Statistica Sinica Preprint No: SS-2025-0151	
Title	Variable Selection and Minimax Prediction in High-dimensional Functional Linear Models
Manuscript ID	SS-2025-0151
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0151
Complete List of Authors	Xingche Guo, Yehua Li and Tailen Hsing
Corresponding Authors	Yehua Li
E-mails	yehuali@ucr.edu
Notice: Accepted author version.	

Variable Selection and Minimax Prediction in High-dimensional Functional Linear Models

Xingche Guo and Yehua Li and Tailen Hsing

University of Connecticut, University of California Riverside and University of Michigan

Abstract: High-dimensional functional data have become increasingly prevalent in modern applications such as high-frequency financial data and neuroimaging data analysis. We investigate a class of high-dimensional linear regression models, where each predictor is a random element in an infinite-dimensional function space, and the number of functional predictors p can potentially be ultra-high. Assuming that each of the unknown coefficient functions belongs to some reproducing kernel Hilbert space (RKHS), we regularize the fitting of the model by imposing a group elastic-net type of penalty on the RKHS norms of the coefficient functions. We show that our loss function is Gateaux sub-differentiable, and our functional elastic-net estimator exists uniquely in the product RKHS. Under suitable sparsity assumptions and a functional version of the irrepresentable condition, we derive a non-asymptotic tail bound for variable selection consistency of our method. Allowing the number of true functional predictors q to diverge with the sample size, we also show a post-selection refined estimator can achieve the oracle minimax optimal prediction rate. The proposed methods are illustrated through simulation studies and a real-data

application from the Human Connectome Project.

Key words: Elastic-net penalty; Functional linear regression; Minimax optimality; Model selection consistency; Reproducing kernel Hilbert space; Sparsity.

1. Introduction

Modern science and technology give rise to large data sets with high-frequency repeated measurements, resulting in random trajectories that can be modeled as functional data (Ramsay and Silverman, 2005). There has been a large volume of literature on scalar-on-function regression models, where the most studied model is the functional linear model (FLM); see James (2002); Müller and Stadtmüller (2005); Cai and Hall (2006); Reiss and Ogden (2007); Crambes et al. (2009); Cai and Yuan (2012); Lei (2014); Shang and Cheng (2015); Liu et al. (2022), among others. With functional data belonging to an infinite-dimensional function space (Hsing and Eubank, 2015), the sequence of eigenvalues of the covariance operator decays to zero, rendering the covariance operator non-invertible and hence the inference of the FLM a challenging inverse problem.

There has been a recent surge in applications of high-dimensional functional data analysis due to new developments in neuroimaging (e.g. fMRI and TDI), electroencephalogram (EEG), and high-frequency stock exchange data. For example, Qiao et al. (2019) modeled EEG activity data from different

nodes as high-dimensional functional data and proposed a functional Gaussian graphical model to study the connectivity between the nodes. Lee et al. (2023) considered a class of conditional functional graphical models to model the connectivity between different regions of interest (ROI) of the brain using fMRI data.

It is also natural to consider regression models with high-dimensional functional predictors. Fan et al. (2015) studied variable selection procedures for linear and non-linear regression models with high-dimensional functional predictors. Their approach was to reduce the dimension of each functional predictor by representing it as a linear combination of some known basis functions and to apply a group-lasso type of penalty in model fitting. As pointed out in Xue and Yao (2021), the results in Fan et al. (2015) relied heavily on the assumption that the minimum eigenvalues of the design matrices are bounded away from zero, which ignored the infinite-dimensional nature of functional data and essentially limited their methods to functional data reside in a finite-dimensional function subspace. Xue and Yao (2021), on the other hand, focused on hypothesis testing issues in high-dimensional FLMs rather than variable selection consistency. As Fan et al. (2015), Xue and Yao (2021) also based their approach on representing functional predictors on pre-selected basis functions and minimizing a penalized least square loss function, where

the group penalty can be flexibly chosen from lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001) or MCP (Zhang, 2010). To the best of our knowledge, the variable selection consistency property for the high-dimensional FLM in a general functional-data setting remains an open problem to date.

We propose to conduct variable selection in high-dimensional FLMs under the RKHS framework using a double-penalty approach, where the first penalty resembles the group-lasso type penalty in Xue and Yao (2021), which encourages sparsity, and the second penalty is on the squared RKHS norms of the functional coefficients to regularize the smoothness of the fit. As shown in Cai and Yuan (2012), the RKHS approach can outperform the principal component regression approach when the coefficient functions are not directly spanned by the eigenfunctions of the functional predictors. Many of the existing high-dimensional functional regression approaches including Fan et al. (2015) and Xue and Yao (2021) are similar in spirit to the principal component regression in which both the functional predictors and the coefficient functions are expressed using the same set of basis functions. Our approach offers the extra flexibility of picking the reproducing kernel based on the application and thus can outperform the existing methods when the coefficient functions are "misaligned" with the functional predictors as described by Cai and Yuan (2012). Our double penalization method resembles a group-penalized version

of the elastic-net (Zou and Hastie, 2005), where the two penalties enforces sparsity and stabilizes the solution paths, respectively. It is well known that the lasso alone tends not to work well when the predictors are highly correlated, while the elastic-net may offer a more stable solution path and better prediction performance under high collinearity.

One of the main contributions of the present paper is providing a theory that addresses variable selection consistency for high-dimensional FLMs. In the scalar case that they considered, Zou and Zhang (2009) established a variable selection consistency result for the elastic-net. However, the noninvertibility of the design matrices of the functional predictors in our problem makes it necessary to create a completely new proof. Another important contribution of our paper is that we develop the minimax optimal prediction rate for the high-dimensional FLMs, where the number of true functional predictors q is allowed to grow to infinity with the sample size n. We show that a post-selection, refined estimation of the high-dimensional FLM using our RKHS approach can achieve such a minimax optimal rate.

2. Functional Elastic-Net Regression

2.1 Model Assumptions

Let $\mathbb{L}_2[0,1]$ be the L_2-space of square-integrable, measurable functions on $[0,1]$, equipped with the inner product $\langle f, g\rangle_2 = \int_0^1 f(t)g(t)dt$ and functional norm

$\|f\|_2 = \langle f, f\rangle_2^{1/2}$, for any $f, g \in \mathbb{L}_2[0,1]$. We will also be concerned with the p-fold product space of $\mathbb{L}_2^p[0,1]$ containing elements $\boldsymbol{f} = (f_1, \ldots, f_p)^\top$ with each $f_j \in \mathbb{L}_2[0,1]$, $\|\boldsymbol{f}\|_2 \equiv (\sum_{j=1}^p \|f_j\|_2^2)^{1/2} < \infty$ and inner product $\langle \boldsymbol{f}, \boldsymbol{g}\rangle_2 \equiv \sum_{j=1}^p \langle f_j, g_j\rangle_2$ for $\boldsymbol{f} = (f_1, \ldots, f_p)^\top, \boldsymbol{g} = (g_1, \ldots, g_p)^\top$. Let $\otimes$ be the outer product associated with either inner product such that $f \otimes g$ defines an operator $(f \otimes g)h = f\langle g, h\rangle_2$. In this paper, we consider a high-dimensional FLM:

$$Y_i = \sum_{j=1}^{p} \langle X_{ij}, \beta_j\rangle_2 + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where the functional predictors $X_{ij}(\cdot)$ are random elements in $\mathbb{L}_2[0,1]$, $\beta_j(\cdot)$ are unknown coefficient functions in $\mathbb{L}_2[0,1]$, and ε_i are iid zero-mean random errors with variance σ^2. Without loss of generality, assume that both Y_i and $X_{ij}(t)$ are centered at 0, i.e., $\mathbb{E}Y_i = 0$ and $\mathbb{E}X_{ij}(t) = 0$ for $t \in [0,1]$, $j = 1, \ldots, p$, so that no intercept is needed in (2.1).

Consider $\boldsymbol{X}_{i\bullet} = (X_{i1}, \ldots, X_{ip})^\top$, $i = 1, \ldots, n$, as iid zero-mean random vectors, with the covariance operator $\mathscr{C}$ defined as $\mathscr{C} = \mathbb{E}(X_{i1}, \ldots, X_{ip})^\top \otimes (X_{i1}, \ldots, X_{ip})$. Note that we do not assume that the functional predictors are independent. It is convenient to view $\mathscr{C}$ as a $p \times p$ operator-valued matrix $\{\mathscr{C}^{(j,j')}\}$ where $\mathscr{C}^{(j,j')} = \mathbb{E}(X_{ij} \otimes X_{ij'})$ is the cross covariance operators of X_{ij} and $X_{ij'}$. Denote $\boldsymbol{Y}_n = (Y_1, \ldots, Y_n)^\top$, $\boldsymbol{\varepsilon}_n = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ and $\boldsymbol{X}_n = (\boldsymbol{X}_{1\bullet}, \ldots, \boldsymbol{X}_{n\bullet})^\top$ as the $n \times p$ matrix of functional predictors. Then, the

sample covariance operator $\mathscr{C}_n$ is defined as

$$\mathscr{C}_n = \frac{1}{n}\sum_{i=1}^{n}(X_{i1},\ldots,X_{ip})^\top \otimes (X_{i1},\ldots,X_{ip}) = \frac{1}{n}\boldsymbol{X}_n^\top \otimes \boldsymbol{X}_n. \tag{2.2}$$

We further assume that $\beta_j(\cdot) \in \mathbb{H}_j := \mathbb{H}(K_j)$, which is the reproducing kernel Hilbert space (RKHS) with kernel K_j (Wahba, 1990). Recall that a real, symmetric, square-integrable, and nonnegative definite function $K(\cdot,\cdot)$ on $[0,1]^2$ is called a reproducing kernel (RK) for a Hilbert space of functions $\mathbb{H}(K)$ on $[0,1]$ if $K(\cdot,t) \in \mathbb{H}(K)$ for any $t \in [0,1]$ and $\mathbb{H}(K)$ is equipped with the inner product such that $\langle \beta, K(\cdot,t)\rangle_{\mathbb{H}(K)} = \beta(t)$ for any $\beta \in \mathbb{H}(K)$ and any $t \in [0,1]$; the Hilbert space $\mathbb{H}(K)$ is then called an RKHS. With a proper choice of RK, an RKHS provides a flexible class of functions which can also be naturally regularized using the RKHS norm. As such, the RKHS is a useful framework in nonparametric estimation (Wahba, 1990) and functional data analysis (Cai and Yuan, 2012; Hsing and Eubank, 2015; Sun et al., 2018; Lee et al., 2023).

Remark 1. *The choice of kernel K determines the smoothness class. Sobolev kernels of order m (Hsing and Eubank, 2015) regulate the m-th derivative, whereas Gaussian kernels yield infinitely differentiable functions. In contrast, total-variation penalties, although successfully applied in scalar-on-image functional regression (Wang et al., 2017) with the benefits of promoting piecewise*

structure and allowing jumps, are not induced by an RKHS norm and therefore lie outside our RKHS-based framework.

We adopt the commonly assumed setting where the total number of functional predictors, p, can be much larger than the sample size n but only a small portion of those have non-zero effects on the response. Denote the signal set as $\mathscr{S} = \{j \in \{1, \ldots, p\} : \mathrm{Var}(\langle X_{1j}, \beta_j \rangle_2) = \langle \beta_j, \mathscr{C}^{(j,j)} \beta_j \rangle_2 \neq 0\}$ and the non-signal set as $\mathscr{S}^c = \{1, \ldots, p\} \backslash \mathscr{S}$, and write $q := |\mathscr{S}|$.

2.2 Functional Elastic-Net Based on RKHS

In order to regularize the solution as well as to enforce sparsity in $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$, we assume $\boldsymbol{\beta} \in \mathbb{H} := \otimes_{j=1}^p \mathbb{H}_j$, which is the direct product of the RKHS (Hsing and Eubank, 2015), and estimate it by

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{H}}{\operatorname{argmin}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} \langle X_{ij}, \beta_j \rangle_2 \right)^2 + \sum_{j=1}^{p} \mathrm{Pen}(\beta_j; \boldsymbol{\lambda}) \right\} \tag{2.3}$$

where $\mathrm{Pen}(\beta_j; \boldsymbol{\lambda})$ is the functional elastic-net penalty to be specified below with $\boldsymbol{\lambda}$ denoting a vector of tuning parameters.

Following Cai and Yuan (2012), for any symmetric positive semi-definite kernel $R(\cdot, \cdot)$, denote $\mathscr{L}_R$ as the integral operator $(\mathscr{L}_R f)(\cdot) = \int_0^1 R(s, \cdot) f(s) ds$, $f \in \mathbb{L}_2[0, 1]$. Suppose R has a spectral decomposition $R(s, t) = \sum_{j=1}^{\infty} \theta_j^R \varphi_j^R(s) \varphi_j^R(t)$. Then its square root is defined as $R^{1/2}(s, t) = \sum_{j=1}^{\infty} (\theta_j^R)^{1/2} \varphi_j^R(s) \varphi_j^R(t)$, and $\mathscr{L}_{R^{1/2}}$ is the associated square-root integral operator. For a matrix of kernel

functions $\boldsymbol{R} = (R_{ij})_{i,j=1}^{k,m}$, let $\mathscr{L}_{\boldsymbol{R}} : \mathbb{L}_2^m \to \mathbb{L}_2^k$ be the corresponding matrix of operators such that $\mathscr{L}_{\boldsymbol{R}}\boldsymbol{f} = \left(\sum_{j=1}^m \mathscr{L}_{R_{ij}} f_j\right)_{i=1}^k$ for any $\boldsymbol{f} = (f_1, \ldots, f_m)^\top \in \mathbb{L}_2^m$. By Wahba (1990) and Cai and Yuan (2012), for any positive semi-definite kernel K and any $\beta \in \mathbb{H}(K)$, there exists an $f \in \mathbb{L}_2[0,1]$ such that $\beta = \mathscr{L}_{K^{1/2}} f$. If K is not strictly positive definte, then multiple f's satisfy this relationship. However, there is always a unique f satisfying $\|\beta\|_{\mathbb{H}(K)} = \|f\|_2$. The ridge regularization term in our objective (introduced later) guarantees the identifiability of this representative. Without causing any confusion, we use $\|\cdot\|_2$ to denote the norm of $\mathbb{L}_2$ functions or vectors of $\mathbb{L}_2$ functions as well as the Euclidean norm in $\mathbb{R}^p$.

Let $\beta_j = \mathscr{L}_{K_j^{1/2}} f_j$ for all j and denote $\boldsymbol{f} = (f_1, \ldots, f_p)^\top$. Then $\boldsymbol{\beta} = \mathscr{L}_{\boldsymbol{K}^{1/2}}\boldsymbol{f}$ where $\boldsymbol{K}(s,t) = \mathrm{diag}(K_1, \ldots, K_p)(s,t)$. Define $\widetilde{X}_{ij} = \mathscr{L}_{K_j^{1/2}} X_{ij}$, $\widetilde{\boldsymbol{X}}_{i\bullet} = (\widetilde{X}_{i1}, \ldots, \widetilde{X}_{ip})^\top$, and $\widetilde{\boldsymbol{X}}_n = (\widetilde{\boldsymbol{X}}_{1\bullet}, \ldots, \widetilde{\boldsymbol{X}}_{n\bullet})^\top$. Thus, the theoretical and empirical covariance of $\widetilde{\boldsymbol{X}}_{i\bullet}$ are $\mathscr{T} = \mathbb{C}\mathrm{ov}(\widetilde{\boldsymbol{X}}_{i\bullet}) = \mathscr{L}_{\boldsymbol{K}^{1/2}}\mathscr{C}\mathscr{L}_{\boldsymbol{K}^{1/2}}$ and $\mathscr{T}_n = \mathscr{L}_{\boldsymbol{K}^{1/2}}\mathscr{C}_n\mathscr{L}_{\boldsymbol{K}^{1/2}} = n^{-1}\widetilde{\boldsymbol{X}}_n^\top \otimes \widetilde{\boldsymbol{X}}_n$. Define $\mathbb{M}_{nj} = \mathrm{Span}\{\widetilde{X}_{ij}(\cdot), i = 1, \ldots, n\}$ and $\mathbb{M}_{nj}^\perp$ the orthogonal complement of $\mathbb{M}_{nj}$. With the above $\mathbb{L}_2$ representation $\boldsymbol{f}$ of $\boldsymbol{\beta}$, the loss function in (2.3) can be rewritten as

$$\ell(\boldsymbol{f}) = \frac{1}{2}\langle \mathscr{T}_n\boldsymbol{f}, \boldsymbol{f}\rangle_2 - \left\langle \frac{1}{n}\widetilde{\boldsymbol{X}}_n^\top \boldsymbol{Y}_n, \boldsymbol{f}\right\rangle_2 + \frac{1}{2n}\|\boldsymbol{Y}_n\|_2^2 + \sum_{j=1}^p \mathrm{Pen}(f_j; \boldsymbol{\lambda}). \quad (2.4)$$

We propose to use the following functional elastic-net penalty

$$\text{Pen}(f_j;\lambda_1,\lambda_2)=\lambda_1\|\Psi_j f_j\|_2+\frac{\lambda_2}{2}\|f_j\|_2^2,\quad \lambda_1,\lambda_2>0,$$

where Ψ_j is an operator on $\mathbb{L}_2[0,1]$ satisfying the following condition.

C.1. *For $j=1,\ldots,p$, Ψ_j is a self-adjoint operator such that $\Psi_j f\in\mathbb{M}_{nj}$ for all $f\in\mathbb{M}_{nj}$. Assume that there exist positive constants $0<C_{\min}<C_{\max}<\infty$ such that, uniformly for all j, the eigenvalues of Ψ_j are in the interval $[C_{\min},C_{\max}]$.*

Remark 2. *(i) The $\mathbb{L}_2$-norm $\|f_j\|_2$ in $\text{Pen}(f_j;\lambda_1,\lambda_2)$ corresponds to the RKHS norm $\|\beta_j\|_{\mathbb{H}_j}$, a commonly used norm in functional regression problems (cf. Cai and Yuan, 2012).*

(ii) A simple choice for Ψ_j is $\Psi_j=\mathscr{I}$, the identity operator, based on which the penalty $\text{Pen}(f_j;\lambda_1,\lambda_2)$ includes both $\|f_j\|_2$ and $\|f_j\|_2^2$ and resembles an elastic-net (cf. Zou and Hastie, 2005) version of the group lasso (Yuan and Lin, 2006). In the high-dimensional functional regression setting, Xue and Yao (2021) considered a penalty that focused on the amount of variation X_j explains rather than the norm of f_j. Their penalty translates in our setting to $\lambda_1 n^{-1/2}(\sum_{i=1}^n\langle X_{ij},\beta_j\rangle_2^2)^{1/2}=\lambda_1\|\{\mathscr{T}_n^{(j,j)}\}^{1/2}f_j\|_2$ where $\mathscr{T}_n^{(j,j)}$ is the empirical covariance of $\widetilde{\boldsymbol{X}}_{\bullet j}=(\widetilde{X}_{1j},\cdots,\widetilde{X}_{nj})^\top$ or the (j,j)th entry of $\boldsymbol{\mathscr{T}}_n$. The approach in Xue and Yao (2021) does not penalize the squared norm, but both X_j and β_j are represented by a growing but finite number of basis functions, which

effectively sets a lower bound on the smallest eigenvalue of $\mathscr{T}_n^{(j,j)}$. *In our setting, we can achieve similar effects by setting* $\Psi_j = (\mathscr{T}_n^{(j,j)} + \theta\mathscr{I})^{1/2}$, *where* $\theta > 0$ *provides a floor to the smallest eigenvalue of* Ψ_j *and is treated as a tuning parameter.*

Note that the functional estimator, $\widehat{\boldsymbol{f}}$, is defined as the solution that minimizes (2.4) over an infinite-dimensional space $\mathbb{L}_2^p[0,1]$. The following proposition establishes that the minimization problem is indeed well defined and any minimizer must be in a finite-dimensional subspace.

Proposition 1. *Suppose that Condition C.1 holds. Then, for each* $j = 1, \ldots, p$, *any minimizer* $\widehat{f}_j$ *of* (2.4) *must be in the space* $\mathbb{M}_{nj}$.

The proof of Proposition 1 uses the ideas of the well-known representer theorem for smoothing splines (Wahba, 1990). The fact that the minimizer of (2.4) is in a finite-dimensional subspace allows us to establish its uniqueness in Proposition 2 below.

Next, we develop the convex programming conditions in the functional space that characterize the optimizer of (2.4). For the classical lasso problem (Tibshirani, 1996), the Karush-Kuhn-Tucker (KKT) condition is used to characterize the solution (cf. Zhao and Yu, 2006; Wainwright, 2009), where subgradients are used in place of gradients due to the nondifferentiability of the lasso objective function. Similarly, in the function space, the objective func-

tion (2.4) is not always differentiable because of the group-lasso-type penalty on $\|\Psi_j f_j\|_2$. In Section S.1.1, we review the definition of Gateaux differentiability and define the corresponding notion of sub-differential. With these in mind, we state the following result.

Proposition 2. *Let $\boldsymbol{\beta}_0$ be the true value of $\boldsymbol{\beta}$ in Model (2.1), and $\boldsymbol{f}_0 = (f_{01}, \ldots, f_{0p})^\top$ be the corresponding $\mathbb{L}_2^p$ surrogate such that $\boldsymbol{\beta}_0 = \mathscr{L}_{\boldsymbol{K}^{1/2}} \boldsymbol{f}_0$. Suppose Condition C.1 holds. Then, for all $\lambda_1, \lambda_2 > 0$, the solution $\widehat{\boldsymbol{f}}$ for (2.4) exists uniquely and satisfies*

$$\mathscr{T}_n(\widehat{\boldsymbol{f}} - \boldsymbol{f}_0) - \boldsymbol{g}_n + \lambda_2 \widehat{\boldsymbol{f}} + \lambda_1 \boldsymbol{\omega} = 0, \tag{2.5}$$

where $\boldsymbol{g}_n = n^{-1} \widetilde{\boldsymbol{X}}_n^\top \boldsymbol{\varepsilon}_n$, and $\omega_j = \frac{\Psi_j^2 \widehat{f}_j}{\|\Psi_j \widehat{f}_j\|_2}$ if $\widehat{f}_j \neq 0$ and $\omega_j = \Psi_j \eta_j$ for some η_j with $\|\eta_j\|_2 \leq 1$ if $\widehat{f}_j = 0$.

Equation (2.5) is referred to as the functional KKT condition for the optimization problem (2.4) and plays a central role in establishing Theorem 1 .

3. Theoretical Results

3.1 Consistency property of variable selection

In this section, we establish the consistency property of variable selection using our approach. Even though the normality assumption is not essential to our methodology, in order to get sharp results that are comparable with those in

the literature, we assume that the rows of $\boldsymbol{X}_{i\bullet}$, $i = 1, \ldots, n$, are iid zero-mean Gaussian random vectors with each element lies in $\mathbb{L}_2[0,1]$, and $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. Note that Gaussianity is invoked only to obtain exponential concentration; our current theory does not cover genuinely heavy-tailed designs or errors. Recall the definitions of $\mathscr{S}$ and $\widehat{\boldsymbol{f}} = (\widehat{f}_1, \ldots, \widehat{f}_p)^\top$ in Sections 2.1 and 2.2, respectively, and define $\widehat{\mathscr{S}} = \{j \in \{1, \ldots, p\} : \widehat{f}_j \neq 0\}$. Then, variable selection consistency is achieved when $\widehat{\mathscr{S}} = \mathscr{S}$.

We collect here some notation used throughout the paper. Let $\mathbb{H}_1$ and $\mathbb{H}_2$ be two Hilbert spaces and $\mathscr{A} : \mathbb{H}_1 \to \mathbb{H}_2$ be a compact linear operator mapping from $\mathbb{H}_1$ to $\mathbb{H}_2$. Then the $\mathbb{L}_2$ operator norm is defined as $\|\mathscr{A}\|_2 = \sup_{f \in \mathbb{H}_1} \|\mathscr{A} f\|_2 / \|f\|_2$ which is the maximum singular value of $\mathscr{A}$; if $\mathbb{H}_1 = \mathbb{H}_2$ and $\mathscr{A}$ is self-adjoint, the trace of $\mathscr{A}$ is $\mathrm{tr}(\mathscr{A}) = \sum_{j \geq 1} \Lambda_j(\mathscr{A})$, which is the sum of all eigenvalues. For any $\boldsymbol{f} \in \mathbb{L}_2^p[0,1]$, $\|\boldsymbol{f}\|_\infty := \max_j \|f_j\|_2$; for any $r \times s$ operator-valued matrix $\boldsymbol{\mathscr{A}} = (\mathscr{A}_{ij})_{i,j=1}^{r,s}$, where each $\mathscr{A}_{ij}$ maps from $\mathbb{L}_2[0,1]$ to $\mathbb{L}_2[0,1]$, define the norm $|||\boldsymbol{\mathscr{A}}|||_{a,b} := \sup_{\|\boldsymbol{f}\|_a \leq 1} \|\boldsymbol{\mathscr{A}} \boldsymbol{f}\|_b$ for $a, b \in \{2, \infty\}$. For any index sets $\mathscr{S}_1$ and $\mathscr{S}_2$, $\boldsymbol{\mathscr{A}}^{(\mathscr{S}_1, \mathscr{S}_2)}$ is the submatrix of $\boldsymbol{\mathscr{A}}$ with rows in $\mathscr{S}_1$ and columns in $\mathscr{S}_2$. This notation is used for matrices of operators, such as $\boldsymbol{\mathscr{C}}$, $\boldsymbol{\mathscr{T}}$, and $\boldsymbol{\mathscr{T}}_n$. Consistent with this notation, $\mathscr{T}^{(j,j)} = \mathbb{C}\mathrm{ov}(X_j)$ is the jth diagonal element of $\boldsymbol{\mathscr{T}}$, and define $\mathscr{T}_\lambda^{(j,j)} = \mathscr{T}^{(j,j)} + \lambda \mathscr{I}$ for any $\lambda > 0$ where $\mathscr{I}$ is the identity operator. Let $\boldsymbol{\mathscr{Q}}^{(\mathscr{S},\mathscr{S})} = \mathrm{diag}\{\mathscr{T}^{(j,j)}, j \in \mathscr{S}\}$ be the

operator-valued matrix that only contains the diagonal terms of $\boldsymbol{\mathscr{T}}^{(\mathscr{S},\mathscr{S})}$, and let $\boldsymbol{\mathscr{Q}}_{\lambda}^{(\mathscr{S},\mathscr{S})} = \boldsymbol{\mathscr{Q}}^{(\mathscr{S},\mathscr{S})} + \lambda \boldsymbol{\mathscr{I}}$.

In addition to Condition C.1, we need the following conditions.

C.2. *Each* $\mathscr{T}^{(j,j)}$ *is standardized such that* $\|\mathscr{T}^{(j,j)}\|_2 = 1$, *with its trace uniformly bounded by a finite constant* τ, *i.e.,* $\sup_{j\in\{1,\ldots,p\}} \mathrm{tr}(\mathscr{T}^{(j,j)}) \leq \tau$.

C.3. *Define* $\varkappa(\lambda_2) := \left|\left|\left| \boldsymbol{\mathscr{T}}^{(\mathscr{S},\mathscr{S})} (\boldsymbol{\mathscr{T}}_{\lambda_2}^{(\mathscr{S},\mathscr{S})})^{-1} \right|\right|\right|_{\infty,\infty}$. *Assume that for some* $\gamma \in (0,1]$, *we have* $\varkappa(\lambda_2) \cdot \left|\left|\left| \boldsymbol{\mathscr{T}}^{(\mathscr{S}^c,\mathscr{S})} (\boldsymbol{\mathscr{T}}^{(\mathscr{S},\mathscr{S})})^{-} \right|\right|\right|_{\infty,\infty} \leq (C_{\min}/C_{\max})(1-\gamma)$, *where* $(\boldsymbol{\mathscr{T}}^{(\mathscr{S},\mathscr{S})})^{-}$ *is the Moore-Penrose generalized inverse of* $\boldsymbol{\mathscr{T}}^{(\mathscr{S},\mathscr{S})}$.

C.4. $\aleph(\lambda_2) := \left|\left|\left| \left(\boldsymbol{\mathscr{T}}^{(\mathscr{S},\mathscr{S})} - \boldsymbol{\mathscr{Q}}^{(\mathscr{S},\mathscr{S})}\right) \left(\boldsymbol{\mathscr{Q}}_{\lambda_2}^{(\mathscr{S},\mathscr{S})}\right)^{-1} \right|\right|\right|_{\infty,\infty} < 1.$

Remark 3. *(i) Condition C.2 places a mild constraint on the decay rate of the eigenvalues for* $\mathscr{T}^{(j,j)}$, *which is equivalent to* $\sup_{j\in\{1,\ldots,p\}} \mathbb{E}\|\widetilde{X}_j\|_2^2 \leq \tau$.

(ii) Condition C.3 controls the correlation between functional predictors in the true signal set $\mathscr{S}$ *and those in the non-signal set* $\mathscr{S}^c$. *This assumption is related to the so-called "irrepresentable condition" on model selection consistency of the classical lasso (Zhao and Yu, 2006; Wainwright, 2009), the classical elastic-net (Jia and Yu, 2010), and the sparse additive models (Ravikumar et al., 2009). Condition C.3 becomes harder to fulfill when* $\varkappa(\lambda_2)$ *is large or when* $C_{\min}/C_{\max}$ *is small. However, when the predictors in* $\mathscr{S}$ *and in* $\mathscr{S}^c$ *are uncorrelated, then* $\left|\left|\left| \boldsymbol{\mathscr{T}}^{(\mathscr{S}^c,\mathscr{S})} (\boldsymbol{\mathscr{T}}^{(\mathscr{S},\mathscr{S})})^{-} \right|\right|\right|_{\infty,\infty} = 0$ *and the assumption holds trivially.*

(iii) Condition C.4 puts constraints on the correlations between the predictors in the true signal set $\mathscr{S}$, so that none of the true predictors can be represented by other predictors in $\mathscr{S}$. When the predictors in $\mathscr{S}$ are uncorrelated, then $\aleph(\lambda_2) = 0$ and C.4 trivially holds.

To gain a deeper understanding of Conditions C.2-C.4, an example will be provided in Section S.1.2 where the functional predictors have a partially separable covariance structure (Zapata et al., 2021). To state the variable selection consistency properties of our approach, we further assume without loss of generality that $\|\boldsymbol{f}_{0\mathscr{S}}\|_\infty = 1$ below. Also, the symbol D^* and similar symbols below will denote universal constants in $(0, \infty)$ that arise from inequalities, whose values change from line to line but do not depend on the model parameters, sample size, or regularization parameters. The specific expressions of universal constants may be complicated and do not add to the understanding of the results. With these in mind, define the following conditions on λ_1, λ_2:

$$\lambda_1/\lambda_2 > \left(\frac{3}{\gamma} - 2\right) C_{\max}^{-1}, \quad D_{1,1}^* > \lambda_1 > D_{1,2}^* \frac{\tau^{1/2}(1+\sigma)}{C_{\min}\gamma} \sqrt{\frac{\log(p-q)}{n}},$$
$$D_{2,1}^* > \lambda_2 > D_{2,2}^* \frac{\tau(1+\sigma)(\rho_1+1)}{(C_{\min}/C_{\max})^2\gamma^2} \max\left(\frac{q\log(p-q)}{n}, \sqrt{\frac{q^2}{n}}\right). \tag{3.6}$$

where ρ_1 denotes the largest eigenvalue of $\boldsymbol{\mathcal{T}}^{(\mathscr{S},\mathscr{S})}$ and $D_{1,1}^*, D_{1,2}^*, D_{2,1}^*, D_{2,2}^*$ are universal constants. It is worth emphasizing that by carefully separating the model/regularization parameters with universal constants, our nonasymptotic

results below can be readily used to state asymptotic results for which some or all of the parameters could change with n. An example of that is provided in Corollary 1 below.

Finally, define the signal set containing predictors with "substantial" predictive power $\mathscr{S}_G := \{j \in \mathscr{S} : \left\|(\mathscr{T}^{(j,j)})^{1/2} f_{0j}\right\|_2 > G\}$, where $G \in (0, \infty)$; recall $\|(\mathscr{T}^{(j,j)})^{1/2} f_{0j}\|_2^2 = \mathbb{E}\langle X_j, \beta_j\rangle_2^2$. The variable selection consistency of our functional elastic-net approach is given in the following result.

Theorem 1. *Consider the functional elastic-net problem* (2.4). *Suppose that Conditions C.1-C.3 and* (3.6) *hold. Then* $\widehat{\mathscr{S}}$ *exists uniquely, and (i) and (ii) below hold with probability at least*

$$1 - \exp\left(-D\frac{\lambda_2^2 n}{q}\right), \quad \textit{where} \quad D = D^*\left(\frac{(C_{\min}/C_{\max})\gamma}{\tau^{1/2}(\rho_1+1)(\sigma+1)}\right)^2, \tag{3.7}$$

for some universal constant D^*.

(i) The estimated signal set is contained in the true signal set, i.e. $\widehat{\mathscr{S}} \subset \mathscr{S}$.

(ii) Under the additional Condition C.4, we have $\widehat{\mathscr{S}} \supset \mathscr{S}_G$ *for*

$$G = \frac{12 - 8\aleph(\lambda_2)}{1 - \aleph(\lambda_2)}\left(C_{\max}\sqrt{\lambda_1^2/\lambda_2} + 2\sqrt{\lambda_2}\right),$$

and, in particular, if $\mathscr{S}_G = \mathscr{S}$, *then* $\widehat{\mathscr{S}} = \mathscr{S}$ *and variable selection consistency is achieved.*

Remark 4. *(i) Part (i) of Theorem 1 guarantees a sparse solution for the functional elastic-net where all predictors in the non-signal set are eliminated. By examining* (3.6) *and* (3.7)*, we can see that increasing* λ_2 *(and, consequently,* λ_1*) leads to a higher probability of eliminating the non-signals. Condition* (3.6) *also implies that, as the correlation of predictors between the signal and non-signal sets increases (i.e., decreasing value of* γ*), larger values of* $\lambda_1, \lambda_2, \lambda_1/\lambda_2$ *are required. Moreover, larger values of* γ*, smaller values of* τ*, and reduced* σ^2 *(resulting in a decreased correlation between* $\mathscr{S}$ *and* $\mathscr{S}^c$*, faster eigenvalue decay for each* $\mathscr{T}^{(j,j)}$*, and a higher signal-to-noise ratio, respectively) enhance the functional elastic-net's ability to accurately identify the signal set.*

(ii) Part (ii) of Theorem 1 provides conditions that prevent the functional elastic-net from removing the true signals and thus guarantees that the predictors identified by the functional elastic-net are not overly sparse. Large values of λ_1*,* λ_1/λ_2*, and* $\aleph(\lambda_2)$ *result in a larger gap* G*, making signal detection more challenging. This is understandable because a large sparsity penalty can lead to the removal of true signals, especially when there is a strong correlation.*

(iii) Condition (3.6) *requires that the lower bound of* λ_1 *must be of the rate* $\sqrt{\frac{\log(p-q)}{n}}$ *to control sparsity. This is similar to the lower bound of the regularization parameter of the lasso (see Theorem 3 of Wainwright, 2009). Our theory also requires a lower bound for* λ_2 *to control both the smoothness and vari-*

ance of $\widehat{f}_j$. *The roles of* λ_2 *in functional linear regression have been discussed by many (see, e.g., Cai and Yuan, 2012). The classical (finite-dimensional) elastic-net optimization (Zou and Hastie, 2005) includes lasso as a special case, with* $\lambda_2 = 0$. *However, this is not feasible in the infinite-dimensional functional setting. To understand it, consider classical high-dimensional data (in the scalar setting) and let* $\mathbf{\Sigma}_{\mathscr{S}}$ *be the* $q \times q$ *covariance matrix of the true predictors. A common assumption to avoid collinearity in that setting is to bound the minimum eigenvalue of* $\mathbf{\Sigma}_{\mathscr{S}}$ *away from zero (Zhao and Yu, 2006; Wainwright, 2009), which is why* λ_2 *could be taken as zero. We cannot bound the eigenvalues of* $\boldsymbol{\mathscr{T}}^{(\mathscr{S},\mathscr{S})}$ *that way in the functional setting because it contradicts the intrinsic infinite dimensionality of functional data; in fact, the sequence of eigenvalues for* $\boldsymbol{\mathscr{T}}^{(\mathscr{S},\mathscr{S})}$ *shrinks to zero even if all the predictors in* $\mathscr{S}$ *are uncorrelated.*

Following Cai and Yuan (2012), we also study the excess risk as a metric to measure the prediction accuracy of the estimator $\mathscr{R}(\boldsymbol{f}) = \mathbb{E}\left(\sum_{j=1}^{p}\langle \widetilde{X}_j^*, f_{0j} - f_j\rangle_2\right)^2$, where $\widetilde{\boldsymbol{X}}_{\bullet}^*$ is a copy of $\widetilde{\boldsymbol{X}}_{i\bullet}$. The excess prediction risk of our estimator, $\widehat{\boldsymbol{f}}$, is obtained by plugging $\widehat{\boldsymbol{f}}$ in $\mathscr{R}(\boldsymbol{f})$. The following result describes the excess prediction risk of the functional elastic-net estimator.

Theorem 2. *Assume that Conditions C.1-C.3 and* (3.6) *hold. Then, the excess risk satisfies* $\mathscr{R}(\widehat{\boldsymbol{f}}) < q\,(4C_{\max}\lambda_1 + 4\lambda_2 + C_{\max}^2\lambda_1^2/\lambda_2)$ *with probability*

bounded below by the expression in (3.7).

Next, we discuss asymptotic results readily derived from Theorems 1 and 2 by allowing p, q as well as the model/regularization parameters to vary with the sample size n. To facilitate the discussion, denote $a_k \asymp b_k$ for two positive sequences $\{a_k\}_{k=1}^{\infty}$ and $\{b_k\}_{k=1}^{\infty}$, if $c_1 < a_k/b_k < c_2$ for some $0 < c_1 < c_2 < \infty$ and for all k. The following corollary is a direct result of Theorem 2, the proof of which is in the Supplementary Material.

Corollary 1. *Assume that Conditions C.1-C.3 and* (3.6) *hold, where* $C_{\min}$ *and* γ *are bounded away from* 0, *and* ρ_1, σ^2, τ, *and* $C_{\max}$ *bounded away from* ∞. *Let* $\alpha(p,q,n) := \max\left(q, \sqrt{\log(p-q)}, \sqrt{q\log n}\right)$ *and assume that* $q\alpha(p,q,n) = o(n^{1/2})$. *Then, for some sufficiently large constant* D, *the probability that* $\mathscr{R}\left(\widehat{\boldsymbol{f}}\right) > Dn^{-1/2}q\alpha(p,q,n)$ *infinitely often is* 0.

Remark 5. *Consider a high dimension FLM setting where* $q \asymp n^{\varsigma}$ *for some* $0 < \varsigma < 1/4$, *and suppose all functional predictor in the signal set have about the same contribution to the variation of the response such that* $G = \min_{j\in\mathscr{S}} \|(\mathscr{T}^{(j,j)})^{1/2} f_{0j}\|_2 \asymp 1/\sqrt{q}$. *By Theorem 1 (ii), we can choose* $\lambda_1 \asymp \lambda_2 \asymp (1/q)$ *to guarantee recovery of the signal set* $\mathscr{S}_G$. *Condition* (3.6) *is also satisfied if* $\log p = O(n^{1-2\varsigma})$, *which is an ultra-high-dimensional FLM setting. Under this setting and with the choice of tuning parameters described above, the probability bound in* (3.7) *goes to 1 which ensures variable selection consis-*

tency; the condition $q\alpha(p,q,n) = o(n^{1/2})$ *in Corollary 1 is also satisfied, and we can conclude* $\mathscr{R}\left(\widehat{\boldsymbol{f}}\right) \to 0$ *almost surely.*

3.2 Oracle minimax optimal rate and a post-selection refined estimator

Cai and Yuan (2012) established the minimax lower bound of the excess prediction risk for univariate FLM with $q = 1$. Such a lower bound is yet to be established for high-dimensional FLMs. In this subsection, we first investigate the minimax lower bound of the excess prediction risk under the oracle model, where $\mathscr{S}$ is known and the true number of functional predictors q is allowed to diverge with the sample size n. We need the following conditions for our results.

C.5. *For each* $j \in \mathscr{S}$*, the* k*-th eigenvalue of* $\mathscr{T}^{(j,j)}$ *is bounded by* ck^{-2r} *for some* $c \in (0,\infty)$ *and* $r > 1/2$*. For some* $b \in (0,\infty)$*,*

$$\sup_{\alpha>0} \left|\!\left|\!\left| \left(\boldsymbol{\mathcal{Q}}_{\alpha}^{(\mathscr{S},\mathscr{S})}\right)^{-1/2} \boldsymbol{\mathcal{T}}_{\alpha}^{(\mathscr{S},\mathscr{S})} \left(\boldsymbol{\mathcal{Q}}_{\alpha}^{(\mathscr{S},\mathscr{S})}\right)^{-1/2} \right|\!\right|\!\right|_{2,2} \leq b \tag{3.8}$$

Condition C.5 requires that the eigenvalues of each $\mathscr{T}^{(j,j)}$, $j \in \mathscr{S}$, to decay in a polynomial rate, which is the same assumption made in Cai and Yuan (2012). By requiring $r > 1/2$, each $\mathscr{T}^{(j,j)}$ is a linear operator that belongs to the trace class, which includes the Hilbert-Schmidt operators. Condition C.5 trivially holds when $\boldsymbol{\mathcal{T}}^{(\mathscr{S},\mathscr{S})} = \boldsymbol{\mathcal{Q}}^{(\mathscr{S},\mathscr{S})}$ meaning that the functional predictors

are uncorrelated. When the functional predictors have a partially separable covariance structure, (3.8) holds in mild conditions (see supplementary materials). The following proposition and its corollary further illustrate what Condition C.5 entails.

Proposition 3. *Assume* (3.8) *holds, we have* $\Lambda_k(\boldsymbol{\mathcal{T}}^{(\mathscr{S},\mathscr{S})}) \leq b\Lambda_k(\boldsymbol{\mathcal{Q}}^{(\mathscr{S},\mathscr{S})})$, *where* $\Lambda_k(\boldsymbol{\mathcal{T}}^{(\mathscr{S},\mathscr{S})})$ *and* $\Lambda_k(\boldsymbol{\mathcal{Q}}^{(\mathscr{S},\mathscr{S})})$ *denote the* k*-th largest eigenvalues of* $\boldsymbol{\mathcal{T}}^{(\mathscr{S},\mathscr{S})}$ *and* $\boldsymbol{\mathcal{Q}}^{(\mathscr{S},\mathscr{S})}$, *respectively.*

Corollary 2. *Assume Condition C.5 holds, let* $\{\rho_l = \Lambda_l(\boldsymbol{\mathcal{T}}^{(\mathscr{S},\mathscr{S})})\}_{l\geq 1}$ *be the eigenvalues of* $\boldsymbol{\mathcal{T}}^{(\mathscr{S},\mathscr{S})}$ *in a decreasing order, then* $\rho_{q(k-1)+j} \leq bc \cdot k^{-2r}$ *for any* $k \geq 1$ *and* $j = 1, \ldots, q$.

Corollary 2 is a direct result of Proposition 3 and is essential in deriving the minimax lower bound in the following theorem.

Theorem 3. *Let* $\boldsymbol{\mathscr{P}}(r)$ *be the class of covariance operators satisfying Condition C.5. Then*

$$\lim_{a\to 0}\lim_{n\to\infty}\inf_{\widetilde{f}_{\mathscr{S}}}\sup_{\boldsymbol{\mathcal{T}}^{(\mathscr{S},\mathscr{S})}\in\boldsymbol{\mathscr{P}}(r)}\sup_{\boldsymbol{f}_{0\mathscr{S}}\in\mathbb{L}_2^q}\mathbb{P}\left(\mathscr{R}(\widetilde{f}_{\mathscr{S}}) \geq a(n/q)^{-\frac{2r}{2r+1}}\right) = 1,$$

where the infimum is taken over all possible predictors $\widetilde{f}_{\mathscr{S}}$ *based on the training data* $\{(\boldsymbol{X}_{i\mathscr{S}}, Y_i), i = 1, \ldots, n\}$.

Theorem 3 provides the oracle minimax lower bound for the excess prediction risk of the high dimensional FLM, which reduces to the lower bound

of Cai and Yuan (2012) if $q = 1$. By comparing this result with Corollary 1, we can see that the excess risk of the functional elastic-net, $\mathscr{R}(\widehat{\boldsymbol{f}})$, is at a rate slower than $(n/q)^{-1/2}$, which in turn is slower than the oracle minimax rate in Theorem 3 when $r > 1/2$. This is understandable, since the primary goal of functional elastic-net is to perform variable selection. Suppose all assumptions in Theorem 1 hold and $\mathscr{S} = \mathscr{S}_G$, the functional elastic-net estimator enjoys variable selection consistency and can help us find an estimated signal set $\widehat{\mathscr{S}}$ that satisfies the following condition.

C.6. $\lim_{n\to\infty} \sup_{\boldsymbol{\mathcal{F}}^{(\mathscr{S},\mathscr{S})}\in\boldsymbol{\mathscr{P}}(r)} \sup_{\boldsymbol{f}_{0,\mathscr{S}}\in\mathbb{L}_2^q} \mathbb{P}\left(\widehat{\mathscr{S}} \neq \mathscr{S}\right) = 0.$

This motivates us to refine our FLM estimator within the selected signal set with the goal of improving the excess prediction risk,

$$\widehat{\boldsymbol{f}}_{\widehat{\mathscr{S}}} = \underset{f_j\in\mathbb{L}_2}{\operatorname{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \sum_{j\in\widehat{\mathscr{S}}}\langle \widetilde{X}_{ij}, f_j\rangle_2\right)^2 + \lambda_3 \sum_{j\in\widehat{\mathscr{S}}} \|f_j\|_2^2 \right\}. \tag{3.9}$$

The refined estimator (3.9) is a special case of the functional elastic-net estimator in Section 2.2 by including functional predictors in $\widehat{\mathscr{S}}$ only and setting the ℓ_1 penalty to 0, as the focus has shifted away from variable selection. As such, $\widehat{\boldsymbol{f}}_{\widehat{\mathscr{S}}}$ can be calculated the same way as the functional elastic-net with a minimum modification to the algorithm.

Theorem 4. *Assume Conditions C.5-C.6 hold, and the number of true signals satisfies* $q = o\left(n^{\frac{2r-1}{4r}}\right)$. *Then*

$$\lim_{A\to\infty}\lim_{n\to\infty}\sup_{\mathscr{T}^{(\mathscr{S},\mathscr{S})}\in\mathscr{P}(r)}\sup_{\boldsymbol{f}_{0,\mathscr{S}}\in\mathbb{L}_2^q}\mathbb{P}\left(\mathscr{R}(\widehat{\boldsymbol{f}}_{\mathscr{S}})\geq A(n/q)^{-\frac{2r}{2r+1}}\right)=0,$$

provided that $\lambda_3 \asymp (n/q)^{-2r/(2r+1)}$.

Theorem 4 shows that our refined estimator (3.9) achieves the oracle the minimax rate in Theorem 3, which is determined by the rate of decay of the eigenvalues of the operator $\mathscr{T}^{(\mathscr{S},\mathscr{S})}$. When q is a constant that does not grow with n, the minimax rate for the excess risk is on the order of $n^{-2r/(2r+1)}$, consistent with the findings in Cai and Yuan (2012).

4. Implementation and Numerical Studies

4.1 Practical Implementation

Proposition 1 provides an expression for the exact solution to the optimization problem (2.4), where each $\widehat{f}_j$ is a linear combination of $\widetilde{\boldsymbol{X}}_{\bullet j}$. However, such a solution is not scalable to big data and ultra-high dimensions, since there are a total of np parameters to estimate. In this subsection, we propose a computationally-efficient algorithm to fit the model based on the idea of reduced-rank approximations, which has been widely used in semiparametric regression (Ruppert et al., 2003) and spline smoothing (Ma et al., 2015). Our low-rank approximation shares a similar spirit as the eigensystem truncation approach proposed by Xu and Wang (2021) for a low-rank approximation of smoothing splines.

Since $\widehat{f}_j$ falls in the subspace spanned by $\widetilde{\boldsymbol{X}}_{\bullet j}$, it can be well approximated by the eigenfunctions of $\mathscr{T}_n^{(j,j)}$, which is the empirical covariance of $\widetilde{\boldsymbol{X}}_{\bullet j}$. Let $\boldsymbol{\varphi}_j(t) = (\varphi_{j1}, \ldots, \varphi_{jM_j})^\top(t)$ be the first M_j eigenfunctions of $\mathscr{T}_n^{(j,j)}$, such that $\int_0^1 \boldsymbol{\varphi}_j(t)\boldsymbol{\varphi}_j^\top(t)dt = \boldsymbol{I}_{M_j}$, and we approximate f_j with $\widetilde{f}_j(t) = \boldsymbol{\varphi}_j^\top(t)\boldsymbol{c}_j$. Define $\boldsymbol{\Gamma}_j = \int_0^1 \widetilde{\boldsymbol{X}}_{\bullet j}(t)\boldsymbol{\varphi}_j^\top(t)dt$ and $\boldsymbol{H}_j = \int_0^1 (\Psi_j\boldsymbol{\varphi}_j)(t)(\Psi_j\boldsymbol{\varphi}_j)^\top(t)dt$. We reparameterize the coefficient vectors as $\boldsymbol{d}_j = \boldsymbol{H}_j^{1/2}\boldsymbol{c}_j$, and solve the group elastic-net problem (2.4) iteratively using a block coordinate-descent algorithm. At coordinate j, we fix $\boldsymbol{d}_{j'}$ for $j' \neq j$, define $\widetilde{\boldsymbol{Y}}_n^{(j)} = \boldsymbol{Y}_n - \sum_{j'\neq j}\boldsymbol{\Gamma}_{j'}\boldsymbol{H}_{j'}^{-1/2}\boldsymbol{d}_{j'}$, and update $\boldsymbol{d}_j$ by

$$\widehat{\boldsymbol{d}}_j = \underset{\boldsymbol{d}_j\in\mathbb{R}^{M_j}}{\operatorname{argmin}} \left\{ \frac{1}{2}\boldsymbol{d}_j^\top \boldsymbol{\Omega}_j \boldsymbol{d}_j - \boldsymbol{\varrho}_j^\top \boldsymbol{d}_j + \lambda_1 \|\boldsymbol{d}_j\|_2 \right\}, \tag{4.10}$$

where $\boldsymbol{\Omega}_j = \boldsymbol{H}_j^{-1/2}\left(\frac{1}{n}\boldsymbol{\Gamma}_j^\top\boldsymbol{\Gamma}_j + \lambda_2\boldsymbol{I}_{M_j}\right)\boldsymbol{H}_j^{-1/2}$ and $\boldsymbol{\varrho}_j = n^{-1}\boldsymbol{H}_j^{-1/2}\boldsymbol{\Gamma}_j^\top\widetilde{\boldsymbol{Y}}_n^{(j)}$.

Proposition 4 provides the solution to the minimization problem (4.10).

Proposition 4. *For $\lambda_1 > 0$, the solution $\widehat{\boldsymbol{d}}_j$ for (4.10) exists. Furthermore, if $\|\boldsymbol{\varrho}_j\|_2 \leq \lambda_1$, then $\widehat{\boldsymbol{d}}_j = 0$; if $\|\boldsymbol{\varrho}_j\|_2 > \lambda_1$, then $\widehat{\boldsymbol{d}}_j \neq 0$ and $\widehat{\boldsymbol{d}}_j$ is the solution to $\boldsymbol{\Omega}_j\boldsymbol{d}_j - \boldsymbol{\varrho}_j + \lambda_1\boldsymbol{d}_j\|\boldsymbol{d}_j\|_2^{-1} = \boldsymbol{0}$.*

We can solve $\widehat{\boldsymbol{d}}_j$ by iteratively updating $\boldsymbol{d}_j \leftarrow \left(\boldsymbol{\Omega}_j + \lambda_1\|\boldsymbol{d}_j\|_2^{-1}\boldsymbol{I}_{M_j}\right)^{-1}\boldsymbol{\varrho}_j$ until convergence. Since the objective function (4.10) is the combination of a convex and differentiable least squares loss and a convex penalty, the block coordinate-wise algorithm is guaranteed to converge to the global minimum

(Friedman et al., 2007).

For the refined estimator in (3.9), no iteration is needed since there is no ℓ_1 penalty involved. Write $\widehat{f}_j(t) = \boldsymbol{\varphi}_j^\top(t)\widehat{\boldsymbol{c}}_j$ for each $j \in \widehat{\mathscr{S}} \equiv \{j_1, j_2, \ldots, j_{\widehat{q}}\}$. Then, the coefficient vectors can be calculated as

$$\left(\widehat{\boldsymbol{c}}_{j_1}^\top, \ldots, \widehat{\boldsymbol{c}}_{j_{\widehat{q}}}^\top\right)^\top = \frac{1}{n}\left(\frac{1}{n}\boldsymbol{\Gamma}_{\widehat{\mathscr{S}}}^\top \boldsymbol{\Gamma}_{\widehat{\mathscr{S}}} + \lambda_3 \boldsymbol{I}\right)^{-1} \boldsymbol{\Gamma}_{\widehat{\mathscr{S}}}^\top \boldsymbol{Y}_n,$$

where $\boldsymbol{\Gamma}_{\widehat{\mathscr{S}}} = \left(\boldsymbol{\Gamma}_{j_1}, \ldots, \boldsymbol{\Gamma}_{j_{\widehat{q}}}\right)$ is the design matrix for functional predictors in the estimated signal set.

In most applications, the functional predictors are observed on N equally spaced points, the kernel functions K_j are evaluated as $N \times N$ matrices, $K_j^{1/2}$ are computed via the spectral decomposition of K_j, and all integrals can be approximated using Riemann sums on the observed discrete points. As discussed in Zhou et al. (2023), the incurred errors by these approximations are negligible when N is sufficiently large.

4.2 Simulation Studies

We simulate the functional predictors as $X_{ij}(t) = \sqrt{2}\sum_{k\geq 1} z_{ijk}\sqrt{\nu_k}\cos(k\pi t)$, $(i = 1, \ldots, n, j = 1, \ldots, p)$, where $\boldsymbol{z}_{i\cdot k} = (z_{i1k}, \ldots, z_{ipk})^\top \sim$ i.i.d. Normal$(\boldsymbol{0}, \boldsymbol{\Sigma}_p)$, and $\boldsymbol{\Sigma}_p$ is an autoregressive correlation matrix with the (j,k)th entry being $\rho^{|j-k|}$, $1 \leq k, j \leq p$. We generate the response Y by the high-dimensional functional linear regression model (2.1), using coefficient functions under one

of the three scenarios described below and setting $\epsilon_i \sim \text{Normal}(0, \sigma^2 = 0.5^2)$. For each scenario, we consider three correlation levels between the functional predictors, $\rho = 0$, 0.3 and 0.75, and three settings for the problem size: a high dimension and high sample size setting with $(n, p, q) = (500, 50, 5)$, a high dimension and low sample size setting with $(n, p, q) = (200, 100, 5)$, and an ultra-high dimension setting with $(n, p, q) = (100, 200, 10)$. For simplicity, we set the signal set to be $\mathscr{S} = \{1, \ldots, q\}$, and set $\beta_{0j}(t) = 4\sum_{k\geq 1}(-1)^{u_{jk}} r_k \phi_k(t)$, for $j \in \mathscr{S}$, where the basis functions $\phi_k(t)$ and coefficients r_k are to be specified below, u_{jk} are i.i.d. Bernoulli random variables with $P(u_{jk} = 1) = 0.5$. Inspired by Cai and Yuan (2012), we consider the following three scenarios for $\{\phi_k(t), r_k, \nu_k\}$:

Scenario I: $\phi_k(t) = \sqrt{2}\cos(k\pi t)$, and $\nu_k = r_k = \exp(-k/4)$, for $k \geq 1$;

Scenario II: $\phi_k(t) = \sqrt{2}\sin(k\pi t)$, and $\nu_k = r_k = \exp(-k/4)$, for $k \geq 1$;

Scenario III: $\phi_k(t) = \sqrt{2}\cos(k\pi t)$, $r_k = k^{-2}$, and $\nu_k = (|k - k_0| + 1)^{-2}$ for $k \geq 1$, where we set $k_0 = 10$.

Scenario I represents a case where the functional predictors and the coefficient functions are perfectly aligned. Not only they are spanned by the same set of cosine functions, but the eigenvalues ν_k and the coefficients r_k both monotonically decay with k. In other words, the signals most important to X_{ij} also contribute the most to Y_i. As shown by Cai and Yuan (2012), β_{0j} under

this scenario belong to an RKHS with the RKHS norm $\|\beta\|_{\mathbb{H}} = \{\int (\beta'')^2\}^{1/2}$, and the reproducing kernel $K(s,t) = -\frac{1}{3}\big[B_4(|s-t|/2) + B_4\{(s+t)/2\}\big]$, where B_k is the kth Bernoulli polynomial.

Scenarios Ⅱ and Ⅲ represent various cases of misalignment. Under Scenario Ⅱ, X_{ij} and β_{0j} are spanned by different bases. Using similar derivations as Cai and Yuan (2012), we can show β_{0j} belong to an RKHS with the reproducing kernel $K(s,t) = -\frac{1}{3}\big[B_4(|s-t|/2) - B_4\{(s+t)/2\}\big]$. Under Scenario Ⅲ, the maximum mode of variation in X_{ij} is contributed from a high-frequency cosine function with $k = k_0$, however, these high-frequency signals do not contribute much to the response because the corresponding r_k's are small. Even though the polynomial decay of the coefficient $r_k = k^{-2}$ in Scenario Ⅲ is slower than the exponential series $r_k = \exp(-k/4)$ in the asymptotic sense, as it turns out $\exp(-k/4) \geq k^{-2}$ for $k \leq 26$. As such, there are practically more random components that contribute to the variations in X_{ij} and the response Y_i under Scenarios I and Ⅱ.

We repeat the simulation 200 times for each scenario, each level of correlation, and each problem size. For each simulated data set, we also simulate an additional sample of 100 data pairs of $(\boldsymbol{X}, Y)$ as testing data to evaluate the prediction performance. We apply our proposed functional elastic-net (fEnet) method to each simulated data set and make a comparison with the

method proposed by Xue and Yao (2021), which is to equip high-dimensional functional linear regression with a SCAD penalty (Fan and Li, 2001) and thus termed FLR-SCAD. For FLR-SCAD, there are two tuning parameters, the SCAD penalty parameter λ and the number of basis functions s_1 to represent both the functional predictor and the coefficient functions. For a fair comparison, we set the basis of FLR-SCAD to be the true basis $\phi_k(t)$ as described above. For the proposed fEnet, we set $\Psi_j = (\mathscr{T}_n^{(j,j)} + \theta\mathscr{I})^{1/2}$ and hence end up with four tuning parameters (λ, α, s, and θ), where $\lambda_1 = \alpha\lambda$, $\lambda_2 = (1-\alpha)\lambda$, and s is the number of eigenfunctions used in the reduced rank approximation described in Section 4.1. For both methods, the tuning parameters are selected based on a grid search that minimizes the averaged mean square prediction error using the testing sample so that the results reported here represent the best possible performance of the two. For a single tuning configuration, fEnet matches FLR–SCAD in runtime (around 3 seconds at the optimal tuning parameter in the ultra–high–dimensional case). Since fEnet is insensitive to the basis size s, we fix s and parallelize cross-validation to keep total cost low. We use false positive rate (FPR) and false negative rate (FNR), defined as FPR$= |\widehat{\mathscr{S}} \cap \mathscr{S}^c|/|\mathscr{S}^c|$ and FNR$= |\widehat{\mathscr{S}^c} \cap \mathscr{S}|/|\mathscr{S}|$, to assess the variable selection performance, and we use the maximum norm difference (MND) to gauge the signal recovery performance, where MND is defined as the maximum of

the $\mathbb{L}_2$ norm of $\widehat{\beta}_j - \beta_{0j}$ for $j = 1, \ldots, p$. In order to make results from the three scenarios more comparable, we measure prediction error by the relative excess risk (RER): $\mathbb{E}\{\sum_{j=1}^{p}\langle X_j^*, (\hat{\beta}_j - \beta_{j0})\rangle\}^2 / \mathbb{E}\{\sum_{j=1}^{p}\langle X_j^*, \beta_{j0}\rangle\}^2$, which is a standardized version of the excess risk.

Table 1: Simulation Scenario I: summary of estimation, prediction, and variable selection performance of the proposed fEnet method versus FLR-SCAD under different problem sizes.

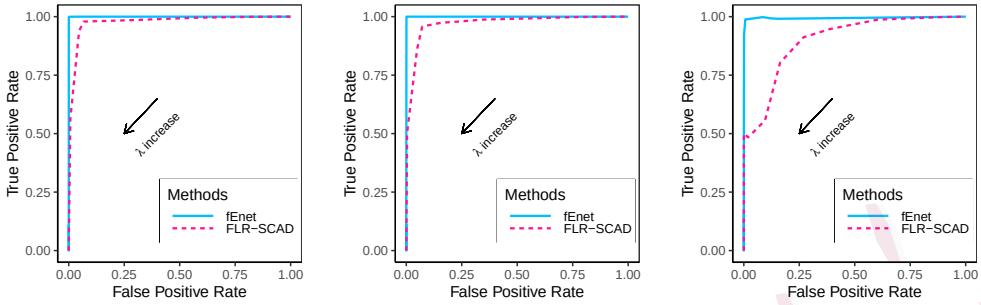
n	p	q	Method	FPR (%)	FNR (%)	MND	RER
				$\rho = 0$			
500	50	5	fEnet	0 (0, 0)	0 (0, 0)	0.36 (0.30, 0.45)	0.0006 (0.0003, 0.0009)
			FLR-SCAD	0 (0, 0)	0 (0, 0)	0.54 (0.37, 0.82)	0.0009 (0.0005, 0.0019)
200	100	5	fEnet	0 (0, 0)	0 (0, 0)	0.53 (0.42, 0.68)	0.0018 (0.0011, 0.0029)
			FLR-SCAD	0 (0, 0)	0 (0, 0)	0.75 (0.58, 1.19)	0.0035 (0.0017, 0.0106)
100	200	10	fEnet	0 (0, 1.1)	0 (0, 0)	1.31 (1.06, 1.65)	0.0179 (0.0094, 0.0399)
			FLR-SCAD	4.7 (1.6, 8.4)	0 (0, 30)	4.89 (3.97, 5.00)	0.5280 (0.3206, 0.7734)
				$\rho = 0.3$			
500	50	5	fEnet	0 (0, 0)	0 (0, 0)	0.37 (0.31, 0.47)	0.0007 (0.0004, 0.0011)
			FLR-SCAD	0 (0, 0)	0 (0, 0)	0.59 (0.41, 1.03)	0.0012 (0.0006, 0.0027)
200	100	5	fEnet	0 (0, 0)	0 (0, 0)	0.58 (0.45, 0.73)	0.0025 (0.0015, 0.0044)
			FLR-SCAD	0 (0, 0)	0 (0, 0)	0.78 (0.58, 1.51)	0.0044 (0.0021, 0.0146)
100	200	10	fEnet	0 (0, 1.6)	0 (0, 0)	1.39 (1.08, 1.92)	0.0192 (0.0103, 0.0441)
			FLR-SCAD	4.7 (1.6, 9.5)	10 (0, 40)	5.00 (4.37, 5.05)	0.5319 (0.3665, 0.7523)
				$\rho = 0.75$			
500	50	5	fEnet	0 (0, 0)	0 (0, 0)	0.53 (0.42, 0.67)	0.0012 (0.0007, 0.0019)
			FLR-SCAD	0 (0, 0)	0 (0, 0)	0.98 (0.67, 1.78)	0.0018 (0.0008, 0.0049)
200	100	5	fEnet	0 (0, 0)	0 (0, 0)	0.85 (0.72, 1.03)	0.0035 (0.0021, 0.0056)
			FLR-SCAD	0 (0, 0)	0 (0, 0)	1.28 (0.76, 4.61)	0.0066 (0.0029, 0.1287)
100	200	10	fEnet	0 (0, 4.2)	0 (0, 10)	2.04 (1.49, 5.00)	0.0175 (0.0078, 0.1329)
			FLR-SCAD	2.1 (0, 4.2)	50 (30, 70)	5.86 (5.00, 7.91)	0.2895 (0.1932, 0.3894)

Simulation results under Scenario I are summarized in Table 1, where we compare the median FPR, FNR, MND, and RER as well as their 2.5% and 97.5% quantiles for the two competing methods. As we can see, both methods accurately choose the correct model under the first two problem sizes and for all correlation levels, although our method shows some small advantages in terms of estimation (MND) and prediction (RER). We now focus on the ultra-high dimension setting with $(n, p, q) = (100, 200, 10)$, where our method

shows an overwhelming advantage over FLR-SCAD in all criteria considered for variable selection, estimation, and prediction. Note that under the high correlation setting ($\rho = 0.75$), not only $\{X_{ij},\ j \in \mathscr{S}\}$ are strongly correlated among themselves, but they are also strongly correlated with some of the predictors in $\mathscr{S}^c$. In this case, even though FLR-SCAD mistakes some of the non-signals with some real signals, its prediction performance may not be as bad as when $\rho = 0$ or 0.3.

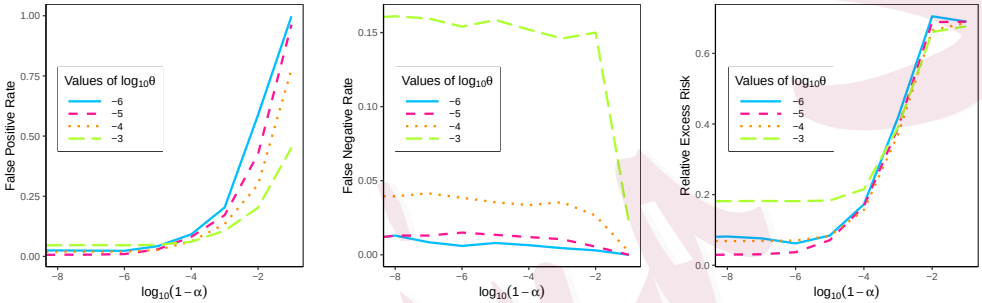
To further investigate the variable selection performance under the ultra-high dimension setting, we plot the receiver operating characteristic (ROC) curves for the two methods in Figure 1, where the false positive rate and true positive rate (TPR), i.e. $1-$FNR, are calculated under different values of λ while holding other tuning parameters fixed at their optimal values. As such, both FPR and TPR become functions of λ. As λ increases, all coefficient functions are shrunk to 0 and hence both FPR and TPR decrease to 0. The ROC of our method yielding a higher area under the curve (AUC) than FLR-SCAD, especially when there is a high correlation between the functional predictors, means that our method has a better variable selection performance.

To investigate the effect of $\alpha = \lambda_1/(\lambda_1+\lambda_2)$ and θ on the variable selection and prediction performance, we revisit the ultra-high dimension setting with $\rho = 0.75$. We calculate the average FPR, FNR, and RER at various values of



(a) $\rho = 0$ **(b)** $\rho = 0.3$ **(c)** $\rho = 0.75$

Figure 1: Simulation Scenario I: The ROC curves of fEnet and FLR-SCAD under the ultra-high-dimension setting $(n, p, q) = (100, 200, 10)$. The ROC curves are obtained by changing the value of λ and holding other hyperparameters at optimal.



(a) FPR **(b)** FNR **(c)** RER

Figure 2: Simulation Scenario I: The plots of FPR, FNR, and RER versus $\log_{10}(1-\alpha)$ for different values of θ under the ultra-high-dimensional case and $\rho = 0.75$.

α and θ while keeping λ and s fixed at their optimal values. In Figure 2 we plot the averaged FPR, FNR, and RER against $\log_{10}(1-\alpha)$ for different values of θ. These plots suggest that for any fixed θ, FPR is a decreasing function of α while FNR increases with α. This observation corroborates our remarks for Theorem 1 that a larger ratio between λ_1 and λ_2 means more predictors will be removed from the model and hence the decreased FPR and increased FNR. There should be an optimal α, which is neither 0 nor 1, providing the

best trade-off between FPR and FNR. The plot of RER against $\log(1-\alpha)$ also suggests the existence of a non-trivial optimal value for α, which in turn suggests that we need both components in the elastic-net penalty for the best performance. By comparing curves across different values of θ, we can see that FPR decreases with θ, FNR increases with θ, and RER is not monotone with θ. All of these point to the conclusion that there is non-zero optimal value for θ.

To save space, results under Scenarios Ⅱ and Ⅲ are deferred to the supplementary material. When there is a misalignment between the functional predictor and the coefficient functions, particularly under Scenario Ⅲ with a high correlation between the functional predictors, we observe better FPR and FNR from the proposed fEnet method not only for the ultra-high dimension setting but all the other problem sizes as well.

Table 2: Relative efficiency (RE) between the functional elastic-net estimate and the two-stage estimate under Scenario I

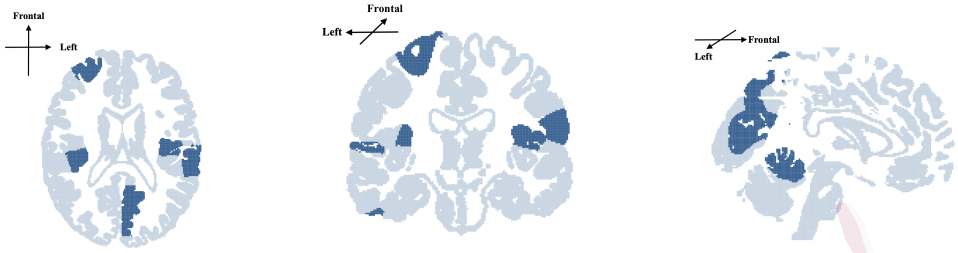
n	p	q	$\rho = 0$	$\rho = 0.3$	$\rho = 0.75$
500	50	5	1.04	1.06	1.29
200	100	5	1.30	1.44	1.51
100	200	10	1.63	1.68	1.95

Next, we demonstrate the efficiency gain of the refined estimator (3.9) in prediction performance. Focusing on Scenario I, we refit FLM to the simulated data as described in (3.9) using the predictors selected by fEnet only. The tun-

ing parameter λ_3 is selected by a grid search that minimizes the averaged mean square prediction error using the testing sample. Table 2 presents a summary of the relative efficiency (RE) between the fEnet estimator $\widehat{\boldsymbol{f}}$ and the refined estimator $\widehat{\boldsymbol{f}}_{\widehat{\mathscr{J}}}$, where $\mathrm{RE}(\widehat{\boldsymbol{f}}, \widehat{\boldsymbol{f}}_{\widehat{\mathscr{J}}}) = \mathrm{RER}(\widehat{\boldsymbol{f}})/\mathrm{RER}(\widehat{\boldsymbol{f}}_{\widehat{\mathscr{J}}})$. The reported REs are based on the average over 200 replicates, and a value of RE greater than 1 indicates an improved prediction performance in the refined estimator. These results demonstrate improved prediction performance of the refined estimator across all problem sizes and correlation levels, particularly in the case of ultra-high-dimension and high correlation between functional predictors, where the refined estimator is almost twice as efficient as the original fEnet.

4.3 Real Data Application

We now demonstrate our methodology using a dataset obtained from the Human Connectome Project (HCP) (Van Essen et al., 2013). The data comprise resting-state fMRI scans from $n = 549$ individuals, where each brain was repeatedly scanned over 1200 time points. These 3-dim fMRI images were pre-processed and parcellated into 268 brain regions-of-interest (ROI) using a whole-brain, functional atlas defined in Finn et al. (2015). Since the raw ROI level fMRI time series are quite noisy, we instead treat the smoothed periodograms at different ROI's as high-dimensional functional data. Specifically, we apply Fast Fourier Transform to the fMRI time series at each ROI,



(a) Top view **(b)** Front view **(c)** Side view

Figure 3: The orthographic projections of a brain (light blue), where the 33 selected ROIs using the HCP data are marked in dark blue.

smooth the resulting periodogram using the 'smooth.spline' function in R, and keep the most informative segment from 1 to 300 Hz as a functional predictor. In addition to the fMRI, each subject in the study also undertook the Penn Progressive Matrix (PPM) test, the score of which is commonly used as a surrogate for fluid intelligence (Greene et al., 2018).

This dataset was previously analyzed by Lee et al. (2023), who used the raw fMRI time series as functional data and the PPM score as a covariate to study functional connectivity between the ROI's. We instead treat the smoothed periodograms from the 268 ROI's as high-dimensional functional predictors and the PPM score as the response. By fitting a high-dimensional functional linear model using the proposed fEnet method, our goal is to identify brain regions that are associated with fluid intelligence.

To ensure the robustness of our results, we randomly divide the 549 individuals into a training set (80%) and a validation set (20%) for a total of 200

times. We select the optimal tuning parameters of our model by minimizing the averaged mean squared prediction error (MSPE) on the 200 validation sets. We find 33 ROIs that are consistently selected by our proposed method across all 200 repetitions. In Figure 3, we provide three projection views of the brain and mark the physical locations of the selected ROIs. Our results suggest that fluid-intelligence-related ROIs are distributed in multiple brain regions, including those on the prefrontal and parietal cortices. These findings agree with the literature (Duncan et al., 2000; Jung and Haier, 2007) that fluid intelligence, considered a complex cognitive ability that involves various cognitive processes, is typically associated with multiple brain regions.

5. Summary

Our RKHS-based functional elastic-net method is different from existing high-dimensional functional linear regression methods in two important ways. First, we do not express the functional predictors and the coefficient functions using the same set of basis functions, which offers the extra flexibility to choose the reproducing kernel based on the application and better numerical performance when the functional predictors and the coefficient functions are misaligned. Second, our penalty consists of two parts: a lasso-type penalty on the normal of the prediction error to enforce sparsity and a ridge penalty that reg-

ularizes the smoothness of the coefficient function for better prediction. Our simulations show that both penalties are important and that the best performance in terms of variable selection, estimation, and prediction is achieved by finding the best trade-off between the two penalties. We also derived a sharp non-asymptotic probability bound on the event of our method achieving variable selection consistency, while assuming the functional predictors are non-degenerative random elements in infinite-dimensional Hilbert spaces. Our theory also suggests a bound for the smallest signal size that can be detected by the functional elastic-net method. Our investigation of the minimax optimal rate for high-dimensional FLM is completely new, and we show that our post-selection refined RKHS estimator achieves the oracle minimax optimal excessive risk. The efficiency gain from using the refined estimator is also demonstrated through simulation studies.

Handling sparsely or irregularly observed functional covariates is an important yet nontrivial extension. There has been some recent work addressing functional linear regression with a single discretely observed predictor under an FPCA framework (Zhou et al., 2023). However, to the best of our knowledge, analogous results within an RKHS framework, particularly for high-dimensional settings, remain unavailable. We note this gap as a promising direction for future research.

Acknowledgement

The authors thank the editor, the associate editor, and two anonymous referees for their many helpful and constructive comments, which led to significant improvements to our paper.

References

Cai, T. T. and Hall, P. (2006). Prediction in functional linear regression. *The Annals of Statistics*, 34:2159–2179.

Cai, T. T. and Yuan, M. (2012). Minimax and adaptive prediction for functional linear regression. *Journal of the American Statistical Association*, 107:1201–1216.

Crambes, C., Kneip, A., and Sarda, P. (2009). Smoothing spline estimators for functional linear regression. *The Annals of Statistics*, 37:35–72.

Duncan, J., Seitz, R. J., Kolodny, J., Bor, D., Herzog, H., Ahmed, A., Newell, F. N., and Emslie, H. (2000). A neural basis for general intelligence. *Science*, 289(5478):457–460.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360.

Fan, Y., James, G. M., and Radchenko, P. (2015). Functional additive regression. *The Annals of Statistics*, 43:2296–2325.

Finn, E. S., Shen, X., Scheinost, D., Rosenberg, M. D., Huang, J., Chun, M. M., Papademetris, X., and Constable, R. T. (2015). Functional connectome fingerprinting: identifying individuals using patterns of brain connec-

tivity. *Nature Neuroscience*, 18(11):1664–1671.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332.

Greene, A. S., Gao, S., Scheinost, D., and Constable, R. T. (2018). Task-induced brain state manipulation improves prediction of individual traits. *Nature Communications*, 9(1):2807.

Hsing, T. and Eubank, R. (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operators*, volume 997. John Wiley & Sons.

James, G. (2002). Generalized linear models with functional predictor variables. *Journal of the Royal Statistical Society, Series B*, 64:411–432.

Jia, J. and Yu, B. (2010). On model selection consistency of the elastic net when $p \gg n$. *Statistica Sinica*, 20:595–611.

Jung, R. E. and Haier, R. J. (2007). The parieto-frontal integration theory (p-fit) of intelligence: converging neuroimaging evidence. *Behavioral and Brain Sciences*, 30(2):135–154.

Lee, K.-Y., Ji, D., Li, L., Constable, T., and Zhao, H. (2023). Conditional functional graphical models. *Journal of the American Statistical Association*, 118(541):257–271.

Lei, J. (2014). Adaptive global testing for functional linear models. *Journal of the American Statistical Association*, 109:624–634.

Liu, Y., Li, Y., Carroll, R. J., and Wang, N. (2022). Predictive functional linear models with diverging number of semiparametric single-index interactions. *Journal of Econometrics*, 230(2):221–239.

Ma, P., Huang, J. Z., and Zhang, N. (2015). Efficient computation of smoothing splines via adaptive basis sampling. *Biometrika*, 102(3):631–645.

Müller, H. G. and Stadtmüller, U. (2005). Generalized functional linear models. *The Annals of Statistics*, 33:774–805.

Qiao, X., Guo, S., and James, G. M. (2019). Functional graphical models. *Journal of the American Statistical Association*, 114(525):211–222.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer, New York, 2nd edition.

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society: Series B*, 71(5):1009–1030.

Reiss, P. T. and Ogden, R. T. (2007). Functional principal component regression and functional partial least squares. *Journal of the American Statistical Association*, 102:984–996.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge university press.

Shang, Z. and Cheng, G. (2015). Nonparametric inference in generalized functional linear models. *The Annals of Statistics*, 43:1742–1773.

Sun, X., Du, P., Wang, X., and Ma, P. (2018). Optimal penalized function-on-function regression under a reproducing kernel hilbert space. *Journal of the American Statistical Association*, 113:1601–1611.

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

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., Consortium, W.-M. H., et al. (2013). The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79.

Wahba, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy

sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202.

Wang, X., Zhu, H., and Initiative, A. D. N. (2017). Generalized scalar-on-image regression models via total variation. *Journal of the American Statistical Association*, 112(519):1156–1168.

Xu, D. and Wang, Y. (2021). Low-rank approximation for smoothing spline via eigensystem truncation. *Stat*, 10(1):e355.

Xue, K. and Yao, F. (2021). Hypothesis testing in large-scale functional linear regression. *Statistica Sinica*, 31:1101 – 1123.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68(1):49–67.

Zapata, J., Oh, S. Y., and Petersen, A. (2021). Partial separability and functional graphical models for multivariate Gaussian processes. *Biometrika*, 109(3):665–681.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38:894–942.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563.

Zhou, H., Yao, F., and Zhang, H. (2023). Functional linear regression for discretely observed data: from ideal to reality. *Biometrika*, 110(2):381–393.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2):301–320.

Zou, H. and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics*, 37:1733 – 1751.