Variable Selection for Global Fréchet Regression

Danielle C. Tucker, Yichao Wu and Hans-Georg Müller

Abstract

Global Fréchet regression is an extension of linear regression to cover more general types of responses, such as distributions, networks and manifolds, which are becoming more prevalent. In such models, predictors are Euclidean while responses are metric space valued. Predictor selection is of major relevance for regression modeling in the presence of multiple predictors but has not yet been addressed for Fréchet regression. Due to the metric space valued nature of the responses, Fréchet regression models do not feature model parameters, and this lack of parameters makes it a major challenge to extend existing variable selection methods for linear regression to global Fréchet regression. In this work, we address this challenge and propose a novel variable selection method that overcomes it and has good practical performance. We provide theoretical support and demonstrate that the proposed variable selection method achieves selection consistency. We also explore the finite sample performance of the proposed method with numerical examples and data illustrations.

Key words: Distributions, Euclidean predictors, Important predictors, Metric space valued data, Ridge regression, Spherical data.

1 Introduction

1.1 Background and Motivation

Regression has been a central technique for analyzing data, especially to study how a response variable depends on one or more predictor variables. It is one of the most studied statistical methods. In classical regression, responses are limited to be scalars. More general types of data, which are situated in a generic metric space, are readily available in the new era of big data. Recently, Petersen and Müller (2019) extended classical regression to Fréchet regression, making it possible to handle these more general types of responses, including distributions, symmetric positive definite matrices and data on Riemannian manifolds. Fréchet regression was formulated as a weighted Fréchet mean, with weights depending on the predictors. When one has multivariate predictors in a

regression model, an important question is which of the predictors are relevant for the response. This leads to the problem of variable selection, which is a very active research direction for regression, not least due to recent technological advances that have made it feasible to collect and store large amounts of data. The goal of variable selection for regression is to select important predictors that explain the variation of the response variable. For a review of variable selection methods, see Fan and Lv (2010) and Desboulets (2018), among many others. For more general regression models where responses are not in a linear or linearizable space, the important issue of predictor selection is difficult and has not been addressed yet. This provides the motivation to develop a practically feasible and theoretically supported method in the context of global Fréchet regression.

Petersen and Müller (2019) introduced global Fréchet regression as an extension of classical linear regression. As the name suggests, global Fréchet regression entails a global model, but it does not involve any global model parameters, in contrast to linear regression models. However, most, if not all, of the existing variable selection methods for linear regression succeed by using a sparsity-encouraging penalty on the regression coefficients. Since the global Fréchet regression model is defined without relying on any model parameter, it is therefore a major challenge to enact variable selection for global Fréchet regression. In this paper, we propose a novel variable selection approach that is shown to work for global Fréchet regression by extending the ridge selection operator that was studied in Wu (2020) for standard linear regression. We refer to this new method as Fréchet ridge selection operator (FRiSO). As the name suggests, it is based on a ridge version of global Fréchet regression and constitutes the first approach for variable selection for metric-space valued responses.

The remainder of this paper is organized as follows: Section 1.2 sets the stage and introduces the basic set-up for Fréchet regression. A brief review of global Fréchet regression is given in Section 2, followed by a review of the ridge selection operator for linear regression in Section 2.4. Section 3 introduces individually penalized ridge Fréchet regression, which is the basic building block of the proposed variable selection method (FRiSO) for global Fréchet regression, with details in Section 4. Selection consistency for FRiSO is derived in Section 4.2. A refitting procedure is presented in Section 4.3 with the goal to remedy the ridge bias. Simulation examples in Section 5 and three real data examples in Section 6 are used to illustrate the finite sample performance of FRiSO. In Section 7, we present concluding remarks and opportunities for future work and applications. All technical proofs are collected in the appendix.

1.2 Preliminaries

For a metric space denoted by (Ω, d), where d denotes the metric, we consider a random process $(\mathbf{X}, Y) \sim F$ on the product space $\mathcal{X} \times \Omega$, where $\mathcal{X} \subset \mathfrak{R}^p$. Here $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$ and Y take values in $\mathcal{X}$ and Ω, respectively, and F denotes the joint distribution of $(\mathbf{X}, Y)$ on $\mathcal{X} \times \Omega$. Denote the marginal distributions of $\mathbf{X}$ and Y by $F_{\mathbf{X}}$ and F_Y, respectively. The conditional distributions $F_{\mathbf{X}|Y}$ and $F_{Y|\mathbf{X}}$ are assumed to exist and to be well defined. This is the same scenario as in Petersen and Müller (2019), where Y is referred to as a random object.

The conventional definitions of mean and variance for Euclidean random variables are not applicable for random objects in metric spaces. Fréchet (1948) generalized the concepts of mean and variance from Euclidean data to random objects by defining the Fréchet mean and Fréchet variance of a random object Y as

$$\omega_\oplus = \arg\min_{\omega \in \Omega} E(d^2(Y, \omega)) \text{ and } V_\oplus = E(d^2(Y, \omega_\oplus)),$$

respectively.

To study the relationship between a random object and multivariate random variables, Petersen and Müller (2019) introduced the general concept of a Fréchet regression function of Y given $\mathbf{X} = \mathbf{x}$ with $\mathbf{x} = (x_1, x_2, \ldots, x_p)^T$ as

$$m_\oplus(\mathbf{x}) = \arg\min_{\omega \in \Omega} M_\oplus(\omega, \mathbf{x}),$$

where $M_\oplus(\cdot, \mathbf{x}) = E(d^2(Y, \cdot)|\mathbf{X} = \mathbf{x})$. Thus Fréchet regression can be interpreted as an implementation of the notion of conditional Fréchet means.

2 Global Fréchet Regression

2.1 Linear Regression

As a special case of Fréchet regression, global Fréchet regression was designed to extend the classical multiple linear regression to cover responses that are random objects (Petersen and Müller 2019). We proceed to review pertinent features of linear regression. The classical linear regression model is given by

$$Z = \beta_0 + \mathbf{X}^T \boldsymbol{\beta} + \epsilon \tag{1}$$

with a p-dimensional predictor vector $\mathbf{X} \in \mathcal{X} \subset \mathfrak{R}^p$ and random errors ϵ with mean zero and finite variance that are independent of $\mathbf{X}$. Of central interest is the estimation of

the unknown regression coefficients β_0 and $\boldsymbol{\beta}$ based on a random sample $\{(\mathbf{x}_i, z_i) : i = 1, 2, \ldots, n\}$ from model (1) and to make a prediction for a future observation at any $\mathbf{x}$ in the domain of interest $\mathcal{X}$. We write $\mathbf{z} = (z_1, z_2, \ldots, z_n)^T$, and $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^T$, slightly abusing the notation $\mathbf{X}$.

By incorporating the intercept term, we denote the augmented data by $\widetilde{\mathbf{x}}_i = (1, \mathbf{x}_i^T)^T$ and $\widetilde{\mathbf{X}} = (\widetilde{\mathbf{x}}_1, \widetilde{\mathbf{x}}_2, \ldots, \widetilde{\mathbf{x}}_n)^T \equiv (\mathbf{1}_{n\times 1}, \mathbf{X})$. Here $\mathbf{1}_{n\times 1}$ denotes an $n\times 1$ vector of ones. Writing $\widetilde{\boldsymbol{\beta}} = (\beta_0, \boldsymbol{\beta}^T)^T$, the ordinary least squares estimate is given by

$$\widehat{\widetilde{\boldsymbol{\beta}}} = \left(\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}\right)^{-1}\widetilde{\mathbf{X}}^T\mathbf{z}. \tag{2}$$

Let $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$ and $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$ denote the sample mean and sample covariance matrix, respectively. By using the decomposition $\mathbf{X} = (\mathbf{X} - \mathbf{1}_{n\times 1}\bar{\mathbf{x}}^T) + \mathbf{1}_{n\times 1}\bar{\mathbf{x}}^T$ and noting that $\widetilde{\mathbf{X}} = (\mathbf{1}_{n\times 1}, \mathbf{X})$, we have

$$\begin{aligned}\left(\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}\right)^{-1} &= \frac{1}{n}\begin{pmatrix} 1 & \bar{\mathbf{x}}^T \\ \bar{\mathbf{x}} & \widehat{\boldsymbol{\Sigma}} + \bar{\mathbf{x}}\bar{\mathbf{x}}^T \end{pmatrix}^{-1} \\ &= \frac{1}{n}\begin{pmatrix} 1 + \bar{\mathbf{x}}^T\widehat{\boldsymbol{\Sigma}}^{-1}\bar{\mathbf{x}} & -\bar{\mathbf{x}}^T\widehat{\boldsymbol{\Sigma}}^{-1} \\ -\widehat{\boldsymbol{\Sigma}}^{-1}\bar{\mathbf{x}} & \widehat{\boldsymbol{\Sigma}}^{-1} \end{pmatrix}\end{aligned}$$

and

$$\left(\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}\right)^{-1}\widetilde{\mathbf{X}}^T = \frac{1}{n}\begin{pmatrix} \mathbf{1}_{n\times 1}^T - \bar{\mathbf{x}}^T\widehat{\boldsymbol{\Sigma}}^{-1}(\mathbf{X}^T - \bar{\mathbf{x}}\mathbf{1}_{n\times 1}^T) \\ \widehat{\boldsymbol{\Sigma}}^{-1}(\mathbf{X}^T - \bar{\mathbf{x}}\mathbf{1}_{n\times 1}^T) \end{pmatrix}.$$

Then the prediction for a future observation at predictor level $\mathbf{x}$ is given by

$$\begin{aligned}(1, \mathbf{x}^T)\left(\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}\right)^{-1}\widetilde{\mathbf{X}}^T\mathbf{z} &= \frac{1}{n}\left(\mathbf{1}_{n\times 1}^T + (\mathbf{x} - \bar{\mathbf{x}})^T\widehat{\boldsymbol{\Sigma}}^{-1}(\mathbf{X}^T - \bar{\mathbf{x}}\mathbf{1}_{n\times 1}^T)\right)\mathbf{z} \\ &= \frac{1}{n}\sum_{i=1}^n \left[1 + (\mathbf{x} - \bar{\mathbf{x}})^T\widehat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})\right] z_i, \end{aligned} \tag{3}$$

which is nothing but a weighted average of the observed responses z_i by noting that the weights sums up to one since $\sum_{i=1}^n (\mathbf{x} - \bar{\mathbf{x}})^T\widehat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}) = 0$. The prediction (3) can be equivalently interpreted as the minimizer of

$$\min_z \sum_{i=1}^n \left[1 + (\mathbf{x} - \bar{\mathbf{x}})^T\widehat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})\right](z_i - z)^2. \tag{4}$$

By replacing the squared difference with the squared metric distance $d^2(z_i, z)$, Petersen and Müller (2019) arrived at the global Fréchet regression.

In the population perspective, (4) becomes

$$\begin{aligned} & \arg\min_{z} E_{(\mathbf{X},Z)} \left[1 + (\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{X}-\boldsymbol{\mu})\right] (Z-z)^2 \\ = \; & E_{(\mathbf{X},Z)} \left[1 + (\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{X}-\boldsymbol{\mu})\right] Z, \end{aligned}$$

where $\boldsymbol{\mu} = E(\mathbf{X})$ and $\boldsymbol{\Sigma} = \mathrm{cov}(\mathbf{X})$.

For the linear regression model (1), we have

$$E_{(\mathbf{X},Z)} \left[1 + (\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{X}-\boldsymbol{\mu})\right] Z = \beta_0 + \mathbf{x}^T \boldsymbol{\beta} = E(Z|\mathbf{X}=\mathbf{x}).$$

Consequently, the linear regression model (1) can be equivalently formulated as

$$E_{(\mathbf{X},Z)} \left[1 + (\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{X}-\boldsymbol{\mu})\right] Z = E(Z|\mathbf{X}=\mathbf{x}) \text{ for any } \mathbf{x} \in \mathcal{X}.$$

This alternative formulation for linear regression does not use any regression parameter, and it is precisely this feature that facilitates the extension from the linear regression model to the global Fréchet regression model.

2.2 Global Fréchet Regression

Motivated by the above alternative formulation of linear regression, Petersen and Müller (2019) introduced global Fréchet regression as follows.

Definition 1. *The global Fréchet regression model is characterized by*

$$m_{\oplus}(\mathbf{x}) = L_{\oplus}(\mathbf{x}) \text{ for any } \mathbf{x} \in \mathcal{X}, \tag{5}$$

where

$$L_{\oplus}(\mathbf{x}) = \arg\min_{\omega \in \Omega} E_{(\mathbf{X},Y)} \left\{ \left[1 + (\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{X}-\boldsymbol{\mu})\right] d^2(Y,\omega) \right\}$$

is assumed to be well-defined.

Assume that $\{(\mathbf{x}_i, Y_i) : i = 1, 2, \ldots, n\}$ is a random sample from F. Let $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$ and $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$ denote the sample mean and sample covariance matrix, respectively, as in the above linear regression setting. For any $\mathbf{x}$ in the domain of interest $\mathcal{X}$, the global Fréchet regression estimator of $m_{\oplus}(\mathbf{x})$ is defined as

$$\widehat{L}_{\oplus}(\mathbf{x}) \;=\; \arg\min_{\omega \in \Omega} \sum_{i=1}^n \left[1 + (\mathbf{x}-\bar{\mathbf{x}})^T \widehat{\boldsymbol{\Sigma}}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})\right] d^2(Y_i, \omega), \tag{6}$$

which can be interpreted as a weighted Fréchet mean.

Note that the global Fréchet regression estimator in (6) is a direct generalization of the fitted model and prediction (3) and (4), corresponding to the ordinary least squares estimator (2) for linear regression. Yet in contrast to the linear regression model (1) it does not involve regression coefficients. This is a key feature of the global Fréchet regression model that reflects the fact that random objects do not live in a linear space and cannot be multiplied with parameters. This central feature constitutes a major problem for applying parameter-based methods to global Fréchet regression and specifically for variable selection in global Fréchet regression, which is the focus of the current paper. While there are many variable selection methods available for classical linear regression, virtually all of these achieve variable selection by including a sparsity-encouraging penalty in the estimating equation for the regression coefficients. A typical example is the least absolute shrinkage and selection operator (LASSO; Tibshirani 1996), which achieves variable selection by placing an L_1 penalty on the regression coefficients. Since there is no parameter in the above definition of the global Fréchet regression model, it is challenging to extend these existing sparsity-encouraging penalty-based variable selection methods to global Fréchet regression.

The key to the above definition of global Fréchet regression is the use of invidual weights for each observation in (6). Such weights are not part of general sparsity-encouraging penalty-based variable selection methods. Yet such weights are available in ridge regression. In a recent paper (Wu 2020) a new variable selection was proposed for linear regression based on ridge regression. We show in this paper that by adopting ridge regression as a guiding principle, it is possible to overcome the challenge of variable selection without model parameters and to derive a well supported variable selection method for global Fréchet regression. We next review the recent variable selection method of Wu (2020), where a new variable selection method using individually penalized ridge regression was introduced. Individually penalized ridge regression is a generalized version of ordinary ridge regression (Hoerl and Kennard 1970), where one uses different ridge regularization parameters for each regression coefficient component.

2.3 Individually Penalized Ridge Regression

Based on a random sample $\{(\mathbf{x}_i, z_i) : i = 1, 2, \ldots, n\}$ from model (1), for individually penalized ridge regression one obtains the unknown regression coefficients β_0 and $\boldsymbol{\beta}$ by solving

$$\min_{\beta_0, \boldsymbol{\beta}} \frac{1}{2n} \sum_{i=1}^{n} (z_i - \beta_0 - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \frac{1}{2} \sum_{j=1}^{p} \nu_j \beta_j^2 \tag{7}$$

with ridge regularization parameters $\nu_j \geq 0$ for $j = 1, 2, \ldots, p$. The solution of (7) is easily found to be

$$\widehat{\widetilde{\boldsymbol{\beta}}}_R = \left[\frac{1}{n}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}} + \text{diag}((0, \boldsymbol{\nu}^T)^T)\right]^{-1}\left(\frac{1}{n}\widetilde{\mathbf{X}}^T\mathbf{z}\right),$$

where $\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots, \nu_p)^T$ and $\text{diag}(\boldsymbol{\nu})$ denotes a diagonal matrix with elements of $\boldsymbol{\nu}$ sitting on the diagonal.

2.4 Ridge Selection Operator

Following Wu (2020) when substituting $\lambda_j = 1/\nu_j$, $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_p)^T$, $\boldsymbol{\lambda}^{-1} = (1/\lambda_1, 1/\lambda_2, \ldots, 1/\lambda_p)^T$ and $\boldsymbol{\nu} = \boldsymbol{\lambda}^{-1}$, the solution of (7) is given by

$$\widehat{\widetilde{\boldsymbol{\beta}}}_R = \left[\frac{1}{n}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}} + \text{diag}((0, (\boldsymbol{\lambda}^{-1})^T)^T)\right]^{-1}\left(\frac{1}{n}\widetilde{\mathbf{X}}^T\mathbf{z}\right)$$

and the corresponding hat matrix is

$$\mathbf{H}(\boldsymbol{\lambda}) = \frac{1}{n}\widetilde{\mathbf{X}}\left[\frac{1}{n}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}} + \text{diag}((0, (\boldsymbol{\lambda}^{-1})^T)^T)\right]^{-1}\widetilde{\mathbf{X}}^T.$$

Variable selection for model (1) then proceeds by solving

$$\begin{aligned} \min_{\boldsymbol{\lambda}} \quad & \langle \mathbf{z} - \mathbf{H}(\boldsymbol{\lambda})\mathbf{z}, \mathbf{z} - \mathbf{H}(\boldsymbol{\lambda})\mathbf{z}\rangle, && (8) \\ \text{subject to} \quad & \lambda_j \geq 0,\ j = 1, 2, \ldots, p; && (9) \\ & \sum_{j=1}^{p} \lambda_j \leq \tau, && (10) \end{aligned}$$

for a regularization parameter $\tau \geq 0$, where $\langle \cdot, \cdot \rangle$ denotes the inner product in $\mathfrak{R}^n$. Denote the optimizer by $\widehat{\boldsymbol{\lambda}} = (\widehat{\lambda}_1, \widehat{\lambda}_2, \ldots, \widehat{\lambda}_p)^T$. For an appropriately tuned τ, some components of the corresponding optimizer will be exactly zero and an estimate of the set of important predictors is given by $\{j : \widehat{\lambda}_j > 0\}$.

Based on the above derivation of (3), the hat matrix can be simplified as follows

$$\mathbf{H}(\boldsymbol{\lambda}) = \frac{1}{n}\left\{\mathbf{1}_{n\times 1}\mathbf{1}_{n\times 1}^T + (\mathbf{X} - \mathbf{1}_{n\times 1}\bar{\mathbf{x}}^T)\left[\widehat{\boldsymbol{\Sigma}} + \text{diag}(\boldsymbol{\lambda}^{-1})\right]^{-1}(\mathbf{X} - \mathbf{1}_{n\times 1}\bar{\mathbf{x}}^T)^T\right\}.$$

Here it is possible that within the feasible domain specified by constraints (9) and (10), some components of $\boldsymbol{\lambda}$ are exactly zero, in which case the second term inside the hat matrix

$\mathbf{H}(\boldsymbol{\lambda})$ cannot be evaluated due to the division by 0. This issue can be circumvented by noting that

$$\left[\widehat{\boldsymbol{\Sigma}} + \text{diag}(\boldsymbol{\lambda}^{-1})\right]^{-1} = \text{diag}(\sqrt{\boldsymbol{\lambda}}) \left[\text{diag}(\sqrt{\boldsymbol{\lambda}})\widehat{\boldsymbol{\Sigma}}\text{diag}(\sqrt{\boldsymbol{\lambda}}) + \mathbf{I}\right]^{-1} \text{diag}(\sqrt{\boldsymbol{\lambda}}), \tag{11}$$

where $\sqrt{\boldsymbol{\lambda}} = (\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_p})^T$ and $\mathbf{I}$ denotes the $p \times p$ identity matrix.

It can be shown that in (10) equality will be attained at the optimal solution. Consequently, it is equivalent to replace constraint (10) with the equality constraint $\sum_{j=1}^{p} \lambda_j = \tau$. The corresponding optimization problem with $\sum_{j=1}^{p} \lambda_j = \tau$ can be efficiently solved by using the modified coordinate descent algorithm introduced in Stefanski et al. (2014).

3 Individually Penalized Ridge Fréchet Regression

A key observation for the extension to global Fréchet regression is that for individually penalized ridge regression, the corresponding prediction of a future observation with covariates $\mathbf{x}$ can be written as a weighted average as follows

$$\begin{aligned}
& (1, \mathbf{x}^T) \left(\frac{1}{n}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}} + \text{diag}((0, \boldsymbol{\nu}^T)^T)\right)^{-1} \frac{1}{n}\widetilde{\mathbf{X}}^T\mathbf{z} \\
= \; & \frac{1}{n}\left(\mathbf{1}_{n\times 1}^T + (\mathbf{x} - \bar{\mathbf{x}})^T \left[\widehat{\boldsymbol{\Sigma}} + \text{diag}(\boldsymbol{\nu})\right]^{-1} (\mathbf{X}^T - \bar{\mathbf{x}}\mathbf{1}_{n\times 1}^T)\right)\mathbf{z} \\
= \; & \frac{1}{n}\sum_{i=1}^{n}\left\{1 + (\mathbf{x} - \bar{\mathbf{x}})^T \left[\widehat{\boldsymbol{\Sigma}} + \text{diag}(\boldsymbol{\nu})\right]^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})\right\} z_i,
\end{aligned}$$

which is equivalent to

$$\arg\min_{z} \sum_{i=1}^{n}\left\{1 + (\mathbf{x} - \bar{\mathbf{x}})^T \left[\widehat{\boldsymbol{\Sigma}} + \text{diag}(\boldsymbol{\nu})\right]^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})\right\} (z_i - z)^2,$$

where the weights sum up to one again since $\sum_{i=1}^{n}(\mathbf{x}-\bar{\mathbf{x}})^T \left[\widehat{\boldsymbol{\Sigma}} + \frac{1}{n}\text{diag}(\boldsymbol{\nu})\right]^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) = 0$.

This equivalent optimization formulation makes it possible to extend the above individually penalized ridge regression to individually penalized ridge Fréchet regression with ridge regularization parameters $\boldsymbol{\nu}$, where we introduce

$$\widehat{R}_{\oplus}(\mathbf{x}; \boldsymbol{\nu}) = \arg\min_{\omega\in\Omega} \sum_{i=1}^{n}\left\{1 + (\mathbf{x} - \bar{\mathbf{x}})^T \left[\widehat{\boldsymbol{\Sigma}} + \text{diag}(\boldsymbol{\nu})\right]^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})\right\} d^2(Y_i, \omega) \tag{12}$$

for the prediction of the Fréchet regression function $m_{\oplus}(\mathbf{x})$ at any location $\mathbf{x}$ in the domain of interest $\mathcal{X}$.

The population target of the above individually penalized ridge Fréchet regression (12) is given by

$$R_{\oplus}(\mathbf{x};\boldsymbol{\nu}) = \underset{\omega\in\Omega}{\arg\min}\, E_{(\mathbf{X},Y)}\left\{\left[1+(\mathbf{x}-\boldsymbol{\mu})^T\left[\boldsymbol{\Sigma}+\operatorname{diag}(\boldsymbol{\nu})\right]^{-1}(\mathbf{X}-\boldsymbol{\mu})\right]d^2(Y,\omega)\right\}. \quad (13)$$

For the following asymptotic result, we assume that $\mathcal{X} \subset \mathfrak{R}^p$ is compact, and that there exists a constant $B > 0$ such that

$$\sup_{\mathbf{x}\in\mathcal{X}} \|x\| \leq B. \quad (14)$$

Proposition 1. *Assume that (U0)-(U2) of Theorem 2 in Petersen and Müller (2019) hold for B in (14). Then*

$$\sup_{\mathbf{x}\in\mathcal{X}} d(\widehat{R}_{\oplus}(\mathbf{x};\boldsymbol{\nu}), R_{\oplus}(\mathbf{x};\boldsymbol{\nu})) = o_p(1)$$

for any $\boldsymbol{\nu} \in \mathfrak{R}_+^p$, where $\mathfrak{R}_+ = [0,\infty)$.

Note that Proposition 1 states the uniform consistency of the ridge Fréchet regression estimate. The convergence rate is found to be $O_p(n^{-\frac{1}{2(\alpha-1)}})$ for a constant $\alpha > 1$ that is connected to assumptions (U0)-(U2); for details we refer to Petersen and Müller (2019). For variable selection consistency, which is developed in Section 4.2, it is sufficient for the individually penalized ridge Fréchet regression estimate in (12) to be uniformly consistent.

4 Variable Selection for Global Fréchet Regression

4.1 Proposed Selector

We are now ready to present the proposed variable selection method for the global Fréchet regression based on the individually penalized ridge Fréchet regression (12). Using $\boldsymbol{\nu} = \boldsymbol{\lambda}^{-1}$ and (11), we solve

$$\min_{\omega\in\Omega} \sum_{i=1}^{n} \left\{1+(\mathbf{x}-\bar{\mathbf{x}})^T \operatorname{diag}(\sqrt{\boldsymbol{\lambda}}) \left[\operatorname{diag}(\sqrt{\boldsymbol{\lambda}})\widehat{\boldsymbol{\Sigma}}\operatorname{diag}(\sqrt{\boldsymbol{\lambda}}) + \mathbf{I}\right]^{-1} \operatorname{diag}(\sqrt{\boldsymbol{\lambda}})(\mathbf{x}_i-\bar{\mathbf{x}})\right\} d^2(Y_i,\omega)$$

and denote the corresponding optimizer by $\widehat{R}_{\oplus}(\mathbf{x};\boldsymbol{\lambda}^{-1})$. Then variable selection is implemented by solving

$$\min_{\boldsymbol{\lambda}} \quad \frac{1}{n}\sum_{i=1}^{n} d^2(Y_i, \widehat{R}_{\oplus}(\mathbf{x}_i;\boldsymbol{\lambda}^{-1})) \quad (15)$$

$$\text{subject to} \quad \lambda_j \geq 0, j = 1,2,\ldots,p; \quad (16)$$

$$\sum_{j=1}^{p} \lambda_j = \tau \quad (17)$$

for $\tau \geq 0$. Denote the solution by $\widehat{\boldsymbol{\lambda}}(\tau) = (\hat{\lambda}_1(\tau), \hat{\lambda}_2(\tau), \ldots, \hat{\lambda}_p(\tau))^T$. For an appropriately tuned τ, some components of the corresponding optimizer $\widehat{\boldsymbol{\lambda}}(\tau)$ will be exactly zero. Then an estimate of the set of important predictors is given by $\widehat{\mathcal{I}}(\tau) = \{j : \widehat{\lambda}_j(\tau) > 0\}$, just as in the linear regression case. We refer to this proposed variable selection method as Fréchet ridge selection operator (FRiSO).

The proposed FRiSO involves a tuning parameter $\tau > 0$. If there are enough data, it is recommended use an independent validation set to tune τ. Denote the independent validation data set by $\{(\tilde{\mathbf{x}}_i, \tilde{Y}_i) : i = 1, 2, \ldots, \tilde{n}\}$. Then τ may be selected by minimizing $\sum_{i=1}^{\tilde{n}} d^2(\tilde{Y}_i, \widehat{R}_{\oplus}(\tilde{\mathbf{x}}_i; (\boldsymbol{\lambda}(\tau))^{-1}))$ with respect to $\tau > 0$. A grid search can be used to implement this optimization. If the data are not rich enough for this approach, another option is cross-validation. Details for both of these tuning methods can be found in the supplementary materials for this paper.

4.2 Selection Consistency

We establish here selection consistency for FRiSO, the proposed variable selection method. A first step is to define the so-called important predictor set for global Fréchet regression.

Definition 2. *A set $\mathcal{I} \subseteq \{1, 2, \ldots, p\}$ is called the important predictor set for global Fréchet regression of random objects Y on multivariate random vectors $\mathbf{X}$, if $\mathcal{I}$ is the smallest set satisfying $Y \perp\!\!\!\perp X_{\mathcal{I}^c} | X_{\mathcal{I}}$, i.e., Y is conditionally independent of $X_{\mathcal{I}^c}$ given $X_{\mathcal{I}}$.*

For any set $\mathcal{A} \subseteq \{1, 2, \ldots, p\}$, $\mathbf{x}_{\mathcal{A}}$ denotes the subvector of $\mathbf{x}$ with indices in $\mathcal{A}$ and $\boldsymbol{\Sigma}_{\mathcal{A},\mathcal{A}}$ denotes the submatrix of $\boldsymbol{\Sigma}$ with row and column indices in $\mathcal{A}$. Define

$$L_{\oplus}^{\mathcal{A}}(\mathbf{x}) = \arg\min_{\omega \in \Omega} E_{(\mathbf{X}_{\mathcal{A}}, Y)} \left[1 + (\mathbf{x}_{\mathcal{A}} - \boldsymbol{\mu}_{\mathcal{A}})^T \boldsymbol{\Sigma}_{\mathcal{A},\mathcal{A}}^{-1} (\mathbf{X}_{\mathcal{A}} - \boldsymbol{\mu}_{\mathcal{A}})\right] d^2(Y, \omega).$$

Then $L_{\oplus}(\mathbf{x}) = L_{\oplus}^{\{1,2,\ldots,p\}}(\mathbf{x})$ by definition. While in the global Fréchet regression model (5) $m_{\oplus}(\mathbf{x}) = L_{\oplus}^{\mathcal{A}}(\mathbf{x})$ for any $\mathbf{x} \in \mathcal{X}$ with $\mathcal{A} = \{1, 2, \ldots, p\}$, the next result shows that this also holds for any $\mathcal{A}$ satisfying $\mathcal{I} \subseteq \mathcal{A} \subseteq \{1, 2, \ldots, p\}$, under the following technical condition.

Condition [A]: For any $\mathcal{A}$ satisfying $\mathcal{I} \subseteq \mathcal{A} \subseteq \{1, 2, \ldots, p\}$, we have

$$\left\{E(\boldsymbol{\Sigma}_{\mathcal{A},\mathcal{A}}^{-1}(\mathbf{X}_{\mathcal{A}} - \boldsymbol{\mu}_{\mathcal{A}}) | \mathbf{X}_{\mathcal{I}})\right\}_{\mathcal{I}} = \boldsymbol{\Sigma}_{\mathcal{I},\mathcal{I}}^{-1}(\mathbf{X}_{\mathcal{I}} - \boldsymbol{\mu}_{\mathcal{I}})$$

and

$$\left\{E(\boldsymbol{\Sigma}_{\mathcal{A},\mathcal{A}}^{-1}(\mathbf{X}_{\mathcal{A}} - \boldsymbol{\mu}_{\mathcal{A}}) | \mathbf{X}_{\mathcal{I}})\right\}_{\mathcal{A}\setminus\mathcal{I}} = \mathbf{0},$$

where $\mathcal{A} \setminus \mathcal{I} = \{j : j \in \mathcal{A} \text{ and } j \notin \mathcal{I}\}$.

This condition holds for example if $\mathbf{X}$ follows a multivariate ellipical distribution; see Theorem 2.18 of Fang et al. (1990).

Lemma 1. *If the global Fréchet regression model (5) holds and $\mathcal{I} \subseteq \{1, 2, \ldots, p\}$ is the important predictor set for the global Fréchet regression of random objects Y on multivariate random vectors $\mathbf{X}$, then under Condition [A] we have*

$$m_{\oplus}(\mathbf{x}) = L_{\oplus}^{\mathcal{A}}(\mathbf{x}) \text{ for any } \mathbf{x} \in \mathcal{X} \text{ as long as } \mathcal{A} \text{ satisfies } \mathcal{I} \subseteq \mathcal{A} \subseteq \{1, 2, \ldots, p\}.$$

Note that by definition the important predictor set $\mathcal{I}$ is the smallest set satisfying $Y \perp\!\!\!\perp X_{\mathcal{I}^c} | X_{\mathcal{I}}$. It is reasonable to assume the following additional condition to hold.

Condition [B]: $E_{\mathbf{X}} d^2(m_{\oplus}(\mathbf{X}), L_{\oplus}^{\mathcal{A}}(\mathbf{X})) \equiv \int_{\mathcal{X}} d^2(m_{\oplus}(\mathbf{x}), L_{\oplus}^{\mathcal{A}}(\mathbf{x})) F_{\mathbf{X}}(d\mathbf{x}) > 0$ for any set $\mathcal{A}$ satisfying $\mathcal{I} \setminus \mathcal{A} \neq \emptyset$.

This condition asserts that global Fréchet regression will change if any important predictor is removed from the predictor set. This assumption is satisfied in the case of Euclidean responses as verified in the next proposition and needs to be verified on a case-by-case basis.

Proposition 2. *Condition [B] is satisisfied by the linear regression model $Y = \beta_0 + \mathbf{X}^T \boldsymbol{\beta} + \epsilon$ with $E(\mathbf{X}) = \boldsymbol{\mu}$, $\text{cov}(\mathbf{X}) = \boldsymbol{\Sigma}$, $E(\epsilon) = 0$ and $\text{var}(\epsilon) < \infty$.*

The following technical conditions are additionally needed to show that the proposed variable selection method is selection consistent. We also verify that they are satisfied in the case of Euclidean responses. In the following 6 instances and also later there is $L_{\oplus}(\mathbf{X})$ which is however not defined. It needs a definition which means the set over which it is taken needs to be defined.

Condition [C]: Assume that $E_{\mathbf{X}} d^2(L_{\oplus}(\mathbf{X}), R_{\oplus}(\mathbf{X}; \boldsymbol{\nu})) > 0$ for any nonnegative vector $\boldsymbol{\nu}$ satisfying $\|\boldsymbol{\nu}_{\mathcal{I}}\| > 0$.

Condition [D]: Assume that the gradient $\frac{\partial}{\partial \boldsymbol{\nu}_{\mathcal{I}}} E_{\mathbf{X}} d^2(L_{\oplus}(\mathbf{X}), R_{\oplus}(\mathbf{X}; \boldsymbol{\nu})) = \mathbf{0}$ and the Hessian $\frac{\partial^2}{\partial \boldsymbol{\nu}_{\mathcal{I}} \partial \boldsymbol{\nu}_{\mathcal{I}}^T} E_{\mathbf{X}} d^2(L_{\oplus}(\mathbf{X}), R_{\oplus}(\mathbf{X}; \boldsymbol{\nu}))$ is strictly positive definite at any $\boldsymbol{\nu}$ satisfying $\nu_j = 0$ for any $j \in \mathcal{I}$.

Condition [C] essentially means that for global Fréchet regression, the model bias (quantified in terms of $E_{\mathbf{X}} d^2(L_{\oplus}(\mathbf{X}), R_{\oplus}(\mathbf{X}; \boldsymbol{\nu}))$) of the individually penalized ridge Fréchet regression is positive as long as the ridge parameters corresponding to important predictors are nonzero. Further Condition [D] refines the local behavioral of

$E_{\mathbf{X}} d^2(L_\oplus(\mathbf{X}), R_\oplus(\mathbf{X}; \boldsymbol{\nu}))$ at any $\boldsymbol{\nu}$ with $\nu_j = 0$ for any j in $\mathcal{I}$. It essentially meants that for any nonnegative vector $\boldsymbol{\nu}_{\mathcal{I}^c}$, $E_{\mathbf{X}} d^2(L_\oplus(\mathbf{X}), R_\oplus(\mathbf{X}; \boldsymbol{\nu}))$ as a function of $\boldsymbol{\nu}_{\mathcal{I}}$ has a zero gradient and a strictly positive definite Hessian at $\boldsymbol{\nu}_{\mathcal{I}} = \mathbf{0}$. These are reasonable assumptions that hold for the case of Euclidean responses, as verified by the next two propositions. It is better to combine these into one prop. just saying [C] and [D] are satisfied for...

Proposition 3. *Condition [C] is satisfied by the linear regression model* $Y = \beta_0 + \mathbf{X}^T\boldsymbol{\beta} + \epsilon$ *with* $E(\mathbf{X}) = \boldsymbol{\mu}$, $\mathrm{cov}(\mathbf{X}) = \boldsymbol{\Sigma}$, $E(\epsilon) = 0$ *and* $\mathrm{var}(\epsilon) < \infty$.

Proposition 4. *Condition [D] is satisfied by the linear regression model* $Y = \beta_0 + \mathbf{X}^T\boldsymbol{\beta} + \epsilon$ *with* $E(\mathbf{X}) = \boldsymbol{\mu}$, $\mathrm{cov}(\mathbf{X}) = \boldsymbol{\Sigma}$, $E(\epsilon) = 0$ *and* $\mathrm{var}(\epsilon) < \infty$.

Under these technical conditions, the next theorem establishes the selection consistency of the proposed FRiSO.

Theorem 1. *Assume that (U0)-(U2) of Theorem 2 in Petersen and Müller (2019) hold for B in (14). Under Conditions [A-C], when* $\tau = \tau_n \to \infty$ *as* $n \to \infty$, *the solution* $\widehat{\boldsymbol{\lambda}}(\tau_n)$ *of (15) satisfies* $\hat{\lambda}_j(\tau_n) \xrightarrow{p} \infty$ *for* $j \in \mathcal{I}$ *and* $\hat{\lambda}_{j'}(\tau_n) \xrightarrow{p} 0$ *for* $j' \notin \mathcal{I}$ *as* $n \to \infty$.

4.3 Refitting

The result in Theorem 1 that $\hat{\lambda}_j(\tau_n) \xrightarrow{p} \infty$ for $j \in \mathcal{I}$ implies that the corresponding ridge penalty parameter $\nu_j = 1/\hat{\lambda}_j(\tau_n) \xrightarrow{p} 0$ in (12) for $j \in \mathcal{I}$. Thus the ridge term corresponding to any important predictor in the individually penalized ridge Fréchet regression disappears asymptotically. Yet in the finite sample case with a finite n, $\hat{\lambda}_j(\tau_n)$ is always finite, $\hat{\lambda}_j(\tau_n) < \infty$, and consequently, $\nu_j = 1/\hat{\lambda}_j(\tau_n) > 0$. In this case, the corresponding ridge term does not disappear, causing finite-sample ridge bias. This issue was also discussed in Wu (2020) for the linear regression case, where a refitting step was proposed to mitigate the finite-sample ridge bias. This refitting step can be extended as follows.

For every τ with optimal solution $\widehat{\boldsymbol{\lambda}}(\tau)$ of (15), we obtain $\widehat{\mathcal{I}}(\tau) = \{j : \widehat{\lambda}_j(\tau) > 0\}$. Then the refitted estimate of the Fréchet regression function $m_\oplus(\mathbf{x})$ is given by

$$\hat{m}_\oplus^{\mathrm{refit}}(\mathbf{x}; \tau) = \arg\min_{\omega \in \Omega} \sum_{i=1}^{n} \left[1 + (\mathbf{x}_{\widehat{\mathcal{I}}(\tau)} - \bar{\mathbf{x}}_{\widehat{\mathcal{I}}(\tau)})^T \widehat{\boldsymbol{\Sigma}}^{-1}_{\widehat{\mathcal{I}}(\tau), \widehat{\mathcal{I}}(\tau)} (\mathbf{x}_{i, \widehat{\mathcal{I}}(\tau)} - \bar{\mathbf{x}}_{\widehat{\mathcal{I}}(\tau)})\right] d^2(Y_i, \omega).$$

5 Simulation Studies

5.1 Overview

In the following, we discuss three types of Fréchet regression examples to demonstrate the finite-sample performance of the proposed FRiSO variable selection method for global

Fréchet regression. Fréchet regression is an abstract concept as also has been pointed out by reviewers. The level of abstractness is necessitated by the lack of linearity in the response space which makes regression a much more difficult concept than it is in linear spaces. Regardless of the level of abstractness, Fréchet regression is immediately applicable in practice. How one solves the optimization problems of global Fréchet regression (6) and ridge Fréchet regression (12) is specific for each space and the selected metric in the space. To facilitate the readers' understanding, we provide further explanatory details in the supplement, specifically for the following three implementation examples. We also provide complete implementation codes in the supplement.

Three relevant examples for responses for which Fréchet regression is applicable are one-dimensional probability distributions with the 2-Wasserstein metric, symmetric positive definitive matrices such as covariances matrices or graph representations, and data on the sphere such as directional data. We will explore these special cases in both the following simulations and the data illustrations below.

5.2 Fréchet Regression for Probability Distributions With the Wasserstein Metric

The 2-Wasserstein metric distance between two distributions with cumulative distribution functions $H(\cdot)$ and $G(\cdot)$ is defined as $W_2(H,G) = \sqrt{\int_0^1 (H^{-1}(t) - G^{-1}(t))^2 dt}$.

Data were generated by adapting the simulation example in Petersen and Müller (2019) and correlated scalar predictors $X_j \sim \mathcal{U}(-1,1)$, $j = 1,2,\ldots,p$, generated in two steps: (1) $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_p)^T$ multivariate Gaussian with $E(Z_j) = 0$ and $\mathrm{cov}(Z_j, Z_{j'}) = \rho^{|j-j'|}$; (2) $X_j = 2\Phi(Z_j) - 1$ for $j = 1,\ldots,p$, where Φ is the standard normal distribution function. We set $p = 10$ and $\rho = 0.5$.

Example 5.2.1. *The Fréchet regression function is given by*

$$m_{\oplus}(\mathbf{x}) = E(Y(\cdot)|\mathbf{X} = \mathbf{x}) = \mu_0 + \beta x_4 + (\sigma_0 + \gamma x_4)\Phi^{-1}(\cdot).$$

Conditional on $\mathbf{X}$*, the random response* Y *is generated by adding noise as follows:* $Y = \mu + \sigma\Phi^{-1}$ *with* $\mu|\mathbf{X} \sim N(\mu_0 + \beta X_4, \nu_1)$ *and* $\sigma|\mathbf{X} \sim Gamma((\sigma_0 + \gamma X_4)^2/\nu_2, \nu_2/(\sigma_0 + \gamma X_4))$ *being independently sampled. It is then obvious that only* X_4 *is important. The additional parameters were chosen as* $\mu_0 = 0$*,* $\sigma_0 = 3$*,* $\beta = 3$*,* $\gamma = 0.5$*,* $\nu_1 = 1$*, and* $\nu_2 = 2$*.*

Example 5.2.2. *The Fréchet regression function is given by*

$$m_{\oplus}(\mathbf{x}) = E(Y(\cdot)|\mathbf{X} = \mathbf{x}) = \mu_0 + \beta(x_4 + x_8) + (\sigma_0 + \gamma x_1)\Phi^{-1}(\cdot).$$

Conditional on $\mathbf{X}$*, the random response* Y *is generated by adding noise as follows:* $Y = \mu + \sigma\Phi^{-1}$ *with* $\mu|\mathbf{X} \sim N(\mu_0 + \beta(X_4 + X_8), \nu_1)$ *and* $\sigma|\mathbf{X} \sim Gamma((\sigma_0 + \gamma X_1)^2/\nu_2, \nu_2/(\sigma_0 +$

$\gamma X_1))$ *being independently sampled. For this example, important predictors are* X_1, X_4, *and* X_8. *The additional parameters are set as* $\mu_0 = 0$, $\sigma_0 = 3$, $\beta = 3/4$, $\gamma = 1$, $\nu_1 = 1$, *and* $\nu_2 = 0.5$.

Training samples of size $n = 200$ were used for both examples, and an independent validation set of the same size and generated in the same way was used to tune the regularization parameter τ via minimizing the squared prediction error over the independent validation set as discussed in Section 4.1. The selection frequencies obtained over 100 repetitions are reported for each predictor in Table 1, where we also report the frequency of consistent solution paths (Yuan and Lin 2007). In our implementation, we obtain the optimal solution $\widehat{\boldsymbol{\lambda}}(\tau)$ over a prespecified grid for τ, say $\{\tau_1 < \tau_2 < \ldots < \tau_K\}$. A solution path is considered to be consistent if the optimal solution $\widehat{\boldsymbol{\lambda}}(\tau_k)$ leads to the same sparsity pattern as the truth for some $k \in \{1, 2, \ldots, K\}$.

Table 1: Simulation results for variable selection with FRiSO for global Fréchet regression when the random objects are probability distributions.

Example	selection frequency										Path
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	consistency
5.2.1	11	16	26	100	31	26	15	16	18	17	100
5.2.2	100	17	21	100	26	22	20	100	20	12	96
5.2.1	0	4	0	100	0	1	0	7	2	3	with refitting
5.2.2	100	0	0	100	4	1	0	100	0	0	

An independent test set of size $\tilde{n} = 100n$, denoted by $\{(\tilde{\mathbf{X}}_i, \tilde{Y}_i) : i = 1, 2, \ldots, \tilde{n}\}$, was generated in the same way to evaluate the performance of the estimated Fréchet regression function. The Wasserstein discrepancy $D_1 = \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} [W_2(m_\oplus(\tilde{\mathbf{X}}_i), \tilde{Y}_i)]^2$ can be interpreted as model error due to randomness; it cannot be predicted. For every sample $\{(\mathbf{X}_i, Y_i) : i = 1, 2, \ldots, n\}$, we denote the corresponding estimated Fréchet regression function by $\widehat{m}_\oplus(\cdot)$, corresponding to $\widehat{R}_\oplus(\cdot; \widehat{\boldsymbol{\lambda}}(\hat{\tau})^{-1})$, using the above notation. Here $\hat{\tau}$ denotes the tuned optimal tuning parameter τ. Defining $D_2 = \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} [W_2(m_\oplus(\tilde{\mathbf{X}}_i), \widehat{m}_\oplus(\tilde{\mathbf{X}}_i))]^2$, $D_3 = \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} [W_2(\widehat{m}_\oplus(\tilde{\mathbf{X}}_i), \tilde{Y}_i)]^2$, $D_4 = \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} [W_2(m_\oplus(\tilde{\mathbf{X}}_i), \hat{m}^*)]^2$ and $D_5 = \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} [W_2(\hat{m}^*, \tilde{Y}_i)]^2$, where $\hat{m}^* = \arg\min_\omega \sum_{i=1}^{n} [W_2(\omega, Y_i)]^2$ denotes the sample Fréchet mean, D_2 calibrates how well the proposed method behaves in terms of estimating the true Fréchet regression function, D_3 evaluates the prediction performance of the estimated Fréchet regression

function, and D_4 and D_5 are the counterparts of D_2 and D_3 if we do not use the predictors to perform Fréchet regression.

In Table 2, we report D_1 and the means of D_2, D_3, D_4, D_5 over 100 repetitions with standard error in parentheses. One finds that D_2 is much smaller than D_1, indicating that the estimated Fréchet regression function estimates the true Fréchet regression function quite well. For one random repetition in Example 5.2.2, we plot the corresponding solution path in Figure 1. In this example, X_1, X_4, and X_8 are important predictors. We find that as the regularization parameter τ increases, the optimal $\hat{\lambda}_j$ corresponding to the important predictors increase far above those of the unimportant predictors, and the solution path is seen to be path consistent.

Table 2: Simulation results (prediction) of Fréchet regression for probability distributions with the Wasserstein metric.

Example	Mean (standard error) over 100 repetitions				D_1
	D_2	D_3	D_4	D_5	
5.2.1	0.1416(0.0050)	2.5332(0.0051)	2.9360(0.0022)	5.3292(0.0022)	2.3912
5.2.2	0.1215(0.0028)	1.4330(0.0027)	0.6164(7e-04)	1.9290(8e-04)	1.3111
5.2.1	0.0259(0.0023)	2.4177(0.0023)	with refitting		
5.2.2	0.0266(0.0018)	1.3379(0.0018)			

We introduced a refitting step in Section 4.3 to mitigate the ridge bias. The corresponding results with a refitting step are shown in the bottom halves of Tables 1 and 2 and indicate significant improvements in terms of both variable selection and Fréchet regression function estimation.

Upon the suggestion of a reviewer, we provide an additional simulation example for a case where all predictors are important. We set $p = 6$ and the predictors are generated in the same way as described above.

Example 5.2.3. *The Fréchet regression function is given by*

$$m_{\oplus}(\mathbf{x}) = E(Y(\cdot)|\mathbf{X} = \mathbf{x}) = \mu_0 + \beta(x_1 + x_3 + x_4 + x_6) + (\sigma_0 + \gamma(x_2 + x_5))\Phi^{-1}(\cdot).$$

Conditional on $\mathbf{X}$*, the random response* Y *is generated by adding noise as follows:* $Y = \mu + \sigma\Phi^{-1}$ *with* $\mu|\mathbf{X} \sim N(\mu_0 + \beta(X_1 + X_3 + X_4 + X_6), \nu_1)$ *and* $\sigma|\mathbf{X} \sim$ *Gamma*$((\sigma_0 + \gamma(X_2 + X_5))^2/\nu_2, \nu_2/(\sigma_0 + \gamma(X_2 + X_5)))$ *being independently sampled. For this example, all six predictors are important predictors.*

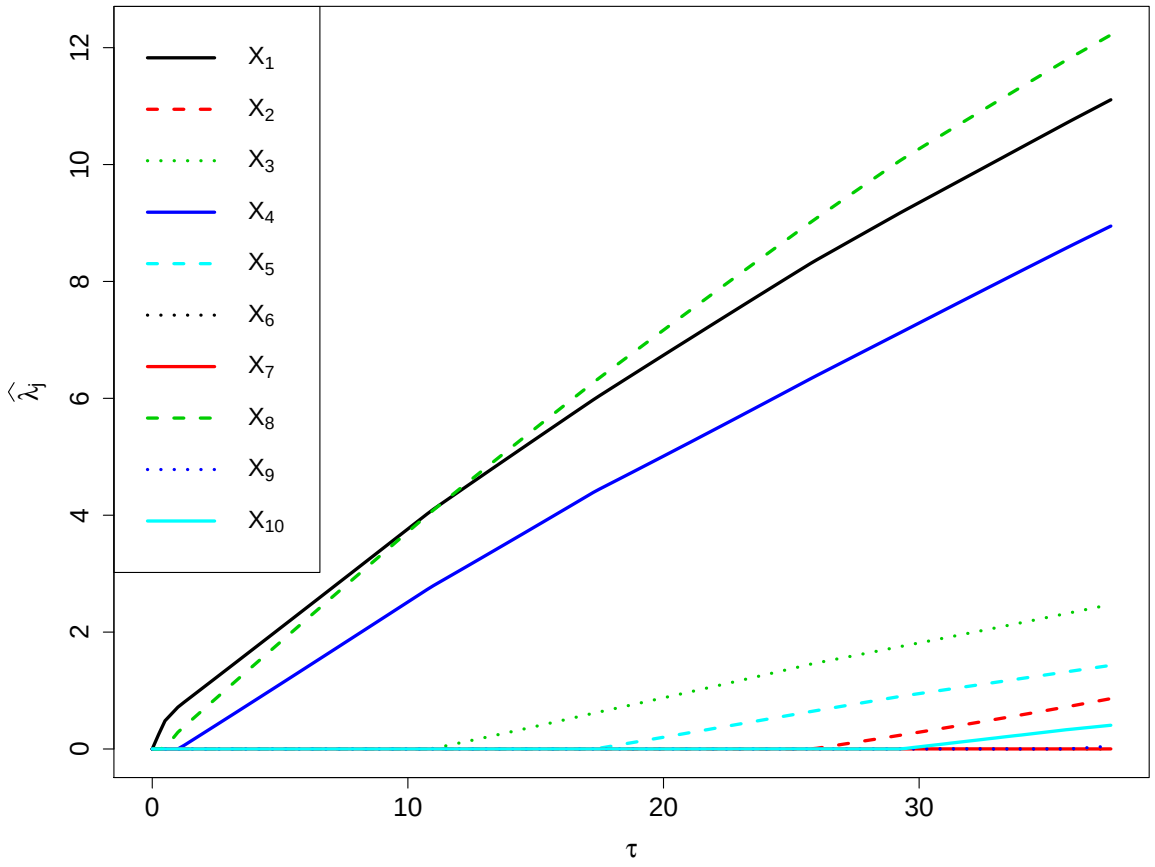

Figure 1: Solution path for a random repetition of Example 5.2.2, where X_1, X_4, and X_8 are important predictors.

The additional parameters are set as $\mu_0 = 0$, $\sigma_0 = 3$, $\beta = 3/4$, $\gamma = 1$, $\nu_1 = 1$, *and* $\nu_2 = 0.5$.

The remaining procedure is exactly the same with training sets of size 200, an independent tuning set of size 200, and an independent test set of size 20000. Corresponding results are reported in Tables 3 and 4. An equally good performance is observed. In particular, the results show that all predictors are selected over all 100 repetitions and the proposed FRiSO selection is seen to work very well when the responses are distributions.

Table 3: Simulation results for variable selection in Example 5.2.3.

selection frequency						Path
X_1	X_2	X_3	X_4	X_5	X_6	consistency
100	100	100	100	100	100	100
100	100	100	100	100	100	with refitting

Table 4: Simulation results for prediction in Example 5.2.3.

Mean (standard error) over 100 repetitions				D_1
D_2	D_3	D_4	D_5	
0.3711(0.0045)	1.6723(0.0044)	1.6911(0.0017)	2.9784(0.0017)	1.307068
0.0484(0.0023)	1.3561(0.0023)	with refitting		

5.3 Fréchet regression for symmetric positive definite matrices with a Cholesky decomposition metric

Consider Ω to be the set of symmetric, positive definite (SPD) matrices. Let $\mathbf{P}_1$ and $\mathbf{P}_2$ be two SPD matrices. Then, under the Cholesky decomposition, we can write $\mathbf{P}_1 = (\mathbf{P}_1^{1/2})^T\mathbf{P}_1^{1/2}$ and $\mathbf{P}_2 = (\mathbf{P}_2^{1/2})^T\mathbf{P}_2^{1/2}$, where $\mathbf{P}_1^{1/2}$ and $\mathbf{P}_2^{1/2}$ are upper triangle matrices with positive diagonal components. Then we define the Cholesky decomposition distance between $\mathbf{P}_1$ and $\mathbf{P}_2$ as $||\mathbf{P}_1^{1/2} - \mathbf{P}_2^{1/2}||_F$, where $||\cdot||_F$ is the Frobenius norm. That is, the Cholesky decomposition metric between two SPD matrices, $\mathbf{P}_1$ and $\mathbf{P}_2$, is given by

$$d_C(\mathbf{P}_1, \mathbf{P}_2) = \sqrt{\text{trace}\left((\mathbf{P}_1^{1/2} - \mathbf{P}_2^{1/2})^T(\mathbf{P}_1^{1/2} - \mathbf{P}_2^{1/2})\right)}.$$

We note that this distance is the same as the square root distance that has been considered in various statistical applications and is a special case of the Box-Cox class of matrix distances (Pigoli et al. 2014; Petersen and Müller 2016; Tavakoli et al. 2019).

For the following examples, data were generated as correlated scalar predictors $X_j \sim \mathcal{U}(0,2)$, $j = 1, 2, \ldots, p$, in two steps: (1) $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_p)^T$ multivariate Gaussian with $E(Z_j) = 0$ and $\mathrm{cov}(Z_j, Z_{j'}) = \rho^{|j-j'|}$; (2) $X_j = 2\Phi(Z_j)$ for $j = 1, \ldots, p$, where Φ is the standard normal distribution function. We set $p = 10$ and $\rho = 0.5$. Y, the random object of interest, is an $M \times M$ SPD matrix, $\mathbf{I}$ denotes an $M \times M$ identity matrix and $\mathbf{U} = (U_{i,j})$ denotes an $M \times M$ matrix where $U_{i,j} = I_{\{i<j\}}$.

Example 5.3.1. *The Fréchet regression function is given by*

$$m_{\oplus}(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x}) \quad = \quad E(\mathbf{A})^T E(\mathbf{A})$$

where $E(\mathbf{A}) = \{\mu_0 + \beta(x_1 + x_4) + \sigma_0 + \gamma x_4\}\mathbf{I} + \{\sigma_0 + \gamma x_4\}\mathbf{U}$.

Conditional on $\mathbf{X}$*, the random response* Y *is generated by adding noise as follows:* $Y = \mathbf{A}^T\mathbf{A}$*, where* $\mathbf{A} = (\mu + \sigma)\mathbf{I} + \sigma\mathbf{U}$ *and with* $\mu|\mathbf{X} \sim N(\mu_0 + \beta(X_1 + X_4), \nu_1)$ *and* $\sigma|\mathbf{X} \sim Gamma((\sigma_0 + \gamma X_4)^2/\nu_2, \nu_2/(\sigma_0 + \gamma X_4))$ *being independently sampled. Thus,* X_1 *and* X_4 *are the only important predictors in this example. The additional parameters are set as* $M = 3$*,* $\mu_0 = 3$*,* $\sigma_0 = 3$*,* $\beta = 3$*,* $\gamma = 2$*,* $\nu_1 = 1$*, and* $\nu_2 = 2$*.*

In the following two examples, we consider the effect of the parameter M, the dimension of the SPD matrices. For these examples, the Fréchet regression function is given by

$$m_{\oplus}(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x}) \quad = \quad E(\mathbf{A})^T E(\mathbf{A})$$

where $E(\mathbf{A}) = \{\mu_0 + \beta(x_1 + x_3) + \sigma_0 + \gamma(x_5 + x_7 + x_9)\}\,\mathbf{I} + \{\sigma_0 + \gamma(x_5 + x_7 + x_9)\}\mathbf{U}$.

Conditional on $\mathbf{X}$, the random response Y is generated by adding noise as follows: $Y = \mathbf{A}^T\mathbf{A}$ where $\mathbf{A} = (\mu + \sigma)\mathbf{I} + \sigma\mathbf{U}$ and with $\mu|\mathbf{X} \sim N(\mu_0 + \beta(X_1 + X_3), \nu_1)$ and $\sigma|\mathbf{X} \sim \mathrm{Gamma}((\sigma_0+\gamma(X_5+X_7+X_9))^2/\nu_2, \nu_2/(\sigma_0+\gamma(X_5+X_7+X_9)))$ being independently sampled. Thus, X_1, X_3, X_5, X_7 and X_9 are important predictors. We set $\mu_0 = 3$, $\sigma_0 = 3$, $\beta = 2$, $\gamma = 3$, $\nu_1 = 1$, and $\nu_2 = 2$ and choose the dimension of Y as follows.

Example 5.3.2. $M = 3$.

Example 5.3.3. $M = 5$.

We fixed $n = 200$ and used an independent validation data set of size n to tune the regularization parameter, as well as an independent test set of size $100n$ to evaluate the

performance of the final Fréchet regression function estimates. Results for 100 repetitions are reported in Tables 5 and 6 in the same way as for the previous simulation. The results suggest that FRiSO with refitting is an effective variable selection technique when the output is an SPD matrix.

Table 5: Simulation results (variable selection) of Fréchet regression for SPD matrix data

Example	selection frequency										Path
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	consistency
5.3.1	100	0	0	100	0	0	0	0	0	0	100
5.3.2	100	1	100	0	100	1	100	0	100	0	100
5.3.3	100	1	100	0	100	1	100	0	100	0	100
5.3.1	100	0	0	100	0	0	0	0	0	0	
5.3.2	100	1	99	0	100	1	100	0	100	0	with refitting
5.3.3	100	0	99	0	100	1	100	0	100	0	

Table 6: Simulation results on prediction for Fréchet regression for SPD matrix data

Example	Mean (standard error) over 100 repetitions				D_1
	D_2	D_3	D_4	D_5	
5.3.1	16.748 (1.648)	23.623 (1.575)	52.819 (1.701)	70.620 (1.605)	3.789
5.3.2	25.243 (1.746)	31.385 (1.760)	101.398 (1.764)	115.735 (1.753)	4.287
5.3.3	60.177 (4.336)	73.128 (4.373)	241.451 (4.332)	267.346 (4.331)	8.135
5.3.1	14.614 (1.598)	13.164 (1.535)			
5.3.2	15.060 (1.707)	14.132 (1.738)	with refitting		
5.3.3	35.312 (4.267)	35.160 (4.338)			

However, the prediction performance as quantified by D_2 is less compelling when comparing D_2 with D_1. This is likely due to the fact that the model bias $Ed^2(L_\oplus(\mathbf{X}), m_\oplus(\mathbf{X}))$ for the global Fréchet regression model could be big in this example because the data generation does not imply any "linear" dependence on the predictors.

For one random repetition in Example 5.3.3, we plot the corresponding solution path in Figure 2. Recall that in this example, X_1, X_3, X_5, X_7 and X_9 are important predictors. We find that as the regularization parameter τ increases, the optimal $\hat{\lambda}_j$ estimates

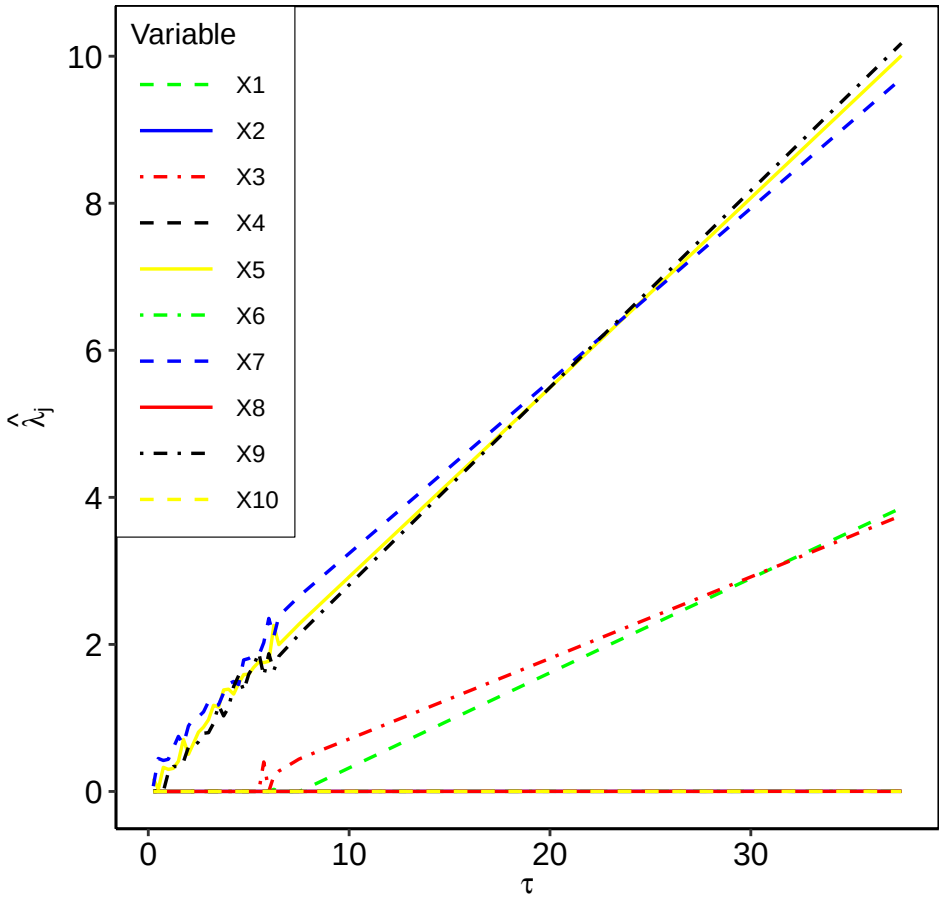

Figure 2: Solution path for a random repetition of Example 5.3.3, where X_1, X_3, X_5, X_7 and X_9 are important predictors.

corresponding to the important predictors increase far above those of the unimportant predictors, and the solution path is seen to be path consistent. Also, we note that predictors X_5, X_7, and X_9 are chosen before X_1 and X_3. This seems reasonable, as the simulation setup for Example 5.3.3 creates different patterns of dependency for these two sets of predictors.

5.4 Fréchet Regression for Spherical Data

In this example, we consider $\Omega = \mathcal{S}^2$, the unit sphere in $\mathfrak{R}^3$, with the geodesic distance $d(y, y') = \arccos(y^T y')$ for any $y, y' \in \mathcal{S}^2$. Correlated scalar predictors were generated in the same way as in the previous examples by first generating multivariate Gaussian

vectors $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_p)^T$ with $E(Z_j) = 0$ and $\text{cov}(Z_j, Z_{j'}) = \rho^{|j-j'|}$ and then applying the transformation $X_j = \Phi(Z_j)$ for $j = 1, \ldots, p$. The true Fréchet regression function was chosen as

$$m_{\oplus}(\mathbf{x}) = (\sqrt{1 - x_3^2}\cos(\pi(x_5 + x_7)), \sqrt{1 - x_3^2}\sin(\pi(x_5 + x_7)), x_3),$$

which implies that important predictors are X_3, X_5 and X_7.

Adopting an approach of Petersen and Müller (2019) to generate the responses, we obtained random observations $(\mathbf{X}_i, Y_i)$ by first generating a predictor vector $\mathbf{X}_i$ as above and a bivariate normal random vector $\mathbf{U}_i$ on the tangent space $T_{m_{\oplus}(\mathbf{X}_i)}\Omega$ and then a random response Y_i by

$$Y_i = \text{Exp}_{m_{\oplus}(\mathbf{X}_i)}(\mathbf{U}_i) = \cos(\|\mathbf{U}_i\|_E) m_{\oplus}(\mathbf{X}_i) + \sin(\|\mathbf{U}_i\|_E)\frac{\mathbf{U}_i}{\|\mathbf{U}_i\|_E},$$

where $\|\mathbf{U}_i\|_E$ denotes the Euclidean norm of $\mathbf{U}_i$ and Exp_p the exponential map on the manifold for the tangent plane at point $p \in \mathcal{S}^2$. The components of $\mathbf{U}_i$ were generated as independent random variables with standard deviation σ_U, for which we considered two different levels, as follows.

Example 5.4.1. $\sigma_U = 0.2$.

Example 5.4.2. $\sigma_U = 0.35$.

We fixed $n = 100$ and $p = 8$, and used an independent validation data set of size n to tune the regularization parameter, as well as an independent test set of size $100n$ to evaluate the performance of the final Fréchet regression function estimates. Results for 100 repetitions are reported in Tables 7 and 8 in the same way as for the previous two simulations.

We find that the proposed variable selection method again shows very good performance and the refitting step is seen to improve the variable selection performance. Similar to examples 5.3.1, 5.3.2 and 5.3.3, the prediction performance as quantified by D_2 is less compelling when comparing D_2 with D_1, which may be due to the fact that the data generation does not imply a globally linear dependence on the predictors.

6 Illustrations with Real Data

6.1 Bike Rental Distribution Regression

We first demonstrate the performance of FRiSO applied to real bike rental data originally collected by Capital Bikeshare in Washington D.C. This data set spans the years 2011

Table 7: Simulation results (variable selection) for Fréchet regression for spherical data

Example	selection frequency								Path consistency
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	
5.4.1	11	10	100	12	100	4	100	18	96
5.4.2	16	17	99	16	100	14	100	20	90
5.4.1	7	9	100	3	100	4	100	6	with refitting
5.4.2	6	12	95	6	100	11	100	8	

Table 8: Simulation results on prediction for Fréchet regression for spherical data

Example	Mean (standard error) over 100 repetitions				D_1
	D_2	D_3	D_4	D_5	
5.4.1	0.5000 (0.0390)	0.5727 (0.0400)	1.0796 (0.0196)	1.1712 (0.0188)	0.0798
5.4.2	0.5050 (0.0456)	0.7290 (0.0465)	1.0817 (0.0215)	1.3231 (0.0194)	0.2444
5.4.1	0.4847 (0.0417)	0.5540 (0.0420)	with refitting		
5.4.2	0.4881 (0.0509)	0.7079 (0.0503)			

and 2012 for a total of 731 days. For each day, there are 24 observations of bike rental counts as well as the following 8 predictors (Fanaee-T and Gama 2013):

- BW: Indicator of bad weather (misty and/or cloudy), standardized
- RBW: Indicator of really bad weather (snowy and/or rainy), standardized
- Holiday: Indicator of a public holiday celebrated in Washington D.C., standardized
- Work: Indicator of neither the weekend nor a holiday, standardized
- 2012: Indicator of the year 2012 , standardized
- Humid: Daily mean humidity, standardized
- Temp: Daily mean temperature, standardized
- Wind: Daily mean windspeed, standardized

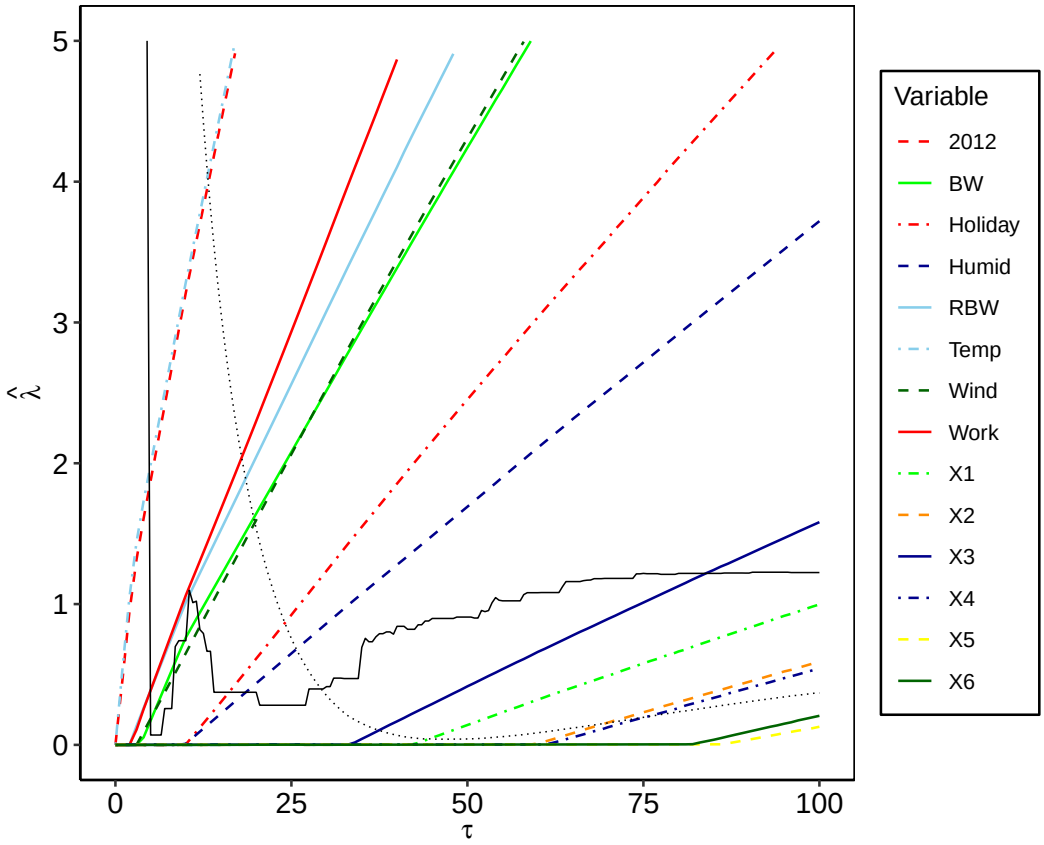

Figure 3: FRiSO solution path for the bike rental data.

We construct the response for each day to be the 24 observed quantiles for an underlying distribution of bike rental counts. To show the variable selection accuracy of FRiSO, we create 6 additional noise variables in the following way: $X_1 \sim N(0,1)$, $X_2 \sim \Gamma(\alpha = 1, \beta = 2)$, $X_3 \sim \Gamma(2,2)$, $X_4 \sim \Gamma(15,1)$, $X_5 \sim \Gamma(15,2)$ and $X_6 \sim N(0,1) + \sqrt{\Gamma(35,2)}$. All 14 predictors are standardized to have mean zero and variance one before applying FRiSO.

The FRiSO solution path is shown in Figure 3. To tune the regularization parameter τ, we use 10-fold cross validation. After appropriate rescaling, the 10-fold cross validation error is shown as the black dotted line, while the 10-fold cross validation error with refitting is shown as the black solid line in Figure 3. Thus, without refitting, we select all original predictors as well as two noise predictors. However, with refitting, we select only the 2012 indicator, temperature, workday indicator, really bad weather indicator, windspeed, and bad weather indicator variables.

We note that the variable selection result with refitting is quite reasonable. Capital

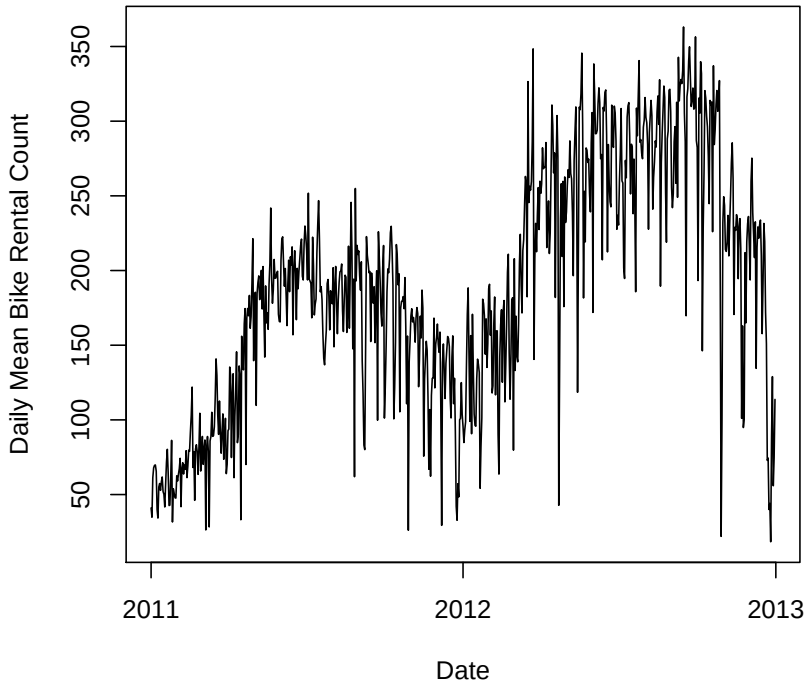

Figure 4: Average daily bike rentals from January 1, 2011 to December 31, 2012.

Bikeshare was launched in 2010. From 2011 to 2012, the company experienced a detectable shift in mean bike rentals per day (see Figure 4). This could explain why the year is an important variable to include in the predictor set. Further, temperature, windspeed, and bad or really bad weather likely all determine whether individuals are comfortable to be outside on a bicycle; workdays will have an impact on bike rentals. The exclusion of the holiday variable may be due to the workday variable capturing enough of this information, and the exclusion of the humidity variable is likely due to the other weather variables capturing closely related information.

Finally, the generalized R^2 (for details see Petersen and Müller 2019) for the global Fréchet regression with the predictors selected after 10-fold cross validation with refitting was found to be a very respectable $R^2 = .708$. (The generalized R^2 with the predictors selected after 10-fold cross validation wihtout refitting was found to be $R^2 = .713$). Figure 5 depicts the path of R^2 as τ increases.

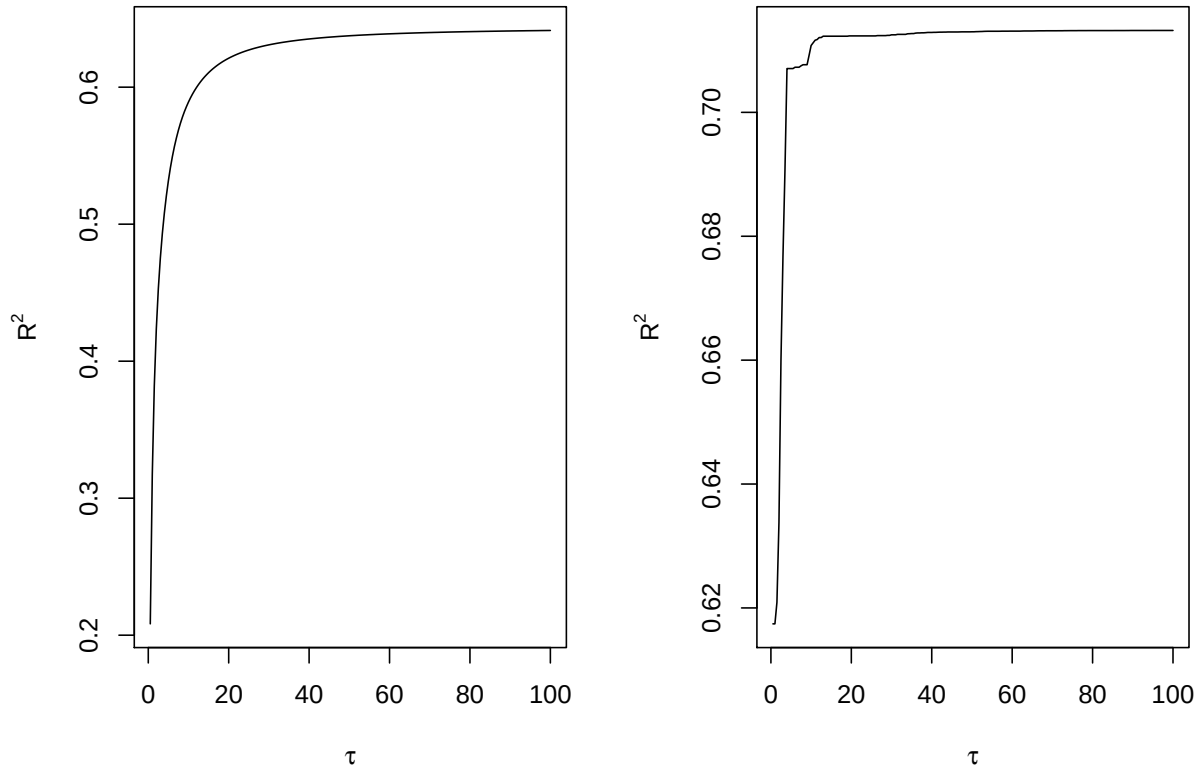

Figure 5: Generalized R^2 values for the bike rental data across τ for without refitting (left) and with refitting (right).

6.2 New York Taxi Network Regression

The New York City Taxi and Limousine Commission provides records on pick-up and drop-off dates and times, pick-up and drop-off locations, trip distances, itemized fares, and driver-reported passenger counts for yellow taxis. The data are available from `https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page`. We transform these data into network or graph data, where neighborhoods are nodes and edges are weighted by the number of taxi rides which picked up in one neighborhood and dropped off in another within a single hour. After proper transformation, these graphs can lie in a metric space of SPD matrices equipped with the Choleksy decomposition distance, as in Section 5.3.

To engineer SPD matrices from the taxi data, we specifically do the following:

1. We filter the data on the month of January 2016 due to resource restrictions.
2. We further filter on observations with both pick-up and drop-off occuring in Manhattan.
3. We then label the corresponding neighborhood for each pick-up and drop-off in the same manner as Dubey and Müller (2020). For specific details, see section 3.2.4 in the supplementary materials.
4. For each hour, we collect the number of pairwise connections between nodes based on taxi pick-ups and drop-offs. These correspond to weights between nodes on a graph.

This yields 723 weighted adjacency matrices of dimension 10×10 for these data (removing a small handful of observations due to their sparsity). To ensure that these outputs are truly SPD matrices, we further square them.

From the taxi data, we also collect the following nine potential predictors, with values averaged over each hour:

- Ave. Distance: Mean distance travelled, 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
- Vendor: Sum of the vendor indicators, standardized (in the original data, the vendors for the recording devices installed in each taxi are coded as 0 = Creative Mobile Technologies, LLC; 1 = VeriFone Inc.)

From https://www.wunderground.com/history/daily/us/ny/new-york-city/KLGA/date, we further collect New York City weather history for January 2016. The following six weather variables were included as potential predictors:

- Day's Ave. Humid.: Daily mean humidity, standardized
- Day's Ave. Press.: Daily mean barometric pressure, standardized
- Day's Ave. Temp.: Daily mean temperature, standardized
- Day's Ave. Wind: Daily mean windspeed, standardized
- Day's Total Precip.: Daily total precipitation, standardized

This then leads to a total of fifteen potential predictors.

To tune the regularization parameter, we randomly split the data into a training set of size 361 and a validation set of size 362. In Figure 6, we plot the FRiSO solution path applied to the training data across $\tau = \{0.5, 1, \ldots, 24.5, 25\}$. After appropriate rescaling, the validation error is shown as the black dotted line, while the validation error with refitting is shown as the black solid line in Figure 6. Thus, with refitting, we select Credit, Cash, Late Hour, Ave. Distance, Day's Ave. Humidity, and Day's Ave. Temp.

This variable selection seems reasonable. Excluding Day's Ave Wind and Day's Total Precip. could be due to the fact that windspeed and precipitation change rapidly. Therefore, daily measurements of these features may be too smooth to explain hourly fluctuations in the taxi networks. Excluding Day's Ave. Press. is intuitive, as barometric pressure likely has less impact than humidity and precipitation. Next, disputed and free rides occur very infrequently in the raw data and thus may not have a strong impact on a large taxi network, explaining the exclusion of Dispute and Free. Not selecting Ave. Fare,

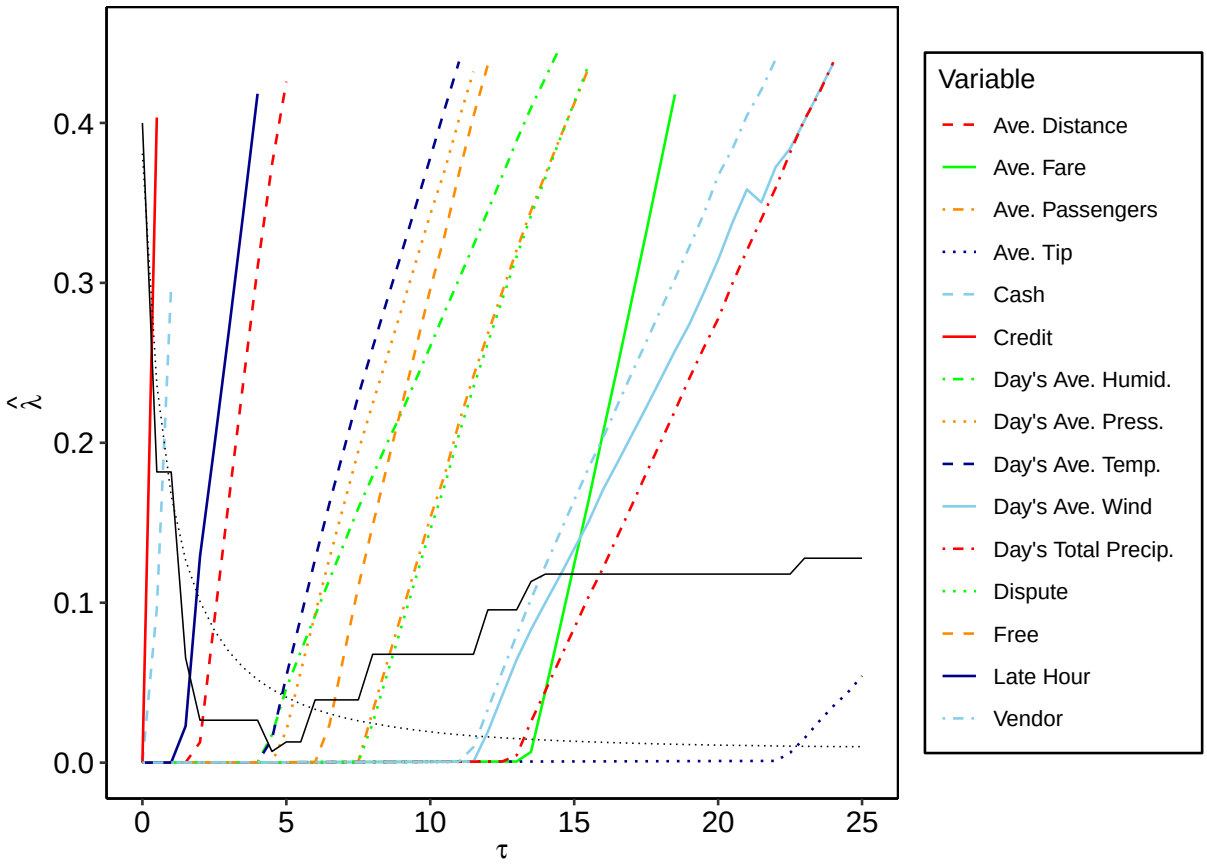

Figure 6: FRiSO solution path for the taxi network data.

Ave. Passengers, and Ave. Tip could be because Ave. Distance captures much of the information these variables provide. Also, the variation in Ave. Passengers is very low. Leaving Vendor out seems reasonable, as the choice of installed recording device may be mostly random and may not any capture underlying differences in driver behavior. Finally, including Late Hour is intuitive, as the neighborhoods connected between 11pm and 5am may differ quite a bit from the neighborhoods connected between 5am and 11pm.

If we analyze the dotted error curve which is computed without refitting, we see that it does not perform any variable selection as desired. Recall that in the bike rental example, the error without refitting selected all of the authentic predictors as well and did not begin to increase until noise variables were included. This is likely due to inherent ridge bias introduced when we do not refit as well as the very strong correlations that exist between the potential predictors in both data sets. Thus, we further emphasize our recommendation to apply the proposed FRiSO method for variable selection with the refitting option.

Finally, the generalized R^2 for the global Fréchet regression with the predictors selected after validation with refitting was found to be a respectable $R^2 = 0.492$. (The generalized R^2 with all predictors selected after validation without refitting was found to be $R^2 = 0.499$.)

6.3 Vectorcardiogram Spherical Regression

We finally demonstrate the effectiveness of FRiSO applied to a dataset derived from vectorcardiogram measurements. The vectorcardiogram is a method used to record the magnitude and direction of the electrical forces generated during heartbeats. It connects three leads to the torso to generate a time-dependent vector that traces three-dimensional approximately closed curves, each representing a heartbeat cycle. As a summary to aid clinical diagnosis, a unit vector defined as the directional components of the vector at a particular extremum across the cycle is often used (Paine et al. 2020). The vectorcardiogram data consist of such defined unit vectors or directional components using two different lead placement systems, the Frank system and the McFee system, and have been obtained from vectorcardiogram measurements of the cardiac electrical activity of 98 children of different age and gender (Paine et al. 2020). This data set has been previously analyzed by Chang (1986) and Paine et al. (2020). We consider here responses that are the unit vectors obtained with the Frank system and view these unit vector summaries that are reported for each subject as spherical data on the sphere S^2.

When applying the proposed FRiSO approach, the spherical response vector derived

from the Frank system is related to the predictors that were recorded in the data set and include age, gender, and X_3, X_4, X_5, which are the three components of the unit vector obtained with the McFee system. While X_3, X_4, X_5 are continuous predictors, X_1 and X_2 are coded as binary variables, with $X_1 = 1$ indicating membership in the age group 11-19 and $X_1 = 0$ indicating membership in the age group 2-10. Males and females are coded as $X_2 = 1$ and $X_2 = 0$, respectively. The predictors are then standardized to have mean zero and variance one before applying FRiSO.

The FRiSO solution path is shown in Figure 7. To tune the regularization parameter τ, we use leave-one-out cross validation. After appropriate rescaling, the leave-one-out cross validation error is shown as the black dotted line in Figure 7 and thus the leave-one-out cross validation selects a final model with predictors X_1, X_3, X_4 and X_5. The generalized R^2 (for details see Petersen and Müller 2019) for the global Fréchet regression with the selected predictors X_1, X_3, X_4 and X_5 was found to be a sizable $R^2 = .462$. This indicates that all predictors except gender are relevant for the response.

7 Concluding Remarks

In this work, we have proposed a novel variable selection method for global Fréchet regression and have demonstrated that it achieves selection consistency, affirming its effectiveness. We have further provided simulation and real data examples to demonstrate its competitive finite sample performance. The inference capability of this method as gleaned from the resulting selection paths as well as the generalized R^2 for random object responses was found to be reasonable. We note that there is great potential to develop further inference, such as a generalized variable importance measures or post-selection inference (Berk et al. 2013). Finally, we emphasize that FRiSO extends the capability of global Fréchet regression to be applied not only to complex data, but also to complex data coupled with high-dimensional predictors where p is large. Such data are becoming increasingly common.

Supplementary materials

Online supplementary materials contains the implementation details of the proposed FRiSO as well as implementation codes in R and Matlab.

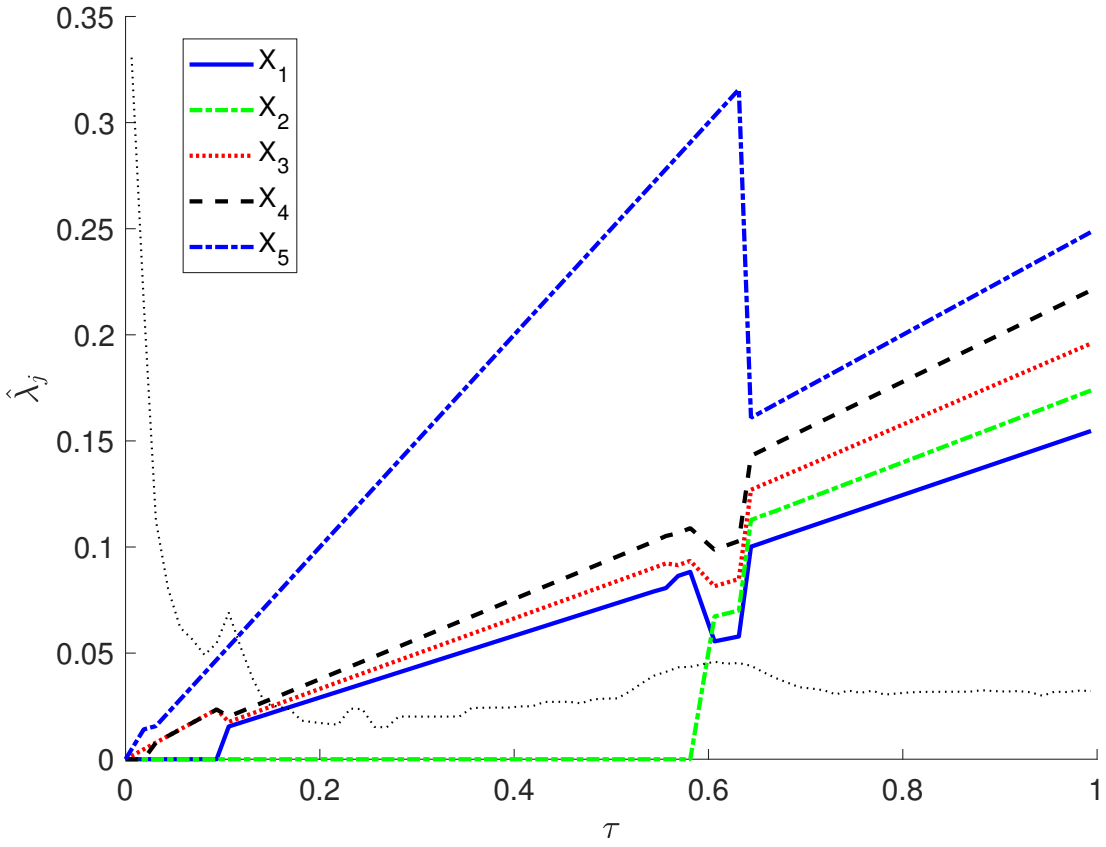

Figure 7: FRiSO solution path for the vectorcardiogram data.

Acknowledgments

The authors would like to thank two anonymous referees, an Associate Editor, and co-Editor Prof. Marina Vannucci for their constructive comments, leading to substantial improvements in the paper.

Appendix

Proof of Proposition 1. The proof is analogous to the proof of Theorem 2 in Petersen and Müller (2019). The only difference is that Lemma 1 in the current context refers to the individually penalized ridge Fréchet regression while Theorem 2 in Petersen and Müller (2019) refers to the global Fréchet regression. We skip the details. □

Proof of Lemma 1. For any $\mathcal{A}$ satisfying $\mathcal{I} \subseteq \mathcal{A} \subseteq \{1, 2, \ldots, p\}$, we have

$$\begin{aligned}
& E_{(\mathbf{X}_\mathcal{A}, Y)} \left[1 + (\mathbf{x}_\mathcal{A} - \boldsymbol{\mu}_\mathcal{A})^T \boldsymbol{\Sigma}_{\mathcal{A},\mathcal{A}}^{-1} (\mathbf{X}_\mathcal{A} - \boldsymbol{\mu}_\mathcal{A})\right] d^2(Y, \omega) \\
= \; & E\left\{E(\left[1 + (\mathbf{x}_\mathcal{A} - \boldsymbol{\mu}_\mathcal{A})^T \boldsymbol{\Sigma}_{\mathcal{A},\mathcal{A}}^{-1} (\mathbf{X}_\mathcal{A} - \boldsymbol{\mu}_\mathcal{A})\right] d^2(Y, \omega) | \mathbf{X}_\mathcal{I})\right\} \\
= \; & E\left\{E(\left[1 + (\mathbf{x}_\mathcal{A} - \boldsymbol{\mu}_\mathcal{A})^T \boldsymbol{\Sigma}_{\mathcal{A},\mathcal{A}}^{-1} (\mathbf{X}_\mathcal{A} - \boldsymbol{\mu}_\mathcal{A})\right] | \mathbf{X}_\mathcal{I}) E(d^2(Y, \omega) | \mathbf{X}_\mathcal{I})\right\} \\
= \; & E\left\{\left[1 + (\mathbf{x}_\mathcal{I} - \boldsymbol{\mu}_\mathcal{I})^T \boldsymbol{\Sigma}_{\mathcal{I},\mathcal{I}}^{-1} (\mathbf{X}_\mathcal{I} - \boldsymbol{\mu}_\mathcal{I})\right] E(d^2(Y, \omega) | \mathbf{X}_\mathcal{I})\right\} \\
= \; & E\left\{E(\left[1 + (\mathbf{x}_\mathcal{I} - \boldsymbol{\mu}_\mathcal{I})^T \boldsymbol{\Sigma}_{\mathcal{I},\mathcal{I}}^{-1} (\mathbf{X}_\mathcal{I} - \boldsymbol{\mu}_\mathcal{I})\right] d^2(Y, \omega) | \mathbf{X}_\mathcal{I})\right\} \\
= \; & E_{(\mathbf{X}_\mathcal{I}, Y)} \left[1 + (\mathbf{x}_\mathcal{I} - \boldsymbol{\mu}_\mathcal{I})^T \boldsymbol{\Sigma}_{\mathcal{I},\mathcal{I}}^{-1} (\mathbf{X}_\mathcal{I} - \boldsymbol{\mu}_\mathcal{I})\right] d^2(Y, \omega) \text{ for any } \mathbf{x} \in \mathcal{X},
\end{aligned}$$

which implies

$$\begin{aligned}
L_\oplus^\mathcal{A}(\mathbf{x}) &= \arg\min_{\omega \in \Omega} E_{(\mathbf{X}_\mathcal{A}, Y)} \left[1 + (\mathbf{x}_\mathcal{A} - \boldsymbol{\mu}_\mathcal{A})^T \boldsymbol{\Sigma}_{\mathcal{A},\mathcal{A}}^{-1} (\mathbf{X}_\mathcal{A} - \boldsymbol{\mu}_\mathcal{A})\right] d^2(Y, \omega) \\
&= \arg\min_{\omega \in \Omega} E_{(\mathbf{X}_\mathcal{I}, Y)} \left[1 + (\mathbf{x}_\mathcal{I} - \boldsymbol{\mu}_\mathcal{I})^T \boldsymbol{\Sigma}_{\mathcal{I},\mathcal{I}}^{-1} (\mathbf{X}_\mathcal{I} - \boldsymbol{\mu}_\mathcal{I})\right] d^2(Y, \omega) \\
&= L_\oplus^\mathcal{I}(\mathbf{x}) \text{ for any } \mathbf{x} \in \mathcal{X}.
\end{aligned}$$

Consequently

$$m_\oplus(\mathbf{x}) = L_\oplus(\mathbf{x}) = L_\oplus^{\{1,2,\ldots,p\}}(\mathbf{x}) = L_\oplus^\mathcal{I}(\mathbf{x}) = L_\oplus^\mathcal{A}(\mathbf{x}) \text{ for any } \mathbf{x} \in \mathcal{X}$$

for any $\mathcal{A}$ satisfying $\mathcal{I} \subseteq \mathcal{A} \subseteq \{1, 2, \ldots, p\}$. □

Proof of Proposition 2. With the squared error distance $d^2(Y,\omega) = (Y-\omega)^2$, we have

$$\begin{aligned}
R_\oplus(\mathbf{x};\boldsymbol{\nu}) &= \arg\min_{\omega\in\Omega} E_{(\mathbf{X},Y)} \left\{ \left[1 + (\mathbf{x}-\boldsymbol{\mu})^T \left[\boldsymbol{\Sigma} + \mathrm{diag}(\boldsymbol{\nu})\right]^{-1} (\mathbf{X}-\boldsymbol{\mu})\right] d^2(Y,\omega) \right\}. \\
&= \arg\min_{\omega\in\Omega} E_{(\mathbf{X},Y)} \left\{ \left[1 + (\mathbf{x}-\boldsymbol{\mu})^T \left[\boldsymbol{\Sigma} + \mathrm{diag}(\boldsymbol{\nu})\right]^{-1} (\mathbf{X}-\boldsymbol{\mu})\right] (Y-\omega)^2 \right\}. \\
&= \arg\min_{\omega\in\Omega} \left\{ \omega^2 - 2\omega(\beta_0 + \boldsymbol{\mu}^T\boldsymbol{\beta} + (\mathbf{x}-\boldsymbol{\mu})^T \left[\boldsymbol{\Sigma} + \mathrm{diag}(\boldsymbol{\nu})\right]^{-1} \mathrm{cov}(\mathbf{X},Y)) + \mathrm{constant} \right\} \\
&= \beta_0 + \boldsymbol{\mu}^T\boldsymbol{\beta} + (\mathbf{x}-\boldsymbol{\mu})^T \left[\boldsymbol{\Sigma} + \mathrm{diag}(\boldsymbol{\nu})\right]^{-1} \mathrm{cov}(\mathbf{X},Y) \\
&= \beta_0 + \boldsymbol{\mu}^T\boldsymbol{\beta} + (\mathbf{x}-\boldsymbol{\mu})^T \left[\boldsymbol{\Sigma} + \mathrm{diag}(\boldsymbol{\nu})\right]^{-1} \boldsymbol{\Sigma}\boldsymbol{\beta} \qquad (18)
\end{aligned}$$

and $L_\oplus(\mathbf{x}) = R_\oplus(\mathbf{x};\mathbf{0}) = \beta_0 + \mathbf{x}^T\boldsymbol{\beta}$, and furthermore

$$\begin{aligned}
L_\oplus^{\mathcal{A}}(\mathbf{x}) &= \beta_0 + \boldsymbol{\mu}^T\boldsymbol{\beta} + (\mathbf{x}_\mathcal{A} - \boldsymbol{\mu}_\mathcal{A})^T \left[\boldsymbol{\Sigma}_{\mathcal{A},\mathcal{A}}\right]^{-1} (\boldsymbol{\Sigma}_{\mathcal{A},\mathcal{A}}\boldsymbol{\beta}_\mathcal{A} + \boldsymbol{\Sigma}_{\mathcal{A},\mathcal{A}^c}\boldsymbol{\beta}_{\mathcal{A}^c}) \\
&= \beta_0 + \mathbf{x}_\mathcal{A}^T\boldsymbol{\beta}_\mathcal{A} + \boldsymbol{\mu}_{\mathcal{A}^c}^T\boldsymbol{\beta}_{\mathcal{A}^c} + (\mathbf{x}_\mathcal{A} - \boldsymbol{\mu}_\mathcal{A})^T \left[\boldsymbol{\Sigma}_{\mathcal{A},\mathcal{A}}\right]^{-1} \boldsymbol{\Sigma}_{\mathcal{A},\mathcal{A}^c}\boldsymbol{\beta}_{\mathcal{A}^c}.
\end{aligned}$$

This implies

$$\begin{aligned}
Ed^2(L_\oplus(\mathbf{X}), L_\oplus^{\mathcal{A}}(\mathbf{X})) &= E\left((\mathbf{X}_{\mathcal{A}^c} - \boldsymbol{\mu}_{\mathcal{A}^c}^T)\boldsymbol{\beta}_{\mathcal{A}^c} - (\mathbf{X}_\mathcal{A} - \boldsymbol{\mu}_\mathcal{A})^T \left[\boldsymbol{\Sigma}_{\mathcal{A},\mathcal{A}}\right]^{-1} \boldsymbol{\Sigma}_{\mathcal{A},\mathcal{A}^c}\boldsymbol{\beta}_{\mathcal{A}^c}\right)^2 \\
&= \boldsymbol{\beta}_{\mathcal{A}^c}^T \left(\boldsymbol{\Sigma}_{\mathcal{A}^c,\mathcal{A}^c} - \boldsymbol{\Sigma}_{\mathcal{A}^c,\mathcal{A}} \left[\boldsymbol{\Sigma}_{\mathcal{A},\mathcal{A}}\right]^{-1} \boldsymbol{\Sigma}_{\mathcal{A},\mathcal{A}^c}\right) \boldsymbol{\beta}_{\mathcal{A}^c}.
\end{aligned}$$

One verifies that Condition [B] is satisfied by noting that $\boldsymbol{\Sigma}_{\mathcal{A}^c,\mathcal{A}^c} - \boldsymbol{\Sigma}_{\mathcal{A}^c,\mathcal{A}} \left[\boldsymbol{\Sigma}_{\mathcal{A},\mathcal{A}}\right]^{-1} \boldsymbol{\Sigma}_{\mathcal{A},\mathcal{A}^c}$ is positive definite as long as $\boldsymbol{\Sigma}$ is positive definite. □

Proof of Proposition 3. With the squared error distance $d^2(Y,\omega) = (Y-\omega)^2$, we have

$$\begin{aligned}
R_\oplus(\mathbf{x};\boldsymbol{\nu}) &= \arg\min_{\omega\in\Omega} E_{(\mathbf{X},Y)} \left\{ \left[1 + (\mathbf{x}-\boldsymbol{\mu})^T \left[\boldsymbol{\Sigma} + \mathrm{diag}(\boldsymbol{\nu})\right]^{-1} (\mathbf{X}-\boldsymbol{\mu})\right] d^2(Y,\omega) \right\}. \\
&= \arg\min_{\omega\in\Omega} E_{(\mathbf{X},Y)} \left\{ \left[1 + (\mathbf{x}-\boldsymbol{\mu})^T \left[\boldsymbol{\Sigma} + \mathrm{diag}(\boldsymbol{\nu})\right]^{-1} (\mathbf{X}-\boldsymbol{\mu})\right] (Y-\omega)^2 \right\}. \\
&= \arg\min_{\omega\in\Omega} \left\{ \omega^2 - 2\omega(\beta_0 + \boldsymbol{\mu}^T\boldsymbol{\beta} + (\mathbf{x}-\boldsymbol{\mu})^T \left[\boldsymbol{\Sigma} + \mathrm{diag}(\boldsymbol{\nu})\right]^{-1} \mathrm{cov}(\mathbf{X},Y)) + \mathrm{constant} \right\} \\
&= \beta_0 + \boldsymbol{\mu}^T\boldsymbol{\beta} + (\mathbf{x}-\boldsymbol{\mu})^T \left[\boldsymbol{\Sigma} + \mathrm{diag}(\boldsymbol{\nu})\right]^{-1} \mathrm{cov}(\mathbf{X},Y) \\
&= \beta_0 + \boldsymbol{\mu}^T\boldsymbol{\beta} + (\mathbf{x}-\boldsymbol{\mu})^T \left[\boldsymbol{\Sigma} + \mathrm{diag}(\boldsymbol{\nu})\right]^{-1} \boldsymbol{\Sigma}\boldsymbol{\beta} \qquad (19)
\end{aligned}$$

and $L_\oplus(\mathbf{x}) = R_\oplus(\mathbf{x};\mathbf{0}) = \beta_0 + \mathbf{x}^T\boldsymbol{\beta}$.

Consequently

$$\begin{aligned}
& E_\mathbf{X} d^2(L_\oplus(\mathbf{X}), R_\oplus(\mathbf{X};\boldsymbol{\nu})) \\
&= E_\mathbf{X}\left(\beta_0 + \mathbf{X}^T\boldsymbol{\beta} - \left\{\beta_0 + \boldsymbol{\mu}^T\boldsymbol{\beta} + (\mathbf{X}-\boldsymbol{\mu})^T \left[\boldsymbol{\Sigma} + \mathrm{diag}(\boldsymbol{\nu})\right]^{-1} \boldsymbol{\Sigma}\boldsymbol{\beta}\right\}\right)^2 \\
&= E_\mathbf{X}\left((\mathbf{X}-\boldsymbol{\mu})^T \left[\boldsymbol{\Sigma} + \mathrm{diag}(\boldsymbol{\nu})\right]^{-1} \mathrm{diag}(\boldsymbol{\nu})\boldsymbol{\beta}\right)^2 \\
&= \boldsymbol{\beta}^T \mathrm{diag}(\boldsymbol{\nu}) \left[\boldsymbol{\Sigma} + \mathrm{diag}(\boldsymbol{\nu})\right]^{-1} \boldsymbol{\Sigma} \left[\boldsymbol{\Sigma} + \mathrm{diag}(\boldsymbol{\nu})\right]^{-1} \mathrm{diag}(\boldsymbol{\nu})\boldsymbol{\beta}.
\end{aligned}$$

Note $\| \text{ diag}(\boldsymbol{\nu})\boldsymbol{\beta} \| > 0$ for any $\boldsymbol{\nu}$ satisfying $\| \boldsymbol{\nu}_{\mathcal{I}} \| > 0$ due to the definition of the important set $\mathcal{I}$. As a result $\boldsymbol{\beta}^T \text{diag}(\boldsymbol{\nu}) \left[\boldsymbol{\Sigma} + \text{diag}(\boldsymbol{\nu})\right]^{-1} \boldsymbol{\Sigma} \left[\boldsymbol{\Sigma} + \text{diag}(\boldsymbol{\nu})\right]^{-1} \text{diag}(\boldsymbol{\nu})\boldsymbol{\beta} > 0$, by noting that both $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma} + \text{diag}(\boldsymbol{\nu})$ are strictly positive definite. □

Proof of Proposition 4. From the above proof of Proposition 3, we have

$$E_{\mathbf{X}} d^2(L_{\oplus}(\mathbf{X}), R_{\oplus}(\mathbf{X}; \boldsymbol{\nu})) = \boldsymbol{\beta}^T \text{diag}(\boldsymbol{\nu}) \left[\boldsymbol{\Sigma} + \text{diag}(\boldsymbol{\nu})\right]^{-1} \boldsymbol{\Sigma} \left[\boldsymbol{\Sigma} + \text{diag}(\boldsymbol{\nu})\right]^{-1} \text{diag}(\boldsymbol{\nu})\boldsymbol{\beta}.$$

Let $\mathbf{E}_j$ be a $p \times p$ matrix, whose elements are all zero except the (j, j) element being 1. Taking the first order derivative, we have

$$\begin{aligned}
& \frac{d}{d\nu_j} \left\{ \boldsymbol{\beta}^T \text{diag}(\boldsymbol{\nu}) \left[\boldsymbol{\Sigma} + \text{diag}(\boldsymbol{\nu})\right]^{-1} \boldsymbol{\Sigma} \left[\boldsymbol{\Sigma} + \text{diag}(\boldsymbol{\nu})\right]^{-1} \text{diag}(\boldsymbol{\nu})\boldsymbol{\beta} \right\} \\
= \; & \boldsymbol{\beta}^T \mathbf{E}_j \left[\boldsymbol{\Sigma} + \text{diag}(\boldsymbol{\nu})\right]^{-1} \boldsymbol{\Sigma} \left[\boldsymbol{\Sigma} + \text{diag}(\boldsymbol{\nu})\right]^{-1} \text{diag}(\boldsymbol{\nu})\boldsymbol{\beta} \\
& - \boldsymbol{\beta}^T \text{diag}(\boldsymbol{\nu}) \left[\boldsymbol{\Sigma} + \text{diag}(\boldsymbol{\nu})\right]^{-1} \mathbf{E}_j \left[\boldsymbol{\Sigma} + \text{diag}(\boldsymbol{\nu})\right]^{-1} \boldsymbol{\Sigma} \left[\boldsymbol{\Sigma} + \text{diag}(\boldsymbol{\nu})\right]^{-1} \text{diag}(\boldsymbol{\nu})\boldsymbol{\beta} \\
& - \boldsymbol{\beta}^T \text{diag}(\boldsymbol{\nu}) \left[\boldsymbol{\Sigma} + \text{diag}(\boldsymbol{\nu})\right]^{-1} \boldsymbol{\Sigma} \left[\boldsymbol{\Sigma} + \text{diag}(\boldsymbol{\nu})\right]^{-1} \mathbf{E}_j \left[\boldsymbol{\Sigma} + \text{diag}(\boldsymbol{\nu})\right]^{-1} \text{diag}(\boldsymbol{\nu})\boldsymbol{\beta} \\
& + \boldsymbol{\beta}^T \text{diag}(\boldsymbol{\nu}) \left[\boldsymbol{\Sigma} + \text{diag}(\boldsymbol{\nu})\right]^{-1} \boldsymbol{\Sigma} \left[\boldsymbol{\Sigma} + \text{diag}(\boldsymbol{\nu})\right]^{-1} \mathbf{E}_j \boldsymbol{\beta} \\
= \; & 2\boldsymbol{\beta}^T \boldsymbol{\Sigma} \left[\boldsymbol{\Sigma} + \text{diag}(\boldsymbol{\nu})\right]^{-1} \mathbf{E}_j \left[\boldsymbol{\Sigma} + \text{diag}(\boldsymbol{\nu})\right]^{-1} \boldsymbol{\Sigma} \left[\boldsymbol{\Sigma} + \text{diag}(\boldsymbol{\nu})\right]^{-1} \text{diag}(\boldsymbol{\nu})\boldsymbol{\beta}. \qquad (20)
\end{aligned}$$

Taking another layer of partial derivatives, we get

$$\begin{aligned}
& \frac{d^2}{d\nu_k d\nu_j} \left\{ \boldsymbol{\beta}^T \text{diag}(\boldsymbol{\nu}) \left[\boldsymbol{\Sigma} + \text{diag}(\boldsymbol{\nu})\right]^{-1} \boldsymbol{\Sigma} \left[\boldsymbol{\Sigma} + \text{diag}(\boldsymbol{\nu})\right]^{-1} \text{diag}(\boldsymbol{\nu})\boldsymbol{\beta} \right\} \\
= \; & -2\boldsymbol{\beta}^T \boldsymbol{\Sigma} \left[\boldsymbol{\Sigma} + \text{diag}(\boldsymbol{\nu})\right]^{-1} \mathbf{E}_k \left[\boldsymbol{\Sigma} + \text{diag}(\boldsymbol{\nu})\right]^{-1} \mathbf{E}_j \left[\boldsymbol{\Sigma} + \text{diag}(\boldsymbol{\nu})\right]^{-1} \boldsymbol{\Sigma} \left[\boldsymbol{\Sigma} + \text{diag}(\boldsymbol{\nu})\right]^{-1} \text{diag}(\boldsymbol{\nu})\boldsymbol{\beta} \\
& - 2\boldsymbol{\beta}^T \boldsymbol{\Sigma} \left[\boldsymbol{\Sigma} + \text{diag}(\boldsymbol{\nu})\right]^{-1} \mathbf{E}_j \left[\boldsymbol{\Sigma} + \text{diag}(\boldsymbol{\nu})\right]^{-1} \mathbf{E}_k \left[\boldsymbol{\Sigma} + \text{diag}(\boldsymbol{\nu})\right]^{-1} \boldsymbol{\Sigma} \left[\boldsymbol{\Sigma} + \text{diag}(\boldsymbol{\nu})\right]^{-1} \text{diag}(\boldsymbol{\nu})\boldsymbol{\beta} \\
& - 2\boldsymbol{\beta}^T \boldsymbol{\Sigma} \left[\boldsymbol{\Sigma} + \text{diag}(\boldsymbol{\nu})\right]^{-1} \mathbf{E}_j \left[\boldsymbol{\Sigma} + \text{diag}(\boldsymbol{\nu})\right]^{-1} \boldsymbol{\Sigma} \left[\boldsymbol{\Sigma} + \text{diag}(\boldsymbol{\nu})\right]^{-1} \mathbf{E}_k \left[\boldsymbol{\Sigma} + \text{diag}(\boldsymbol{\nu})\right]^{-1} \text{diag}(\boldsymbol{\nu})\boldsymbol{\beta} \\
& + 2\boldsymbol{\beta}^T \boldsymbol{\Sigma} \left[\boldsymbol{\Sigma} + \text{diag}(\boldsymbol{\nu})\right]^{-1} \mathbf{E}_j \left[\boldsymbol{\Sigma} + \text{diag}(\boldsymbol{\nu})\right]^{-1} \boldsymbol{\Sigma} \left[\boldsymbol{\Sigma} + \text{diag}(\boldsymbol{\nu})\right]^{-1} \mathbf{E}_k \boldsymbol{\beta}. \qquad (21)
\end{aligned}$$

Recalling the definition of important set $\mathcal{I} = \{j : \beta_j \neq 0\}$ in the case of a classical linear regression, for any $\boldsymbol{\nu}$ with $\nu_j = 0$ for $j \in \mathcal{I}$, we have $\text{diag}(\boldsymbol{\nu})\boldsymbol{\beta} = \mathbf{0}$. Then (20) implies for the gradient $\frac{\partial}{\partial \boldsymbol{\nu}_{\mathcal{I}}} E_{\mathbf{X}} d^2(L_{\oplus}(\mathbf{X}), R_{\oplus}(\mathbf{X}; \boldsymbol{\nu})) = \mathbf{0}$ for any $\boldsymbol{\nu}$ with $\nu_j = 0$ for $j \in \mathcal{I}$, and (21) implies for the Hessian

$$\begin{aligned}
& \frac{\partial^2}{\partial \boldsymbol{\nu}_{\mathcal{I}} \partial \boldsymbol{\nu}_{\mathcal{I}}^T} E_{\mathbf{X}} d^2(L_{\oplus}(\mathbf{X}), R_{\oplus}(\mathbf{X}; \boldsymbol{\nu})) \\
= \; & \left(2\boldsymbol{\beta}^T \boldsymbol{\Sigma} \left[\boldsymbol{\Sigma} + \text{diag}(\boldsymbol{\nu})\right]^{-1} \mathbf{E}_j \left[\boldsymbol{\Sigma} + \text{diag}(\boldsymbol{\nu})\right]^{-1} \boldsymbol{\Sigma} \left[\boldsymbol{\Sigma} + \text{diag}(\boldsymbol{\nu})\right]^{-1} \mathbf{E}_k \boldsymbol{\beta} \right)
\end{aligned}$$

with $j, k \in \mathcal{I}$ for any $\boldsymbol{\nu}$ with $\nu_j = 0$ for $j \in \mathcal{I}$.

For the rest of the proof, we assume without loss of generality that $q = |\mathcal{I}|$, the cardinality of $\mathcal{I}$, and $\mathcal{I} = \{1, 2, \ldots, q\}$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_q, 0, \ldots, 0)^T$ and $\boldsymbol{\nu} = (0, 0, \ldots, 0, \nu_{q+1}, \nu_{q+2}, \ldots, \nu_p)^T$ are then sparsely populated vectors with zero components. In addition, $\nu_{q+1} \geq 0$, $\nu_{q+2} \geq 0$, ..., $\nu_p \geq 0$.

Note that $\boldsymbol{\Sigma}$ is a strictly positive definite $p \times p$ matrix. Let $\mathrm{diag}(\boldsymbol{\nu})$ denote the diagonal matrix with the elements of the vector $\boldsymbol{\nu}$ sitting on the diagonal and $\mathbf{M}$ be the $q \times q$ matrix with (j, k)-th element

$$M(j,k) = \boldsymbol{\beta}^T \boldsymbol{\Sigma} \left[\boldsymbol{\Sigma} + \mathrm{diag}(\boldsymbol{\nu})\right]^{-1} \mathbf{E}_j \left[\boldsymbol{\Sigma} + \mathrm{diag}(\boldsymbol{\nu})\right]^{-1} \boldsymbol{\Sigma} \left[\boldsymbol{\Sigma} + \mathrm{diag}(\boldsymbol{\nu})\right]^{-1} \mathbf{E}_k \boldsymbol{\beta},$$

where $\mathbf{E}_j$ is a $p \times p$ matrix, whose elements are all zero except the (j, j) element being 1. To complete the proof, it will suffice to show that $\mathbf{M}$ is strictly positive definite. We essentially need to prove that

$$\sum_{j=1}^{q} \sum_{k=1}^{q} c_j M(j,k) c_k > 0 \text{ as long as } \| \mathbf{c} \| \neq 0,$$

where $\mathbf{c} = (c_1, c_2, \ldots, c_q)^T$. Denoting $D_{\mathbf{c}} = \mathrm{diag}(\mathbf{c})$,

$$\begin{aligned} & \sum_{j=1}^{q} \sum_{k=1}^{q} c_j M(j,k) c_k \qquad (22) \\ = & \sum_{j=1}^{q} \sum_{k=1}^{q} c_j \boldsymbol{\beta}^T \boldsymbol{\Sigma} \left[\boldsymbol{\Sigma} + \mathrm{diag}(\boldsymbol{\nu})\right]^{-1} \mathbf{E}_j \left[\boldsymbol{\Sigma} + \mathrm{diag}(\boldsymbol{\nu})\right]^{-1} \boldsymbol{\Sigma} \left[\boldsymbol{\Sigma} + \mathrm{diag}(\boldsymbol{\nu})\right]^{-1} \mathbf{E}_k \boldsymbol{\beta} c_k \\ = & \boldsymbol{\beta}^T \boldsymbol{\Sigma} \left[\boldsymbol{\Sigma} + \mathrm{diag}(\boldsymbol{\nu})\right]^{-1} \begin{pmatrix} D_{\mathbf{c}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \left[\boldsymbol{\Sigma} + \mathrm{diag}(\boldsymbol{\nu})\right]^{-1} \boldsymbol{\Sigma} \left[\boldsymbol{\Sigma} + \mathrm{diag}(\boldsymbol{\nu})\right]^{-1} \begin{pmatrix} D_{\mathbf{c}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \boldsymbol{\beta}. \end{aligned}$$

Next we work on the detailed matrix products: Using the partition $p = q + (p - q)$, we partition $\boldsymbol{\Sigma}$ as follows

$$\begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

Letting $\tilde{\boldsymbol{\nu}} = (\nu_{q+1}, \nu_{q+2}, \ldots, \nu_p)^T$ and $D_{\tilde{\boldsymbol{\nu}}} = \mathrm{diag}(\tilde{\boldsymbol{\nu}})$, denote the inverse of $[\boldsymbol{\Sigma} + \mathrm{diag}(\boldsymbol{\nu})]$ as

$$\left[\boldsymbol{\Sigma} + \mathrm{diag}(\boldsymbol{\nu})\right]^{-1} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} + D_{\tilde{\boldsymbol{\nu}}} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix}. \qquad (23)$$

Consequently we have

$$\begin{aligned}
& \boldsymbol{\Sigma}\left[\boldsymbol{\Sigma}+\operatorname{diag}(\boldsymbol{\nu})\right]^{-1}\begin{pmatrix} D_{\mathbf{c}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \\
= & \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}\begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix}\begin{pmatrix} D_{\mathbf{c}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \\
= & \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}\begin{pmatrix} \mathbf{B}_{11}D_{\mathbf{c}} & \mathbf{0} \\ \mathbf{B}_{21}D_{\mathbf{c}} & \mathbf{0} \end{pmatrix} \\
= & \begin{pmatrix} (\boldsymbol{\Sigma}_{11}\mathbf{B}_{11}+\boldsymbol{\Sigma}_{12}\mathbf{B}_{21})D_{\mathbf{c}} & \mathbf{0} \\ (\boldsymbol{\Sigma}_{21}\mathbf{B}_{11}+\boldsymbol{\Sigma}_{22}\mathbf{B}_{21})D_{\mathbf{c}} & \mathbf{0} \end{pmatrix} \\
= & \begin{pmatrix} D_{\mathbf{c}} & \mathbf{0} \\ -D_{\tilde{\boldsymbol{\nu}}}\mathbf{B}_{21}D_{\mathbf{c}} & \mathbf{0} \end{pmatrix}. \qquad (24)
\end{aligned}$$

The very last step follows from the fact that

$$\begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}+D_{\tilde{\boldsymbol{\nu}}} \end{pmatrix}\begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix} = \mathbf{I}$$

implies

$$\boldsymbol{\Sigma}_{11}\mathbf{B}_{11}+\boldsymbol{\Sigma}_{12}\mathbf{B}_{21} = \mathbf{I}$$

and

$$\boldsymbol{\Sigma}_{21}\mathbf{B}_{11}+(\boldsymbol{\Sigma}_{22}+D_{\tilde{\boldsymbol{\nu}}})\mathbf{B}_{21} = \mathbf{0},$$

which further implies

$$\boldsymbol{\Sigma}_{21}\mathbf{B}_{11}+\boldsymbol{\Sigma}_{22}\mathbf{B}_{21} = -D_{\tilde{\boldsymbol{\nu}}}\mathbf{B}_{21}.$$

Here $\mathbf{I}$ denotes idenity matrix of an approppriate size.

Let $\tilde{\boldsymbol{\beta}} = (\beta_1, \beta_2, \ldots, \beta_q)^T$ and recall that

$\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_q, 0, 0, \ldots, 0)^T = (\tilde{\boldsymbol{\beta}}^T, 0, 0, \ldots, 0)^T$. Plugging (24) into (22) leads to

$$
\begin{aligned}
& \boldsymbol{\beta}^T \boldsymbol{\Sigma} \left[\boldsymbol{\Sigma} + \operatorname{diag}(\boldsymbol{\nu})\right]^{-1} \begin{pmatrix} D_{\mathbf{c}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \left[\boldsymbol{\Sigma} + \operatorname{diag}(\boldsymbol{\nu})\right]^{-1} \boldsymbol{\Sigma} \left[\boldsymbol{\Sigma} + \operatorname{diag}(\boldsymbol{\nu})\right]^{-1} \begin{pmatrix} D_{\mathbf{c}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \boldsymbol{\beta} \\
= \; & \boldsymbol{\beta}^T \begin{pmatrix} D_{\mathbf{c}} & \mathbf{0} \\ -D_{\tilde{\boldsymbol{\nu}}} \mathbf{B}_{21} D_{\mathbf{c}} & \mathbf{0} \end{pmatrix} \left[\boldsymbol{\Sigma} + \operatorname{diag}(\boldsymbol{\nu})\right]^{-1} \begin{pmatrix} D_{\mathbf{c}} & \mathbf{0} \\ -D_{\tilde{\boldsymbol{\nu}}} \mathbf{B}_{21} D_{\mathbf{c}} & \mathbf{0} \end{pmatrix} \boldsymbol{\beta} \\
= \; & (\tilde{\boldsymbol{\beta}}^T D_{\mathbf{c}}, 0, 0, \ldots, 0) \left[\boldsymbol{\Sigma} + \operatorname{diag}(\boldsymbol{\nu})\right]^{-1} \begin{pmatrix} D_{\mathbf{c}} \tilde{\boldsymbol{\beta}} \\ -D_{\tilde{\boldsymbol{\nu}}} \mathbf{B}_{21} D_{\mathbf{c}} \tilde{\boldsymbol{\beta}} \end{pmatrix} \\
= \; & (\tilde{\boldsymbol{\beta}}^T D_{\mathbf{c}}, 0, 0, \ldots, 0) \left[\boldsymbol{\Sigma} + \operatorname{diag}(\boldsymbol{\nu})\right]^{-1} \begin{pmatrix} \mathbf{I} \\ -D_{\tilde{\boldsymbol{\nu}}} \mathbf{B}_{21} \end{pmatrix} D_{\mathbf{c}} \tilde{\boldsymbol{\beta}} \\
= \; & (\tilde{\boldsymbol{\beta}}^T D_{\mathbf{c}}, 0, 0, \ldots, 0) \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{I} \\ -D_{\tilde{\boldsymbol{\nu}}} \mathbf{B}_{21} \end{pmatrix} D_{\mathbf{c}} \tilde{\boldsymbol{\beta}} \\
= \; & (\tilde{\boldsymbol{\beta}}^T D_{\mathbf{c}}, 0, 0, \ldots, 0) \begin{pmatrix} \mathbf{B}_{11} - \mathbf{B}_{12} D_{\tilde{\boldsymbol{\nu}}} \mathbf{B}_{21} \\ \mathbf{B}_{21} - \mathbf{B}_{22} D_{\tilde{\boldsymbol{\nu}}} \mathbf{B}_{21} \end{pmatrix} D_{\mathbf{c}} \tilde{\boldsymbol{\beta}} \\
= \; & \tilde{\boldsymbol{\beta}}^T D_{\mathbf{c}} \left(\mathbf{B}_{11} - \mathbf{B}_{12} D_{\tilde{\boldsymbol{\nu}}} \mathbf{B}_{21} \right) D_{\mathbf{c}} \tilde{\boldsymbol{\beta}} \\
= \; & \mathbf{c}^T D_{\tilde{\boldsymbol{\beta}}} \left(\mathbf{B}_{11} - \mathbf{B}_{12} D_{\tilde{\boldsymbol{\nu}}} \mathbf{B}_{21} \right) D_{\tilde{\boldsymbol{\beta}}} \mathbf{c},
\end{aligned}
$$

which is positive as long as $\| \mathbf{c} \| \neq 0$ since $\mathbf{B}_{11} - \mathbf{B}_{12} D_{\tilde{\boldsymbol{\nu}}} \mathbf{B}_{21}$ is strictly positive definite, as we verify next. Since (23) implies $\mathbf{B}_{22}^{-1} = \boldsymbol{\Sigma}_{22} + D_{\tilde{\boldsymbol{\nu}}} - \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12}$ and $\boldsymbol{\Sigma}_{11}^{-1} = \mathbf{B}_{11} - \mathbf{B}_{12} \mathbf{B}_{22}^{-1} \mathbf{B}_{21}$ implies

$$
\boldsymbol{\Sigma}_{11}^{-1} = \mathbf{B}_{11} - \mathbf{B}_{12} \mathbf{B}_{22}^{-1} \mathbf{B}_{21} = \mathbf{B}_{11} - \mathbf{B}_{12} \left(\boldsymbol{\Sigma}_{22} + D_{\tilde{\boldsymbol{\nu}}} - \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12} \right) \mathbf{B}_{21},
$$

it follows that

$$
\mathbf{B}_{11} - \mathbf{B}_{12} D_{\tilde{\boldsymbol{\nu}}} \mathbf{B}_{21} = \boldsymbol{\Sigma}_{11}^{-1} + \mathbf{B}_{12} \left(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12} \right) \mathbf{B}_{21},
$$

which is strictly positive definite since the strictly positive definiteness of $\boldsymbol{\Sigma}$ implies that both $\boldsymbol{\Sigma}_{11}$ and $\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12}$ to be strictly positive definite. This completes the proof. □

Proof of Theorem 1. Note that a metric distance is non-negative and subadditive, i.e., $d(Y, Y'') \leq d(Y, Y') + d(Y', Y'')$ for any $Y, Y', Y'' \in \Omega$. Recalling that $m_{\oplus}(\mathbf{x}) = L_{\oplus}(\mathbf{x})$ for any $\mathbf{x} \in \mathcal{X}$ due to the global Fréchet regression model assumption, one obtains for the

objective function (15)

$$\begin{aligned} f(\boldsymbol{\lambda}) &\triangleq \frac{1}{n}\sum_{i=1}^{n} d^2(Y_i, \widehat{R}_{\oplus}(\mathbf{x}_i; \boldsymbol{\lambda}^{-1})) \\ &\leq \frac{1}{n}\sum_{i=1}^{n} \Big[d(Y_i, m_{\oplus}(\mathbf{x}_i)) + d(L_{\oplus}(\mathbf{x}_i), R_{\oplus}(\mathbf{x}_i; \boldsymbol{\lambda}^{-1})) + d(R_{\oplus}(\mathbf{x}_i; \boldsymbol{\lambda}^{-1}), \widehat{R}_{\oplus}(\mathbf{x}_i; \boldsymbol{\lambda}^{-1})) \Big]^2 \\ &= T_1 + T_2(\boldsymbol{\lambda}) + T_3(\boldsymbol{\lambda}) + T_4(\boldsymbol{\lambda}) + T_5(\boldsymbol{\lambda}) + T_6(\boldsymbol{\lambda}), \end{aligned}$$

where

$$\begin{aligned} T_1 &= \frac{1}{n}\sum_{i=1}^{n} d^2(Y_i, m_{\oplus}(\mathbf{x}_i)) \\ T_2(\boldsymbol{\lambda}) &= \frac{1}{n}\sum_{i=1}^{n} d^2(L_{\oplus}(\mathbf{x}_i), R_{\oplus}(\mathbf{x}_i; \boldsymbol{\lambda}^{-1})) \\ T_3(\boldsymbol{\lambda}) &= \frac{1}{n}\sum_{i=1}^{n} d^2(R_{\oplus}(\mathbf{x}_i; \boldsymbol{\lambda}^{-1}), \widehat{R}_{\oplus}(\mathbf{x}_i; \boldsymbol{\lambda}^{-1})) \\ T_4(\boldsymbol{\lambda}) &= \frac{2}{n}\sum_{i=1}^{n} d(Y_i, m_{\oplus}(\mathbf{x}_i)) d(L_{\oplus}(\mathbf{x}_i), R_{\oplus}(\mathbf{x}_i; \boldsymbol{\lambda}^{-1})) \\ T_5(\boldsymbol{\lambda}) &= \frac{2}{n}\sum_{i=1}^{n} d(Y_i, m_{\oplus}(\mathbf{x}_i)) d(R_{\oplus}(\mathbf{x}_i; \boldsymbol{\lambda}^{-1}), \widehat{R}_{\oplus}(\mathbf{x}_i; \boldsymbol{\lambda}^{-1})) \\ T_6(\boldsymbol{\lambda}) &= \frac{2}{n}\sum_{i=1}^{n} d(L_{\oplus}(\mathbf{x}_i), R_{\oplus}(\mathbf{x}_i; \boldsymbol{\lambda}^{-1})) d(R_{\oplus}(\mathbf{x}_i; \boldsymbol{\lambda}^{-1}), \widehat{R}_{\oplus}(\mathbf{x}_i; \boldsymbol{\lambda}^{-1})). \end{aligned}$$

The first term $T_1 \xrightarrow{p} E_{(\mathbf{X},Y)} d^2(Y, m_{\oplus}(\mathbf{X}))$ does not depend on $\boldsymbol{\lambda}$, is due to random error and cannot be controlled; the second term $T_2(\boldsymbol{\lambda}) \xrightarrow{p} E_{\mathbf{X}} d^2(L_{\oplus}(\mathbf{X}), R_{\oplus}(\mathbf{X}; \boldsymbol{\lambda}^{-1}))$ is due to the model bias of the individually penalized ridge Fréchet regression; the third term $T_3(\boldsymbol{\lambda})$ is due to the estimation error of the individually penalized ridge Fréchet regression; and the remaining three terms are due to the cross-products of these main terms. Without loss of generality, we assume in the following that $\widehat{\boldsymbol{\lambda}}(\tau_n)$ converges; otherwise we consider a convergent subsequence.

We first prove by contradiction that $\hat{\lambda}_j(\tau_n) \xrightarrow{p} \infty$ for any $j \in \mathcal{I}$. Suppose that $\hat{\lambda}_j(\tau) \not\to \infty$ in probability for some $j \in \mathcal{I}$. In this case, Conditions [B] and [C] imply that $T_2(\widehat{\boldsymbol{\lambda}}(\tau_n))$ will converge in probability to some positive number due to the bias of the individually penalized ridge Fréchet regression since $\hat{\lambda}_j(\tau_n)$ does not diverge to infinity in probability (the corresponding ridge parameter $\nu_j = 1/\hat{\lambda}_j(\tau_n) \not\to 0$ in probability so that the ridge term does not disappear asymptotically). We consider another possible sequence

$\tilde{\boldsymbol{\lambda}}(\tau_n) = (\tilde{\lambda}_1(\tau_n), \tilde{\lambda}_2(\tau_n), \ldots, \tilde{\lambda}_p(\tau_n))^T$ with $\tilde{\lambda}_j(\tau_n) = \tau_n/|\mathcal{I}|$ for $j \in \mathcal{I}$ and $\tilde{\lambda}_{j'}(\tau_n) = 0$ for $j \notin \mathcal{I}$. It can be checked easily that as $\tau_n \to \infty$ along with $n \to \infty$, we have $T_2(\tilde{\boldsymbol{\lambda}}(\tau_n)) \xrightarrow{p} 0$ since the ridge bias goes away and unimportant predictors are left out, $T_3(\tilde{\boldsymbol{\lambda}}(\tau_n)) \xrightarrow{p} 0$ due to Lemma 1, and similarly $T_4(\tilde{\boldsymbol{\lambda}}(\tau_n)) \xrightarrow{p} 0$, $T_5(\tilde{\boldsymbol{\lambda}}(\tau_n)) \xrightarrow{p} 0$, $T_6(\tilde{\boldsymbol{\lambda}}(\tau_n)) \xrightarrow{p} 0$. Consequently $f(\tilde{\boldsymbol{\lambda}}) \xrightarrow{p} E_{(\mathbf{X},Y)} d^2(Y, m_{\oplus}(\mathbf{X}))$, which is strictly less than the limit $f(\hat{\boldsymbol{\lambda}}(\tau_n))$. As a result $\hat{\boldsymbol{\lambda}}(\tau_n)$ is sub-optimal, which proves that $\hat{\lambda}_j(\tau_n) \xrightarrow{p} \infty$ for any $j \in \mathcal{I}$ since $\hat{\boldsymbol{\lambda}}(\tau_n)$ is the optimizer of (15).

The next part is to prove that $\hat{\lambda}_{j'}(\tau_n) \xrightarrow{p} 0$ for $j' \notin \mathcal{I}$ as $n \to \infty$. If this is not the case, we consider another sequence of $\check{\boldsymbol{\lambda}}(\tau_n) = (\check{\lambda}_1(\tau_n), \check{\lambda}_2(\tau_n), \ldots, \check{\lambda}_p(\tau_n))^T$ with $\check{\lambda}_{j'}(\tau_n) = 0$ for $j' \notin \mathcal{I}$ and $\check{\lambda}_j(\tau_n) = \frac{\tau_n}{\tau_n - \sum_{j' \notin \mathcal{I}} \hat{\lambda}_{j'}(\tau_n)} \hat{\lambda}_j(\tau_n)$ for $j \in \mathcal{I}$, which also satisfies the constraints (16) and (17). Note that $\check{\lambda}_j(\tau_n) > \hat{\lambda}_j(\tau_n)$ for $j \in \mathcal{I}$ when n is large enough. Consequently the bias term $T_2(\check{\boldsymbol{\lambda}}(\tau_n))$ converges in probability to zero faster than $T_2(\hat{\boldsymbol{\lambda}}(\tau_n))$ does, due to Condition [D], which implies that $\hat{\boldsymbol{\lambda}}(\tau_n)$ is sub-optimal. As a result, $\hat{\lambda}_{j'}(\tau_n) \xrightarrow{p} 0$ for $j' \notin \mathcal{I}$, as $n \to \infty$.

□

REFERENCES

Berk, R., L. Brown, A. Buja, K. Zhang, and L. Zhao (2013). Valid post-selection inference. *Annals of Statistics 41*(2), 802–837.

Chang, T. (1986). Spherical regression. *The Annals of Statistics 14*(3), 907–924.

Desboulets, L. D. D. (2018). A Review on Variable Selection in Regression Analysis. *Econometrics 6*(4), 1–27.

Dubey, P. and H.-G. Müller (2020). Functional models for time-varying random objects. *J.R. Statist. Soc. B 82*(Part2), 1–35.

Fan, J. and J. Lv (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica 20*(1), 101–148.

Fanaee-T, H. and J. Gama (2013). Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 1–15.

Fang, K., S. Kotz, and K. Ng (1990). *Symmetric multivariate and related distributions*. Monographs on statistics and applied probability. London: Chapman & Hall.

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

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*, 55–67.

Paine, P. J., S. P. Preston, M. Tsagris, and A. T. A. Wood (2020). Spherical regression models with general covariates and anisotropic errors. *Statistics and Computing 30*, 153–165.

Petersen, A. and H.-G. Müller (2016). Fréchet integration and adaptive metric selection for interpretable covariances of multivariate functional data. *Biometrika 103*, 103–120.

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

Pigoli, D., J. A. Aston, I. L. Dryden, and P. Secchi (2014). Distances and inference for covariance operators. *Biometrika 101*, 409–422.

Stefanski, L. A., Y. Wu, and K. R. White (2014). Variable selection in nonparametric classification via measurement error model selection likelihoods. *Journal of the American Statistical Association 109*, 574–589.

Tavakoli, S., D. Pigoli, J. A. Aston, and J. S. Coleman (2019). A spatial modeling approach for linguistic object data: Analyzing dialect sound variations across great britain. *Journal of the American Statistical Association 114*(527), 1081–1096.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 58*, 267–288.

Wu, Y. (2020). Can't ridge regression perform variable selection? *Technometrics*, in press.

Yuan, M. and Y. Lin (2007). On the non-negative garrotte estimator. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69*(2), 143–161.