MODEL AVERAGING FOR GLOBAL FRÉCHET REGRESSION

DAISUKE KURISU AND TAISUKE OTSU

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 long history and is widely applied in statistics literature. This paper generalizes the notion of model averaging for global Fréchet regressions and establishes 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, 2014, for an overview). A fundamental object to describe distributions of non-Euclidean random objects is the so-called Fréchet mean (Fréchet, 1948), 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, 2015, for a survey). Recently, in a seminal paper, Petersen and Müller (2019) 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, 2008, for a survey). Indeed Tucker, Wu and Müller (2021) developed a model selection method for global Fréchet regressions by extending the ridge selection operator to the present context, and established its selection consistency.[1] This paper addresses another open question, model averaging of regression models for non-Euclidean response objects.

In this paper, we generalize the notion of model averaging for global Fréchet regressions and establishes an optimal out-of-sample prediction property of the cross validation to select the averaging weights in terms of the final prediction error (Akaike, 1970). First of all, it is not trivial how to conduct model averaging for global Fréchet regressions which reside in non-Euclidean spaces. By adapting 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, to the best of our knowledge, this is the first paper that builds and studies the notions of the final prediction

Our research is supported by JSPS KAKENHI (JP23K12456) (Kurisu).

[1]See also Ying and Yu (2022) for sufficient dimension reduction on non-Euclidean random objects using Euclidean predictors.

error for out-of-sample predictions and cross validation for regression analyses on non-Euclidean random objects. In contrast to Tucker, Wu and Müller (2021) 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.

2. Model averaging estimator

Let (Ω, d) be a totally bounded metric space. We consider a random process $(X, Y) \sim F$, where X and Y take values in $\mathbb{R}^p$ and Ω, respectively, and F is the joint distribution of (X, Y) on $\mathbb{R}^p \times \Omega$. We are concerned with the situation where Y is a complex random object so that the space Ω 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 $\omega_\oplus = \arg\min_{\omega\in\Omega} \mathbb{E}[d^2(Y, \omega)]$, and there is rich literature on statistical inference for $\omega_\oplus$.

In a seminal paper, Petersen and Müller (2019) extended the notion of the Fréchet mean to regression problems and proposed the Fréchet regression function $\omega_\oplus(x) = \arg\min_{\omega\in\Omega} \mathbb{E}[d^2(Y, \omega)|X = x]$. Furthermore, Petersen and Müller (2019) generalized the idea of global least squares regression and developed the global Fréchet regression:

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

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

We now introduce our setup for model averaging of global Fréchet regressions. Hereafter $x = (x_1, x_2, \ldots)'$ takes values in $\mathbb{R}^\infty$. Let $X^{(m)} = (X_1, X_2, \ldots, X_{k_m})' \in \mathbb{R}^{k_m}$ $(m = 1, \ldots, M)$ be a nested sequence of predictors for $0 \le k_1 < k_2 < \ldots. < k_M$, $x^{(m)} = (x_1, x_2, \ldots, x_{k_m})' \in \mathbb{R}^{k_m}$, and for $m = 1, \ldots, M$,

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

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. In order to build the notion of model averaging for the global Fréchet regressions $\{L_\oplus^{(m)}(x)\}_{m=1}^M$, we note that in the d-dimensional Euclidean space, the weighted average $\bar{l}_{\boldsymbol{w}} = \sum_{m=1}^M w_m l^{(m)}$ of points $l^{(m)} \in \mathbb{R}^d$ can be defined as

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

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

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

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

$$\begin{aligned}\hat{L}_\oplus^{(m)}(x) &= \arg\min_{\omega\in\Omega} \frac{1}{n}\sum_{i=1}^n \{1 + (x^{(m)} - \bar{X}^{(m)})'(\hat{\Sigma}^{(m)})^{-1}(X_i^{(m)} - \bar{X}^{(m)})\} d^2(Y_i, \omega),\\ \hat{m}_\oplus(\boldsymbol{w}, x) &= \arg\min_{\omega\in\Omega} \sum_{m=1}^M w_m d^2(\hat{L}_\oplus^{(m)}(x), \omega),\end{aligned}$$

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

As a criterion to evaluate model averaging weights, we extend the notion of the final prediction error (Akaike, 1970) 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 $(X_i^{(M)}, Y_i)$. 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, Wu and Müller (2021), which focuses on 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 $\hat{m}_{\oplus,-i}(\boldsymbol{w}, x) = \arg\min_{\omega\in\Omega} \sum_{m=1}^M w_m d^2(\hat{L}_{\oplus,-i}^{(m)}(x), \omega)$, and $\hat{L}_{\oplus,-i}^{(m)}(x)$ is defined as $\hat{L}_\oplus^{(m)}(x)$ with the i-th observation deleted. Letting $\mathbb{W} = \{\boldsymbol{w} = (w_1, \ldots, w_M)' \in [0,1]^M : \sum_{m=1}^M w_m = 1\}$, our model averaging estimator for global Fréchet regressions is defined as

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

3. Optimality

We now present our main result, optimality of the model averaging estimator $\hat{m}_\oplus(\hat{\boldsymbol{w}}, x)$ in terms of the final prediction error. Let $\|x\|_{\ell^2} = \left(\sum_{j=1}^\infty x_j^2\right)^{1/2}$ be a norm of the ℓ^2 space, $\|z\|_{\ell^1} = \sum_{m=1}^M |z_m|$ for $z \in \mathbb{R}^M$, and

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

We impose the following assumptions.

Assumption.

(1) (Ω, d) *is a totally bounded metric space,* $\mathbb{P}(\|X^{(M)}\|_{\ell^2} \le B) = 1$ *for some* $B > 0$, $L_\oplus^{(m)}(x)$ *is continuous at* x *with* $\|x^{(M)}\|_{\ell^2} \le B$, *and the global Fréchet regression estimators*

$\{\hat{L}_{\oplus}^{(m)}(x)\}_{m=1}^{M}$ *are uniformly consistent in the sense that*

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

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

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

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

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

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

Assumption (1) contains conditions 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 can be found in Petersen and Müller (2019, Theorem 1). Assumptions (2) is an additional condition to guarantee uniform consistency of the model averaging estimator $\hat{m}_{\oplus}(\boldsymbol{w},x)$, which is an analog of Petersen and Müller (2019, Condition (U0)) and is commonly imposed to derive the consistency of M-estimators (see, e.g., van der Vaart and Wellner, 1996). Assumptions (3)-(4) are additional conditions to establish the asymptotic optimality of our model averaging estimator $\hat{m}_{\oplus}(\hat{\boldsymbol{w}},x)$ using the cross validation. Assumption (3) is a Lispschitz-type condition for weights to derive uniform convergence of $\frac{1}{n}\sum_{i=1}^{n} d^2(Y_i, m_{\oplus}(\boldsymbol{w},X_i))$. Assumption (4) says that all the global Fréchet regressions and their averaged versions are misspecified so that it is natural to evaluate model averaging weights by out-of-sample predictions.

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

Theorem.

(1) *Under Assumptions 1-2, it holds*

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

(2) *Under Assumptions 1-4, it holds*

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

Theorem (1) shows uniform consistency of the model averaging estimator $\hat{m}_{\oplus}(\boldsymbol{w},x)$ over the weights $\boldsymbol{w}$ and values of predictors x. Theorem (2) 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 $w\in\mathbb{W}$.

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

Example 1. [Symmetric positive-definite matrices with the Frobenius norm] Let Ω be the set of symmetric positive-definite matrices with the Frobenius norm. For this example, Petersen

and Müller (2019, Proposition 2 and Theorem 1) guarantee Assumption (1). Let $L_{\oplus}^{(m)}(x)$ be the global Fréchet regression function of the m-th model. The model average global Fréchet regression function $m_{\oplus}(\boldsymbol{w}, x)$ is given by $m_{\oplus}(\boldsymbol{w}, x) = \sum_{m=1}^{M} w_m L_{\oplus}^{(m)}(x)$. Applying a similar argument in the proof of Petersen and Müller (2019, Proposition 2), one can see that Assumptions (2) and (3) are satisfied with $\beta_B = 1$.

Example 2. [Functional data with L_2 metric] Let $\Omega = \left\{ f : [0,1] \mapsto \mathbb{R}, \int_0^1 f^2(t)dt < \infty \right\}$ equipped with the L_2 metric d_{L_2} defined as

$$d_{L_2}(f, g) = \sqrt{\int_0^1 (f(t) - g(t))^2 dt},$$

for any $f, g \in \Omega$. For this example, Petersen and Müller (2019, Corollary2) guarantee Assumption (1). Let $L_{\oplus}^{(m)}(x)$ be the global Fréchet regression function of the m-th model. The model average global Fréchet regression function $m_{\oplus}(\boldsymbol{w}, x)$ is given by $m_{\oplus}(\boldsymbol{w}, x) = \sum_{m=1}^{M} w_m L_{\oplus}^{(m)}(x)$ and Assumptions (2) and (3) are satisfied with $\beta_B = 1$.

Example 3. [Probability distributions with Wasserstein metric] Let Ω be the set of probability distributions F on $\mathbb{R}$ such that $\int_{\mathbb{R}} x^2 dF(x) < \infty$ 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, Petersen and Müller (2019, Proposition 1 and Theorem 1) guarantee Assumption (1). Let $L_{\oplus}^{(m)}(x)$ be the global Fréchet regression function of the m-th model, which is a distribution function on $\mathbb{R}$, and let $L_{\oplus}^{(m)-1}(x)$ be the quantile function of $L_{\oplus}^{(m)}(x)$. The quantile function of the model average global Fréchet regression function $m_{\oplus}^{-1}(\boldsymbol{w}, x)$ is given by $m_{\oplus}^{-1}(\boldsymbol{w}, x) = \sum_{m=1}^{M} w_m L_{\oplus}^{(m)-1}(x)$. Applying a similar argument in the proof of Petersen and Müller (2019, Proposition 1), one can see that Assumptions (2) and (3) are satisfied with $\beta_B = 1$.

Example 4. [Bounded Riemann manifold with geodesic distance] This example considers spherical data. Let $\Omega = \mathbb{S}^2$, 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 (2019) and Tucker, Wu and Müller (2021) considered the following Fréchet regression model. Let $\omega_{\oplus}(x) \in \mathbb{S}^2$ be a regression function and V be a random vector on the tangent space $T_{\omega_{\oplus}(X)}$. Define Y as an exponential map of V at $\omega_{\oplus}(X)$, i.e.,

$$Y = \mathrm{Exp}_{\omega_{\oplus}(X)}(V) = \cos(\|V\|_E)\omega_{\oplus}(X) + \sin(\|V\|_E)\frac{V}{\|V\|_E},$$

where $\|\cdot\|_E$ is the Euclidean norm. Petersen and Müller (2019, Proposition 3) gives sufficient conditions of Assumption (1).

4. Simulation

4.1. **Data generating process.** We consider the set of symmetric positive-definite (SPD) matrices as Ω. For SPD matrices A_1 and A_2, the Cholesky decomposition yields $A_1 = (A_1^{1/2})' A_1^{1/2}$

and $A_2 = (A_2^{1/2})' A_2^{1/2}$, 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

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

For predictors $X_i = (X_{i,1}, \ldots, X_{i,p})'$, we consider two kinds of designs. (i) Generate p-dimensional multivariate Gaussian random variables $Z_i = (Z_{i,1}, \ldots, Z_{i,p})'$ 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 $[0, 2]$. We set the random object Y_i as $T \times T$ SPD matrix and consider the following Fréchet regression function: $\omega_\oplus(x) = \mathbb{E}[Y|X = x] = \mathbb{E}[A]'\mathbb{E}[A]$, where

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

with the $T \times T$ identity matrix I_T and a $T \times T$ matrix $V = (I_{\{j<k\}})$. Conditional on $X = (X_1, \ldots, X_9)'$, the random response Y is generated by $Y = A'A$, where $A = (\mu + \sigma)I_T + \sigma V$ with

$$\begin{aligned}\mu|X &\sim N\left(\mu_0 + \beta\left(X_1 + \frac{X_3}{3} + \frac{X_5}{5} + \frac{X_7}{7} + \frac{X_9}{9}\right), \nu_1\right), \\ \sigma|X &\sim \mathrm{Gamma}\left(\nu_2^{-1}\left(\sigma_0 + \gamma\left(\frac{X_2}{2} + \frac{X_4}{4} + \frac{X_6}{6} + \frac{X_8}{8}\right)\right)^2, \frac{\nu_2}{\sigma_0 + \gamma\left(\frac{X_2}{2} + \frac{X_4}{4} + \frac{X_6}{6} + \frac{X_8}{8}\right)}\right).\end{aligned}$$

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

Let $L_\oplus^{(m)}(x)$ be the global Fréchet regression function of the m-th model and let $(L_\oplus^{(m)1/2}(x))' L_\oplus^{(m)1/2}(x)$ be the Cholesky decomposition of $L_\oplus^{(m)}(x)$. In this case, the model average global Fréchet regression function $m_\oplus(\boldsymbol{w}, x)$ is given by

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

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.

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

$$\begin{aligned}\mathrm{AIC}_m &= n \log\left(\frac{1}{n}\sum_{i=1}^n d_C^2(Y_i, \hat{L}_\oplus^{(m)}(X_i))\right) + 2k_m, \\ \mathrm{BIC}_m &= n \log\left(\frac{1}{n}\sum_{i=1}^n d_C^2(Y_i, \hat{L}_\oplus^{(m)}(X_i))\right) + k_m \log n.\end{aligned}$$

Then the AIC- and BIC-type model average estimators are defined as

$$\hat{m}_{\oplus}(\hat{\boldsymbol{w}}^{\mathrm{AIC}}, x) = \arg\min_{\omega\in\Omega} \sum_{m=1}^{M} \hat{w}_m^{\mathrm{AIC}} d_C^2(\hat{L}_{\oplus}^{(m)}(x), \omega) \text{ with } \hat{w}_m^{\mathrm{AIC}} = \frac{\exp(-\mathrm{AIC}_m/2)}{\sum_{j=1}^{M}\exp(-\mathrm{AIC}_j/2)},$$

$$\hat{m}_{\oplus}(\hat{\boldsymbol{w}}^{\mathrm{BIC}}, x) := \arg\min_{\omega\in\Omega} \sum_{m=1}^{M} \hat{w}_m^{\mathrm{BIC}} d_C^2(\hat{L}_{\oplus}^{(m)}(x), \omega) \text{ with } \hat{w}_m^{\mathrm{BIC}} = \frac{\exp(-\mathrm{BIC}_m/2)}{\sum_{j=1}^{M}\exp(-\mathrm{BIC}_j/2)},$$

respectively.

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 r-th replication, the final prediction error is calculated as

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

where $\hat{\boldsymbol{w}}$ is chosen by one of the three methods. Then we average the out-of-sample prediction error over $R = 200$ replications: $\mathrm{FPE} = \frac{1}{R}\sum_{r=1}^{R}\mathrm{FPE}(r)$. We consider 5 nested models M_k that use predictors $\{X_{i,1}, \ldots, X_{i,k}\}$ for $k = 1, \ldots, 5$ and compute FPEs of 4 model averaging estimators that use $\{M_1, \ldots, M_k\}$ for $k = 2, \ldots, 5$ for the three averaging methods.

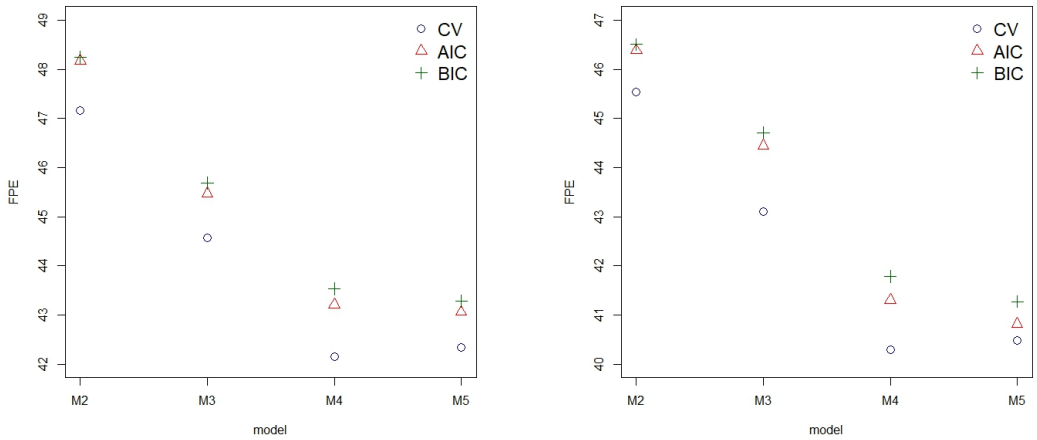

FIGURE 1. FPE of CV, AIC, and BIC for $n = 50$ (left) and $n = 100$ (right) with correlated predictors. $M2$, $M3$, $M4$, and $M5$ correspond to the model averaging estimator that use $\{M_1, \ldots, M_k\}$, $k = 2, 3, 4, 5$.

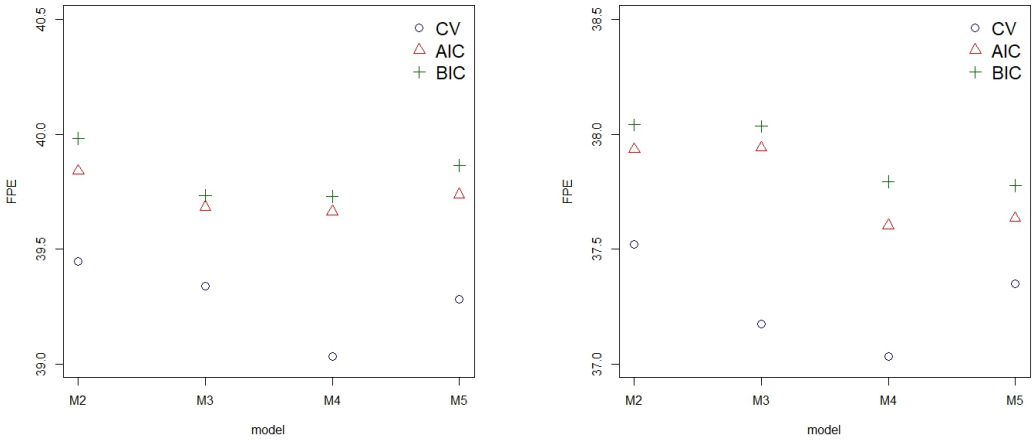

FIGURE 2. FPE of CV, AIC, and BIC for $n = 50$ (left) and $n = 100$ (right) with independent predictors. $M2$, $M3$, $M4$, and $M5$ correspond to the model averaging estimator that use $\{M_1, \ldots, M_k\}$, $k = 2, 3, 4, 5$.

Figures 1-2 present the FPEs for 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 Figure 1.

Appendix A. Mathematical appendix

A.1. **Proof of Theorem (1).** First, we show the pointwise convergence:

$$d(\hat{m}_{\oplus}(\boldsymbol{w}, x), m_{\oplus}(\boldsymbol{w}, x)) \xrightarrow{p} 0 \quad \text{for each } \boldsymbol{w} \in \mathbb{W} \text{ and } x \in \ell^2. \tag{1}$$

Pick any $\boldsymbol{w} \in \mathbb{W}$ and $x \in \ell^2$. By van der Vaart and Wellner (1996, Corollary 3.2.3), it is sufficient for (1) to show $\sup_{\omega \in \Omega} |\hat{R}(\boldsymbol{w}, x, \omega) - R(\boldsymbol{w}, x, w)| \xrightarrow{p} 0$. For this, we show that $\hat{R}(\boldsymbol{w}, x, \cdot)$ converges weakly to $R(\boldsymbol{w}, x, \cdot)$ in $\ell^\infty(\Omega)$, and then apply van der Vaart and Wellner (1996, Theorem 1.3.6). By van der Vaart and Wellner (1996, Theorem 1.5.4), this weak convergence follows by showing that

(i): $\hat{R}(\boldsymbol{w}, x, \omega) - R(\boldsymbol{w}, x, \omega) \xrightarrow{p} 0$ for each $\omega \in \Omega$.

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

$$\limsup_{n \to \infty} \mathbb{P}\left(\sup_{d(\omega_1, \omega_2) < \delta} |\hat{R}(\boldsymbol{w}, x, \omega_1) - \hat{R}(\boldsymbol{w}, x, \omega)| > \varepsilon \right) < \eta.$$

Pick any $\omega \in \Omega$. For (i), observe that

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

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

Pick any $\omega_1, \omega_2 \in \Omega$. For (ii), a similar argument yields

$$\begin{aligned}
& |\hat{R}(\boldsymbol{w}, x, \omega_1) - \hat{R}(\boldsymbol{w}, x, \omega_2)| \\
\leq & \sum_{m=1}^{M} w_m |\{d(\hat{L}_{\oplus}^{(m)}(x), \omega_1) + d(\hat{L}_{\oplus}^{(m)}(x), \omega_2)\}\{d(\hat{L}_{\oplus}^{(m)}(x), \omega_1) - d(\hat{L}_{\oplus}^{(m)}(x), \omega_2)\}| \\
\leq & 2\mathrm{diam}(\Omega) \sum_{m=1}^{M} w_m |d(\hat{L}_{\oplus}^{(m)}(x), \omega_1) - d(\hat{L}_{\oplus}^{(m)}(x), \omega_2)| \\
\leq & 2\mathrm{diam}(\Omega) d(\omega_1, \omega_2),
\end{aligned}$$

which implies $\sup_{d(\omega_1, \omega_2) < \delta} |\hat{R}(\boldsymbol{w}, x, \omega_1) - \hat{R}(\boldsymbol{w}, x, \omega_2)| = O_p(\delta)$ so that we obtain (ii). Therefore, we obtain (1).

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

(1996, Theorem 1.5.4), it is sufficient to show that for each $S > 0$,

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

as $\delta \to 0$. Since

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

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

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

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 $x_1, x_2 \in \ell^2$ with $\|x_1^{(M)} - x_2^{(M)}\|_{\ell^2} < \delta$. Note that Assumption (1) guarantees uniformly continuity of $L_\oplus^{(m)}(x)$ over $\|x^{(M)}\|_{\ell^2} \le B$ for $m = 1, \ldots, M$. Then due to the form of $R(\boldsymbol{w}, x, \omega)$, we have

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

for some $C > 0$. Thus, Assumption (2) 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 $d(\hat{m}_\oplus(\boldsymbol{w}_1,x_1), m_\oplus(\boldsymbol{w}_2,x_2)) > \varepsilon$ with $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{W}$ and $\|x_1^{(M)}\|_{\ell^2}, \|x_2^{(M)}\|_{\ell^2} \le B$. Observe that

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

as $\delta \to 0$, where the second inequality follows from the triangle inequality, third inequality follows from the uniform convergence of $\hat{L}_{\oplus}^{(m)}(x)$ in Assumption (1), and the equality follows from uniform continuity of $L_{\oplus}^{(m)}(x)$ over $\|x^{(M)}\|_{\ell^2} \leq B$.

Therefore, we obtain (3), and the conclusion of the theorem follows.

A.2. **Proof of Theorem (2).** 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

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

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

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

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 $h_{\boldsymbol{w}}(y, z) = d^2(y, m_{\oplus}(\boldsymbol{w}, z))$ and $\mathcal{F}_{\mathbb{W}} = \{h_{\boldsymbol{w}}(y, z) : \boldsymbol{w} \in \mathbb{W}\}$. An envelop function of $\mathcal{F}_{\mathbb{W}}$ is $F_{\mathbb{W}} = \mathrm{diam}(\Omega)^2$. By Assumption (3), we have

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

Thus, from van der Vaart and Wellner (1996, Theorems 2.14.2 and 2.7.11), it holds

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

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

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

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

where the convergence follows from Theorem (1).

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 Assumption (4), we obtain the conclusion.

References

[1] Akaike, H. (1970) Statistical predictor identification, *Annals of the Institute of Statistical Mathematics*, 22, 203-17.

[2] Claeskens, G. and N. L. Hjort (2008) *Model Selection and Model Averaging*, Cambridge University Press.

[3] Fréchet, M. (1948) Les éléments aléatoires de nature quelconque dans un espace distancié, *Annales de l'institut Henri Poincaré*, 10, 215-310.

[4] Marron, J. S. and A. M. Alonso (2014) Overview of object oriented data analysis, *Biometrical Journal*, 56, 732-753.

[5] Patrangenaru, V. and L. Ellingson (2015) *Nonparametric Statistics on Manifolds and Their Applications to Object Data Analysis*, CRC Press.

[6] Petersen, A. and H.-G. Müller (2019) Fréchet regression for random objects with Euclidean predictors, *Annals of Statistics*, 47, 691-719.

[7] Tucker, D. C., Wu, Y. and H.-G. Müller (2021) Variable selection for global Fréchet regression, *Journal of the American Statistical Association*, 118, 1023-1037.

[8] van der Vaart, A. and J. Wellner (1996) *Weak Convergence and Empirical Processes*, Springer.

[9] Ying, C. and Z. Yu (2022) Fréchet sufficient dimension reduction for random objects, *Biometrika*, 109, 975-992.

Center for Spatial Information Science, The University of Tokyo, 5-1-5, Kashiwanoha, Kashiwa-shi, Chiba 277-8568, Japan.

Email address: daisukekurisu@csis.u-tokyo.ac.jp

Department of Economics, London School of Economics, Houghton Street, London, WC2A 2AE, UK.

Email address: t.otsu@lse.ac.uk