Statistics and Its Interface Volume 15 (2022) 161–170

Distribution free prediction intervals for multiple functional regression

Ryan Kelly and Kehui Chen*

This paper applies conformal prediction techniques to the problem of constructing prediction intervals in a multiple functional regression setting. After a short introduction to the Signature expansion and its favorable properties, a method utilizing this feature set is developed with great modeling flexibility. With minimal assumptions, the resulting algorithm produces a closed form solution for a prediction set with guaranteed coverage. The good performance of the proposed method is illustrated using simulations and data examples.

Keywords and phrases: Multiple functional regression, Prediction intervals, Conformal prediction, Signature expansion.

1. INTRODUCTION

One area of research in functional data analysis is the seemingly simple task of using functional data to make predictions. Construction of prediction intervals in regression settings is a classical problem in statistics with wide ranging applications. From neuroscience to climatology, researchers wish to use the functional data they have collected to predict the future value of some other variable. Moreover, it is often valuable to obtain a prediction interval for this variable rather than a single "most likely" estimate. Few methods to create these intervals exist in functional data analysis, and those that do exist require fairly strict conditions on the true nature of the data. The research on model free or distribution free prediction intervals has gained increasing interest in recent years, because it becomes more challenging to specify a correct model or rigorously check the modeling assumptions when the data is so complex. In this research, we focused on developing computationally efficient methods for conformal prediction intervals in multiple functional regression settings. The prediction intervals constructed by the conformal method have guaranteed coverage (confidence) without the heavy restrictions on the error distribution and on the regression function, while the efficiency (implied by the length of the intervals) will depend on the representation and information compression of the functional predictors. To accommodate flexible regression relationships, we developed multiple functional regression approaches based on the Signature extraction, which is a mathematical tool to represent the information contained in the functions by a collection of iterated integrals. Then, we were able to derive a closed form expression for the conformal prediction set using the Signature-based conformity score. Numerical studies as well as a data application related to corn and soy yield in Kansas confirmed the efficacy of the proposed method.

The rest of the paper is organized as follows. Section 2 introduces the conformal prediction method in the context of a multiple functional regression problem. Section 3 presents the proposed conformity score using the Signature-based functional regression approach, and derives the algorithm for constructing the exact conformal prediction intervals for multiple functional regression. Section 4 focuses on application of this algorithm in simulated data settings, and Section 5 applies it to a meteorological dataset. Some further details are given in the Appendix.

2. CONFORMAL PREDICTION FOR MULTIPLE FUNCTIONAL REGRESSION

Let us first look at the prediction problem in general. In the prediction problem, i.i.d. (X, Y) pairs of data are observed. A new X_{n+1} is then obtained, and we wish to predict a range of likely values for Y_{n+1}. More precisely, we would like to construct a prediction set C which contains Y_{n+1} with probability at least $1-\alpha$, for some given $\alpha \in (0,1)$. In the simplest case, $Y = \beta X + \epsilon$, $X \in \mathbb{R}^p$, $Y \in \mathbb{R}$, $\epsilon \sim N(0, \sigma^2)$, this process results in the standard t interval: $\hat{Y} \pm t_{n-p,\alpha/2}\sqrt{\text{MSE}(1 + X'_{n+1}(X'X)^{-1}X_{n+1})}$. Among other nice properties, this interval is easy to calculate and has correct finite sample coverage, which makes it a natural choice in this setting.

These properties begin to disappear once we start to generalize the model. If we instead consider the model $Y = m(X) + \epsilon$, m an unknown function, $X \in \mathbb{R}^p$, $Y \in \mathbb{R}$, $\epsilon \sim N(0, \sigma^2)$, then we must first use some nonparametric method to estimate m, then construct an interval based on the residuals and normal quantiles. If we further generalize to a symmetric, but not necessarily normal ϵ, then we must additionally estimate this distribution using nonparametric methods. Not only do these methods only provide asymptotic coverage rather than finite sample coverage, but estimating m is difficult when p is larger than 2 or 3, and even

*Corresponding author.

more challenging when we consider multiple functional predictors, where X_i contains $\{x_{i1}(t), x_{i2}(t), \ldots, x_{ip}(t), t \in \mathcal{T}\}$.

Other methods of approaching this problem include nonparametric conditional distribution or density estimation of Y given X ([7], [10]), or nonparametric quantile regression ([14]) in the form of $f_\tau(x)$. If quantile regression assumptions hold for both $\tau = \alpha/2$ and $\tau = 1 - \alpha/2$, one can have asymptotically valid prediction intervals $(\hat{f}_{\alpha/2}(x), \hat{f}_{1-\alpha/2}(x))$. However, these nonparametric methods are difficult to generalize to functional data, where X itself is in a functional space. There are a few functional quantile regression methods with a single functional predictor ([3], [4]), but all of them need modeling assumptions on the true relationships.

2.1 Background on conformal prediction

[22] introduced the idea of conformal prediction as a method of generating prediction intervals with finite sample coverage with minimal assumptions. Conformal prediction is based on the simple observation that if $U_1, \ldots, U_{n+1}$ is a sequence of i.i.d. random variables, then the rank of U_{n+1} will be uniform over $\{1, 2, \ldots, n+1\}$. Therefore, $P(\text{rank}(U_{n+1}) \leq \lceil (n+1)(1-\alpha) \rceil) \geq 1 - \alpha$ for any $\alpha \in (0,1)$ and we can define the sample quantile based on the order statistics $U_{(1)} \leq U_{(2)} \leq \cdots \leq U_{(n+1)}$ as

$$\hat{q}_{1-\alpha} = \begin{cases} U_{(\lceil (n+1)(1-\alpha) \rceil)} & \text{if } \lceil (n+1)(1-\alpha) \rceil \leq n \\ \infty & \text{otherwise} \end{cases}$$

By this definition $P(U_{n+1} \leq \hat{q}_{1-\alpha}) \geq 1 - \alpha$ for all α.

Returning to our goal of constructing prediction intervals, let $\sigma_i(y) = \sigma(\{(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1} = y)\}, (X_i, Y_i))$, $i = 1, \ldots, (n+1)$ be a *conformity score*, where σ is some function symmetric in the entries in its first argument. This conformity score measures how similar (X_i, Y_i) is to the rest of the data. Although there are many reasonable choices for σ, in the context of a regression problem a natural choice of conformity score would be based on residuals from the regression model. For the rest of the paper, we let $\sigma_i(y) = |Y_i - \hat{m}(X_i)|$, where $\hat{m}$ is some regression function trained on the augmented data $\{(X_1, Y_1), \ldots, (X_{n+1}, y)\}$. Note that a *larger* σ_i indicates a *less* similar observation. Using the idea from above, we can say if (X_{n+1}, y) is from the same distribution as $(X_1, Y_1), \ldots, (X_n, Y_n)$ then $P(\sigma_{n+1} \leq \sigma_{(\lceil (n+1)(1-\alpha) \rceil)}) \geq 1 - \alpha$ (Although we omit the argument y to simplify notation, the conformity scores σ_i do depend on the value of y). Thus, our prediction set $C(X_{n+1})$ of level $1 - \alpha$ consists of all values of y such that

$$\sum_{i=1}^{n+1} 1[\sigma_i \leq \sigma_{n+1}] \leq \lceil (n+1)(1-\alpha) \rceil \tag{1}$$

Conformal inference has a few useful properties. Perhaps most important among them is the finite sample coverage guarantee. [15] not only showed that $P(Y_{n+1} \in C(X_{n+1})) \geq 1-\alpha$, but also that $P(Y_{n+1} \in C(X_{n+1})) \leq 1-\alpha+\frac{1}{n+1}$ under the weak assumption that the residuals have a joint continuous distribution. Crucially, this result holds even when the model $\hat{m}(x)$ is not the true form of the data. Therefore, regardless of choice of the regression model, we are guaranteed that the prediction set is neither too conservative nor anti-conservative.

In general, a grid search must be used to construct the conformal prediction interval $C(X_{n+1})$ as defined in (1). One has to check every potential value y and the pair (X_{n+1}, y) to determine if its conformity score meets the cutoff. Such an approach would be unusable in most real world applications. To address this problem, [15] proposes a split conformal prediction method which reduces computational cost by dividing the data into a training set and a ranking set. However, this algorithm often results in larger intervals than the exact prediction set. Fortunately, a closed form expression for the prediction set may be obtained with appropriate choice of conformity score.

The finite sample coverage guarantee mentioned earlier $P(Y_{n+1} \in C(X_{n+1})) \geq 1 - \alpha$ is over the joint distribution of $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$. A stronger claim of interest would be $P(Y_{n+1} \in C(x) | X_{n+1} = x) \geq 1 - \alpha$ (we will call this type of coverage *conditional coverage*). Unfortunately, [16] showed a non-trivial (not infinite length) finite sample guarantee for conditional coverage is impossible in the nonparametric setting. At best, we can achieve asymptotic conditional validity: $\sup_x [P(Y_{n+1} \notin C(x) | X_{n+1} = x) - \alpha]_+ \xrightarrow{p} 0$. Later [15] shows that for i.i.d. (X_i, Y_i) with homogenous and symmetric noise, and a base estimator $\hat{m}(x)$ which is consistent and stable under small perturbations, the conformal prediction sets will have near optimal length and location. Therefore, the choice and estimation of the conformity score σ_i plays a crucial role in the application of conformal prediction methods.

2.2 Multiple functional regression and the conformity score

Assume that we observe i.i.d pairs $(\boldsymbol{X}_1, Y_1), \ldots (\boldsymbol{X}_n, Y_n)$, where $\boldsymbol{X}_i = \{X_{ik}(t), t \in \mathcal{T} \subset \mathcal{R}\}_{k=1,\ldots,p}$ are multiple functional predictors and $Y_1, \ldots, Y_n$ are scalar responses. We are provided new observed functions $\boldsymbol{X}_{n+1}$ and wish to predict the value of Y_{n+1}.

In the case of a single functional predictor, i.e., $p = 1$, many common approaches to the functional regression problem are based on principal components analysis (FPCA) ([21], [9]).

In particular, several authors have published work on FPCA based nonparametric functional regression for single functional predictor. Early work by [8] focused on the Nadaraya-Watson type estimator $\hat{m}(x) = \frac{\sum_{i=1}^n K(((x-X_i))/h) Y_i}{\sum_{i=1}^n K(((x-X_i))/h)}$, where $((\cdot))$ is a semi-metric on $L^2(T)$, K is a univariate kernel, and h a scalar bandwidth.

Given that the principal component functions ϕ_j form an orthonormal basis, the function

$$((x - X_i))_D = \sqrt{\sum_{j=1}^{D} \langle x - X_i, \phi_j \rangle^2}$$

is a semi-norm. This function is simply the Euclidean distance between the first D principal component scores of x and X_k. [1] extended the model to the local linear model minimizing the weighted square error $\sum_{i=1}^{n}(Y_i - a - \langle \beta, x - X_i)\rangle^2 K(((x - X_i))/h)$ over a and $\beta \in L^2$. Based on FPCA, one can express β and $x - X_i$ in terms of ϕ_j, and truncate the infinite sum after D terms, thus having to minimize the sum $\sum_{i=1}^{n}(Y_i - a - \sum_{j=1}^{D} b_j(\xi_{ij} - \xi_j))^2 K(((x - X_i))/h)$ over $a, b_1, \ldots, b_D$.

When we have multiple functional predictors, we can naturally use multivariate FPCA methods to extract principal component scores. Formally, consider a set of random functions $\boldsymbol{X} = \{X_k(t)\}_{k=1,\ldots,p}$, $t \in \mathcal{T} \subset \mathcal{R}$ in Hilbert space $\mathbb{H}$, with each X_k square integrable, means $\mu_1(t) \ldots \mu_p(t)$ and covariance function $\boldsymbol{G}(s,t) = \{G_{kl}(s,t)\}_{1 \le k,l \le p}$, $G_{kl}(s,t) = cov(X_k(s), X_l(t))$, $\boldsymbol{G}_k = (G_{k1}, \ldots, G_{kp})^T$. The autocovariance operator is

$$(A\boldsymbol{f})(t) = \int_{s \in \mathcal{T}} \boldsymbol{f}(s)\boldsymbol{G}(s,t)\, ds = \begin{pmatrix} \langle \boldsymbol{G}_1(s,\cdot), \boldsymbol{f} \rangle_{\mathbb{H}} \\ \vdots \\ \langle \boldsymbol{G}_p(s,\cdot), \boldsymbol{f} \rangle_{\mathbb{H}} \end{pmatrix}$$

with orthonormal eigenfunctions $\boldsymbol{\phi}_j = (\phi_{1j}, \ldots, \phi_{pj})^T$, $j \ge 1$ and ordered eigenvalues $\lambda_1 \ge \lambda_2 \ge \ldots$. Then for $j > 1$, the j-th functional principal component score is $\xi_j = \sum_{k=1}^{p} \int_{t \in \mathcal{T}} (X_k(t) - \mu_k(t))\phi_{kj}(t)dt$. In practice, we observe our functions $\boldsymbol{X}_i$ at discrete time points $t_1, \ldots, t_d$, for $i = 1, \ldots, n$ (assume that they are centered to have mean zero), and we may estimate the covariance function by the sample covariance matrix $\hat{G} = n^{-1}\boldsymbol{X}^T\boldsymbol{X}$. Existing methods for multivariate PCA may then be used to obtain discrete approximations for the eigen-functions $\hat{\phi}_j(t)$, and principal component scores $\hat{\xi}_{ij}$ are obtained from numerical integration.

Based on the mFPCA, we were able to extend the local linear nonparametric functional regression to the multiple functional regression setting, and derived exact conformal prediction sets (algorithm in the Appendix). The numerical results for this mFPCA based local linear approach are included in the simulations (LL-mFPCA). Unfortunately, local linear regression does not perform well in more than a few dimensions, and it is difficult to express the information in multiple functional predictors with only a few principal components. While marginal coverage is always obtained, these prediction sets may be unnecessarily large.

Papers such as [23] and [13] have tried to develop models with structural assumptions such as additive or single index models for multiple functional predictors. In principle, these models can all be used to construct conformity scores, but the derivation of exact conformal prediction sets (as opposed to a grid search) is case by case, and the efficiency of the constructed prediction intervals depends on the true nature of the model and whether the added restrictions are met.

While functional PCA is a natural choice, other methods of extracting feature sets exist. The signature method, one such approach first described by [5], has recently been applied to rough path theory in areas of machine learning [11, 6]. If we denote the k-th order Signature terms of $\boldsymbol{X}(t)$ as $[S(X_t)]^k$ (details given in the next section), and $S(X_t)$ the collection of all orders, we have the following approximate equalities:[1]

$$\begin{aligned} & f(\boldsymbol{X}(t)) \\ \approx & \tilde{f}(S(X_t)) \\ = & L(S(X_t)) \\ \approx & L([S(X_t)]^1, [S(X_t)]^2, \ldots, [S(X_t)]^k) \\ = & L([S(X_t)]^k \setminus S(X_t)^{1,1,\ldots,1}) \end{aligned}$$

where f and $\tilde{f}$ are unknown regression functions, and L represents an unknown linear function. The ability to transform any nonlinear function of multiple functional predictors to a linear function of its Signature terms makes this approach appealing in our problem.

In the next section, we describe the details of constructing a Signature-based conformity score for multiple functional regression, which offers modeling flexibility and algorithmic efficiency.

3. SIGNATURE-BASED CONFORMITY SCORE

3.1 Background on signature expansion

In the field of rough path research, a d-dimensional path is defined as a continuous mapping from $[a, b]$ to $\mathbb{R}^d$. A single function $X(t)$ can be considered a specific case of a 2-dimensional path $X_t = \{X_t^1, X_t^2\} = \{t, X(t)\}$. Note that a 1-dimensional path alone is insufficient to capture the information contained within $X(t)$, as the Signature of the 1-dimensional path only depends on the image of the mapping. Similarly, the p functions $X_1(t), X_2(t), \ldots X_p(t)$, if defined on the same domain, can be considered a $p+1$-dimensional path $X_t = \{t, X_1(t), \ldots, X_p(t)\}$.

Before defining the Signature of a path, we will first define the path integral. The path integral of a 1-dimensional path Y_t against another 1-dimensional path X_t is defined as the integral

$$\int_a^b Y_t dX_t = \int_a^b Y_t \frac{dX_t}{dt} dt.$$

[1] By Uniqueness of the signature, the shuffle property, the finite order approximation, and the shuffle property in the functional data context, respectively.

Noting that the path integral $Z_s = \int_a^s Y_t dX_t$ is itself a 1-dimensional path, we can define the Signature of a d-dimensional path iteratively.

For all $i \in \{1, 2, \ldots, d\}$:

$$S(X_t)^i_{a,t} = \int_{a<s<t} dX^i_s,$$

and for $k \geq 2$, $i_1, i_2, \ldots, i_k \in \{1, 2, \ldots, d\}$, the k-fold iterated integral is defined as:

$$S(X_t)^{i_1,i_2,\ldots,i_k}_{a,t} = \int_{a<s<t} S(X_t)^{i_1,i_2,\ldots,i_{k-1}}_{a,s} dX^{i_k}_s$$

Defining $S(X_t)^0_{a,b}$ to be 1, the Signature of a d-dimensional path X_t, denoted by $S(X_t)_{a,b}$ is the infinite sequence

$$\begin{aligned} S(X_t)_{a,b} = \{1,& \\ & S(X_t)^1_{a,b}, \ldots, S(X_t)^d_{a,b}, \\ & S(X_t)^{1,1}_{a,b}, S(X_t)^{1,2}_{a,b}, \ldots, S(X_t)^{d,d}_{a,b}, \\ & S(X_t)^{1,1,1}_{a,b}, \ldots, S(X_t)^{d,d,d}_{a,b} \\ & \ldots \\ & \} \end{aligned}$$

where all d^k k-fold iterated integrals with unique superscripts are included for $k = 1, 2, \ldots$. If we use $[S(X_t)]^k_{a,b}$ to denote all k-fold iterated integrals with unique superscripts (corresponding to one line in the above equation), then the Signature $S(X_t)_{a,b} = (1, [S(X_t)]^1_{a,b}, [S(X_t)]^2_{a,b}, \ldots)$.

[5] provides a derivation and analysis of iterated integrals, while [11] show how this definition of the Signature naturally arises from the study of controlled differential equations. Although it may be difficult to intuit the meaning of higher order terms in the Signature, the terms of order 1 and 2 have fairly straightforward geometric interpretations. The first order terms $S(X_t)^i_{a,b}$ are simply the displacements $X^i_b - X^i_a$ in each dimension. The second order terms $S(X_t)^{i,i}_{a,b}$ are half the square displacement $(X^i_b - X^i_a)^2/2$ in each dimension. The cross second order terms $S(X_t)^{i,j}_{a,b}$ satisfy the equation $A_{ij} = (S(X_t)^{i,j}_{a,b} - S(X_t)^{j,i}_{a,b})/2$, where A_{ij} is the signed area enclosed by the two-dimensional path $\{X_i(t), X_j(t)\}$ and the chord connecting the endpoints.

3.2 Properties of the signature in functional regression setting

Without loss of generality, we will consider functions on the domain $[0, 1]$. For convenience, we will ignore the domain and path and write a Signature term $S(X_t)^{i_1,\ldots,i_k}_{a,b}$ as $S^{i_1,\ldots,i_k}$ when there is no confusion. Also, we use $[S(X_t)]^k_{a,b}$ or $[S]^k$ to represent all k-th order Signature terms.

Immediately clear from its definition is that the Signature is invariant under translation. The following corollary was first shown by [5] for all continuously differentiable functions and later expanded to paths of bounded variance by [11].

Corollary: If, for two piecewise regular continuous paths X_t and Y_t in $\mathbb{R}^d$, $S(X_t)_{a,b} = S(Y_t)_{a,b}$ then the irreducible path of Y_t can be obtained from the irreducible path of X_t by translation and change of parameter.

In the context of functional regression, our functions cannot cross themselves; therefore the corresponding paths are irreducible. Furthermore, change of a path's parameter does not change the function. Thus, each unique Signature corresponds to a unique family of functions which only differ by vertical and horizontal translation. If all functions in a data set have the same domain, then vertical location is the only information lost.

The Signature method is a promising approach for functional nonparametric regression because of the *shuffle property*. First proved by [19], the shuffle property states that any product of two terms $S^{i_1,i_2,\ldots,i_k}$ and $S^{j_1,j_2,\ldots,j_n}$ can be expressed as the sum of terms with the multi-indices $i_1, i_2, \ldots, i_k, j_1, j_2, \ldots, j_n$. Specifically, for $I = (i_1, i_2, \ldots, i_k)$ and $J = (j_1, j_2, \ldots, j_n)$,

$$S^I S^J = \sum_{K \in I \sqcup\!\sqcup J} S^K$$

where $I \sqcup\!\sqcup J$ is the set of all $\frac{(k+n)!}{k!n!}$ ways to interleave the elements of I and J. This result allows us to express any nonlinear function of the Signature as a linear combination of the Signature instead. Consider a setting with two functional predictors $X_1(t)$ and $X_2(t)$ on $t \in [0, 1]$. The first two orders of our 3 dimensional path's Signature would be the sequence $\{1, S^1, S^2, S^3, S^{1,1}, S^{1,2}, S^{1,3}, S^{2,1}, S^{2,2}, S^{2,3}, S^{3,1}, S^{3,2}, S^{3,3}\}$. This shuffle property states that, for example:

$$\begin{aligned} S^{\mathbf{1},\mathbf{2}} * S^{2,3} &= S^{\mathbf{1},\mathbf{2},2,3} + S^{1,\mathbf{2},\mathbf{2},3} + S^{\mathbf{1},2,3,\mathbf{2}} + S^{2,\mathbf{1},\mathbf{2},3} \\ &\quad + S^{2,\mathbf{1},3,\mathbf{2}} + S^{2,3,\mathbf{1},\mathbf{2}} \\ S^{1,2} * S^{2,3} &= 2S^{1,2,2,3} + S^{1,2,3,2} + S^{2,1,2,3} + S^{2,1,3,2} \\ &\quad + S^{2,3,1,2}, \end{aligned}$$

where indices corresponding to those from the first term are bolded in the top equality to help demonstrate an interleaving of superscripts.

The Signature as defined above is an infinite sequence, which is not practical for regression problems. We must therefore build the regression function based on a finite order of Signature terms. As higher order Signature terms in some sense correspond to more complex features of the original function, this choice seems reasonable.

In functional data regression, we can further reduce the number of terms needed for regression. Since we assume the domain of the sample functions does not change, the Signature terms $S^1_i, S^{11}_i, S^{111}_i, \ldots$, which only depend on the first 1-dimensional path $\{t\}$, will take the same value for all sample functions, and thus are unnecessary to include in our

feature set. This fact, combined with the *shuffle property*, allows us to express all Signature terms of order $1, 2, \ldots, k-1$ as linear combinations of Signature terms of order k. For example, the shuffle property states $S^1 S^2 = S^{12} + S^{21}$. Since $S^1 \equiv c$, we can rearrange the equation to say $S^2 = \frac{S^{12}+S^{21}}{c}$. Therefore, in our regression, we only need to include the $d^k - 1$ non-$S^{11\ldots1}$ terms of order k, i,e., $[S]^k \setminus S^{1,1,\ldots,1}$, rather than all $d^{k+1} - 1$ terms of orders $1, 2, \ldots, k$, where p is the dimension of the multiple functional predictors and $d = p+1$ is the dimension of the path.

To summarize, we can approximate any unknown regression relationship between Y and p-dimensional functional predictors $\boldsymbol{X}(t)$ through a linear function of the k-th order Signature terms of the $p+1$-dimensional path. Of course, any dependence the response has on vertical location will not be captured by this feature set. As such, additional predictors capturing the vertical location of each predictor function should be added to the Signature expansion.

Let $[S]^k$ be the k-th order Signature expansion for $X_t = \{t, X_1(t), \ldots, X_p(t)\}$, and Z_j be scalar predictors which capture the vertical location of each predictor function $X_j(t)$, $j = 1, \ldots, p$. For example, Z_j can be the sum of $X_j(t)$ over all t. One possibility for the semiparametric model would be a multiple index model, i.e., $Y = f([S]^k\beta, Z_1, \ldots, Z_p) + \epsilon$, where f is an unknown function. This model is very flexible, but models of this sort are often fit iteratively, and we could not find a closed-form solution for conformal prediction interval. One could combine this model with the sample-splitting method described in [15] to obtain an approximate conformal prediction set based on this model.

We propose to use the Signature-based multiple partial linear model

$$Y_i = [S_i]^k\beta + \sum_{j=1}^{p} g_j(Z_{ji}) + \epsilon_i, \tag{2}$$

where g_js are unknown functions to be estimated nonparametrically. This model makes the additional assumption that the vertical locations $\boldsymbol{Z}$ have additive effects on the response. We view this model as balance between model flexibility and simplicity.

3.3 Derivation of the exact prediction intervals

To derive the conformal prediction interval, we need to fit the model (2) on the augmented data set $\{(\boldsymbol{X}_1(t), Y_1), \ldots, (\boldsymbol{X}_{n+1}(t), Y_{n+1})\}$, and obtain the residuals $\{r_1, \ldots, r_{n+1}\}$. Let $Y = (Y_1, \ldots, Y_{n+1})^T$ and $[S]^k = ([S]^k_1, \ldots, [S]^k_{n+1})^T$. Extending the work of [20], we first fit the additive models $Y_i \sim \alpha_1 + g_{y1}(Z_{1i}) + g_{y2}(Z_{2i}) + \cdots + g_{yp}(Z_{pi})$ and $[S_i]^k \sim \alpha_2 + g_{s1}(Z_{1i}) + g_{s2}(Z_{2i}) + \cdots + g_{sp}(Z_{pi})$ using the local constant approach with an exact equation derived by [18]. These residuals will retain their linear relationship while having no remaining dependence on $\boldsymbol{Z}$. Thus, we can use OLS to regress the residuals from the first regression on the residuals from the second. Using the notation from the [18], the equations for the Y-residuals and S-residuals are:

$$\begin{aligned}
\hat{r}_{yi} &= Y_i - \hat{g}_y(\boldsymbol{Z}_i) = Y_i - \boldsymbol{W}_{M_i,}Y \\
&= Y_i - \sum_{j=1}^{n} \boldsymbol{W}_{M_{i,j}} Y_j - \boldsymbol{W}_{M_{i,n+1}} Y_{n+1} \\
\hat{r}_{si} &= [S_i]^k - \hat{g}_s(\boldsymbol{Z}_i) = [S_i]^k - \boldsymbol{W}_{M_i,}[S]^k
\end{aligned}$$

As defined by [18], $\boldsymbol{W}_M$ is the smoother matrix satisfying $\hat{g}_y = \boldsymbol{W}_M Y$ in the regression of Y on $\boldsymbol{Z}$ and $\hat{g}_s = \boldsymbol{W}_M[S]^k$ in the regression of $[S]^k$ on $\boldsymbol{Z}$. Here, $\boldsymbol{W}_{M_{i,}}$ denotes the i-th row of $\boldsymbol{W}_M$ and $\boldsymbol{W}_{M_{i,j}}$ denotes the entry of $\boldsymbol{W}_M$ in the i-th row and j-th column, and $\hat{r}_{si}$ is a vector as $[S_i]^k$ is a vector.

Our estimate for β in (2) is

$$\begin{aligned}
\hat{\beta} &= (\sum_{i=1}^{n+1} \hat{r}_{si}\hat{r}'_{si})^{-1} \sum_{i=1}^{n+1} \hat{r}_{si}\hat{r}_{yi} \\
&= (\sum_{i=1}^{n+1} \hat{r}_{si}\hat{r}'_{si})^{-1} \sum_{i=1}^{n+1} \hat{r}_{si}(Y_i - \sum_{j=1}^{n} \boldsymbol{W}_{M_{i,j}} Y_j \\
&\quad - \boldsymbol{W}_{M_{i,n+1}} Y_{n+1}) \\
&= (\sum_{i=1}^{n+1} \hat{r}_{si}\hat{r}'_{si})^{-1} \sum_{i=1}^{n} \hat{r}_{si}(Y_i - \sum_{j=1}^{n} \boldsymbol{W}_{M_{i,j}} Y_j) \\
&\quad - (\sum_{i=1}^{n+1} \hat{r}_{si}\hat{r}'_{si})^{-1} \hat{r}_{s,n+1} \sum_{j=1}^{n} \boldsymbol{W}_{M_{n+1,j}} Y_j \\
&\quad - (\sum_{i=1}^{n+1} \hat{r}_{si}\hat{r}'_{si})^{-1} \sum_{i=1}^{n} \hat{r}_{si} \boldsymbol{W}_{M_{i,n+1}} Y_{n+1} \\
&\quad + (\sum_{i=1}^{n+1} \hat{r}_{si}\hat{r}'_{si})^{-1} \hat{r}_{sn+1}(1 - \boldsymbol{W}_{M_{n+1,n+1}}) Y_{n+1} \\
&= c_1 + c_2 Y_{n+1}
\end{aligned}$$

where c_1 and c_2 are terms not dependent on i or Y_{n+1}.

Thus, the equation for the i^{th} residual is

$$\begin{aligned}
r_i &= Y_i - \hat{Y}_i \\
&= Y_i - \hat{r}_{si}\hat{\beta} - \hat{g}_y(\boldsymbol{Z}_i) \\
&= Y_i - \hat{r}_{si}(c_1 + c_2 Y_{n+1}) - (\sum_{j=1}^{n} \boldsymbol{W}_{M_{ij}} Y_j + \boldsymbol{W}_{M_{i,n+1}} Y_{n+1}) \\
&= Y_i - a_i - b_i Y_{n+1}
\end{aligned} \tag{3}$$

where a_i and b_i are terms not dependent on Y_{n+1}.

We wish to include in our prediction set every y for which $\sum_{i=1}^{n+1} 1[\sigma_i \leq \sigma_{n+1}] \leq \lceil (n+1)(1-\alpha) \rceil$, or equivalently $\sum_{i=1}^{n+1} 1[|r_i| \leq |r_{n+1}|] \leq \lceil (n+1)(1-\alpha) \rceil$. As we just derived, $|r_{n+1}| > |r_i| \Leftrightarrow |Y_{n+1} - a_{n+1} - b_{n+1}Y_{n+1}| > |Y_i - a_i - b_i Y_{n+1}|$.

Solving this inequality leads to the following four inequalities:

Inequalities to determine upper bound:

$$\begin{cases}(1-b_{n+1}+b_i)Y_{n+1} < Y_i - a_i + a_{n+1}\\ (1-b_{n+1}-b_i)Y_{n+1} < -Y_i + a_i + a_{n+1}\end{cases} \tag{4}$$

Inequalities to determine lower bound:

$$\begin{cases}(1-b_{n+1}+b_i)Y_{n+1} > Y_i - a_i + a_{n+1}\\ (1-b_{n+1}-b_i)Y_{n+1} > -Y_i + a_i + a_{n+1}\end{cases} \tag{5}$$

This results in the following algorithm for constructing the exact conformal prediction intervals using the Signature-based multiple partial linear conformity score:

Signature-based MPL Algorithm (Algorithm 1)

1. Calculate the k-th order Signature $[S_i]^k$ of each X_i, $i = 1,\ldots,n+1$. Construct the matrix $[\boldsymbol{S}]^k = ([S_1]^k,\ldots,[S_{n+1}]^k)^T$. You may append any linear scalar covariates to this matrix.
2. Calculate Z_{ij}, for each $X_{ij}(t)$, $i = 1,\ldots,n+1, j = 1,\ldots,p$. Construct the matrix $\boldsymbol{Z} = (\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_{n+1})^T$. You may append any additive nonlinear scalar covariates to this matrix.
3. Calculate $\hat{r}_{yi}$, $\hat{r}_{Si}$, $\hat{\beta}$, and a_i, b_i as described in Eq. (3).
4. For each $i \in \{1,2,\ldots,n\}$ calculate $U_i = \max((Y_i - a_i + a_{n+1})/(1-b_{n+1}+b_i), (-Y_i + a_i + a_{n+1})/(1-b_{n+1}-b_i))$ and $L_i = \min((Y_i - a_i + a_{n+1})/(1-b_{n+1}+b_i), (-Y_i + a_i + a_{n+1})/(1-b_{n+1}-b_i))$
5. Construct the $100(1-\alpha)\%$ prediction interval for $Y_{n+1} = (L_{(\lfloor \alpha(n+1) \rfloor)}, U_{(\lceil (1-\alpha)(n+1) \rceil)})$

The conformity score we use here is constructed on the augmented data, and is symmetric in the entries in its first argument. Therefore, we have guaranteed finite sample coverage following the general proof in Theorem 2.1 in [15].

Lemma 3.1. *If the observations $(\boldsymbol{X}_1, Y_1),\ldots(\boldsymbol{X}_n, Y_n)$ are i.i.d, the prediction interval produced by Algorithm 1 will have coverage at least $100(1-\alpha)\%$ for the a new i.i.d pair $(\boldsymbol{X}_{n+1}, Y_{n+1})$. If the residuals have a joint continuous distribution, the coverage will be at most $100(1-\alpha+\frac{1}{n+1})\%$.*

To ensure the resulting set is a contiguous interval, this algorithm assumes a **Contiguity Condition**

$$1 - b_{n+1} \pm b_i > 0, \forall i. \tag{6}$$

To better understand this condition, let us consider a simpler setting. It is possible to derive a similar condition in the simple linear regression problem. In that case, the analogous condition is $1 - H_{n+1,n+1} \pm H_{i,n+1} > 0, \forall i$, where $H_{i,j}$ is the entry in the i-th row and j-th column of the augmented hat matrix $H = X(X'X)^{-1}X'$. As $H_{i,i}$ is the leverage of the i-th observation, $H_{n+1,n+1}$ will be large when X_{n+1} is far from the other data points. Additionally, one can see that the magnitude of $H_{i,n+1}$ will be large when X_i and X_{n+1} are both (approximately equally) far from the other data points. Therefore, the restriction roughly translates into a requirement that X_{n+1} not be "too far" away from the rest of the X_i. Back to the multiple partial linear approach, b_i similarly measures the leverage of the i-th observation, and the restriction translates into a requirement that X_{n+1} not be "too far" away from the rest of the X_i. If this condition is not satisfied, an extract conformal prediction set can still be constructed with a more complex algorithm, although the set is not a contiguous interval (see **Algorithm 1b** in the Appendix).

Remark: [15] contains four assumptions necessary for results pertaining to asymptotic conditional coverage and interval efficiency. While readers are referred to the source paper for detailed assumptions; we briefly discuss the implication of these assumptions in the context of the Signature-based multiple function regression. As noted in the source paper, Assumption 1 (i.i.d. data) and Assumption 2 (symmetric noise) are relatively weak assumptions, and the symmetric noise assumption can even be dropped, but is included for convenience.

Following the discussion after Eq. (6), it is obvious that b_i is the degree to which the value of Y_{n+1} affects r_i and $\hat{y}_i$. Thus, the contiguity condition is similar to Assumption 3 (the perturb one sensitivity) from that paper.

Assumption 4 (regarding consistency of the regression estimator) does not always hold for our Signature-based MPL algorithm. While conformal sets have guaranteed marginal coverage regardless, the conditional coverage and length of the prediction interval approaches optimality in some respect when the additive modeling assumption on the vertical locations (Z_j) as well as the finite order approximation hold.

There are not many results concerning the finite order approximation of the Signature. The few results which do appear are within the context of Controlled Differential Equations. Expanding on earlier work in [17], focusing only on Linear Controlled Differential Equations, [2] studied Controlled Differential Equations of the form $dY_t = f(Y_t)dX_t$, $Y_0 = y_0$, and derived the finite order approximation error.

In the functional regression setting, this result implies that the difference $Y - \langle [a_i]^{1:K}, [S]^{1:K} \rangle$ for some coefficients $[a_i]^{1:K}$ is approximately proportional to $\frac{1}{K!}$ for large K. Since the conformity score depends on the Signature terms in a linear form, we are able to include a relatively large number of order k and therefore the constructed prediction intervals have empirical efficiency (measured by the length of the intervals).

3.4 Computational aspects

The ESig python package provides various tools relating to the signature. Most importantly, it includes a

function to convert any path into its signature expansion, up to any order (https://github.com/kormilitzin/the-signature-method-in-machine-learning). [24] discusses some practical considerations of computing the signature in Section 3.2.

The R package for the Signature-based MPL algorithm developed in this paper will be publicly available on authors website and all simulation and data code are available upon request. The various algorithms discussed in this paper require selection of a few tuning parameters. The mFPCA methods require the user to select the number of principal components retained, while the signature based method requires selection of signature order. Additionally, the nonparametric and semiparametric methods also necessitate bandwidth selection. We have found best performance when using the K-fold cross validation, with the "one standard-error" rule, as described in section 7.10.1 of [12]. This preference for slightly simpler models increases the likelihood of satisfying the Contiguity Condition and thus producing contiguous prediction intervals. Since we derive the exact conformal prediction set, the algorithm is very fast, and is much more computationally efficient than any exhaustive grid search algorithm otherwise required for conformal prediction.

4. SIMULATIONS

To test the performance and robustness of the proposed Signature-based multiple partial linear algorithm, we performed multiple simulations under a variety of conditions. Foremost, we wanted to empirically confirm the finite sample marginal coverage guarantee. To accomplish this, we performed 1000 simulations at each combination of settings and recorded the exact coverage of the prediction interval constructed at a random X_{n+1} for each simulation. Note that in simulations, we know the theoretical conditional distribution of Y_{n+1} given X_{n+1}, so that we can compute the exact coverage for a prediction interval constructed by our proposed algorithms. In the tables below are the average coverage under each combination of settings. The second goal was to explore the efficiency of the algorithms' resulting intervals. As noted previously, the algorithms will produce near-optimally efficient intervals under certain conditions. To measure efficiency, the lengths of the prediction intervals were recorded. The following tables include median interval length for each combination of settings. We performed simulations using different functions for the relationship between Y and multiple predictors, using different sample sizes, and using different error distributions. We included the results for the Signature-based partial linear conformity score, as well as the mFPCA-based local linear conformity score. As a benchmark, we also included the asymptotic prediction interval $(\hat{m}(x_{n+1}) - Q_{\alpha/2}, \hat{m}(x_{n+1}) + Q_{1-\alpha/2})$, where the function $m(x)$ and the quantiles Q were estimated as if we knew the simulation settings. Therefore, we call this a naive "oracle".

Table 1. mFPCA based linear relationship: We report the mean coverage (SD) of the constructed intervals from 1000 simulations for Signature-based multiple partial linear conformity score, mFPCA-based local linear conformity score, and naive "oracle"

Error	N	MPL-Sig	LL-mFPCA	Naive oracle
N(0,1)	200	0.949 (0.063)	0.946 (0.116)	0.939 (0.022)
	800	0.947 (0.058)	0.955 (0.089)	0.947 (0.009)
t_3	200	0.949 (0.050)	0.947 (0.091)	0.943 (0.021)
	800	0.953 (0.026)	0.951 (0.085)	0.949 (0.008)

We extended the simulation setting from [23] to a setting with 4 functional predictors. The predictor $\boldsymbol{X}$ consisted of the following four functions:

$$X_{i1} = t + \sin(t) + \sum_{k=1}^{20} \xi_{ik}\psi_k^{(1)}(t),$$
$$X_{i2} = t + \cos(t) + \sum_{k=1}^{20} \xi_{ik}\psi_k^{(2)}(t),$$
$$X_{i3} = -t + \sin(t) + \sum_{k=1}^{20} \xi_{ik}\psi_k^{(3)}(t),$$
$$X_{i4} = -t + \cos(t) + \sum_{k=1}^{20} \xi_{ik}\psi_k^{(4)}(t),$$

where $\xi_{ik} \sim N(0, 28.96k^{-2})$, $\psi_k^{(1)} = \frac{1}{\sqrt{10}}\sin(\pi kt/10 + \pi/4)$, $\psi_k^{(2)} = \frac{1}{\sqrt{10}}\sin(\pi kt/10 + 3\pi/4)$, $\psi_k^{(3)} = \frac{1}{\sqrt{10}}\sin(\pi kt/10)$, and $\psi_k^{(4)} = \frac{1}{\sqrt{10}}\sin(\pi kt/10 + \pi/2)$, for $t \in \mathcal{T} = [0, 10]$. Independent, normally distributed measurement error with standard deviation $\sqrt{0.2}$ was added to the functional predictors on the regular grid of 100 points in $\mathcal{T} = [0, 10]$.

We generated Y in two different manners. In the first setting, we let $Y_i = 1.4 - \sum_{j=1}^{10} \frac{(-1)^j j}{25}\xi_{ij} + \epsilon_i$, and we used a functional linear regression model based on mFPCA as the naive "oracle" method. In the second setting, we let $Y_i = -0.1 + 3\zeta_{i1} + \sin(2\pi(\zeta_{i2} - 1/2)) + 8(\zeta_{i4}^2 - \frac{2}{3}\zeta_{i4}) + \epsilon_i$, where $\zeta_{ik} = \Phi(\frac{\xi_{ik}}{\sqrt{28.96k^{-2}}})$. The second setting is adapted from [23], where the authors proposed a multivariate FPCA based functional additive model for multiple functional data regression, and their PLFAM method was used to obtain $\hat{m}(x)$ in the naive "oracle". In each case, the ϵ_i either $\sim N(0, 1)$ or $\sim t_3/\sqrt{2}$. Bandwidths and number of predictors for the algorithms were chosen via 10-fold cross validation. The one standard error rule was utilized to prevent overfitting, thereby reducing the frequency of the contiguity conditions being violated.

As expected, the conformal prediction method produces prediction intervals with the desired coverage, using both the MPL-Signature and LL-mFPCA conformity scores. The naive "oracles" prediction intervals do tend to undercover,

Table 2. mFPCA based additive relationship: We report the mean coverage (SD) of the constructed intervals from 1000 simulations for Signature-based multiple partial linear conformity score, mFPCA-based local linear conformity score, and naive "oracle"

Error	N	MPL-Sig	LL-mFPCA	Naive oracle
N(0,1)	200	0.951 (0.086)	0.949 (0.100)	0.945 (0.048)
	800	0.950 (0.070)	0.951 (0.087)	0.946 (0.017)
t_3	200	0.951 (0.062)	0.950 (0.069)	0.941 (0.043)
	800	0.948 (0.055)	0.950 (0.074)	0.948 (0.012)

Table 3. mFPCA based linear relationship: We report the median length (MAD) of the constructed intervals from 1000 simulations for Signature-based multiple partial linear conformity score, mFPCA-based local linear conformity score, and naive "oracle"

Error	N	MPL-Sig	LL-mFPCA	Naive oracle
N(0,1)	200	4.85 (0.48)	6.44 (0.49)	3.90 (0.27)
	800	4.60 (0.44)	6.36 (0.21)	3.93 (0.14)
t_3	200	5.37 (0.80)	6.80 (0.60)	4.53 (0.59)
	800	5.16 (0.57)	6.71 (0.29)	4.54 (0.30)

Table 4. mFPCA based additive relationship: We report the median length (MAD) of the constructed intervals from 1000 simulations for Signature-based multiple partial linear conformity score, mFPCA-based local linear conformity score, and naive "oracle"

Error	N	MPL-Sig	LL-mFPCA	Naive oracle
N(0,1)	200	5.68 (0.58)	5.95 (0.44)	4.55 (0.32)
	800	5.21 (0.47)	5.77 (0.21)	4.03 (0.13)
t_3	200	6.08 (0.73)	6.31 (0.57)	4.92 (0.56)
	800	5.66 (0.60)	6.27 (0.27)	4.59 (0.29)

especially in smaller samples, but still serves as a good lower bound for interval length.

As we can see, the local linear mFPCA based method performs decently in the mFPCA based additive setting, although slightly worse than the Signature based method. In the mFPCA based linear setting, we see much worse performance from this method, due the limited number of components it can utilize. We note that in the additive setting, the relationship between Y and $\boldsymbol{X}$ is completely determined through the first four functional principal components of $\boldsymbol{X}$, and indeed mostly captured by the first two components. So this is a low dimensional case for the mFPCA-based nonparametric regression. In the first simulation setting, however, the relationship between Y and $\boldsymbol{X}$ can not be well captured if we only represent $\boldsymbol{X}$ using the first few functional principal components. As a result, the nonparametric regression based on mFPCA produced larger intervals that still had valid coverage.

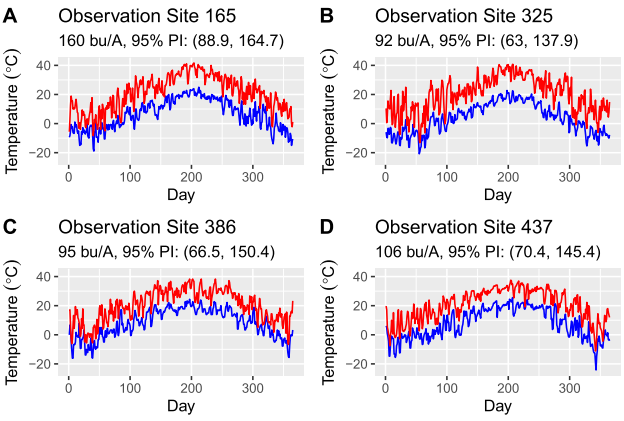

Figure 1. Daily minimum (blue) and maximum (red) temperatures for 4 randomly selected observation sites. Actual corn yield in bushels per acre reported alongside 95% out of sample prediction interval.

The Signature based partial linear algorithm performs well. The length of intervals is a bit larger than the "oracle", but with more accurate coverage. It has great flexibility to capture the relationship between Y and $\boldsymbol{X}$, performs decently in small samples, and deals with fat-tailed error fine. Overall we saw a small reduction in interval size when increasing n, which likely corresponds to more accurately estimating the mean function $\hat{m}(x)$. Regardless of setting, the finite sample coverage is guaranteed for conformal methods.

5. CROP YIELD DATA

To illustrate our method, we analyzed the crop yield dataset described in [23]. This dataset consists of several county-level corn and soybean yield related variables from 1999 to 2011, as well as annual averaged precipitation, daily maximum temperature, and daily minimum temperature. The raw dataset was from the National Agricultural Statistics Agency (https://quickstats.nass.usda.gov/) and the National Climatic Data Center (https://www.ncdc.noaa.gov/data-access). Following the source paper, we let Y be the average crop yield per acre for a specific year and county, $X_1(t)$ and $X_2(t)$ be the daily maximum and minimum temperatures for the same year and county, and following the source paper, we also added additional scalar covariates including the proportion of irrigated land in that county and for that particular type of crop, averaged annual precipitation, the interaction between the two, and a year indicator. In Figure 1 we display both predictor functions of four randomly selected observations.

Using our Signature-based multiple partial linear conformity score, we constructed the out of sample 95% prediction interval for each observation in the corn and soy datasets. In line with expectations, 95.1% of the actual

corn and soy yields fell within the corresponding interval. The median interval length was 75.7 for the corn data and 31.6 for the soy data. To put these numbers in perspective, the [23] paper produced a weighted mean square prediction error of 298.43 for the corn data and 35.64 for the soy data (weights correspond to size of harvested land). We also tried the mFPCA-based local linear conformity score for comparison. This method also produced intervals with 95.1% empirical coverage, but the median intervals for corn and soy were longer, at 116.7 and 37.4, respectively.

APPENDIX A

Signature-based MPL Algorithm for Non-contiguous Prediction Sets (Algorithm 1b)

1. Calculate the k-th order Signature $[S_i]^k$ of each X_i, $i = 1, \ldots, n+1$. Construct the matrix $[\boldsymbol{S}]^k = ([S_1]^k, \ldots, [S_{n+1}]^k)^T$
2. Calculate Z_{ij}, for each $X_{ij}(t)$, $i = 1, \ldots, n+1, j = 1, \ldots, p$. Construct the matrix $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{n+1})^T$
3. Calculate $\hat{r}_{yi}$, $\hat{r}_{Si}$, $\hat{\beta}$, and a_i, b_i as described in Eq. (3).
4. For each $i \in \{1, 2, \ldots, n\}$ calculate $c_i = (Y_i - a_i + a_{n+1})/(1 - b_{n+1} + b_i)$ and $d_i = (-Y_i + a_i + a_{n+1})/(1 - b_{n+1} - b_i)$
5. Construct n regions in the following manner:
   If $\text{sgn}(1 - b_{n+1} + b_i) = +$ and $\text{sgn}(1 - b_{n+1} - b_i) = +$, then your region is $(\min(c_i, d_i), \max(c_i, d_i)]$.
   If $\text{sgn}(1 - b_{n+1} + b_i) = -$ or $\text{sgn}(1 - b_{n+1} - b_i) = -$, then your region is $(-\infty, \min(c_i, d_i)] \cup [\max(c_i, d_i), \infty)$.
6. Take the intersection of unions of all combinations of $\lceil (n+1)(1-\alpha) \rceil$ of these sets. This resulting set is the $100(1-\alpha)\%$ prediction interval for Y_{n+1}.

mFPCA-based Local Linear Algorithm (Algorithm 2)

1. Calculate the D eigenfunctions $\hat{\phi}_1, \ldots, \hat{\phi}_D$ corresponding to the D largest eigenvalues where $1 \leq D \leq rank(X)$. Let $\phi = (\hat{\phi}_1, \ldots, \hat{\phi}_D)^T$.
2. Calculate $\xi = (\hat{\xi}_1, \ldots, \hat{\xi}_D)^T = X\phi^T$, the scores of $X_1, \ldots, X_{n+1}$ with respect to the basis $\hat{\phi}_1, \ldots, \hat{\phi}_D$.
3. Calculate A_i and B_i for all i, where A and B are defined below.
4. For all $i \in 1, 2, \ldots, n$, calculate
   $L_i = \min(\frac{y_i - A_i + A_{n+1}}{1 - B_{n+1} + B_i}, \frac{-y_i + A_i + A_{n+1}}{1 - B_{n+1} - B_i})$ and
   $U_i = \max(\frac{y_i - A_i + A_{n+1}}{1 - B_{n+1} + B_i}, \frac{-y_i + A_i + A_{n+1}}{1 - B_{n+1} - B_i})$.
5. Construct the $100(1-\alpha)\%$ prediction interval for $Y_{n+1} = (L_{(\lfloor (n+1)\alpha \rfloor)}, U_{(\lceil (1-\alpha)(n+1) \rceil)})$

To calculate A and B in step 3, first construct the matrices $Z_x = [\mathbf{1} \;\; \xi_i - \xi_x]$ and $W_x = diag(K(||\xi_i - \xi_x||/h))$, where $((\cdot))$ is a semi-metric on $L^2(T)$, K is a univariate kernel, and h a scalar bandwidth. Then $\hat{m}(x) = e_1(Z_x' W_x Z_x)^{-1} Z_x' W_x Y$, where e_1 is the $D+1$ dimensional vector with a 1 followed by D zeroes. Then

$$A_i = e_1(Z_{x_i}' W_{x_i} Z_{x_i})^{-1} X_{x_i[1:n,]}' W_{x_i[1:n,1:n]} Y_{[1:n]}$$
$$B_i = e_1(Z_{x_i}' W_{x_i} Z_{x_i})^{-1} X_{x_i, n+1} W_{x_i, n+1}$$

Received 25 February 2020

REFERENCES

[1] Baíllo, A. and Grané, A. (2009). Local linear regression for functional predictor and scalar response. *Journal of Multivariate Analysis* **100** 102–111. MR2460480

[2] Boedihardjo, H., Lyons, T., Yang, D. et al. (2015). Uniform factorial decay estimates for controlled differential equations. *Electronic Communications in Probability* **20**. MR3438739

[3] Cardot, H., Crambes, C. and Sarda, P. (2005). Quantile regression when the covariates are functions. *Nonparametric Statistics* **17** 841–856. MR2180369

[4] Chen, K. and Müller, H.-G. (2012). Conditional quantile analysis when covariates are functions, with application to growth data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74** 67–89. MR2885840

[5] Chen, K.-s. (1958). Integration of paths—a faithful representation of paths by non-commutative formal power series. *Transactions of the American Mathematical Society* **89** 395–407. MR0106258

[6] Chevyrev, I. and Kormilitzin, A. (2016). A primer on the signature method in machine learning. *arXiv preprint arXiv:1603.03788*.

[7] Fan, J., Yao, Q. and Tong, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika* **83** 189–206. MR1399164

[8] Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media. MR2229687

[9] Hall, P., Müller, H.-G., Wang, J.-L. et al. (2006). Properties of principal component methods for functional and longitudinal data analysis. *The Annals of Statistics* **34** 1493–1517. MR2278365

[10] Hall, P., Wolff, R. C. and Yao, Q. (1999). Methods for estimating a conditional distribution function. *Journal of the American Statistical Association* **94** 154–163. MR1689221

[11] Hambly, B. and Lyons, T. (2010). Uniqueness for the signature of a path of bounded variation and the reduced path group. *Annals of Mathematics* 109–167. MR2630037

[12] Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media. MR2722294

[13] Jiang, F., Baek, S., Cao, J. and Ma, Y. (2018). A Functional Single Index Model. *Statistica Sinica*.

[14] Koenker, R., Chesher, A. and Jackson, M. (2005). *Quantile regression. Econometric Society Monographs*. Cambridge University Press. MR2268657

[15] Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J. and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association* **113** 1094-1111. MR3862342

[16] Lei, J. and Wasserman, L. (2014). Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76** 71–96. MR3153934

[17] Levin, D., Lyons, T. and Ni, H. (2013). Learning from the past, predicting the statistics for the future, learning an evolving system. *arXiv preprint arXiv:1309.0260*.

[18] Opsomer, J. D. (2000). Asymptotic properties of backfitting estimators. *Journal of Multivariate Analysis* **73** 166–179. MR1763322

[19] Ree, R. (1958). Lie elements and an algebra associated with shuffles. *Annals of Mathematics* 210–220. MR0100011

[20] Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society* 931–954. MR0951762
[21] Silverman, B. W. et al. (1996). Smoothed functional principal components analysis by choice of norm. *The Annals of Statistics* **24** 1–24. MR1389877
[22] Vovk, V., Gammerman, A. and Shafer, G. (2005). *Algorithmic learning in a random world*. Springer Science & Business Media. MR2161220
[23] Wong, R. K., Li, Y. and Zhu, Z. (2019). Partially linear functional additive models for multivariate functional data. *Journal of the American Statistical Association* **114** 406–418. MR3941264
[24] Yang, W., Lyons, T., Ni, H., Schmid, C., Jin, L. and Chang, J. (2017). Leveraging the path signature for skeleton-based human action recognition. *arXiv preprint arXiv:1707.03993*.

Ryan Kelly
University of Pittsburgh
Department of Statistics
USA
E-mail address: rmk79@pitt.edu

Kehui Chen
University of Pittsburgh
Department of Statistics
USA
E-mail address: khchen@pitt.edu