Electronic Journal of Statistics
Vol. 12 (2018) 4602–4639
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1509

Simultaneous variable selection and smoothing for high-dimensional function-on-scalar regression

Alice Parodi

MOX - Department of Mathematics
Politecnico di Milano
Milano, Italy
e-mail: alice.parodi89@gmail.com

Matthew Reimherr*

Department of Statistics
Pennsylvania State University
University Park, PA, USA
e-mail: mreimherr@psu.edu

Abstract: We present a new methodology, called *FLAME*, which simultaneously selects important predictors and produces smooth estimates in a function-on-scalar linear model with a large number of scalar predictors. Our framework applies quite generally by viewing the functional outcomes as elements of an arbitrary real separable Hilbert space. To select important predictors while also producing smooth parameter estimates, we utilize operators to define subspaces that are imbued with certain desirable properties as determined by the practitioner and the setting, such as smoothness or periodicity. In special cases one can show that these subspaces correspond to Reproducing Kernel Hilbert Spaces, however our methodology applies more broadly. We provide a very fast algorithm for computing the estimators, which is based on a functional coordinate descent, and an `R` package, `flm`, whose backend is written in `C++`. Asymptotic properties of the estimators are developed and simulations are provided to illustrate the advantages of FLAME over existing methods, both in terms of statistical performance and computational efficiency. We conclude with an application to childhood asthma, where we find a potentially important genetic mutation that was not selected by previous functional data based methods.

MSC 2010 subject classifications: 62G08, 62G20.
Keywords and phrases: Nonlinear regression, variable selection, functional data analysis, reproducing kernel Hilbert space, minimax convergence.

Received June 2018.

Contents

*Supported by NSF DMS 1712826 and NIDA P50 DA039838.

1. Introduction

High-dimensional regression and functional data analysis are currently central research areas in statistics and machine learning. The rising interest in both areas reflects the difficult realities of "big data" that many scientists are now facing in their work. Increasingly complex studies and data gathering technologies require sophisticated methods that are mathematically sound, computationally efficient, and practically interpretable. This work concerns a new approach for function-on-scalar regression:

$$Y_n = \sum_{i=1}^{I} X_{n,i}\beta_i^\star + \varepsilon_n \qquad n = 1, \ldots, N,$$

where the parameters, $\beta_i^\star$, and errors, ε_n, lie in a real separable Hilbert space (most commonly $L^2[0,1]$), but the predictors are real-valued scalars. However, we are interested in the case when the number of predictors, I, is much larger than than number of statistical units, N. Such data is especially common in genetic studies where one encounters large numbers of scalar predictors. These studies are also now increasingly likely to contain sophisticated phenotypic measurements that are suitable for functional data analysis. For example, in Section 4 we discuss an application concerning lung growth in children as they age. There the goal is to determine which genetic mutations are impacting this growth; it is assumed that this impact varies smoothly with age. Our methodology simultaneously exploits the smoothness of the underlying data and functional parameters, as well as the sparsity of the genetic effects. For short, we call this framework *FLAME*, for *functional linear adaptive mixed estimation*. The "mixed" here refers to the mixing of functional norms to simultaneously select significant predictors and smooth their corresponding effect on the functional outcome.

Currently, very little work has been done in this area, but there are several key recent papers which have made substantial in-roads into this problem. For

scalar-on-function regression, there are a few recent works [25, 24, 14, 12], but this is the opposite of the problem we consider here as it is our outcomes that are functional, not the predictors. Concerning function-on-scalar regression, Chen et al. [8] proposed combining functional least squares with a sparsity inducing penalty. There they took the penalty to be the *group minimax concave penalty*, MCP [38]. In addition, the authors used a pre-whitening technique to more fully exploit the within curve dependence. Unfortunately, the method is computationally expensive and cannot be applied when the number of predictors, I, is greater than the sample size, N, meaning that it cannot be applied to our intended high-dimensional applications. As we shall see in Section 4.2.2, the pre-whitening can also be counter productive when working with densely sampled functional data. Barber et al. [1] proposed the function-on-scalar lasso, FSL, which uses penalized functional least squares. In their approach they assumed the data and parameters were from an arbitrary Hilbert space, but to induce sparsity, the penalty was taken to be a type of induced ℓ_1 norm on the product space of Hilbert spaces where the parameters and data lie. Their approach is computationally efficient since it is a convex optimization problem, and achieves optimal rates of convergence for the parameter estimates even when the number of predictors, I, is much larger than the sample size N ($I \gg N$). However, the method, like traditional lasso, does not achieve the functional oracle property due to a non-negligible asymptotic bias. To that end, in a follow up paper Fan and Reimherr [13] developed an adaptive version, AFSL, and showed it achieves, what we call here, the *strong functional oracle property*, which we will discuss in further detail in Section 5. Furthermore, this method can be implemented at nearly the same computational cost as FSL. However, in all of these cases, the methods only shrink the estimates, there is no ability to include additional structure, such as smoothness or periodicity.

A related area of research we should mention concerns variable selection in time varying effect models (or varying coefficient models), TVEMs. There is a great deal of overlap when the TVEMs incorporate repeated measures on the same unit/subject, i.e. longitudinal data. These models have traditionally been designed for univariate functions that are observed very sparsely and with predictors that also vary over time (also called a concurrent model in functional data analysis). Wang et al. [35] considered variable selection in TVEMs using a B-splines expansion combined with a group SCAD penalty, but their focus was the low-dimensional setting, i.e. a fixed number of predictors. [26] extended this work for a more general basis and provide a more efficient computational technique for finding the minimizer. Wei et al. [36] then considered the high-dimensional case, where one has a large number of time-varying predictors. They also used an arbitrary basis alongside a group lasso penalty. Xue and Qu [37] also considered the low-dimensional setting using splines, but with a TLP penalty. In a fairly different direction [34] considered having both scalar predictors and time varying predictors that are observed with measurement error, but did not consider repeated measures or having a large number of predictors. Each of these works provides asymptotic theory in the form of Oracle properties. However, these works also focus on the case where the curves are sampled fairly sparsely,

while we focus on the functional case. In addition, we control the smoothness of the estimates via the RKHS norm, while these works control the smoothness by choosing the number of basis functions.

The major contributions of this work are as follows. We develop a new high-dimensional functional regression methodology that simultaneously selects important predictors and provides smooth estimates of their effects; previous approaches focused on selection only. Using convex analysis over Hilbert spaces, we provide a coordinate descent algorithm for model fitting and a very fast `R` package, `flm`, whose backend is written `C++`; previous methods "piggybacked" off of existing multivariate tools while ours is customized for functional data, resulting in substantial gains in computational efficiency. As part of this computational efficiency, we also avoid the use of the "Representer Theorem" of RKHSs for expressing parameter estimates, which, while theoretically convenient, is often not computationally efficient [31]. Instead we utilize the eigenfunctions of the kernel to expand the parameters, which can dramatically improve computational efficiency. We also provide asymptotic theory, which demonstrates that FLAME achieves a functional version of the oracle property. This theory requires substantial advances over the theory for FSL as we are mixing Hilbert space and RKHS norms, which are not equivalent (in a mathematical sense). Lastly, our framework allows one to build in a variety of structures into the parameters, including smoothness and periodicity. As can be seen in Section 4 this can result in dramatic gains in statistical efficiency.

The remainder of the paper is organized as follows. In Section 2 we outline several important concepts from FDA as well as the modeling assumptions of the data. In Section 3 we detail our approach, presenting a coordinate descent algorithm which allows FLAME to be computed very efficiently. In Sections 4 we present numerical illustrations including simulations and an application to a longitudinal genetic association study. We conclude with Section 5 which presents an asymptotic justification for our approach in the form of an oracle property.

2. Background and methodology

For a detailed introduction to FDA we refer the interested reader to Ramsay and Silverman [28], Graves et al. [15], Horváth and Kokoszka [18], Hsing and Eubank [19], Kokoszka and Reimherr [22]. For an introduction on machine learning and high dimensional regression we refer the reader to [16, 6, 21, 17]. Let $\mathbb{H}$ be a real separable Hilbert space, with norm $\|\cdot\|_{\mathbb{H}}$; our theory will hold quite generally for data from an arbitrary real separable Hilbert space. In this way, our methodology is quite broad covering typical spaces such as $L^2[0,1]$, as well as product spaces, Sobolev spaces, etc.

To produce a framework for selecting important predictors and producing smooth estimates which applies as broadly as possible, we introduce a linear operator, $K : \mathbb{H} \to \mathbb{H}$. We assume that it is positive definite and self-adjoint, i.e., $\langle Kx, x\rangle \geq 0$ and $\langle Kx, y\rangle = \langle x, Ky\rangle$. We also assume that the operator has

a spectral decomposition

$$K = \sum_{i=1}^{\infty} \theta_i v_i \otimes v_i,$$

where $\infty > \theta_1 \geq \theta_2 \geq \cdots > 0$ are the ordered eigenvalues and $v_i \in \mathbb{H}$ are the corresponding eigenfunctions, which are orthonormal since K is self-adjoint. This expansion would hold, for example, if K were compact [11], however the expansion can exist for noncompact operators, as well, for example the identity operator. The eigenfunctions $\{v_i\}$ form an orthonormal basis in $\mathbb{H}$. The tensor product $x \otimes y$ is used to denote the operator $(x \otimes y)(h) := \langle y, h \rangle x$. We define a subspace of $\mathbb{H}$, denoted $\mathbb{K}$, as follows:

$$\mathbb{K} := \left\{ h \in \mathbb{H} : \sum_{i=1}^{\infty} \frac{\langle h, v_i \rangle^2}{\theta_i} := \|h\|_{\mathbb{K}}^2 < \infty \right\}.$$

If we equip $\mathbb{K}$ with the norm $\|h\|_{\mathbb{K}}$ then this space also a Hilbert space. When $\mathbb{H}$ is $L^2[0,1]$ and K is an integral operator with a bivariate kernel function, $k(t,s)$:

$$(Kf)(t) = \int_0^1 k(t,s) f(s)\, ds,$$

then $\mathbb{K}$ is also a reproducing kernel Hilbert space with reproducing kernel $k(t,s)$ [4]. In this way, our setting is quite broad.

We now make the following modeling assumption about the response functions, $Y_n \in \mathbb{H}$, and the predictors $X_{n,i} \in \mathbb{R}$. In Section 5 additional assumptions will be presented when providing our asymptotic theory.

Assumption 1. *Let $Y_1, \ldots, Y_N$ be elements of $\mathbb{H}$, satisfying the functional linear model*

$$Y_n = \sum_{i=1}^{I} X_{n,i} \beta_i^{\star} + \varepsilon_n,$$

where $\mathbf{X} = \{X_{n,i}\} \in \mathbb{R}^{N \times I}$ is the deterministic design matrix with standardized columns, and ε_n are i.i.d. Gaussian random elements of $\mathbb{H}$ with mean function 0 and covariance operator C. We assume that there exists $0 \leq I_0 \leq I$ such that only $\beta_1^{\star}, \ldots, \beta_{I_0}^{\star}$ are nonzero. This means that, for notational simplicity, the first I_0 of the predictors are significant in the model. We will use the notation $\mathbf{X} = (\mathbf{X}_1\ \mathbf{X}_2)$ to partition the predictors into the significant predictors, $\mathbf{X}_1$, and the null predictors $\mathbf{X}_2$.

Note that any Gaussian process in $\mathbb{H}$ will necessarily have a mean function in $\mathbb{H}$ and a covariance operator C which is compact, symmetric, and positive definite [23]. In our theory, the normality is only used to derive functional concentration inequalities. These inequalities determine the rate at which I can grow with N. When the errors are Gaussian, one has that I can grow exponentially fast relative to N, and the assumptions (as given in Assumption 2) are easier to interpret. Our arguments can be readily generalized to the non-normal

case, but the rates will change and the assumptions will be more complicated, we thus do not pursue that direction presently.

Once one has the operator K, the FLAME target function is defined using a penalized least squares, a common approach in functional regression:

$$L(\beta) = \frac{1}{2N}\sum_{n=1}^{N} ||Y_n - X_n^\top \beta||_{\mathbb{H}}^2 + \lambda \sum_{i=1}^{I} \tilde{\omega}_i ||\beta_i||_{\mathbb{K}} = \frac{1}{2N}||Y - \mathbf{X}\beta||_{\mathbb{H}}^2 + \lambda \sum_{i=1}^{I} \tilde{\omega}_i ||\beta_i||_{\mathbb{K}},$$

with $Y \in \mathbb{H}^N$, $\mathbf{X} \in \mathbb{R}^{N\times I}$ and $X_n = \mathbf{X}_{(n,\cdot)} \in \mathbb{R}^I$, $\beta \in \mathbb{K}^I$. Throughout, we use notation such as $\mathbb{H}^N$ to denote product spaces. For the sake of simplicity, we abuse notation a bit by letting $\|\cdot\|_{\mathbb{H}}$ also denote the induced Hilbert space norm on product spaces such as $\mathbb{H}^N$. There are at least a few data driven ways one can choose the weights $\tilde{\omega}_i$. One option is to use marginal regressions to get initial parameter estimates, then the weights would be one over the norms of those estimates [20]. Another option is to run FSL first and then use its corresponding estimates. This has the advantage of also dramatically reducing the dimension of the problem, and is the approach we take for developing our asymptotic theory in Section 5. Lastly, one could first run the nonadaptive version of FLAME (i.e. with $\tilde{\omega} \equiv 1$) to obtain preliminary estimates, $\tilde{\beta}_i$, and then compute the weights as $\tilde{\omega}_i = \|\hat{\beta}_{i,N}\|_{\mathbb{K}}^{-1}$. This is the approach we take for our empirical work in Section 4. Our reasoning is that we wanted a more pure comparison between the different methods to compare their performances. Since all of the methods, except FSL, utilize a preliminary run to different degrees, opening the door to mixing and matching would create a huge number of potential options, and is beyond the scope of this paper.

In our approach we use the norm $\|\cdot\|_{\mathbb{K}}$ to both induce sparsity and smooth the parameter estimates. Previous approaches have focused only on one or the other. Furthermore, by allowing for a general K, we provide a framework which is very flexible and can accommodate a variety of underlying assumptions about the parameters, such as periodicity and boundary conditions. The purpose of the data driven weights is to penalize "smaller" parameters more, and thus not shrink the larger ones as much. This allows the estimator to be asymptotically unbiased and achieve an oracle property. We now discuss several examples of popular kernels. In general, we recommend using the Sobolev kernels with either one or two derivatives as the default, however, for those settings with additional structure other choices are viable as well. In particular, for periodic data we would recommend the periodic kernel, or for very smooth data we would recommend the Gaussian kernel.

Example 1 (Sobelev Space)**.** *Consider $\mathbb{H} = L^2(\mathcal{D})$, where $\mathcal{D}$ is a compact subset of $\mathbb{R}^d$. Define $\mathbb{K}$ to be the subset of functions in $L^2(\mathcal{D})$ that have up to and including m^{th} order derivatives that are also in $L^2(\mathcal{D})$. A family of norms can be defined on $\mathbb{K}$ as*

$$\|x\|_{\mathbb{K}}^2 = \sum_{|\alpha|\le m} \frac{1}{\sigma_\alpha^2} \int_{\mathcal{D}} |x^{(\alpha)}(\mathbf{s})|^2 \, d\mathbf{s}.$$

Here α is a d-dimensional vector of nonnegative integers whose sum is less than or equal to m, while the σ_α are nonzero weights. Equipped with this norm, $\mathbb{K}$ is an RKHS if and only if $m > d/2$. In the case where $\mathcal{D} = [0,1]$ and $m = 1$, we have that

$$k(t,s) = \begin{cases} \frac{\sigma}{\sinh(\sigma)} \cosh(\sigma(1-s)) \cosh(\sigma t) & t \leq s \\ \frac{\sigma}{\sinh(\sigma)} \cosh(\sigma(1-t)) \cosh(\sigma s) & t > s \end{cases}.$$

One can then numerically solve for the eigenfunctions and eigenvalues of K. These details can be found on Page 281 of Berlinet and Thomas-Agnan [4].

Example 2 (Gaussian Kernel)**.** *Let $\mathbb{H} = L^2(\mathcal{D})$, then the Gaussian kernel is given by*

$$k(\mathbf{s}, \mathbf{s}') = \exp\left\{-\sigma|\mathbf{s} - \mathbf{s}'|^2\right\}.$$

While the Sobelev spaces contain functions which are differentiable up to a given order, the space $\mathbb{K}$ here contains functions which are infinitely differentiable. When used in FLAME, such a kernel will produce very smooth estimates.

Example 3 (Exponential Kernel)**.** *The exponential kernel is on the other end of the "smoothness" spectrum compared to the Gaussian kernel, producing functions with only one derivative. In this case we have*

$$k(\mathbf{s}, \mathbf{s}') = \exp\left\{-\sigma|\mathbf{s} - \mathbf{s}'|\right\}.$$

This seemingly minor adjustment to the power in the exponent produces a space consisting of continuous functions with only one derivative. Using this kernel will typically produce "rougher" looking FLAME estimates than the Gaussian kernel. In practice, they are usually quite similar to the Sobolev kernel when $\mathcal{D} = [0,1]$ and $m = 1$, as the two produce equivalent norms (in a mathematical sense).

Example 4 (Periodic Kernel)**.** *A very useful feature of working with an RKHS is that one can incorporate structures such as periodicity and boundary conditions into the parameter estimates. This may be useful, for example, when the domain represents time over the course of a year. In that case, one might expect the parameters to be periodic. In this case one may use the periodic kernel with period $p = 1$ for yearly periodicity, $p = 1/2$ for semestral periodicity, or $p = 1/4$ for seasonal. The periodic kernel with period p is defined as*

$$k(\mathbf{s}, \mathbf{s}') = \sigma^2 \exp\left\{-2/\sigma \sin^2\left(\frac{\pi|\mathbf{s} - \mathbf{s}'|}{p}\right)\right\}.$$

More general boundary conditions can be worked into Sobelev spaces and norms, but we refrain from printing the details here, since we will not explore them in our simulations. An interested reader is referred to, for example, Section 4 of Chapter 7 in Berlinet and Thomas-Agnan [4] who list many examples of kernels that can work in different structures.

3. Implementation and computational details

In this section we develop a coordinate descent algorithm for quickly finding the FLAME estimator. Previous methods relied on rephrasing the problem so that group lasso for scalars could be used, and avoided developing a computational framework specifically for functional data. Those approaches work, however, one can obtain a more efficient algorithm which applies quite broadly by utilizing the functional structure of the problem. These methods are implemented in an accompanying `R` package `flm`. The computationally intensive functions in this package are coded in `C++`, so that the methodology can be implemented very quickly even for very large datasets.

The algorithm is based on utilizing functional subgradients so that, at each step, individual parameter estimates can be updated very quickly in a nearly closed form. An interested reader is referred to Boyd and Vandenberghe [5], Bauschke and Combettes [3], Barbu and Precupanu [2], Shor [30] for more details on subgradients and subdifferentials. Subgradients generalize derivatives (in this case Fréchet derivatives) to convex functionals (mappings from $\mathbb{H}$ to $\mathbb{R}$) which are not necessarily differentiable. At any point where the functional is differentiable, the two notions coincide, but subgradients are defined more broadly to convex functionals that need not be differentiable, and have become a staple of convex analysis. Let $f : \mathbb{H} \to \mathbb{R}$ be a convex functional. We say that $h \in \mathbb{H}$ is a subgradient of f at $x \in \mathbb{H}$ if for all $y \in \mathbb{H}$ we have $f(x+y) - f(x) \geq \langle h, y \rangle$. We denote by $\partial f(x)$ the collection of all subgradients of f at x, called the subdifferential. Trivially, x is a minimizer of f if and only if $0 \in \partial f(x)$. We show in the appendix that the subgradient for FLAME is given by

$$\frac{\partial}{\partial \beta_i} L_\lambda(\beta) = -\frac{1}{N}\sum_{n=1}^{N} X_{n,i} K(Y_n - X_n^\top \beta) + \lambda \tilde{\omega}_i \begin{cases} ||\beta_i||_{\mathbb{K}}^{-1}\beta_i, & \beta_i \neq 0 \\ \{h \in \mathbb{K} : ||h||_{\mathbb{K}} \leq 1\}, & \beta_i = 0 \end{cases}. \tag{1}$$

At each step of the coordinate descent we can use (1) to update our estimates. In particular, suppose that $\hat{\beta}$ is our current estimate and we aim to update the i^{th} coordinate, $\hat{\beta}_i$, treating all other coordinates as fixed. Since the X_i are assumed to be standardized, the least squares (in terms of the $\mathbb{H}$ norm) estimator would be

$$\check{\beta}_i = \frac{1}{N}\sum_{n=1}^{N} X_{n,i} E_n \qquad \text{where} \qquad E_n = Y_n - \sum_{j \neq i} X_{n,j}\hat{\beta}_j.$$

We can then express the subgradient as

$$\frac{\partial}{\partial \beta_i} L(\beta) = -K(\check{\beta}_i) + K(\beta_i) + \lambda \tilde{\omega}_i \begin{cases} ||\beta_i||_{\mathbb{K}}^{-1}\beta_i, & \beta_i \neq 0 \\ \{h \in \mathbb{K} : ||h||_{\mathbb{K}} \leq 1\}, & \beta_i = 0 \end{cases}.$$

Recall that a point $\hat{\beta}_i$ minimizes $L(\beta_i)$ (in the ith coordinate) if 0 is in the subdifferential at $\hat{\beta}_i$. We can thus use this expression to determine if $\hat{\beta}_i = 0$, which would mean that 0 is in the subdifferential when evaluated at $\beta_i = 0$.

Every subgradient at $\beta_i = 0$ is of the form $K(\check{\beta}_i) + \lambda w_i h$, where h is any function that satisfies $\|h\|_{\mathbb{K}} \leq 1$. Setting this equal to zero, we get $h = -(\lambda w_i)^{-1} K(\check{\beta}_i)$, which means 0 is in the subdifferential as long as

$$||K(\check{\beta}_i)||_{\mathbb{K}} \leq \lambda \tilde{\omega}_i \Longrightarrow \hat{\beta}_i = 0. \tag{2}$$

Note this also indicates a useful starting value of λ for the algorithm; if we take

$$\lambda = \max_{i=1,\ldots,I} \{\tilde{\omega}_i^{-1} \|N^{-1} \sum X_{ni} K(Y_n)\|_{\mathbb{K}}\}, \tag{3}$$

then the solution will always be $\hat{\beta}_i = 0$ when updating any coordinate. Since all quantities in (3) are known, we can compute this quantity to determine a starting value for λ, and then compute a sequence of solutions as we decrease λ. This is an interesting contrast with nonparametric smoothing, where it is often challenging to specify a range of tuning parameters that works in every setting. When $\hat{\beta}_i \neq 0$, we can solve for it in a nearly closed form. In particular, we have

$$-K(\check{\beta}_i) + K(\hat{\beta}_i) + \frac{\lambda \tilde{\omega}_i}{||\hat{\beta}_i||_{\mathbb{K}}} \hat{\beta}_i = 0 \quad \Longrightarrow \quad \hat{\beta}_i = \left(K + \frac{\lambda \tilde{\omega}_i}{\|\hat{\beta}_i\|_{\mathbb{K}}} I \right)^{-1} K(\check{\beta}_i). \tag{4}$$

The only unknown quantity at this point is $\|\hat{\beta}_i\|_{\mathbb{K}}$. Unfortunately, its expression does not have a closed form solution (unlike FLS or AFSL). However, if we take the $\mathbb{K}$-norm of the expression in (4) we arrive at the following scalar equation that can be solved numerically

$$1 = \sum_{j=1}^{\infty} \frac{\theta_j \langle \check{\beta}_i, v_j \rangle^2}{(\theta_j ||\hat{\beta}_i||_{\mathbb{K}} + \lambda \omega_i)^2}.$$

Our coordinate descent algorithm therefore proceeds iteratively, defining a sequence of $\beta^{(t)}$ for $t = 1, \ldots, T$ which converges to the desired approximation $\hat{\beta}$. We set the maximum number of iterations T and a stopping criteria based on the improvement in the estimation of the β coefficients (i.e. the $\mathbb{K}$-norm of the increment should be higher than a fixed tolerance).

Regarding the weights, $\tilde{\omega}_i$, we run the algorithm twice. The first one (*the non-adaptive step*) is run with weights set to 1, and the second time (*adaptive step*) we take $\tilde{\omega}_j = \|\hat{\beta}_{j,N}\|_{\mathbb{K}}^{-1}$ with $\|\hat{\beta}_{j,N}\|_{\mathbb{K}}$ the norm of the β estimated in the *non-adaptive step*. In particular the *adaptive step* is run to improve the statistical performance of the estimates for the significant predictors by reducing their bias; without the adaptive step one does not have the oracle property (Section 5). The algorithm is then run only on the non-zero predictors isolated in the *non-adaptive step*. These steps must be run for a sequence of λ and we have to identify a proper λ which maximizes some selection criterion; we choose λ to minimize the cross validation error, once we have isolated a training and a test set (randomly sampled as the 25% of the entire data set).

We mention two features we have built into the code which help increase its computational efficiency. The first is a *warm start*, which means when moving to

the next λ value, we use the previous $\hat{\beta}$ as the initial value for β. Since λ usually changes very little with each step, this means that the new $\hat{\beta}$ can be computed very quickly (usually with just a few iterations). In this way, one can obtain the solutions for an entire sequence of λ with only marginally more computation time than with a single λ, and is a commonly used device in optimization. The second feature is what we call a *kill switch*, which will stop the algorithm from considering subsequent λ values if the number of active predictors moves past a certain threshold. This allows the user to set the maximum size for the number of predictors selected by the model. When the algorithm moves past this threshold, no further λ are considered after calculating its next parameter estimate (so the final estimate might have a few more predictors than specified by the kill switch). In certain applications, one can make a good guess as to the maximum number of predictors that could conceivably be selected by the model. In these settings, this bound can be used for the *kill switch*. For example, most genetic studies involving complex traits, even with hundreds of thousands of predictors, result in only a handful of significant and reproducible associations. If one takes the kill switch too small, then the number of predictors chosen in the final solution will be very close to number indicated in the kill switch, and thus the user should increase the kill switch and then rerun the algorithm. However, in practice this switch need not be used if there is no previous knowledge about the expected sparsity of the solution. The algorithm slows down as more predictors enter the model, thus this has the potential to provide a substantial computational savings.

Lastly, all functional data methods of this type require some preprocessing of the raw data into functional units. This is now a fairly well developed step and a more detailed discussion can be found in [18]. For implementing FSL and AFSL we utilize a penalized cubic B-splines expansion, where the penalty is chosen by generalized cross validation. The number of B-splines in our simulations and application is taken to be 100 so that the smoothing is determined entirely by the penalty. We then rotate to the FPCA basis so that less that 100 basis functions can be used, thus gaining computational efficiency. However, for FLAME, we take a slightly different approach and instead only use the eigenfunctions of the kernel K, which we compute numerically on a fine grid. This allows us to quickly compute both $\mathbb{H}$ norms and $\mathbb{K}$ norms. We choose the number of basis functions, J, so that $\sum_{j=1}^{J} \theta_j \geq 0.99 \sum_{j=1}^{\infty} \theta_j$, where θ_j are the eigenvalues of the kernel K. This formulation is similar to explaining 99% of the variability in FPCA. We use such a high mark because dimension reduction is not our goal; we aim to approximate the data nearly exactly.

4. Empirical study

In this section we introduce several simulation schemes to analyze the performance of FLAME with different RKHS (Section 4.1) and to compare this method with AFSL and MCP (Section 4.2). We conclude (Section 4.3) with an application to a large genetic dataset. For all simulations we assume the data is

in $L^2[0,1]$. The kernels we consider are three popular kernels, the Exponential, the Sobolev, and the Gaussian. In Section C of the appendix we show additional results using a periodic kernel. In Figure 7 of the appendix, the first four eigenfunctions associated to the Exponential, the Sobolev, and the Gaussian kernel are plotted and the explained variance is shown. These three kernels show different structure and complexity; in Section 4.1 the consequences of the different smoothness levels required to functions embedded in these kernels are presented.

All simulations used 100 runs on a Intel quad-core i7 desktop with 8GB of ram with the vecLib linear algebra library of R and measured in terms of:

- *computational time*: median of the computational time (sec.) of the runs.
- *number of true positive predictors*: average number of correctly non-zero predictors identified (i.e. $\#\{i : \beta_i^\star \neq 0 \wedge \hat{\beta}_i \neq 0\}$).
- *number of false positive predictors*: average number of wrongly identified non-zero predictors (i.e. $\#\{i : \beta_i^\star = 0 \wedge \hat{\beta}_i \neq 0\}$).
- *prediction error*: average of the prediction error, both for data and derivatives,
  $\sum_{n=1}^N \| \mathbf{X}_n\beta^\star - \mathbf{X}_n\hat{\beta} \|_{L^2}$ and $\sum_{n=1}^N \| \mathbf{X}_n\beta^{\star\prime} - \mathbf{X}_n\hat{\beta}' \|_{L^2}$.

Lastly, we implement FLAME by first running its non-adaptive version ($\tilde{\omega}_i \equiv 1$) and then use the resulting estimates in a second adaptive step. In Appendix D we provide two additional sets of simulations. The first set repeats one of our settings using FSL to compute the weights, but using FLAME for the adaptive step (Figure 5). The results end up being nearly the same, indicating that the procedure is not too sensitive to how the adaptive weights are chosen. The second set of additional simulations avoids using the kill switch and runs the algorithim for all of the specified λ values (Figure 6 and Table 2). The result is almost identical in terms of statistical performance, but the computation time increases by a factor of 5-20 depending on the setting.

4.1. Comparison between different kernels

In this section we compare the performance of FLAME using different kernels. We show how the variation of the kernel can influence the identification of the number of correctly identified predictors and the prediction error. Two high-dimensional simulation settings are introduced: with rough and smooth $\beta^\star$ coefficients.

The simulations consist of the random generation of a sample of size $N = 500$ and $I = 1000$ predictors, with $I_0 = 10$ significant ones. The predictor matrix $\mathbf{X}$ is the standardized version of a matrix randomly sampled from a N dimension Gaussian distribution with 0 mean and identity covariance matrix. For the rough case, the true coefficients $\beta^\star(t)$ are sampled from a Matérn process with 0 average and parameters ($\nu = 2.5, \text{range} = 1/4, \sigma^2 = 1$), while for the smooth setting the range parameter of the Matérn process is set to 1 and ν is set to 3.5. In Figure 8 of the appendix an example of the true coefficients in the two settings is shown. The outcomes, $Y_n(t)$, are obtained as the sum of the

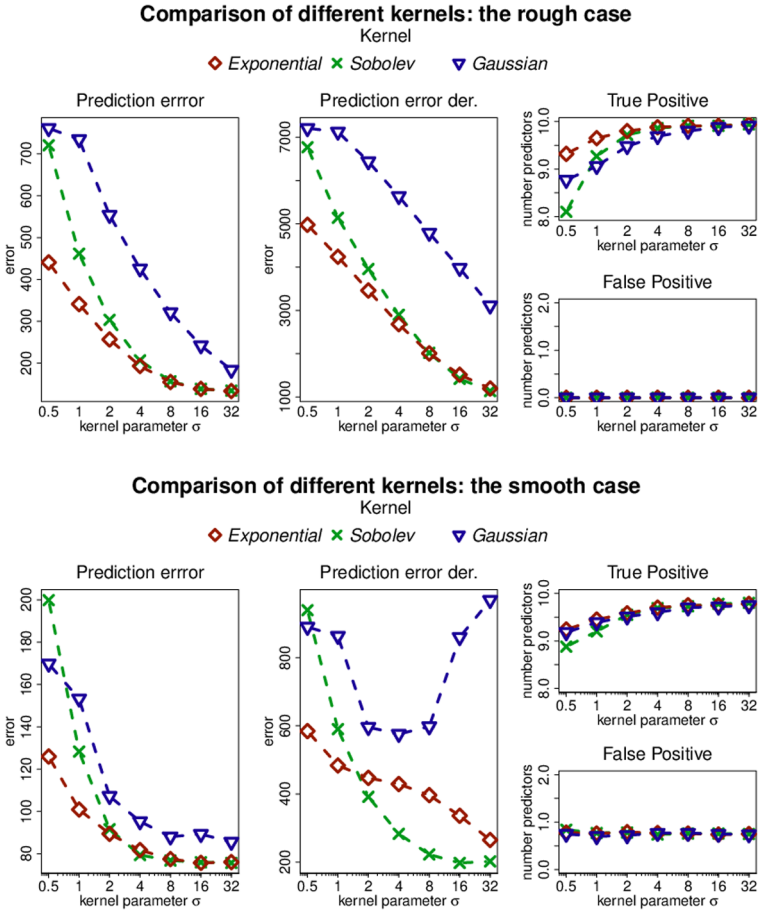

FIG 1. *Summary of the simulations varying kernel for the rough (top) and smooth (bottom) cases. From the left, the prediction error, the prediction error on derivatives, and the number of true and false positive predictors.*

contribution of all the predictors and a random noise, a 0-mean Matérn process with parameters $(\nu = 1.5, \text{range} = 1/4, \sigma^2 = 1)$. Functions are sampled on an evenly spaced grid between 0 and 1 with $m = 50$ points.

For these simulations the *kill switch* parameter is set to $2I_0 = 20$ and λ is chosen via cross-validation from a 100-point gird of values that are equally spaced on the logarithmic scale. The range of this grid is from $\lambda_{\max}$, as given in (3), to $r_\lambda \lambda_{\max}$ with $r_\lambda = 0.01$ for the rough case and $r_\lambda = 0.001$ for the smooth setting. A summary of the results is shown in Figure 1 for both the rough and smooth cases.

Focusing on the rough setting we notice that the Gaussian kernel always performs worse than other kernels in terms of prediction error both for data and derivatives: it imposes on the functions a structure (infinitely differentiable) they

don't possess. Moreover, increasing the σ parameter of the kernels, which results in a rougher estimates, reduces the prediction error and more true non zeros predictors are identified. In fact, with a too strong smoothness level, imposed by the Gaussian kernel or by a small value for the σ parameter, some true predictors are forced to be zero throughout the domain and this reduces the number of true positives and increases the prediction error. The rough structure of the parameters allows all methods considered to avoid the identification of non significant predictors as the number of false positives is always zero.

A slightly different behavior can be observed in the smooth case. The performance of the Gaussian kernel, while still worse, is now much closer in performance to the other two kernels. The strange behavior of the prediction error of derivatives for the gaussian and the exponential kernel is due to an instability in the estimation of the derivatives of the eigenfunctions of these kernels at the boundaries of the time domain (not shown here). The number of false positives in this setting is different from zero (but it remains on average smaller than one per simulation).

A final remark regarding the high dimensional setting is the computational cost of the estimation and variable selection procedure. As presented in Table 1 in appendix, the computational time is almost invariant with respect to the kernel and parameter, while increasing the smoothness level of the predictors increases the computational time. In the next section we present how competitive FLAME is compared to different methods.

4.2. Comparison with previous methods

4.2.1. The high dimensional setting

In this section we apply AFLS to the simulation setting we've introduced in Section 4.1 and in Table 4 of the appendix we present the results of AFSL estimation in terms of prediction error, computation time and number of predictors identified (true positives and false positives).

An advantage of FLAME is the reduction of the computation time: FLAME takes much less than AFSL to run and it also achieves better statistical performance. Mainly in the rough case, the Exponential and the Sobolev kernel (with $\sigma > 1$) perform better in terms of prediction error on data, derivatives and in the number of true positive and false positive predictors.

4.2.2. The small dimensional setting

In this section we reduce the simulation size to make the application of MCP possible; this method is suitable just for $N > I$ schemes. We present the results of FLAME, MCP, and AFSL with the same rough and smooth settings introduced in Section 4.1, but with $N = 50$, $I = 20$ and $I_0 = 5$. Moreover we focus on the number of points per curve m to detect whether these three methods are

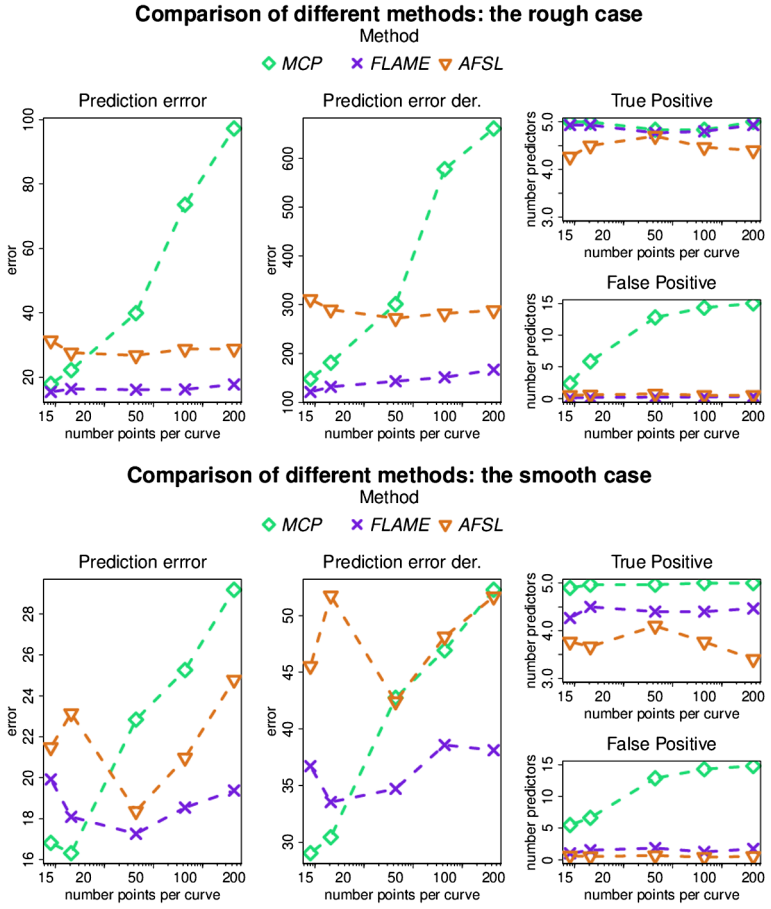

FIG 2. *Summary of the simulations varying the method for the rough (top) and smooth (bottom) cases. From the left, the prediction error, the prediction error on derivatives, and the number of true and false positive predictors.*

affected by m. For FLAME we focus on the Sobolev kernel with $\sigma = 8$, since, from Section 4.1, it is shown to be a suitable kernel for both these two settings.

In Figure 2 the results for the three methods varying m are shown. We notice that both FLAME and AFSL estimations are almost invariant with respect to m, while MCP is strongly affected by variations of m, becoming very unreliable when the number of points per curve is large. However, if the number of points is small, MCP performs better than FLAME and AFSL in terms of prediction error and selecting true predictors, mainly in the smooth setting, but still often has trouble in terms of false positives. Focusing on the computational efficiency, presented in Table 3 of the appendix we notice that FLAME and AFSL are comparable, with the well known higher efficiency of FLAME in the rough case with respect to the smooth, and they both are almost invariant with the change

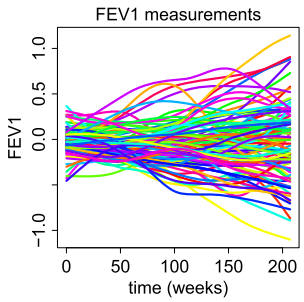

FIG 3. *FEV1 curves of* 100 *randomly selected children measured on 4 years of follow up. The contribution of age, gender and treatment have already been removed.*

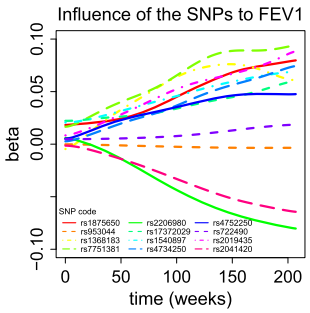

FIG 4. *Coefficients of the influent SNPs detected and estimated by FLAME.*

of m. They globally perform significantly better than MCP, which in addition becomes slower and slower with the increase of m. The difference in the efficiency of FLAME and AFSL is due to the method used to solve the problem: the coordinate descent method of FLAME is faster than ADMM of AFSL in the high dimensional setting since it is not based on matrix algebra operations, while in the small setting both coordinate descent and ADMM are efficient.

4.3. Childhood asthma management program

In this section we present the application of FLAME to a large genetic dataset collected from [33]. The Childhood Asthma Management Project, CAMP, is a longitudinal trial to analyze the longterm impacts of several daily treatments on children with asthma. It includes 439 Caucasian children, ages 5-12, affected by asthma and monitored for 4 years. These data are freely available from the dbGaP, Study Accession phs000166.v2.p1 ([10]).

Genotypic informations consists of approximately $670,000$ SNPs with minor allele frequency larger than 5%. We first apply a screening tool from [9] to isolate a subset of $I = 10,000$ SNPs, on which we apply FLAME. The focus of our analysis is, then, the detection of the significant SNPs among these $10,000$.

Each child is given one of three treatments: Budesonide, Nedocromil, or a placebo. We account for age at the beginning of the study and gender. To quantify the lung strength of children we consider 16 longitudinal measurements of the Forced Expiratory Volume in one second (FEV1), a common proxy for lung strength. The lung capacity is the response function of our linear model and we convert it into a functional data object with a cubic B-spline basis projection with penalty on the second derivative and smoothing parameter chosen via generalized cross-validation. As a first preprocessing step we remove the influence of gender, age, and treatment from FEV1 and then we apply FLAME to evaluate the impact of the SNPs to the residual functions shown in Figure 3. In Figure 4 the FLAME estimation is presented; for this analysis we use the Sobolev kernel with $\sigma = 8$, a 200 points grid for λ with the ratio $r_\lambda = 0.01$. We identify the presence of 12 significant SNPs, 9 with a positive effect in the lung development and 3 (rs2206980, rs2041420 and rs953044) with a negative contribution. Table 6 of the appendix lists the identified SNPs with the comparison with the ones identified by AFSL: we notice that FLAME identifies two more SNPs, one with positive effect (rs722490) and one with negative effect (rs2041420). While the additional positive effect is fairly small, the negative effect is actually quite sizable, making it a bit surprising that AFSL missed it.

To add a further comparison with AFSL we identify a test (80% of data) and a training set (20%) to compute the prediction error of data as $\sum_{n=1}^{N} \| Y_n - \mathbf{X}_n \hat{\beta} \|_{L^2}$. We replicate this analysis 10 times, to mimic a 10-fold cross-validation, to present a robust conclusion. The average prediction error for FLAME is 0.200, while for AFSL is slightly higher at 0.205. Moreover measuring the computational time we have for FLAME a median of 172.01 sec. and for AFSL 365.07 sec. showing the advantage of FLAME in terms of computational time, with also a slight improvement in term of prediction error.

As a last point, the SNP selected by FLAME but not by AFSL, rs2041420, is located on the gene MACROD2. This gene has been associated with a number of negative health outcomes including Autism, Celiac disease, Crohn's disease, and Parkinson's disease (`http://www.gwascentral.org`). It has also been linked to FEV1 and lung development [32, 29]. However, neither of these previous studies were based on CAMP, and therefore helps validate this finding.

5. Theoretical properties

In this section we provide several theoretical guarantees for FLAME. While this theory provides a strong justification for using FLAME, there are still several interesting theoretical questions which remain open and will be discussed below. We begin by making the following assumption concerning the various terms in the model. Very similar assumptions can also be found in Fan and Reimherr [13].

Assumption 2. *The regression problem satisfies the following.*

1. ***Minimum Signal:*** *Let* $b_N = \min_{i\in\mathcal{S}} \|K(\beta_i^\star)\|_{\mathbb{K}}$*, then we assume that* $b_N^2 \gg I_0^2 \log(I) N^{-1}$.
2. ***Tuning Parameter:*** *The tuning parameter* λ *satisfies the following lower and upper bounds* $I_0^{1/2}\log(I)N^{-1} \ll \lambda \ll b_N I_0^{-1/2} N^{-1/2}$.
3. ***Design Matrix:*** *Let* $\hat{\Sigma}_{11} = N^{-1}\mathbf{X}_1^\top \mathbf{X}_1$*, be the design matrix for only the true predictors. We assume the minimum eigenvalue* $\sigma_{\min}(\hat{\Sigma}_{11})$ *and maximum eigenvalue* $\sigma_{\max}(\hat{\Sigma}_{11})$ *satisfy:* $\nu_1^{-1} \leq \sigma_{\min}(\hat{\Sigma}_{11}) \leq \sigma_{\max}(\hat{\Sigma}_{11}) \leq \nu_1$.
4. ***Irrepresentable Condition*** *Let* $\hat{\Sigma}_{21} = N^{-1}\mathbf{X}_2^\top \mathbf{X}_1$*, be the cross covariance between the null and true predictors. We assume that* $\|\hat{\Sigma}_{21}\hat{\Sigma}_{11}^{-1}\|_{op} \leq \phi < 1$*, where* ϕ *is a fixed scalar and* $\|\cdot\|_{op}$ *the operator norm.*

The first assumption is called a minimum signal condition and indicates the minimum magnitude (of the signals) required for detecting the relevant predictors. Notice that this condition is placed on $\beta^\star$ relative to K, which means that if K wipes out a signal, the algorithm will not be able to detect it. The second condition concerns the rate for λ, and takes a fairly familiar form [1, 13]. Since our FLAME formulation normalizes the sum of squares by N, the λ needs to tend to zero. The lower bound, indicates that it cannot go to zero too quickly, otherwise one cannot guarantee that all of the null predictors are dropped. Conversely, the upper bound actually indicates two things, first if λ goes to zero too slowly then some of the significant predictors may also be dropped. Second, the upper bound on λ also ensures the bias is asymptotically negligible for establishing an oracle property. The third condition on the design matrix simply says that the design matrix for the true predictors, must be well behaved. This ensures that the oracle estimate as well as the FLAME estimate are well behaved when restricted to the set of true predictors. The last condition is interpreted as requiring that the true predictors and the null predictors are not too correlated. This condition is essentially necessary to obtain an oracle property [39].

Under these conditions, we can now state our primary theorem, which states that FLAME recovers the true support with probability tending to 1, and that its projections are asymptotically normal.

Theorem 1. *If the regression problem satisfies Assumptions 1 and 2, the solution of the FLAME problem,* $\hat{\beta}$*, asymptotically*

1. *has the same support of the true solution of the regression problem*

$$P(\hat{\beta} \overset{\text{S}}{=} \beta^\star) \to 1,$$

2. *and is equivalent to the Oracle estimator in the sense that, for any sequence* $h_n = \{h_{i,n}\} \in \mathbb{K}^I$ *that satisfies* $\|h_n\|_{\mathbb{K}} \leq M_1$ *and* $\sum \|C^{1/2}h_{i,n}\|_{\mathbb{H}}^2 \geq M_2 > 0$ *we have*

$$\frac{\sqrt{N}\langle \hat{\beta} - \beta^\star, h_n\rangle}{\sigma_n} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1) \quad \textit{where} \quad \sigma_n^2 = \sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}} \hat{\Sigma}_{11;ij}^{-1}\langle C^{1/2}h_i, C^{1/2}h_j\rangle,$$

and $\mathcal{S}$ *denotes support of* $\beta^\star$.

The first part of the theorem is a fairly standard result; we are showing that our method is variable selection consistent. The second result shows that the estimators are consistent and are asymptotically normal, but there is a serious caveat to this, namely the projections are normal only when projected onto an element of $\mathbb{K}$, not $\mathbb{H}$. If the Y_n were finite dimensional, then the two would be equivalent, but not in the functional setting.

In the context of functional data, we call Theorem 1 a *weak oracle property* because the normality occurs in the weak topology (i.e. on projections). Such results are not uncommon in functional data analysis [7]. Our next result shows that one can actually obtain a stronger result, namely, that the FLAME and oracle estimates are asymptotically equivalent in the *strong* topology. For this reason, we say that the following theorem is a *strong oracle property*. First let us define the oracle estimate, namely

$$\hat{\beta}_O = \{(\mathbf{X}_1^\top \mathbf{X}_1)^{-1}\mathbf{X}_1^\top Y, 0\},$$

where 0 a vector of zero functions of length $I - I_0$.

Theorem 2. *Suppose Assumptions 1 and 2 are satisfied, but that I_0 is fixed. Furthermore, assume there exists a $\delta > 0$ and a constant $0 < B < \infty$ such that for all $i \in \mathcal{S}$*

$$\sum_{j=1}^{\infty} \frac{\langle \beta_i^\star, v_j \rangle^2}{\theta_j^{1+\delta}} \leq B < \infty.$$

If λ is such that

$$\lambda \ll \frac{b_N}{N^{1/2[1+1/(1+\delta)]}},$$

then one also has that

$$\sqrt{N}\|\hat{\beta} - \hat{\beta}_O\|_{\mathbb{H}} = o_P(1).$$

Notice that we have introduced slightly stronger assumptions to achieve a strong oracle property. In particular, we needed a more explicit assumption on the rate at which the coordinates of $\beta^\star$ decrease. If $\delta = 0$ this simply implies that $\beta^\star$ is in $\mathbb{K}$. Lastly, we require a tighter control of the λ which depends on how quickly the coordinates of $\beta^\star$ decrease. If the coordinates actually terminate (i.e. are zero) at a certain point or if they decrease exponentially fast, then our assumption on λ is the same as before. The assumption that I_0 is fixed allows us to simplify the results. Using our techniques it is possible to allow I_0 to grow, but we would need additional assumptions on the behavior of the trace of the covariance operator of the errors with respect to the $\{v_i\}$ basis, and so do not pursue it here.

6. Conclusions and future work

In this work we have provided a new tool for simultaneous variable selection, parameter estimation, and parameter smoothing for function-on-scalar regression. We have provided theoretical guarantees as well as an efficient computational

algorithm for carrying out our procedure. While other methods are available for selection and estimation, none of them incorporate smoothing as well. While our work is fairly complete, there are still many opportunities for improvement.

The first area from improvement concerns the penalty. Our approach is convenient in that we have a fairly simple penalty that has only one tuning parameter. The downside is that we have to use this parameter to both select parameters and smooth their estimates. However, there is no reason to think that the same parameter is optimal for accomplishing both, and in fact, the tendency for the algorithm to want to choose larger σ values in the kernels seems to indicate this, namely, the choice of λ is driven more by selection than smoothing. We believe that altering the penalty can alleviate much of this, especially if one allows for a second tuning parameter (one for selection and one for smoothing).

The second area for improvement involves the coordinate descent algorithm we employed. This approach works fairly well, especially when combined with *RCPP*, however, there are at least two reasons to consider other approaches. The first is that the coordinate descent does not have closed form updates (they are *almost* closed form). We believe that by using something like ADMM, this problem could be avoided since one can "decouple" the least squares term and penalty. The second reason is that the optimization community is actively researching tools for solving very large sparse problems, which could potentially allow one to handle millions of predictors.

The last key area we see for improvement concerns our statistical theory. We believe that our results can be tightened, especially the additional assumptions needed to achieve Theorem 2. Maybe the major obstacle is obtaining a good control of $\|\hat{\beta}\|_{\mathbb{K}}$. This quantity shows up when updating via coordinate descent and when trying to control the bias of the FLAME estimate. However, unlike FSL, we do not have a closed form expression for this quantity in terms of the least squares estimator. If one can obtain a tighter control of this quantity, it should be easier to relax the assumptions of Theorem 2. Lastly, it might be interesting to study the asymptotic properties of $\hat{\beta}$ under the $\mathbb{K}$ norm, instead of the $\mathbb{H}$ norm. For example, it might be of interest to study the estimated derivatives of the parameters. However, since this is a much stronger norm, clearly additional assumptions will be needed. Furthermore, the oracle estimate would not be the least squares estimator as this need not even live in the space $\mathbb{K}$. We thus believe there are many open and exciting questions concerning the behaviors of such functional estimators and their necessary assumptions.

Appendix A: Subgradient equations for FLAME

Before deriving (1) we state the following Lemma which can found in any of the discussed references on convex analysis.

Lemma 1. *Let $f_1 : \mathbb{H} \to \mathbb{R}$ be $f_2 : \mathbb{H} \to \mathbb{R}$ be two convex functionals over a real separable Hilbert space $\mathbb{H}$. Then we have the following.*

1. *If the Fréchet derivative of f_1 exists at a point $x \in \mathbb{H}$, then the subdifferential of f_1 at x consists of single point which is the derivative of f_1 at x.*

2. *The subdifferential of $f_1 + f_2$ is the sum of their respective subdifferentials: $\partial(f_1 + f_2) = \partial f_1 + \partial f_2$. Where the sum is understood as the Minkowski sum between two sets.*

We now state a lemma concerning the subdifferential of the norm and the norm squared which follows from standard arguments, see for example Section A.1 of Fan and Reimherr [13].

Lemma 2. *1. Consider the functional $f(x) = \|x\|_{\mathbb{H}}^2$. Then f is convex and everywhere differentiable with*

$$\partial f(x) = 2x.$$

2. Consider the functional $f(x) = \|x\|_{\mathbb{H}}$. Then f is convex and differentiable when $x \neq 0$ in which case

$$\partial f(x) = \|x\|_{\mathbb{H}}^{-1} x \qquad x \neq 0.$$

When $x = 0$ we have

$$\partial f(0) = \{x \in \mathbb{H} : \|x\| \leq 1\}.$$

Proof. The proof of the second claim can be found in Section A.1 of [13]. To see the first part consider

$$f(x+h) = \|x+h\|_{\mathbb{H}}^2 = \|x\|^2 + 2\langle x, h\rangle_{\mathbb{H}} + \|h\|_{\mathbb{H}}^2 \geq f(x) + \langle 2x, h\rangle_{\mathbb{H}}.$$

Thus, by the definition of the subgradient, it follows that $2x$ is the subgradient of $f(x) = \|x\|_{\mathbb{H}}^2$. □

We now derive the FLAME subgradient equations. First, we rewrite them using a common norm:

$$L(\beta) = \frac{1}{2N}\|K^{1/2}(Y - \mathbf{X}\beta)\|_{\mathbb{K}}^2 + \lambda \sum_{i=1}^{I} \tilde{\omega}_i \|\beta_i\|_{\mathbb{K}}.$$

So L is a convex function from $\mathbb{K}^I \to \mathbb{R}$. Here it is also understood that $K^{1/2}(Y)$ is applied coordinate wise to each function. Since $\mathbb{K}$ is a real separable Hilbert space we have by Lemma 2.1 and the chain rule that

$$\frac{\partial}{\partial \beta_i}\frac{1}{2N}\|K^{1/2}(Y - \mathbf{X}\beta)\|_{\mathbb{K}}^2 = -\frac{1}{N}\sum_{n=1}^{N} X_{n,i}(K^{1/2}(Y - \mathbf{X}\beta)).$$

By Lemma 2.2 we have that

$$\frac{\partial}{\partial \beta_i}\lambda \sum_{j=1} \tilde{\omega}_j \|\beta_j\|_{\mathbb{K}} = \lambda \tilde{\omega}_j \begin{cases} \|\beta_j\|_{\mathbb{K}}^{-1}\beta_j & \beta_j \neq 0 \\ \{h \in \mathbb{H} : \|h\|_{\mathbb{K}} \leq 1\} & \beta_j = 0 \end{cases}.$$

Applying Lemma 1 we can combine the two subdifferentials to obtain (1).

Appendix B: Proofs

The following two lemmas follow from Theorem 4 and Lemma 9 (respectively) of Barber et al. [1].

Lemma 3. *If Assumption 2 holds, the FSL estimate $\tilde{\beta}$, computed with all the weights set to 1, satisfies*

$$sup_{\imath\in\mathcal{S}}\|\beta_i^\star - \tilde{\beta}_i\|_{\mathbb{H}} = O_P(r_N^{1/2}) \qquad where \qquad r_N = \frac{\log(I)I_0}{N}.$$

Lemma 4. *Let X be an $\mathbb{H}$ valued Gaussian process with mean zero and covariance operator C. Then we have the bound*

$$P\left\{\|X\|_{\mathbb{H}}^2 \geq \|C\|_1 + 2\|C\|_2\sqrt{t} + 2\|C\|_\infty t\right\} \leq \exp(-t)$$

where $\|C\|_1$ the sum of the eigenvalues of C, $\|C\|_2^2$ the sum of the squared eigenvalues and $\|C\|_\infty$ the largest one.

Corollary 1. *Given the Gaussian process X, with zero mean and covariance operator C, and given the kernel operator K (represented by the eigenvalues θ_j: $\theta_1 \geq \theta_2 \geq \ldots \geq 0$, and the eigenvectors v_j which define an orthogonal basis for $\mathbb{H}$ and $\mathbb{K}$), we have that*

$$P\left\{\|K(X)\|_{\mathbb{K}}^2 \geq \theta_1(\|C\|_1 + 2\|C\|_2\sqrt{t} + 2\|C\|_\infty t)\right\} \leq \exp(-t)$$

Proof From the definition of the $\mathbb{K}$ norm and recalling that v_j is an eigenfunction of K with eigenvalue θ_j, we obtain that

$$\|K(X)\|_{\mathbb{K}}^2 = \sum_{j=1}^{\infty}\frac{\langle\theta_j X, v_j\rangle^2}{\theta_j} = \sum_{j=1}^{\infty}\theta_j\langle X, v_j\rangle^2 \leq \theta_1\sum_{j=1}^{\infty}\langle X, v_j\rangle^2 = \theta_1\|X\|_{\mathbb{H}}^2.$$

We then apply Lemma 4 to obtain

$$P\left\{\|K(X)\|_{\mathbb{K}}^2 \geq \theta_1(\|C\|_1 + 2\|C\|_2\sqrt{t} + 2\|C\|_\infty t)\right\} \leq$$
$$P\left\{\|X\|_{\mathbb{H}}^2 \geq \|C\|_1 + 2\|C\|_2\sqrt{t} + 2\|C\|_\infty t\right\} \leq \exp(-t),$$

as desired. ■

Proof of Theorem 1.1

We begin by partitioning the set of the estimated parameters into $\hat{\mathcal{S}}$ and $\hat{\mathcal{S}}^C$ where

$$\hat{\mathcal{S}} = \left\{i \in \{1, \ldots, I\} : \hat{\beta}_i \neq 0\right\}.$$

Our aim for this section is then to prove that, with high probability, $\mathcal{S} = \hat{\mathcal{S}}$, that is $\hat{\beta}$ has $\mathcal{S}$ as support.

Suppose, for the moment, that $\hat{\mathcal{S}} = \mathcal{S}$, then from the subgradient equation (1) we have that

$$\mathbf{X}_1^\top K\left(Y - \mathbf{X}_1\hat{\beta}_1\right) = \lambda\tilde{s}_1 \qquad \text{where} \qquad \tilde{s}_1 = \left\{N\tilde{\omega}_i\hat{\beta}_i\|\hat{\beta}_i\|_{\mathbb{K}}^{-1} : \ i \in \mathcal{S}\right\}, \quad (5)$$

and $\hat{\beta}_1 = \{\hat{\beta}_i : \ i \in \mathcal{S}\}$ is the estimate of the non-zero predictors. This then implies that

$$K(\hat{\beta}_1) = \left(\mathbf{X}_1^\top\mathbf{X}_1\right)^{-1}\left(\mathbf{X}_1^\top K(Y) - \lambda\tilde{s}_1\right) = K(\beta_1^\star) + \left(\mathbf{X}_1^\top\mathbf{X}_1\right)^{-1}\left(\mathbf{X}_1^\top K(\varepsilon) - \lambda\tilde{s}_1\right).$$

To prove that $\beta^\star$ and $\hat{\beta}$ have the same support ($\mathcal{S} = \hat{\mathcal{S}}$) we have to verify the following.

- If $i \in \mathcal{S}$, $\hat{\beta}_1 \overset{\text{S}}{=} \beta_1^\star$, i.e. the true non-zero predictors are correctly identified. This condition can be also written as
$$\|K(\beta_i^\star) - K(\hat{\beta}_i)\|_{\mathbb{K}} < \|K(\beta_i^\star)\|_{\mathbb{K}}. \quad (6)$$
- If $i \notin \mathcal{S}$, $\hat{\beta}_i$ is set to zero, so that the zero predictors are correctly detected. That means
$$\left\|\frac{1}{N}\mathbf{X}_i^\top K\left(Y - \mathbf{X}_1\hat{\beta}_1\right)\right\|_{\mathbb{K}} < \lambda\tilde{\omega}_i \quad (7)$$

To achieve a better definition of (6) and (7) we introduce the definition of Y and find, for all $i \in \mathcal{S}$

$$\begin{aligned}&\|K(\beta_i^\star) - K(\hat{\beta}_i)\|_{\mathbb{K}} < \|K(\beta_i^\star)\|_{\mathbb{K}}\\ &\quad\Longrightarrow \left\|e_i^\top\left[N^{-1}\hat{\Sigma}_{11}^{-1}(\mathbf{X}_1^\top K(\varepsilon) - \lambda\tilde{s}_1)\right]\right\|_{\mathbb{K}} < \|K(\beta_i^\star)\|_{\mathbb{K}}\end{aligned}$$

with e_i a I-size vector with all zero coefficient but the i^{th} which is 1 and $\hat{\Sigma}_{11}$ the estimated covariance matrix of $\mathbf{X}_1$: $\hat{\Sigma}_{11} = N^{-1}\mathbf{X}_1^\top\mathbf{X}_1$. While, for all $i \notin \mathcal{S}$

$$\begin{aligned}&\left\|\frac{1}{N}\mathbf{X}_i^\top K\left(Y - \mathbf{X}_1\hat{\beta}_1\right)\right\|_{\mathbb{K}} < \lambda\tilde{\omega}_i\\ &\quad\Longrightarrow \left\|\mathbf{X}_i^\top N^{-1}\left[HK(\varepsilon) + \lambda\mathbf{X}_1(\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\tilde{s}_1\right]\right\|_{\mathbb{K}} < \lambda\tilde{\omega}_i\end{aligned}$$

with $H = (I - \mathbf{X}_1(\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\mathbf{X}_1^\top)$.
Considering the event $\left\{\mathcal{S} = \hat{\mathcal{S}}\right\}$, we observe that

$$\left\{\mathcal{S} \neq \hat{\mathcal{S}}\right\} \subset B_1 \cup B_1 \cup B_3 \cup B_4$$

with

$$\begin{aligned}B_1 &= \left\{\frac{1}{N}\|e_i^\top\hat{\Sigma}_{11}^{-1}\mathbf{X}_1^\top K(\varepsilon)\|_{\mathbb{K}} \geq \frac{\|K(\beta_i^\star)\|_{\mathbb{K}}}{2} : \ \text{for some } i \in \mathcal{S}\right\}\\ B_2 &= \left\{\frac{\lambda}{N}\|e_i^\top\hat{\Sigma}_{11}^{-1}\tilde{s}_1\|_{\mathbb{K}} \geq \frac{\|K(\beta_i^\star)\|_{\mathbb{K}}}{2} : \ \text{for some } i \in \mathcal{S}\right\}\end{aligned}$$

$$B_3 = \left\{ \frac{1}{N}\|\mathbf{X}_i^\top H K(\varepsilon)\|_{\mathbb{K}} \geq \frac{\lambda\tilde{\omega}_i}{2} : \text{ for some } i \notin \mathcal{S} \right\}$$
$$B_4 = \left\{ \frac{1}{N^2}\|\mathbf{X}_i^\top \mathbf{X}_1 \hat{\Sigma}_{11}^{-1}\tilde{s}_1\|_{\mathbb{K}} \geq \frac{\tilde{\omega}_i}{2} : \text{ for some } i \notin \mathcal{S} \right\}.$$

We will show that with N increasing $P(B_l) \to 0$ for $l = 1, \dots 4$ and then $P(\hat{\mathcal{S}} \neq \mathcal{S}) \to 0$.

Step 1: $P(B_1) \to 0$

Recall that

$$B_1 = \left\{ \frac{1}{N}\|e_i^\top \hat{\Sigma}_{11}^{-1}\mathbf{X}_1^\top K(\varepsilon)\|_{\mathbb{K}} \geq \frac{\|K(\beta_i^\star)\|_{\mathbb{K}}}{2} : \text{ for some } i \in \mathcal{S} \right\}.$$

We can express $B_1 = \cup_{i\in\mathcal{S}} A_i$ where

$$A_i = \left\{ \frac{1}{N}\|e_i^\top \hat{\Sigma}_{11}^{-1}\mathbf{X}_1^\top K(\varepsilon)\|_{\mathbb{K}} \geq \frac{\|K(\beta_i^\star)\|_{\mathbb{K}}}{2} \right\}$$
$$= \left\{ \frac{1}{N^2}\|e_i^\top \hat{\Sigma}_{11}^{-1}\mathbf{X}_1^\top K(\varepsilon)\|_{\mathbb{K}}^2 \geq \frac{\|K(\beta_i^\star)\|_{\mathbb{K}}^2}{4} \right\}.$$

This then implies that that $P(B_1) \leq \sum_{i\in\mathcal{S}} P(A_i)$. Thus if we can find a suitable bound for $P(A_i)$ we can show that $P(B_1) \to 0$. For each i we have that

$$\frac{1}{N^2}\|e_i^\top \hat{\Sigma}_{11}^{-1}\mathbf{X}_1^\top K(\varepsilon)\|_{\mathbb{K}}^2 = \|K(T_i)\|_{\mathbb{K}}^2$$

where $T_i = N^{-1}e_i^\top \hat{\Sigma}_{11}^{-1}\mathbf{X}_1^\top \varepsilon$ is a Gaussian process (in $\mathbb{H}$) with zero mean and covariance operator C_T

$$C_T = N^{-1}e_i^\top \hat{\Sigma}_{11}^{-1}\mathbf{X}_1^\top \mathbf{X}_1 \left(\hat{\Sigma}_{11}^{-1}\right)^\top e_i N^{-1} C$$
$$= N^{-1}e_i^\top \hat{\Sigma}_{11}^{-1} N\hat{\Sigma}_{11}\hat{\Sigma}_{11}^{-1} e_i N^{-1} C = N^{-1}e_i^\top \hat{\Sigma}_{11}^{-1} e_i C.$$

Recall that C is the covariance operator of the errors, ε_i. Now that we have the form of the C_T, we can apply Corollary 1 to obtain

$$P\left\{ \|K(T_i)\|_{\mathbb{K}}^2 \geq \theta_1 N^{-1} e_i^\top \hat{\Sigma}_{11}^{-1} e_i (\|C\|_1 + 2\|C\|_2\sqrt{t} + 2\|C\|_\infty t) \right\} \leq \exp(-t).$$

As long as $t \geq 1$, we can find a constant $c > 0$ (which depends only on C) such that

$$\theta_1 N^{-1} e_i^\top \hat{\Sigma}_{11}^{-1} e_i (\|C\|_1 + 2\|C\|_2\sqrt{t} + 2\|C\|_\infty t) \leq \frac{\theta_1 v_1 c}{N} t.$$

As t was arbitrary, we now choose a specific value, $\tilde{t}$. In particular, we will take

$$\tilde{t} = \frac{N b_N^2}{4\theta_1 v_1 c} \implies \frac{\theta_1 v_1 c}{N}\tilde{t} = \frac{b_N^2}{4}.$$

By Assumption 2, $\tilde{t}$ tends to ∞, so for N large, it will trivially be larger than 1. This then implies that

$$\begin{aligned} P(A_i) &= P\left(\|K(T_i)\|_{\mathbb{K}} \geq \frac{\|K(\beta_i^\star)\|_{\mathbb{K}}}{2}\right) \\ &\leq P\left(\|K(T_i)\|_{\mathbb{K}}^2 \geq \frac{b_N^2}{4}\right) \leq \exp\left(-\tilde{t}\right) = \exp\left\{-\frac{Nb_N^2}{4\theta_1 v_1 c}\right\}. \end{aligned}$$

Coming back to B_1, we can apply Assumption 2.1 to conclude that

$$P(B_1) \leq \sum_{i\in S} P(A_i) \leq I_0 \exp\left(-\frac{Nb_N^2}{4\nu_1\theta_1 c}\right) = \exp\left(-\frac{Nb_N^2}{4\theta_1\nu_1 c} + \log(I_0)\right) \to 0.$$

Step 2: $P(B_2) \to 0$

Recall that

$$B_2 = \left\{\frac{\lambda}{N}\|e_i^\top \hat{\Sigma}_{11}^{-1}\tilde{s}_1\|_{\mathbb{K}} \geq \frac{\|K(\beta_i^\star)\|_{\mathbb{K}}}{2} : \text{ for some } i \in \mathcal{S}\right\}$$

with $\tilde{s}_1 = \left\{N\tilde{\omega}_i\hat{\beta}_i\|\hat{\beta}_i\|_{\mathbb{K}}^{-1} \;\; i \in \mathcal{S}\right\}$. The $\mathbb{K}$ norm of $\tilde{s}_1$ is given by

$$\|\tilde{s}_1\|_{\mathbb{K}}^2 = \sum_{i\in\mathcal{S}} N^2{\tilde{\omega}_i}^2\frac{\|\hat{\beta}_i\|_{\mathbb{K}}^2}{\|\hat{\beta}_i\|_{\mathbb{K}}^2} = N^2\sum_{i\in\mathcal{S}}\tilde{\omega}_i^2 = N^2\left(\sum_{i\in\mathcal{S}}\omega_i^2 + \sum_{i\in\mathcal{S}}(\tilde{\omega}_i^2 - \omega_i^2)\right),$$

where $\tilde{w}_i = \|\tilde{\beta}_i\|_{\mathbb{H}}^{-1}$ is computed using FSL and $w_i = \|\beta_i^\star\|_{\mathbb{H}}^{-1}$. We have that

$$\begin{aligned} \tilde{\omega}_i^2 - \omega_i^2 &= \|\tilde{\beta}_i\|_{\mathbb{H}}^{-2} - \|\beta_i^\star\|_{\mathbb{H}}^{-2} \\ &= \frac{\|\beta_i^\star\|_{\mathbb{H}}^2 - \|\tilde{\beta}_i\|_{\mathbb{H}}^2}{\|\beta_i^\star\|_{\mathbb{H}}^2\|\tilde{\beta}_i\|_{\mathbb{H}}^2} \\ &= \frac{(\|\beta_i^\star\|_{\mathbb{H}} - \|\tilde{\beta}_i\|_{\mathbb{H}})(\|\beta_i^\star\|_{\mathbb{H}} + \|\tilde{\beta}_i\|_{\mathbb{H}})}{\|\beta_i^\star\|_{\mathbb{H}}^2\|\tilde{\beta}_i\|_{\mathbb{H}}^2}. \end{aligned}$$

From the definition of the rate r_N of Lemma (3), uniformly in i

$$\|\beta_i^\star - \tilde{\beta}_i\|_{\mathbb{H}} \leq \sup_{i\in\mathcal{S}}\|\beta_i^\star - \tilde{\beta}_i\|_{\mathbb{H}} = O_P(r_N^{1/2}),$$

and from Assumption 2 we have that $r_N^{1/2}/b_N \to 0$. Using these facts we have that

$$\|\tilde{\beta}_i\|_{\mathbb{H}} \leq \|\beta_i^\star\|_{\mathbb{H}} + \|\tilde{\beta}_i - \beta_i^\star\|_{\mathbb{H}} = \|\beta_i^\star\|_{\mathbb{H}}(1 + o_P(1)),$$

uniformly in i. Using the reverse triangle inequality we get that

$$\|\tilde{\beta}_i\|_{\mathbb{H}} \geq \|\beta_i^\star\|_{\mathbb{H}} - \|\tilde{\beta}_i - \beta_i^\star\|_{\mathbb{H}} = \|\beta_i^\star\|_{\mathbb{H}}(1 + o_P(1)).$$

One last application of the reverse triangle inequality implies $|\|\beta_i^\star\|_{\mathbb{H}} - \|\tilde{\beta}_i\|_{\mathbb{H}}| \leq \|\beta_i^\star - \tilde{\beta}_i\|_{\mathbb{H}}$ and we thus obtain that

$$|\tilde{\omega}_i^2 - \omega_i^2| = \frac{\|\beta_i^\star - \tilde{\beta}_i\|_{\mathbb{H}}}{\|\beta_i^\star\|_{\mathbb{H}}^3}(2 + o_P(1)),$$

uniformly in i. We replace one of the terms in the denominator by noticing that for all $i \in \mathcal{S}$

$$b_N \leq \|K(\beta_i^\star)\|_{\mathbb{K}} \leq \theta_1^{1/2}\|\beta_i^\star\|_{\mathbb{H}}.$$

Then, uniformly in $i \in \mathcal{S}$

$$|\tilde{\omega}_i^2 - \omega_i^2| \leq \frac{O_P(1)}{\|\beta_i^\star\|_{\mathbb{H}}^2}\frac{\theta_1^{1/2}}{b_N}\|\beta_i^\star - \tilde{\beta}_i\|_{\mathbb{H}} \leq \frac{\theta_1^{1/2}}{b_N}O_P(r_N^{1/2})\omega_i^2.$$

Again by Assumption 2, $r_N^{1/2}/b_N \to 0$, and so we conclude

$$\begin{aligned}\|\tilde{s}_1\|_{\mathbb{K}}^2 &\leq N^2\left(\sum_{i\in\mathcal{S}}\omega_i^2\right)(1 + o_p(1)) = N^2\left(\sum_{i\in\mathcal{S}}\frac{1}{\|\beta_i^\star\|_{\mathbb{H}}^2}\right)(1 + o_p(1)) \\ &\leq N^2\frac{I_0\theta_1^2}{b_N^2}(1 + o_p(1)). \qquad (8)\end{aligned}$$

Then for the original object we have for each $i \in \mathcal{S}$

$$\frac{\lambda}{N}\frac{\|e_i^\top\hat{\Sigma}_{11}^{-1}\tilde{s}_1\|_{\mathbb{K}}}{\|K(\beta_i^\star)\|_{\mathbb{K}}} \leq \frac{\lambda}{N}\frac{\|e_i^\top\hat{\Sigma}_{11}^{-1}\|\,\|\tilde{s}_1\|_{\mathbb{K}}}{\|K(\beta_i^\star)\|_{\mathbb{K}}}$$

with $\|e_i^\top\hat{\Sigma}_{11}^{-1}\| \leq \|e_i\|\|\hat{\Sigma}_{11}^{-1}\|_{op} \leq \nu_1$ from Assumption 2, and in the end

$$\frac{\lambda}{N}\frac{\|e_i^\top\hat{\Sigma}_{11}^{-1}\|\,\|\tilde{s}_1\|_{\mathbb{K}}}{\|K(\beta_i^\star)\|_{\mathbb{K}}} \leq \frac{\lambda\nu_1\sqrt{I_0}N}{Nb_Nb_N}(1 + o_p(1)) \to 0,$$

as desired.

Step 3

From the previous definition of B_3:

$$B_3 = \left\{\frac{1}{N}\|\mathbf{X}_i^\top HK(\varepsilon)\|_{\mathbb{K}} \geq \frac{\lambda\tilde{\omega}_i}{2} : \text{ for some } i \notin \mathcal{S}\right\}$$

we define A_i s.t. for $i \notin \mathcal{S}$

$$A_i = \left\{\frac{1}{N}\|\mathbf{X}_i^\top HK(\varepsilon)\|_{\mathbb{K}} \geq \frac{\lambda\tilde{\omega}_i}{2}\right\}$$

and $B_3 = \cup_{i \notin \mathcal{S}} A_i$. We can define the gaussian process $\mathbf{X}_i H\varepsilon$, which has zero mean and as covariance operator $\mathbf{X}_i^\top HH^\top \mathbf{X}_i C = \mathbf{X}_i^\top H\mathbf{X}_i C$, since H is symmetric and idempotent, with C the covariance operator of the zero mean gaussian process ε. Moreover, since we have that $\sup_{i\notin\mathcal{S}} \|\tilde{\beta}_i\|_{\mathbb{H}} = O_P\left(r_N^{1/2}\right)$ we can notice that $\tilde{\omega}_i \leq 1/\sup_{i\notin\mathcal{S}}(\|\tilde{\beta}_i\|_{\mathbb{H}})$ and then

$$A_i \subseteq \left\{ O_P\left(r_N^{1/2}\right) \|\mathbf{X}_i^\top HK(\varepsilon)\|_{\mathbb{K}} \geq \frac{N\lambda}{2} \right\}.$$

Then for any $\epsilon > 0$ we can find a $T = T(\epsilon) > 0$ s.t.

$$P(A_i) \leq \frac{\epsilon}{2(I - I_0)} + P\left(\|\mathbf{X}_i^\top HK(\varepsilon)\|_{\mathbb{K}} \geq \frac{N\lambda}{2Tr_N^{1/2}} \right).$$

As we discussed before, to apply Corollary 1, we need to choose $\tilde{t}$ such that

$$\mathbf{X}_i^\top H\mathbf{X}_i(\|C\|_1 + 2\|C\|_2\sqrt{\tilde{t}} + 2\|C\|_\infty \tilde{t}) \leq \left(\frac{N\lambda}{2Tr_N^{1/2}}\right)^2. \tag{9}$$

Focusing on the left side of the inequality we know that

$$\mathbf{X}_i^\top H\mathbf{X}_i(\|C\|_1 + 2\|C\|_2\sqrt{\tilde{t}} + 2\|C\|_\infty \tilde{t}) \leq N\tilde{t}c.$$

Since H is a projection matrix we have

$$\mathbf{X}_i H\mathbf{X}_i = \sum_{t=1}^{N} \left(\sum_{n=1}^{N} \mathbf{X}_{i,n} H_{n,t} \right)^2 = \sum_{t=1}^{N} 1 = N,$$

and again there exists a constant c such that $\forall t, ct \geq (\|C\|_1 + 2\|C\|_2\sqrt{t} + 2\|C\|_\infty t)$, so we define $\tilde{t}$:

$$\tilde{t}cN \leq \left(\frac{N\lambda}{2Tr_N^{1/2}}\right)^2 \Rightarrow \tilde{t} = \frac{\lambda^2 N}{4T^2 c r_N}.$$

Applying corollary 1 we have

$$P\left(\|\mathbf{X}_i^\top HK(\varepsilon)\|_{\mathbb{K}} \geq \frac{N\lambda}{2Tr_N^{1/2}} \right) \leq \exp\left(-\frac{\lambda^2 N}{4T^2 c r_N}\right) \leq \exp\left(-\frac{I_0 \log^2(I)}{N 4T^2 c r_N}\right)$$

and then we can compute the probability of B_3

$$\begin{aligned} P(B_3) &\leq \sum_{i\notin\mathcal{S}} P(A_i) \leq (I - I_0)\exp\left(-\frac{I_0\log^2(I)}{4NT^2 c r_N}\right) + \frac{\epsilon}{2} \\ &\leq \exp\left(-\frac{I_0\log^2(I)}{4NT^2 c r_N} + \log(I - I_0)\right) + \frac{\epsilon}{2}. \end{aligned}$$

Since $r_N \ll (I_0 \log^2(I))/N$, we can take N large enough to make the first term smaller then $\epsilon/2$ and have the convergence of the probability to 0.

Step 4

Recall that B_4 is defined as

$$B_4 = \left\{ \frac{1}{N^2} \|\mathbf{X}_i^\top \mathbf{X}_1 \hat{\Sigma}_{11}^{-1} \tilde{s}_1\|_{\mathbb{K}} \geq \frac{\tilde{\omega}_i}{2} : \text{ for some } i \notin \mathcal{S} \right\}.$$

Recall from (8)

$$\|\tilde{s}_1\|_{\mathbb{K}}^2 \leq N^2 \theta_1^2 \frac{I_0}{b_N^2} (1 + o_p(1)),$$

as well as

$$\sup_{i \notin \mathcal{S}} \tilde{\omega}_i{}^{-1} = O_P(r_N^{1/2}).$$

The irrepresentable condition (Assumption 2.4) implies

$$\forall i \notin \mathcal{S}, \ \ \|\mathbf{X}_i^\top \mathbf{X}_1 \hat{\Sigma}_{11}^{-1}\|_{op} \leq \|\hat{\Sigma}_{21} \hat{\Sigma}_{11}^{-1}\|_{op} \leq \phi < 1.$$

Then we consider the inequality of B_4 for a fixed $i \notin \mathcal{S}$

$$\frac{2\|\mathbf{X}_i^\top \mathbf{X}_1 \hat{\Sigma}_{11}^{-1} \tilde{s}_1\|_{\mathbb{K}}}{N^2 \tilde{\omega}_i} \leq \frac{2\|\mathbf{X}_i^\top \mathbf{X}_1 \hat{\Sigma}_{11}^{-1}\|_{op} \|\tilde{s}_1\|_{\mathbb{K}}}{N^2 \tilde{\omega}_i} \leq \frac{2\phi r_N^{1/2} I_0^{1/2} \theta_1}{N b_N} O_P(1) \to 0,$$

which finishes Step 4 and completes the proof.

Proof of Theorem 1.2

The strategy of this proof is to show that the (projected) FLAME estimate is equal to the oracle estimator (least squares when the true predictors are known) plus a bias term. We then show how the adaptive step allows for the bias to be asymptotically negligible. Let $h_n = \{h_{i,n}\} \in \mathbb{K}^I$ be a bounded sequence: $\|h_n\|_{\mathbb{K}} < M_1$. We will show that

$$\frac{\sqrt{N}\langle h_n, \hat{\beta} - \beta^\star \rangle_{\mathbb{H}}}{\sigma_n} \xrightarrow{D} \mathcal{N}(0,1) \qquad \text{where} \qquad \sigma_n^2 = \sum_{i=1}^{I_0} \sum_{i=1}^{I_0} \hat{\Sigma}_{11;ij}^{-1} \langle h_{i,n}, C h_{j,n} \rangle,$$

assuming that the $h_{i,n}$ are chosen such that $\sum_{i \in \mathcal{S}} \langle C^{1/2} h_i, C^{1/2} h_i \rangle \geq M_2 > 0$ for some fixed M_2. Recall that the oracle estimator is

$$\hat{\beta}_O^{\mathcal{S}} = (\mathbf{X}_1^\top \mathbf{X}_1)^{-1} \mathbf{X}_1^\top Y \qquad \text{and} \qquad \hat{\beta}_O = \{\hat{\beta}_O^{\mathcal{S}}, 0\},$$

where 0 here is the zero function in $\mathbb{K}^{I - I_0}$. Since we assume that the Y are Gaussian, we have that

$$\sqrt{N}\langle h_n, \hat{\beta}_O - \beta_1^\star \rangle_{\mathbb{H}} \sim \mathcal{N}(0, \sigma_n^2).$$

By Assumption 2.3 we have that

$$\sigma_n^2 \geq \nu_1^{-1} \sum_{i \in \mathcal{S}} \langle C^{1/2} h_i, C^{1/2} h_i \rangle \geq \nu_1 M_2,$$

and so is bounded from below, so we need only to show that

$$\sqrt{N} \langle h_n, \hat{\beta}_O - \hat{\beta}_1 \rangle_{\mathbb{H}} = o_P(1).$$

From equation 5, when $\hat{S} = S$ we have that

$$\begin{aligned} \mathbf{X}_1^\top K(Y - \mathbf{X}_1 \hat{\beta}_1) = \lambda \tilde{s}_1 \implies \hat{\beta}_1 &= (\mathbf{X}_1^\top \mathbf{X}_1)^{-1} \mathbf{X}_1^\top Y - \lambda (\mathbf{X}_1^\top \mathbf{X}_1)^{-1} K^{-1}(\tilde{s}_1) \\ &= \hat{\beta}_O - \lambda (\mathbf{X}_1^\top \mathbf{X}_1)^{-1} K^{-1}(\tilde{s}_1). \end{aligned}$$

Let $h_n^S = \{h_{i,n} : i \in S\}$. We then have that the difference, projected onto h is given by

$$\begin{aligned} \sqrt{N} \langle h, \hat{\beta}_O - \hat{\beta} \rangle_{\mathbb{H}} &= \sqrt{N} \lambda \langle (\mathbf{X}_1^\top \mathbf{X}_1)^{-1} K^{-1}(\tilde{s}_1), h_n^{\mathcal{S}} \rangle_{\mathbb{H}} \\ &= \frac{\lambda}{\sqrt{N}} \langle \hat{\Sigma}_{11}^{-1} K^{-1/2}(\tilde{s}_1), K^{-1/2} h_n^{\mathcal{S}} \rangle \\ &\leq \frac{\lambda}{\sqrt{N}} \| \hat{\Sigma}_{11}^{-1} K^{-1/2}(\tilde{s}_1) \|_{\mathbb{H}} \| h_n \|_{\mathbb{K}}, \end{aligned}$$

where the last step follows from the Cauchy-Schwarz inequality. Applying Assumption 2.3 along with an operator inequality we have that

$$\frac{\lambda}{\sqrt{N}} \| \hat{\Sigma}_{11}^{-1} K^{-1/2}(\tilde{s}_1) \|_{\mathbb{H}} \| h_n \|_{\mathbb{K}} \leq \frac{\lambda}{\sqrt{N} \nu_1} \| \tilde{s}_1 \|_{\mathbb{K}} \| h_n \|_{\mathbb{K}}.$$

From the equation (8) we have

$$\| \tilde{s}_1 \|_{\mathbb{K}} \leq \frac{\sqrt{I_0} N \theta_1}{b_N} (1 + o_p(1))$$

and then

$$|\sqrt{N} \langle h, \hat{\beta}_O - \hat{\beta}_1 \rangle_{\mathbb{H}}| \leq \frac{\lambda \theta_1 \sqrt{I_0} \sqrt{N} \| h_n \|_{\mathbb{K}}}{\nu_1 b_N} (1 + o_P(1)) = o_P(1),$$

by Assumption 2. Since $P\left(\hat{\mathcal{S}} = \mathcal{S}\right) \to 1$ the proof is complete.

Proof of Theorem 2

We begin by partitioning the problem into two pieces. Let $e_i \otimes v_j$ denote the tensor product between e_i, a vector of zeros except in the ith coordinate which is 1, and v_j, the jth eigenfunction of K. In other words, $e_i \otimes v_j \in \mathbb{K}^I$, and is the zero function in all coordinates except the ith where it is equal to v_j. Now

fix a positive integar J and apply Parseval's identity to obtain

$$\begin{aligned} N\|\hat{\beta}-\hat{\beta}_O\|^2 &= N\sum_{i=1}^{I}\|\hat{\beta}_i-\hat{\beta}_{O;i}\|^2 \\ &= N\sum_{i=1}^{I}\sum_{j=1}^{\infty}\langle\hat{\beta}_i-\hat{\beta}_{O;i},v_j\rangle^2 \\ &= N\sum_{i=1}^{I}\sum_{i=1}^{J}\langle\hat{\beta}-\hat{\beta}_O,e_i\otimes v_j\rangle^2 \qquad (10) \\ &+ N\sum_{i=1}^{I}\sum_{i=J+1}^{\infty}\langle\hat{\beta}-\hat{\beta}_O,e_i\otimes v_j\rangle^2. \qquad (11) \end{aligned}$$

Bounding (10) follows the similar arguments as in the proof of Theoerem 1.2, namely, when $\hat{S}=S$, the summands are zero unless $i\in S$. For $i\in S$ we then have

$$\begin{aligned} \langle\hat{\beta}-\hat{\beta}_O,e_i\otimes v_j\rangle^2 &= \frac{\lambda^2}{N^2\theta_j}\langle\hat{\Sigma}_{11}^{-1}K^{-1/2}(\tilde{s}_1),e_i\otimes v_j\rangle^2 \\ &\leq \frac{\lambda^2}{N^2\theta_j\nu_1^2}\langle K^{-1/2}(\tilde{s}_1),e_i\otimes v_j\rangle^2. \end{aligned}$$

This gives the bound

$$\begin{aligned} N\sum_{i=1}^{I}\sum_{j=1}^{J}\langle\hat{\beta}-\hat{\beta}_O,e_i\otimes v_j\rangle^2 = N\sum_{i\in S}\sum_{j=1}^{J}\langle\hat{\beta}-\hat{\beta}_O,e_i\otimes v_j\rangle^2 &\leq \frac{\lambda^2}{\theta_J\nu_1^2N}\|\tilde{s}_1\|_{\mathbb{K}}^2 \\ &\leq \frac{\lambda^2NI_0}{\theta_J\nu_1^2b_N^2}(1+o_P(1)). \end{aligned}$$

Turning to the second term, we express $\hat{\beta}$ using a different form. Notice that we can write

$$\tilde{s}_1=\Lambda\hat{\beta}_1,$$

where Λ is a diagonal matrix of the terms $\{N\tilde{w}_i\|\hat{\beta}_i\|_K^{-1}\}$. We therefore have that

$$\mathbf{X}_1^\top K(Y)-(\mathbf{X}_1^\top\mathbf{X}_1)K(\hat{\beta})-\lambda\Lambda\hat{\beta}_1=0.$$

We can re-express this equation as

$$\hat{\beta}_O-\hat{\beta}_1+\lambda(\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\Lambda K^{-1}(\hat{\beta}_1)=0\Longrightarrow\hat{\beta}_1=(I+\lambda(\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\Lambda K^{-1})^{-1}\hat{\beta}_O.$$

The above shrinks (all operators above are positive definite) every coordinate of $\hat{\beta}_O$ to obtain $\hat{\beta}_1$ and thus we have that

$$N\sum_{i\in S}\sum_{j=J+1}^{\infty}\langle\hat{\beta}-\hat{\beta}_O, e_i\otimes v_j\rangle^2 \le 4N\sum_{i\in S}\sum_{j=J+1}^{\infty}\langle\hat{\beta}_O, e_i\otimes v_j\rangle^2.$$

To bound the above in probability, notice that it is positive, thus we can use Markov's inequality. Computing the expected value (recall that $\mathbf{X}$ is not random, only the error terms are) of the projected oracle estimate we obtain

$$\mathrm{E}\langle\hat{\beta}_O, e_i\otimes v_j\rangle^2 = \langle\beta^\star, e_i\otimes v_j\rangle^2 + (\mathbf{X}_1^\top\mathbf{X}_1)^{-1}_{i,i}\langle Cv_j, v_j\rangle.$$

This implies that

$$\begin{aligned}&4N\sum_{i\in S}\sum_{j=J+1}^{\infty}\langle\hat{\beta}_O, e_i\otimes v_j\rangle^2\\ &\quad= O_P(1)N\left[\sum_{i\in S}\sum_{j=J+1}^{\infty}\langle\beta^\star, e_i\otimes v_j\rangle^2 + \operatorname{trace}((\mathbf{X}_1^\top\mathbf{X}_1)^{-1})\sum_{j=J+1}^{\infty}\langle Cv_j, v_j\rangle\right].\end{aligned}$$

By Assumption 2.3 we have $\operatorname{trace}((\mathbf{X}_1^\top\mathbf{X}_1)^{-1}) \le I_0\nu_1/N$ and since C is trace class, it follows that $\sum_{j=J+1}^{\infty}\langle Cv_j, v_j\rangle \to 0$ as $J\to\infty$. Lastly, by the additional assumptions of Theorem 2 we have

$$\sum_{i\in S}\sum_{j=J+1}^{\infty}\langle\beta^\star, e_i\otimes v_j\rangle^2 = \sum_{i\in S}\sum_{j=J+1}^{\infty}\theta_j^{1+\delta}\frac{\langle\beta^\star, e_i\otimes v_j\rangle^2}{\theta_j^{1+\delta}} \le \theta_J^{1+\delta}I_0B.$$

Putting everything together, we get the bound

$$4N\sum_{i\in S}\sum_{j=J+1}^{\infty}\langle\hat{\beta}_O, e_i\otimes v_j\rangle^2 = O_P(1)\left[NI_0\theta_J^{1+\delta}B^2 + I_0\nu_1 o(1)\right].$$

In this case, I_0 is fixed, so the second term is just $o(1)$.

To ensure both (10) and (11) go to zero, we require that J is such that

$$N\theta_J^{1+\delta}\to 0 \qquad \text{and} \qquad \frac{\lambda^2 N}{\theta_J b_N^2}\to 0.$$

So we need to be able to choose J such that

$$\theta_J \ll N^{-1/(1+\delta)} \quad \text{and} \quad \theta_J \gg \frac{\lambda^2 N}{b_N^2}.$$

This is possible if

$$\frac{\lambda^2 N}{b_N^2} \ll N^{-1/(1+\delta)} \iff \lambda \ll \frac{b_N}{N^{1/2[1+1/(1+\delta)]}},$$

as desired.

Appendix C: The periodic setting

In this section we focus on a distinctive feature of FLAME: the possibility of adapting the choice of the kernel to the prior knowledge on the data. For example in Figure 9 we plot several periodic coefficients $\beta^\star$. When using FLAME with a periodic kernel, the resulting estimates will also be periodic. In Figure 10 of the Appendix, for example, the eigenfunctions of the periodic kernel with period $1/2$ are shown. This kernel is general enough to be used for the estimations in a simulation setting where $\beta^\star$ functions are sampled as periodic functions with period varying in $\{1/2, 1/4, 1/8\}$. AFSL and MCP, on the contrary, don't allow this characterizations of the coefficients.

The design matrix $\mathbf{X}$ is the standardized realization of a multivariate normal distribution with 0 average and identity covariance structure and the errors are sampled from a Matérn process with parameter $(\nu = 1.5, \text{range} = 1/4, \sigma^2 = 1)$. The aim is to compare the results of FLAME, MCP, and AFSL. In this particular case, a kernel with period $\{1/2\}$ allows FLAME to estimate all the predictors identifying also their periodicity. MCP and AFSL, in contrast, are estimated in the general L^2 space, without any further specifications. In Table 5 of the appendix we present a summary of the average results across 100 replications for the three methods; where we see a fairly dramatic increase in statistical performance for FLAME. An example of the estimates produced by the different methods, based on $\beta^\star$ from Figure 9, is given in Figure 11 of the appendix, where we see a again a fairly dramatic advantage when using FLAME.

Appendix D: Additional simulation settings

In this section we provide some additional simulations to complement the ones found in the main body of the paper. In Figure 5 we repeat the simulations from Section 4.1 with the rougher β coefficient functions ($\nu = 2.5$). This plot should be compared to Figure 1 as the only difference is in how the adaptive weights are computed. We see that the results are nearly the same, with the only noticeable difference being the performance of the exponential kernel. When using FSL to compute the weights, results for the exponential kernel level off a bit earlier (as a function of the kernel parameter). Otherwise the results are the same, suggesting that FLAME is fairly robust against method used in the non-adaptive step for finding the weights, though our asymptotic theory currently requires that these weights be based on consistent estimates.

In Figure 6 and Table 2 one can find results using the exponential kernel, the rougher β, but without using the kill switch. We can see that while the statistical performance is nearly the same, the computation time increases dramatically. In essense, the kill switch has saved us from computing solutions that had no chance of being selected in the cross validation. In practice, if a solution is chosen near the kill switch, it is a signal that one might want to increase it and rerun the procedure to make sure the best solution wasn't missed.

Appendix E: Additional figures and tables

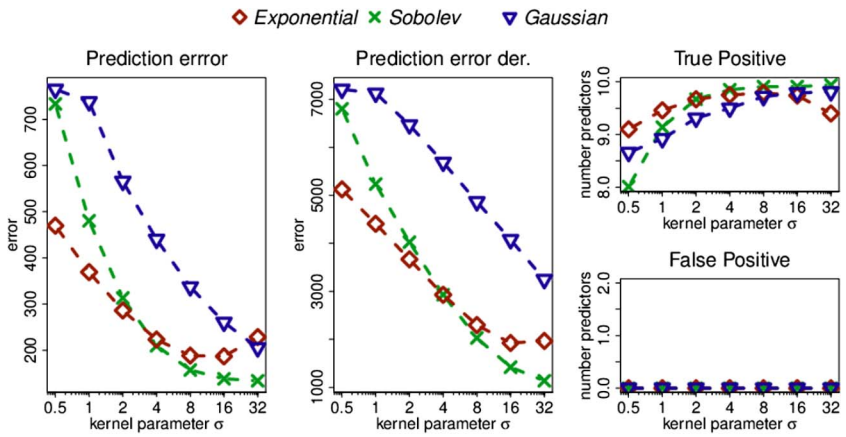

FIG 5. *A recreation of Figure 1, but using FSL to compute the weights* $\tilde{\omega}_i$.

TABLE 1
Median time for the simulations varying kernel for the rough (left panel) and smooth case (right panel).

	FLAME		
	Kernel		
σ	Gaus.	Sob.	Exp.
0.5	29.30	30.75	41.14
1	21.64	36.62	48.17
2	28.87	43.92	58.67
4	32.34	39.14	61.48
8	32.61	42.99	47.29
16	33.67	42.59	39.95
32	35.47	33.47	40.83

	FLAME		
	Kernel		
σ	Gaus.	Sob.	Exp.
0.5	80.38	85.81	95.00
1	77.67	81.33	94.66
2	72.23	87.59	97.95
4	66.69	76.18	91.18
8	58.46	79.12	99.08
16	61.14	80.36	92.98
32	63.23	70.22	69.97

TABLE 2
Median time for simulations in the rough case using the exponential kernel and **no kill switch**. *Compare to Table 1.*

σ	0.5	1	2	4	8	16	32
time	607.78	650.23	555.42	167.40	148.68	172.29	312.86

TABLE 3
Median time (sec.) for the simulations varying method for the rough (left panel) and smooth case (right panel) in the small dimensional setting.

m	MCP	FLAME	AFSL
15	36.00	12.90	7.34
20	32.20	12.56	7.20
50	92.35	13.00	7.28
100	126.58	12.08	7.15
200	377.36	13.95	6.54

m	MCP	FLAME	AFSL
15	12.84	76.85	7.75
20	13.89	60.39	6.92
50	66.30	45.106	8.11
100	139.86	92.57	7.00
200	221.36	85.45	6.14

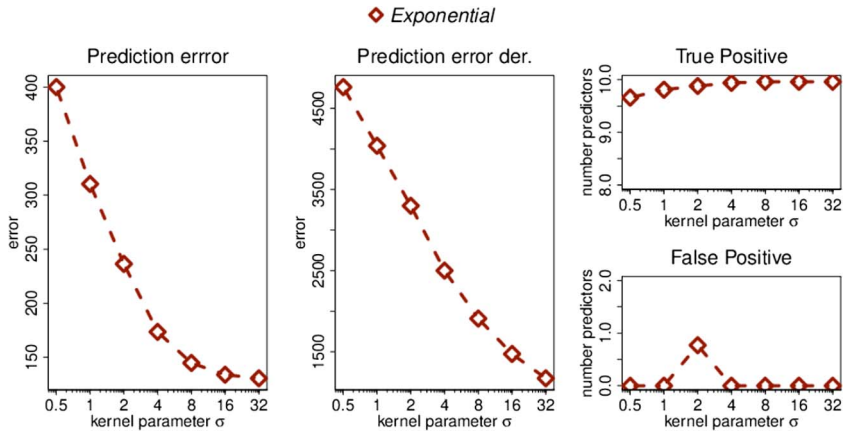

FIG 6. *A recreation of Figure 1, but without using a kill switch.*

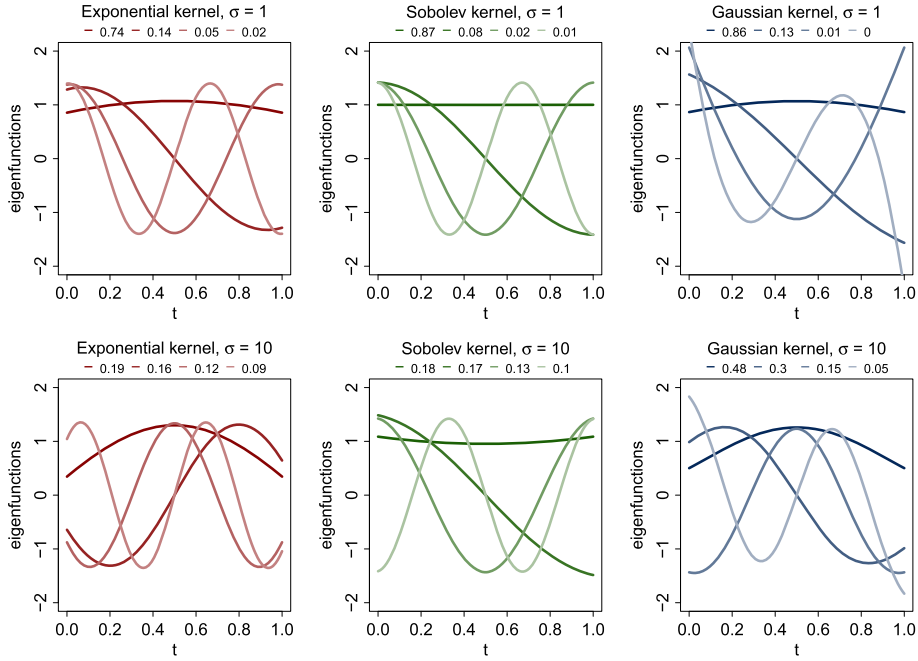

FIG 7. *Representation of the first four eigenfunctions for each kernel with different σ. From the left: the Exponential, the Sobolev and the Gaussian kernel. The legend at the top of each panel denotes proportion of the explained variability for each eigenfunction.*

TABLE 4

AFSL results for the rough and smooth high-dimensional simulation setting. Prediction error, computation time and number of correctly and wrongly identified predictors are presented. This results have to be compared with Figure 1 for estimation error and with Table 1 of the appendix for the computational efficiency.

	prediction error	prediction error der.	True Positives	False Positives	Time (sec.)
rough setting	352.51	4664.2	9.92	0.08	1031.01
smooth setting	95.43	382.17	9.64	0.41	812.24

TABLE 5

Comparison of the results of the three methods on simulations in the periodic setting. Average prediction error on data, derivatives, average number of true positive, false positive and the median computational time are shown.

	prediction error	prediction error der.	True Positives	False Positives	Time
FLAME	24.99	666.15	4.93	0.03	25.99
MCP	162.24	4055.37	5	5	924.98
AFSL	54.54	2081.90	4.87	0.53	8.04

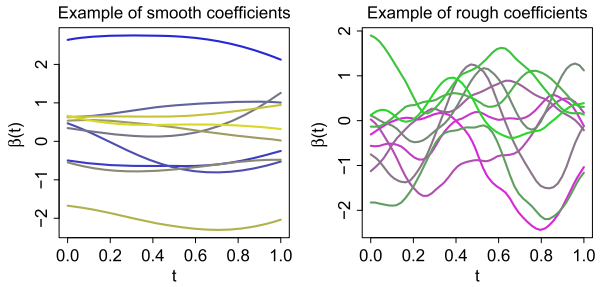

FIG 8. *Example of 10 $\beta^\star$ coefficients for the smooth (left panel) and rough (right panel) simulation setting.*

TABLE 6

List of the identified SNPs with AFSL and FLAME. + identifies the SNPs with positive effect and - the SNPs with negative effect, empty cells identify non detected SNPs. Informations on the chromosome location of SNPs and further details can be found in the ALFRED database ([27]).

SNP		AFSL	FLAME
chr	name		
1	rs1875650	+	+
2	rs953044	-	-
5	rs1368183	+	+
6	rs7751381	+	+
6	rs2206980	-	-
7	rs17372029	+	+
8	rs1540897	+	+
8	rs4734250	+	+
10	rs4752250	+	+
11	rs722490		+
15	rs2019435	+	+
20	rs2041420		-

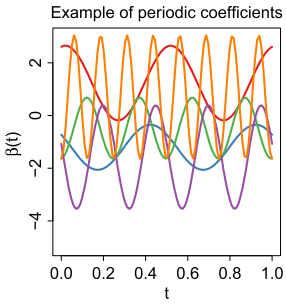

FIG 9. *Example of 5 $\beta^\star$ periodic coefficients, two have period 0.5, two 0.25 and one 0.125.*

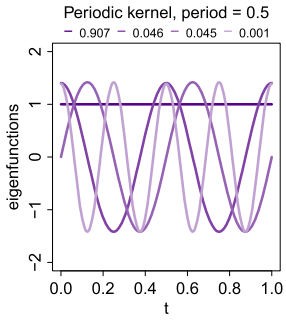

FIG 10. *First four eigenfunctions of the periodic kernel with period 0.5. Correspondent explained variability is shown in the top legend*

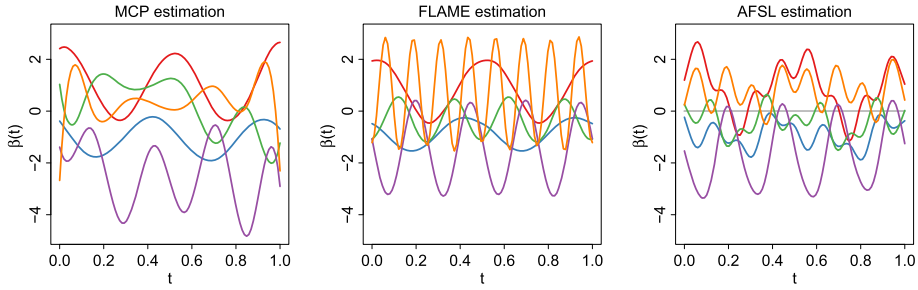

FIG 11. *Example of the estimation of the functions of Figure 9 with, from the left, FLAME, MCP and AFSL.*

References

[1] R. Barber, M. Reimherr, and T. Schill. The function-on-scalar lasso with applications to longitudinal GWAS. *Electronic Journal of Statistics*, 11(1):1351–1389, 2017. MR3635916

[2] V. Barbu and T. Precupanu. *Convexity and optimization in Banach spaces*. Springer Science & Business Media, 2012. MR3025420

[3] H. H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer Science & Business Media, 2011. MR2798533

[4] A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011. MR2239907

[5] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004. MR2061575

[6] P. Bühlmann, M. Kalisch, and M. Maathuis. Variable selection in high-dimensional linear models: partially faithful distributions and the PC–simple algorithm. *Biometrika*, 97(2):261–278, 2010. MR2650737

[7] H. Cardot, A. Mas, and P. Sarda. CLT in functional linear regression models. *Probability Theory and Related Fields*, 138(3-4):325–361, 2007. MR2299711

[8] Y. Chen, J. Goldsmith, and T. Ogden. Variable selection in function-on-scalar regression. *Stat*, 5:88–101, 2016. MR3478799

[9] W. Chu, R. Li, and M. Reimherr. Feature screening for time-varying coefficient models with ultrahigh-dimensional longitudinal data. *Ann. Appl. Stat.*, 10(2):596–617, 06 2016. URL http://dx.doi.org/10.1214/16-AOAS912. MR3528353

[10] dbGaP. SHARP - national heart, lung, and blood institute snp health association asthma resource project. http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000166.v2.p1, 2009.

[11] N. Dunford and J. Schwartz. *Linear operators. Part 2: Spectral theory. Self adjoint operators in Hilbert space*. Interscience Publishers, 1963. MR0188745

[12] Y. Fan, G. James, and P. Radchenko. Functional Additive Regression. *Annals of Statistics*, 43:2296–2325, 2015. MR3396986

[13] Z. Fan and M. Reimherr. Adaptive function-on-scalar regression. *Econometrics and Statistics*, 1:167–183, 2017. MR3669995

[14] J. Gertheiss, A. Maity, and A. Staicu. Variable selection in generalized functional linear models. *Stat*, 2(1):86–101, 2013.

[15] S. Graves, G. Hooker, and J. Ramsay. *Functional Data Analysis with R and MATLAB*. Springer, 2009. MR2815593

[16] T. Hastie, R. Tibshirani, and J. Friedman. *Elements of Statistical Learning*. Springer, 2001. MR1851606

[17] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC Press, 2015. MR3616141

[18] L. Horváth and P. S. Kokoszka. *Inference for Functional Data with Applications*. Springer, 2012. MR2920735

[19] T. Hsing and R. Eubank. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. John Wiley & Sons, 2015. MR3379106

[20] J. Huang, S. Ma, and C.-H. Zhang. Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, pages 1603–1618, 2008. MR2469326

[21] G. James, T. Hastie, D. Witten, and R. Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. Springer Texts in Statistics. Springer London, Limited, 2013. ISBN 9781461471370. URL http://books.google.com/books?id=at1bmAEACAAJ. MR3100153

[22] P. Kokoszka and M. Reimherr. *Introduction to Functional Data Analysis*. Chapman & Hall, 2017. MR3793167

[23] R. Laha and V. Rohatgi. *Probability Theory*. Wiley, New York, 1979. MR0534143

[24] H. Lian. Shrinkage estimation and selection for multiple functional regression. *Statistica Sinica*, 23:51–74, 2013. MR3076158

[25] H. Matsui and S. Konishi. Variable selection for functional regression models via the L1 regularization. *Computational Statistics & Data Analysis*, 55(12):3304–3310, 2011. MR2825412

[26] H. S. Noh and B. U. Park. Sparse varying coefficient models for longitudinal data. *Statistica Sinica*, pages 1183–1202, 2010. MR2730179

[27] H. Rajeevan, U. Soundararajan, S. Stein, K. K. Kidd, and P. Miller. ALFRED - the allele frequency database. https://alfred.med.yale.edu/alfred/AboutALFRED.asp, 1999. Yale University.

[28] J. O. Ramsay and B. Silverman. *Functional data analysis*. Wiley Online Library, 2006. MR2168993

[29] E. Repapi, I. Sayers, L. V. Wain, P. R. Burton, T. Johnson, M. Obeidat, J. H. Zhao, A. Ramasamy, G. Zhai, V. Vitart, et al. Genome-wide association study identifies five loci associated with lung function. *Nature genetics*, 42(1):36–44, 2010.

[30] N. Z. Shor. *Minimization methods for non-differentiable functions*, volume 3. Springer Science & Business Media, 2012. MR0775136

[31] B. Sriperumbudur and Z. Szabó. Optimal rates for random fourier features. In *Advances in Neural Information Processing Systems*, pages 1144–1152, 2015.

[32] D. P. Strachan, A. R. Rudnicka, C. Power, P. Shepherd, E. Fuller, A. Davis, I. Gibb, M. Kumari, A. Rumley, G. J. Macfarlane, et al. Lifecourse influences on health among british adults: effects of region of residence in childhood and adulthood. *International journal of epidemiology*, 36(3):522–531, 2007.

[33] The Childhood Asthma Management Program Research Group. The childhood asthma management program (CAMP): design, rationale, and methods. *Controlled Clinical Trials*, 20:91–120, 1999.

[34] H. Wang, G. Zou, and A. T. Wan. Adaptive lasso for varying-coefficient partially linear measurement error models. *Journal of Statistical Planning and Inference*, 143(1):40–54, 2013. MR2969009

[35] L. Wang, H. Li, and J. Huang. Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association*, 103(484):1556–1569, 2008. MR2504204

[36] F. Wei, J. Huang, and H. Li. Variable selection and estimation in high-dimensional varying-coefficient models. *Statistica Sinica*, 12(4):1515–1540, 2011. MR2895107

[37] L. Xue and A. Qu. Variable selection in high-dimensional varying-coefficient models with global optimality. *Journal of Machine Learning Research*, 13(Jun):1973–1998, 2012. MR2956349

[38] C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, pages 894–942, 2010. MR2604701

[39] P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7(Nov):2541–2563, 2006. MR2274449