entropy

Article

Multivariate Functional Kernel Machine Regression and Sparse Functional Feature Selection

Joseph Naiman and Peter Xuekun Song *

Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA; jnaiman@umich.edu
* Correspondence: pxsong@umich.edu

Abstract: Motivated by mobile devices that record data at a high frequency, we propose a new methodological framework for analyzing a semi-parametric regression model that allow us to study a nonlinear relationship between a scalar response and multiple functional predictors in the presence of scalar covariates. Utilizing functional principal component analysis (FPCA) and the least-squares kernel machine method (LSKM), we are able to substantially extend the framework of semi-parametric regression models of scalar responses on scalar predictors by allowing multiple functional predictors to enter the nonlinear model. Regularization is established for feature selection in the setting of reproducing kernel Hilbert spaces. Our method performs simultaneously model fitting and variable selection on functional features. For the implementation, we propose an effective algorithm to solve related optimization problems in that iterations take place between both linear mixed-effects models and a variable selection method (e.g., sparse group lasso). We show algorithmic convergence results and theoretical guarantees for the proposed methodology. We illustrate its performance through simulation experiments and an analysis of accelerometer data.

Keywords: functional principal component analysis; functional predictor; linear mixed-effects model; mobile device; sparse group regularization; wearable device data

check for updates

Citation: Naiman, J.; Song, P.X. Multivariate Functional Kernel Machine Regression and Sparse Functional Feature Selection. *Entropy* **2022**, *24*, 203. https://doi.org/10.3390/e24020203

Academic Editors: S. Ejaz Ahmed and Farouk Nathoo

Received: 4 January 2022
Accepted: 26 January 2022
Published: 28 January 2022

1. Introduction

Data captured by mobile devices have lately received much attention in the data science community. Such data are typically recorded at a high frequency, giving rise to an ample volume of information at a very fine scale, and thus present many methodological challenges in statistical modeling and data analyses. In this paper, we plan to utilize the strength of the classical kernel machine method that enjoys fast computing speed via the linear mixed-effects model to deal with such high-frequency data using a functional data analysis approach. The motivation for our proposed framework come from data collected from a tri-axis accelerometer. Accelerometers, worn on the hip or wrist as a way of monitoring physical activity, are becoming more and more common [1–4]. There are several different accelerometers available such as ActiGraph GT3X+ (ActiGraph, Pensacola, FL, USA) and Actical (Phillips Respironics, Bend, OR). Raw accelerometer data are often collected in high-resolution signals with a sampling frequency ranging from 30–100 Hz. The commercial software on these devices provides activity counts (ACs) [2,4], which are calculated from the raw accelerometer data using proprietary algorithms. As an example from our motivating dataset, Figure 1 displays a three-dimensional time series of ACs per minute, each on one axis, from one subject wearing the GT3X+ over a period of 7 days (d).

Oftentimes, different types of summaries of the tri-axis ACs are suggested in the literature as opposed to the utility of all three raw functionals [5–8]. These summary-data-based approaches may be regarded as a quick and dirty dimension reduction strategy that comes up with summarized data with computationally manageable volumes, which would be then analyzed by existing methods and software. One concern with the use of summarized data would be the loss of potential fine features that can only be captured in data of high resolution. Recently, some researchers have attempted to use the entire functional AC curve through functional data analysis techniques [6,9,10]. Further details on current methods being used to retrieve and interpret accelerometer data can be found in [11]. Our contribution in this paper pertains to a new framework in that tri-axis accelerometer data are used as three-dimensional correlated functional predictors in an association analysis with a potential health outcome such as the Body Mass Index (BMI). The relationship between physical activities and childhood obesity has long been a central interest of public health sciences, and our new scalar-on-functional regression model can provide some new insights into this important scientific problem.

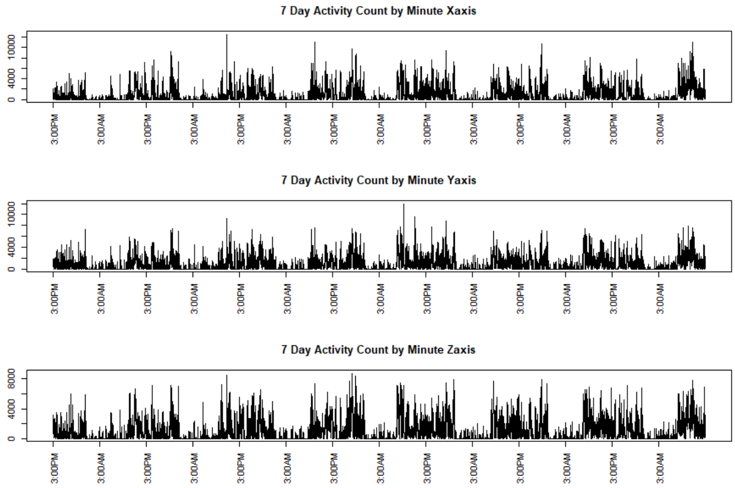

Figure 1. Activity counts over 7 d from a tri-axis (X-, Y- and Z-axis) accelerometer of a subject.

We begin with a brief review of existing functional data models, the least-squares kernel machine model, and different variable selection techniques, which prelude the framework for this paper.

1.1. Functional Regression

There has been much attention in recent years given to functional data analysis (FDA) where either covariates, or response, or both are functional as opposed to scalar in nature [12–17]. In this paper, we focused on the methodology that allows us to relate multiple functional covariates to a scalar outcome in a nonlinear way in the presence of other scalar covariates. To proceed, let us introduce some notation. Let $L^2(\mathcal{T})$ be the class of square-integrable functions on a compact set $\mathcal{T}$. This is a separable Hilbert space with inner product $< f\,, g >:= \int_{\mathcal{T}} fg$ for $f, g \in L^2(\mathcal{T})$. Consider a probability space $(\Omega, \mathcal{F}, P)$, where Z denotes a functional random variable that maps into $L^2(\mathcal{T})$, namely $Z : \Omega \mapsto L^2(\mathcal{T})$. Define $L^2(\Omega) := \{Z : (\int_\Omega \|Z\|^2 dP)^{\frac{1}{2}} < \infty\}$, where P is a certain probability measure, $\|Z\|^2 = < Z\,, Z >$, and assume $Z \in L^2(\Omega)$ in the rest of this paper. For convenience, we also assume that Z is mean centered, namely $E(Z) = 0$.

The class of functional linear models (FLM) (e.g., [13–15]) is proposed to relate a functional covariate Z with a mean-centered scalar outcome y, which is also known as scalar-on-functional regression: $y = < b, Z > + \epsilon$, where the error term ϵ is a mean zero random variable uncorrelated with Z. An optimal solution of the unknown functional parameter $b \in L^2(\mathcal{T})$ is typically obtained by minimizing the mean-squared error: $\inf_{b \in L^2(\mathcal{T})} E(y - < b, Z >)^2$. Moreover, the mean model for the mean-centered scalar y takes the form $E(y|Z) = \int_{\mathcal{T}} Z(t)b(t)dt$.

As suggested in the literature, we may obtain an optimal estimator of b by expanding functional predictor Z under certain basis functions. In this paper, we focus on the utility of functional principal component analysis (FPCA) to perform the decomposition of the functional Z. By the Karhunen–Loève expansion (e.g., [18–20]), we may write

$Z(t) = \sum_{k=1}^{\infty} \sqrt{\varsigma_k}\xi_k\phi_k(t)$, where $\varsigma_k > 0$ are the eigenvalues, and the loadings are given by $\xi_k := \frac{1}{\sqrt{\varsigma_k}} < Z, \phi_k >$. These coefficients satisfy (i) mean zero, $E(\xi_k) = 0$; (ii) variance one, $E(\xi_k^2) = 1$; (iii) uncorrelated, $E(\xi_k\xi_j) = 0$ for $k \neq j$. Then, the mean model may be rewritten as follows,

$$E(y|Z) = \sum_{k=1}^{\infty} \beta_k\xi_k, \tag{1}$$

where coefficients $\beta_k =< b, \sqrt{\varsigma_k}\phi_k >, k = 1, \cdots$, which are unknown due to the unknown b. Equation (1) presents a linear projection of scalar outcome y on the space spanned by the standardized principal components (PCs) ξ_k's of functional predictor Z. On these lines of research, Müller and Yao (2008) proposed a class of functional additive models (FAMs) that extends Equation (1) by allowing a nonparametric form of the projection:

$$E(y|Z) = \sum_{k=1}^{\infty} f_k(\xi_k), \tag{2}$$

where f_k is a fully unspecified nonlinear smooth function to be estimated. It is obvious that Müller and Yao's extension given in (2) takes an additive model on individual coefficient (or feature) components ξ_k's. Regularization is often needed for both (1) and (2) in order to deal with these infinite-dimensional unknowns. One of the challenges concerning regularization for (2) lies in the technical treatment in the functional space. Müller and Yao (2008) [21] proposed truncation (or a hard threshold) of the eigenspace to retain only the leading components that explain the majority of the total variation in Z. Zhu, Yao, and Zhang (2014) [15] proposed another regularization for the functions f_k using the powerful COSSO method [22]. One advantage for this kind of regularization method is that sums of higher-order functional principal components are allowed to be potentially included in the fit model, if they make stronger contributions to the functional relationship than the leading functional principal components. This regularization method [15] begins with an additive model $E(y|Z) = \sum_{k=1}^{s} f_k(\xi_k)$, where s represents some initial degrees of truncation to specify the total number of additive components to be considered. Then, COSSO helps simultaneously regularize and select important functional components among the s functions f_k. Although the above discussion is based on a single functional predictor Z in mind, it is appealing to extend such a framework with multiple functional predictors for a broad range of problems.

When multiple functional predictors, *say* $Z^1, \ldots, Z^p$, are considered, it is not clear if the above additive model specification remains suitable to handle the complexity, especially a non-additive relationship (e.g., interactions) may be of interest to understand the association between a scalar outcome and multiple functional predictors. In effect, from both the perspectives of theoretical advances and application needs, relaxing the additive relationship is an important task in functional data analysis. Alternatively, there are some methods (e.g., [16,17]) in the literature that do not use the strategy of decomposing Z into its functional components. In this paper, we adopt the framework of kernel machine regression models to extend the methodologies with non-additive relationships between multiple functional predictors and the scalar outcome.

1.2. Least-Squares Kernel Machine

Liu, Lin, and Ghosh (2007) [23] proposed a semi-parametric regression model $y_i = \mathbf{x}_i^{\top}\boldsymbol{\beta} + h(\mathbf{z}_i) + \epsilon_i$ for subject $i = 1, \ldots, n$, where they used the least-squares kernel machine (LSKM) to analyze multidimensional genetic pathways denoted by a vector $\mathbf{z}_i$. The key feature of this model is the nonlinear relationship between the outcome y_i and a vector of gene expressions $\mathbf{z}_i$, which is characterized by a nonparametric smooth function h. Under the theory of smoothing splines, function h is assumed to lie in a reproducing kernel Hilbert space (RKHS), $\mathcal{H}_{\mathcal{K}}$, generated by a positive-definite kernel function $\mathcal{K}(\cdot, \cdot)$. For the ease of exposition, we suppress the bandwidth for the kernel $\mathcal{K}$ in the following discussion.

Then, both parameter $\boldsymbol{\beta}$ and function h are estimated by maximizing the scaled penalized likelihood function:

$$J(h, \boldsymbol{\beta}) = -\frac{1}{2}\sum_{i=1}^{n}\{y_i - \mathbf{x}_i^\top \boldsymbol{\beta} - h(\mathbf{z}_i)\}^2 - \frac{1}{2}\lambda_1 \|h\|^2_{\mathcal{H}_\mathcal{K}}, \quad (3)$$

where $\lambda_1 > 0$ is the tuning parameter and $\|\cdot\|_{\mathcal{H}_\mathcal{K}}$ is the norm of the RKHS. For a function $h \in L^2(\mathcal{H}_\mathcal{K})$, we have $h(\cdot) = \sum_{i=1}^n \alpha_i \mathcal{K}(\cdot, \mathbf{z}_i)$. Then, $\|h\|^2_{\mathcal{H}_\mathcal{K}} = \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha}$, where $\mathbf{K}$ is an $n \times n$ matrix whose (i, j) entry is $\mathcal{K}(\mathbf{z}_i, \mathbf{z}_j)$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^\top$.

It is known in the literature (e.g., [23,24]) that maximizing $J(h, \boldsymbol{\beta})$ in (3) turns out to be equivalent to solving the normal equations from the following linear mixed-effects model (LMM): $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{h} + \boldsymbol{\epsilon}$, where $\mathbf{h}$ is an $n \times 1$ vector of random effects with distribution $N(\mathbf{0}, \tau\mathbf{K})$ and an n-dimensional vector error term $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, with $\tau = \lambda_1^{-1}\sigma^2 > 0$. One remarkable advantage of solving (3) through the existing numerical procedure of the LMM is most advocated in the literature [25], where we can determine the smoothing parameter λ_1 as part of the estimation of the variance components of the LMM. Therefore, instead of using cross-validation or other information-based tuning methods on λ_1, we can solve simultaneously for all the model parameters in (3), as shown in [23]. Utilizing this numerical strength of the kernel machine regression model, we propose a semi-parametric regression model by incorporating functional principal components of functional predictors (i.e., the $\mathbf{z}_i$) to evaluate a nonlinear relationship of a scalar outcome with multiple functional covariates in a non-additive way. Assuming that function h belongs to an RKHS, we can use existing software packages for solving LMMs to obtain estimates of all model parameters and the smoothing parameter.

1.3. Feature Selection

To deal with high-dimensional functional principal components from functional covariates, we invoked the sparse regularization approach in the kernel machine regression model. Note that for both mean models (1) and (2), one needs to truncate the series from the Karhunen–Loève expansion. Regularization helps reduce from an infinite number of terms to a sum of finite terms. To introduce some notations, here we present a brief review on the group lasso (GL) [26], sparse group lasso (SGL) [27], and non-negative garrote [28]. See also the series of work originated by COSSO [22]. Yuan and Lin (2007) [26] proposed the group lasso, which solves the convex optimization problem: $\min_{\boldsymbol{\beta} \in \mathcal{R}^p} \left\|\mathbf{Y} - \sum_{\ell=1}^L \mathbf{X}^\ell \boldsymbol{\beta}^\ell\right\|_2^2 + \lambda \sum_{\ell=1}^L \left\|\boldsymbol{\beta}^\ell\right\|_2$, where L is the total number of groups of covariates and $\mathbf{X}^\ell$ refers to a subset of covariates associated with group ℓ. Friedman, Hastie, and Tibshirani [27] extended the group lasso to allow within-group sparsity, namely SGL, given as $\min_{\boldsymbol{\beta} \in R^p} \left\|\mathbf{Y} - \sum_{\ell=1}^L \mathbf{X}^\ell \boldsymbol{\beta}^\ell\right\|_2^2 + \lambda(1 - \delta) \sum_{\ell=1}^L \left\|\boldsymbol{\beta}^\ell\right\|_2 + \lambda\delta\|\boldsymbol{\beta}\|_1$, where $\delta \in [0, 1]$. The additional ℓ_1-norm penalty term on $\boldsymbol{\beta}$ encourages individual sparsity, while the first penalty targets sparsity at the group level. It is easy to see that group lasso is a special case of the SGL when $\delta = 0$.

The non-negative garrote proposed by Breiman (1995) [28] is another useful means of variable selection. It invokes a scaled version of least-squares estimation given by: $\arg\min_{\mathbf{d}} \frac{1}{2}\|\mathbf{Y} - \tilde{\mathbf{X}}\mathbf{d}\|_2^2 + \lambda \sum_{j=1}^p d_j$, subject to $d_j \geq 0, j = 1, \ldots, p$. Here, $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_p)$ is an $n \times p$ matrix with columns $\tilde{\mathbf{x}}_j = \mathbf{x}_j \hat{\beta}_j^{OLS}$, with $\hat{\beta}_j^{OLS}$ being the least-squares estimates from $\arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$ with no constraints. Obviously, estimate $\hat{d}_j = 0$ implies that covariate x_j would be excluded from the fit model. Breiman's formulation that turns a variable selection problem into a parameter estimation problem will be applied for the development of feature selection on functional principal components in this paper.

This paper is organized as follows. Section 2 introduces our proposed high-dimensional kernel machine regression. Section 3 outlines a simple step-by-step algorithm that is used to implement the sparse estimation method. Section 4 concerns asymptotic properties for our proposed sparse kernel machine regression. Section 5 provides simulation results to examine the performance of our method, with comparisons with existing methods. Section 6 illustrates the proposed method by an association analysis of the relationship between the BMI and functional accelerometer data. Section 7 includes our conclusions. The Appendix A contains some key technical details, including the proofs of the theoretical results, while Appendix B presents a discussion on the model identifiability issue.

2. Model and Estimation

Consider a regression analysis of a scalar outcome y on p functional covariates, Z^ℓ, $\ell = 1, \dots, p$. Let $\mathbf{z}_i^\ell = (\xi_1^\ell, \dots, \xi_{s_\ell}^\ell)_i^\top$ be the s_ℓ-element vector of functional principal component (FPC) features from the i^{th} observation of the ℓth functional covariate Z^ℓ, and let $\vec{\mathbf{z}}_i = [(\mathbf{z}_i^1)^\top, \dots, (\mathbf{z}_i^p)^\top]^\top$ be the grand vector of all FPC features from all p functional covariates for subject i, $i = 1, \dots, n$. Clearly, the set of FPC features from each functional covariate forms a group, and in total, there are p groups with $s = \sum_{\ell=1}^p s_\ell$ many FPC features and $\vec{\mathbf{z}}_i \in \mathcal{R}^s$. The high dimensionality of FPC features presents the key methodological challenge in the analysis. We consider the following functional kernel machine regression (FKMR) model:

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + h(\vec{\mathbf{z}}_i) + \epsilon_i,\ i = 1, \cdots, n, \tag{4}$$

where $\boldsymbol{\beta} \in \mathcal{R}^q$ is a set of parameters for the effects of q scalar covariates $\mathbf{x} = (x_1, \dots, x_q)^\top$, $h \in \mathcal{H}_\mathcal{K}$ is an s-variate smooth nonparametric function with $\mathcal{H}_\mathcal{K}$ being the functional space generated by a *Mercer kernel* $\mathcal{K}$ and error terms $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. The FKMR model (4) allows for not only nonlinear, but also non-additive relationships with multiple functional covariates Z^ℓ via their FPC features, $\ell = 1, \dots, p$, and a scalar outcome, y. The statistical task is to estimate and select important functional covariates that are related to the outcome of interest through regularizing the FPC features within each functional covariate. To proceed, following Beiman's [28] non-negative garrote method, we here introduce a new s-dimensional scaling vector $\gamma \in \mathcal{R}^s$, $\gamma = (\gamma_1, \dots, \gamma_{s_1}, \dots, \gamma_s)^\top$, by which we can set $\gamma \circ \vec{\mathbf{z}}_i = (\gamma_1 \xi_1^1, \dots, \gamma_{s_1} \xi_{s_1}^1, \dots, \gamma_s \xi_{s_p}^p)_i^\top$ a new vector of weighted FPC features by γ via the Hadamard product (i.e., elementwise product). Note that γ is grouped and denoted by $\gamma = ((\gamma^1)^\top, \dots, (\gamma^p)^\top)^\top$ where γ^ℓ is an s_ℓ-element vector of FPC features $\mathbf{z}^\ell$ of the ℓ^{th} functional covariate Z^ℓ. When the element, say γ_j, is equal to zero, the corresponding FPC feature ξ_j will not be selected in the set of important FPCs, and moreover, functional covariate Z^ℓ is excluded from the FKMR model when the entire vector $(\gamma^\ell)^\top = 0$.

We estimate the unknowns in the FKMR model (4), as well as the scaling parameters γ by minimizing the penalized objective function $J_1(h, \boldsymbol{\beta}, \gamma)$, whose expression is given on the right-hand side of the following Equation (5):

$$\min_{h,\boldsymbol{\beta},\gamma} J_1(h, \boldsymbol{\beta}, \gamma) = \min_{h,\boldsymbol{\beta},\gamma} \frac{1}{2n} \sum_{i=1}^n \{y_i - \mathbf{x}_i^\top \boldsymbol{\beta} - h(\gamma \circ \mathbf{z}_i)\}^2 + \frac{1}{2}\lambda_1 \|h\|_{\mathcal{H}_K}^2 + \lambda_2 \rho(\gamma; \delta), \tag{5}$$

where $\lambda_1 > 0$ and $\lambda_2 > 0$ are two tuning parameters, and penalty $\rho(\gamma; \delta)$ may be specified according to a certain regularization method. For the case of sparse group lasso (SGL), we take $p(\gamma; \delta) = (1 - \delta) \sum_{\ell=1}^p \left\| \gamma^\ell \right\|_2 + \delta \|\gamma\|_1$, $\delta \in [0, 1]$. Typically, δ is predetermined and set to 0.95 or 0.05 depending on the trade-off between group and within-group sparsity, while the factor $(1 - \delta)$ controls the relative group sparsity to individual sparsity of each functional predictor Z^ℓ. Meanwhile, a large tuning parameter for λ_2 would remove a certain group of FPC features from the FKMR model when all elements in the vector γ^ℓ are zero. Given $h \in \mathcal{H}_\mathcal{K}$, an equivalent optimization to the above (5) can be formulated as follows:

$$\begin{aligned}\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\gamma} J_2(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma) = \min_{\boldsymbol{\alpha},\boldsymbol{\beta},\gamma} \frac{1}{2n} \sum_{i=1}^n \left\{ y_i - \mathbf{x}_i^\top \boldsymbol{\beta} - \sum_{k=1}^n \alpha_k \mathcal{K}(\gamma \circ \vec{\mathbf{z}}_i, \gamma \circ \vec{\mathbf{z}}_k) \right\}^2 \\ + \frac{1}{2}\lambda_1 \boldsymbol{\alpha}^\top \mathbf{K}(\gamma; Z) \boldsymbol{\alpha} + \lambda_2 \rho(\gamma; \delta),\end{aligned} \tag{6}$$

where $\mathbf{K}(\gamma; Z)$ is an $n \times n$ matrix whose (i,k)th element is $[\mathbf{K}(\gamma; \mathbf{Z})]_{ik} = \mathcal{K}(\gamma \circ \vec{\mathbf{z}}_i, \gamma \circ \vec{\mathbf{z}}_k)$. Lemma 1 below establishes the equivalency of optimization solutions between (5) and (6), which is crucial in our estimation procedure.

Lemma 1. *A solution $(\hat{h}, \hat{\boldsymbol{\beta}}, \hat{\gamma})$ is a minimizer of* (5) *if and only if $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\gamma})$ is a minimizer of* (6)*, where $\hat{h}(\hat{\gamma} \circ \vec{\mathbf{z}}) = \sum_{k=1}^{n} \hat{\alpha}_k \mathcal{K}(\hat{\gamma} \circ \vec{\mathbf{z}}, \hat{\gamma} \circ \vec{\mathbf{z}}_k)$.*

The proof of Lemma 1 is given in Appendix A.1.

Theorem 1 (Existence of optimizers)**.** *If the kernel $\mathcal{K}(\cdot, \gamma \circ \vec{\mathbf{z}})$ is continuous with respect to $\gamma \in \mathcal{R}^s$, then there exists a global minimizer $(\hat{h}, \hat{\boldsymbol{\beta}}, \hat{\gamma})$ for the optimization problem* (5)*.*

The proof of Theorem 1 is given in Appendix A.3. Note that there may exist multiple optimal minimizers for (5); Theorem 1 ensures only the existence of optimal solutions, but provides no guarantees for uniqueness due to the fact that (5) or (6) is a nonlinear and non-convex optimization problem. It is worth noting that in both (5) and (6), we set the bandwidth for the kernel at a fixed value due to the identifiability issue with respect to the scaling parameters γ. Refer to Appendix B for more detailed discussions on the issue of parameter identifiability.

3. Implementation and Algorithm

We propose an iterative algorithm to implement our proposed estimation procedure in which we require the differentiability of the kernel with respect to the scaling factor γ and some additional assumptions presented below in order to ensure algorithmic convergence. One part of the algorithm solving (5) is carried out under fixed γ, where the resulting minimization problem reduces to the equivalent maximization problem in the least-squares kernel machine (3) with the FPC features, $\vec{\mathbf{z}}_i$, being replaced by $\gamma \circ \vec{\mathbf{z}}_i$. As pointed out in Section 1.2, the step of numerical calculation can be easily executed in the same fashion as the solution from the linear mixed model, including the REML estimation of the smoothing parameter λ_1. The other part of the algorithm is performed under fixed $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and λ_1, where we solve the nonlinear and non-convex optimization problem to update estimates of γ. Lemma 2 below helps us solve for the scaling parameter γ.

Lemma 2. *For fixed $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda_1)$, minimizing* (6) *over γ is equivalent to minimizing over γ the following objective function:*

$$\frac{1}{2n}\left\|\mathbf{F}(\gamma) - \tilde{\mathbf{Y}}\right\|_2^2 + \lambda_2 \rho(\gamma; \delta), \text{ for } \lambda_2 > 0, \tag{7}$$

where $\mathbf{F}(\gamma) = \mathbf{K}(\gamma; Z)\boldsymbol{\alpha}$ and $\tilde{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \frac{n}{2}\lambda_1 \boldsymbol{\alpha}$.

The proof of Lemma 2 is given in Appendix A.2. Linearizing the function $\mathbf{F}(\gamma)$ in (7) leads to an equivalent form:

$$\min_{\gamma} \frac{1}{2n}\left\|\tilde{\mathbf{Y}} - \sum_{\ell=1}^{p} \nabla_\gamma \mathbf{F}^{(\ell)}(\tilde{\gamma})\gamma^{\ell}\right\|_2^2 + \lambda_2 \rho(\gamma; \delta), \tag{8}$$

where $\tilde{\mathbf{Y}} = \left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \frac{n}{2}\lambda_1 \boldsymbol{\alpha}\right) - \mathbf{F}(\tilde{\gamma}) + \nabla_\gamma \mathbf{F}(\tilde{\gamma})\tilde{\gamma}$, with $\nabla_\gamma \mathbf{F}(\tilde{\gamma})$ being the gradient of the function $\mathbf{F}$ with respect to γ evaluated at $\tilde{\gamma}$ for some $\tilde{\gamma}$, and $\nabla_\gamma \mathbf{F}^{(\ell)}(\tilde{\gamma})$ being the columns of $\nabla_\gamma \mathbf{F}(\tilde{\gamma})$ associated with the ℓth group of γ^{ℓ}. This is precisely the form of the standard sparse group regularization problem: $\min_{\beta \in \mathcal{R}^p} \frac{1}{2n}\left\|\mathbf{Y} - \sum_{\ell=1}^{p} \mathbf{X}^{\ell} \boldsymbol{\beta}^{\ell}\right\|_2^2 + \lambda_2 \rho(\gamma; \delta)$. This implies that (8) presents a standard sparse group regularization problem with a specific choice of penalty function $\rho(\gamma; \delta)$.

The convergence of the above iterative search algorithm for updating $\tilde{\gamma}$ for fixed $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda_1)$ can be justified by the proximal Gauss–Newton method [29]. Readers are referred to [30] for details on the proximal Gauss–Newton method. One of the key assumptions of the proximal Gauss–Newton method is the existence of a local minimizer. This condition is satisfied in the above (8). This is because according to Theorem 1, there exists a global minimizer.

Algorithm 1 summarizes these iterative steps, which is showed to satisfy a descent property: $J_2(\boldsymbol{\alpha}^{(r+1)}, \boldsymbol{\beta}^{(r+1)}, \gamma^{(r+1)}) \leq J_2(\boldsymbol{\alpha}^{(r)}, \boldsymbol{\beta}^{(r)}, \gamma^{(r)})$ under the convergence of the proximal Gauss–Newton algorithm for Step 2.2.

Algorithm 1 An iterative algorithm for optimization in FKMR.

```
1.1  Perform FPCA (e.g., the R package fdapace) to extract the functional component features for the $p$ functional predictors, and store them in a grand vector for each individual subject $\vec{\mathbf{z}}_i = [(\mathbf{z}_i^1)^\top, \ldots, (\mathbf{z}_i^p)^\top)]^\top, i = 1, \cdots, n;$
1.2  Initialize $\gamma$ to be a vector of ones. which translates to mapping the original component scores to itself. Set up a grid of possible tuning parameters for $\lambda_1$ and $\lambda_2$, respectively. Set the kernel bandwidth parameter, which may depend on $\lambda_1$. For each pair of $(\lambda_1, \lambda_2)$ from our grid, perform Steps 2.1-2.3 and 3.1 below.
2.1  At the $(r+1)$-th step in the algorithm, first solve the LSKM problem with fixed $(\gamma^{(r)}, \lambda_1)$ (based on a closed-form solution) to update $\boldsymbol{\beta}^{(r+1)}$ and $\boldsymbol{\alpha}^{(r+1)}$.
2.2  Solve the group regularity problem (8) with fixed $\tilde{\gamma} = \gamma^{(r)}$ and fixed $(\boldsymbol{\alpha}^{(r+1)}$, $\boldsymbol{\beta}^{(r+1)}$, $\lambda_1$, $\lambda_2)$ using the $r+1$ updates from the previous iteration. At this step, the proximal Gauss–Newton algorithm produces an update $\gamma^{(r+1)}$ at convergence.
2.3  Repeat Steps 2.1–2.2 until convergence.
3.1  Perform cross-validation over all pairs of $(\lambda_1, \lambda_2)$ to determine the final $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma)$.
```

To speed up Algorithm 1, we propose the following operational schemes that avoid setting up the pairs of (λ_1,λ_2) and performing Step 3.1. Here are a few remarks on the two algorithms. (i) Algorithm 2 depends on good starting values in order to enjoy a fast search. (ii) The main difference between Algorithms 1 and 2 is that λ_2 is fixed in Algorithm 1, while it is changing in Algorithm 2. Some similar algorithms with changing tuning parameters have been proposed in the literature, such as the single index model [31]. (iii) There is no guarantee that both algorithms converge to a global minimizer, and the proximal Gauss–Newton method used in the implementation can only find stationary points. Numerical solvers for the optimization problem in (5) or in (6) indeed remain an open problem in the field of nonlinear and nonconvex optimization.

Algorithm 2 A fast operational scheme of Algorithm 1.

```
1.   Step 2.1 of Algorithm 1 is performed by running the linear mixed model with our initial fixed $\gamma$ from Step 1.2 of Algorithm 1 to obtained updated values of $\lambda_1$, $\boldsymbol{\beta}$, and $\boldsymbol{\alpha}$.
2.   Step 2.2 is performed with solving the group regularity problem (8) through the Gauss–Newton algorithm using cross-validation-based tuning (e.g., R package oem).
3.   Rerun Step 2.1 using the updated $\gamma$ from Step 2.2 to obtain the estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$.
```

4. Theoretical Guarantees

Our theoretical analysis focuses on the finite-sample L_2 error bounds for the estimators $(\hat{h}, \hat{\gamma})$ obtained by (5) or (6). Consequently, we are able to establish the estimation consistency. For simplicity, we set $\boldsymbol{\beta} = \mathbf{0}$ and consider a general setting of random vectors $\mathbf{z}_1, \ldots, \mathbf{z}_n$ so that the FPC features $\vec{\mathbf{z}}_1, \ldots, \vec{\mathbf{z}}_n$ correspond to a special case. Along similar lines as those of [15,32], the estimation consistency is proven in the case of the SGL penalty function. We define a map Γ with an s-element vector $\gamma \in \mathcal{R}^s$, which gives rise to a collection of all scaling map functions: $\mathcal{A} = \{\Gamma : \mathcal{R}^s \mapsto \mathcal{R}^s \mid \Gamma(\mathbf{z}) = \gamma \circ \mathbf{z}, \mathbf{z} \in \mathcal{R}^s \text{ and } \gamma \in \mathcal{R}^s\}$. Since Γ is a linear (and bounded) operator, $\mathcal{A}$ is a real vector space where $(c_1\Gamma_1 + c_2\Gamma_2)(\mathbf{z}) = c_1\Gamma_1(\mathbf{z}) + c_2\Gamma_2(\mathbf{z})$ with any $c_1, c_2 \in \mathcal{R}$ and $\Gamma_1, \Gamma_2 \in \mathcal{A}$. To perform a group regularization estimation, we define an SGL penalty by a norm on $\mathcal{A}$ for a fixed $\delta \in [0,1]$ as follows:

$$\|\Gamma\|_{SGL} = \delta \sum_{\ell=1}^{p} \left\|\gamma^{\ell}\right\|_2 + (1-\delta)\|\gamma\|_1. \tag{9}$$

Consequently, the SGL regularization estimation requires the following constrained optimization:

$$\min_{\Gamma\in\mathcal{A},\, h\in\mathcal{H}_{\mathcal{K}}} J_3(\Gamma, h) = \min_{\Gamma\in\mathcal{A},\, h\in\mathcal{H}_{\mathcal{K}}} \|\mathbf{Y} - h\circ\Gamma\|_n^2 + \lambda_1\|h\|_{H_K}^2 + \lambda_2\|\Gamma\|_{SGL}, \tag{10}$$

where $\|\mathbf{Y} - h\circ\Gamma\|_n^2 = \frac{1}{n}\sum_{i=1}^{n}\{y_i - (h\circ\Gamma)(\mathbf{z}_i)\}^2$. Lemma 3 below provides the essential finite-sample inequalities that lead to the estimation consistency.

Lemma 3 (Basic inequality)**.** *Let $\hat{h}\circ\hat{\Gamma}$ be the minimizer of* (10)*. Let $h_0\circ\Gamma_0$ be the true function. Then, we have:*

$$J_3(\hat{\Gamma}, \hat{h}) \le 2(\boldsymbol{\epsilon}, \hat{h}\circ\hat{\Gamma} - h_0\circ\Gamma_0)_n + \lambda_1\|h_0\|_{\mathcal{H}_{\mathcal{K}}}^2 + \lambda_2\|\Gamma_0\|_{SGL}, \tag{11}$$

where $2(\boldsymbol{\epsilon}, \hat{h}\circ\hat{\Gamma} - h_0\circ\Gamma_0)_n = \frac{2}{n}\sum_{i=1}^{n}\epsilon_i\Big\{(\hat{h}\circ\hat{\Gamma})(\mathbf{z}_i) - (h_0\circ\Gamma_0)(\mathbf{z}_i)\Big\}$.

We need the following notation before presenting our theoretical guarantees. Let $\mathcal{N}(\delta, M, P_n)$ denote the minimal δ covering number of the function set $\mathcal{M}$ under the empirical metric P_n based on the random vectors $\mathbf{z}_1, \cdots, \mathbf{z}_n$. Let $N = \mathcal{N}(\delta, M, P_n)$ be a shorthand notation. This means that there exist functions $m_1, \cdots, m_N$ (not necessarily in the set $\mathcal{M}$) such that for every function $m \in \mathcal{M}$, there exists a $j \in \{1, \cdots, N\}$ such that $\|m - m_j\|_{P_n} \le \delta$, with $\|m - m_j\|_{P_n} := \sqrt{\frac{1}{n}\sum_{i=1}^{n}\{m(\mathbf{z}_i) - m_j(\mathbf{z}_i)\}^2}$. Define the δ-entropy of $\mathcal{M}$ for the empirical metric, P_n, as $H(\delta, \mathcal{M}, P_n) := log(\mathcal{N}(\delta, \mathcal{M}, P_n))$. Consider a functional space of the form:

$$\mathcal{B} = \left\{ b := b(h, \Gamma) = \frac{h\circ\Gamma - h_0\circ\Gamma_0}{\|h\|_{\mathcal{H}_{\mathcal{K}}}^2 + \|h_0\|_{\mathcal{H}_{\mathcal{K}}}^2 + \|\Gamma\|_{SGL}^2 + \|\Gamma_0\|_{SGL}^2} | h \in \mathcal{H}_{\mathcal{K}}, \Gamma \in \mathcal{A} \right\}.$$

We postulate the following assumptions.

Assumption 1. *The error term $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$ is uniformly sub-Gaussian; that is, for constants C_1 and C_2,*

$$\max_{n\ge1}\max_{i=1,\cdots,n} C_1^2\left[E\left\{exp\left(\frac{{\epsilon_i}^2}{C_1^2}\right)\right\} - 1\right] \le C_2.$$

Clearly, the moment condition is bounded below from zero.

Assumption 2. $\|\Gamma_0\|_{SGL}^2 + \|h_0\|_{\mathcal{H}_{\mathcal{K}}}^2 > 0$*, and the entropy of space $\mathcal{B}$ with respect to the empirical metric P_n is bounded as follows:*

$$H(\delta, \mathcal{B}, P_n) \le C_3\delta^{-2\psi},$$

where C_3 is some constant and $\psi \in (0,1)$.

Assumption 3. $\sup_{b\in\mathcal{B}}\|b\|_{P_n} \le C_4$ *for some constant C_4.*

Theorem 2. *(Consistency) Under Assumptions 1-3 above, if tuning parameters λ_1 and λ_2 satisfy*

$$\lambda_2^{-1} = n^{\frac{1}{1+\psi}}\left(\|h_0\|_{\mathcal{H}_{\mathcal{K}}}^2 + \|\Gamma_0\|_{SGL}\right)^{\frac{1-\psi}{1+\psi}}, \text{ and } \lambda_1 = O_p(1)\lambda_2,$$

then we have

$$\left\|\hat{h}\circ\hat{\Gamma}-h_0\circ\Gamma_0\right\|_n=O_p(n^{-\frac{1}{2+2\psi}})\left(\|h\|^2_{\mathcal{H}_{\mathcal{K}}}+\|\Gamma\|_{SGL}\right)^{\frac{\psi}{1+\psi}},\ \textit{and} \tag{12}$$

$$\left\|\hat{h}\right\|^2_{\mathcal{H}_{\mathcal{K}}}+\|\hat{\Gamma}\|_{SGL}=O_p(1)\left(\|h_0\|^2_{\mathcal{H}_{\mathcal{K}}}+\|\Gamma_0\|_{SGL}\right). \tag{13}$$

Theorem 2 implies estimation consistency under the right rates for the two tuning parameters λ_1 and λ_2. Due to the potential identifiability issues explained in detail in Appendix B, although the estimator $(\hat{h},\hat{\Gamma})$ may not be unique, the sum of $\hat{h}$ and $\hat{\Gamma}$ is not too far away from the sum of the true h_0 and Γ_0.

Corollary 1. *If the RKHS, $\mathcal{H}_{\mathcal{K}}$, contains differentiable functions $\nabla h(\mathbf{z})$ whose norm $\|\nabla h(\mathbf{z})\|_{\mathcal{H}_{\mathcal{K}}}$ is uniformly bounded for all functions $h\in\mathcal{H}_{\mathcal{K}}$ and $\mathbf{z}\in R^s$, then Assumption 2 holds when Theorem 2 is replaced by $H(\delta,\mathcal{H}_{\mathcal{K}},P_n)\leq C_1\delta^{-2\psi}$, for all $\delta\geq 0$.*

The proofs of Theorem 2 and Corollary 1 are given in Appendices A.4 and A.5, respectively. Often, when we are only interested in a subset of functions in the RKHS (e.g., functions with norm less than one), we can substitute the full space $\mathcal{H}_{\mathcal{K}}$ in Corollary 1 with the subspace of interest. Refer to [15] or [32], where both considered an RKHS (i.e., Sobolev space) with functions of norm less than or equal to one.

5. Simulation Experiments

We performed extensive simulation to investigate the performance of our proposed procedure, including the performance of SGL variable selection and its overall accuracy. Due to the limitations of space, we include results from two simulation experiments in this section, and more results may be found in the first author's Ph.D. dissertation [30].

5.1. Setup

In the evaluation of the performance accuracy, following [15], we used both quasi-R^2 and adjusted quasi-R^2 defined as follows:

$$R^2_Q:=1-\frac{\sum_{i=1}^n(y_i-\hat{y}_i)^2}{\sum_{i=1}^n(y_i-\bar{y}_i)^2},\ \text{and}\ R^2_{AQ}:=1-\left(1-R^2_Q\right)\left(\frac{n-1}{n-(k+1)}\right).$$

The latter is known to be appealing for the comparison of the estimation sparsity. There is another performance metric of interest in addition to model accuracy. Performance in variable selection is summarized in terms of the stability measured by sensitivity and specificity for both functional and variable selections under these simulation experiments. Our algorithm uses existing R packages, including `emmreml`, `kspm`, and `oem`.

Specifically, we designed the following two simulation settings.

Scenario 1: A single functional predictor with sparsity in the FPC features.
Scenario 2: Multiple functional predictors with sparsity in the functional predictors and with sparsity in the FPC features of important functional predictors.

Each of these two scenarios would be handled using certain suitable penalty functions to address the designed sparsity; for example, in Scenario 2 we used a two-level variable selection penalty (e.g., SGL) to deal with two types of sparsity in the true model. In all analyses, we used the Gaussian kernel $\mathcal{K}(u,v)=\exp(-\frac{1}{p}\|u-v\|^2)$ in our estimation, where p was set as the number of features, which is equivalent to dividing the γ vector by $\sqrt{p}$. This scaling parameter may be either estimated or set to the number of features to overcome the identifiability issue according to [33], where theoretical justification was given for the use of the number of features for the bandwidth parameter in the case of the Gaussian kernel.

According to [23], due to the difficulty of the graphical display for the estimated s-dimensional function $h(\cdot)$ of $\mathbf{z}$, we summarized the goodness-of-fit by regressing the true h on the estimated $\hat{h}$, with both being evaluated at the design points. From this concordance regression analysis, we may measure the goodness-of-fit on $\hat{h}$ through the average intercepts, slopes, and R-squared (also known as the coefficient of determination) obtained over the number of replications. Clearly, a high-quality fit is reflected by (i) the intercept being close to zero, (ii) the slope being close to one, and (iii) the R-squared being close to one. Moreover, we graphically display the estimated function $\hat{h}$ by setting all variables equal to 0.5 except the one of interest over a grid of 100 equally spaced points on the interval $[0,1]$. Such visualization of the functional estimation at each margin further facilitates the evaluation of the proposed algorithm in addition to the results obtained from the concordance regression analyses.

In all scenarios, we generated 1000 IID functional paths, of which 750 paths were assigned to the training set and 250 paths were assigned to the test set for an external performance evaluation. It is the test set that we used to display the performance accuracy. We used a one-dimensional covariate x_i to show the flexibility of our model in a semi-parametric setting, with independent copies of $x_i \sim N(0,1)$. We chose the true coefficients in the kernel machine model similar to those given in [23].

5.2. Simulation in Scenario 1

In this simple scenario with a single functional predictor, we simulated data from a model with sparsity in its FPC features. To do so, we generated a single functional predictor based on the first 15 eigenbasis of the Fourier basis functions over the interval $[0,1]$: $Z(t) = \sum_{j=1}^{15} \sqrt{\varsigma_j}\xi_j\phi_j(t)$. That is, a functional predictor was created as a linear combination of the 15 basis functions, where $\phi_j(\cdot)$ is the j^{th} Fourier basis function, ς_j is the jth eigenvalue of Z, and ξ_j is the jth FPC feature that is simulated from a normal distribution detailed as follows.

There were 100 sampled points that were first equally spaced in the interval $[0,1]$ and then varied with certain small deviations drawn from $\nu \sim N(0, 0.001)$. Set $\varsigma_j = 45 \times 0.64^j$ and $\xi_j \sim N(0,1)$ independently over $j = 1, \ldots, 15$. As was done in [17], instead of directly using ξ_j, we used $\zeta_j = \Phi(\xi_j)$, where Φ is the CDF of the standard normal. This resulted in $\vec{\mathbf{z}} = (\zeta_1, \ldots, \zeta_{15})^\top$. We chose the second, ζ_2, and ninth, ζ_9, features as important features in the following true nonlinear non-additive model:

$$y_i = 2x_i + 20\,\cos(2\pi\zeta_{i2}) - 10\,\sin(2\pi\zeta_{i9}) + \zeta_{i2}\zeta_{i9} + \epsilon_i,$$

with $\epsilon_i \overset{iid}{\sim} N(0,1)$. FPCA was performed by the R package `PACE` [34], producing the estimated FPC scores, $\hat{\xi}_j$, as well as the estimated eigenvalues, $\hat{\varsigma}_j$, which in turn enabled us to compute $\hat{\zeta}_j$, $j = 1, \ldots, 15$.

We applied both LASSO and MCP penalty functions in our implementation, termed as $FKMR_{Lasso}$ and $FKMR_{MCP}$, respectively. We compared the results of our method with the standard linear approach with both LASSO and MCP under the assumption of linear functional relationships, as well as the COSSO method for functional additive regression [15] using the R package `COSSO` [15,34]. Since the COSSO package is built for nonparametric regression (and not partial linear models), we adopted the backfitting strategy and regressed the residuals with our estimated effect of x_i removed.

In addition, we compared our method with an oracle FKMR estimator, called $FKMR^{oracle}$, that assumed the full knowledge of the true ζ_j containing two true nonzero signals, ζ_2 and ζ_9. We also considered two oracle versions of our proposed algorithm, $FKMR^{oracle}_{Lasso}$ and $FKMR^{oracle}_{MCP}$, both of which used the knowledge of true ζ_j in order to evaluate the performance of the FPCA procedure. This evaluation is important as our proposed procedure can be in principle used in simpler cases that do not involve functional covariates. Note that once we used FPCA to obtain $\hat{\zeta}_j$ features, our algorithm essentially works in a standard regression setting with the sparsity of covariates. Thus, our proposed procedure can be in principle used in simpler cases with scalar covariates. In Scenario 1, due to the highly nonlinear relationships between the FPC features and the outcome, as expected, the naive linear model performed poorly in terms of both model selection and model consistency. The detailed simulation results for Scenario 1 can be found in the first author's Ph.D. dissertation [30]. In brief, our proposed method worked well in all aspects. In this setting, COSSO also worked well in terms of model fit, but it tended to select noisy features more frequently than our proposed method, leading to more false positives.

5.3. Simulation in Scenario 2

Now, we generated four functional predictors of the form: $Z^\ell(t) = \sum_{j=1}^{9} \sqrt{\varsigma_j^\ell}\xi_j^\ell\phi_j^\ell(t)$, $\ell = 1, \ldots, 4$, where ϕ_j^ℓ, ς_j^ℓ, and ξ_j^ℓ were set in the same way as those given in Scenario 1. It follows that $\vec{\mathbf{z}} = (\zeta_1^1, \ldots, \zeta_9^1, \ldots, \zeta_1^4, \ldots, \zeta_9^4)^\top$, where ζ_j^ℓ is the jth Φ-transformed feature for the ℓth functional covariate. Sparsity was specified as follows: the first and second functional covariates, Z^1 and Z^2, were chosen as important signals in which these transformed FPC features, $\{\zeta_1^1, \zeta_3^1, \zeta_4^1, \zeta_2^2, \zeta_7^2\}$, are five important features (three features from the Z^1 and two features from Z^2) that are related to the outcome:

$$
\begin{aligned}
y_i = {} & 2x_i + \zeta_{i1}^1 + \zeta_{i3}^1 + \zeta_{i4}^1 + \zeta_{i2}^2 + \zeta_{i7}^2 + 10\cos(2\pi\zeta_{i1}^1) - 10\left(\zeta_{i2}^2\right)^2 + 10\left(\zeta_{i7}^2\right)^2 - 10\left(\zeta_{i3}^1\right)^2 \\
& + 10\exp(-\zeta_{i3}^1)\zeta_{i4}^1 - 8\sin(2\pi\zeta_{i7}^2)\cos(2\pi\zeta_{i3}^1) + 20\zeta_{i1}^1\zeta_{i7}^2 + \epsilon_i,\ i = 1, \ldots, n,
\end{aligned}
$$

where $\epsilon_i \overset{iid}{\sim} N(0,1)$. This model specifies both group sparsity (two of the four functional predictors) and within-group sparsity (three of the nine FPC features in Z^1 and two of the nine FPC features in Z^2). In addition, we specified non-additive relationships in the true model across multiple functional covariates.

We fit the data using the proposed methods, including $FKMR_{GMCP}^{oracle}$, $FKMR_{Lasso}$, $FKMR_{GLasso}$, $FKMR_{SGL}$, $FKMR_{MCP}$, and $FKMR_{GMCP}$, and the results based on 100 replicates are summarized in Table 1. For comparison, we also fit the simulated data by existing methods, including the linear model (denoted by LM + penalty), COSSO functional additive regression, and the oracle method using the knowledge of true important features in the analysis, as done in the above simulation of Scenario 1. From Table 1 regarding the goodness-of-fit, we see that all of our FKMR estimators outperformed the standard linear estimators in terms of R^2_{AQ} among all of our penalty functions, and they outperformed COSSO for penalties that accounted for group sparsity. In the concordance regression analysis, we see that all intercepts were close to zero, all slopes close to one, and all R^2 close to one, indicating a high goodness-of-fit for functional estimation. COSSO tended to perform on par for penalties that did not account for group sparsity (LASSO and MCP). It is evident that using a group sparsity penalty function (SGL, GLasso, and GMCP) clearly outperformed the methods that did not regularize the grouping of covariates (Lasso and MCP). In addition, our FKMR estimators (except $FKMR_{Lasso}$) performed as well as the oracle estimator $FKMR_{GMCP}^{oracle}$ both in terms of R^2_{AQ} and in terms of our estimate of functional h. The results also indicated that there were little differences between using a concave (MCP or GMCP) penalty function or using a convex (GLasso or SGL) penalty function.

As regards the group sparsity, Table 2 indicates that the all methods had a high sensitivity of detecting functional signals, while the proposed FKMR methods had better specificity than both sparse linear models and COSSO. Concerning the within-group sparsity, it is interesting to note that a bigger difference was seen in terms of what type of penalty function was being used in feature selection. As shown in Tables 3 and 4, using a general penalty (e.g., Lasso and MCP) that does not take the grouping structure into account tended to under-select important features within a group. COSSO tended to perform well within group sparsity. Moreover, Figure 2 shows that the FKMR method estimated the five signal functions (Z^1 and Z^2) well.

Table 1. Goodness-of-fit and the concordance regression for Scenario 2.

Model	R^2_{AQ}	β	**Reg of h on $\hat{h}$**		
			Intercept	**Slope**	R^2
$FKMR_{Lasso}$	0.830	2.00	−0.062	1.01	0.848
$FKMR_{GLasso}$	0.937	1.99	−0.055	1.01	0.972
$FKMR_{SGL}$	0.928	2.00	−0.051	1.01	0.955
$FKMR_{MCP}$	0.835	2.01	−0.062	1.01	0.856
$FKMR_{GMCP}$	0.935	1.99	−0.056	1.01	0.970
$FKMR^{oracle}_{GMCP}$	0.911	1.99	−0.049	1.01	0.937
COSSO	0.832	–	–	–	–
LM + Lasso	0.453	–	–	–	–
LM + GLasso	0.324	–	–	–	–
LM + SGL	0.450	–	–	–	–
LM + MCP	0.513	–	–	–	–
LM + GMCP	0.307	–	–	–	–

Table 2. Sensitivity and specificity of functional selection for Scenario 2.

Model	**Selection Frequency**			
	$\hat{Z}^1$	$\hat{Z}^2$	$\hat{Z}^3$	$\hat{Z}^4$
$FKMR_{Lasso}$	100	100	0	0
$FKMR_{GLasso}$	100	100	4	4
$FKMR_{SGL}$	100	100	0	0
$FKMR_{MCP}$	100	100	0	0
$FKMR_{GMCP}$	100	100	3	4
COSSO	100	100	5	6
LM + Lasso	100	100	19	21
LM + GLasso	94	99	7	8
LM + SGL	100	100	19	18
LM + MCP	100	100	20	19
LM + GMCP	93	99	7	8

Table 3. FPC feature selection for signal functional Z^1 in Scenario 2.

Model	**Selection Frequency**								
	$\hat{\zeta}^1_1$	$\hat{\zeta}^1_2$	$\hat{\zeta}^1_3$	$\hat{\zeta}^1_4$	$\hat{\zeta}^1_5$	$\hat{\zeta}^1_6$	$\hat{\zeta}^1_7$	$\hat{\zeta}^1_8$	$\hat{\zeta}^1_9$
$FKMR_{Lasso}$	100	1	97	0	0	0	0	0	0
$FKMR_{GLasso}$	100	100	100	100	100	100	100	100	100
$FKMR_{SGL}$	100	21	100	71	26	20	17	16	15
$FKMR_{MCP}$	100	1	99	1	0	0	0	0	0
$FKMR_{GMCP}$	100	100	100	100	100	100	100	100	100
COSSO	100	2	100	93	1	0	0	1	0
LM + Lasso	100	10	100	100	10	8	7	10	5
LM + GLasso	94	94	94	94	94	94	94	94	94
LM + SGL	100	12	100	100	10	8	8	11	5
LM + MCP	100	10	100	100	9	8	9	7	5
LM + GMCP	93	93	93	93	93	93	93	93	93

Table 4. FPC feature selection for signal functional Z^2 in Scenario 2.

Model	**Selection Frequency**								
	$\hat{\zeta}^2_1$	$\hat{\zeta}^2_2$	$\hat{\zeta}^2_3$	$\hat{\zeta}^2_4$	$\hat{\zeta}^2_5$	$\hat{\zeta}^2_6$	$\hat{\zeta}^2_7$	$\hat{\zeta}^2_8$	$\hat{\zeta}^2_9$
$FKMR_{Lasso}$	0	3	0	0	0	0	100	0	0
$FKMR_{GLasso}$	100	100	100	100	100	100	100	100	100
$FKMR_{SGL}$	16	100	14	7	16	23	100	15	7
$FKMR_{MCP}$	0	11	0	0	0	1	100	0	0
$FKMR_{GMCP}$	100	100	100	100	100	100	100	100	100
COSSO	8	97	5	5	5	15	100	3	3
LM + Lasso	17	100	14	7	16	23	100	15	6
LM + GLasso	99	99	99	99	99	99	99	99	99
LM + SGL	17	100	14	7	16	23	100	15	7
LM + MCP	17	100	13	6	16	23	100	15	8
LM + GMCP	99	99	99	99	99	99	99	99	99

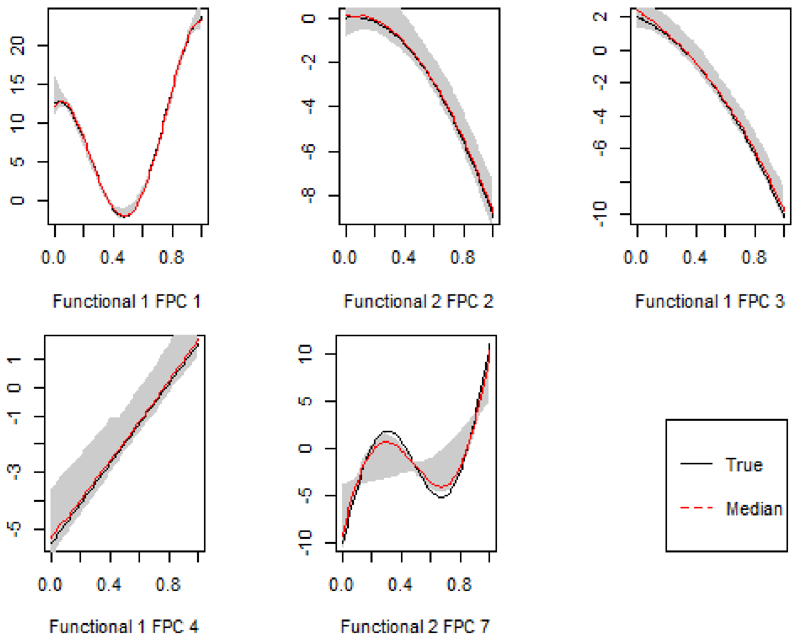

Figure 2. Five marginal estimates of important feature functions with 95% shaded confidence bands evaluated at 100 grid points while holding all other components equal to 0.5 in Scenario 2.

6. Data Example

To show the usefulness of our proposed methodology, we analyzed data of 550 children recruited by the ELEMENTS study [35], who had consent to wear an actigraph (ActiGraph GT3X+; ActiGraph LLC. Pensacola, FL, USA). This wearable was to be placed on their non-dominant wrist for five to seven days with no interruption. The actigraph measured tri-axis accelerometer data sampled at 30 Hz, which captured three different directions of a person's movement. The BMI was the outcome of interest as it is biomarker of obesity. Sex and age were confounding factors used in the analysis. Due to some missing data, our analysis only included children who wore the device properly for 85% or more over the study period, which resulted in 395 participants, consisting of 189 males and 206 females. Other studies such as [36] have excluded days of accelerometer data with more than five percent missing. The mean ± SD BMI of the study cohort was 21.5 ± 4.1. The mean age of the study participants was 14.3 ± 2.1 y. A more detailed description of the dataset used for this paper can be found in [37]. Our primary interest was to see if the BMI is associated with physical activity in the presence of other covariates, specifically sex and age. We preprocessed the activity counts over the 7 d of wear by taking the median in the 1 min epoch over the entire 7 d of wear. For example, since all the participants started wearing the device at 3 p.m., the first data point for each individual was a median of 7 ACs (each for one day) for the 1 min epoch of 3:00–3:01 p.m. This procedure that takes the medians across the minutes from different days has been considered in other applications such as [36]. See Figure 3 as an example of the resulting time series of medians derived from the AC data displayed in Figure 1.

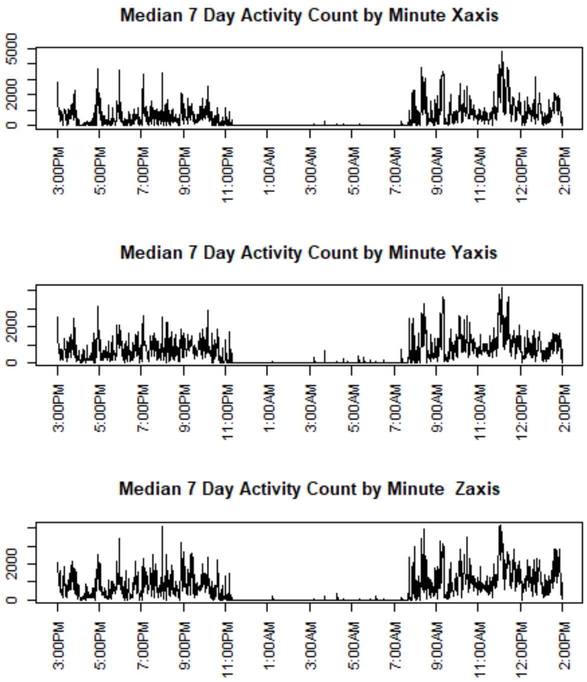

Figure 3. The 24 h minute-by-minute medians of 7 d ACs for one subject.

We applied the following five models, labeled as M0–M4 for convenience, to analyze the data with the 24 h median ACs as functional predictors. Let ξ_{ij}^k be the ith person's kth FPC score for functional predictor j.

M0: Linear model (LM) with only the fixed features: $BMI_i \sim \beta_0 + \beta_1 Age_i + \beta_2 Sex_i$;

M1: Linear model with SGL penalty (LM+SGL) using the FPCA features: $BMI_i \sim \beta_0 + \beta_1 Age_i + \beta_2 Sex_i + \sum_{j=1}^{3} \sum_{k=1}^{s_k} \beta_j^k \xi_{ij}^k$;

M2: LSKM using the FPCA features: $BMI_i \sim \beta_0 + \beta_1 Age_i + \beta_2 Sex_i + h(\mathbf{z}_i)$;

M3: FKMR model with SGL penalty ($FKMR_{SGL}$) using the FPCA features: $BMI_i \sim \beta_0 + \beta_1 Age_i + \beta_2 Sex + h(\gamma \circ \mathbf{z}_i)$;

M4: COSSO using the FPCA features: $res(BMI_i)|\mathbf{z}_i \sim \sum_{j=1}^{3} \sum_{k=1}^{s_k} f_{ij}(\xi_{ij}^k)$. In order for a direct application of the COSSO R package, we used residuals $res(BMI_i) = BMI_i - \hat{\beta}_0 + \hat{\beta}_1 Age_i + \hat{\beta}_2 Sex_i$ in the COSSO model fit, with $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$ being the estimates of the coefficients from Model M0.

The BMI and age were mean centered and scaled to be a standard deviation of one, so β_0 was absent in the models. Here are some key findings from the data analyses. First, in terms of the goodness-of-fit, Table 5 suggests that M3, i.e., our proposed model FKMR with the SGL penalty, gave the best performance, where the adjusted R^2 of M3 was nearly twice as big as all the other four models. Second, it is interesting to note that both the COSSO and the $FKMR_{SGL}$ did not select the FPC scores associated with the Z-axis. Third, as shown in Table 6, all of the FPC components chosen by COSSO were also chosen by the $FKMR_{SGL}$. It is worth noting that the linear model together with the SGL penalty selected the highest number of FPC components, yet performed the worst in terms of the model fit.

Table 5. Goodness-of-fit for the five models used in the data analysis.

Model	Adjusted R^2
M0: LM	0.07
M1: LM + SGL	0.13
M2 : LSKM	0.18
M3: $FKMR_{SGL}$	0.30
M4: COSSO	0.14

Table 6. Axis-specific FPC feature selection.

Model	*X*-Axis						*Y*-Axis					*Z*-Axis			
	$\hat{\zeta}_1^1$	$\hat{\zeta}_2^1$	$\hat{\zeta}_3^1$	$\hat{\zeta}_4^1$	$\hat{\zeta}_5^1$	$\hat{\zeta}_6^1$	$\hat{\zeta}_1^2$	$\hat{\zeta}_2^2$	$\hat{\zeta}_3^2$	$\hat{\zeta}_4^2$	$\hat{\zeta}_5^2$	$\hat{\zeta}_1^3$	$\hat{\zeta}_2^3$	$\hat{\zeta}_3^3$	$\hat{\zeta}_4^3$
$FKMR_{SGL}$		✓	✓	✓		✓	✓		✓		✓				
COSSO				✓			✓		✓						
LM + SGL	✓			✓	✓	✓	✓	✓	✓						✓

7. Conclusions

In this paper, we proposed a method to model the nonlinear relationship between multiple functional predictors and a scalar outcome in the presence of other scalar confounders. We used the FPCA to decompose the functional predictors for feature extraction and used the LSKM framework to model the functional relationship between the outcome and principal components. We developed a simultaneous procedure to select important functional predictors and important features within selected functionals. We proposed a computationally efficient algorithm to implement our regularization method, which was easily programmed in R with the utility of multiple existing R packages. It should be noted that although we focused on functional regression in this paper, the method proposed can be applied to non-functional predictors. In effect, by using functional principal components, we essentially bypassed the infinite-dimensional problem and worked effectively in a non-functional framework with the FPC features. Through simulation and using data from the ELEMENT dataset, we demonstrated how the FKMR estimator outperformed existing methods in terms of both variable selection and model fit. It should be noted that the existing COSSO method did perform well in terms of variable selection, as shown in Section 5.

A technical issue pertains to identifiability limitations with regard to the bandwidth parameter and to the RKHS estimator. To overcome this, we suggested fixing the bandwidth parameter; see the detailed discussion in Section 3. We established key theoretical guarantees for our proposed estimator. In the case where there are multiple proposed estimators (and thus the identifiability issues arise), the established theoretical properties in Section 4 apply to any of those estimators.

Variable section on functional predictors presents many technical challenges, and there are many methodological problems that remain unsolved. This paper demonstrated a possible framework to regularize estimation with a bi-level sparsity of functional group sparsity and within-group sparsity. In the LSKM paper [23], it was briefly mentioned that if the relationship between the scalar outcome and p genetic pathways is additive, we can tweak the model as $y_i = \boldsymbol{x}_i^\top \beta + h_1(z_i^1) + \cdots + h_p(z_i^p) + \epsilon_i$ where each h_j belongs to its own RKHS. It is easy to extend our method and algorithms to handle this case. For future research, an extension on longitudinal outcomes may be considered via a mixed-effects model $y_{ij} = \boldsymbol{x}_i^\top \beta + h(z_{ij}) + \boldsymbol{u}_{ij}^\top v_i + \epsilon_{ij}$ where $\boldsymbol{u}_{ij}^\top v_i$ are the random effects. Other useful extensions to the proposed paradigm would be on the lines of generalized linear models and Cox regression models.

Author Contributions: Conceptualization, P.X.S. and J.N.; Formal analysis, J.N.; Methodology, J.N. and P.X.S.; Supervision, P.X.S.; Writing—original draft, J.N.; Writing—review & editing, P.X.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by NSF DMS#2113564.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The used data of physical activity counts, BMI and demographic variables (sex and age) are available upon request through a formal data request procedure outlined by the ELEMENT Cohort Study. Contact the corresponding author of this paper for the detail.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Technical Assumptions and Proofs

Appendix A.1. Proof of Lemma 1

It suffices to show that for any $J_1(h, \boldsymbol{\beta}, \gamma)$ in (5) we can always find $\boldsymbol{\alpha} \in \mathcal{R}^n$ such that $J_1(\tilde{h} = \sum_{i=1}^n \alpha_i \mathcal{K}(\cdot, \gamma \circ \vec{\mathbf{z}}_i), \gamma, \boldsymbol{\beta}) \leq J_1(h, \boldsymbol{\beta}, \gamma)$ where $\tilde{h}$ is the projection of h onto the linearly spanned space given by $span\{\mathcal{K}(\cdot, \gamma \circ \vec{\mathbf{z}}_i), \cdots, \mathcal{K}(\cdot, \gamma \circ \vec{\mathbf{z}}_n)\}$. For any h we can write $h = h^\perp + \tilde{h}$ where $h^\perp \in span\{\mathcal{K}(\cdot, \gamma \circ \vec{\mathbf{z}}_1), \cdots, \mathcal{K}(\cdot, \gamma \circ \vec{\mathbf{z}}_n)\}^\perp$. Since $\mathcal{H}_k$ is a reproducing kernel Hilbert space we can rewrite (5) as follows:

$$J_1(h, \gamma, \boldsymbol{\beta}) = \frac{1}{2n} \sum_{i=1}^n \{y_i - \mathbf{x}_i^\top \boldsymbol{\beta} - < h, \mathcal{K}(\cdot, \gamma \circ \vec{\mathbf{z}}_i) >\}^2 + \frac{1}{2}\lambda_1 \|h\|^2_{\mathcal{H}_k} + \lambda_2 \rho(\gamma; \delta).$$

Since $< h^\perp, \mathcal{K}(\cdot, \gamma \circ \vec{\mathbf{z}}_i) >= 0$ for every i, we obtain

$$\begin{aligned}
J_1(h, \gamma, \boldsymbol{\beta}) &= \frac{1}{2n} \sum_{i=1}^n \left\{ y_i - \mathbf{x}_i^\top \boldsymbol{\beta} - \sum_{k=1}^n \alpha_k \mathcal{K}(\gamma \circ \vec{\mathbf{z}}_i, \gamma \circ \vec{\mathbf{z}}_k)) \right\}^2 + \frac{1}{2}\lambda_1 \left\| h^\perp + \tilde{h} \right\|^2_{\mathcal{H}_k} + \lambda_2 \rho(\gamma; \delta) \\
&\geq \frac{1}{2n} \sum_{i=1}^n \left\{ y_i - \mathbf{x}_i^\top \boldsymbol{\beta} - \sum_{k=1}^n \alpha_k \mathcal{K}(\gamma \circ \vec{\mathbf{z}}_i, \gamma \circ \vec{\mathbf{z}}_k)) \right\}^2 + \frac{1}{2}\lambda_1 \|\tilde{h}\|^2_{\mathcal{H}_k} + \lambda_2 \rho(\gamma; \delta) \\
&= J_1(\tilde{h}, \gamma, \boldsymbol{\beta}).
\end{aligned}$$

Appendix A.2. Proof of Lemma 2

The equivalence of forms become clear once we rewrite (6) in the matrix notation. Equation (6) can be written as follows:

$$\min_{\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma} J_2(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma) = \min_{\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma} \frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{K}(\gamma; \mathbf{Z})\boldsymbol{\alpha}\|_2^2 + \frac{1}{2}\lambda_1 \boldsymbol{\alpha}^\top \mathbf{K}(\gamma; Z)\boldsymbol{\alpha} + \lambda_2 \rho(\gamma; \delta). \tag{A1}$$

For fixed $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and λ_1, minimizing the function in (A1) with respect to γ is equivalent to

$$\min_{\gamma} \left\{ \frac{1}{2n} \left\| \left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \frac{n}{2}\lambda_1 \boldsymbol{\alpha} \right) - \mathbf{K}(\gamma; Z)\boldsymbol{\alpha} \right\|_2^2 + \lambda_2 \rho(\gamma; \delta) \right\}. \tag{A2}$$

Appendix A.3. Proof of Theorem 1

With loss of the generality we use the penalty function for sparse group lasso but this proof can easily be modified for other penalty functions. Also, we fix $\lambda_1 = \lambda_2 = \delta = 1$, and consider $\beta \in \mathcal{R}$ as well as set the design matrix $\mathbf{X}$ (or vector in this case) scaled to have norm 1. The case of $\boldsymbol{\beta} \in \mathcal{R}^q$ will follow along similar lines of arguments. Let $\gamma \in D_3$ with $D_3 = \{\gamma : \|\gamma\|_1 \leq \frac{1}{2n}\|\mathbf{Y}\|_2^2\}$. Define $f(\gamma) = \|\mathbf{K}(\gamma; Z)\| = \eta_{max}(\mathbf{K}(\gamma; Z)) \geq 0$, where $\eta_{max}(\mathbf{K}(\gamma; Z))$ denotes the largest eigenvalue of $\mathbf{K}(\gamma; Z)$ with the operator norm (the norm of $\mathbf{K}(\gamma; Z)$) defined in its usual way $\|\mathbf{K}(\gamma; Z)\| = sup\{\|\mathbf{K}(\gamma; Z)\mathbf{x}\|_2^2 : \|\mathbf{x}\|_2^2 = 1\}$. Since D_3

is compact and $\mathbf{K}(\gamma; Z)$ is continuous with respect to γ it achieves its maximum over D_3. Thus, we define $\eta^\star = sup_{\gamma \in D_3} f(\gamma) \geq 0$. Define $D_2 = \{\beta :\mid \beta \mid \leq (1+\eta^\star)\|\mathbf{Y}\|_2\}$, where the upper bound is denoted by $b^\star = (1+\eta^\star)\|\mathbf{Y}\|_2 \geq 0$. Moreover, define $D_1 = \{\boldsymbol{\alpha} : \|\boldsymbol{\alpha}\|_2 \leq \sqrt{n}(\|\mathbf{Y}\|_2 + b^\star)\}$.

Since D_1, D_2 and D_3 are compact there exists a $(\boldsymbol{\alpha}^\star, \beta^\star, \gamma^\star)$ such that $J_2(\boldsymbol{\alpha}^\star, \beta^\star, \gamma^\star) \leq J_2(\boldsymbol{\alpha}, \beta, \gamma)$ for all $(\boldsymbol{\alpha}, \beta, \gamma) \in D_1 \times D_2 \times D_3$. Note that $J_2(\mathbf{0}, 0, \mathbf{0}) = \frac{1}{2n}\|\mathbf{Y}\|_2^2$ and $(\mathbf{0}, 0, \mathbf{0}) \in D_1 \times D_2 \times D_3$. We claim that $(\boldsymbol{\alpha}^\star, \beta^\star, \gamma^\star)$ is a global minimizer, which is proved below by contradiction.

Suppose that there exists $(\tilde{\boldsymbol{\alpha}}, \tilde{\beta}, \tilde{\gamma}) \notin D_1 \times D_2 \times D_3$ where $J_2(\tilde{\boldsymbol{\alpha}}, \tilde{\beta}, \tilde{\gamma}) < J_2(\boldsymbol{\alpha}^\star, \beta^\star, \gamma^\star)$. We must have that $\tilde{\gamma} \in D_3$; if not, we have $J_2(\tilde{\boldsymbol{\alpha}}, \tilde{\beta}, \tilde{\gamma}) \geq \|\tilde{\gamma}\|_1 \geq J_2(\mathbf{0}, 0, \mathbf{0}) \geq J_2(\boldsymbol{\alpha}^\star, \beta^\star, \gamma^\star)$. Let $q_1, \cdots, q_n$ be the orthonormal vectors of $\mathbf{K}(\tilde{\gamma}; Z)$ with its associated eigenvalues $\eta_1 \geq \cdots \geq \eta_n \geq 0$. We can write out $\tilde{\boldsymbol{\alpha}}, \mathbf{X}, \mathbf{Y}$ in terms of these basis functions where $\tilde{\boldsymbol{\alpha}} = \sum_{i=1}^n < \tilde{\boldsymbol{\alpha}}, q_i > q_i$, $\mathbf{Y} = \sum_{i=1}^n < \mathbf{Y}, q_i > q_i$ and $\mathbf{X} = \sum_{i=1}^n < \mathbf{X}, q_i > q_i$. Let $C_i^{\tilde{\boldsymbol{\alpha}}} =< \tilde{\boldsymbol{\alpha}}, q_i >$, $C_i^{\mathbf{Y}} =< \mathbf{Y}, q_i >$ and $C_i^{\mathbf{X}} =< \mathbf{X}, q_i >$. It follows that

$$J_2(\tilde{\boldsymbol{\alpha}}, \tilde{\beta}, \tilde{\gamma}) \geq \frac{1}{2n}\left\| \sum_{i=1}^n C_i^{\mathbf{Y}} q_i - \sum_{i=1}^n C_i^{\mathbf{X}} \tilde{\beta} q_i - \sum_{i=1}^n C_i^{\tilde{\boldsymbol{\alpha}}} \eta_i q_i \right\|_2^2 + \frac{1}{2}\sum_{i=1}^n (C_i^{\tilde{\boldsymbol{\alpha}}})^2 \eta_i,$$

which is equal to $\frac{1}{2n}\sum_{i=1}^n (C_i^{\mathbf{Y}} - C_i^{\mathbf{X}}\tilde{\beta} - C_i^{\tilde{\boldsymbol{\alpha}}}\eta_i)^2 + \frac{1}{2}\sum_{i=1}^n (C_i^{\tilde{\boldsymbol{\alpha}}})^2 \eta_i$. We can minimize the above objective function with respect to $C_i^{\tilde{\boldsymbol{\alpha}}}$ and $\tilde{\beta}$. First, note that for any $\eta_i = 0$ we can let $C_i^{\tilde{\boldsymbol{\alpha}}} = 0$ as it will not affect the expression above. It is sufficient to consider $\eta_i > 0$. Taking the first derivative and setting it equal to zero, we obtain the score equations the minimizer must satisfy, for our minimum $\tilde{\beta}$ and $C_i^{\tilde{\boldsymbol{\alpha}}}$

$$\beta = \sum_{i=1}^n C_i^{\mathbf{X}}(C_i^{\mathbf{Y}} - C_i^{\tilde{\boldsymbol{\alpha}}}\eta_i) \tag{A3}$$

$$C_i^{\tilde{\boldsymbol{\alpha}}} = \frac{1}{n+\eta_i}(C_i^{\mathbf{Y}} - C_i^{\mathbf{X}}\tilde{\beta}). \tag{A4}$$

In the above derivation we used the fact that $1 = \|\mathbf{X}\|_2^2 = \sum_{i=1}^n (C_i^{\mathbf{X}})^2$. Plugging (A4) into (A3), we obtain

$$\beta = \frac{\sum_{i=1}^n C_i^X C_i^Y (1 - \frac{\eta_i}{n+\eta_i})}{1 - \sum_{i=1}^n (C_i^X)^2 \frac{\eta_i}{n+\eta_i}}. \tag{A5}$$

It follows that

$$\beta \leq \frac{\sum_{i=1}^n \mid C_i^{\mathbf{X}} C_i^{\mathbf{Y}} \mid}{1 - \sum_{i=1}^n (C_i^{\mathbf{X}})^2 \frac{\eta^\star}{n+\eta^\star}} \leq \frac{\|\mathbf{X}\|_2 \|\mathbf{Y}\|_2}{\|\mathbf{X}\|_2^2 (1 - \frac{\eta^\star}{n+\eta^\star})} \leq \frac{\|\mathbf{Y}\|_2}{(1 - \frac{\eta^\star}{1+\eta^\star})} = b^\star.$$

Thus, the β that minimizes J_2 for a given $\gamma \in D_3$ is in D_2. Also, (A4) implies that $\mid C_i^{\tilde{\boldsymbol{\alpha}}} \mid \leq (\|\mathbf{Y}\|_2 + \|\mathbf{X}\|_2 \|\beta\|_2)$; consequently, the optimal $\boldsymbol{\alpha}$ for the given $\tilde{\gamma} \in D_3$ and $\beta \in D_2$ that minimizes J_2 satisfies $\|\boldsymbol{\alpha}\|_2 \leq \sqrt{n}(\|\mathbf{Y}\|_2 + b^\star)$. As a result, $\alpha \in D_2$. This suggests that for any $(\tilde{\boldsymbol{\alpha}}, \tilde{\beta}, \tilde{\gamma}) \notin D_1 \times D_2 \times D_3$ we can find an $(\boldsymbol{\alpha}, \beta, \gamma) \in D_1 \times D_2 \times D_3$ such that $J_2(\tilde{\boldsymbol{\alpha}}, \tilde{\beta}, \tilde{\gamma}) \geq J_2(\boldsymbol{\alpha}, \beta, \gamma)$.

Appendix A.4. Proof of Theorem 2

By Lemma 8.4 on page 129 in [32], Assumptions 1, 2, and 3 imply:

$$P\left(\sup_{b \in \mathcal{B}} \frac{\frac{1}{\sqrt{n}} |\sum_{i=1}^n \epsilon_i b(\mathbf{z}_i)|}{\|b\|_{P_n}^{1-\psi}} \geq T \right) \leq c \exp\left(-\frac{T^2}{c^2} \right),\ T \geq c \tag{A6}$$

where the constant c is dependent on C_1, C_2, C_3, C_4, and ψ. It follows that

$$\sup_{b \in \mathcal{B}} \frac{\frac{1}{\sqrt{n}}|\sum_{i=1}^{n} \epsilon_i b(\mathbf{z}_i)|}{\|b\|_{P_n}^{1-\psi}} = O_p(1). \tag{A7}$$

Therefore, for any $h \in \mathcal{H}_{\mathcal{K}}$ and a scaling map function $\Gamma \in \mathcal{A}$, we obtain

$$\frac{\sqrt{n}(\epsilon, h \circ \Gamma - h_0 \circ \Gamma_0)_n \left(\|h\|_{\mathcal{H}_{\mathcal{K}}}^2 + \|h_0\|_{\mathcal{H}_{\mathcal{K}}}^2 + \|\Gamma\|_{SGL}^2 + \|\Gamma_0\|_{SGL}^2\right)^{-\psi}}{\|h \circ \Gamma - h_0 \circ \Gamma_0\|_{P_n}^{1-\psi}} = O_p(1). \tag{A8}$$

For our estimators, $\hat{h}$ and $\hat{\Gamma}$, it is easy to see that

$$\begin{aligned}&(\epsilon, \hat{h} \circ \hat{\Gamma} - h_0 \circ \Gamma_0)_n = \\ &O_p(n^{-\frac{1}{2}})\left\|\hat{h} \circ \hat{\Gamma} - h_0 \circ \Gamma_0\right\|_n^{1-\psi}\left(\left\|\hat{h}\right\|_{\mathcal{H}_{\mathcal{K}}}^2 + \|h_0\|_{\mathcal{H}_{\mathcal{K}}}^2 + \|\hat{\Gamma}\|_{SGL}^2 + \|\Gamma_0\|_{SGL}^2\right)^{\psi}.\end{aligned} \tag{A9}$$

From (A9), we obtain the following inequality:

$$\begin{aligned}&\left\|\hat{h} \circ \hat{\Gamma} - h_0 \circ \Gamma_0\right\|_n^2 + \lambda_1\left\|\hat{h}\right\|_{\mathcal{H}_{\mathcal{K}}}^2 + \lambda_2\|\hat{\Gamma}\|_{SGL}^2 \leq \\ &O_p(n^{-\frac{1}{2}})\left\|\hat{h} \circ \hat{\Gamma} - h_0 \circ \Gamma_0\right\|_n^{1-\psi}\left(\left\|\hat{h}\right\|_{\mathcal{H}_{\mathcal{K}}}^2 + \|h_0\|_{\mathcal{H}_{\mathcal{K}}}^2 + \|\hat{\Gamma}\|_{SGL}^2 + \|\Gamma_0\|_{SGL}^2\right)^{\psi} \\ &+ \lambda_1\|h_0\|_{\mathcal{H}_{\mathcal{K}}}^2 + \lambda_2\|\Gamma_0\|_{SGL}^2.\end{aligned} \tag{A10}$$

We require $\lambda_1 = O_p(1)\lambda_2$, namely λ_2 and λ_1 go to zero at the same rate. We will show at the end of the proof what happens if they are not of the same order. Therefore, without loss of generality, we set $\lambda_1 = \lambda_2$, denoted by λ. In what follows, we divide (A10) into two cases.

Case 1: Suppose that

$$\begin{aligned}O_p(n^{-\frac{1}{2}})\left\|\hat{h} \circ \hat{\Gamma} - h_0 \circ \Gamma_0\right\|_n^{1-\psi}\left(\left\|\hat{h}\right\|_{\mathcal{H}_{\mathcal{K}}}^2 + \|h_0\|_{\mathcal{H}_{\mathcal{K}}}^2 + \|\hat{\Gamma}\|_{SGL}^2 + \|\Gamma_0\|_{SGL}^2\right)^{\psi} \\ \geq \lambda\left(\|h_0\|_{\mathcal{H}_{\mathcal{K}}}^2 + \|\Gamma_0\|_{SGL}^2\right).\end{aligned}$$

In this case, we have

$$\begin{aligned}&\left\|\hat{h} \circ \hat{\Gamma} - h_0 \circ \Gamma_0\right\|_n^2 + \lambda\left(\left\|\hat{h}\right\|_{\mathcal{H}_{\mathcal{K}}}^2 + \|\hat{\Gamma}\|_{SGL}^2\right) \leq \\ &O_p(n^{-\frac{1}{2}})\left\|\hat{h} \circ \hat{\Gamma} - h_0 \circ \Gamma_0\right\|_n^{1-\psi}\left(\left\|\hat{h}\right\|_{\mathcal{H}_{\mathcal{K}}}^2 + \|h_0\|_{\mathcal{H}_{\mathcal{K}}}^2 + \|\hat{\Gamma}\|_{SGL}^2 + \|\Gamma_0\|_{SGL}^2\right)^{\psi}.\end{aligned} \tag{A11}$$

Above (A11) is further discussed separately in two sub-cases.

Case 1a: If $\|h_0\|_{\mathcal{H}_{\mathcal{K}}}^2 + \|\Gamma_0\|_{SGL}^2 \leq \left\|\hat{h}\right\|_{\mathcal{H}_{\mathcal{K}}}^2 + \|\hat{\Gamma}\|_{SGL}^2$, then we have

$$\begin{aligned}\left\|\hat{h} \circ \hat{\Gamma} - h_0 \circ \Gamma_0\right\|_n^2 + \lambda\left(\left\|\hat{h}\right\|_{\mathcal{H}_{\mathcal{K}}}^2 + \|\hat{\Gamma}\|_{SGL}^2\right) \leq \\ O_p(n^{-\frac{1}{2}})\left\|\hat{h} \circ \hat{\Gamma} - h_0 \circ \Gamma_0\right\|_n^{1-\psi}\left(\left\|\hat{h}\right\|_{\mathcal{H}_{\mathcal{K}}}^2 + \|\hat{\Gamma}\|_{SGL}^2\right)^{\psi}.\end{aligned} \tag{A12}$$

Therefore,

$$\left(\left\|\hat{h}\right\|_{H_{\mathcal{K}}}^2 + \|\hat{\Gamma}\|_{SGL}^2\right)^{\psi} \leq O_p(n^{-\frac{\psi}{2(1-\psi)}})\left\|\hat{h} \circ \hat{\Gamma} - h_0 \circ \Gamma_0\right\|_n^{\psi}\lambda^{-\frac{\psi}{1-\psi}}. \tag{A13}$$

It follows that

$$\begin{aligned}\left\|\hat{h}\circ\hat{\Gamma}-h_0\circ\Gamma_0\right\|_n &= O_p(n^{-\frac{1}{2(1-\psi)}})O_p(\lambda^{-\frac{\psi}{1-\psi}}),\\ \left\|\hat{h}\right\|_{H_K}^2+\|\hat{\Gamma}\|_{SGL}^2 &= O_p(n^{-\frac{1}{1-\psi}})O_p(\lambda^{-\frac{1+\psi}{1-\psi}}).\end{aligned} \tag{A14}$$

Case 1b: If $\|h_0\|_{\mathcal{H}_K}^2+\|\Gamma_0\|_{SGL}^2 \geq \left\|\hat{h}\right\|_{H_K}^2+\|\hat{\Gamma}\|_{SGL}^2$, then:

$$\left\|\hat{h}\right\|_{\mathcal{H}_K}^2+\|\hat{\Gamma}\|_{SGL}^2 = O_p(\|h_0\|_{\mathcal{H}_K}^2+\|\Gamma_0\|_{SGL}^2)O_p(1).$$

Therefore,

$$\left\|\hat{h}\circ\hat{\Gamma}-h_0\circ\Gamma_0\right\|_n = O_p(n^{-\frac{1}{2(1+\psi)}})\left(\|h_0\|_{\mathcal{H}_K}^2+\|\Gamma\|_{SGL}^2]\right)^{\frac{\psi}{1+\psi}}.$$

Consequently, we obtain

$$\begin{aligned}\left\|\hat{h}\circ\hat{\Gamma}-h_0\circ\Gamma_0\right\|_n &= O_p(n^{-\frac{1}{2(1-\psi)}})O_p(\lambda^{-\frac{\psi}{1-\psi}}),\\ \left\|\hat{h}\right\|_{\mathcal{H}_K}^2+\|\hat{\Gamma}\|_{SGL}^2 &= O_p(n^{-\frac{1}{1-\psi}})O_p(\lambda^{-\frac{1+\psi}{1-\psi}}).\end{aligned} \tag{A15}$$

Both terms in (A15) are the same rates as those in (A14).
Case 2: Suppose that

$$\begin{aligned}O_p(n^{-\frac{1}{2}})\left\|\hat{h}\circ\hat{\Gamma}-h_0\circ\Gamma_0\right\|_n^{1-\psi}\left(\left\|\hat{h}\right\|_{\mathcal{H}_K}^2+\|h_0\|_{\mathcal{H}_K}^2+\|\hat{\Gamma}\|_{SGL}^2+\|\Gamma_0\|_{SGL}^2\right)^{\psi}&\\ \leq \lambda(\|h_0\|_{\mathcal{H}_K}^2+\|\Gamma_0\|_{SGL}^2).&\end{aligned}$$

Then, we have

$$\left\|\hat{h}\circ\hat{\Gamma}-h_0\circ\Gamma_0\right\|_n^2+\lambda\left(\left\|\hat{h}\right\|_{\mathcal{H}_K}^2+\|\hat{\Gamma}\|_{SGL}^2\right) \leq 2\lambda\left(\|h_0\|_{\mathcal{H}_K}^2+\|\Gamma_0\|_{SGL}^2\right).$$

This implies that

$$\begin{aligned}\left\|\hat{h}\circ\hat{\Gamma}-h_0\circ\Gamma_0\right\|_n &= O_p(\lambda^{\frac{1}{2}})\left(\|h_0\|_{\mathcal{H}_K}^2+\|\Gamma_0\|_{SGL}^2\right)^{\frac{1}{2}},\\ \left\|\hat{h}\right\|_{\mathcal{H}_K}^2+\|\hat{\Gamma}\|_{SGL}^2 &= O_p(1)\left(\|h_0\|_{\mathcal{H}_K}^2+\|\Gamma_0\|_{SGL}^2\right).\end{aligned} \tag{A16}$$

In order to make (A14) and (A16) have the same rates we first equate the two term $O_p(\lambda^{\frac{1}{2}})\left(\|h\|_{\mathcal{H}_K}^2+\|\Gamma\|_{SGL}^2\right)^{\frac{1}{2}}$ and $O_p(n^{-\frac{1}{2(1-\psi)}})O_p(\lambda^{-\frac{\psi}{1-\psi}})$, and then solve for a common λ. The solution is given as follows:

$$\lambda^{-1} = n^{\frac{1}{1+\psi}}\left(\|h\|_{\mathcal{H}_K}^2+\|\Gamma\|_{SGL}^2\right)^{\frac{1-\psi}{1+\psi}}.$$

Under this λ value we obtain that (A14)–(A16) as of the form:

$$\left\|\hat{h}\circ\hat{\Gamma}-h_0\circ\Gamma_0\right\|_n = O_p(n^{-\frac{1}{2(1+\psi)}})\left(\|h_0\|_{\mathcal{H}_K}^2+\|\Gamma_0\|_{SGL}^2\right)^{\frac{\psi}{1+\psi}}, \tag{A17}$$

$$\left\|\hat{h}\right\|_{\mathcal{H}_K}^2+\|\hat{\Gamma}\|_{SGL}^2 = O_p(1)\left(\|h_0\|_{\mathcal{H}_K}^2+\|\Gamma_0\|_{SGL}^2\right). \tag{A18}$$

This completes the proof of Theorem 2.

Now we discuss the situation where the tuning parameters λ_1 and λ_2 are not of the same order. As seen blow, the selection consistency may not be guaranteed. Take Case 2 as an example. Suppose that

$$O_p(n^{-\frac{1}{2}})\left\|\hat{h}\circ\hat{\Gamma}-h_0\circ\Gamma_0\right\|_n^{1-\psi}\left(\left\|\hat{h}\right\|_{\mathcal{H}_\mathcal{K}}^2+\|h_0\|_{\mathcal{H}_\mathcal{K}}^2+\|\hat{\Gamma}\|_{SGL}^2+\|\Gamma_0\|_{SGL}^2\right)^{\psi} \\ \leq \lambda_1\|h_0\|_{\mathcal{H}_\mathcal{K}}^2+\lambda_2\|\Gamma_0\|_{SGL}^2.$$

Let us consider two cases.
Case 2a: If $\lambda_1\|h_0\|_{\mathcal{H}_\mathcal{K}}^2 \leq \lambda_2\|\Gamma_0\|_{SGL}^2$, following the same arguments above, we have

$$\begin{aligned}\left\|\hat{h}\circ\hat{\Gamma}-h_0\circ\Gamma_0\right\|_n &= O_p(\lambda_2^{\frac{1}{2}})\|\Gamma_0\|_{SGL}),\\ \left\|\hat{h}\right\|_{\mathcal{H}_\mathcal{K}}^2 &= O_p(\frac{\lambda_2}{\lambda_1})\|\Gamma_0\|_{SGL}^2,\\ \|\hat{\Gamma}\|_{SGL}^2 &= O_p(1)\|\Gamma_0\|_{SGL}^2.\end{aligned} \tag{A19}$$

Case 2b: If $\lambda_1\|h_0\|_{\mathcal{H}_\mathcal{K}}^2 \geq \lambda_2\|\Gamma_0\|_{SGL}^2$, then following the same logic as before:

$$\begin{aligned}\left\|\hat{h}\circ\hat{\Gamma}-h_0\circ\Gamma_0\right\|_n &= O_p(\lambda_1^{\frac{1}{2}})\|h_0\|_{\mathcal{H}_\mathcal{K}}),\\ \|\hat{\Gamma}\|_{SGL}^2 &= O_p(\frac{\lambda_1}{\lambda_2})\|h_0\|_{\mathcal{H}_\mathcal{K}}^2,\\ \left\|\hat{h}\right\|_{\mathcal{H}_\mathcal{K}}^2 &= O_p(1)\|h_0\|_{\mathcal{H}_\mathcal{K}}^2.\end{aligned} \tag{A20}$$

Both terms involve $O_p(\frac{\lambda_1}{\lambda_2})$ and $O_p(\frac{\lambda_2}{\lambda_1})$, indicating that these two tuning parameters λ_1 and λ_2 should go to zero at the same rates. Moreover, we can think of our estimator $\hat{h}\circ\hat{\Gamma}$ as one operational object. See Appendix B for more details on this, which can further explain the need of one rate for the two penalties.

Appendix A.5. Proof of Corollary 1

For convenience, we present the following lemma proved by [32] (on page 20).

Lemma A1. *(Geer's Lemma) A d dimensional ball of radius R, $B_d(R)$, in $\mathcal{R}^d$ with Euclidean metric can be covered by $(\frac{4R+\delta}{\delta})^d$ balls of radius δ.*

We have shown in the proof of Theorem 1 that the optimal γ vector is restricted to be within a ball of a radius that depends on the norm of $\mathbf{Y}$. For the sake of simplicity let us confine our γ to be within a norm ball of radius 1, $\gamma \in \mathcal{G} = \{\gamma : \|\gamma\|_2^2 \leq 1\}$. We then confine our set which we called $\mathcal{A}$ to be restricted to those γ, that is $\mathcal{A} = \{\Gamma : \Gamma(\mathbf{z}) = \gamma\circ\mathbf{z}, \gamma \in \mathcal{G}\}$. Since our $\gamma \in R^s$, we can use above Lemma A1 and cover our set $\mathcal{A}$ with $N_1 = \left(\frac{4+\delta}{\delta}\right)^s$ number of functions in the following sense. The ball of radius 1 in R^s can be covered (using the Euclidean metric) by $\{\gamma_1, \cdots \gamma_{N_1}\}$. Since there is a one to one relationship between the functions Γ and γ, take the set $\{\Gamma_1, \ldots, \Gamma_{N_1}\}$ and define the metric between some Γ_j and Γ_k in the set $\mathcal{A}$ as $d(\Gamma_j, \Gamma_k) = \left\|\gamma_j - \gamma_k\right\|_2$. Then, the set of functions $\{\Gamma_1, \ldots, \Gamma_{N_1}\}$ is a δ-covering for $\mathcal{A}$ under this metric with entropy $s\ log(\frac{4+\delta}{\delta})$. For each Γ_j we have an induced RKHS, $\mathcal{H}_{\mathcal{K}\circ\Gamma_j} = \{h\circ\Gamma_j : h \in \mathcal{H}_\mathcal{K}\}$ with entropy no larger than that of $\mathcal{H}_\mathcal{K}$, which according to the assumption, has entropy $\leq A\delta^{-2\psi}$ for some $\psi \in (0,1)$ and $A \in \mathcal{R}$. Therefore, the covering number $N_2 = N(\delta, \mathcal{H}_{\mathcal{K}\circ\Gamma_j}, P_n) \leq \exp\{A\delta^{-2\psi}\}$. This implies that for every Γ_j there exists a set $\{h_{j_1}\circ\Gamma_j, \cdots, h_{j_{N_2}}\circ\Gamma_j\}$ such that for every $h\circ\Gamma_j \in \mathcal{H}_{\mathcal{K}\circ\Gamma_j}$ there exists an integer $i \in \{1, \ldots, N_2\}$ we have $\left\|h\circ\Gamma_j - h_{j_i}\circ\Gamma_j\right\|_{P_n} \leq \delta$. Set $\mathcal{B}$ is essentially the union of the different Hilbert spaces of the form $\mathcal{H}_{\mathcal{K}\circ\Gamma}$. Under the setup, a natural estimate of the *delta*-covering number of this set would be approximately of size $N_1 \times N_2$ where functions take the form of $\{h_{1_1} \circ \Gamma_1, \cdots, h_{1_{N_2}} \circ \Gamma_1, \cdots, h_{N_{11}} \circ \Gamma_{N_1}, \cdots, h_{N_{1N_2}} \circ \Gamma_{N_1}\}$. In addition, we add N_2 functions from the set $\{h_1 \circ \Gamma_0, \cdots, h_{N_2} \circ \Gamma_0\}$ where Γ_0 is the true Γ_0 (or one of the true Γ_0). Since $\mathcal{H}_{\mathcal{K}\circ\Gamma_j}$ is a Hilbert space for every j, if $h \circ \Gamma_j \in \mathcal{H}_{\mathcal{K}\circ\Gamma_j}$ so is $\frac{h\circ\Gamma_j}{\|h\|^2_{\mathcal{H}_\mathcal{K}}+\|h_0\|^2_{\mathcal{H}_\mathcal{K}}+\left\|\Gamma_j\right\|^2_{SGL}+\|\Gamma_0\|^2_{SGL}}$. We can simply ignore the denominator and substitute $\frac{h\circ\Gamma_j}{\|h\|^2_{\mathcal{H}_\mathcal{K}}+\|h_0\|^2_{\mathcal{H}_\mathcal{K}}+\left\|\Gamma_j\right\|^2_{SGL}+\|\Gamma_0\|^2_{SGL}}$ with $\tilde{h} \circ \Gamma_j \in H_{K\circ\Gamma_j}$ where $\tilde{h} = \frac{h}{\|h\|^2_{\mathcal{H}_\mathcal{K}}+\|h_0\|^2_{\mathcal{H}_\mathcal{K}}+\left\|\Gamma_j\right\|^2_{SGL}+\|\Gamma_0\|^2_{SGL}}$.

We now prove Corollary 1.

Proof. Set $M = \sup_h < \nabla h(\mathbf{z}), \nabla h(\mathbf{z}) >$ where the inner product is the standard Euclidean inner product. This is for a fixed $\mathbf{z}$, or under the assumption that the gradient is uniformly bounded, we can take the $\sup_{h\in\mathcal{H}_\mathcal{K},\mathbf{z}\in R^s} < \nabla h(\mathbf{z}), \nabla h(\mathbf{z}) >$. Let $N_1 = \frac{4+\left(\frac{\delta}{3M^{\frac{1}{2}}}\right)^s}{\left(\frac{\delta}{3M^{\frac{1}{2}}}\right)}$ which is the number of balls needed to provide a $\left(\frac{\delta}{3M^{\frac{1}{2}}}\right)$ covering for a norm 1 ball in $\mathcal{R}^s$. Let $N_2 = \exp\left\{\left(A(\frac{\delta}{3})^{-2\psi}\right)\right\}$ which is the covering number needed to provide a $\frac{\delta}{3}$ cover of our space $\mathcal{H}_\mathcal{K}$. Let:

$$
\begin{aligned}
&\tilde{\hat{h}} \circ \hat{\Gamma} - \tilde{h}_0 \circ \Gamma_0 = \\
&\frac{\hat{h} \circ \hat{\Gamma}}{\left\|\hat{h}\right\|^2_{\mathcal{H}_\mathcal{K}} + \|h_0\|^2_{\mathcal{H}_\mathcal{K}} + \left\|\hat{\Gamma}\right\|^2_{SGL} + \|\Gamma_0\|^2_{SGL}} - \frac{h_0 \circ \Gamma_0}{\left\|\hat{h}\right\|^2_{\mathcal{H}_\mathcal{K}} + \|h_0\|^2_{\mathcal{H}_\mathcal{K}} + \left\|\hat{\Gamma}\right\|^2_{SGL} + \|\Gamma_0\|^2_{SGL}}
\end{aligned}
$$

be an arbitrary function in the set $\mathcal{B}$. There exists a Γ_j where $j \in \{1, \ldots, N_1\}$ such that $d(\Gamma_j, \hat{\Gamma}) \leq \frac{\delta}{3\max\limits_{i=1,\cdots,n} \|\mathbf{z}_i\|_2 \sqrt{M}}$, and there exists an i where $i \in \{1, \ldots, N_2\}$ such that $\left\|\tilde{\hat{h}} \circ \Gamma_j - h_{j_i} \circ \Gamma_j\right\|_{P_n} \leq \frac{\delta}{3}$.

Similarly, there exists a $t \in \{1, \ldots, N_2\}$ such that $\left\|\tilde{h}_0 \circ \Gamma_0 - h_t \circ \Gamma_0\right\|_{P_n} \leq \frac{\delta}{3}$. We construct our approximating function of $\tilde{\hat{h}} \circ \hat{\Gamma} - \tilde{h}_0 \circ \Gamma_0$ as $h_{j_i} \circ \Gamma_j - h_t \circ \Gamma_0$. We now show that this function is within δ of our arbitrary function $\tilde{\hat{h}} \circ \hat{\Gamma} - \tilde{h}_0 \circ \Gamma_0$. Applying the mean value theorem for multivariate functions, $\tilde{\hat{h}} \circ \hat{\Gamma}(\mathbf{z}) = \tilde{\hat{h}} \circ \Gamma_j(\mathbf{z}) + \nabla\tilde{\hat{h}}(C(\mathbf{z}))(\Gamma\hat{(\mathbf{z})} - \Gamma_j(\mathbf{z}))$, we have:

$$
\begin{aligned}
&\left\|(\tilde{\hat{h}} \circ \hat{\Gamma} - \tilde{h}_0 \circ \Gamma_0) - (h_{j_i} \circ \Gamma_j - h_t \circ \Gamma_0)\right\|_{P_n} \\
&\leq \left\|\tilde{\hat{h}} \circ \hat{\Gamma} - h_{j_i} \circ \Gamma_j\right\|_{P_n} + \left\|\tilde{h}_0 \circ \Gamma_0 - h_t \circ \Gamma_0\right\|_{P_n} \\
&\leq \left\|\tilde{\hat{h}} \circ \hat{\Gamma} - h_{j_i} \circ \Gamma_j\right\|_{P_n} + \frac{\delta}{3} \\
&= \left\|\tilde{\hat{h}} \circ \Gamma_j - h_{j_i} \circ \Gamma_j + \nabla\tilde{\hat{h}}(C(\cdot))(\hat{\Gamma} - \Gamma_j)\right\|_{P_n} + \frac{\delta}{3}
\end{aligned}
$$

where vector $\mathbf{z} \in \mathcal{R}^s$ lies in the segment from $\gamma_j \circ \mathbf{z}$ and $\hat{\gamma} \circ \mathbf{z}$, and $C(\cdot)$ is an unknown function that maps from $\mathcal{R}^s$ into $\mathcal{R}^s$ that allows for the formula to hold. Continuing our chain of inequalities, we obtain:

$$\begin{aligned}
\left\|\tilde{\hat{h}} \circ \Gamma_j - h_{j_i} \circ \Gamma_j + \nabla\tilde{\hat{h}}(C(\cdot))(\hat{\Gamma} - \Gamma_j)\right\|_{P_n} + \frac{\delta}{3} \leq \\
\left\|\nabla\tilde{\hat{h}}(C(\cdot))(\hat{\Gamma} - \Gamma_j)\right\|_{P_n} + \frac{\delta}{3} + \frac{\delta}{3} = \\
\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\nabla\tilde{\hat{h}}(C(\mathbf{z}_i))(\hat{\Gamma}(\mathbf{z}_i) - \Gamma_j(\mathbf{z}_i))\right)^2} + \frac{\delta}{3} + \frac{\delta}{3} \leq \\
\sqrt{\frac{1}{n}\sum_{i=1}^{n} M\|\hat{\gamma} \circ \mathbf{z_i} - \gamma_j \circ \mathbf{z_i}\|_2^2} + \frac{\delta}{3} + \frac{\delta}{3} \leq \\
\sqrt{M\left(\frac{\delta}{3\max\limits_{i=1,\cdots,n}\|\mathbf{z}_i\|_2\sqrt{M}}\right)^2 \max_{i=1,\cdots,n}\|\mathbf{z}_i\|_2^2} + \frac{\delta}{3} + \frac{\delta}{3} = \\
\frac{\delta}{3} + \frac{\delta}{3} + \frac{\delta}{3} = \delta.
\end{aligned}$$

Therefore, to provide a δ cover we need $N_1 \times N_2 + N_2$ number of functions or:

$$\begin{aligned}
\exp\{\left(A(\frac{\delta}{3})^{-2\psi}\right)\}\left(\frac{4 + \left(\frac{\delta}{3M^{\frac{1}{2}}}\right)}{\left(\frac{\delta}{3M^{\frac{1}{2}}}\right)}\right)^s + \exp\left\{\left(A\left(\frac{\delta}{3}\right)^{-2\psi}\right)\right\} = \\
\exp\{\tilde{A}\delta^{-2\psi}\}\left(\frac{C+\delta}{\delta}\right)^s + \exp\{\tilde{A}\delta^{-2\psi}\},
\end{aligned}$$

where $\tilde{A} = \frac{A}{3^{-2\psi}}$ and $C = 12M^{\frac{1}{2}}$. Taking the log we see the entropy is $\leq \tilde{A}\delta^{-2\psi} + \log\left((\frac{C+\delta}{\delta})^s + 1\right)$ which is of the same order as $\leq \tilde{A}\delta^{-2\psi}$ (the *log* term is dominated by the first term). Therefore a sufficient (but not necessary) condition for our set $\mathcal{B}$ to have the same entropy as that of the original RKHS $\mathcal{H}_{\mathcal{K}}$ is for the $\sup_h < \nabla h(\mathbf{z}), \nabla h(\mathbf{z}) >$ to be bounded. Having bounded derivatives is reasonable for any RKHS since every RKHS satisfies the Lipschitz condition of the form:

$$|h(X) - h(Y)| = |< h, \mathcal{K}_X > - < h, \mathcal{K}_Y > | \leq \|h\|_{\mathcal{H}_{\mathcal{K}}} < \mathcal{K}_X, \mathcal{K}_Y >^{\frac{1}{2}} = \|h\|_{\mathcal{H}_{\mathcal{K}}} d(X, Y),$$

where the distance metric in $\mathcal{R}^s$ is defined as $d(X,Y)^2 = \mathcal{K}(X,X) - 2\mathcal{K}(X,Y) + \mathcal{K}(Y,Y)$. If we restrict our functions in the RKHS of norm $\leq C$ for some constant C then we have a universal Lipschitz constant C to ensure bounded derivatives. □

Appendix B. Discussion about the FKMR Estimator

We introduce γ as a way of performing variable selection on our vector of FPC features. We want to illustrate this technical trick with some concrete examples and discuss identifiability issues with the resulting estimator. There are two ways of looking at the estimation of the unknown functions h_0 and Γ_0. The first way is to view our feature vector, $\mathbf{z}$, as being related to the dependent variable y through the composite function $h \circ \Gamma$, as explained in Section 4. The second and equivalent way is to view our features as unknown. The true features take the form of $\gamma \circ \mathbf{z}$, where in this case the $\circ$ denotes the Hadamard product. We are given $\mathbf{z}$ and need to estimate the "true" features $\gamma \circ \mathbf{z}$. In addition, we need to estimate the relationship between $\gamma \circ \mathbf{z}$ and y, which is done through the function $h \in \mathcal{H}_{\mathcal{K}}$.

The first way is to estimate the function $h_0 \circ \Gamma_0$. The function belongs to the RKHS $\mathcal{H}_{\mathcal{K}\circ\Gamma}$. We essentially consider many different function spaces to construct our estimator. The intersection between the function spaces is not necessarily empty, implying that our estimator may not be unique. We proceed this discussion more formally. Let $\mathcal{K} : \mathcal{R}^s \times \mathcal{R}^s \mapsto \mathcal{R}$ be a positive definite function. Let $\Gamma : \mathcal{R}^s \mapsto \mathcal{R}^s$. We define $\mathcal{K} \circ \Gamma : \mathcal{R}^s \times \mathcal{R}^s \mapsto \mathcal{R}$ as the function given by $\mathcal{K} \circ \Gamma(\mathbf{s}, \mathbf{t}) = \mathcal{K}(\Gamma(\mathbf{s}), \Gamma(\mathbf{t}))$. This new function, $\mathcal{K} \circ \Gamma$ is positive definite. There is a relationship between the original RKHS, $\mathcal{H}_{\mathcal{K}}$ and the new RKHS, $\mathcal{H}_{\mathcal{K}\circ\Gamma}$. This results in $\mathcal{H}_{\mathcal{K}\circ\Gamma} = \{h \circ \Gamma : h \in \mathcal{H}_{\mathcal{K}}\}$. For any vector $u \in H_{\mathcal{K}\circ\Gamma}$, we have that $\|u\|_{\mathcal{H}_{\mathcal{K}\circ\Gamma}} = inf\{\|h\|_{\mathcal{H}_{\mathcal{K}}} : u = h \circ \Gamma\}$. In general, $\mathcal{H}_{\mathcal{K}\circ\Gamma} \not\subset \mathcal{H}_{\mathcal{K}}$. In (5), we take the norm with respect to the original space $\mathcal{H}_{\mathcal{K}}$. Our iterative procedure essentially presents the second way in which the true features are unknown, whereas our theoretical arguments are justified through the first way. Given the knowledge of the features (which translates to fixing a γ), we are confined to just one RKHS, $\mathcal{H}_{\mathcal{K}}$. Take the linear kernel, $\mathcal{K}(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1^\top \mathbf{x}_2$ as an example. Suppose the truth is that y is related to a one-dimensional feature $\mathbf{z}_0$ through the following formulation: $y = h_0(z_0) + \varepsilon$ where $h_0 \in \mathcal{H}_{\mathcal{K}_1}$, where $\mathcal{K}_1$ is the kernel that maps from $\mathcal{R} \times \mathcal{R} \mapsto \mathcal{R}$. Therefore, if we knew the feature z_1, we would proceed to optimize (6) using the standard LSKM. However, when each y is associated with a two-dimensional vector $\mathbf{z} = (z_1, z_2)$, where z_2 is a "noisy" feature and unrelated to y. Suppose that *a priori* we do not know this information. Typically we use a model $y = h(z_1, z_2) + \varepsilon$ where $h \in \mathcal{H}_{\mathcal{K}}$, where $\mathcal{K}$ is the kernel that maps from $\mathcal{R}^2 \times \mathcal{R}^2 \mapsto \mathcal{R}$. In this case, we introduce our $\boldsymbol{\gamma}$ vector (γ_1, γ_2) and formulate $y = h(\gamma_1 z_1, \gamma_2 z_2) + \epsilon$. All functions, h in the space $\mathcal{H}_{\mathcal{K}}$, are of the form $h(\mathbf{z}) = \mathbf{x}^\top \mathbf{z}$ for some two-dimensional vector $\mathbf{x} = (x_1, x_2)$. There is a one-to-one relationship between h and $\mathbf{x}$. The true function, h_0, has an associated real number c where $h_1(z_1) = cz_1$. We can recover $h_1 \in \mathcal{H}_{\mathcal{K}_1}$ from our estimation of h and γ if we set $\boldsymbol{\gamma} = (1, 0)$ and $\mathbf{x} = (c, \star)$, where "$\star$" is any real number. Equivalently, we can recover h_1 under $\boldsymbol{\gamma} = (1, 1)$ where $\mathbf{x} = (c, 0)$. There are many functions that may recover the original function in the RKHS corresponding to the linear space kernel. Formulating our problem in the first way, through function composition, we can estimate Γ_0 with the γ being $(1, 0)$ or $(1, 1)$.

We can now see that in the intersection between $\mathcal{H}_{\mathcal{K}\circ\Gamma_1}$ and $\mathcal{H}_{\mathcal{K}\circ\Gamma_2}$, where Γ_1 has associated $\boldsymbol{\gamma_1} = (1, 0)$ and Γ_2 has associated $\boldsymbol{\gamma_2} = (1, 1)$, lies our estimate of h_1. In truth, for the linear space RKHS, there is no need to apply our method since $h_0 \in \mathcal{H}_{\mathcal{K}_1}$ can be estimated directly from the larger space $\mathcal{H}_{\mathcal{K}}$ where we set $h(\mathbf{z}) = \mathbf{x}^\top \mathbf{z}$ where $\mathbf{x} = (c, 0)$. We can never hope to have variable selection consistency nor can we hope to have identifiability of our estimator for these types of spaces. However, from a goodness-of-fit standpoint, we are able to do just as good a job with many types of function compositions. Our hope is that we can glean some variable selection by penalizing the γ vector with the $\rho(\gamma; \delta)$ term which, going back to the above scenario, should give preference to $\gamma = (1, 0)$ over $\gamma = (1, 1)$. For the RKHS associated with the Gaussian Kernel, the "larger dimensional space", a Gaussian Kernel mapping from higher dimensions, does not necessarily contain the functions from a "lower dimensional space", a Gaussian Kernel mapping from lower dimensions. However through the introduction of the γ transformation of the features, we can recover the equivalent functions of the "lower dimensional space".

References

1. Chandler, J.L.; Brazendale, K.; Beets, M.W.; Mealing, B.A. Classification of Physical Activity Intensities Using a Wrist-worn Accelerometer in 8–12-Year-old Children. *Pediatric Obes.* **2016**, *11*, 120–127. [CrossRef] [PubMed]
2. Chen, K.Y.; Bassett, D.R. The Technology of Accelerometry-based Activity Monitors: Current and Future. *Med. Sci. Sport. Exerc.* **2005**, *37*, S490–S500. [CrossRef] [PubMed]
3. Bai, J.; Di, C.; Xiao, L.; Evenson, K.R.; LaCroix, A.Z.; Crainiceanu, C.M.; Buchner, D.M. An Activity Index for Raw Accelerometry Data and Its Comparison with Other Activity Metrics. *PLoS ONE* **2016**, *11*, e0160644. [CrossRef] [PubMed]
4. John, D.; Freedson, P. ActiGraph and Actical Physical Activity Monitors: A Peek under the Hood. *Med. Sci. Sport. Exerc.* **2012**, *44*, S86–S89. [CrossRef]
5. Kim, Y.; Lee, J.M.; Peters, B.P.; Gaesser, G.A.; Welk, G.J. Examination of Different Accelerometer Cut-points for Assessing Sedentary Behaviors in Children. *PLoS ONE* **2014**, *9*, e90630. [CrossRef]

6. Bai, J.; Sun, Y.; Schrack, J.A.; Crainiceanu, C.M.; Wang, M.C. A Two-stage Model for Wearable Device Data. *Biometrics* **2018**, *74*, 744–752. [CrossRef]
7. Sasaki, J.E.; Hickey, A.M.; Staudenmayer, J.W.; John, D.; Kent, J.A.; Freedson, P.S. Performance of Activity Classification Algorithms in Free-Living Older Adults. *Med. Sci. Sport. Exerc.* **2016**, *48*, 941–950. [CrossRef]
8. Di, C.Z.; Crainiceanu, C.M.; Caffo, B.S.; Punjabi, N.M. Multilevel Functional Principal Component Analysis. *Ann. Appl. Stat.* **2009**, *3*, 458–488. [CrossRef]
9. Goldsmith, J.; Liu, X.; Rundle, A.; Jacobson, J. New Insights into Activity Patterns in Children, Found Using Functional Data Analyses. *Med. Sci. Sport. Exerc.* **2016**, *48*, 1723–1729. [CrossRef]
10. Li, H.; Keadle, S.K.; Staudenmayer, J.; Assaad, H.; Huang, J.Z.; Carroll, R.J. Methods to Assess An Exercise Intervention Trial Based on 3-Level Functional Data. *Biostatistics* **2015**, *16*, 754–771. [CrossRef]
11. Zhang, Y.; Li, H.; Keadle, S.K.; Matthews, C.E.; Carroll, R.J. A Review of Statistical Analyses on Physical Activity Data Collected from Accelerometers. *Stat. Biosci.* **2019**, *11*, 465–476. [CrossRef] [PubMed]
12. Ramsay, J.O.; Silverman, B.W. *Functional Data Analysis*; Springer Series in Statistics; Springer: Berlin/Heidelberg, Germany, 2005.
13. Cardot, H.; Ferraty, F.; Sarda, P. Spline Estimators for the Functional Linear model. *Stat. Sin.* **2003**, *13*, 571–591.
14. Cardot, H.; Ferraty, F.; Sarda, P. Functional Linear Model. *Stat. Probab. Lett.* **1999**, *45*, 11–22. [CrossRef]
15. Zhu, H.; Yao, F.; Zhang, H.H. Structured Functional Additive Regression in Reproducing Kernel Hilbert Spaces. *J. R. Stat. Soc. Ser. (Stat. Methodol.)* **2014**, *76*, 581–603. [CrossRef]
16. Ferraty, F.; Mas, A.; Vieu, P. Nonparametric Regression on Functional Data: Inference and Practical Aspects. *Aust. N. Z. J. Stat.* **2007**, *49*, 267–286. [CrossRef]
17. McLean, M.W.; Hooker, G.; Staicu, A.M.; Scheipl, F.; Ruppert, D. Functional Generalized Additive Models. *J. Comput. Graph. Stat.* **2014**, *23*, 249–269. [CrossRef]
18. Bosq, D. *Linear Processes in Function Spaces*; Lecture Notes in Statistics; Springer: New York, NY, USA, 2000; Volume 149.
19. Hall, P.; Müller, H.G.; Wang, J.L. Properties of Principal Component Methods for Functional and Longitudinal Data Analysis. *Ann. Stat.* **2006**, *34*, 1493–1517. [CrossRef]
20. Hall, P.; Hosseini-Nasab, M. On Properties of Functional Principal Components Analysis. *J. R. Stat. Soc. Ser. (Stat. Methodol.)* **2006**, *68*, 109–126. [CrossRef]
21. Müller, H.G.; Yao, F. Functional Additive Models. *J. Am. Stat. Assoc.* **2008**, *103*, 1534–1544. [CrossRef]
22. Lin, Y.; Zhang, H.H. Component Selection and Smoothing in Multivariate Nonparametric Regression. *Ann. Stat.* **2006**, *34*, 2272–2297. [CrossRef]
23. Liu, D.; Lin, X.; Ghosh, D. Semiparametric Regression of Multidimensional Genetic Pathway Data: Least-Squares Kernel Machines and Linear Mixed Models. *Biometrics* **2007**, *63*, 1079–1088. [CrossRef] [PubMed]
24. Wood, S.N. *Generalized Additive Models: An Introduction with R*; Chapman and Hall: London, UK, 2006.
25. Lin, X.; Zhang, D. Inference in Generalized Additive Mixed Models by Using Smoothing Splines. *J. R. Stat. Soc. Ser. (Stat. Methodol.)* **1999**, *61*, 381–400. [CrossRef]
26. Yuan, M.; Lin, Y. Model Selection and Estimation in Regression with Grouped Variables. *J. R. Stat. Soc. Ser. (Stat. Methodol.)* **2006**, *68*, 49–67. [CrossRef]
27. Simon, N.; Friedman, J.; Hastie, T.; Tibshirani, R. A Sparse-Group Lasso. *J. Comput. Graph. Stat.* **2013**, *22*, 231–245. [CrossRef]
28. Breiman, L. Better Subset Regression Using the Nonnegative Garrote. *Technometrics* **1995**, *37*, 373–384. [CrossRef]
29. Salzo, S.; Villa, S. Convergence Analysis of a Proximal Gauss–Newton Method. *Comput. Optim. Appl.* **2012**, *53*, 557–589. [CrossRef]
30. Naiman, J. Multivariate Functional Kernel Machine Regression and Feature Selection with Applications to Accelerometer Mobile Health Devices. Ph.D. Dissertation, University of Michigan, Ann Arbor, MI, USA, 2020.
31. Peng, H.; Huang, T. Penalized Least Squares for Single Index Models. *J. Stat. Plan. Inference* **2011**, *141*, 1362–1379. [CrossRef]
32. Geer, S.A. *Empirical Processes in M-Estimation*; Cambridge Series in Statistical and Probabilistic Mathematics; Cambridge University Press: Cambridge, UK, 2000.
33. Hainmueller, J.; Hazlett, C. Kernel Regularized Least Squares: Reducing Misspecification Bias with a Flexible and Interpretable Machine Learning Approach. *Political Anal.* **2014**, *22*, 143–168. [CrossRef]
34. Yao, F.; Müller, H.G.; Wang, J.L. Functional Data Analysis for Sparse Longitudinal Data. *J. Am. Stat. Assoc.* **2005**, *100*, 577–590. [CrossRef]
35. Lewis, R.C.; Meeker, J.D.; Peterson, K.E.; Lee, J.M.; Pace, G.G.; Cantoral, A.; Téllez-Rojo, M.M. Predictors of Urinary Bisphenol A and Phthalate Metabolite Concentrations in Mexican Children. *Chemosphere* **2013**, *93*, 2390–2398. [CrossRef]
36. Schrack, J.A.; Zipunnikov, V.; Goldsmith, J.; Bai, J.; Simonsick, E.M.; Crainiceanu, C.; Ferrucci, L. Assessing the Physical Cliff: Detailed Quantification of Age-related Differences in Daily Patterns of Physical Activity. *J. Gerontol. Ser. Biol. Sci. Med. Sci.* **2014**, *69*, 973–979. [CrossRef] [PubMed]
37. Jansen, E.C.; Dunietz, G.L.; Chervin, R.D.; Baylin, A.; Baek, J.; Banker, M.; Song, P.X.K.; Cantoral, A.; Tellez Rojo, M.M.; Peterson, K.E. Adiposity in Adolescents: The Interplay of Sleep Duration and Sleep Variability. *J. Pediatr.* **2018**, *203*, 309–316. [CrossRef] [PubMed]