. Secondly, disputed and free rides are infrequent events in the raw data, and as such, their influence on the taxi ride network is consequently limited. This helps explain the relatively low importance assigned to Dispute ($X^{(7)}$) and Free ($X^{(8)}$) among the 14 features. Finally, it can be observed that the four most crucial features are Ave. Distance ($X^{(1)}$), Cash ($X^{(5)}$), Late Hour ($X^{(9)}$), and Ave. Passengers ($X^{(3)}$). This result can also be explained intuitively. Travel distance and passenger count are typically the two main considerations when people decide whether to choose a taxi for their trip. Cash, as a primary mode of payment, plays a pivotal role in people's travel decisions. Late Hour emerges as a critical factor shaping taxi ride networks, as late-night travel activity is significantly different from other times.

Based on the established variable importance ranking, we further illustrate how to perform variable selection. Here we judge which features can be omitted by comparing the performance of RFWLCFR with different feature subsets on the testing set. However, evaluating the performance of Fréchet regression separately for all possible subsets of the 14 features is a cumbersome task. If we leverage the feature importance ranking, this task will be greatly simplified. Specifically, we first remove the four least important features ($X^{(11)}$ to $X^{(14)}$). Then we initiate the selection process with the most important feature, and add other features, one by one, to the candidate feature subset in order of their importance among the remaining 10 variables. We conduct testing with each candidate subset. In this way, we only need 10 experiments to determine the most appropriate feature subset as the result of our selection. The entire data set consisting of 1416 samples is divided into training, validation, and testing sets in the same way as before. In the jth experiment, we intercept the most important j features to implement RFWLCFR, tune the hyperparameters using the validation set, and calculate the testing error using the testing set. The results of 10 experiments are summarized in Figure 10. Notably, the Fréchet regression with the first seven most important features achieves the lowest testing error. Adding additional features does not improve prediction accuracy but instead increases computational complexity and complicates interpretation. Conse-

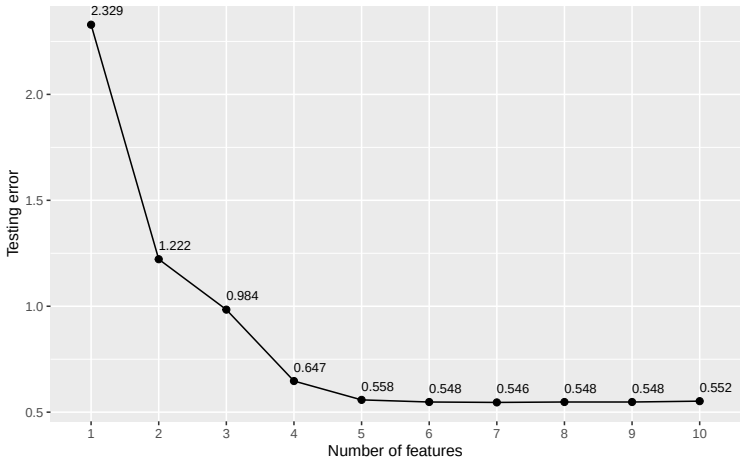


Figure 10: The testing errors with the different number of features.

quently, the final selected seven variables are Ave. Distance, Cash, Late Hour, Ave. Passengers, Credit, Ave. Tip, and Ave. Fare. In Figure 10, the testing error does not exhibit a clear upward trend as the number of features increases. This once again illustrates, to some extent, that our random forest based method can adaptively identify valuable features, and the presence of less relevant features will not interfere too much with its accuracy.

7.2 Mortality Data

Taking distribution as the outcome of interest allows us to obtain more information than summary statistics. In this part, we apply the two proposed methods to deal with the Fréchet regression problem for human mortality distribution. Just like Zhang et al. (2023), we also consider the following 9 predictor variables that have been standardized: (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 variables involve population, economy, health, and geography factors in 2015, which are closely related to human mortality. 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 life table considered here contains the number of deaths for each single age group from 162 countries in 2015. We treat the life table data as histograms of death versus age, with bin width equal to one year (the results of the subsequent analysis are similar when bin width is set to five years). Then the package "frechet" helps to transform the histograms into smoothed probability density functions. For comparison, we try to consider GFR, LFR, IFR. Similarly, before using LFR, WIRE (Ying and Yu, 2022) is adopted to achieve sufficient dimension reduction. The first four largest singular values of the WIRE matrix (Ying and Yu, 2022) are $4.486, 0.785, 0.101, 0.066$, and the structural dimension of central space is determined to be 2. The first two sufficient dimension reduction directions

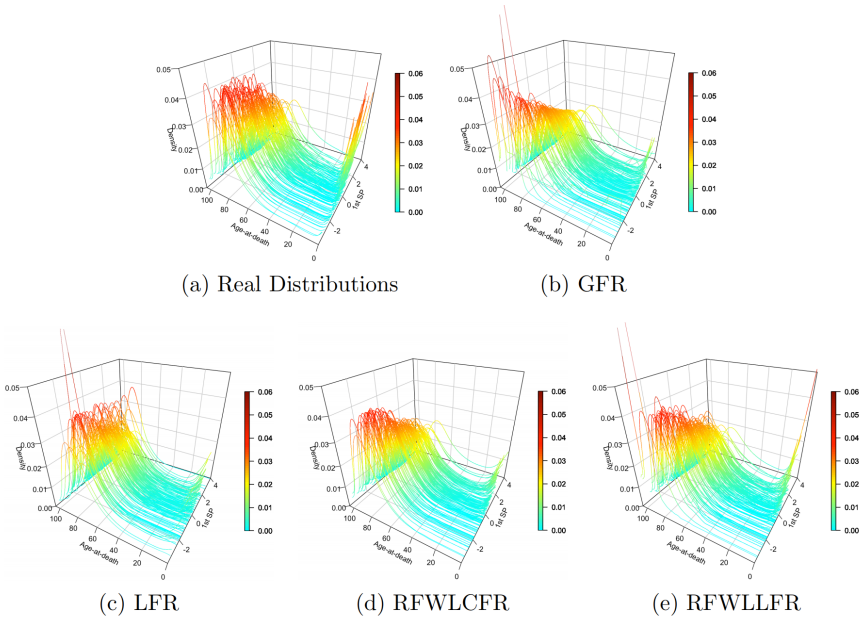
obtained by WIRE are

$$\hat{\beta}_1 = (-0.092, -0.084, 0.009, -0.429, 0.806, 0.048, 0.104, -0.364, -0.076)^{\mathrm{T}};$$
$$\hat{\beta}_2 = (0.103, 0.079, 0.641, 0.566, 0.415, -0.017, 0.130, 0.243, 0.066)^{\mathrm{T}}.$$

The two directions transform the 9-dimensional feature X into a 2-dimensional vector $(\hat{\beta}_1^{\mathrm{T}} X, \hat{\beta}_2^{\mathrm{T}} X)$ as the input of LRF. This also implies that the single index method is inappropriate. Here we drop the use of the IFR method.

We then perform 9-fold testing to evaluate the performance of all Fréchet regression methods. Specifically, we divide the 162 countries into 9 parts evenly and conduct 9 training runs. For each run, one of the 9 parts is chosen as the testing set and the rest as the training set. The test errors (mean squared errors based on the Wasserstein distance) obtained for nine runs are averaged for each method under the best choice of hyperparameters. The average test errors are recorded as 56.51 for GFR, 41.66 for LFR, 31.79 for RFWLCFR, and 36.20 for RFWLLFR. RFWLCFR has the best performance. These large errors reflect that inadequate sample size and large variation across countries increase the difficulty of the 9-dimensional Fréchet regression problem.

To show the performance of each method more vividly, we aid the analysis by plotting the mortality density predictions against the first sufficient predictor $\hat{\beta}_1^{\mathrm{T}} X$ (see Fig. 11). For reference, plot (a) of Fig. 11 is the smooth real density functions fitted according to the human mortality data. Observations reveal that the first sufficient predictor represents the development degree of a country. Countries with large $\hat{\beta}_1^{\mathrm{T}} X$ have backward medical levels and indigent living conditions, resulting in higher infant mortality and lower life expectancy. It is clear from plots that the distribution in the elderly age region ($80 \sim 100$ years old) and the infant age region (near 0 years old) is challenging to



(a) Real Distributions (b) GFR

(c) LFR (d) RFWLCFR (e) RFWLLFR

Figure 11: The plot (a) is the real mortality distributions against $\hat{\beta}_1^{\mathrm{T}} X$, and the remaining plots are the distributions predicted by each method against $\hat{\beta}_1^{\mathrm{T}} X$.

be well estimated. GFR performs poorly in both regions. It underestimates the infant mortality rate but successfully exhibits a tendency for the distribution to concentrate towards the elderly age region as the first sufficient predictor decreases. Compared with GFR, LFR based solely on two sufficient dimension reduction directions improves predictions of the elderly age region. However, the shape of the predicted mortality density function does not change significantly with the first sufficient predictor. In terms of the overall visual effect, the predictions of RFWLCFR are closest to the real distributions, but still suffer from large deviations for the infant age region in density functions. Among all local methods, RFWLLFR has the best performance in the infant age region. While RFWLLFR performs slightly inferior to RFWLCFR in the elderly age region, but much better than LFR. Overall, RFWLCFR has a remarkable advantage in this real data application. It again reflects the fact that RFWLLRF is not always the best choice, especially for complex regression problems with a small amount of data. RFWLCFR tends to be more robust and accurate in these cases.

8. Discussion

We propose two highly flexible and complementary locally weighted Fréchet regression methods for random object responses residing in a general metric space coupled with relatively high-dimensional Euclidean predictors. These methods employ adaptive random forest weights that effectively mitigate the curse of dimensionality, leading to significantly improved prediction accuracy compared to classical kernel weights. The two methods certainly extend random forests to the case with metric space valued responses. In addition, our theoretical findings include the most up-to-date result for random forests with Euclidean responses as a special case. Our proposals are supported by strong numerical performance, as demonstrated in both simulation studies and real data applications. In the present work, we focus on the theory of random forest weighted local constant Fréchet regression estimator. Theoretical investigation into the random forest weighted local linear Fréchet regression estimator including convergence rate and asymptotic normality is challenging for future research.

For Fréchet regression, our two methods only use the most basic information of a metric space. So our methods have 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 intrinsic local polynomial regression (Yuan et al., 2012) and the manifold additive model (Lin et al., 2022). For metric space being a specific Hilbert space, vector operations and an inner product structure are available, which inspires several promising nonparametric Hilbertian regressions such as Jeon and Park (2020) and Jeon et al. (2022). For the above two types of responses, we can consider a nonparametric regression framework based on the random forest kernel in the future. When more information of the output space is considered, models are expected to be more specific and targeted.

Acknowledgments

We sincerely thank the action editor and two reviewers for their valuable comments and constructive suggestions. The research of Rui Qiu and the corresponding author Zhou Yu is supported by the National Key R&D Program of China (Grant No. 2021YFA1000100 and 2021YFA1000101), the National Natural Science Foundation of China (Grant No. 12371289) and the Shanghai Pilot Program for Basic Research (Grant No. TQ20220105). The research of Ruoqing Zhu is supported by NSF grant 2210657.

Appendix A. More Explanation of Some Concepts

This section serves to provide a more detailed explanation of certain concepts mentioned in the main text.

A.1 Random Forest Kernel

Nadaraya-Watson Fréchet regression (Hein, 2009) and local Fréchet regression (Petersen and Müller, 2019) are based on the rationality that (X, Y) should be informative for $m_{\oplus}(x)$ if X is close to x (assume the function $m_{\oplus}$ has some degree of smoothness). The smoothing kernel $K_h(X_i - x)$ is exactly used to weight the contribution of each (X_i, Y_i) to the estimation of $m_{\oplus}(x)$ according to the proximity of X_i to x. But if the predictor contains some irrelevant variables, using the classical kernel smoothing functions often has unsatisfactory performance.

Different from the above kernel, the random forest kernel has a different mechanism for generating local weights. Fréchet trees produce local relationships among samples by recursively partitioning the input space. In addition to helping combat the curse of dimensionality, the random forest kernel can be adaptive if the partition process makes use of the information from responses Y, for example, the variance reduction splitting criterion introduced above. For the sample points divided into the same child node (or leaf), it is required that not only the distance of X is close to each other, but also the sample Fréchet variance of Y is small. It encourages Fréchet trees to select more relevant variables to divide the sample space. So the contribution value of (X_i, Y_i) given by the random forest kernel $\alpha_i(x)$ is jointly determined by both the information of X_i and Y_i. The random forest kernel prefers to assign a high weight to sample points that share a similar value of the response.

Assuming each tree is trained with at most k sample points per leaf node, and the full training data is used for each tree, Lin and Jeon (2006) introduced a paradigm to understand Euclidean random forests by considering any sample points falling into the leaf $L(x)$ be a k potential nearest neighbor (k-PNN) of x. When the splitting scheme of trees depends on the response, k-PNNs are chosen by an adaptive selection scheme. And a k-potential nearest neighbor can be made a k nearest neighbor by choosing a reasonable distance metric but not a simple Euclidean distance. The adaptive nature of the random forest kernel can be reflected in this way. A similar analysis can be extended to non-Euclidean cases.

A.2 Honest Tree

The honesty assumption is the largest divergence between the theory and applications of random forests. However, it is necessary for pointwise asymptotic theoretical analysis as it can help to eliminate bias. Similar assumptions have been used in many literatures (Friedberg et al., 2020; Bloniarz et al., 2016; Denil et al., 2013; Biau, 2012). The training examples whose Y_i's are used for

prediction are called prediction points, while the training examples whose Y_i's are 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 prediction process are independent. This brings great convenience to the analysis of the theoretical property of random forests. Wager and Athey (2018) achieved both consistency and the central limit theorem of honest random forests provided that the honesty assumption guarantees the following critical relationship for the bias analysis

$$E\big\{T_b(x;\mathcal{D}_n^b,\xi_b)\big\} = E\big[E\big\{Y \mid X \in L_b(x;\mathcal{D}_n^b,\xi_b)\big\}\big],$$

where $T_b(x;\mathcal{D}_n^b,\xi_b)$ is the prediction at x of the tree T_b constructed by a subsample $\mathcal{D}_n^b$ and a random draw $\xi_b \sim \Xi$, and $L_b(x;\mathcal{D}_n^b,\xi_b)$ is the corresponding leaf node containing x of T_b. Moreover, their theoretical results can apply to a wide range of random forest algorithms, including the classical variance reduction splitting criterion.

Any method of constructing an honest tree can be applied exactly to the Fréchet trees. The simplest way to achieve honesty is that the splitting rule of trees only depends on the predictor X like purely random forests (Arlot and Genuer, 2014; Genuer, 2012). When treating the random forest as a local weighting generator, all training examples in $L_b(x)$ can be used to calculate the random forest kernel. If the information of Y is also considered, then the double-sample tree (outlined in Procedure 1 of Wager and Athey (2018)) is a common approach to generate an honest tree. It divides the training sample into two non-overlapping parts: the structure set $\mathcal{J}_b$ and the prediction set $\mathcal{I}_b$ satisfying $|\mathcal{J}_b| = \lceil s_n/2 \rceil$ and $|\mathcal{I}_b| = \lfloor s_n/2 \rfloor$. During the tree growing process, the splits are chosen using any data from the sample $\mathcal{J}_b$ and X-observations from the sample $\mathcal{I}_b$, but without using Y-observations from the sample $\mathcal{I}_b$. And the estimation of leaf-wise responses only adopts Y-observations from the sample $\mathcal{I}_b$. When treating the random forest as a local weighting generator, we only consider the subset $\{(X_i, Y_i) : (X_i, Y_i) \in L_b(x)\}$ of $\mathcal{I}_b$ to calculate the random forest kernel; For more discussion, please refer to section 2.4 and Appendix B of Wager and Athey (2018) or section 2.3 of Friedberg et al. (2020). It is worth noting that if honesty is achieved by double-sample trees, another condition α-regular mentioned in the paper should be satisfied for the sample $\mathcal{I}_b$. This is why X-observations from the sample $\mathcal{I}_b$ may be used during the construction of trees.

A.3 A Little Remark about RFWLLFR

Consider a special case with $p = 1$, then RFWLLFR estimator $\hat{l}_{\oplus}(x)$ has the following equivalent expression.

$$\begin{aligned}
\hat{l}_{\oplus}(x) &= \underset{y\in\Omega}{\operatorname{argmin}} \sum_{i=1}^{n} e_1^{\mathrm{T}} (\tilde{X}^{\mathrm{T}} A \tilde{X})^{-1} \begin{pmatrix} 1 \\ X_i - x \end{pmatrix} \alpha_i(x) d^2(Y_i, y) \\
&= \underset{y\in\Omega}{\operatorname{argmin}} \sum_{i=1}^{n} e_1^{\mathrm{T}} \begin{pmatrix} \sum_{i=1}^n \alpha_i(x) & \sum_{i=1}^n \alpha_i(x)(X_i - x) \\ \sum_{i=1}^n \alpha_i(x)(X_i - x) & \sum_{i=1}^n \alpha_i(x)(X_i - x)^2 \end{pmatrix}^{-1} \\
&\qquad\qquad \begin{pmatrix} 1 \\ X_i - x \end{pmatrix} \alpha_i(x) d^2(Y_i, y) \\
&= \underset{y\in\Omega}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} e_1^{\mathrm{T}} \begin{pmatrix} \hat{\mu}_0 & \hat{\mu}_1 \\ \hat{\mu}_1 & \hat{\mu}_2 \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ X_i - x \end{pmatrix} \alpha_i(x) d^2(Y_i, y)
\end{aligned}$$

$$
\begin{aligned}
&= \underset{y\in\Omega}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} e_1^{\mathrm{T}} \frac{1}{\hat{\mu}_0\hat{\mu}_2 - \hat{\mu}_1^2} \begin{pmatrix} \hat{\mu}_2 & -\hat{\mu}_1 \\ -\hat{\mu}_1 & \hat{\mu}_0 \end{pmatrix} \begin{pmatrix} 1 \\ X_i - x \end{pmatrix} \alpha_i(x) d^2(Y_i, y) \\
&= \underset{y\in\Omega}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\hat{\mu}_0\hat{\mu}_2 - \hat{\mu}_1^2} \left\{\hat{\mu}_2 - \hat{\mu}_1 (X_i - x)\right\} \alpha_i(x) d^2(Y_i, y) \\
&= \underset{y\in\Omega}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} t_{in}(x) d^2(Y_i, y),
\end{aligned}
$$

where $t_{in}(x) = \frac{1}{\hat{\mu}_0\hat{\mu}_2 - \hat{\mu}_1^2}\alpha_i(x)\left\{\hat{\mu}_2 - \hat{\mu}_1 (X_i - x)\right\}$, $\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n}\alpha_i(x)(X_i - x)^j$. Apart from the weight generating function, this form coincides with the local Fréchet regression proposed by Petersen and Müller (2019). Acquiring local Fréchet regression estimators directly from corresponding explicit Euclidean forms as we do may be more straightforward.

Appendix B. Asymptotic Normality

This section is dedicated to deriving the asymptotic normality of RFWLCFR. We first generalize the theory of M_m-estimator in Bose and Chatterjee (2018) to the case that m tends to infinity along with n.

Definition 11 *Let $Z_1, Z_2, \ldots, Z_{m_n}$ be i.i.d. $\mathcal{Z}$-valued random variables and $\theta \in \mathcal{R}^q$. A real-valued measurable function $f_n(z_1, z_2, \ldots, z_{m_n}, \theta)$ is symmetric in the arguments $z_1, z_2, \ldots, z_{m_n}$ for each n. Define*

$$Q_n(\theta) = E f_n\left(Z_1, Z_2, \ldots, Z_{m_n}, \theta\right)$$

and

$$\theta_n = \underset{\theta\in\mathcal{R}^q}{\operatorname{argmin}} Q_n(\theta).$$

θ_n is called the M_{m_n}-parameter.

Definition 12 *Let $Z_1, Z_2, \ldots, Z_n$ be a sequence of i.i.d. observations. Define*

$$\hat{Q}_n(\theta) = \binom{n}{m_n}^{-1} \sum_{1\le i_1<i_2<\ldots<i_{m_n}\le n} f_n\left(Z_{i_1}, Z_{i_2}, \ldots, Z_{i_{m_n}}, \theta\right)$$

and

$$\hat{\theta}_n = \underset{\theta\in\mathcal{R}^q}{\operatorname{argmin}} \hat{Q}_n(\theta).$$

$\hat{\theta}_n$ is called the M_{m_n}-estimator of θ_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 θ. Since $\hat{Q}_n(\theta)$ is the sample analogue of $Q_n(\theta)$, $\hat{\theta}_n$ is a reasonable estimator of θ_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 Bose and Chatterjee (2018). 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 Bose and Chatterjee (2018) for more details.

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

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

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

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

(i) $f_n\left(z_1, z_2, \ldots, z_{m_n}, \theta\right)$ is measurable in $\left(z_1, z_2, \ldots, z_{m_n}\right)$ and convex in θ.

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

(iii) θ_n exists and is unique, and $f_n\left(z_1, z_2, \ldots, z_{m_n}, \theta\right)$ is twice differentiable on an appropriate neighborhood of θ_n.

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

(v) $H_n = \nabla^2 Q\left(\theta_n\right)$ exists and is positive definite and $\lambda_{\min}(H_n) \nrightarrow 0$.

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 $\hat{\theta}_n$. Mentch and Hooker (2016) gave sufficient conditions for the asymptotic normality of such U-statistics. However, these conditions can not hold simultaneously. DiCiccio and Romano (2022) then developed conditions that can be verified on the basis of Mentch and Hooker (2016). But both of their results require that the order of the infinite order U-statistics is $o(\sqrt{n})$. Peng et al. (2022) further improved this result when the order of the infinite order U-statistics is $o(n)$. Wager and Athey (2018) 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 Peng et al. (2022). And assumptions (i), (ii), (iii) and (v) are adaptions of that for studying M_m estimator in Bose and Chatterjee (2018).

Theorem 13 *Suppose that assumptions (i)-(v) hold, then for any sequence of measurable minimizers* $\{\hat{\theta}_n, n \ge 1\}$,

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

(b) $\sqrt{n}\Lambda_n^{-1/2}\left(\hat{\theta}_n - \theta_n\right) \xrightarrow{d} \mathcal{N}(0, I)$, *where*

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

Theorem 13 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 Bose and Chatterjee (2018). Next, we derive the asymptotic normality of RFWLCFR by applying this result. Here we continue to adopt the expressions (8) and (9). Let

$$h_n\left(Z_{i_{k,1}}, Z_{i_{k,2}}, \ldots, Z_{i_{k,sn}}, y\right) = E_{\xi \sim \Xi}\left\{\frac{1}{N(L(x; \mathcal{D}_n^k, \xi))} \sum_{i: X_i \in L(x; \mathcal{D}_n^k, \xi)} d^2\left(Y_i, y\right)\right\}$$

where $Z_i = (X_i, Y_i)$ and $\mathcal{D}_n^k = \left(Z_{i_{k,1}}, Z_{i_{k,2}}, \ldots, Z_{i_{k,s_n}}\right)$. Then

$$\hat{r}_\oplus(x) = \underset{y\in\Omega}{\operatorname{argmin}} \hat{R}_n(x,y) = \underset{y\in\Omega}{\operatorname{argmin}} \binom{n}{s_n}^{-1} \sum_{1\le i_1<i_2<\ldots<i_{s_n}\le n} h_n\left(Z_{i_1}, Z_{i_2}, \ldots, Z_{i_{s_n}}, y\right),$$

$$\tilde{r}_\oplus(x) = \underset{y\in\Omega}{\operatorname{argmin}} \tilde{R}_n(x,y) = \underset{y\in\Omega}{\operatorname{argmin}} E h_n\left(Z_1, Z_2, \ldots, Z_{s_n}, y\right).$$

Unfortunately, since y is not in the Euclidean space, the derivative about y can't be computed and $\hat{r}_\oplus(x) - \tilde{r}_\oplus(x)$ has no sense. In order to apply the Theorem 13, 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 Bhattacharya and Lin (2017); Bhattacharya and Patrangenaru (2003, 2005). We assume the following conditions to establish the central limit theorem of the proposed RFWLCFR.

(A13) $h_n\left(z_1, z_2, \ldots, z_{s_n}, y\right)$ is measurable in $(z_1, z_2, \ldots, z_{s_n})$ and $\tilde{R}_n(x,y) < \infty$ for each y.

(A14) $\tilde{r}_\oplus(x)$ exists and is unique.

(A15) $\tilde{r}_\oplus(x) \in G$ for large n, where G is a measurable subset of Ω. And there is a homeomorphism $\phi : G \to U$, where U is an open subset of $\mathcal{R}^q$ for some $q \ge 1$, and G is given its relative topology on Ω. Also

$$u \mapsto f_n(z_1, z_2, \ldots, z_{s_n}, u) = h_n\left(z_1, z_2, \ldots, z_{s_n}, \phi^{-1}(u)\right)$$

is twice differentiable on an appropriate neighborhood of $\phi(\tilde{r}_\oplus(x))$.

(A16) $f_n(z_1, z_2, \ldots, z_{s_n}, u)$ is convex in u.

(A17) Let g_n be a measurable sub-gradient of f_n about u. Define

$$K_n = \operatorname{Var}\left\{E\left[g_n\left(Z_1, Z_2, \ldots Z_{s_n}, \phi(\tilde{r}_\oplus(x))\right) \mid Z_1\right]\right\},$$

then $E\left|g_n\left(Z_1, Z_2, \ldots, Z_{s_n}, \phi(\tilde{r}_\oplus(x))\right)\right|^2 < C$ for some constant C, and $s_n \lambda_{\min}(K_n) \nrightarrow 0$.

(A18) $H_n = \nabla^2 E f_n(Z_1, Z_2, \ldots, Z_{s_n}, \phi(\tilde{r}_\oplus(x)))$ exits and is positive definite, and $\lambda_{\min}(H_n) \nrightarrow 0$.

The assumption (A15) is crucial just like the assumption for establishing the asymptotic normality of sample Fréchet mean as suggested in Bhattacharya and Lin (2017). For example, when (Ω, d_g) is a q-dimensional complete Riemannian manifold with metric tensor g and geodesic distance d_g, we can choose the Riemannian logarithmic map at $\tilde{r}_\oplus(x)$ as the homeomorphism ϕ, which is defined on a neighborhood of $\tilde{r}_\oplus(x)$ onto its image U in the tangent space at $\tilde{r}_\oplus(x)$. Other assumptions are adaptions of that of Theorem 13.

Theorem 14 *Suppose that for a fixed $x \in [0,1]^p$, (A1), (A4), (A13)–(A18) hold, and the Fréchet trees are symmetric. Then $\phi(\hat{r}_\oplus(x))$ is asymptotically normal, i.e.,*

$$\sqrt{n}\Lambda_n^{-1/2}\left\{\phi(\hat{r}_\oplus(x)) - \phi(\tilde{r}_\oplus(x))\right\} \to \mathcal{N}(0, I),$$

where $\Lambda_n = s_n^2 H_n^{-1} K_n H_n^{-1}$ with K_n, H_n defined in assumption (A17) and (A18).

If the homeomorphism ϕ in Theorem 14 is further Lipschitz-continuous, by Lemma 6,

$$\|\phi(\tilde{r}_\oplus(x)) - \phi(m_\oplus(x))\| \le L d(\tilde{r}_\oplus(x), m_\oplus(x)) = O\left(s_n^{-\frac{1}{2}\frac{\log(1-\alpha)}{\log(\alpha)}\frac{\pi}{p}\frac{1}{\beta_1-1}}\right),$$

where L is the Lipschitz constant. Under a suitable range of orders for s_n with respect to n, $\|\phi(\tilde{r}_{\oplus}(x)) - \phi(m_{\oplus}(x))\|/\|\Lambda_n^{1/2}/\sqrt{n}\|$ can converge to zero. Then we can get the asymptotic normality of $\phi(\hat{r}_{\oplus}(x))$ about $\phi(m_{\oplus}(x))$ by Theorem 14 and Slutsky's theorem. The following takes the Euclidean case as a simple example to illustrate it.

Remark 15 *Consider the special case when $\Omega \subseteq \mathcal{R}$. For responses that are Euclidean, we naturally choose ϕ as the identity mapping. Then we have*

$$\begin{aligned} f_n(Z_{i_{k,1}}, \ldots, Z_{i_{k,s_n}}, u) &= h_n\left(Z_{i_{k,1}}, \ldots, Z_{i_{k,s_n}}, u\right) \\ &= E_{\xi\sim\Xi}\left\{ \frac{1}{N(L(x;\mathcal{D}_n^k,\xi))} \sum_{i:X_i\in L(x;\mathcal{D}_n^k,\xi)} (Y_i - u)^2 \right\}. \end{aligned}$$

And we further get

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

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

In addition,

$$u_n = \underset{u\in U}{\operatorname{argmin}} Q_n(u) = E\left\{ \frac{1}{N(L(x;\mathcal{D}_n^k,\xi))} \sum_{i:X_i\in L(x;\mathcal{D}_n^k,\xi)} Y_i \right\},$$

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

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

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

Let $\zeta_{1,n} = \operatorname{Var}\left[E\left\{ E_{\xi\sim\Xi}\left(\frac{1}{N(L(x;\mathcal{D}_n^*,\xi))} \sum_{i:X_i\in L(x;\mathcal{D}_n^*,\xi)} Y_i \right) \mid Z_1 \right\} \right]$ with $\mathcal{D}_n^* = (Z_1, \ldots, Z_{s_n})$, 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 $s_n\zeta_{1,n} \nrightarrow 0$ and the assumption (A17) holds, by Theorem 14 we have

$$\frac{\sqrt{n}\left(\hat{u}_n - u_n\right)}{\sqrt{s_n^2\zeta_{1,n}}} \xrightarrow{d} \mathcal{N}(0,1),$$

i.e.,

$$\frac{\sqrt{n}\left\{\hat{r}_{\oplus}(x) - \tilde{r}_{\oplus}(x)\right\}}{\sqrt{s_n^2\zeta_{1,n}}} \xrightarrow{d} \mathcal{N}(0,1). \tag{15}$$

Here $\hat{r}_{\oplus}(x)$ is exactly the prediction of Euclidean random forest at x, and (15) *is the standard results of Euclidean random forests (Mentch and Hooker, 2016; Peng et al., 2022; DiCiccio and Romano, 2022). Since $s_n\zeta_{1,n}$ is bounded and $s_n\zeta_{1,n} \nrightarrow 0$, it holds that $\hat{r}_{\oplus}(x) - \tilde{r}_{\oplus}(x) = O_p\big((s_n/n)^{1/2}\big)$.*

For the Euclidean case, $\beta_1 = 2$ and Lemma 4 gives

$$|\tilde{r}_{\oplus}(x) - m_{\oplus}(x)| = O\left(s_n^{-\frac{1}{2}\frac{\log(1-\alpha)}{\log(\alpha)}\frac{\pi}{p}}\right).$$

Let $s_n = n^{\beta}$, then

$$\frac{|\tilde{r}_{\oplus}(x) - m_{\oplus}(x)|}{\sqrt{s_n^2\zeta_{1,n}/n}} = O\left(n^{\frac{1}{2}\left[1-\beta\left\{1+\frac{\log(1-\alpha)}{\pi^{-1}p\log(\alpha)}\right\}\right]}\right).$$

The right-hand-side converges to zero provided that

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

Then by (15) *and Slutsky's theorem,*

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

when

$$s_n \asymp n^{\beta} \quad \textit{for some} \quad \beta_{\min} := 1 - \left\{1 + \frac{p}{\pi}\frac{\log(\alpha)}{\log(1-\alpha)}\right\}^{-1} < \beta < 1.$$

This result coincides with Theorem 1 of Wager and Athey (2018). Therefore, our asymptotic normality established for Fréchet regression with metric space valued responses includes their result for Euclidean random forests as a special case.

Appendix C. Additional Simulations

Here we specifically introduce the simplified splitting criterion used by the Fréchet tree construction in our simulations, and add more simulations for responses being distributions or symmetric positive-definite matrices.

C.1 Simplified Adaptive Splitting Criterion for Fréchet Trees

The process introduced in Section 2.1.2 to find the optimal split is accurate but computationally intensive. In all simulation experiments of this paper, we adopt another efficient way introduced by Capitaine et al. (2019). A split on an internal node A along the direction of feature j is any couple of distinct elements $(c_{j,l}, c_{j,r})$. The partition associated with elements $(c_{j,l}, c_{j,r})$ is defined by

$$A_{j,l} = \left\{x \in A : |x^{(j)} - c_{j,l}| \leq |x^{(j)} - c_{j,r}|\right\}$$

and

$$A_{j,r} = \left\{x \in A : |x^{(j)} - c_{j,r}| < |x^{(j)} - c_{j,l}|\right\},$$

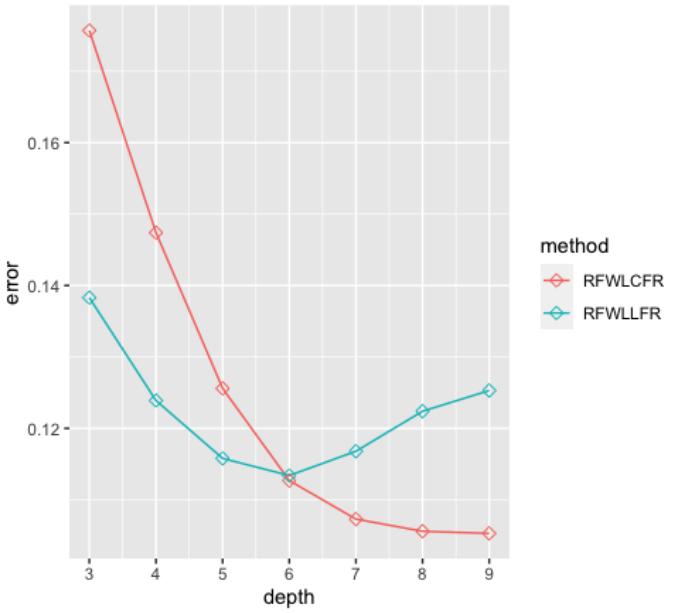


Figure 12: The influence of depth of Fréchet trees on average MSE for (10,500) from setting I-2.

which generate the left and right child nodes of the node A. Again, let

$$\mathcal{H}_n(j) = \frac{1}{N_n(A)} \left\{ \sum_{i:X_i \in A} d^2\left(Y_i, \bar{Y}_A\right) - \sum_{i:X_i \in A_{j,l}} d^2\left(Y_i, \bar{Y}_{A_{j,l}}\right) - \sum_{i:X_i \in A_{j,r}} d^2\left(Y_i, \bar{Y}_{A_{j,r}}\right) \right\}.$$

Then the optimal split $(c_{j_n^*,l}, c_{j_n^*,r})$ is decided by

$$j_n^* = \underset{j}{\operatorname{argmax}}\, \mathcal{H}_n(j).$$

To determinate the representatives $(c_{j,l}, c_{j,r})$, the 2-means algorithm (k-means with k = 2) can be implemented on the jth component of the sample points falling into the node A.

C.2 Fréchet Regression for Distributions

RFWLCFR is similar in nature to random forests, so it prefers using deeper Fréchet trees. As for RFWLLFR, knowing that a more powerful local linear regression will be used for the final model fitting, it is not reasonable to capture too much signal from the data during the construction of the Fréchet trees. So shallower trees are often used to avoid overfitting for RFWLLFR. Figure 12 shows the effect of the depth of Fréchet trees on the performance of our two methods based on setting I-2 with $p = 10$ and $n = 500$. RFWLLFR achieves the optimal performance when the depth is six, while RFWLCFR prefers deeper Fréchet trees.

We also select several combinations of (n, p) from setting I-2 to study the effect of noise size σ on the performance of each method. The results are summarized in Table 5. It can be seen that GFR is almost unaffected by noise. LFR performs poorly when the noise level is high. And our proposed RFWLCFR and RFWLLFR are still better than GFR and LFR in general.

(p, n)	σ	GFR	LFR	RFWLCFR/FRF	RFWLLFR
(2,200)	$\sigma = 0.1$	0.3026 (0.0283)	0.0261 (0.0198)	0.0158 (0.0036)	**0.0084** (0.0041)
	$\sigma = 0.2$	0.3023 (0.0278)	0.0745 (0.1550)	0.0254 (0.0050)	**0.0186** (0.0073)
	$\sigma = 0.5$	0.3032 (0.0274)	0.3341 (0.3412)	0.0895 (0.0181)	**0.0786** (0.0209)
(5,500)	$\sigma = 0.1$	0.2331 (0.0240)	NA	0.0515 (0.0080)	**0.0437** (0.0074)
	$\sigma = 0.2$	0.2335 (0.0241)	NA	0.0557 (0.0087)	**0.0502** (0.0083)
	$\sigma = 0.5$	0.2363 (0.0245)	NA	**0.0850** (0.0128)	0.0946 (0.0144)
(10,1000)	$\sigma = 0.1$	0.2434 (0.0295)	NA	**0.0870** (0.0175)	0.0879 (0.0145)
	$\sigma = 0.2$	0.2438 (0.0297)	NA	**0.0901** (0.0174)	0.0927 (0.0148)
	$\sigma = 0.5$	0.2462 (0.0302)	NA	**0.1103** (0.0195)	0.1251 (0.0171)
(20,2000)	$\sigma = 0.1$	0.2452 (0.0285)	NA	**0.1227** (0.0225)	0.1300 (0.0191)
	$\sigma = 0.2$	0.2456 (0.0286)	NA	**0.1257** (0.0238)	0.1337 (0.0194)
	$\sigma = 0.5$	0.2479 (0.0288)	NA	**0.1401** (0.0260)	0.1654 (0.0235)

Table 5: Average MSE (standard deviation) of different methods for (2,200), (5,500), (10,1000), (20,2000) from setting I-2 with different σ over 100 simulation runs. Bold-faced numbers indicate the best performers.

Model	(p, n)	GFR	LFR	RFWLCFR/FRF	RFWLLFR
I-3	$(2, 100)$	0.2495 (0.2002)	0.0869 (0.3530)	0.1179 (0.1745)	**0.0423** (0.0551)
	$(2, 200)$	0.2302 (0.3131)	0.0390 (0.1558)	0.0944 (0.2843)	**0.0347** (0.1355)
	$(5, 200)$	0.0893 (0.0641)	NA	0.0429 (0.0432)	**0.0342** (0.0261)
	$(5, 500)$	0.0827 (0.0551)	NA	0.0248 (0.0272)	**0.0164** (0.0107)
	$(10, 500)$	0.0896 (0.1006)	NA	0.0368 (0.0660)	**0.0307** (0.0293)
	$(10, 1000)$	0.0810 (0.0888)	NA	0.0226 (0.0623)	**0.0187** (0.0201)
	$(20, 1000)$	0.0445 (0.0197)	NA	**0.0181** (0.0119)	0.0210 (0.0085)
	$(20, 2000)$	0.0502 (0.0275)	NA	**0.0155** (0.0110)	0.0172 (0.0081)

Table 6: Average MSE (standard deviation) of different methods for setting I-3 over 100 simulation runs. Bold-faced numbers indicate the best performers.

Since the components of the predictor X are independent in all previous simulation settings, we here add another setting to cover the case that components are correlated.

Setting I-3: We generate X by a multivariate normal distribution

$$X \sim \mathcal{N}(0, \Sigma),$$

where the ij-th element of Σ is $0.5^{|i-j|}$. Then Y is generated by

$$Y = \mathcal{N}\left(\mu_Y, \sigma_Y^2\right),$$

where

$$\mu_Y \sim \mathcal{N}\left(0.1(e_1^{\mathrm{T}} X)^2 \left(2\beta^{\mathrm{T}} X - 1\right), 0.2^2\right) \quad \text{and} \quad \sigma_Y = 1.$$

The above e_i is a vector of zeros with 1 in the ith element. Consider the following four kinds of dimensions

(i) For $p = 2$: $\beta = (0.75, 0.25)$.
(ii) For $p = 5, 10$: $\beta = (0.1, 0.2, 0.3, 0.4, 0, \ldots, 0)$.
(iii) For $p = 20$: $\beta = (0.1, 0.2, 0.3, 0.4, 0, \ldots, 0, 0.1, 0.2, 0.3, 0.4)/2$.

From the results in Table 6, we can find 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. RFWLLFR is the most stable among all methods. As the dimension of X becomes larger, the performance of RFWLCFR gets closer to that of RFWLLFR. Especially in the high-dimensional case, RFWLCFR begins to outperform RFWLLFR. Overall, all methods under the current setting behave similarly to the cases when the components of X are independent.

C.3 Fréchet Regression for Symmetric Positive-definite Matrices

For responses being symmetric positive definite matrices, the intrinsic local polynomial regression (ILPR) (Yuan et al., 2012) and the manifold additive model (MAM) (Lin et al., 2022) are two promising tools that take advantage of the geometric structure of the Riemannian manifold. We plan to include the two methods for comparisons for Fréchet regression with symmetric positive-definite matrices.

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. Lin et al. (2022) further proposed an additive model for the regression of symmetric positive-definite matrix valued responses 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, Yuan et al. (2012)), especially when the underlying model is fully additive. However, Lin et al. (2022) only considered $p = 3, 4$ in their simulation studies. In the next, we adopt the settings in Lin et al. (2022) to make a comprehensive comparison among MAM, ILPR, GFR, FRF, RFWLCFR, and RFWLLFR. MAM can be implemented with the R-package "matrix-manifold" (Lin, 2020).

Setting II: Let $X \sim \mathcal{U}([0,1]^p)$. The response Y is generated via

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

where μ is the 3×3 identity matrix, $w(X) = \mathfrak{exp}\tau_{\mu,e} f(X)$, e is the identity element of the group, $\tau_{\mu,e}$ denotes the parallel transport from μ to e, $\mathfrak{exp}(\cdot)$ denotes the Lie exponential map, $\oplus$ denotes the group operation, and ζ is the random noise. The noise ζ is generated according to $\mathfrak{log}\zeta = \sum_{j=1}^{6} Z_j v_j$, where $\mathfrak{log}(\cdot)$ denotes the Lie log map, $Z_1, \ldots, Z_6$ are independently sampled from $\mathcal{N}\left(0, \sigma^2\right)$, and $v_1, \ldots, v_6$ are an orthonormal basis of the tangent space $T_e\mathcal{S}_3^+$. The signal-to-ratio (SNR) is measured by SNR $= E\left\|\mathfrak{log}w(X)\right\|_e^2 / E\|\mathfrak{log}\zeta\|_e^2$. Take the value of the parameter σ^2 to cover two choices for the SNR, namely, SNR $= 2$ and SNR $= 4$. Refer to Lin et al. (2022) for the details of the notations and concepts here. We consider the following setting about $f(X)$.

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

II-4: $f(X) = f_{12}(x_1, x_2)\prod_{k=3}^{q} f_k(x_k)$, where $f_{12}(x_1, x_2)$ is an 3×3 matrix whose (j,l)-entry is $\exp\{-(j+l)(x_1+x_2)\}$, and $f_k(x_k)$ is an 3×3 matrix whose (j,l)-entry is $\sin(2\pi x_k)$.

To maintain the consistency of the simulations, we use the same way as Lin et al. (2022) to measure the quality of the estimation. For settings II-3 and II-4, we consider $p = 3, 4, 10, 20$

Model	(p, n)	MAM	ILPR	GFR	RFWLCFR/FRF	RFWLLFR
	$(3, 100)$	**0.415** (0.023)	0.922 (0.126)	0.970 (0.012)	0.714 (0.018)	0.794 (0.023)
	$(3, 200)$	**0.299** (0.017)	0.796 (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)
II-3	$(4, 200)$	**0.357** (0.019)	0.916 (0.021)	0.970 (0.010)	0.748 (0.014)	0.832 (0.018)
(SNR$=2$)	$(10, 500)$	NA	NA	0.967 (0.008)	**0.774** (0.014)	0.929 (0.015)
	$(10, 1000)$	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)
	$(3, 100)$	0.744 (0.054)	0.672 (0.189)	0.773 (0.051)	**0.501** (0.051)	0.540 (0.058)
	$(3, 200)$	0.713 (0.049)	0.481 (0.087)	0.761 (0.049)	**0.439** (0.037)	0.465 (0.041)
	$(4, 100)$	0.841 (0.065)	0.834 (0.146)	0.855 (0.064)	**0.676** (0.078)	0.788 (0.125)
II-4	$(4, 200)$	0.835 (0.063)	0.758 (0.113)	0.841 (0.060)	**0.601** (0.073)	0.684 (0.092)
(SNR$=2$)	$(10, 500)$	NA	NA	0.838 (0.061)	**0.681** (0.076)	0.773 (0.068)
	$(10, 1000)$	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)
	$(3, 100)$	**0.346** (0.022)	0.916 (0.136)	0.965 (0.011)	0.693 (0.017)	0.755 (0.019)
	$(3, 200)$	**0.229** (0.011)	0.774 (0.058)	0.954 (0.008)	0.589 (0.012)	0.616 (0.015)
	$(4, 100)$	**0.449** (0.033)	0.948 (0.030)	0.979 (0.012)	0.789 (0.017)	0.884 (0.025)
II-3	$(4, 200)$	**0.284** (0.012)	0.902 (0.026)	0.966 (0.010)	0.734 (0.013)	0.802 (0.016)
(SNR$=4$)	$(10, 500)$	NA	NA	0.964 (0.008)	**0.764** (0.013)	0.894 (0.013)
	$(10, 1000)$	NA	NA	0.958 (0.009)	**0.732** (0.012)	0.847 (0.011)
	$(20, 1000)$	NA	NA	0.964 (0.009)	**0.771** (0.010)	0.919 (0.013)
	$(20, 2000)$	NA	NA	0.957 (0.008)	**0.744** (0.010)	0.886 (0.011)
	$(3, 100)$	0.736 (0.054)	0.655 (0.199)	0.770 (0.051)	**0.466** (0.054)	0.477 (0.059)
	$(3, 200)$	0.709 (0.049)	0.439 (0.084)	0.760 (0.048)	0.404 (0.040)	**0.401** (0.043)
	$(4, 100)$	0.841 (0.066)	0.853 (0.162)	0.851 (0.066)	**0.656** (0.083)	0.746 (0.126)
II-4	$(4, 200)$	0.834 (0.063)	0.758 (0.131)	0.839 (0.060)	**0.582** (0.076)	0.648 (0.093)
(SNR$=4$)	$(10, 500)$	NA	NA	0.836 (0.061)	**0.675** (0.076)	0.749 (0.069)
	$(10, 1000)$	NA	NA	0.829 (0.061)	**0.636** (0.080)	0.696 (0.074)
	$(20, 1000)$	NA	NA	0.835 (0.061)	**0.734** (0.080)	0.822 (0.078)
	$(20, 2000)$	NA	NA	0.830 (0.060)	**0.695** (0.076)	0.769 (0.070)

Table 7: Average MSE (standard deviation) of different methods for setting II-3,4 with SNR = 2, 4 and Log-Cholesky metric over 100 simulation runs. Bold-faced numbers indicate the best performers.

and two choices of n for each p. For $p = 3, 4$, the setting is the same as Lin et al. (2022). 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 7 shows the results. For setting II-3 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 II-4, RFWLCFR/FRF tends to perform the best. This setting also indicates that there are indeed cases where RFWLLFR will perform worse than RFWLCFR. It may be more efficient for complex settings to use RFWLCFR, whose mechanism is similar to that of random forests. Since both MAM and GFR make specific model assumptions, setting II-4 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 thus they are not feasible when the dimension of X is large. Through this experiment, we again demonstrate the outstanding performance of our methods for relatively high-dimensional 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.

Appendix D. Additional Materials for New York Taxi Data

The map for Manhattan excluding the islands is delimited in Figure 13 (from Dubey and Müller (2020)). Manhattan can be further grouped into ten distinct zones, as detailed in Table 8, and the center point of each zone is selected as in Figure 13.

Zone	Towns
1	Inwood, Fort George, Washington Heights, Hamilton Heights, Harlem, East Harlem
2	Upper West Side, Morningside Heights, Central Park
3	Yorkville, Lenox Hill, Upper East Side
4	Lincoln Square, Clinton, Chelsea, Hell's Kitchen
5	Garment District, Theatre District
6	Midtown
7	Midtown South
8	Turtle Bay, Murray Hill, Kips Bay, Gramercy Park, Sutton, Tudor, Medical City, Stuy Town
9	Meat packing district, Greenwich Village, West Village, Soho, Little Italy, China Town, Civic Center, Noho
10	Lower East Side, East Village, ABD Park, Bowery, Two Bridges, Southern tip, White Hall, Tribecca, Wall Street

Table 8: 10 zones in Manhattan defined for the New York taxi data analysis

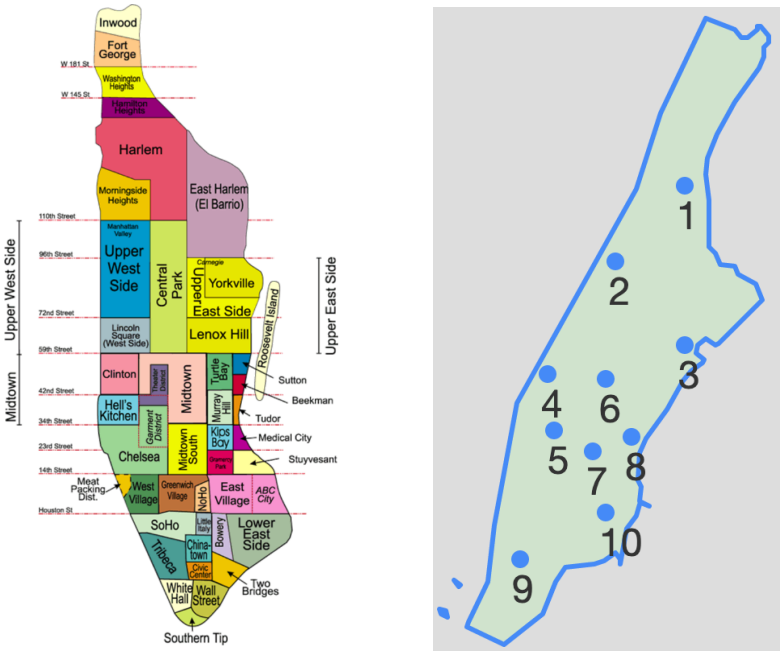


Figure 13: Towns in Manhattan (left) and center points for zones (right)

Appendix E. Proofs

In what follows, we prove our main results.

E.1 Some Preparation

To facilitate the proof of important results in Section 3, we give some preparatory work as follows. Recall

$$\begin{aligned}\hat{r}_{\oplus}(x) &= \underset{y\in\Omega}{\operatorname{argmin}}\hat{R}_n(x,y)\\ &= \underset{y\in\Omega}{\operatorname{argmin}}\sum_{i=1}^{n}\bar{\alpha}_i\left(x\right)d^2\left(Y_i,y\right) \qquad (16)\\ &= \underset{y\in\Omega}{\operatorname{argmin}}\binom{n}{s_n}^{-1}\sum_{k}E_{\xi\sim\Xi}\left\{\frac{1}{N(L(x;\mathcal{D}_n^k,\xi))}\sum_{i:X_i\in L(x;\mathcal{D}_n^k,\xi)}d^2\left(Y_i,y\right)\right\}. \qquad (17)\end{aligned}$$

and

$$\begin{aligned}\tilde{r}_{\oplus}(x) &= \underset{y\in\Omega}{\operatorname{argmin}}\tilde{R}_n(x,y)\\ &= \underset{y\in\Omega}{\operatorname{argmin}}\, nE\big\{\bar{\alpha}_i\left(x\right)d^2\left(Y_i,y\right)\big\} \qquad (18)\\ &= \underset{y\in\Omega}{\operatorname{argmin}}E\left\{\frac{1}{N(L(x;\mathcal{D}_n^k,\xi))}\sum_{i:X_i\in L(x;\mathcal{D}_n^k,\xi)}d^2\left(Y_i,y\right)\right\}. \qquad (19)\end{aligned}$$

The goal of the Fréchet regression is

$$m_{\oplus}(x)=\underset{y\in\Omega}{\operatorname{argmin}}M_{\oplus}(x,y)=\underset{y\in\Omega}{\operatorname{argmin}}E\big\{d^2(Y,y)\mid X=x\big\}.$$

We conduct asymptotic analysis by separating $d(\hat{r}_{\oplus}(x),m_{\oplus}(x))$ into the bias term $d(\tilde{r}_{\oplus}(x),m_{\oplus}(x))$ and the variance term $d(\hat{r}_{\oplus}(x),\tilde{r}_{\oplus}(x))$. As $\hat{r}_{\oplus}(x)$ have two different expressions, choosing a suitable form will bring great convenience for our theoretical developments. When we adopt (17) and (19), the theory of infinite order U-statistics and U-processes can be applied. And when we adopt (16) and (18), the perspective of the weighted average is helpful.

Under the honesty assumption, we have

$$E\left\{\frac{1}{N(L(x;\mathcal{D}_n^k,\xi))}\sum_{i:X_i\in L(x;\mathcal{D}_n^k,\xi)}d^2\left(Y_i,y\right)\right\}=E\big[E\big\{d^2(Y,y)\mid X\in L(x)\big\}\big],$$

where $L(x)$ is the leaf node containing x of any honest Fréchet tree satisfying the assumption (A3). We emphasize here that $N(L(x;\mathcal{D}_n^k,\xi))$ and $X_i\in L(x;\mathcal{D}_n^k,\xi)$ don't involve sample points whose responses have been used to construct the Fréchet tree. Then (19) can be further rewritten as

$$\tilde{r}_{\oplus}(x)=\underset{y\in\Omega}{\operatorname{argmin}}\tilde{R}_n(x,y)=\underset{y\in\Omega}{\operatorname{argmin}}E\big[E\big\{d^2(Y,y)\mid X\in L(x)\big\}\big].$$

E.2 Proofs of Results in Section 3.2

To prove Theorem 1, we need to prove the convergence of the bias term $d(\tilde{r}_{\oplus}(x), m_{\oplus}(x))$ and variance term $d(\hat{r}_{\oplus}(x), \tilde{r}_{\oplus}(x))$, respectively.

Lemma 16 *Suppose that for a fixed $x \in [0,1]^p$, (A1)–(A4) hold and the Fréchet trees are honest. Then,*

$$d(\tilde{r}_{\oplus}(x), m_{\oplus}(x)) = o(1).$$

Proof [Proof of Lemma 16] By the proof of Theorem 3 in Petersen and Müller (2019), we have the relationship that

$$\mathrm{d}F_{Y|X}(x,y)/\mathrm{d}F_Y(y) = g_y(x)/f(x)$$

for all x such that $f(x) > 0$. To prevent notation confusion, we consider a fixed $x_0 \in [0,1]^p$ and $y_0 \in \Omega$, then

$$\begin{aligned} M_{\oplus}(x_0, y_0) = E\big\{d^2(Y, y_0) \mid X = x_0\big\} &= \int_{\Omega} d^2(y, y_0) dF_{Y|X}(x_0, y) \\ &= \int_{\Omega} d^2(y, y_0) \frac{g_y(x_0)}{f(x_0)} dF_Y(y) \end{aligned}$$

and

$$\begin{aligned} E\big\{d^2(Y, y_0) \mid X \in L(x_0)\big\} &= \int_{L(x_0)} \left(\int_{\Omega} d^2(y, y_0) \frac{1}{P\{X \in L(x_0)\}} dF_{Y|X}(x, y) \right) f(x) dx \\ &= \int_{L(x_0)} \int_{\Omega} d^2(y, y_0) \frac{g_y(x)}{P\{X \in L(x_0)\}} dF_Y(y) dx \\ &= \int_{\Omega} d^2(y, y_0) \left(\int_{L(x_0)} \frac{g_y(x)}{P\{X \in L(x_0)\}} dx \right) dF_Y(y). \end{aligned}$$

It is obvious to have

$$\int_{L(x_0)} \frac{g_y(x)}{P\{X \in L(x_0)\}} dx = \frac{\int_{L(x_0)} g_y(x) dx}{\int_{L(x_0)} f(x) dx}.$$

Since $g_y(x)$ is continuous by the assumption (A2), for every $(x_0, y) \in [0,1]^p \times \Omega$, $\forall\, \epsilon > 0$, $\exists\, \delta_{\epsilon}^1 > 0$ such that when $x \in B(x_0, \delta_{\epsilon}^1) = \big\{x : \|x - x_0\| \le \delta_{\epsilon}^1\big\}$, we have

$$|g_y(x) - g_y(x_0)| \le \epsilon.$$

Thus, $\exists\, \delta_{\epsilon}^1 > 0$ such that

$$\begin{aligned} &\left| \int_{L(x_0)} \{g_y(x) - g_y(x_0)\} dx \right| \\ \le &\int_{L(x_0)} |g_y(x) - g_y(x_0)| dx \\ = &\int_{L(x_0) \cap B(x_0, \delta_{\epsilon}^1)} |g_y(x) - g_y(x_0)| \, dx + \int_{L(x_0) \setminus B(x_0, \delta_{\epsilon}^1)} |g_y(x) - g_y(x_0)| \, dx \end{aligned}$$

$$\begin{aligned}
&\leq \epsilon \operatorname{Vol}\left(L(x_0) \cap B(x_0, \delta_\epsilon^1)\right) + 2\|g_y\|_\infty \operatorname{Vol}\left(L(x_0) \backslash B(x_0, \delta_\epsilon^1)\right) \\
&\leq \epsilon \operatorname{Vol}\left(L(x_0)\right) + 2\|g_y\|_\infty \operatorname{Vol}\left(L(x_0) \backslash B(x_0, \delta_\epsilon^1)\right) \qquad (20)
\end{aligned}$$

where $\|g_y\|_\infty$ denotes the supremum norm of g_y and Vol denotes the volume of subsets in the $[0,1]^p$. Since the marginal density f is also continuous, using the same argument, for above ϵ, $\exists\ \delta_\epsilon^2$ such that

$$\int_{L(x_0)} \{f(x) - f(x_0)\} dx \leq \epsilon \operatorname{Vol}(L(x_0)) + 2\|f\|_\infty \operatorname{Vol}\left(L(x_0) \backslash B(x_0, \delta_\epsilon^2)\right) \qquad (21)$$

We define $\delta_\epsilon = \min(\delta_\epsilon^1, \delta_\epsilon^2)$. For two sequences of positive functions $\{f_n\}$, $\{g_n\}$ and for another two positive functions $\{f_n'\}$ and $\{g_n'\}$, we have

$$\left|\frac{f_n}{g_n} - \frac{f_n'}{g_n'}\right| = \left|\frac{f_n}{g_n} - \frac{f_n'}{g_n} + \frac{f_n'}{g_n} - \frac{f_n'}{g_n'}\right| = \left|\frac{f_n - f_n'}{g_n} - f_n' \frac{g_n' - g_n}{g_n' g_n}\right| \leq \frac{|f_n - f_n'|}{g_n} + f_n' \frac{|g_n - g_n'|}{g_n' g_n}. \qquad (22)$$

We take

$$f_n(x_0, y) = \int_{L(x_0)} g_y(x) dx; \qquad f_n'(x_0, y) = \int_{L(x_0)} g_y(x_0) dx;$$

$$g_n(x_0) = \int_{L(x_0)} f(x) dx; \qquad g_n'(x_0) = \int_{L(x_0)} f(x_0) dx.$$

By (20), for above ϵ, we have

$$\begin{aligned}
\frac{|f_n(x_0, y) - f_n'(x_0, y)|}{g_n(x_0)} &= \frac{\left|\int_{L(x_0)} g_y(x) dx - \int_{L(x_0)} g_y(x_0) dx\right|}{\int_{L(x_0)} f(x) dx} \\
&\leq \frac{\left|\int_{L(x_0)} \{g_y(x) - g_y(x_0)\} dx\right|}{f_{\min} \operatorname{Vol}\left(L(x_0)\right)} \\
&\leq \frac{\epsilon \operatorname{Vol}\left(L(x_0)\right) + 2\|g_y\|_\infty \operatorname{Vol}\left(L(x_0) \backslash B\left(x_0, \delta_\epsilon\right)\right)}{f_{\min} \operatorname{Vol}\left(L(x_0)\right)} \\
&= \frac{\epsilon}{f_{\min}} + \frac{2\|g_y\|_\infty \operatorname{Vol}\left(L(x_0) \backslash B\left(x_0, \delta_\epsilon\right)\right)}{f_{\min} \operatorname{Vol}\left(L(x_0)\right)}.
\end{aligned}$$

And similarly by (21), we have

$$\frac{|g_n(x_0) - g_n'(x_0)|}{g_n(x_0)} = \frac{\left|\int_{L(x_0)} f(x) dx - \int_{L(x_0)} f(x_0) dx\right|}{\int_{L(x_0)} f(x) dx} \leq \frac{\epsilon}{f_{\min}} + \frac{2\|f\|_\infty \operatorname{Vol}\left(L(x_0) \backslash B\left(x_0, \delta_\epsilon\right)\right)}{f_{\min} \operatorname{Vol}\left(L(x_0)\right)}.$$

By the assumption (A3), $\operatorname{diam}\left(L(x_0)\right) \to 0$ in probability. Hence, for δ_ϵ defined above,

$$\lim_{n \to +\infty} P\left\{\operatorname{diam}\left(L(x_0)\right) < \delta_\epsilon\right\} = 1.$$

Obviously when $\operatorname{diam}\left(L(x_0)\right) < \delta_\epsilon$, $\operatorname{Vol}\left(L(x_0) \backslash B\left(x_0, \delta_\epsilon\right)\right) = 0$ holds. Therefore

$$P\left\{\operatorname{diam}\left(L(x_0)\right) < \delta_\epsilon\right\} \leq P\left\{\frac{\operatorname{Vol}\left(L(x_0) \backslash B\left(x_0, \delta_\epsilon\right)\right)}{\operatorname{Vol}\left(L(x_0)\right)} = 0\right\}.$$

Take the limit on both sides of the above formula, we have $\frac{\mathrm{Vol}(L(x_0)\setminus B(x_0,\delta_\epsilon))}{\mathrm{Vol}(L(x_0))} \to 0$ in probability. Combine with the assumption (A2), then

$$\frac{|f_n(x_0,y)-f_n'(x_0,y)|}{g_n(x_0)} \xrightarrow{P} \frac{\epsilon}{f_{\min}} \quad \text{and} \quad \frac{|g_n(x_0)-g_n'(x_0)|}{g_n(x_0)} \xrightarrow{P} \frac{\epsilon}{f_{\min}}. \tag{23}$$

Finally, combine (22) and (23), then the following formula holds

$$\begin{aligned}\left|\frac{\int_{L(x_0)} g_y(x)dx}{\int_{L(x_0)} f(x)dx} - \frac{g_y(x_0)}{f(x_0)}\right| &= \left|\frac{f_n(x_0,y)}{g_n(x_0)} - \frac{f_n'(x_0,y)}{g_n'(x_0)}\right| \\ &\le \frac{|f_n(x_0,y)-f_n'(x_0,y)|}{g_n(x_0)} + f_n'(x_0,y)\frac{|g_n(x_0)-g_n'(x_0)|}{g_n'(x_0)g_n(x_0)} \\ &\xrightarrow{P} \frac{\epsilon}{f_{\min}} + \frac{\epsilon}{f_{\min}}\frac{g_y(x_0)}{f(x_0)}.\end{aligned}$$

Let $\epsilon \to 0$, we can get for each $y \in \Omega$

$$\left|\frac{\int_{L(x_0)} g_y(x)dx}{\int_{L(x_0)} f(x)dx} - \frac{g_y(x_0)}{f(x_0)}\right| \xrightarrow{P} 0.$$

Moreover,

$$\left|\frac{\int_{L(x_0)} g_y(x)dx}{\int_{L(x_0)} f(x)dx}\right| \le \frac{\|g_y\|_\infty \mathrm{Vol}(L(x_0))}{f_{\min}\mathrm{Vol}(L(x_0))} < \infty.$$

By the dominated convergence theorem and the assumption (A1), we conclude

$$\begin{aligned}&\sup_{y_0\in\Omega}\left|E\left\{d^2(Y,y_0) \mid X \in L(x_0)\right\} - M_\oplus(x_0,y_0)\right| \\ &= \sup_{y_0\in\Omega}\left|\int_\Omega d^2(y,y_0)\left\{\frac{\int_{L(x_0)} g_y(x)dx}{\int_{L(x_0)} f(x)dx} - \frac{g_y(x_0)}{f(x_0)}\right\}dF_Y(y)\right| \\ &\le \sup_{y_0\in\Omega}\int_\Omega d^2(y,y_0)\left|\frac{\int_{L(x_0)} g_y(x)dx}{\int_{L(x_0)} f(x)dx} - \frac{g_y(x_0)}{f(x_0)}\right|dF_Y(y) \\ &\le \int_\Omega \sup_{y_0\in\Omega} d^2(y,y_0)\left|\frac{\int_{L(x_0)} g_y(x)dx}{\int_{L(x_0)} f(x)dx} - \frac{g_y(x_0)}{f(x_0)}\right|dF_Y(y) \\ &\xrightarrow{P} 0.\end{aligned} \tag{24}$$

Under the honest condition, we have

$$\tilde{R}_n(x_0,y_0) = E\left\{\frac{1}{N(L(x_0;\mathcal{D}_n^k,\xi))}\sum_{i:X_i\in L(x_0;\mathcal{D}_n^k,\xi)} d^2(Y_i,y_0)\right\} = E\left[E\left\{d^2(Y,y_0) \mid X \in L(x_0)\right\}\right].$$

By the dominated convergence theorem, we take the expectation about $L(x_0)$ on both sides of (24) and get

$$\sup_{y_0\in\Omega}\left|\tilde{R}_n(x_0,y_0) - M_\oplus(x_0,y_0)\right|$$

$$
\begin{aligned}
&= \sup_{y_0 \in \Omega} \left| E\left[E\{ d^2(Y, y_0) | X \in L(x_0) \} - M_{\oplus}(x_0, y_0) \right] \right| \\
&\leq \sup_{y_0 \in \Omega} E \left| E\{ d^2(Y, y_0) | X \in L(x_0) \} - M_{\oplus}(x_0, y_0) \right| \\
&\leq E \left(\sup_{y_0 \in \Omega} \left| E\{ d^2(Y, y_0) | X \in L(x_0) \} - M_{\oplus}(x_0, y_0) \right| \right) \\
&\to 0.
\end{aligned}
$$

By the assumption (A4), we then get

$$d(\tilde{r}_{\oplus}(x_0), m_{\oplus}(x_0)) = o(1).$$

■

Lemma 17 *Suppose that for a fixed $x \in [0,1]^p$, (A1), (A4) hold and the Fréchet trees are symmetric. Then,*

$$d(\hat{r}_{\oplus}(x), \tilde{r}_{\oplus}(x)) = o_p(1).$$

Before giving the proof of Lemma 17, we need to prove another lemma first. It is the generalization of Corollary 3.2.3 of van der Vaart and Wellner (1996).

Lemma 18 *Let $\mathbb{M}_n$ be stochastic processes indexed by a metric space Θ, and let $M_n : \Theta \mapsto \mathcal{R}$ be deterministic functions. Suppose that $\|\mathbb{M}_n - M_n\|_{\Theta} \to 0$ in probability and that there exists a sequence θ_n such that*

$$\liminf_n \inf_{d(\theta, \theta_n) > \varepsilon} \{ M_n(\theta) - M_n(\theta_n) \} > 0$$

for every $\varepsilon > 0$. Then any sequence $\hat{\theta}_n$, such that $\mathbb{M}_n(\hat{\theta}_n) \leq \inf_{\theta} \mathbb{M}_n(\theta) + o_P(1)$, satisfies $d(\hat{\theta}_n, \theta_n) \to 0$ in probability.

Proof [Proof of Lemma 18] By the requirement of $\hat{\theta}_n$, $\mathbb{M}_n(\hat{\theta}_n) \leq \mathbb{M}_n(\theta_n) + o_P(1)$. Since $\|\mathbb{M}_n - M_n\|_{\Theta} \to 0$ in probability, $\mathbb{M}_n(\theta_n) - M_n(\theta_n) \to 0$ in probability. So we have

$$\mathbb{M}_n(\hat{\theta}_n) \leq M_n(\theta_n) + o_P(1).$$

Therefore, again by the uniform convergence, we have

$$
\begin{aligned}
M_n(\hat{\theta}_n) - M_n(\theta_n) &\leq M_n(\hat{\theta}_n) - \mathbb{M}_n(\hat{\theta}_n) + o_P(1) \\
&\leq \|\mathbb{M}_n - M_n\|_{\Theta} + o_P(1) \\
&\xrightarrow{P} 0.
\end{aligned} \tag{25}
$$

From the requirement of θ_n, given $\epsilon > 0$, there exists $\delta > 0, N \in \mathcal{N}^+$, when $n \geq N$ and $d(\theta, \theta_n) \geq \epsilon$, we have

$$M_n(\theta) - M_n(\theta_n) > \delta.$$

So $\left\{d(\hat{\theta}_n, \theta_n) \geq \epsilon\right\} \subseteq \left\{M_n(\hat{\theta}_n) - M_n(\theta_n) > \delta\right\}$, however, by (25)

$$P\left\{M_n(\hat{\theta}_n) - M_n(\theta_n) > \delta\right\} \to 0.$$

Therefore

$$P\left\{d(\hat{\theta}_n, \theta_n) \geq \epsilon\right\} \to 0.$$

i.e.,

$$d(\hat{\theta}_n, \theta_n) \to 0 \text{ in probability.}$$

■

Proof [Proof of Lemma 17] For a fixed $x \in [0,1]^p$, by Lemma 18 and the assumption (A4), we only need to prove convergence of $\sup_{y\in\Omega}\left|\hat{R}_n(x,y) - \tilde{R}_n(x,y)\right|$ to zero in probability. To implement this, we can show $\hat{R}_n(x,\cdot) - \tilde{R}_n(x,\cdot) \rightsquigarrow 0$ in $l^\infty(\Omega)$ which denotes the space of bounded functions on Ω, and apply Theorem 1.3.6 of van der Vaart and Wellner (1996). Thanks to Theorem 1.5.4 of van der Vaart and Wellner (1996), this weak convergence is equivalent to $\hat{R}_n(x,\cdot) - \tilde{R}_n(x,\cdot)$ is asymptotically tight and the marginals converge weakly. By Theorem 1.5.7 of van der Vaart and Wellner (1996), the asymptotically tight continues to be equivalent to two requirements that $\hat{R}_n(x,y) - \tilde{R}_n(x,y)$ is asymptotically tight in $\mathcal{R}$ for every $y \in \Omega$ and $\hat{R}_n(x,\cdot) - \tilde{R}_n(x,\cdot)$ is asymptotically uniformly d-equicontinuous in probability. So the proof will be finished if the following conditions hold

(i) $\hat{R}_n(x,y) - \tilde{R}_n(x,y) = o_p(1)$ for each $y \in \Omega$,

(ii) For all $\varepsilon, \eta > 0$, there exists $\delta > 0$ such that

$$\limsup_n P\left\{\sup_{d(y_1,y_2)<\delta}\left|\left(\hat{R}_n - \tilde{R}_n\right)(x,y_1) - \left(\hat{R}_n - \tilde{R}_n\right)(x,y_2)\right| > \varepsilon\right\} < \eta.$$

First, prove (i): We consider the expressions of (17) and (19). $\hat{R}_n(x,y)$ is an infinite order U-statistic for each $y \in \Omega$. Since (Ω, d) is a bounded metric space by the assumption (A1), the kernels of $\hat{R}_n(x,y)$ are uniformly bounded. By Remark 3.1 of DiCiccio and Romano (2022), we have $\hat{R}_n(x,y) - \tilde{R}_n(x,y) = o_p(1)$ for each $y \in \Omega$.

Then (ii): We consider the expressions of (16) and (18). For any $y_1, y_2 \in \Omega$,

$$\begin{aligned}
&\left|\left(\hat{R}_n - \tilde{R}_n\right)(x,y_1) - \left(\hat{R}_n - \tilde{R}_n\right)(x,y_2)\right| \\
\leq &\left|\hat{R}_n(x,y_1) - \hat{R}_n(x,y_2)\right| + \left|\tilde{R}_n(x,y_1) - \tilde{R}_n(x,y_2)\right| \\
= &\left|\sum_{i=1}^n \bar{\alpha}_i(x)\left\{d^2(Y_i,y_1) - d^2(Y_i,y_2)\right\}\right| + \left|nE\left[\bar{\alpha}_i(x)\left\{d^2(Y_i,y_1) - d^2(Y_i,y_2)\right\}\right]\right| \\
\leq &\sum_{i=1}^n \left|\bar{\alpha}_i(x)\left\|d(Y_i,y_1) - d(Y_i,y_2)\right\| d(Y_i,y_1) + d(Y_i,y_2)\right| \\
&+ nE\left\{\left|\bar{\alpha}_i(x)\left\|d(Y_i,y_1) - d(Y_i,y_2)\right\| d(Y_i,y_1) + d(Y_i,y_2)\right|\right\}
\end{aligned}$$

$$
\begin{aligned}
\leq & 2\operatorname{diam}(\Omega) d\left(y_1, y_2\right) \sum_{i=1}^n \bar{\alpha}_i(x)+2 \operatorname{diam}(\Omega) d\left(y_1, y_2\right)\left[n E\left\{\bar{\alpha}_i(x)\right\}\right] \\
= & 4 \operatorname{diam}(\Omega) d\left(y_1, y_2\right) \\
= & O_p\left(d\left(y_1, y_2\right)\right)
\end{aligned}
$$

where the O_p term is independent of y_1 and y_2. The second equality is because $\bar{\alpha}_i(x), i=1, \ldots, n$ are identically distributed and a fact

$$
\begin{aligned}
\sum_{i=1}^n \bar{\alpha}_i(x) & =\sum_{i=1}^n\binom{n}{s_n}^{-1} \sum_k E_{\xi \sim \Xi} \frac{1\left\{X_i \in L(x ; \mathcal{D}_n^k, \xi)\right\}}{N(L(x ; \mathcal{D}_n^k, \xi))} \\
& =\binom{n}{s_n}^{-1} \sum_k E_{\xi \sim \Xi} \sum_{i=1}^n \frac{1\left\{X_i \in L(x ; \mathcal{D}_n^k, \xi)\right\}}{N(L(x ; \mathcal{D}_n^k, \xi))} \\
& =1 .
\end{aligned}
$$

Hence

$$
\sup_{d\left(y_1, y_2\right)<\delta}\left|\left(\hat{R}_n-\tilde{R}_n\right)\left(x, y_1\right)-\left(\hat{R}_n-\tilde{R}_n\right)\left(x, y_2\right)\right|=O_p(\delta)
$$

which can deduce (ii). So, $d(\hat{r}_{\oplus}(x), \tilde{r}_{\oplus}(x))=o_p(1)$ ■

Based on Lemma 16 and Lemma 17, we can prove Theorem 1 easily.

Proof [Proof of Theorem 1] Notice that

$$
d(\hat{r}_{\oplus}(x), m_{\oplus}(x)) \leq d(\hat{r}_{\oplus}(x), \tilde{r}_{\oplus}(x))+d(\tilde{r}_{\oplus}(x), m_{\oplus}(x)).
$$

By the results of Lemma 16 and Lemma 17, we complete the proof. ■

Before the proof of Theorem 3, we state the required assumptions (U1)–(U4).

(U1) For any $\|x\| \leq J$, $M_{\oplus}(x, y)$ is equicontinuous, *i.e.*,

$$
\limsup_{\dot{x} \to x} \sup_{y \in \Omega}\left|M_{\oplus}(\dot{x}, y)-M_{\oplus}(x, y)\right|=0.
$$

(U2) The marginal density f of X, as well as the conditional densities g_y of $X \mid Y=y$, exist and are bounded and uniformly continuous, the latter for all $y \in \Omega$. And f is also bounded away from zero such that $0<f_{\min } \leq f$. Additionally, for any open $V \subseteq \Omega$, $\int_V \mathrm{d} F_{Y|X}(x, y)$ is continuous as a function of x.

(U3) $\sup_{\|x\| \leq J} \operatorname{diam}(L(x)) \to 0$ in probability, where $L(x)$ is the leaf node containing x of any Fréchet tree in the random forest.

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

$$
\inf_{\|x\| \leq J} \inf_{d(y, m_{\oplus}(x))>\varepsilon}\left\{M_{\oplus}(x, y)-M_{\oplus}\left(x, m_{\oplus}(x)\right)\right\}>0,
$$

$$\liminf_n \inf_{\|x\|\le J} \inf_{d(y,\tilde r_\oplus(x))>\varepsilon} \left\{\tilde R_n(x,y) - \tilde R_n\left(x,\tilde r_\oplus(x)\right)\right\} > 0,$$

and there exists $\zeta = \zeta(\varepsilon) > 0$ such that

$$P\left\{\inf_{\|x\|\le J} \inf_{d(y,\hat r_\oplus(x))>\varepsilon} \left(\hat R_n(x,y) - \hat R_n\left(x,\hat r_\oplus(x)\right)\right) \ge \zeta\right\} \to 1.$$

Proof [Proof of Theorem 3] We need to prove the following two results:

(i) $\sup_{\|x\|\le J} d(\hat r_\oplus(x), \tilde r_\oplus(x)) = o_p(1)$,

(ii) $\sup_{\|x\|\le J} d(\tilde r_\oplus(x), m_\oplus(x)) = o(1)$.

First prove (i): By Lemma 17, given any $x \in [0,1]^p$, $d(\hat r_\oplus(x), \tilde r_\oplus(x)) = o_p(1)$. Now consider the process $D_n(x) = d(\hat r_\oplus(x), \tilde r_\oplus(x))$ with $\|x\| \le J$. As the proof of Lemma 17, based on Theorems 1.5.4, 1.5.7 and 1.3.6 of van der Vaart and Wellner (1996), it suffices to show that for any $S > 0$, as $\delta \to 0$,

$$\limsup_{n\to\infty} P\left\{\sup_{\substack{\|x_1-x_2\|<\delta \\ \|x_1\|,\|x_2\|\le J}} |D_n(x_1) - D_n(x_2)| > 2S\right\} \to 0.$$

Since

$$\begin{aligned}
&|D_n(x) - D_n(y)| \\
=&|d(\hat r_\oplus(x_1), \tilde r_\oplus(x_1)) - d(\hat r_\oplus(x_2), \tilde r_\oplus(x_2))| \\
=&|d(\hat r_\oplus(x_1), \tilde r_\oplus(x_1)) - d(\hat r_\oplus(x_1), \tilde r_\oplus(x_2)) + d(\hat r_\oplus(x_1), \tilde r_\oplus(x_2)) - d(\hat r_\oplus(x_2), \tilde r_\oplus(x_2))| \\
\le& d(\tilde r_\oplus(x_1), \tilde r_\oplus(x_2)) + d(\hat r_\oplus(x_1), \hat r_\oplus(x_2)),
\end{aligned}$$

it suffices to show that, as $\delta \to 0$,

$$\limsup_{n\to\infty} \sup_{\substack{\|x_1-x_2\|<\delta \\ \|x_1\|,\|x_2\|\le J}} d(\tilde r_\oplus(x_1), \tilde r_\oplus(x_2)) \to 0, \tag{26}$$

and

$$\limsup_{n\to\infty} P\left\{\sup_{\substack{\|x_1-x_2\|<\delta \\ \|x_1\|,\|x_2\|\le J}} d(\hat r_\oplus(x_1), \hat r_\oplus(x_2)) > S\right\} \to 0. \tag{27}$$

Recall that we have proved $\sup_{y\in\Omega} \left|\tilde R_n(x,y) - M_\oplus(x,y)\right| \to 0$ for any $x \in [0,1]^p$ in the proof of Lemma 16. Since the density f and g_f are uniformly continuous by the assumption (U2) and $\sup_{\|x\|\le J} \operatorname{diam}(L(x)) \to 0$ in probability by the assumption (U3), we can get stronger convergence

$$\sup_{\|x\|\le J, y\in\Omega} \left|\tilde R_n(x,y) - M_\oplus(x,y)\right| \to 0. \tag{28}$$

Notice that

$$\sup_{\substack{\|x_1-x_2\|<\delta \\ \|x_1\|,\|x_2\|\le J}} \sup_{y\in\Omega} \left|\tilde R_n(x_1,y) - \tilde R_n(x_2,y)\right|$$

$$\leq \sup_{\substack{\|x_1-x_2\|<\delta \\ \|x_1\|,\|x_2\|\leq J}} \sup_{y\in\Omega} |M_\oplus(x_1,y) - M_\oplus(x_2,y)| + 2 \sup_{\|x\|\leq J, y\in\Omega} \left|\tilde{R}_n(x,y) - M_\oplus(x,y)\right|.$$

Combining the assumption (U1) and (28), we can obtain, as $\delta \to 0$,

$$\limsup_{n\to\infty} \sup_{\substack{\|x_1-x_2\|<\delta \\ \|x_1\|,\|x_2\|\leq J}} \sup_{y\in\Omega} \left|\tilde{R}_n(x_1,y) - \tilde{R}_n(x_2,y)\right| \to 0.$$

Then (26) holds by assumption (U4). Now consider (27), let $\epsilon > 0$ and suppose $d(\hat{r}_\oplus(x_1), \hat{r}_\oplus(x_2)) > \epsilon$ with $\|x_1\|, \|x_2\| \leq J$. Then the assumption (U4) and the form of $\hat{R}_n(x,y)$ imply that

$$\zeta \leq \sup_{\substack{\|x_1-x_2\|<\delta \\ \|x_1\|,\|x_2\|\leq J}} \sup_{y\in\Omega} \left|\hat{R}_n(x_1,y) - \hat{R}_n(x_2,y)\right| = O_p(\delta)$$

and (27) follows as $\delta \to 0$. At this point, the proof of (i) is finished.

Next prove (ii): By Lemma 16, given any $x \in [0,1]^p$, $d(\tilde{r}_\oplus(x), m_\oplus(x)) = o(1)$. Similarly, we consider $F_n(x) = d(\tilde{r}_\oplus(x), m_\oplus(x))$. It suffices to show that, as $\delta \to 0$,

$$\limsup_{n\to\infty} \sup_{\substack{\|x_1-x_2\|<\delta \\ \|x_1\|,\|x_2\|\leq J}} |F_n(x_1) - F_n(x_2)| \to 0.$$

Since

$$|F_n(x_1) - F_n(x_2)| \leq d(m_\oplus(x_1), m_\oplus(x_2)) + d(\tilde{r}_\oplus(x_1), \tilde{r}_\oplus(x_2)),$$

it suffices to show that, as $\delta \to 0$,

$$\sup_{\substack{\|x_1-x_2\|<\delta \\ \|x_1\|,\|x_2\|\leq J}} d(m_\oplus(x_1), m_\oplus(x_2)) \to 0, \tag{29}$$

and

$$\limsup_{n\to\infty} \sup_{\substack{\|x_1-x_2\|<\delta \\ \|x_1\|,\|x_2\|\leq J}} d(\tilde{r}_\oplus(x_1), \tilde{r}_\oplus(x_2)) \to 0. \tag{30}$$

Based on the assumption (U1) and (U4), it is not difficult to prove that $m_\oplus(x)$ is continuous at x and hence uniformly continuous on $\{x : \|x\| \leq J\}$ considering the compactness of $\{x : \|x\| \leq J\}$. Then (29) naturally holds. And (30) has been solved in part (i).

Therefore, based on the results of (i) and (ii), it follows that

$$\begin{aligned}
\sup_{\|x\|\leq J} d(\hat{r}_\oplus(x), m_\oplus(x)) \leq & \sup_{\|x\|\leq J} (d(\hat{r}_\oplus(x), \tilde{r}_\oplus(x)) + d(\tilde{r}_\oplus(x), m_\oplus(x))) \\
\leq & \sup_{\|x\|\leq J} d(\hat{r}_\oplus(x), \tilde{r}_\oplus(x)) + \sup_{\|x\|\leq J} d(\tilde{r}_\oplus(x), m_\oplus(x)) \\
= & o_p(1).
\end{aligned}$$

■

E.3 Proofs of Results in Section 3.3

Proof [Proof of Lemma 4] Since $X \in [0,1]^p$ with a density ρ_X bounded away from 0 and infinity. By Lemma 2 of Wager and Athey (2018), for any $0 < \eta < 1$, and for large enough s_n,

$$P\left\{\text{diam}_j(L(x)) \geq \left(\frac{s_n}{2k-1}\right)^{-\frac{0.99(1-\eta)\log\left((1-\alpha)^{-1}\right)}{\log(\alpha^{-1})}\frac{\pi}{p}}\right\} \leq \left(\frac{s_n}{2k-1}\right)^{-\frac{\eta^2}{2}\frac{1}{\log(\alpha^{-1})}\frac{\pi}{p}}. \tag{31}$$

where $\text{diam}_j(L(x))$ denote the length of the jth dimension of the leaf $L(x)$. If the honest tree is implemented by the double-sample tree, the above argument still holds by simply replacing s_n with $s_n/2$. But it does not affect the final result.

The above bound can derive $\text{diam}\,(L(x)) \to 0$ in probability. Then by Lemma 16, we have $d(\tilde{r}_\oplus(x), m_\oplus(x)) = o(1)$.

Since the Fréchet trees are honest,

$$\begin{aligned}\tilde{R}_n(x,y) - M_\oplus(x,y) =& E\left[E\left\{d^2(Y,y) \mid X \in L(x)\right\}\right] - E\left\{d^2(Y,y) \mid X = x\right\} \\ =& E\left[E\left\{d^2(Y,y) \mid X \in L(x)\right\} - E\left\{d^2(Y,y) \mid X = x\right\}\right].\end{aligned}$$

By the assumption (A5),

$$\left|E\left\{d^2(Y,y) \mid X \in L(x)\right\} - E\left\{d^2(Y,y) \mid X = x\right\}\right| \leq K\,\text{diam}(L(x)).$$

Now take the same approach as the proof of Theorem 3 of Wager and Athey (2018) to bound the diameter of $L(x)$. By plugging in $\eta = \sqrt{\log\left((1-\alpha)^{-1}\right)}$ in the bound from (31). Since $\alpha \leq 0.2$, we see that $\eta \leq 0.48$ and so $0.99 \cdot (1-\eta) \geq 0.51$; thus, a union bound gives us that, for large enough s_n,

$$P\left\{\text{diam}(L(x)) \geq p^{1/2}\left(\frac{s_n}{2k-1}\right)^{-0.51\frac{\log\left((1-\alpha)^{-1}\right)}{\log(\alpha^{-1})}\frac{\pi}{p}}\right\} \leq p\left(\frac{s_n}{2k-1}\right)^{-\frac{1}{2}\frac{\log\left((1-\alpha)^{-1}\right)}{\log(\alpha^{-1})}\frac{\pi}{p}}.$$

The Lipschitz assumption lets us bound $\tilde{R}_n(x,y) - M_\oplus(x,y)$ base on the above result about $\text{diam}(L(x))$. Specifically, let

$$a_1 = p^{1/2}\left(\frac{s_n}{2k-1}\right)^{-0.51\frac{\log\left((1-\alpha)^{-1}\right)}{\log(\alpha^{-1})}\frac{\pi}{p}}, \quad a_2 = p\left(\frac{s_n}{2k-1}\right)^{-\frac{1}{2}\frac{\log\left((1-\alpha)^{-1}\right)}{\log(\alpha^{-1})}\frac{\pi}{p}}.$$

Then

$$P\left\{\text{diam}(L(x)) \geq a_1\right\} \leq a_2.$$

We have

$$\begin{aligned}&\left|\tilde{R}_n(x,y) - M_\oplus(x,y)\right| \\ \leq& \left|E\left[E\left\{d^2(Y,y) \mid X \in L(x)\right\} - E\left\{d^2(Y,y) \mid X = x\right\}\right]\right| \\ \leq& \left|E\left(\left[E\left\{d^2(Y,y) \mid X \in L(x)\right\} - E\left\{d^2(Y,y) \mid X = x\right\}\right] I(\text{diam}(L(x)) \geq a_1)\right)\right|\end{aligned}$$

$$
\begin{aligned}
&+\left|E\left(\left[E\left\{d^2(Y,y)\mid X\in L(x)\right\}-E\left\{d^2(Y,y)\mid X=x\right\}\right]I(\mathrm{diam}(L(x))<a_1)\right)\right| \\
\leq& E\left[\left|E\left\{d^2(Y,y)\mid X\in L(x)\right\}-E\left\{d^2(Y,y)\mid X=x\right\}\right|I(\mathrm{diam}(L(x))\geq a_1)\right] \\
&+E\left[\left|E\left\{d^2(Y,y)\mid X\in L(x)\right\}-E\left\{d^2(Y,y)\mid X=x\right\}\right|I(\mathrm{diam}(L(x))<a_1)\right] \\
\leq& \left(\sup_{x\in[0,1]^p}\left[E\left\{d^2(Y,y)\mid X=x\right\}\right]-\inf_{x\in[0,1]^p}\left[E\left\{d^2(Y,y)\mid X=x\right\}\right]\right)P\left\{\mathrm{diam}(L(x))\geq a_1\right\}+Ka_1 \\
\leq& \left(\sup_{x\in[0,1]^p}\left[E\left\{d^2(Y,y)\mid X=x\right\}\right]-\inf_{x\in[0,1]^p}\left[E\left\{d^2(Y,y)\mid X=x\right\}\right]\right)a_2+Ka_1 \\
\lesssim& \left(\sup_{x\in[0,1]^p}\left[E\left\{d^2(Y,y)\mid X=x\right\}\right]-\inf_{x\in[0,1]^p}\left[E\left\{d^2(Y,y)\mid X=x\right\}\right]\right)a_2,
\end{aligned}
$$

since $a_1/a_2\to 0$.
Due to the Lipschitz condition,

$$
\begin{aligned}
\sup_{x\in[0,1]^p}\left[E\left\{d^2(Y,y)\mid X=x\right\}\right]-\inf_{x\in[0,1]^p}\left[E\left\{d^2(Y,y)\mid X=x\right\}\right]&\leq K\sup_{x_1,x_2\in[0,1]^p}\|x_1-x_2\| \\
&=Kp^{1/2}.
\end{aligned}
$$

So for large enough s_n,

$$
\left|\tilde{R}_n(x,y)-M_\oplus(x,y)\right|=O\left(s_n^{-\frac{1}{2}\frac{\log(1-\alpha)}{\log(\alpha)}\frac{\pi}{p}}\right).
$$

The above bound is uniform over $y\in\Omega$. Let $T_n(x,y)=\tilde{R}_n(x,y)-M_\oplus(x,y)$, then easily we can get, for any $\delta>0$,

$$
\sup_{d(y,m_\oplus(x))<\delta}\left|T_n(x,y)-T_n\left(x,m_\oplus(x)\right)\right|\lesssim c_1\delta s_n^{-\frac{1}{2}\frac{\log(1-\alpha)}{\log(\alpha)}\frac{\pi}{p}} \tag{32}
$$

for some constant $c_1>0$.

Now, set $t_n=s_n^{\frac{1}{4}\frac{\log(1-\alpha)}{\log(\alpha)}\frac{\pi}{p}\frac{\beta_1}{\beta_1-1}}$ and define

$$
S_{j,n}=\left\{y:2^{j-1}<t_n d\left(y,m_\oplus(x)\right)^{\beta_1/2}\leq 2^j\right\}.
$$

Choose δ_1 satisfying the assumption (A6). Set $\tilde{\delta}_1:=(\delta_1)^{\beta_1/2}$. For any integer M,

$$
\begin{aligned}
&I\left\{t_n d\left(\tilde{r}_\oplus(x),m_\oplus(x)\right)^{\beta_1/2}>2^M\right\} \\
=&\sum_{\substack{j>M\\ 2^j<t_n\tilde{\delta}_1}} I\left\{2^{j-1}<t_n d\left(\tilde{r}_\oplus(x),m_\oplus(x)\right)^{\beta_1/2}\leq 2^j\right\} \\
&\qquad+\sum_{\substack{j>M\\ 2^j\geq t_n\tilde{\delta}_1}} I\left\{2^{j-1}<t_n d\left(\tilde{r}_\oplus(x),m_\oplus(x)\right)^{\beta_1/2}\leq 2^j\right\} \qquad (33)\\
\leq&\sum_{\substack{j>M\\ 2^j<t_n\tilde{\delta}_1}} I\left\{2^{j-1}<t_n d\left(\tilde{r}_\oplus(x),m_\oplus(x)\right)^{\beta_1/2}\leq 2^j\right\}+I\left\{2d\left(\tilde{r}_\oplus(x),m_\oplus(x)\right)^{\beta_1/2}>\tilde{\delta}_1\right\}.
\end{aligned}
$$

By the definition of $S_{j,n}(x)$, we have

$$\begin{aligned} I\left\{2^{j-1} < t_n d\left(\tilde{r}_{\oplus}(x), m_{\oplus}(x)\right)^{\beta_1/2} \leq 2^j\right\} =& I\left\{\tilde{r}_{\oplus}(x) \in S_{j,n}(x)\right\} \\ \leq& I\left\{\inf_{y \in S_{j,n}(x)} \left(\tilde{R}_n(x,y) - \tilde{R}_n(x, m_{\oplus}(x))\right) \leq 0\right\}. \end{aligned} \tag{34}$$

In addition, notice that when $y \in S_{j,n}(x)$, $d(y, m_{\oplus}(x)) \leq (\frac{2^j}{t_n})^{\frac{2}{\beta_1}}$. If $2^j < t_n\tilde{\delta}_1$, we have $d(y, m_{\oplus}(x)) < \delta_1$. Then by the assumption (A6),

$$M_{\oplus}(x,y) - M_{\oplus}\left(x, m_{\oplus}(x)\right) \geq C_1 d\left(y, m_{\oplus}(x)\right)^{\beta_1} > C_1\left(\frac{2^{2(j-1)}}{t_n^2}\right).$$

Therefore, if $2^j < t_n\tilde{\delta}_1$,

$$\begin{aligned} & I\left\{\inf_{y \in S_{j,n}(x)} \left(\tilde{R}_n(x,y) - \tilde{R}_n(x, m_{\oplus}(x))\right) \leq 0\right\} \\ \leq & I\left\{\sup_{y \in S_{j,n}(x)} \left|T_n(x,y) - T_n\left(x, m_{\oplus}(x)\right)\right| \geq C_1 \frac{2^{2(j-1)}}{t_n^2}\right\}. \end{aligned} \tag{35}$$

Combine (33), (34) and (35), we get

$$\begin{aligned} I\left\{t_n d\left(\tilde{r}_{\oplus}(x), m_{\oplus}(x)\right)^{\beta_1/2} > 2^M\right\} \leq & I\left\{2d\left(\tilde{r}_{\oplus}(x), m_{\oplus}(x)\right)^{\beta_1/2} > \tilde{\delta}_1\right\} \\ & + \sum_{\substack{j>M \\ 2^j < t_n\tilde{\delta}_1}} I\left\{\sup_{y \in S_{j,n}(x)} \left|T_n(x,y) - T_n\left(x, m_{\oplus}(x)\right)\right| \geq C_1 \frac{2^{2(j-1)}}{t_n^2}\right\}, \end{aligned}$$

where the first term of the right-hand side goes to zero for any $\tilde{\delta}_1 > 0$ since $d(\tilde{r}_{\oplus}(x), m_{\oplus}(x)) = o(1)$. Now we focus on the second term. Obviously,

$$\begin{aligned} & \sum_{\substack{j>M \\ 2^j < t_n\tilde{\delta}_1}} I\left\{\sup_{y \in S_{j,n}(x)} \left|T_n(x,y) - T_n\left(x, m_{\oplus}(x)\right)\right| \geq C_1 \frac{2^{2(j-1)}}{t_n^2}\right\} \\ \leq & \sum_{\substack{j>M \\ 2^j < t_n\tilde{\delta}_1}} \frac{t_n^2}{C_1 2^{2(j-1)}} \sup_{y \in S_{j,n}} \left|T_n(x,y) - T_n\left(x, m_{\oplus}(x)\right)\right|. \end{aligned}$$

As mentioned above, for each j in the sum, we have $d(y, m_{\oplus}(x)) \leq (\frac{2^j}{t_n})^{\frac{2}{\beta_1}} < \delta_1$, then (32) holds with $\delta = (\frac{2^j}{t_n})^{\frac{2}{\beta_1}}$.Therefore, the sum is bounded by

$$4c_1 C_1^{-1} \sum_{\substack{j>M \\ 2^j < t_n\tilde{\delta}_1}} \frac{2^{2j(1-\beta_1)/\beta_1}}{t_n^{2(1-\beta_1)/\beta_1}} s_n^{-\frac{1}{2}\frac{\log(1-\alpha)}{\log(\alpha)}\frac{\pi}{p}} \leq 4c_1 C_1^{-1} \sum_{j>M} \left(\frac{1}{4^{(\beta_1-1)/\beta_1}}\right)^j,$$

which converges since $\beta_1 > 1$. Thus, for some $M > 0$, we have

$$d\left(\tilde{r}_{\oplus}(x), m_{\oplus}(x)\right) \leq 2^{2M/\beta_1} s_n^{-\frac{1}{2}\frac{\log(1-\alpha)}{\log(\alpha)}\frac{\pi}{p}\frac{1}{\beta_1-1}}$$

for large enough n. ■

Before proving the convergence rate of the variance term, we make some notes on the related literature regarding infinite order U-processes. From (17) and (19), let

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

$$\tilde{R}_n(x,y) = E\left\{ \frac{1}{N(L(x;\mathcal{D}_n^k,\xi))} \sum_{i:X_i\in L(x;\mathcal{D}_n^k,\xi)} d^2\left(Y_i,y\right) \right\}.$$

By Theorem 3.2.5 of van der Vaart and Wellner (1996), the key to determining the convergence rate is to establish the upper bound for the expectation of the local maximum value of the centered process $\left\{(\hat{R}_n-\tilde{R}_n)(x,y)-(\hat{R}_n-\tilde{R}_n)(x,\tilde{r}_\oplus(x))\right\}$. Recall that $\hat{R}_n$ is an infinite order U-statistic for any fixed $y\in\Omega$ if Fréchet trees are symmetric. To figure out such an upper bound, 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. Sherman (1994) established the maximal inequality for degenerate U-processes of arbitrary order based on moment inequalities. Arcones and Giné (1993) obtained a similar maximal inequality by exponential inequalities, which is stronger than that of Sherman (1994). However, both results are limited to U-processes of fixed order. Heilig and Nolan (2001) first extended the maximal inequality of Sherman (1994) to infinite order U-processes by the complete sign-symmetrization technique. Unfortunately, they are unable to extend the maximal inequality of Arcones and Giné (1993) to the infinite order case because such results rely on a symmetrization inequality of de la Pena (1992) which incurs upper bounds that grow with the order much too quickly. Chen and Kato (2020) gave a local maximal inequality for degenerate infinite order U-processes in the form of uniform entropy integrals. Heilig (1997) 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.

Proof [Proof of Lemma 6] We consider the expressions of (17) and (19). For the sake of simplicity in notation, let

$$h_n\left(Z_{i_{k,1}},\ldots,Z_{i_{k,s_n}},y\right) = E_{\xi\sim\Xi}\left\{ \frac{1}{N(L(x;\mathcal{D}_n^k,\xi))} \sum_{i:X_i\in L(x;\mathcal{D}_n^k,\xi)} d^2\left(Y_i,y\right) \right\}$$

where $Z_i=(X_i,Y_i)$ and $\mathcal{D}_n^k=\left(Z_{i_{k,1}},\ldots,Z_{i_{k,s_n}}\right)$.
Then

$$\hat{r}_\oplus(x) = \underset{y\in\Omega}{\operatorname{argmin}}\hat{R}_n(x,y) = \underset{y\in\Omega}{\operatorname{argmin}}\binom{n}{s_n}^{-1} \sum_{1\le i_1<\ldots<i_{s_n}\le n} h_n\left(Z_{i_1},\ldots,Z_{i_{s_n}},y\right);$$

$$\tilde{r}_\oplus(x) = \underset{y\in\Omega}{\operatorname{argmin}}\tilde{R}_n(x,y) = \underset{y\in\Omega}{\operatorname{argmin}} E h_n\left(Z_1,\ldots,Z_{s_n},y\right).$$

Consider the centered process $V_n(x,y)=\hat{R}_n(x,y)-\tilde{R}_n(x,y)$, and we have

$$\left|V_n(x,y)-V_n\left(x,\tilde{r}_\oplus(x)\right)\right|$$

$$
\begin{aligned}
=&\left| \binom{n}{s_n}^{-1} \sum_{1\le i_1<\ldots<i_{s_n}\le n} h_n\left(Z_{i_1},\ldots,Z_{i_{s_n}},y\right) - Eh_n\left(Z_1,\ldots,Z_{s_n},y\right) \right.\\
&\left. - \binom{n}{s_n}^{-1} \sum_{1\le i_1<\ldots<i_{s_n}\le n} h_n\left(Z_{i_1},\ldots,Z_{i_{s_n}},\tilde{r}_{\oplus}(x)\right) + Eh_n\left(Z_1,\ldots,Z_{s_n},\tilde{r}_{\oplus}(x)\right) \right|\\
=&\left| \binom{n}{s_n}^{-1} \sum_{1\le i_1<\ldots<i_{s_n}\le n} \left\{h_n\left(Z_{i_1},\ldots,Z_{i_{s_n}},y\right) - h_n\left(Z_{i_1},\ldots,Z_{i_{s_n}},\tilde{r}_{\oplus}(x)\right)\right\} \right.\\
&\left. - E\left\{h_n\left(Z_1,\ldots,Z_{s_n},y\right) - h_n\left(Z_1,\ldots,Z_{s_n},\tilde{r}_{\oplus}(x)\right)\right\} \right|.
\end{aligned}
$$

Let

$$
H_n(Z_{i_1},\ldots,Z_{i_{s_n}},y) = h_n\left(Z_{i_1},\ldots,Z_{i_{s_n}},y\right) - h_n\left(Z_{i_1},\ldots,Z_{i_{s_n}},\tilde{r}_{\oplus}(x)\right).
$$

Then

$$
\begin{aligned}
&\left|V_n(x,y) - V_n\left(x,\tilde{r}_{\oplus}(x)\right)\right|\\
=&\left| \binom{n}{s_n}^{-1} \sum_{1\le i_1<\ldots<i_{s_n}\le n} H_n(Z_{i_1},\ldots,Z_{i_{s_n}},y) - E\left\{H_n(Z_1,\ldots,Z_{s_n},y)\right\} \right|.
\end{aligned}
$$

Next, to control the $\left|V_n(x,y) - V_n\left(x,\tilde{r}_{\oplus}(x)\right)\right|$ uniformly over small $d\left(y,\tilde{r}_{\oplus}(x)\right)$, we consider the function class

$$
\mathcal{H}_\delta = \left\{H_n(z_1,\ldots,z_{s_n},y) : d\left(y,\tilde{r}_{\oplus}(x)\right) < \delta\right\}.
$$

It's not hard to find $\left|V_n(x,y) - V_n\left(x,\tilde{r}_{\oplus}(x)\right)\right|$ is a centered infinite order U-process with index set $\mathcal{H}_\delta$.

Since

$$
\begin{aligned}
&\left|H_n(Z_{i_{k,1}},\ldots,Z_{i_{k,s_n}},y)\right|\\
=&\left| E_{\xi\sim\Xi}\left[\frac{1}{N(L(x;\mathcal{D}_n^k,\xi))} \sum_{i:X_i\in L(x;\mathcal{D}_n^k,\xi)} \left\{d^2\left(Y_i,y\right) - d^2\left(Y_i,\tilde{r}_{\oplus}(x)\right)\right\}\right]\right|\\
\le& E_{\xi\sim\Xi}\left\{\frac{1}{N(L(x;\mathcal{D}_n^k,\xi))} \sum_{i:X_i\in L(x;\mathcal{D}_n^k,\xi)} \left|d\left(Y_i,y\right) - d\left(Y_i,\tilde{r}_{\oplus}(x)\right)\right| \left|d\left(Y_i,y\right) + d\left(Y_i,\tilde{r}_{\oplus}(x)\right)\right|\right\}\\
\le& 2\operatorname{diam}(\Omega) d(y,\tilde{r}_{\oplus}(x)).
\end{aligned}
$$

An envelope function for $\mathcal{H}_\delta$ is $G_\delta(z_1,\ldots,z_{s_n}) = 2\operatorname{diam}(\Omega)\delta$, and the proof of Theorem 4.7 of Heilig (1997) gives that, for small enough δ and any $\epsilon\in(0,1]$,

$$
\begin{aligned}
&E\left\{\sup_{d(y,\tilde{r}_{\oplus}(x))<\delta} \left|V_n(x,y) - V_n\left(x,\tilde{r}_{\oplus}(x)\right)\right|\right\}\\
\le& 2K_1\varepsilon + K_2 G_\delta \sum_{j=1}^{s_n} E\left\{n^{-1}\log N\left(\varepsilon/s_n, d_j, \mathcal{H}_\delta\right)\right\}^{1/2}
\end{aligned}
$$

where K_1, K_2 are absolute constants.
By the assumption (A7), for small δ,

$$\sum_{j=1}^{s_n} E\left\{\log N\left(\varepsilon/s_n, d_j, \mathcal{H}_\delta\right)\right\}^{1/2} \leq s_n(V\log s_n + \log A - V\log\varepsilon)^{1/2}.$$

Then

$$E\left\{\sup_{d(y,\tilde{r}_\oplus(x))<\delta} |V_n(x,y) - V_n\left(x,\tilde{r}_\oplus(x)\right)|\right\} \leq c_2\delta\frac{s_n(logs_n)^{\frac{1}{2}}}{n^{1/2}} \tag{36}$$

for some constant $c_2 > 0$.

Now, set $r_n = (\frac{n}{s_n^2 logs_n})^{\frac{\beta_2}{4(\beta_2-1)}}$ and define

$$S_{j,n}(x) = \left\{y : 2^{j-1} < r_n d\left(y,\tilde{r}_\oplus(x)\right)^{\beta_2/2} \leq 2^j\right\}.$$

Choose δ_2 satisfying the assumption (A8) and such that the assumption (A7) is satisfied for any $\delta < \delta_2$. Set $\tilde{\delta}_2 := (\delta_2)^{\beta_2/2}$. For any integer M,

$$\begin{aligned}
&P\left\{r_n d\left(\hat{r}_\oplus(x),\tilde{r}_\oplus(x)\right)^{\beta_2/2} > 2^M\right\} \\
=&\sum_{\substack{j>M\\ 2^j<r_n\tilde{\delta}_2}} P\left\{2^{j-1} < r_n d\left(\hat{r}_\oplus(x),\tilde{r}_\oplus(x)\right)^{\beta_2/2} \leq 2^j\right\} \\
&\qquad + \sum_{\substack{j>M\\ 2^j\geq r_n\tilde{\delta}_2}} P\left\{2^{j-1} < r_n d\left(\hat{r}_\oplus(x),\tilde{r}_\oplus(x)\right)^{\beta_2/2} \leq 2^j\right\} \\
\leq&\sum_{\substack{j>M\\ 2^j<r_n\tilde{\delta}_2}} P\left\{2^{j-1} < r_n d\left(\hat{r}_\oplus(x),\tilde{r}_\oplus(x)\right)^{\beta_2/2} \leq 2^j\right\} + P\left\{2d\left(\hat{r}_\oplus(x),\tilde{r}_\oplus(x)\right)^{\beta_2/2} > \tilde{\delta}_2\right\}.
\end{aligned} \tag{37}$$

By the definition of $S_{j,n}(x)$, we have

$$\begin{aligned}
P\left\{2^{j-1} < r_n d\left(\hat{r}_\oplus(x),\tilde{r}_\oplus(x)\right)^{\beta_2/2} \leq 2^j\right\} =& P\left\{\hat{r}_\oplus(x) \in S_{j,n}(x)\right\} \\
\leq& P\left\{\inf_{y\in S_{j,n}(x)}\left(\hat{R}_n(x,y) - \hat{R}_n(x,\tilde{r}_\oplus(x))\right) \leq 0\right\}.
\end{aligned} \tag{38}$$

In addition, notice that when $y \in S_{j,n}(x)$, $d(y,\tilde{r}_\oplus(x)) \leq (\frac{2^j}{r_n})^{\frac{2}{\beta_2}}$. If $2^j < r_n\tilde{\delta}_2$, we have $d(y,\tilde{r}_\oplus(x)) < \delta_2$. Then by the assumption (A8), for large enough n,

$$\tilde{R}_n(x,y) - \tilde{R}_n\left(x,\tilde{r}_\oplus(x)\right) \geq C_2 d\left(y,\tilde{r}_\oplus(x)\right)^{\beta_2} > C_2\left(\frac{2^{2(j-1)}}{r_n^2}\right).$$

Therefore, if $2^j < r_n\tilde{\delta}_2$,

$$\begin{aligned}
&P\left\{\inf_{y\in S_{j,n}(x)}\left\{\hat{R}_n(x,y) - \hat{R}_n(x,\tilde{r}_\oplus(x))\right\} \leq 0\right\} \\
\leq& P\left\{\sup_{y\in S_{j,n}(x)} |V_n(x,y) - V_n\left(x,\tilde{r}_\oplus(x)\right)| \geq C_2\frac{2^{2(j-1)}}{r_n^2}\right\}.
\end{aligned} \tag{39}$$

Combine (37), (38) and (39), we get

$$P\left\{r_n d\left(\hat{r}_\oplus(x), \tilde{r}_\oplus(x)\right)^{\beta_2/2} > 2^M\right\} \leq P\left\{2d\left(\tilde{r}_\oplus(x), \hat{r}_\oplus(x)\right)^{\beta_2/2} > \tilde{\delta}_2\right\} + \sum_{\substack{j>M \\ 2^j < r_n\tilde{\delta}_2}} P\left\{\sup_{y\in S_{j,n}(x)} |V_n(x,y) - V_n\left(x, \tilde{r}_\oplus(x)\right)| \geq C_2 \frac{2^{2(j-1)}}{r_n^2}\right\}.$$

where the first term of the right-hand side goes to zero for any $\tilde{\delta}_2 > 0$ by Lemma 17. Now we focus on the second term. By Markov's inequality,

$$\sum_{\substack{j>M \\ 2^j < r_n\tilde{\delta}_2}} P\left\{\sup_{y\in S_{j,n}(x)} |V_n(x,y) - V_n\left(x, \tilde{r}_\oplus(x)\right)| \geq C_2 \frac{2^{2(j-1)}}{r_n^2}\right\}$$
$$\leq \sum_{\substack{j>M \\ 2^j < r_n\tilde{\delta}_2}} \frac{r_n^2}{C_2 2^{2(j-1)}} E\left\{\sup_{y\in S_{j,n}(x)} |V_n(x,y) - V_n\left(x, \tilde{r}_\oplus(x)\right)|\right\}.$$

As mentioned above, for each j in the sum, we have $d(y, \tilde{r}_\oplus(x)) \leq (\frac{2^j}{r_n})^{\frac{2}{\beta_2}} < \delta_2$, then (36) holds with $\delta = (\frac{2^j}{r_n})^{\frac{2}{\beta_2}}$. Therefore, the sum is bounded by

$$4c_2 C_2^{-1} \sum_{\substack{j>M \\ 2^j < r_n\tilde{\delta}_2}} \frac{2^{2j(1-\beta_2)/\beta_2}}{r_n^{2(1-\beta_2)/\beta_2}} \frac{s_n (logs_n)^{\frac{1}{2}}}{n^{1/2}} < 4c_2 C_2^{-1} \sum_{j>M} \left(\frac{1}{4^{(\beta_2-1)/\beta_2}}\right)^j.$$

Because $\beta_2 > 1$, the last series converges, and hence this probability can be made small by choosing M large. Hence

$$d\left(\hat{r}_\oplus(x), \tilde{r}_\oplus(x)\right) = O_p\left(r_n^{-2/\beta_2}\right) = O_p\left(\left(\frac{s_n^2 logs_n}{n}\right)^{\frac{1}{2(\beta_2-1)}}\right).$$

■

E.4 Proofs of Results in Section 4

Proof [Proof of Theorem 9] Let

$$\hat{L}_n(x,y) = e_1^{\mathrm{T}} (\tilde{X}^{\mathrm{T}} A \tilde{X})^{-1} \sum_{i=1}^n \begin{pmatrix} 1 \\ X_i - x \end{pmatrix} \alpha_i(x) d^2(Y_i, y).$$

For a fix $x \in [0,1]^p$, by Corollary 3.2.3 of van der Vaart and Wellner (1996) and the assumption (A12), we only need to prove $\sup_{y\in\Omega} \left|\hat{L}_n(x,y) - M_\oplus(x,y)\right|$ to zero in probability. To do this, we show $\hat{L}_n(x,\cdot) \rightsquigarrow M_\oplus(x,\cdot)$ in $l^\infty(\Omega)$ and apply Theorem 1.3.6 of van der Vaart and Wellner (1996). By Theorem 1.5.4 of van der Vaart and Wellner (1996), this weak convergence is equivalent

to $\hat{L}_n(x,\cdot)$ is asymptotically tight and the marginals converge weakly. By Theorem 1.5.7 of van der Vaart and Wellner (1996), This asymptotically tight is equivalent to $\hat{L}_n(x,y)$ is asymptotically tight in $\mathcal{R}$ for every $y \in \Omega$ and $\hat{L}_n(x,\cdot)$ is asymptotically uniformly d-equicontinuous in probability. So the proof will be finished if the following conditions hold.

(i) $\hat{L}_n(x,y) - M_{\oplus}(x,y) = o_p(1)$ for each $y \in \Omega$,

(ii) For all $\varepsilon, \eta > 0$, there exists $\delta > 0$ such that

$$\limsup_n P\left\{ \sup_{d(y_1,y_2)<\delta} \left| \hat{L}_n(x,y_1) - \hat{L}_n(x,y_2) \right| > \varepsilon \right\} < \eta.$$

First, prove (i): By the fact about local linear estimators,

$$e_1^{\mathrm{T}}(\tilde{X}^{\mathrm{T}}A\tilde{X})^{-1}\left[\sum_{i=1}^n \begin{pmatrix} 1 \\ X_i - x \end{pmatrix} \alpha_i(x) E\left\{d^2(Y,y) \mid X = x\right\}\right] = E\left\{d^2(Y,y) \mid X = x\right\}.$$

So we have

$$\begin{aligned} &\hat{L}_n(x,y) - M_{\oplus}(x,y) \\ =&e_1^{\mathrm{T}}(\tilde{X}^{\mathrm{T}}A\tilde{X})^{-1}\left(\sum_{i=1}^n \begin{pmatrix} 1 \\ X_i - x \end{pmatrix} \alpha_i(x)\left[d^2(Y_i,y) - E\left\{d^2(Y,y) \mid X = x\right\}\right]\right). \end{aligned}$$

Define

$$\bar{L}_n(x,y) = e_1^{\mathrm{T}}(\tilde{X}^{\mathrm{T}}A\tilde{X})^{-1}\left[\sum_{i=1}^n \begin{pmatrix} 1 \\ X_i - x \end{pmatrix} \alpha_i(x) E\left\{d^2(Y,y) \mid X = X_i\right\}\right].$$

Then consider to decompose $\hat{L}_n(x,y) - M_{\oplus}(x,y)$ into a variance-type term

$$\begin{aligned} &\hat{L}_n(x,y) - \bar{L}_n(x,y) \\ =&e_1^{\mathrm{T}}(\tilde{X}^{\mathrm{T}}A\tilde{X})^{-1}\left(\sum_{i=1}^n \begin{pmatrix} 1 \\ X_i - x \end{pmatrix} \alpha_i(x)\left[d^2(Y_i,y) - E\left\{d^2(Y,y) \mid X = X_i\right\}\right]\right) \end{aligned}$$

and a bias-type term

$$\begin{aligned} &\bar{L}_n(x,y) - M_{\oplus}(x,y) \\ =&e_1^{\mathrm{T}}(\tilde{X}^{\mathrm{T}}A\tilde{X})^{-1}\left(\sum_{i=1}^n \begin{pmatrix} 1 \\ X_i - x \end{pmatrix} \alpha_i(x)\left[E\left\{d^2(Y,y) \mid X = X_i\right\} - E\left\{d^2(Y,y) \mid X = x\right\}\right]\right). \end{aligned}$$

Next, we will follow the proof of Theorem 1 in Bloniarz et al. (2016) with a few modifications to show that each of these terms converges to zero in probability for each $y \in \Omega$.

So we begin to prove $\hat{L}_n(x,y) - \bar{L}_n(x,y) \xrightarrow{p} 0$. For convenience in notation, let $\alpha_i(x; \mathcal{D}_n^b, \xi_b) = 1\left\{X_i \in L_b(x; \mathcal{D}_n^b, \xi_b)\right\}$, indicating that a training point X_i belongs to the same leaf node as x in the bth Fréchet tree trained with subsample $\mathcal{D}_n^b$ and random parameter ξ_b. We define the bandwidth matrix of the random forest to be a diagonal matrix with diagonal elements set to be the

largest component-wise distances from x to a training point that has a nonzero weight. Let $X_i = \left(X_i^{(1)}, \ldots, X_i^{(p)}\right)$ and $x = \left(x^{(1)}, \ldots, x^{(p)}\right)$. Then define

$$h_j = \max_{i,b} \left\{ \alpha_i(x; \mathcal{D}_n^b, \xi_b) \left| X_i^{(j)} - x^{(j)} \right| \right\}, \quad H = \operatorname{diag}(1, h_1, \ldots, h_p).$$

By the assumption (A10), the number of training points falling in a leaf node goes to infinity. Under the honesty condition, if we condition on the variables $\alpha_i(x; \mathcal{D}_n^b, \xi_b)$, the subset of the training data falling in $L_b(x; \mathcal{D}_n^b, \xi_b)$ is independent and identically distributed in the rectangle $L_b(x; \mathcal{D}_n^b, \xi_b)$. By definition of H, for the training point that has nonzero weight, we have

$$\left\| H^{-1} \begin{pmatrix} 1 \\ X_i - x \end{pmatrix} \right\|_\infty \leq 1.$$

And notice that $E\left[d^2(Y_i, y) - E\left\{d^2(Y, y) \mid X = X_i\right\}\right] = 0$. Combining the previous discussion, with the weak law of large numbers we can obtain

$$\frac{1}{N(L_b(x; \mathcal{D}_n^b, \xi_b))} \sum_{i=1}^n H^{-1} \begin{pmatrix} 1 \\ X_i - x \end{pmatrix} \alpha_i(x; \mathcal{D}_n^b, \xi_b) \left[d^2(Y_i, y) - E\left\{d^2(Y, y) \mid X = X_i\right\}\right] \xrightarrow{p} 0.$$

Therefore, averaging over total B Fréchet trees, we get

$$\sum_{i=1}^n H^{-1} \begin{pmatrix} 1 \\ X_i - x \end{pmatrix} \alpha_i(x) \left[d^2(Y_i, y) - E\left\{d^2(Y, y) \mid X = X_i\right\}\right] \xrightarrow{p} 0. \tag{40}$$

Below we need to study the matrix $\tilde{X}^{\mathrm{T}} A \tilde{X}$. Since A is a diagonal matrix,

$$\begin{aligned} \tilde{X}^{\mathrm{T}} A \tilde{X} &= \sum_{i=1}^n \alpha_i(x) \begin{pmatrix} 1 \\ X_i - x \end{pmatrix} \left(1, (X_i - x)^{\mathrm{T}}\right) \\ &= \frac{1}{B} \sum_{b=1}^B \left[\frac{1}{N(L_b(x; \mathcal{D}_n^b, \xi_b))} \sum_{i=1}^n \alpha_i(x; \mathcal{D}_n^b, \xi_b) \begin{pmatrix} 1 \\ X_i - x \end{pmatrix} \left(1, (X_i - x)^{\mathrm{T}}\right) \right]. \end{aligned}$$

By the same argument of Theorem 1 in Bloniarz et al. (2016), we can have

$$e_1^{\mathrm{T}} (\tilde{X}^{\mathrm{T}} A \tilde{X})^{-1} H = (O_p(1), \ldots, O_p(1)). \tag{41}$$

which needs the assumption (A9), (A10) and the honesty condition. Combining (40) and (41), then

$$\hat{L}_n(x, y) - \bar{L}_n(x, y) \xrightarrow{p} 0. \tag{42}$$

Now we turn to prove $\bar{L}_n(x, y) - M_\oplus(x, y) \xrightarrow{p} 0$. Similarly, we have

$$\begin{aligned} \bar{L}_n(x, y) - M_\oplus(x, y) =& e_1^{\mathrm{T}} (\tilde{X}^{\mathrm{T}} A \tilde{X})^{-1} H \Bigg(\sum_{i=1}^n H^{-1} \begin{pmatrix} 1 \\ X_i - x \end{pmatrix} \alpha_i(x) \\ & \left[E\left\{d^2(Y, y) \mid X = X_i\right\} - E\left\{d^2(Y, y) \mid X = x\right\}\right] \Bigg). \end{aligned}$$

By the assumption (A11) and the definition of H, we can have

$$\sum_{i=1}^{n} H^{-1} \begin{pmatrix} 1 \\ X_i - x \end{pmatrix} \alpha_i(x) \left[E\left\{ d^2(Y, y) \mid X = X_i \right\} - E\left\{ d^2(Y, y) \mid X = x \right\} \right] \xrightarrow{p} 0.$$

Since $e_1^{\mathrm{T}}(\tilde{X}^{\mathrm{T}} A \tilde{X})^{-1} H = (O_p(1), \ldots, O_p(1))$, we get

$$\bar{L}_n(x, y) - M_{\oplus}(x, y) \xrightarrow{p} 0. \tag{43}$$

Hence combine (42) and (43), it follows that for each $y \in \Omega$,

$$\hat{L}_n(x, y) - M_{\oplus}(x, y) = o_p(1).$$

Then (ii): For any y_1, $y_2 \in \Omega$,

$$\begin{aligned}
&\left| \hat{L}_n\left(x, y_1\right) - \hat{L}_n\left(x, y_2\right) \right| \\
=&\left| e_1^{\mathrm{T}}(\tilde{X}^{\mathrm{T}} A \tilde{X})^{-1} \left[\sum_{i=1}^{n} \begin{pmatrix} 1 \\ X_i - x \end{pmatrix} \alpha_i(x) \left\{ d^2(Y_i, y_1) - d^2(Y_i, y_2) \right\} \right] \right| \\
\leq& \sum_{i=1}^{n} \left| e_1^{\mathrm{T}}(\tilde{X}^{\mathrm{T}} A \tilde{X})^{-1} \begin{pmatrix} 1 \\ X_i - x \end{pmatrix} \alpha_i(x) \right| \left| d\left(Y_i, y_1\right) - d\left(Y_i, y_2\right) \right| \left| d\left(Y_i, y_1\right) + d\left(Y_i, y_2\right) \right| \\
\leq& 2 \operatorname{diam}(\Omega) d\left(y_1, y_2\right) \sum_{i=1}^{n} \left| e_1^{\mathrm{T}}(\tilde{X}^{\mathrm{T}} A \tilde{X})^{-1} \begin{pmatrix} 1 \\ X_i - x \end{pmatrix} \alpha_i(x) \right| \\
=& 2 \operatorname{diam}(\Omega) d\left(y_1, y_2\right) \sum_{i=1}^{n} \left| e_1^{\mathrm{T}}(\tilde{X}^{\mathrm{T}} A \tilde{X})^{-1} H H^{-1} \begin{pmatrix} 1 \\ X_i - x \end{pmatrix} \alpha_i(x) \right| \\
\leq& 2 \operatorname{diam}(\Omega) d\left(y_1, y_2\right) \sum_{i=1}^{n} \alpha_i(x) \left\| e_1^{\mathrm{T}}(\tilde{X}^{\mathrm{T}} A \tilde{X})^{-1} H \right\| \left\| H^{-1} \begin{pmatrix} 1 \\ X_i - x \end{pmatrix} \right\| \\
=& 2 \operatorname{diam}(\Omega) d\left(y_1, y_2\right) O_p(1) \\
=& O_p\left(d\left(y_1, y_2\right)\right),
\end{aligned}$$

where the O_p term is independent of y_1 and y_2. Hence

$$\sup_{d(y_1, y_2) < \delta} \left| \hat{L}_n\left(x, y_1\right) - \hat{L}_n\left(x, y_2\right) \right| = O_p(\delta),$$

which can deduce (ii). So, $d(\hat{l}_{\oplus}(x), m_{\oplus}(x)) = o_p(1)$.

■

E.5 Proofs of Results in Appendix B

Proof [Proof of Theorem 13] Now we generalize the proof of Theorem 2.3 of Bose and Chatterjee (2018). The norm involved in this part is the spectral norm for matrices and the Euclidean norm for vectors.

Since $f_n(x,\theta)$ is convex in θ, for all α,β,

$$f_n(x,\alpha)+(\beta-\alpha)^T g_n(x,\alpha)\le f_n(x,\beta),$$
$$f_n(x,\beta)+(\alpha-\beta)^T g_n(x,\beta)\le f_n(x,\alpha).$$

Hence,

$$(\beta-\alpha)^T g_n(x,\alpha)\le f_n(x,\beta)-f_n(x,\alpha)\le(\beta-\alpha)^T g_n(x,\beta). \tag{44}$$

Then

$$0\le f_n(x,\beta)-f_n(x,\alpha)-(\beta-\alpha)^T g_n(x,\alpha)\le(\beta-\alpha)^T[g_n(x,\beta)-g_n(x,\alpha)]. \tag{45}$$

By the assumption (ii), we know $Eg_n(Z_1,Z_2,\ldots,Z_{m_n},\theta)<\infty$ from (44). Moreover, based on (45), note that $Eg_n(Z_1,Z_2,\ldots,Z_{m_n},\theta)$ serves as a subgradient of $Q_n(\theta)$. Now, when $Q_n(\theta)$ is differentiable, it follows that

$$\nabla Q_n(\theta_n)=Eg_n(Z_1,Z_2,\ldots,Z_{m_n},\theta_n)=0.$$

Let $W_n=\{w_n=(i_1,i_2,\ldots,i_{m_n}):1\le i_1<i_2\cdots<i_{m_n}\le n\}$. For any $w_n\in W_n$, let $Z_{w_n}=\left(Z_{i_1},\ldots,Z_{i_{m_n}}\right)$ and

$$Z_{n,w_n}=f_n\left(Z_{w_n},n^{-1/2}\Lambda_n^{\frac12}\alpha+\theta_n\right)-f_n(Z_{w_n},\theta_n)-n^{-1/2}\alpha^T\Lambda_n^{\frac12}g_n(Z_{w_n},\theta_n)$$

where $\Lambda_n=m_n^2H_n^{-1}K_nH_n^{-1}$. Note that $V_n=\binom{n}{m_n}^{-1}\sum_{w_n\in W_n}Z_{n,w_n}$ is an infinite order U-statistic. Using (45), it follows that

$$\begin{aligned}
\mathrm{Var}(V_n)&\le\frac{m_n}{n}\mathrm{Var}(Z_{n,w_n})\\
&\le\frac{m_n}{n}EZ_{n,w_n}^2\\
&\le\frac{m_n}{n^2}E\left\{\alpha^T\Lambda_n^{\frac12}\left[g_n(Z_{w_n},n^{-1/2}\Lambda_n^{\frac12}\alpha+\theta_n)-g_n(Z_{w_n},\theta_n)\right]\right\}^2\\
&\le\frac{m_n}{n^2}E\left[\|\alpha\|\|\Lambda_n^{\frac12}\|\|g_n(Z_{w_n},n^{-1/2}\Lambda_n^{\frac12}\alpha+\theta_n)-g_n(Z_{w_n},\theta_n)\|\right]^2\\
&=\frac{m_n\|\Lambda_n\|}{n^2}\|\alpha\|^2E\left[\|g_n(Z_{w_n},n^{-1/2}\Lambda_n^{\frac12}\alpha+\theta_n)-g_n(Z_{w_n},\theta_n)\|\right]^2.
\end{aligned}$$

By Taylor's expansion, we have

$$g_n(Z_{w_n},n^{-1/2}\Lambda_n^{\frac12}\alpha+\theta_n)=g_n(Z_{w_n},\theta_n)+\frac{\partial g_n(Z_{w_n},\tilde\theta_n)}{\partial\theta}n^{-1/2}\Lambda_n^{\frac12}\alpha,$$

where $\tilde\theta_n$ is between θ_n and $n^{-1/2}\Lambda_n^{\frac12}\alpha+\theta_n$. Then

$$\begin{aligned}
&E\left[\|g_n(Z_{w_n},n^{-1/2}\Lambda_n^{\frac12}\alpha+\theta_n)-g_n(Z_{w_n},\theta_n)\|\right]^2\\
=&E\left[\|\frac{\partial g_n(Z_{w_n},\tilde\theta_n)}{\partial\theta}n^{-1/2}\Lambda_n^{\frac12}\alpha\|\right]^2\\
\le&\frac1n\|\Lambda_n\|\|\alpha\|^2E\left[\|\frac{\partial g_n(Z_{w_n},\tilde\theta_n)}{\partial\theta}\|\right]^2.
\end{aligned}$$

Since $\lambda_{\min}(H_n) \nrightarrow 0$ by the assumption (v), $\|H_n^{-1}\| = \frac{1}{\lambda_{\min}(H_n)} \nrightarrow \infty$. And due to $m_n K_n \leq \mathrm{Var}\left(g_n\left(Z_1, \ldots, Z_{m_n}, \theta_n\right)\right) < \infty$, we have $\|m_n K_n\| < \infty$. Hence,

$$\frac{\|\Lambda_n\|}{n} = \frac{\|m_n^2 H_n^{-1} K_n H_n^{-1}\|}{n} \leq \frac{m_n \|H_n^{-1}\|^2 \|m_n K_n\|}{n} \to 0.$$

And we get

$$E\left[\|g_n(Z_{w_n}, n^{-1/2}\Lambda_n^{\frac{1}{2}}\alpha + \theta_n) - g_n(Z_{w_n}, \theta_n)\|\right]^2 \to 0.$$

By Markov's inequality, it follows that for each fixed α,

$$\frac{n}{\sqrt{m_n\|\Lambda_n\|}}\left(V_n - E(V_n)\right) \xrightarrow{P} 0.$$

Specifically,

$$\begin{aligned}
&\frac{n}{\sqrt{m_n\|\Lambda_n\|}}\binom{n}{m_n}^{-1}\sum_{w_n \in W_n}\left(Z_{n,w_n} - EZ_{n,w_n}\right) \\
=&\frac{n}{\sqrt{m_n\|\Lambda_n\|}}\hat{Q}_n\left(n^{-1/2}\Lambda_n^{\frac{1}{2}}\alpha + \theta_n\right) - \frac{n}{\sqrt{m_n\|\Lambda_n\|}}\hat{Q}_n(\theta_n) - \frac{\sqrt{n}}{\sqrt{m_n\|\Lambda_n\|}}\alpha^T\Lambda_n^{\frac{1}{2}}U_n \\
&- \frac{n}{\sqrt{m_n\|\Lambda_n\|}}Q_n\left(n^{-1/2}\Lambda_n^{\frac{1}{2}}\alpha + \theta_n\right) + \frac{n}{\sqrt{m_n\|\Lambda_n\|}}Q_n(\theta_n) \\
\xrightarrow{P}&0.
\end{aligned} \tag{46}$$

On the other hand, for each fixed α, by Taylor's expansion and the assumption (iv), (v),

$$\frac{n}{\sqrt{m_n\|\Lambda_n\|}}Q_n\left(n^{-1/2}\Lambda_n^{\frac{1}{2}}\alpha + \theta_n\right) \to \frac{n}{\sqrt{m_n\|\Lambda_n\|}}Q_n(\theta_n) + \frac{\alpha^T\Lambda_n^{\frac{1}{2}}H_n\Lambda_n^{\frac{1}{2}}\alpha}{2\sqrt{m_n\|\Lambda_n\|}}. \tag{47}$$

The reason why there are only two items on the right side of the above formula is

$$\frac{n\|\Lambda_n^{\frac{1}{2}}\|^3 n^{-\frac{3}{2}}}{\sqrt{m_n\|\Lambda_n\|}} = \frac{\|\Lambda_n\|}{\sqrt{m_n}\sqrt{n}} \leq \frac{m_n\|H_n^{-1}\|^2\|m_n K_n\|}{\sqrt{m_n}\sqrt{n}} \to 0.$$

By Lemma 2.1 of Bose and Chatterjee (2018), the convergences in (46) and (47) are uniform on compact sets due to convexity. Thus for every $\epsilon > 0$ and every $M > 0$,

$$\begin{aligned}
\sup_{\|\alpha\| \leq M}\Bigg|&\frac{n}{\sqrt{m_n\|\Lambda_n\|}}\hat{Q}_n\left(n^{-1/2}\Lambda_n^{\frac{1}{2}}\alpha + \theta_n\right) - \frac{n}{\sqrt{m_n\|\Lambda_n\|}}\hat{Q}_n(\theta_n) \\
&- \frac{\sqrt{n}}{\sqrt{m_n\|\Lambda_n\|}}\alpha^T\Lambda_n^{\frac{1}{2}}U_n - \frac{\alpha^T\Lambda_n^{\frac{1}{2}}H_n\Lambda_n^{\frac{1}{2}}\alpha}{2\sqrt{m_n\|\Lambda_n\|}}\Bigg| < \epsilon.
\end{aligned} \tag{48}$$

holds with probability at least $(1 - \epsilon/2)$ for large n.

Define the quadratic form

$$B_n(\alpha) = \frac{\sqrt{n}}{\sqrt{m_n\|\Lambda_n\|}}\alpha^T\Lambda_n^{\frac{1}{2}}U_n + \frac{\alpha^T\Lambda_n^{\frac{1}{2}}H_n\Lambda_n^{\frac{1}{2}}\alpha}{2\sqrt{m_n\|\Lambda_n\|}}.$$

Its minimizer is $\alpha_n = -\Lambda_n^{-\frac{1}{2}}H_n^{-1}n^{1/2}U_n$. Since for any $c \neq 0$,

$$c^TU_n = \binom{n}{m_n}^{-1} \sum_{1\leq i_1<i_2<\ldots<i_{m_n}\leq n} c^T g_n\left(Z_{i_1}, Z_{i_2}, \ldots, Z_{i_{m_n}}, \theta_n\right)$$

And

$$m_n \operatorname{Var}\left[E\left(c^T g_n\left(Z_1, \ldots, Z_{m_n}, \theta_n\right) \mid Z_1\right)\right] = c^T m_n K_n c.$$

By Rayleigh-Ritz theorem and assumption (iv), $c^T m_n K_n c \nrightarrow 0$ for any $c \neq 0$. Utilizing Theorem 1 of Peng et al. (2022), c^TU_n is asymptotically normal. Hence, by Cramér-Wold device,

$$\frac{\sqrt{n}}{m_n}K_n^{-\frac{1}{2}}U_n \xrightarrow{d} \mathcal{N}\left(0, I\right).$$

So

$$\alpha_n = -\Lambda_n^{-\frac{1}{2}}H_n^{-1}n^{1/2}U_n \xrightarrow{d} \mathcal{N}\left(0, I\right).$$

The minimum value of the quadratic form is

$$B_n\left(\alpha_n\right) = -nU_n^TH_n^{-1}U_n/2\sqrt{m_n\|\Lambda_n\|}.$$

Further, α_n is bounded in probability. So we can select an M such that

$$P\left(\|\alpha_n\| < M - 1\right) \geq 1 - \epsilon/2. \tag{49}$$

The rest of the argument is on the intersection of the two events in (48) and (49), which has a probability of at least $1 - \epsilon$.

Consider the convex function

$$A_n(\alpha) = \frac{n}{\sqrt{m_n\|\Lambda_n\|}}\hat{Q}_n\left(n^{-1/2}\Lambda_n^{\frac{1}{2}}\alpha + \theta_n\right) - \frac{n}{\sqrt{m_n\|\Lambda_n\|}}\hat{Q}_n(\theta_n).$$

From (48),

$$A_n\left(\alpha_n\right) \leq B_n\left(\alpha_n\right) + \epsilon = -nU_n^TH_n^{-1}U_n/2\sqrt{m_n\|\Lambda_n\|} + \epsilon. \tag{50}$$

Again by using (48), we know the value of $A_n(\alpha)$ is at least

$$B_n(\alpha) - \epsilon. \tag{51}$$

By simple calculation, the bound in (51) is always strictly larger than the one in (50) when $\|\alpha - \alpha_n\| \geq T(\epsilon\sqrt{m_n\|\Lambda_n\|})^{\frac{1}{2}}/\|\Lambda_n^{\frac{1}{2}}\|$, where $T = 4\left[\lambda_{\min}(H_n)\right]^{-1/2}$ and $\lambda_{\min}$ denotes the minimum eigenvalue.

On the other hand A_n has the minimizer $\sqrt{n}\Lambda_n^{-\frac{1}{2}}(\hat{\theta}_n - \theta_n)$. So, using the fact that A_n is convex, it follows that its minimizer satisfies

$$\|\sqrt{n}\Lambda_n^{-\frac{1}{2}}(\hat{\theta}_n - \theta_n) - \alpha_n\| < T(\epsilon\sqrt{m_n\|\Lambda_n\|})^{\frac{1}{2}}/\|\Lambda_n^{\frac{1}{2}}\|. \tag{52}$$

Note that $\frac{(\sqrt{m_n\|\Lambda_n\|})^{\frac{1}{2}}}{\|\Lambda_n^{\frac{1}{2}}\|} = \frac{m_n^{\frac{1}{4}}}{\|\Lambda_n\|^{\frac{1}{4}}} < \infty$. Since (52) holds with probability at least $(1-\epsilon)$ where ϵ is arbitrary, we get

$$\sqrt{n}\Lambda_n^{-\frac{1}{2}}(\hat{\theta}_n - \theta_n) - \alpha_n = o_p(1).$$

The first part of the theorem has been proved, and the second part follows from the multivariate version of the Central Limit Theorem of U_n. ■

Proof [Proof of Theorem 14] Since the assumption (A1) and assumption (A4) hold, by Lemma 17, we have $d(\hat{r}_\oplus(x), \tilde{r}_\oplus(x)) = o_p(1)$. Then $P(\hat{r}_\oplus(x) \in G) \to 1$ as $n \to \infty$.

For large n, let

$$u_n(x) = \phi(\tilde{r}_\oplus(x)), \quad \hat{u}_n(x) = \phi(\hat{r}_\oplus(x)).$$

Since

$$Ef_n(Z_1, Z_2, \ldots, Z_{s_n}, u) = Eh_n\left(Z_1, Z_2, \ldots, Z_{s_n}, \phi^{-1}(u)\right) = \tilde{R}_n(\phi^{-1}(u)),$$

we have

$$u_n(x) = \underset{u\in U}{\operatorname{argmin}} Ef_n(Z_1, Z_2, \ldots, Z_{s_n}, u) \triangleq \underset{u\in U}{\operatorname{argmin}} Q_n(u).$$

Similarly, since

$$\begin{aligned}
&\binom{n}{s_n}^{-1} \sum_{1\le i_1<i_2<\ldots<i_{s_n}\le n} f_n(Z_{i_1}, Z_{i_2}, \ldots, Z_{i_{s_n}}, u) \\
=&\binom{n}{s_n}^{-1} \sum_{1\le i_1<i_2<\ldots<i_{s_n}\le n} h_n\left(Z_{i_1}, Z_{i_2}, \ldots, Z_{i_{s_n}}, \phi^{-1}(u)\right) \\
=&\hat{R}_n(\phi^{-1}(u)),
\end{aligned}$$

we have

$$\hat{u}_n(x) = \underset{u\in U}{\operatorname{argmin}} \binom{n}{s_n}^{-1} \sum_{1\le i_1<i_2<\ldots<i_{s_n}\le n} f_n(Z_{i_1}, Z_{i_2}, \ldots, Z_{i_{s_n}}, u) \triangleq \underset{u\in U}{\operatorname{argmin}} \hat{Q}_n(u).$$

When we consider M_{s_n}-estimator $\hat{u}_n(x)$ of $u_n(x)$, the assumptions of Theorem 13 are satisfied by the assumptions (A13)–(A18). Hence by Theorem 13, we get

$$\sqrt{n}\Lambda_n^{-1/2}\left\{\hat{u}_n(x) - u_n(x)\right\} \xrightarrow{d} \mathcal{N}(0, I)$$

where

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

■

References

Miguel A Arcones and Evarist Giné. Limit theorems for U-processes. *The Annals of Probability*, pages 1494–1542, 1993.

Sylvain Arlot and Robin Genuer. Analysis of purely random forests bias. *arXiv preprint arXiv:1407.3939*, 2014.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

Satarupa Bhattacharjee and Hans-Georg Müller. Single index Fréchet regression. *The Annals of Statistics*, 51(4):1770–1798, 2023.

Rabi Bhattacharya and Lizhen Lin. Omnibus CLTs for Fréchet means and nonparametric inference on non-Euclidean spaces. *Proceedings of the American Mathematical Society*, 145(1):413–428, 2017.

Rabi Bhattacharya and Vic Patrangenaru. Large sample theory of intrinsic and extrinsic sample means on manifolds. *The Annals of Statistics*, 31(1):1–29, 2003.

Rabi Bhattacharya and Vic Patrangenaru. Large sample theory of intrinsic and extrinsic sample means on manifolds-II. *Annals of statistics*, pages 1225–1259, 2005.

Gérard Biau. Analysis of a random forests model. *Journal of Machine Learning Research*, 13: 1063–1095, 2012.

Gérard Biau, Luc Devroye, and Gäbor Lugosi. Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9(9), 2008.

Adam Bloniarz, Ameet Talwalkar, Bin Yu, and Christopher Wu. Supervised neighborhoods for distributed nonparametric regression. In *Artificial Intelligence and Statistics*, pages 1450–1459. PMLR, 2016.

Arup Bose and Snigdhansu Chatterjee. *U-statistics, M_m-estimators and Resampling*. Springer, 2018.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Louis Capitaine. Frechforest: Frechet random forests. *R package version 0.95*, 2021.

Louis Capitaine, Jérémie Bigot, Rodolphe Thiébaut, and Robin Genuer. Fréchet random forests for metric space valued regression with non Euclidean predictors. *arXiv preprint arXiv:1906.01741*, 2019.

Domagoj Cevid, Loris Michel, Jeffrey Näf, Peter Bühlmann, and Nicolai Meinshausen. Distributional random forests: Heterogeneity adjustment and multivariate distributional regression. *Journal of Machine Learning Research*, 23(333):1–79, 2022.

Xiaohui Chen and Kengo Kato. Jackknife multiplier bootstrap: finite sample approximations to the U-process supremum with applications. *Probability Theory and Related Fields*, 176(3):1097–1163, 2020.

Yaqing Chen and Hans-Georg Müller. Uniform convergence of local Fréchet regression with applications to locating extrema and time warping for metric space valued trajectories. *The Annals of Statistics*, 50(3):1573–1592, 2022.

Yaqing Chen, Alvaro Gajardo, Jianing Fan, Qixian Zhong, Paromita Dubey, Kyunghee Han, Satarupa Bhattacharjee, and Hans-Georg Müller. frechet: Statistical analysis for random objects and non-Euclidean data. *R package version 0.2.0.*, 2020.

Victor H de la Pena. Decoupling and Khintchine's inequalities for U-statistics. *The Annals of Probability*, pages 1877–1892, 1992.

Misha Denil, David Matheson, and Nando Freitas. Consistency of online random forests. In *International conference on machine learning*, pages 1256–1264. PMLR, 2013.

Cyrus DiCiccio and Joseph Romano. CLT for U-statistics with growing dimension. *Statistica Sinica*, 32(1), 2022.

Paromita Dubey and Hans-Georg Müller. Functional models for time-varying random objects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(2):275–327, 2020.

Roxane Duroux and Erwan Scornet. Impact of subsampling and tree depth on random forests. *ESAIM: Probability and Statistics*, 22:96–128, 2018.

Rina Friedberg, Julie Tibshirani, Susan Athey, and Stefan Wager. Local linear forests. *Journal of Computational and Graphical Statistics*, pages 1–15, 2020.

Wei Gao and Zhi-Hua Zhou. Towards convergence rate analysis of random forests for classification. *Advances in Neural Information Processing Systems*, 33, 2020.

Robin Genuer. Variance reduction in purely random forests. *Journal of Nonparametric Statistics*, 24(3):543–562, 2012.

Aritra Ghosal, Wendy Meiring, and Alexander Petersen. Fréchet single index models for object response regression. *Electronic Journal of Statistics*, 17(1):1074–1112, 2023.

Charles Heilig and Deborah Nolan. Limit theorems for the infinite-degree U-process. *Statistica Sinica*, pages 289–302, 2001.

Charles Martin Heilig. *An empirical process approach to U-processes of increasing degree*. University of California, Berkeley, 1997.

Matthias Hein. Robust nonparametric regression with metric-space valued output. In *Advances in Neural Information Processing Systems*, pages 718–726. Citeseer, 2009.

Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, 2008.

Jeong Min Jeon and Byeong U Park. Additive regression with Hilbertian responses. *The Annals of Statistics*, 48(5):2671–2697, 2020.

Jeong Min Jeon, Young Kyung Lee, Enno Mammen, and Byeong U Park. Locally polynomial Hilbertian additive regression. *Bernoulli*, 28(3):2034–2066, 2022.

Jason Klusowski. Sharp analysis of a simple model for random forests. In *International Conference on Artificial Intelligence and Statistics*, pages 757–765. PMLR, 2021.

Yi Lin and Yongho Jeon. Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590, 2006.

Z Lin, HG Müller, and BU Park. Additive models for symmetric positive-definite matrices and lie groups. *Biometrika*, 2022.

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

Zhenhua Lin. matrix-manifold: Basic operations and functions on matrix manifolds. *R package version 0.1.0*, 2020.

Nicolai Meinshausen and Greg Ridgeway. Quantile regression forests. *Journal of Machine Learning Research*, 7(6), 2006.

Lucas Mentch and Giles Hooker. Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *Journal of Machine Learning Research*, 17(1):841–881, 2016.

Maher Moakher. A differential geometric approach to the geometric mean of symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 26(3):735–747, 2005.

Wei Peng, Tim Coleman, and Lucas Mentch. Rates of convergence for random forests via generalized U-statistics. *Electronic Journal of Statistics*, 16(1):232–292, 2022.

Xavier Pennec, Pierre Fillard, and Nicholas Ayache. A Riemannian framework for tensor computing. *International Journal of computer vision*, 66(1):41–66, 2006.

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

Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741, 2015.

Robert P Sherman. Maximal inequalities for degenerate U-processes with applications to optimization estimators. *The Annals of Statistics*, pages 439–459, 1994.

Jon Arni Steingrimsson, Liqun Diao, and Robert L Strawderman. Censoring unbiased regression trees and ensembles. *Journal of the American Statistical Association*, 114(525):370–383, 2019.

Aad van der Vaart and Jon A Wellner. *Weak Convergence and Empirical Processes: with Applications to Statistics*. Springer Science & Business Media, 1996.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Stefan Wager, Trevor Hastie, and Bradley Efron. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *The Journal of Machine Learning Research*, 15(1): 1625–1651, 2014.

Weichi Yao, Halina Frydman, Denis Larocque, and Jeffrey S Simonoff. Ensemble methods for survival function estimation with time-varying covariates. *Statistical Methods in Medical Research*, 31(11):2217–2236, 2022.

Chao Ying and Zhou Yu. Fréchet sufficient dimension reduction for random objects. *Biometrika*, 109(4):975–992, 2022.

Ying Yuan, Hongtu Zhu, Weili Lin, and James Stephen Marron. Local polynomial regression for symmetric positive definite matrices. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(4):697–719, 2012.

Qi Zhang, Lingzhou Xue, and Bing Li. Dimension reduction for Fréchet regression. *Journal of the American Statistical Association*, pages 1–15, 2023.