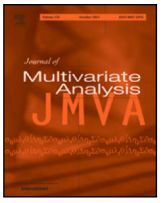
Journal of Multivariate Analysis 207 (2025) 105416

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva



Model averaging for global Fréchet regression

Daisuke Kurisu [a],*, Taisuke Otsu [b]

[a] Center for Spatial Information Science, The University of Tokyo 5-1-5, Kashiwanoha, Kashiwa-shi, Chiba 277-8568, Japan
[b] Department of Economics, London School of Economics Houghton Street, London, WC2A 2AE, UK

ARTICLE INFO

AMS 2020 subject classifications:
primary 62R02
secondary 62J02

Keywords:
Asymptotic optimality
Cross-validation
Global Fréchet regression
Model averaging

ABSTRACT

Non-Euclidean complex data analysis becomes increasingly popular in various fields of data science. In a seminal paper, Petersen and Müller (2019) generalized the notion of regression analysis to non-Euclidean response objects. Meanwhile, in the conventional regression analysis, model averaging has a long history and is widely applied in statistics literature. This paper studies the problem of optimal prediction for non-Euclidean objects by extending the method of model averaging. In particular, we generalize the notion of model averaging for global Fréchet regressions and establish an optimal property of the cross-validation to select the averaging weights in terms of the final prediction error. A simulation study illustrates excellent out-of-sample predictions of the proposed method.

1. Introduction

Non-Euclidean complex data analysis becomes increasingly popular in various fields of data science (see, Marron and Alonso [9] for an overview). A fundamental object to describe distributions of non-Euclidean random objects is the so-called Fréchet mean [5], which is a generalization of the conventional population mean. There is growing literature on statistical inference for the Fréchet means (see, e.g., Patrangenaru and Ellingson [10], for a survey). Recently, in a seminal paper, Petersen and Müller [11] generalized the notion of the Fréchet mean to conditional distributions, and developed nonparametric and least square regression analyses for non-Euclidean random objects, called the local and global Fréchet regressions, respectively.

In the conventional regression analysis, a central question is how to select or combine information from various predictors, and model selection and model averaging are widely applied in the statistics literature (see, Claeskens and Hjort [4], for a survey). Indeed Tucker et al. [12] developed a model selection method for global Fréchet regressions by extending the ridge selection operator to the present context, and established its selection consistency. See also Ying and Yu [14] for sufficient dimension reduction on non-Euclidean random objects using Euclidean predictors. This paper addresses another open question, model averaging of regression models for non-Euclidean response objects. In particular, we focus on optimal prediction for non-Euclidean objects by extending the method of model averaging.

In this paper, we generalize the notion of model averaging for global Fréchet regressions and establish an optimal out-of-sample prediction property of the cross-validation to select the averaging weights in terms of the final prediction error [1]. First of all, it is not trivial how to conduct model averaging for global Fréchet regressions that reside in non-Euclidean spaces. By adapting the construction of the empirical Fréchet mean to weighted averages over a class of global Fréchet regressions, we develop a model averaging scheme as a minimizer of a weighted average of squared metrics of global Fréchet regressions. Second, we introduce and study the notions of the final prediction error for out-of-sample predictions and cross-validation for model averaging of regression models on non-Euclidean random objects. We refer to Bhattacharjee and Müller [2] and Ghosal et al. [6] that study out-of-sample

* Corresponding author.
E-mail address: daisukekurisu@csis.u-tokyo.ac.jp (D. Kurisu).

https://doi.org/10.1016/j.jmva.2025.105416
Received 11 May 2024; Received in revised form 13 January 2025; Accepted 17 January 2025
Available online 25 January 2025

cross-validation criteria in the context of single index modeling of random objects. In contrast to Tucker et al. [12] who studied consistent model selection for global Fréchet regressions, this paper investigates optimal model averaging for the out-of-sample prediction when all global Fréchet regressions are misspecified.

This paper is organized as follows. Section 2 introduces our basic setup and model averaging estimator. Section 3 presents our main result, asymptotic optimality of the cross-validation to select the model averaging weights in terms of the final prediction error. Section 4 illustrates the main result by a simulation study. Technical details to prove our main result are included in Section 5.

2. Model averaging estimator

Let $(\mathbb{Y}, d)$ be a totally bounded metric space and consider a random object Y that takes values in $\mathbb{Y}$. We are concerned with the situation where Y is a complex object so that the space $\mathbb{Y}$ may be non-Euclidean and may not lie in a vector space. In such a situation, a standard notion of mean is the so-called Fréchet mean $\eta_\oplus = \arg\min_{\eta\in\mathbb{Y}} \mathbb{E}[d^2(Y,\eta)]$, and there is rich literature on statistical inference for $\eta_\oplus$.

In a seminal paper, Petersen and Müller [11] extended the notion of the Fréchet mean to regression problems and proposed the Fréchet regression function

$$\eta_\oplus(x) = \arg\min_{\eta\in\mathbb{Y}} \mathbb{E}[d^2(Y,\eta)|X = x]$$

for an Euclidean vector of predictors X. Furthermore, Petersen and Müller [11] generalized the idea of global least squares regression and developed the global Fréchet regression:

$$L_\oplus(x) = \arg\min_{\eta\in\mathbb{Y}} \mathbb{E}[\{1 + (x-\mu)^\top \Sigma^{-1}(X-\mu)\}d^2(Y,\eta)],$$

where $\mu = \mathbb{E}[X]$ and $\Sigma = \mathrm{Var}(X)$. Note that $L_\oplus(x)$ becomes the conventional population least square regression when $\mathbb{Y}$ is Euclidean and d is the Euclidean distance.

We now introduce our setup for model averaging of global Fréchet regressions. To simplify the presentation, we hereafter focus on the case where the researcher conducts model selection based on a nested sequence of predictors $X = (X_1, X_2, \dots, X_M)^\top \in \mathbb{R}^M$. One can also apply our method to average over other subsets of X and analogous theoretical results in Section 3 can be obtained. Although there is no theoretical difficulty to consider a large (but fixed) number of subsets of X, practically it needs to be moderate due to the computational cost. It is beyond the scope of this paper to consider how to select the subsets of X to be averaged under computational constraints.

Let $X^{(m)} = (X_1, \dots, X_m)^\top \in \mathbb{R}^m$, $m \in \{1, \dots, M\}$, be a nested sequence of predictors, $x^{(m)} = (x_1, \dots, x_m)^\top \in \mathbb{R}^m$, and for $m \in \{1, \dots, M\}$, let

$$L_\oplus^{(m)}(x^{(m)}) = \arg\min_{\eta\in\mathbb{Y}} \mathbb{E}[\{1 + (x^{(m)} - \mu^{(m)})^\top (\Sigma^{(m)})^{-1}(X^{(m)} - \mu^{(m)})\}d^2(Y,\eta)]$$

be the global Fréchet regression based on the predictors $X^{(m)}$, where $\mu^{(m)} = \mathbb{E}[X^{(m)}]$ and $\Sigma^{(m)} = \mathrm{Var}(X^{(m)})$. In this paper, M is treated as fixed. An extension allowing M to increase depending on the sample size n, as assumed in model averaging of linear regression models for Euclidean data, necessitates a substantial extension in the theoretical analysis of the global Fréchet regression itself and we left this extension as future research. In order to build the notion of model averaging for the global Fréchet regressions $\{L_\oplus^{(m)}(x^{(m)})\}_{m=1}^M$, we note that in the q-dimensional Euclidean space, the weighted average $\bar{L}_{\boldsymbol{w}} = \sum_{m=1}^M w_m L^{(m)}$ of points $L^{(m)} \in \mathbb{R}^q$ can be defined as

$$\bar{L}_{\boldsymbol{w}} = \arg\min_{\eta\in\mathbb{R}^q} \sum_{m=1}^M w_m d_E^2(L^{(m)}, \eta),$$

for the Euclidean distance d_E. Then the model averaging for global Fréchet regressions can be defined as

$$m_\oplus(\boldsymbol{w}, x^{(M)}) = \arg\min_{\eta\in\mathbb{Y}} \sum_{m=1}^M w_m d^2(L_\oplus^{(m)}(x^{(m)}), \eta).$$

Based on an independent and identically distributed sample $\mathcal{D}_n = \{X_i, Y_i\}_{i=1}^n$ of (X, Y), $L_\oplus^{(m)}(x^{(m)})$ and $m_\oplus(\boldsymbol{w}, x^{(M)})$ can be estimated by their sample counterparts:

$$\hat{L}_\oplus^{(m)}(x^{(m)}) = \arg\min_{\eta\in\mathbb{Y}} \frac{1}{n}\sum_{i=1}^n \{1 + (x^{(m)} - \bar{X}^{(m)})^\top (\hat{\Sigma}^{(m)})^{-1}(X_i^{(m)} - \bar{X}^{(m)})\}d^2(Y_i,\eta),$$

$$\hat{m}_\oplus(\boldsymbol{w}, x^{(M)}) = \arg\min_{\eta\in\mathbb{Y}} \sum_{m=1}^M w_m d^2(\hat{L}_\oplus^{(m)}(x^{(m)}), \eta),$$

where $\bar{X}^{(m)} = n^{-1}\sum_{i=1}^n X_i^{(m)}$ and $\hat{\Sigma}^{(m)} = n^{-1}\sum_{i=1}^n (X_i^{(m)} - \bar{X}^{(m)})(X_i^{(m)} - \bar{X}^{(m)})^\top$.

As a criterion to evaluate model averaging weights, we extend the notion of the final prediction error [1] to the global Fréchet regression as

$$\mathrm{FPE}_n(\boldsymbol{w}) = \mathbb{E}[d^2(\mathcal{Y}, \hat{m}_\oplus(\boldsymbol{w}, \mathcal{X}))|\mathcal{D}_n],$$

where $(\mathcal{X}, \mathcal{Y})$ is an independent copy of (X_i, Y_i). In this paper, we consider the situation where all global Fréchet regressions and their averaging versions are misspecified, and develop a selection rule for the averaging weights to achieve an optimal out-of-sample

prediction property in terms of $\mathrm{FPE}_n(\boldsymbol{w})$. This is a sharp contrast with the approach in Tucker et al. [12], which focuses on the consistent selection of a true model.

As a feasible selection rule for the optimal weights, we propose to minimize the leave-one-out cross-validation criterion:

$$\mathrm{CV}_n(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n} d^2(Y_i, \hat{m}_{\oplus,-i}(\boldsymbol{w}, X_i)),$$

where $\hat{m}_{\oplus,-i}(\boldsymbol{w}, x^{(M)}) = \arg\min_{\eta\in\mathbb{Y}} \sum_{m=1}^{M} w_m d^2(\hat{L}^{(m)}_{\oplus,-i}(x^{(m)}), \eta)$ and $\hat{L}^{(m)}_{\oplus,-i}(x^{(m)})$ is defined as $\hat{L}^{(m)}_{\oplus}(x^{(m)})$ with the ith observation deleted. Letting $\mathbb{W} = \{\boldsymbol{w} = (w_1, \dots, w_M)^\top \in [0,1]^M : \sum_{m=1}^{M} w_m = 1\}$, our model averaging estimator for global Fréchet regressions is defined as

$$\hat{m}_{\oplus}(\hat{\boldsymbol{w}}, x^{(M)}), \quad \text{where } \hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}\in\mathbb{W}} \mathrm{CV}_n(\boldsymbol{w}).$$

Remark 1 (*Connection with the Euclidean Case*)**.** For the conventional Euclidean case of $(\mathbb{Y}, d) = (\mathbb{R}, d_E)$, our model averaging estimator $\hat{m}_{\oplus}(\hat{\boldsymbol{w}}, x^{(M)})$ reduces to the one studied in Hansen [7]. To see this, consider the following linear regression model

$$Y_i = \tilde{X}_i^\top \tilde{\beta} + \varepsilon_i,$$

where $\tilde{X}_i = (X_{i,1}, \dots, X_{i,M_0})^\top$ is an Euclidean vector of predictors, $\tilde{\beta} = (\beta_1, \dots, \beta_{M_0})^\top$ is an unknown vector of parameters, and ε_i is an unobservable error term with $\mathbb{E}[\varepsilon_j|\tilde{X}_i] = 0$. We consider a situation where only a subset of the predictors $X_i = X_i^{(M)} = (X_{i,j_1}, \dots, X_{i,j_M})^\top \in \mathbb{R}^M$ with $1 \le M \le M_0$ are observable, and let $X_i^{(m)} = (X_{i,j_1}, \dots, X_{i,j_m})^\top$ for $m \in \{1, \dots, M\}$. Assume that for each $m \in \{1, \dots, M\}$, $\mathbb{E}[X_1^{(m)} X_1^{(m)\top}]$ is invertible and let $L^{(m)}(x^{(m)}) = x^{(m)\top}\theta^{(m)}$ be the mth model, where $x^{(m)} = (x_{j_1}, \dots, x_{j_m})^\top \in \mathbb{R}^m$ and $\theta^{(m)} = \mathbb{E}[X_1^{(m)} X_1^{(m)\top}]^{-1}\mathbb{E}[X_1^{(m)} Y_1]$. Then the linear prediction of Y_i at $X_i^{(m)} = x^{(m)}$ based on the ordinary least square estimation using $\{X_i\}_{i=1}^n$ is given by $\hat{L}^{(m)}(x^{(m)}) = x^{(m)\top}\hat{\theta}^{(m)}$, where

$$\hat{\theta}^{(m)} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^{(m)} X_i^{(m)\top}\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} X_i^{(m)} Y_i.$$

Then the model averaging estimator of $\mathbb{E}[Y_i|\tilde{X}_i]$ based on $\{\hat{L}^{(m)}(X_i^{(m)})\}_{m=1}^M$ is obtained as

$$\hat{m}(\boldsymbol{w}, X_i^{(M)}) = \sum_{m=1}^{M} w_m X_i^{(m)\top}\hat{\theta}^{(m)} = \sum_{m=1}^{M} w_m \hat{L}^{(m)}(X_i^{(m)}) = \arg\min_{\eta\in\mathbb{R}} \sum_{m=1}^{M} w_m d_E^2(\hat{L}^{(m)}(X_i^{(m)}), \eta).$$

Further, one can see that $\mathrm{FPE}_n(\boldsymbol{w})$ and $\mathrm{CV}_n(\boldsymbol{w})$ correspond to the final prediction error (or the out-of-sample prediction error) in the Euclidean case.

3. Optimality

We now present our main result, the optimality of the model averaging estimator $\hat{m}_{\oplus}(\hat{\boldsymbol{w}}, x^{(M)})$ in terms of the final prediction error. For $z = (z_1, \dots, z_q)^\top \in \mathbb{R}^q$, let $\|z\|_{\ell^2} = \sqrt{\sum_{j=1}^q z_j^2}$ be the ℓ^2 norm and $\|z\|_{\ell^1} = \sum_{j=1}^q |z_j|$ be the ℓ^1 norm, and

$$R(\boldsymbol{w}, x^{(M)}, \eta) = \sum_{m=1}^{M} w_m d^2(L^{(m)}_{\oplus}(x^{(m)}), \eta), \quad \hat{R}(\boldsymbol{w}, x^{(M)}, \eta) = \sum_{m=1}^{M} w_m d^2(\hat{L}^{(m)}_{\oplus}(x^{(m)}), \eta).$$

We impose the following assumptions.

Assumption 1.

A1. $(\mathbb{Y}, d)$ is a totally bounded metric space, $\mathbb{P}(\|X\|_{\ell^2} \le B) = 1$ for some constant $B > 0$, $L^{(m)}_{\oplus}(x)$ is continuous at $x^{(M)} \in \mathbb{R}^M$ with $\|x^{(M)}\|_{\ell^2} \le B$, and the global Fréchet regression estimators $\{\hat{L}^{(m)}_{\oplus}(x^{(m)})\}_{m=1}^M$ are uniformly consistent in the sense that as $n \to \infty$,

$$\max_{1\le m\le M} \sup_{\|x^{(M)}\|_{\ell^2}\le B} d(\hat{L}^{(m)}_{\oplus}(x^{(m)}), L^{(m)}_{\oplus}(x^{(m)})) \xrightarrow{p} 0.$$

A2. Almost surely, for each $\boldsymbol{w} \in \mathbb{W}$ and $\|x^{(M)}\|_{\ell^2} \le B$, $m_{\oplus}(\boldsymbol{w}, x^{(M)})$ and $\hat{m}_{\oplus}(\boldsymbol{w}, x^{(M)})$ exist and are unique. Additionally, for each $\varepsilon > 0$,

$$\inf_{\boldsymbol{w}\in\mathbb{W}, \|x^{(M)}\|_{\ell^2}\le B} \inf_{d(\eta, m_{\oplus}(\boldsymbol{w}, x^{(M)}))>\varepsilon} R(\boldsymbol{w}, x^{(M)}, \eta) - R(\boldsymbol{w}, x^{(M)}, m_{\oplus}(\boldsymbol{w}, x^{(M)})) > 0$$

and there exists $\zeta = \zeta(\varepsilon) > 0$ such that as $n \to \infty$,

$$\mathbb{P}\left(\inf_{\boldsymbol{w}\in\mathbb{W}, \|x^{(M)}\|_{\ell^2}\le B} \inf_{d(\eta, \hat{m}_{\oplus}(\boldsymbol{w}, x^{(M)}))>\varepsilon} \hat{R}(\boldsymbol{w}, x^{(M)}, \eta) - \hat{R}(\boldsymbol{w}, x^{(M)}, \hat{m}_{\oplus}(\boldsymbol{w}, x^{(M)})) \ge \zeta\right) \to 1.$$

A3. There exist constants $\bar{D}_B > 0$ and $0 < \beta_B \leq 1$ such that for each $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{W}$,

$$\sup_{\|x^{(M)}\|_{\ell^2} \leq B} d(m_\oplus(\boldsymbol{w}_1, x^{(M)}), m_\oplus(\boldsymbol{w}_2, x^{(M)})) \leq \bar{D}_B \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_{\ell^1}^{\beta_B}.$$

A4. There exists a constant $\kappa > 0$ such that $\inf_{\boldsymbol{w} \in \mathbb{W}} \mathbb{E}[d^2(Y, m_\oplus(\boldsymbol{w}, X))] \geq \kappa$.

A1 is on the support of Y and X, and a high-level condition on the uniform consistency of the global Fréchet regression estimators whose primitive conditions are given in Theorem 1 of Petersen and Müller [11]. In particular, under Conditions (U0)-(U2) in Petersen and Müller [11], we can show A1. A2 is an additional condition to guarantee uniform consistency of the model averaging estimator $\hat{m}_\oplus(\boldsymbol{w}, x^{(M)})$, which is an analog of Condition (U0) of Petersen and Müller [11] and is commonly imposed to derive the consistency of M-estimators (see, e.g., van der Vaart and Wellner [13]). A3 and A4 are additional conditions to establish the asymptotic optimality of our model averaging estimator $\hat{m}_\oplus(\hat{\boldsymbol{w}}, x^{(M)})$ using the cross-validation. A3 is a Lipschitz-type condition for weights to derive uniform convergence of $n^{-1}\sum_{i=1}^n d^2(Y_i, m_\oplus(\boldsymbol{w}, X_i))$. A4 is imposed to control the approximation error of all the (possibly misspecified) global Fréchet regressions and their averaged versions.

Remark 2. Although our setup accommodates dependent observations, the leave-one-out cross-validation is known to be affected by dependence in finite samples [3]. We expect that an analogous remedy by Chu and Marron [3] to delete more than one observation can be adapted to our context.

Based on these assumptions, our main result is presented as follows.

Theorem 1.

(i) Under A1 and A2, it holds

$$\sup_{\boldsymbol{w} \in \mathbb{W}, \|x^{(M)}\|_{\ell^2} \leq B} d(\hat{m}_\oplus(\boldsymbol{w}, x^{(M)}), m_\oplus(\boldsymbol{w}, x^{(M)})) \xrightarrow{p} 0 \text{ as } n \to \infty.$$

(ii) Under Assumption 1, it holds

$$\frac{\mathrm{FPE}_n(\hat{\boldsymbol{w}})}{\inf_{\boldsymbol{w} \in \mathbb{W}} \mathrm{FPE}_n(\boldsymbol{w})} \xrightarrow{p} 1 \text{ as } n \to \infty.$$

Theorem 1(i) shows uniform consistency of the model averaging estimator $\hat{m}_\oplus(\boldsymbol{w}, x)$ over the weights $\boldsymbol{w}$ and values of predictors x. Compared to the uniform consistency result for the global Fréchet regression (Theorem 1 of Petersen and Müller [11]), a technical challenge is to establish the uniform convergence over the averaging weight $\boldsymbol{w} \in \mathbb{W}$ (in addition to $x^{(M)}$). Since the averaging estimator $\hat{m}_\oplus(\boldsymbol{w}, x^{(M)})$ depends on $\boldsymbol{w}$ and $x^{(M)}$ in different ways, our proof of Theorem 1(i) is non-trivially different from that of Theorem 1 by Petersen and Müller [11].

Theorem 1 (ii) establishes the optimal out-of-sample prediction property of our averaging weights $\hat{\boldsymbol{w}}$ that minimizes the cross-validation criterion $\mathrm{CV}_n(\boldsymbol{w})$. This result says $\mathrm{FPE}_n(\hat{\boldsymbol{w}})$ by using $\hat{\boldsymbol{w}}$ is asymptotically equivalent to the oracle final prediction error to minimize $\mathrm{FPE}_n(\boldsymbol{w})$ over $\boldsymbol{w} \in \mathbb{W}$. To the best of our knowledge, in the literature of statistics for non-Euclidean objects, this is the first optimality result for the weight selection by the cross-validation criterion. We note that our proof strategy is very different from the conventional Euclidean case (see, e.g., Hansen [7] and Li [8]), where the support of Y has a linear structure. Instead, we invoke the empirical process theory to control the difference between the cross-validation and final prediction error criteria uniformly over the weight space.

We close this section by illustrating our main result with some specific examples.

Example 1 (*Symmetric Positive-Definite Matrices with the Frobenius or Cholesky Decomposition Distance*)**.** Let $\mathbb{Y}$ be the set of symmetric positive-definite matrices with the Frobenius norm or Cholesky decomposition distance. For SPD matrices A_1 and A_2, the Cholesky decomposition yields $A_1 = (A_1^{1/2})^\top A_1^{1/2}$ and $A_2 = (A_2^{1/2})^\top A_2^{1/2}$, where $A_1^{1/2}$ and $A_2^{1/2}$ are upper triangle matrices with positive diagonal components. Then define the Cholesky decomposition distance between A_1 and A_2 as

$$d_C(A_1, A_2) = \sqrt{\mathrm{trace}((A_1^{1/2} - A_2^{1/2})^\top (A_1^{1/2} - A_2^{1/2}))}.$$

For these examples, Propositions 2 and Theorem 1 in Petersen and Müller [11] guarantee A1.

Let $L_\oplus^{(m)}(x^{(m)})$ be the global Fréchet regression function of the mth model. For SPD matrices with the Frobenius distance, the model average global Fréchet regression function $m_\oplus(\boldsymbol{w}, x^{(M)})$ is given by $m_\oplus(\boldsymbol{w}, x^{(M)}) = \sum_{m=1}^M w_m L_\oplus^{(m)}(x^{(m)})$. Let $(L_\oplus^{(m)1/2}(x^{(m)}))^\top L_\oplus^{(m)1/2}(x^{(m)})$ be the Cholesky decomposition of $L_\oplus^{(m)}(x^{(m)})$. For SPD matrices with the Cholesky decomposition distance, $m_\oplus(\boldsymbol{w}, x^{(M)})$ is given by

$$m_\oplus(\boldsymbol{w}, x^{(M)}) = \left(\sum_{m=1}^M w_m L_\oplus^{(m)1/2}(x^{(m)})\right)^\top \left(\sum_{m=1}^M w_m L_\oplus^{(m)1/2}(x^{(m)})\right).$$

Applying a similar argument in the proof of Proposition 2 in Petersen and Müller [11], one can see that A2 and A3 are satisfied with $\beta_B = 1$.

Example 2 (*Probability Distributions with the Wasserstein Metric*)**.** Let $\mathbb{Y}$ be the set of univariate probability distributions F on a compact set equipped with the Wasserstein metric d_W defined as

$$d_W(F_1, F_2) = \sqrt{\int_0^1 (F_1^{-1}(t) - F_2^{-1}(t))^2 dt}$$

for the quantile functions F_1^{-1} and F_2^{-1} of probability distributions F_1 and F_2. For this example, Proposition 1 and Theorem 1 in Petersen and Müller [11] guarantee A1. Let $L_{\oplus}^{(m)}(x^{(m)})$ be the global Fréchet regression function of the mth model, which is a distribution function on a compact set, and let $L_{\oplus}^{(m)-1}(x^{(m)})$ be the quantile function of $L_{\oplus}^{(m)}(x^{(m)})$. The quantile function of the model average global Fréchet regression function $m_{\oplus}^{-1}(\boldsymbol{w}, x^{(M)})$ is given by $m_{\oplus}^{-1}(\boldsymbol{w}, x^{(M)}) = \sum_{m=1}^{M} w_m L_{\oplus}^{(m)-1}(x^{(m)})$. Applying a similar argument in the proof of Proposition 1 in Petersen and Müller [11], one can see that A2 and A3 are satisfied with $\beta_B = 1$.

Example 3 (*Spherical Data with the Geodesic Distance*)**.** Let $\mathbb{Y} = \mathbb{S}^2$, the unit sphere in $\mathbb{R}^3$, equipped with the geodesic distance $d_g(x_1, x_2) = \arccos(x_1^\top x_2)$ for $x_1, x_2 \in \mathbb{S}^2$. Specifically, Petersen and Müller [11] and Tucker et al. [12] considered the following Fréchet regression model. Let $\eta_{\oplus}(x) \in \mathbb{S}^2$ be a regression function and V be a random vector on the tangent space $T_{\eta_{\oplus}(X)}$. Define Y as an exponential map of V at $\eta_{\oplus}(X)$, i.e.,

$$Y = \mathrm{Exp}_{\eta_{\oplus}(X)}(V) = \cos(\|V\|_{\ell^2})\eta_{\oplus}(X) + \sin(\|V\|_{\ell^2})\frac{V}{\|V\|_{\ell^2}}.$$

Proposition 3 in Petersen and Müller [11] gives sufficient conditions of A1.

4. Simulation

4.1. Data generating processes

We consider two data generating processes: (i) the set of 5×5 symmetric positive-definite (SPD) matrices with the Cholesky decomposition distance and (ii) the set of univariate probability distributions with the Wasserstein metric.

For predictors $X_i = (X_{i,1}, \ldots, X_{i,p})^\top$ with $p = 9$, we consider the following designs, which are modifications of the data generating processes considered in Tucker et al. [12]: (i) Generate p-dimensional multivariate Gaussian random variables $Z_i = (Z_{i,1}, \ldots, Z_{i,p})^\top$ with $\mathbb{E}[Z_{i,j}] = 0$ and $\mathrm{Cov}(Z_{i,j}, Z_{i,k}) = \rho^{|j-k|}$, and then set $X_{i,j} = 2\Phi(Z_{i,j})$, where $\Phi(\cdot)$ is the standard normal distribution function. (ii) Generate $X_{i,j} = U_{i,j}$, where $\{U_{i,j}\}_{1 \le i \le n, 1 \le j \le p}$ is an array of independent and identically distributed random variables with the uniform distribution on $[0, 2]$.

4.1.1. SPD matrices with the Cholesky decomposition distance

We set the random object Y_i as 5×5 SPD matrix and consider the following Fréchet regression function: $\eta_{\oplus}(x) = \mathbb{E}[Y|X = x] = \mathbb{E}[A]^\top \mathbb{E}[A]$, where

$$\begin{aligned}\mathbb{E}[A] = &\left\{\mu_0 + \beta\left(x_1 + \frac{x_3}{3} + \frac{x_5}{5} + \frac{x_7}{7} + \frac{x_9}{9}\right) + \sigma_0 + \gamma\left(\frac{x_2}{2} + \frac{x_4}{4} + \frac{x_6}{6} + \frac{x_8}{8}\right)\right\} I_T \\ &+ \left\{\sigma_0 + \gamma\left(\frac{x_2}{2} + \frac{x_4}{4} + \frac{x_6}{6} + \frac{x_8}{8}\right)\right\} V,\end{aligned}$$

with the $T \times T$ identity matrix I_T and the $T \times T$ matrix $V = (I_{\{j<k\}})$. Conditional on $X = (X_1, \ldots, X_9)^\top$, the random response Y is generated by $Y = A^\top A$, where $A = (\mu + \sigma) I_T + \sigma V$ with

$$\mu | X \sim N\left(\mu_0 + \beta\left(X_1 + \frac{X_3}{3} + \frac{X_5}{5} + \frac{X_7}{7} + \frac{X_9}{9}\right), \nu_1\right),$$

$$\sigma | X \sim \mathrm{Gamma}\left(\nu_2^{-1}\left(\sigma_0 + \gamma\left(\frac{X_2}{2} + \frac{X_4}{4} + \frac{X_6}{6} + \frac{X_8}{8}\right)\right)^2, \frac{\nu_2}{\sigma_0 + \gamma\left(\frac{X_2}{2} + \frac{X_4}{4} + \frac{X_6}{6} + \frac{X_8}{8}\right)}\right).$$

In our simulation study, we set $n \in \{50, 100\}$, $T = 5$, $p = 9$, $\rho = 0.5$, $\mu_0 = 3$, $\sigma_0 = 3$, $\beta = 2$, $\gamma = 3$, $\nu_1 = 1$, and $\nu_2 = 2$.

4.1.2. Univariate probability distributions with the Wasserstein metric

Consider the following Fréchet regression function:

$$\eta_{\oplus}(x) = \mathbb{E}[Y(\cdot)|X = x] = \beta\left(x_1 + \frac{x_3}{3} + \frac{x_5}{5} + \frac{x_7}{7} + \frac{x_9}{9}\right) + \left\{\sigma_0 + \gamma\left(\frac{x_2}{2} + \frac{x_4}{4} + \frac{x_6}{6} + \frac{x_8}{8}\right)\right\} \Phi^{-1}(\cdot).$$

Conditional on X, the random response Y is generated as follows: $Y = \mu + \sigma\Phi^{-1}$ with

$$\mu | X \sim N\left(\beta\left(X_1 + \frac{X_3}{3} + \frac{X_5}{5} + \frac{X_7}{7} + \frac{X_9}{9}\right), \nu_1\right),$$

$$\sigma | X \sim \mathrm{Gamma}\left(\nu_2^{-1}\left(\sigma_0 + \gamma\left(\frac{X_2}{2} + \frac{X_4}{4} + \frac{X_6}{6} + \frac{X_8}{8}\right)\right)^2, \frac{\nu_2}{\sigma_0 + \gamma\left(\frac{X_2}{2} + \frac{X_4}{4} + \frac{X_6}{6} + \frac{X_8}{8}\right)}\right).$$

In our simulation study, we set $n \in \{50, 100\}$, $p = 9$, $\rho = 0.5$, $\beta = 0.75$, $\sigma_0 = 5$, $\beta = 2$, $\gamma = 0.5$, $\nu_1 = 1$, and $\nu_2 = 0.5$.

4.2. Results

We consider the following three methods to choose the weights in the model averaging: (i) The proposed cross-validation-based model averaging (CV), (ii) AIC-type model averaging, and (iii) BIC-type model averaging. The proposed method is the first and currently only asymptotically optimal selection procedure in terms of the final prediction error, and there is no theoretical justification for the methods (ii) and (iii) in the present setup (see Claeskens and Hjort [4] for a general discussion of the AIC or BIC model averaging).

For the mth model, we define the AIC- and BIC-type information criteria as

$$\mathrm{AIC}_m = n \ln \left(\frac{1}{n} \sum_{i=1}^{n} d^2(Y_i, \hat{L}_{\oplus}^{(m)}(X_i)) \right) + 2m, \ \mathrm{BIC}_m = n \ln \left(\frac{1}{n} \sum_{i=1}^{n} d^2(Y_i, \hat{L}_{\oplus}^{(m)}(X_i)) \right) + m \ln n,$$

where $d \in \{d_C, d_W\}$. Then the AIC- and BIC-type model average estimators are defined as

$$\hat{m}_{\oplus}(\hat{\boldsymbol{w}}^{\mathrm{AIC}}, x^{(M)}) = \arg\min_{\eta \in \mathbb{Y}} \sum_{m=1}^{M} \hat{w}_m^{\mathrm{AIC}} d^2(\hat{L}_{\oplus}^{(m)}(x^{(m)}), \eta), \ \hat{w}_m^{\mathrm{AIC}} = \frac{\exp(-\mathrm{AIC}_m/2)}{\sum_{j=1}^{M} \exp(-\mathrm{AIC}_j/2)},$$

$$\hat{m}_{\oplus}(\hat{\boldsymbol{w}}^{\mathrm{BIC}}, x^{(M)}) := \arg\min_{\eta \in \mathbb{Y}} \sum_{m=1}^{M} \hat{w}_m^{\mathrm{BIC}} d^2(\hat{L}_{\oplus}^{(m)}(x^{(m)}), \eta), \ \hat{w}_m^{\mathrm{BIC}} = \frac{\exp(-\mathrm{BIC}_m/2)}{\sum_{j=1}^{M} \exp(-\mathrm{BIC}_j/2)},$$

respectively, where $d \in \{d_C, d_W\}$.

We evaluate each method using the out-of-sample prediction error. For each Monte Carlo replication, we generate $\{X_s, Y_s\}_{s=1}^{100}$ as out-of-sample observations. For the rth replication, the final prediction error is calculated as

$$\mathrm{FPE}(r) = \frac{1}{100} \sum_{s=1}^{100} d^2(Y_s, \hat{m}_{\oplus}(\hat{\boldsymbol{w}}, X_s)).$$

where $d \in \{d_C, d_W\}$ and $\hat{\boldsymbol{w}}$ is chosen by one of the three methods. Then we average the out-of-sample prediction error over $R = 200$ replications: $\mathrm{FPE} = R^{-1} \sum_{r=1}^{R} \mathrm{FPE}(r)$. We consider six predictors $(X_{i,1}, \ldots, X_{i,6})$ from X_i described in Section 4.1, and compute FPEs by the averaging methods (i)-(iii) for the following cases:

M1 : the model by $X_{i,1}$, M2 : average the models by $X_{i,1}$, $\{X_{i,k}\}_{k=1}^{2}$,
M3 : average the models by $X_{i,1}$, $\{X_{i,k}\}_{k=1}^{2}$, $\{X_{i,k}\}_{k=1}^{3}$,
M4 : average the models by $X_{i,1}$, $\{X_{i,k}\}_{k=1}^{2}$, $\{X_{i,k}\}_{k=1}^{3}$, $\{X_{i,k}\}_{k=1}^{4}$,
M5 : average the models by $X_{i,1}$, $\{X_{i,k}\}_{k=1}^{2}$, $\{X_{i,k}\}_{k=1}^{3}$, $\{X_{i,k}\}_{k=1}^{4}$, $\{X_{i,k}\}_{k=1}^{5}$,
M6 : average the models by $X_{i,1}$, $\{X_{i,k}\}_{k=1}^{2}$, $\{X_{i,k}\}_{k=1}^{3}$, $\{X_{i,k}\}_{k=1}^{4}$, $\{X_{i,k}\}_{k=1}^{5}$, $\{X_{i,k}\}_{k=1}^{6}$.

Figs. 1–2 present the FPEs for SPD matrices with the predictors generated by (i) and (ii), respectively. Our cross-validation weights $\hat{\boldsymbol{w}}$ outperform other averaging weights for all the cases. The improvements in terms of the values of the FPEs are larger for the case of correlated predictors in Fig. 1. When predictors are correlated, M6 shows the best performance in terms of FPE for both $n = 50$ and $n = 100$. This suggests that it is advantageous to average models with a greater number of predictors when making predictions with correlated predictors. On the other hand, when predictors are independent, M2 demonstrates the best performance in terms of FPE for $n = 50$, while M4 exhibits the best performance for $n = 100$. This implies that averaging simpler models may lead to better predictions when using independent or weakly correlated predictors. However, in this scenario as well, with an increase in sample size, there is a tendency for averaging more complex models to improve FPE.

Figs. 3–4 present the FPEs for univariate probability distributions with the predictors generated by (i) and (ii), respectively. One can find that the FPEs of our method are significantly smaller than other methods for all the cases. When predictors are correlated, M5 shows the best performance in terms of FPE for both $n = 50$ and $n = 100$. On the other hand, when predictors are independent, M2 demonstrates the best performance in terms of FPE for both $n = 50$ and $n = 100$.

5. Technical details

Proof of Theorem 1. (i) Define $\mathrm{diam}(\mathbb{Y}) = \sup_{\mu_1, \mu_2 \in \mathbb{Y}} d(\mu_1, \mu_2)$. First, we show the pointwise convergence:

$$d(\hat{m}_{\oplus}(\boldsymbol{w}, x^{(M)}), m_{\oplus}(\boldsymbol{w}, x^{(M)})) \xrightarrow{p} 0, \ \boldsymbol{w} \in \mathbb{W}, \ \|x^{(M)}\|_{\ell^2} \leq B. \tag{1}$$

Pick any $\boldsymbol{w} \in \mathbb{W}$ and $x^{(M)}$ with $\|x^{(M)}\|_{\ell^2} \leq B$. By Corollary 3.2.3 in van der Vaart and Wellner [13], it is sufficient for (1) to show

$$\sup_{\eta \in \mathbb{Y}} |\hat{R}(\boldsymbol{w}, x^{(M)}, \eta) - R(\boldsymbol{w}, x^{(M)}, \eta)| \xrightarrow{p} 0.$$

For this, we show that $\hat{R}(\boldsymbol{w}, x^{(M)}, \cdot)$ converges weakly to $R(\boldsymbol{w}, x^{(M)}, \cdot)$ in $\ell^{\infty}(\mathbb{Y})$, and then apply Theorem 1.3.6 in van der Vaart and Wellner [13]. By Theorem 1.5.4 in van der Vaart and Wellner [13], this weak convergence follows by showing that

(i) $\hat{R}(\boldsymbol{w}, x^{(M)}, \eta) - R(\boldsymbol{w}, x^{(M)}, \eta) \xrightarrow{p} 0$ for each $\eta \in \mathbb{Y}$.

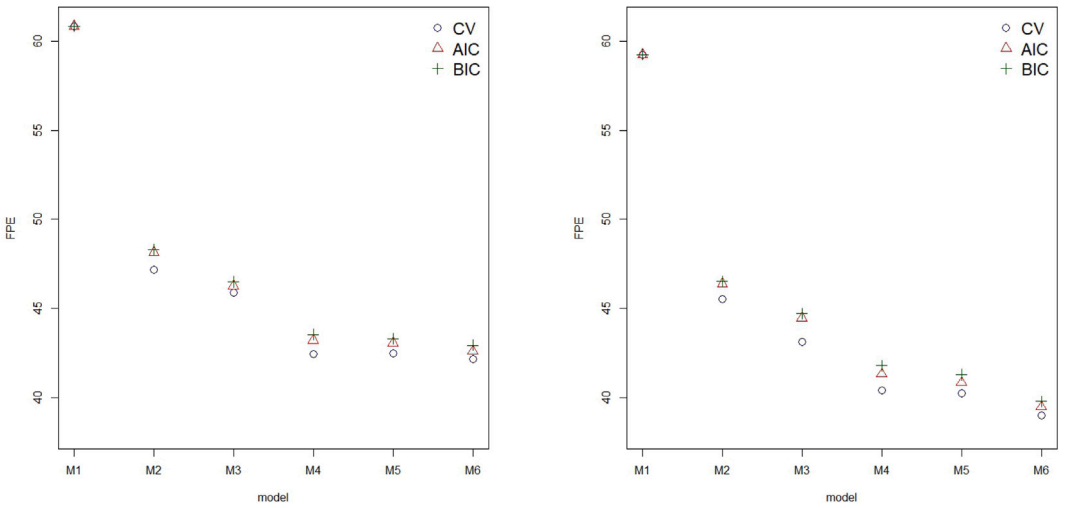


Fig. 1. FPE of CV, AIC, and BIC when $(\mathbb{Y}, d)$ is the space of SPD matrices with the Cholesky decomposition distance for $n = 50$ (left) and $n = 100$ (right) with correlated predictors. The circles correspond to the proposed cross-validation-based method (CV), the triangles to AIC-type model averaging (AIC), and the crosses to BIC-type model averaging (BIC). M1, M2, M3, M4, M5, and M6 correspond to the sets of models averaged.

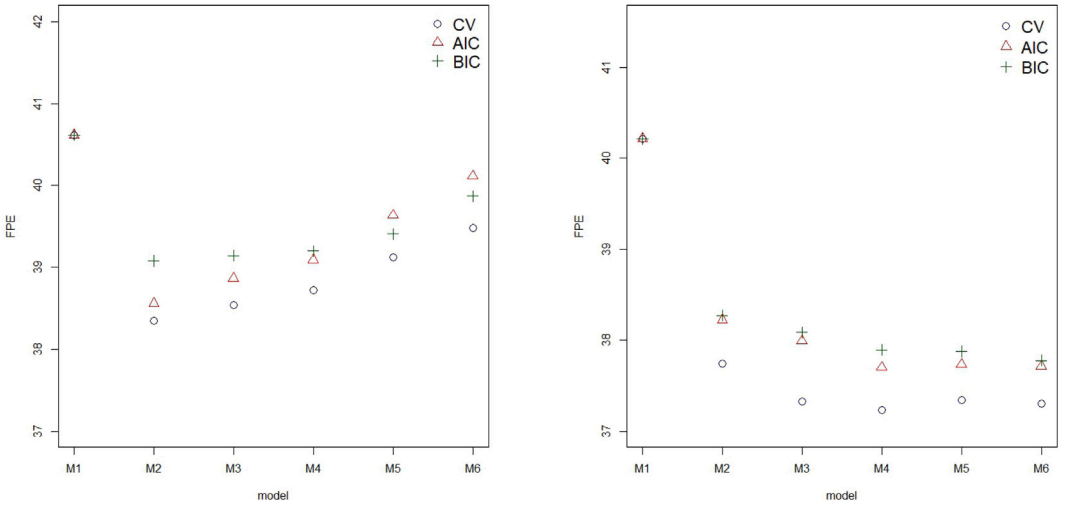


Fig. 2. FPE of CV, AIC, and BIC when $(\mathbb{Y}, d)$ is the space of SPD matrices with the Cholesky decomposition distance for $n = 50$ (left) and $n = 100$ (right) with independent predictors. The circles correspond to the proposed cross-validation-based method (CV), the triangles to AIC-type model averaging (AIC), and the crosses to BIC-type model averaging (BIC). M1, M2, M3, M4, M5, and M6 correspond to the sets of models averaged.

(ii) $\hat{R}(\boldsymbol{w}, x^{(M)}, \eta)$ is asymptotically equicontinuous in probability, i.e., for each $\varepsilon, \zeta > 0$, there exists $\delta > 0$ such that

$$\limsup_{n\to\infty} \mathbb{P}\left(\sup_{d(\eta_1,\eta_2)<\delta} |\hat{R}(\boldsymbol{w}, x^{(M)}, \eta_1) - \hat{R}(\boldsymbol{w}, x^{(M)}, \eta_2)| > \varepsilon \right) < \zeta.$$

Pick any $\eta \in \mathbb{Y}$. For (i), observe that

$$\begin{aligned}
|\hat{R}(\boldsymbol{w}, x^{(M)}, \eta) - R(\boldsymbol{w}, x^{(M)}, \eta)| &\le \sum_{m=1}^{M} w_m |\{d(\hat{L}_{\oplus}^{(m)}(x^{(m)}), \eta) + d(L_{\oplus}^{(m)}(x^{(m)}), \eta)\}\{d(\hat{L}_{\oplus}^{(m)}(x^{(m)}), \eta) - d(L_{\oplus}^{(m)}(x^{(m)}), \eta)\}| \\
&\le 2\mathrm{diam}(\mathbb{Y}) \sum_{m=1}^{M} w_m |d(\hat{L}_{\oplus}^{(m)}(x^{(m)}), \eta) - d(L_{\oplus}^{(m)}(x^{(m)}), \eta)| \\
&\le 2\mathrm{diam}(\mathbb{Y}) \max_{1\le m\le M} d(\hat{L}_{\oplus}^{(m)}(x^{(m)}), L_{\oplus}^{(m)}(x^{(m)})) \xrightarrow{p} 0.
\end{aligned}$$

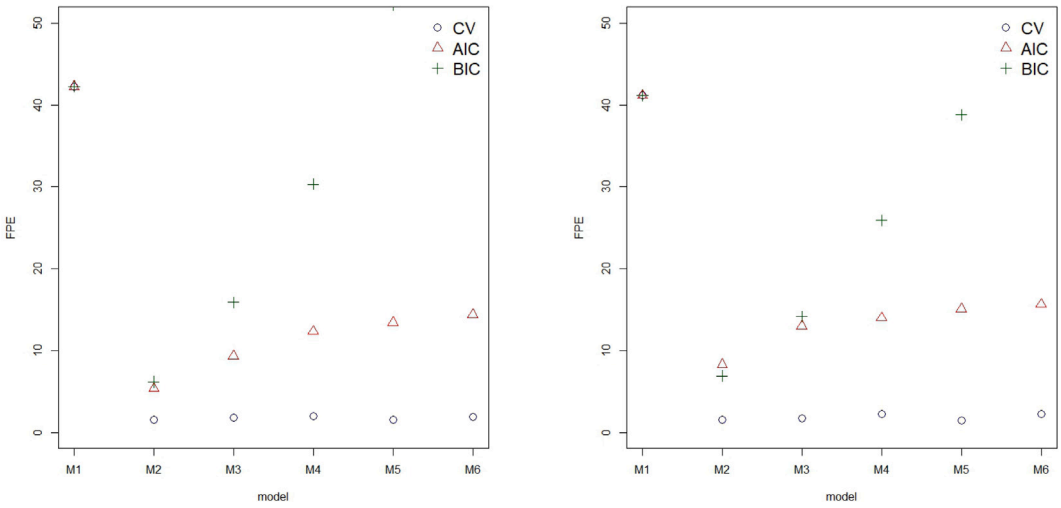


Fig. 3. FPE of CV, AIC, and BIC when $(\mathbb{Y}, d)$ is the space of univariate probability distributions with the Wasserstein metric for $n = 50$ (left) and $n = 100$ (right) with correlated predictors. The circles correspond to the proposed cross-validation-based method (CV), the triangles to AIC-type model averaging (AIC), and the crosses to BIC-type model averaging (BIC). M1, M2, M3, M4, M5, and M6 correspond to the sets of models averaged.

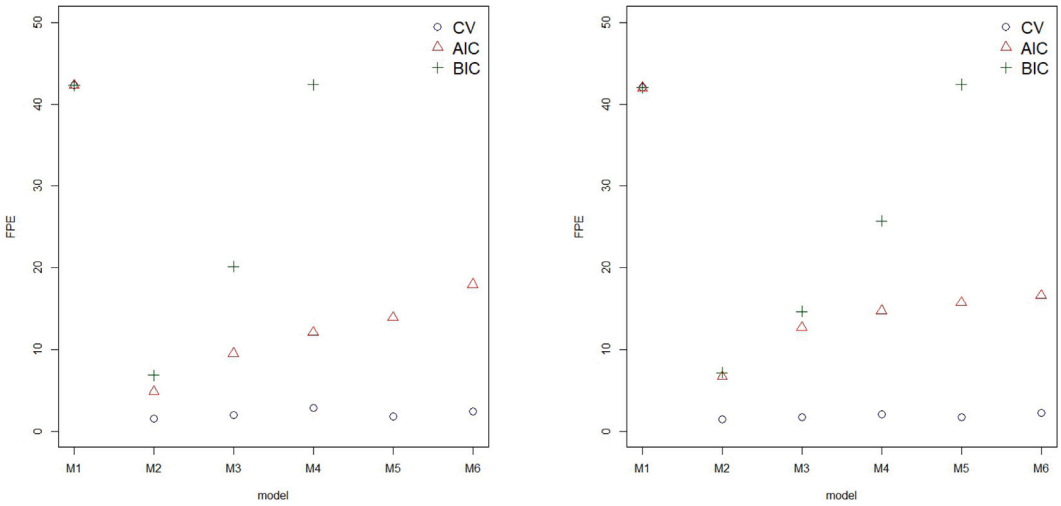


Fig. 4. FPE of CV, AIC, and BIC when $(\mathbb{Y}, d)$ is the space of univariate probability distributions with the Wasserstein metric for $n = 50$ (left) and $n = 100$ (right) with independent predictors. The circles correspond to the proposed cross-validation-based method (CV), the triangles to AIC-type model averaging (AIC), and the crosses to BIC-type model averaging (BIC). M1, M2, M3, M4, M5, and M6 correspond to the sets of models averaged.

where the first inequality follows from the triangle inequality, the second inequality follows from $d(\tilde{\eta}, \eta) \leq \mathrm{diam}(\mathbb{Y})$ for any $\tilde{\eta} \in \mathbb{Y}$, the third inequality follows from the triangle inequality and $\sum_{m=1}^{M} w_m = 1$, and the convergence follows from A1.

Pick any $\eta_1, \eta_2 \in \mathbb{Y}$. For (ii), a similar argument yields

$$\begin{aligned}
|\hat{R}(\boldsymbol{w}, x^{(M)}, \eta_1) - \hat{R}(\boldsymbol{w}, x^{(M)}, \eta_2)| &\leq \sum_{m=1}^{M} w_m |\{d(\hat{L}_{\oplus}^{(m)}(x^{(m)}), \eta_1) + d(\hat{L}_{\oplus}^{(m)}(x^{(m)}), \eta_2)\}\{d(\hat{L}_{\oplus}^{(m)}(x^{(m)}), \eta_1) - d(\hat{L}_{\oplus}^{(m)}(x^{(m)}), \eta_2)\}| \\
&\leq 2\mathrm{diam}(\mathbb{Y}) \sum_{m=1}^{M} w_m |d(\hat{L}_{\oplus}^{(m)}(x^{(m)}), \eta_1) - d(\hat{L}_{\oplus}^{(m)}(x^{(m)}), \eta_2)| \\
&\leq 2\mathrm{diam}(\mathbb{Y}) d(\eta_1, \eta_2),
\end{aligned}$$

which implies $\sup_{d(\eta_1,\eta_2)<\delta} |\hat{R}(\boldsymbol{w}, x^{(M)}, \eta_1) - \hat{R}(\boldsymbol{w}, x^{(M)}, \eta_2)| = O_p(\delta)$ so that we obtain (ii). Therefore, we obtain (1).

Next, we show the uniform convergence. Consider the process $Z_n(\boldsymbol{w}, x^{(M)}) = d(\hat{m}_{\oplus}(\boldsymbol{w}, x^{(M)}), m_{\oplus}(\boldsymbol{w}, x^{(M)}))$. By (1), we have $Z_n(\boldsymbol{w}, x^{(M)}) \xrightarrow{p} 0$ for each $\boldsymbol{w} \in \mathbb{W}$ and $\|x^{(M)}\|_{\ell^2} \leq B$. By Theorem 1.5.4 in van der Vaart and Wellner [13], it is sufficient to

show that for each $S > 0$,

$$\limsup_{n\to\infty} \mathbb{P}\left(\sup_{\substack{\boldsymbol{w}_1,\boldsymbol{w}_2\in\mathbb{W},\|\boldsymbol{w}_1-\boldsymbol{w}_2\|_{\ell^1}<\delta,\\ \|x_1^{(M)}\|_{\ell^2}\le B,\|x_1^{(M)}\|_{\ell^2}\le B,\|x_1^{(M)}-x_2^{(M)}\|_{\ell^2}<\delta}} |Z_n(\boldsymbol{w}_1,x_1^{(M)}) - Z_n(\boldsymbol{w}_2,x_2^{(M)})| > 2S \right) \to 0, \tag{2}$$

as $\delta \to 0$. Since

$$|Z_n(\boldsymbol{w}_1,x_1^{(M)}) - Z_n(\boldsymbol{w}_2,x_2^{(M)})| \le d(m_\oplus(\boldsymbol{w}_1,x_1^{(M)}), m_\oplus(\boldsymbol{w}_2,x_2^{(M)})) + d(\hat{m}_\oplus(\boldsymbol{w}_1,x_1^{(M)}), \hat{m}_\oplus(\boldsymbol{w}_2,x_2^{(M)}))$$

by the triangle inequality, it is sufficient for (2) to show that $m_\oplus(\cdot,\cdot)$ is uniformly continuous over $\boldsymbol{w} \in \mathbb{W}$ and $\|x^{(M)}\|_{\ell^2} \le B$ and that

$$\limsup_{n\to\infty} \mathbb{P}\left(\sup_{\substack{\boldsymbol{w}_1,\boldsymbol{w}_2\in\mathbb{W},\|\boldsymbol{w}_1-\boldsymbol{w}_2\|_{\ell^1}<\delta,\\ \|x_1^{(M)}\|_{\ell^2}\le B,\|x_1^{(M)}\|_{\ell^2}\le B,\|x_1^{(M)}-x_2^{(M)}\|_{\ell^2}<\delta}} d(\hat{m}_\oplus(\boldsymbol{w}_1,x_1^{(M)}), \hat{m}_\oplus(\boldsymbol{w}_2,x_2^{(M)})) > S \right) \to 0, \tag{3}$$

as $\delta \to 0$.

Now, pick any $\delta > 0$ and then pick any $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{W}$ with $\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_{\ell^1} < \delta$, and $x_1^{(M)}, x_2^{(M)} \in \mathbb{R}^M$ with $\|x_1^{(M)} - x_2^{(M)}\|_{\ell^2} < \delta$. Note that A1 guarantees uniformly continuity of $L_\oplus^{(m)}(x^{(m)})$ over $\|x^{(M)}\|_{\ell^2} \le B$ for $m \in \{1,\dots,M\}$. Then due to the form of $R(\boldsymbol{w},x^{(M)},\eta)$, we have

$$\begin{aligned}
\zeta &< \sup_{\eta\in\mathbb{Y}} |R(\boldsymbol{w}_1,x_1^{(M)},\eta) - R(\boldsymbol{w}_2,x_2^{(M)},\eta)| \le \max\{\mathrm{diam}(\mathbb{Y}),2\}^2 \left\{ \|\boldsymbol{w}_1-\boldsymbol{w}_2\|_{\ell^1} + \max_{1\le m\le M} d(L_\oplus^{(m)}(x_1^{(m)}), L_\oplus^{(m)}(x_2^{(m)})) \right\} \\
&\le 2(1+C)\max\{\mathrm{diam}(\mathbb{Y}),2\}^2 (O(\delta)+o(1)),\ \ \delta\to 0,
\end{aligned}$$

for some constant $C > 0$. Thus, A2 implies that $m_\oplus$ is continuous at $(\boldsymbol{w},x)$ and thus uniformly continuous over $(\boldsymbol{w},x^{(M)}) \in \mathbb{W}\times\{x^{(M)} : \|x^{(M)}\|_{\ell^2} \le B\}$. To show (3), pick any $\varepsilon > 0$, and suppose $d(\hat{m}_\oplus(\boldsymbol{w}_1,x_1^{(M)}), \hat{m}_\oplus(\boldsymbol{w}_2,x_2^{(M)})) > \varepsilon$ with $\boldsymbol{w}_1,\boldsymbol{w}_2 \in \mathbb{W}$ and $\|x_1^{(M)}\|_{\ell^2}, \|x_2^{(M)}\|_{\ell^2} \le B$. Then A2 and the form of $\hat{R}(\boldsymbol{w},x^{(M)},\eta)$ imply that

$$\begin{aligned}
\zeta &\le \sup_{\substack{\boldsymbol{w}_1,\boldsymbol{w}_2\in\mathbb{W},\|\boldsymbol{w}_1-\boldsymbol{w}_2\|_{\ell^1}<\delta,\\ \|x_1^{(M)}\|_{\ell^2}\le B,\|x_1^{(M)}\|_{\ell^2}\le B,\|x_1^{(M)}-x_2^{(M)}\|_{\ell^2}<\delta}} |\hat{R}(\boldsymbol{w}_1,x_1^{(M)},\eta) - \hat{R}(\boldsymbol{w}_2,x_2^{(M)},\eta)| \\
&\le \max\{\mathrm{diam}(\mathbb{Y}),2\}^2 \sup_{\substack{\boldsymbol{w}_1,\boldsymbol{w}_2\in\mathbb{W},\|\boldsymbol{w}_1-\boldsymbol{w}_2\|_{\ell^1}<\delta,\\ \|x_1^{(M)}\|_{\ell^2}\le B,\|x_1^{(M)}\|_{\ell^2}\le B,\|x_1^{(M)}-x_2^{(M)}\|_{\ell^2}<\delta}} \left\{ \|\boldsymbol{w}_1-\boldsymbol{w}_2\|_{\ell^1} + \max_{1\le m\le M} d(\hat{L}_\oplus^{(m)}(x_1^{(m)}), \hat{L}_\oplus^{(m)}(x_2^{(m)})) \right\} \\
&\le \max\{\mathrm{diam}(\mathbb{Y}),2\}^2 \sup_{\substack{\boldsymbol{w}_1,\boldsymbol{w}_2\in\mathbb{W},\|\boldsymbol{w}_1-\boldsymbol{w}_2\|_{\ell^1}<\delta,\\ \|x_1^{(M)}\|_{\ell^2}\le B,\|x_1^{(M)}\|_{\ell^2}\le B,\|x_1^{(M)}-x_2^{(M)}\|_{\ell^2}<\delta}} \left\{ \|\boldsymbol{w}_1-\boldsymbol{w}_2\|_{\ell^1} + \max_{1\le m\le M} d(L_\oplus^{(m)}(x_1^{(m)}), L_\oplus^{(m)}(x_2^{(m)})) \right\} + o_p(1) = O(\delta) + o_p(1),
\end{aligned}$$

as $\delta \to 0$, where the second inequality follows from the triangle inequality, the third inequality follows from the uniform convergence of $\hat{L}_\oplus^{(m)}(x^{(m)})$ in A1, and the equality follows from uniform continuity of $L_\oplus^{(m)}(x^{(m)})$ over $\|x^{(M)}\|_{\ell^2} \le B$. Therefore, we obtain (3) and the conclusion of the theorem follows. □

Proof of Theorem 1. (ii) First, we show

$$\sup_{\boldsymbol{w}\in\mathbb{W}} |\mathrm{CV}_n(\boldsymbol{w}) - \mathrm{FPE}_n(\boldsymbol{w})| \xrightarrow{p} 0. \tag{4}$$

Decompose

$$\begin{aligned}
\mathrm{CV}_n(\boldsymbol{w}) - \mathrm{FPE}_n(\boldsymbol{w}) &= \frac{1}{n}\sum_{i=1}^n \{d^2(Y_i,\hat{m}_{\oplus,-i}(\boldsymbol{w},X_i)) - d^2(Y_i,m_\oplus(\boldsymbol{w},X_i))\} \\
&\quad + \frac{1}{n}\sum_{i=1}^n \{d^2(Y_i,m_\oplus(\boldsymbol{w},X_i)) - \mathbb{E}[d^2(Y,m_\oplus(\boldsymbol{w},X))]\} + \{\mathbb{E}[d^2(Y,m_\oplus(\boldsymbol{w},X))] - \mathrm{FPE}_n(\boldsymbol{w})\} \\
&=: T_1(\boldsymbol{w}) + T_2(\boldsymbol{w}) + T_3(\boldsymbol{w}).
\end{aligned}$$

For $T_1(\boldsymbol{w})$, Theorem 1(i) implies

$$\sup_{\boldsymbol{w}\in\mathbb{W}} |T_1(\boldsymbol{w})| \le 2\mathrm{diam}(\mathbb{Y}) \sup_{\boldsymbol{w}\in\mathbb{W},\|x^{(M)}\|_{\ell^2}\le B} d(\hat{m}_\oplus(\boldsymbol{w},x^{(M)}), m_\oplus(\boldsymbol{w},x^{(M)})) \xrightarrow{p} 0. \tag{5}$$

For $T_2(\boldsymbol{w})$, we show

$$\sup_{\boldsymbol{w}\in\mathbb{W}} \left| \frac{1}{n}\sum_{i=1}^n \{d^2(Y_i,m_\oplus(\boldsymbol{w},X_i)) - \mathbb{E}[d^2(Y,m_\oplus(\boldsymbol{w},X))]\} \right| = O_p(n^{-1/2}). \tag{6}$$

Define $h_{\boldsymbol{w}}(y, x^{(M)}) = d^2(y, m_\oplus(\boldsymbol{w}, x^{(M)}))$ and $\mathcal{F}_{\mathbb{W}} = \{h_{\boldsymbol{w}}(y, x^{(M)}) : \boldsymbol{w} \in \mathbb{W}\}$. An envelop function of $\mathcal{F}_{\mathbb{W}}$ is $F_{\mathbb{W}} = \mathrm{diam}(\mathbb{Y})^2$. By A3, we have

$$\begin{aligned}|h_{\boldsymbol{w}_1}(y, x^{(M)}) - h_{\boldsymbol{w}_2}(y, x^{(M)})| &\le |d(y, m_\oplus(\boldsymbol{w}_1, x^{(M)})) + d(y, m_\oplus(\boldsymbol{w}_2, x^{(M)}))||d(y, m_\oplus(\boldsymbol{w}_1, x^{(M)})) - d(y, m_\oplus(\boldsymbol{w}_2, x^{(M)}))| \\ &\le 2\bar{D}_B \mathrm{diam}(\mathbb{Y})\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_{\ell_1}^{\beta_B}.\end{aligned}$$

Thus, from Theorems 2.14.2 and 2.7.11 in van der Vaart and Wellner [13], it holds

$$\begin{aligned}\mathbb{E}\left[\sup_{\boldsymbol{w}\in\mathbb{W}}\left|\frac{1}{n}\sum_{i=1}^n\{d^2(Y_i, m_\oplus(\boldsymbol{w}, X_i)) - \mathbb{E}[d^2(Y, m_\oplus(\boldsymbol{w}, X))]\}\right|\right] &\lesssim \frac{1}{\sqrt{n}}\int_0^1\sqrt{1 + \ln N_{[]}(2\varepsilon\bar{D}_B\mathrm{diam}(\mathbb{Y}), \mathcal{F}_{\mathbb{W}}, \|\cdot\|)}d\varepsilon \\ &\le \frac{1}{\sqrt{n}}\int_0^1\sqrt{1 + \ln N(\varepsilon, \mathbb{W}, \|\cdot\|_{\ell_1})}d\varepsilon \lesssim \frac{1}{\sqrt{n}}\int_0^1\sqrt{1 + \ln(\varepsilon^{-M/\beta_B})}d\varepsilon \\ &\lesssim \frac{1}{\sqrt{n}}\left(1 + \sqrt{M}\int_0^1\sqrt{-\ln\varepsilon}d\varepsilon\right) = O(n^{-1/2}),\end{aligned}$$

where $N_{[]}(\varepsilon, \mathcal{F}_{\mathbb{W}}, \|\cdot\|)$ is the ε-bracketing number of $\mathcal{F}_{\mathbb{W}}$ with respect to any norm $\|\cdot\|$ and $N(\varepsilon, \mathbb{W}, \|\cdot\|_{\ell_1})$ denote the ε-covering number of $\mathbb{W}$ with respect to the norm $\|\cdot\|_{\ell_1}$. This yields (6).

For $T_3(\boldsymbol{w})$, a similar argument to (5) yields

$$\sup_{\boldsymbol{w}\in\mathbb{W}}|\mathbb{E}[d^2(Y, m_\oplus(\boldsymbol{w}, X))] - \mathrm{FPE}_n(\boldsymbol{w})| \le 2\mathrm{diam}(\mathbb{Y})\sup_{\boldsymbol{w}\in\mathbb{W}, \|x^{(M)}\|_E\le B} d(\hat{m}_\oplus(\boldsymbol{w}, x^{(M)}), m_\oplus(\boldsymbol{w}, x^{(M)})) \xrightarrow{p} 0, \tag{7}$$

where the convergence follows from Theorem 1(i). Combining (5)–(7), we obtain (4).

Next, we show

$$\mathrm{FPE}_n(\hat{\boldsymbol{w}}) = \inf_{\boldsymbol{w}\in\mathbb{W}}\mathbb{E}[d^2(Y, m_\oplus(\boldsymbol{w}, X))] + o_p(1). \tag{8}$$

Observe that

$$\begin{aligned}\mathrm{FPE}_n(\hat{\boldsymbol{w}}) &= \mathrm{CV}_n(\hat{\boldsymbol{w}}) + o_p(1) = \inf_{\boldsymbol{w}\in\mathbb{W}}\mathrm{CV}_n(\boldsymbol{w}) + o_p(1) \\ &= \inf_{\boldsymbol{w}\in\mathbb{W}}\mathrm{FPE}_n(\boldsymbol{w}) + o_p(1) = \inf_{\boldsymbol{w}\in\mathbb{W}}\mathbb{E}[d^2(Y, m_\oplus(\boldsymbol{w}, X))] + o_p(1),\end{aligned}$$

where the first and third equalities follow from (4), the second equality follows from the definition of $\hat{\boldsymbol{w}}$, and the last equality follows from (7). Therefore, we obtain (8).

Finally, we complete the proof. From (7), we have

$$\inf_{\boldsymbol{w}\in\mathbb{W}}\mathrm{FPE}_n(\boldsymbol{w}) = \inf_{\boldsymbol{w}\in\mathbb{W}}\mathbb{E}[d^2(Y, m_\oplus(\boldsymbol{w}, X))] + o_p(1). \tag{9}$$

Combining (8), (9), and A4, we obtain the conclusion. □

CRediT authorship contribution statement

Daisuke Kurisu: Conceptualization, Methodology, Writing – original draft. **Taisuke Otsu:** Writing – review & editing.

Acknowledgments

We thank the Editor, Associate Editor and a referee for their constructive comments which led to the improvements of the paper. D. Kurisu is partially supported by JSPS KAKENHI Grant Number 23K12456 . We thank Hans-Georg Müller, Satarupa Bhattacharjee, Yoshiyuki Ninomiya, Ryo Okui and Keisuke Yano for their constructive comments and discussion. We also thank the participants of the seminars at The University of California, Davis, The Institute of Statistical Mathematics, Kyoto University and Osaka University.

References

[1] H. Akaike, Statistical predictor identification, Ann. Inst. Statist. Math. 22 (1) (1970) 203–217.
[2] S. Bhattacharjee, H.-G. Müller, Single index Fréchet regression, Ann. Statist. 51 (4) (2023) 1770–1798.
[3] C.-K. Chu, J.S. Marron, Comparison of two bandwidth selectors with dependent errors, Ann. Statist. 19 (4) (1991) 1906–1918.
[4] G. Claeskens, N.L. Hjort, Model Selection and Model Averaging, Cambridge University Press, 2008.
[5] M. Fréchet, Les éléments aléatoires de nature quelconque dans un espace distancié, Ann. Inst. Henri Poincare 10 (1948) 215–310.
[6] A. Ghosal, W. Meiring, A. Petersen, Fréchet single index models for object response regression, Electron. J. Stat. 17 (1) (2023) 1074–1112.
[7] B.E. Hansen, Least squares model averaging, Econometrica 75 (4) (2007) 1175–1189.
[8] K.-C. Li, Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set, Ann. Statist. 15 (3) (1987) 958–975.
[9] J.S. Marron, A.M. Alonso, Overview of object oriented data analysis, Biom. J. 56 (5) (2014) 732–753.
[10] V. Patrangenaru, L. Ellingson, Nonparametric Statistics on Manifolds and Their Applications to Object Data Analysis, CRC Press, Taylor & Francis Group Boca Raton, 2015.
[11] A. Petersen, H.-G. Müller, Fréchet regression for random objects with Euclidean predictors, Ann. Statist. 47 (2) (2019) 691–719.
[12] D.C. Tucker, Y. Wu, H.-G. Müller, Variable selection for global Fréchet regression, J. Amer. Statist. Assoc. 118 (542) (2023) 1023–1037.
[13] A.W. van der Vaart, J.A. Wellner, Weak Convergence and Empirical Processes, Springer, 2023.
[14] C. Ying, Z. Yu, Fréchet sufficient dimension reduction for random objects, Biometrika 109 (4) (2022) 975–992.