Distribution Free Prediction Intervals for Multiple Functional Regression

by

Ryan Kelly

B.A. in Mathematics and Biology, Washington University in St. Louis, 2015

M.A. in Statistics University of Pittsburgh, 2017

Submitted to the Graduate Faculty of

the Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2020

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Ryan Kelly

It was defended on

July 31st, 2020

and approved by

Kehui Chen, Department of Statistics at University of Pittsburgh

Yu Cheng, Department of Statistics at University of Pittsburgh

Satish Iyengar, Department of Statistics at University of Pittsburgh

Jing Lei, Department of Statistics at Carnegie Mellon University

Distribution Free Prediction Intervals for Multiple Functional Regression

Ryan Kelly, PhD

University of Pittsburgh, 2020

My research aims to establish a method of constructing prediction intervals for a scalar response of interest when predictors are functional data, with minimal distributional and modeling assumptions. To accommodate flexible regression relationships, we integrated nonparametric functional regression based on functional principal component analysis into our conformal prediction method, and we developed nonparametric functional regression approaches based on the signature method, which is a mathematical tool to represent the information contained in the functions by a collection of iterated integrals. 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. Conditions necessary for efficiency and contiguity of the prediction set are discussed. Finally, our methods are illustrated using simulated and real data examples.

Keywords: Multiple Functional Regression, Prediction Intervals, Conformal Prediction, Multiple Functional Principal Components Analysis, Signature Expansion.

Table of Contents

List of Tables

List of Figures

1.0 Introduction

My research aims to establish a method of constructing prediction intervals for a scalar response of interest when multiple predictors are functional data, with minimal distributional and modeling assumptions. With advancements in technology, it has become far easier and more common to continuously or repeatedly measure a variable over an interval in time to track how it changes. If we record one or more curves for each subject in our sample of interest, these data are referred to as functional data. The term is quite general, as even shapes, images, and surfaces may be considered specific types of functional data.

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 multiple sources of functional data they have collected to estimate the unknown 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. 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 are so complex. 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. In this research, we focused on developing computationally efficient methods for conformal prediction intervals in functional regression settings. The prediction intervals constructed by the conformal method have guaranteed coverage (confidence) without 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 integrated nonparametric functional regression based on multivariate functional principal component analysis into our conformal prediction method, and we developed nonparametric functional regression approaches based on the signature method, which is a mathematical tool to represent the information contained in the functions by a collection of iterated integrals.

We then ran numerous simulations to confirm the general efficacy of the algorithms under a variety of settings. We also applied the method to a publicly available meteorological dataset related to air particulate matter (PM) levels in China and an agricultural dataset concerning the effects of temperature and precipitation on crop yield. High PM levels generated via fuel combustion are understood to have significant adverse effects on mortality, and various Chinese cities are infamous for dense smog. Using weather data from the previous day, the algorithms were able to classify the PM levels during rush hour in Beijing, which would allow for citizens and authorities to take necessary precautions on particularly severe days. In the other example, better prediction of the range of likely crop yields can allow farmers to prepare for the likely worst-case and best-case scenarios.

The rest of the thesis is organized as follows. Chapter 2 introduces the conformal prediction method, and describe several of its properties which make it a desirable method for constructing prediction intervals. Chapter 3 derives algorithms for constructing prediction intervals for functional regression based on the conformal prediction method. Additionally, relevant theoretical results are discussed at the end of the chapter. Chapter 4 focuses on application of these algorithms in simulated data settings. Chapter 5 applies these algorithms to both aforementioned publicly

available data sets. We finish with a brief discussion of the results in this paper, and include some supplementary material at the end.

2.0 Conformal Prediction Background

2.1 Motivation for Conformal Prediction

To understand the appeal of the conformal prediction method, 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: $\hat{Y} \pm t_{n-p,\alpha/2}\sqrt{MSE(1 + X'_{n+1}(X'X)^{-1}X_{n+1})}$. 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 start to disappear once we begin generalizing the model. If we instead consider the model $Y = f(X) + \epsilon$, f unknown, $X \in \mathbb{R}^p$, $Y \in \mathbb{R}$, $\epsilon \sim N(0, \sigma^2)$, then we must first use some nonparametric method to estimate f, 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 f is difficult when p is larger than 2 or 3, and even more challenging when we consider one or more functional predictors.

Other methods of approaching this problem include nonparametric conditional distribution or density estimation of Y given X (Fan et al. (1996), Hall et al. (1999)), or nonparametric quantile regression (Koenker et al. (2005)) 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 methods in functional data setting (Cardot et al. (2005), Chen and Müller (2012)), but all of them need modeling assumptions on the true relationships and can be difficult to extend to multiple functional predictors.

Vovk et al. (2005) 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, ..., 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 $U_{(1)} \leq U_{(2)} \leq \cdots \leq U_{(n)}$ 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

$$\sigma_i(y) = \sigma(\{(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1} = y)\}, (X_i, Y_i)), \quad i = 1, \ldots, (n+1)$$

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 section, we let $\sigma_i(y) = |Y_i - \hat{m}(X_i)|$, where $\hat{m}$ is

some regression function trained on the augmented data $\{(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1} = y)\}$. 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 $P(\sigma_{n+1} \leq \sigma_{(\lceil (n+1)(1-\alpha) \rceil)}) \geq 1 - \alpha$. We omit the argument y for notation simplicity, but remember the conformity scores σ are functions of y. Thus, our prediction set $C(X_{n+1})$ of level $1 - \alpha$ consists of all values of y such that

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

2.2 Properties of Conformal Prediction

In addition to making minimal assumptions about the data, conformal inference has a few useful properties. Perhaps most important among them is the finite sample coverage guarantee. Lei et al. (2018) 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 Equation (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, Lei et al. (2018) proposes a split conformal prediction method which re-

duces 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. For example, if the conformity score is linear in y, then it is possible to solve for y and calculate the exact prediction set without searching over all values of y.

Even when an exact prediction set can be calculated directly, there is a possibility this set is not a single contiguous interval. In the regression setting, this happens when the observation X_{n+1} has high leverage relative to the rest of the data. While this is not always a problem, it is still of value to develop a set of conditions under which the prediction set will be an interval.

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 $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$ (we will call this type of coverage *marginal coverage*). 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, Lei and Wasserman (2014) 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:

$$\sup_x [P(Y_{n+1} \notin C(x) | X_{n+1} = x) - \alpha]_+ \xrightarrow{p} 0$$

In their two papers, Lei et al. explore a couple different approaches to achieving this goal. Lei and Wasserman (2014) describes an approach to locally estimate the conditional density of the residuals and use that to construct conditionally valid sets. Lei et al. (2018) later 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, an asymptotically conditionally valid prediction set can be obtained. Politis (2015) also describes a method of forming distribution free prediction sets with asymptotic conditional validity, although this too requires a consistent estimator of the mean.

Even if marginal coverage is sufficient for the analysis, the above results are important due to what they tell us about the efficiency of the prediction sets. In the 2017 paper, Lei shows that under the same regularity assumptions the prediction sets will have near optimal length and location. These results will be extended to cover the new developments in this paper, and practical performance will be explored.

3.0 Conformal Prediction in Multiple Functional Regression

In the multiple functional regression setting we observe i.i.d. pairs $(\boldsymbol{X}_1, Y_1), \ldots (\boldsymbol{X}_n, Y_n)$, where $\boldsymbol{X}_i = \{X_{ij}(t), t \in \mathcal{T} \subset \mathcal{R}\}_{j=1,\ldots,p}$ 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}. The basic model for this setting is $Y = f(X_1, \ldots, X_p) + \epsilon$, with unknown f and minimal assumptions on ϵ. The estimation of the nonparametric regression function $f(X_1, \ldots, X_p)$ is itself a challenging problem due to the curse of dimensionality. Jiang et al. (2018) considers the functional single index model $Y_i = g(\int_0^1 (\sum_{j=1}^p \beta_j X_{ji}(t))\alpha(t)dt) + \epsilon_i$, where β_j are predictor weights, and $\alpha(t)$ is the coefficient function. Meanwhile, Wong et al. (2019) extends the functional additive model originally developed by Müller and Yao (2008) to the multiple functional predictor case $Y_i = \sum_{k=1}^K g_k(\zeta_{ki}) + \epsilon_i$, where ζ_{ki} are standardized principal component scores from multivariate functional principal component analysis. That said, few (if any) of the multiple functional regression models in the literature provide a method by which one can calculate a prediction interval. As such, we will explore linear and nonlinear methods based on functional principal component analysis, as well as a method based on signature extraction. We then integrate these nonparametric regression approaches with the conformal method. We derive close form solutions for conformal prediction intervals based on carefully chosen conformity scores in the functional regression setting, where the residuals are obtained from linear functional principal component regression, nonparametric functional principal component regression, and nonparametric functional regression based on signature extraction. Given the finite sample coverage guarantee, all of the algorithms will produce intervals with the desired coverage regardless of the true

relationship. The algorithms based on nonparametric regressions retain efficiency in various regression settings.

3.1 Multiple Functional PCR-based Conformity Scores

3.1.1 Linear Multiple Functional Regression

Many common approaches to the univariate functional regression problem are based on functional principal components analysis (FPCA) (see Silverman et al. (1996), Hall et al. (2006)). Multivariate FPCA (mFPCA) is a natural extension to the multiple predictor setting. As with multivariate PCA, mFPCA aims to identify the strongest sources of variation between observations, except that the observations are now functions. Formally, consider a set of random functions $\boldsymbol{X} = \{X_j(t)\}_{j=1,\dots,p}$, $t \in \mathcal{T} \subset \mathcal{R}$ in Hilbert space $\mathbb{H}$, with each X_j square integrable, means $\mu_1(t) \dots \mu_p(t)$ and covariance function $\boldsymbol{G}(s,t) = \{G_{kl}(s,t)\}_{1 \le k,l \le p}$, $G_{kl}(s,t) = cov(X_k(s), X_l(t))$, $\boldsymbol{G}_k = (G_{k1}, \dots, G_{kp})^T$. 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 $\boldsymbol{\phi}_k = (\phi_{1k}, \dots, \phi_{pk})^T$, $j \ge 1$ and ordered eigenvalues $\lambda_1 \ge \lambda_2 \ge \dots$. Then for $k > 1$, the k-th functional principal component score is $\xi_k = \sum_{j=1}^{p} \int_{t \in \mathcal{T}} (X_j(t) - \mu_j(t))\phi_{jk}(t)dt$. Note that for each observation $\boldsymbol{X}_i$, this method produces a single principal component score ξ_{ki} associated with the k-th principal component $\boldsymbol{\phi}_k$, rather than the p scores obtained from performing FPCA

on each predictor individually. 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 eigenfunctions $\hat{\phi}_k(t)$, and principal component scores $\hat{\xi}_{ik}$ are obtained from numerical integration.

The first conformity score we wish to examine will be based on the natural extension of univariate linear principal components regression (see Hall et al. (2007)). This model assumes that the data pairs $(\boldsymbol{X}_i, Y_i)$ have the relationship

$$Y_i = \alpha + \int \boldsymbol{\beta}_j(t)\boldsymbol{X}_i(t) + \epsilon_i$$

As the principal component functions $\{\boldsymbol{\phi}_k(t)\}_{k\geq 1}$ form a complete orthonormal basis for $L_2(\mathcal{T})$, we may write $\boldsymbol{\beta}(t) = \sum_{k=1}^{\infty} b_k\boldsymbol{\phi}_k(t)$ for any set of functions $\boldsymbol{\beta}(t) \in L_2(\mathcal{T})$, and therefore:

$$Y_i = \alpha + \int \sum_{k=1}^{\infty} b_k\boldsymbol{\phi}_k(t) \sum_{k=1}^{\infty} \xi_{ik}\boldsymbol{\phi}_k(t)dt + \epsilon_i$$

Since $\boldsymbol{\phi}_k(t)$ are orthonormal, we can simplify this to:

$$Y_i = \alpha + \sum_{k=1}^{\infty} b_k\xi_{ik} + \epsilon_i$$

In practice, we may truncate the infinite sum at a finite order K, and regress Y on the first K principal component scores $\xi_k, 1 \leq k \leq K$.

If we define the matrix ξ such that the i,k-th entry is ξ_{ik} and the matrix $\boldsymbol{\xi} = [\mathbf{1}\ \xi]$, then the estimator for the mean function $m(x)$ is the OLS solution $\hat{m}(X_i) = \boldsymbol{\xi_i}(\boldsymbol{\xi}'\boldsymbol{\xi})^{-1}\boldsymbol{\xi}'Y$.

Using the notation above, our conformity score for the first algorithm will be $\sigma_i = |Y_i - \hat{m}(\boldsymbol{X}_i)|$. We would now like to derive a closed-form algorithm for producing

a conformal prediction set. This means that we would like to derive inequalities equivalent to Equation (1) which depend solely on $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n+1}, Y_1, \ldots, Y_n$. In the linear regression case, this is relatively easy to do. For example, Vovk et al. (2005) has used the closed-form solution for ridge regression, including the special case of OLS. We here include the derivation for completeness.

Notation: if T is a statistic based on the augmented data $(X_1, y_1), \ldots, (X_{n+1}, y_{n+1})$, let T^n be the same statistic based on $(X_1, y_1), \ldots, (X_n, y_n)$.

Applying the Sherman-Morrison formula to $\hat{\beta}^n = (X'X - x_{n+1}x'_{n+1})^{-1}X^{n\prime}Y^n$, we find $\hat{\beta} - \hat{\beta}^n = \frac{(X'X)^{-1}x_{n+1}r_{n+1}}{1-h_{n+1,n+1}}$. From there, we find $r_i = r_i^n - \frac{h_{i,n+1}r_{n+1}}{1-h_{n+1,n+1}}$, where h_{ij} is the i,j-th entry of the standard hat matrix $H = X(X'X)^{-1}X'$. Using the leave-one-out residual equality, $r_i^{(i)} = r_i/(1-h_{ii})$, we have

$$r_i = r_i^n - h_{i,n+1}r_{n+1}^n$$

As per Equation (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 $\sum_{i=1}^{n+1} 1[|r_i| \leq |r_{n+1}|] \leq \lceil (n+1)(1-\alpha) \rceil$. From above, $|r_{n+1}| > |r_i| \Leftrightarrow |(1-h_{n+1,n+1})r_{n+1}^n| > |r_i^n - h_{i,n+1}r_{n+1}^n|$ Solving this inequality leads to the following four inequalities:

Inequalities to determine upper bound:

$$\begin{cases} (1 - h_{n+1,n+1} + h_{i,n+1})r_{n+1}^n > r_i^n \\ (1 - h_{n+1,n+1} - h_{i,n+1})r_{n+1}^n > -r_i^n \end{cases} \tag{2}$$

Inequalities to determine lower bound:

$$\begin{cases} (1 - h_{n+1,n+1} + h_{i,n+1})r_{n+1}^n < r_i^n \\ (1 - h_{n+1,n+1} - h_{i,n+1})r_{n+1}^n < -r_i^n \end{cases} \tag{3}$$

This results in the following algorithm for constructing conformal prediction intervals based on functional PCR linear regression conformity score:

Algorithm 1

1. Calculate the K eigenfunctions $\hat{\boldsymbol{\phi}}_1, \dots \hat{\boldsymbol{\phi}}_K$ of $\boldsymbol{X}^T\boldsymbol{X}$ corresponding to the K largest eigenvalues where $1 \leq K \leq rank(X)$. Let $\boldsymbol{\phi} = (\hat{\boldsymbol{\phi}}_1, \dots, \hat{\boldsymbol{\phi}}_K)^T$.
2. Calculate $\xi = (\hat{\xi}_1, \dots, \hat{\xi}_K)^T = \boldsymbol{X}\boldsymbol{\phi}^T$, the scores of $\boldsymbol{X}_1, \dots, \boldsymbol{X}_{n+1}$ with respect to the basis $\hat{\boldsymbol{\phi}}_1, \dots, \hat{\boldsymbol{\phi}}_K$. Append a column of 1 to the matrix ξ to form $\boldsymbol{\xi}$. Any linear covariates may be appended to this matrix.
3. Calculate the residuals r_i^n for each point $(\boldsymbol{\xi}_i, Y_i)$ using the linear regression estimates from the data $(\boldsymbol{\xi}_1, Y_1), (\boldsymbol{\xi}_2, Y_2), \dots, (\boldsymbol{\xi}_n, Y_n)$.
4. Calculate the hat matrix $H = \boldsymbol{\xi}(\boldsymbol{\xi}'\boldsymbol{\xi})^{-1}\boldsymbol{\xi}'$.
5. Calculate $a_i = \frac{r_i^n}{1-h_{n+1,n+1}+h_{i,n+1}}$ and $a_{n+i} = \frac{-r_i^n}{1-h_{n+1,n+1}-h_{i,n+1}}$ for $i = 1, 2, \dots, n$.
6. Order these values from smallest to largest. Denote the ordered values $a_{(1)}, a_{(2)}, \dots, a_{(2n)}$.
7. Calculate $\hat{y}_{n+1}^n = \boldsymbol{\xi}_{n+1}\hat{\beta}$ using the OLS estimates from the data $(\boldsymbol{\xi}_1, Y_1), (\boldsymbol{\xi}_2, Y_2), \dots, (\boldsymbol{\xi}_n, Y_n)$.
8. Construct the $100(1-\alpha)\%$ prediction interval for $Y_{n+1} = (\hat{y}_{n+1}^n + a_{(\lfloor (n+1)\alpha \rfloor)}, \hat{y}_{n+1}^n + a_{(\lceil (2-\alpha)(n+1)-1 \rceil)})$

This algorithm makes the assumption that $1 - h_{n+1,n+1} \pm h_{i,n+1} > 0 \ \forall i$ (**Contiguity Condition 1**) to ensure the resulting set is a contiguous interval. If this condition is not satisfied, a conformal prediction set may still be constructed (see supplemental section).

3.1.2 Nonparametric Multiple Functional Regression

The main problems with the functional linear regression approach above occur when the model is misspecified. While marginal coverage is always obtained, if the true relationship between $\boldsymbol{X}_i$ and Y_i is not linear, then these prediction sets may be extremely large and have no guarantee of conditional coverage. The natural next step is to consider a nonparametric approach.

In the setting with a single functional predictor, work by Ferraty and Vieu (2006) and Geenens et al. (2011) focused on the Nadaraya-Watson type estimator $\hat{m}(x) = \frac{\sum_{k=1}^{n} K(((x-X_k))/h)Y_k}{\sum_{k=1}^{n} K(((x-X_k))/h)}$, where $((\cdot))$ is a semi-metric on $L^2(T)$, K is a univariate kernel, and h a scalar bandwidth. Many potential kernels K may work, including the standard Gaussian kernel $K(t) = \sqrt{2/\pi}e^{-t^2/2}$ for $t \in (0, \infty)$. This model may additionally be viewed as a local constant model which minimizes the weighted square error $\sum_{i=1}^{n}(Y_i - a)^2 K(((x - X_k))/h)$ over a. Crucially, Geenens showed that if $e_1, e_2, \ldots$ form an orthonormal basis of $L^2(T)$, then the function

$$((X))_p = \sqrt{\sum_{j=1}^{p} \langle X, e_j \rangle^2}$$

with p a fixed positive integer, is a semi-norm on $L^2(T)$. Given that the univariate principal component functions ϕ_j form an orthonormal basis, the function

$$((x - X_k))_p = \sqrt{\sum_{j=1}^{p} \langle x - X_k, \phi_j \rangle^2}$$

is a semi-norm. This function is simply the Euclidean distance between the first p FPC scores of x and X_k.

Baíllo and Grané (2009) 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

$\beta \in L^2$. We can express β and $x - X_i$ in terms of univariate principal component functions ϕ_j and coefficients b_k or FPC scores $\xi_{ik} - \xi_k$ respectively, and truncate the infinite sum after K terms. This results in us having to minimize the sum $\sum_{i=1}^{n}(Y_i - a - \sum_{k=1}^{K} b_k(\xi_{ik} - \xi_k))^2 K(((x - X_i))/h)$ over $a, b_1, \ldots, b_K$.

In the natural extension to the multiple functional regression setting, one would first apply mFPCA to your observations $\boldsymbol{X}_i$ and the new observation $\boldsymbol{x}$ to obtain the principal component scores $\hat{\xi}_i$ and for $\hat{\xi}_x$, respectively. Then construct the matrices $Z_x = [\mathbf{1}\ \xi_i - \xi_x]$ and $W_x = diag(K(||\xi_i - \xi_x||/h))$. Then $\hat{m}(x) = e_1(Z_x' W_x Z_x)^{-1} Z_x' W_x Y$, where e_1 is the $K+1$ dimensional vector with a 1 followed by K zeroes. This will be the estimated mean function used in the next conformity score $\sigma_i = |Y_i - \hat{m}(x)|$. In the functional linear regression based conformity score, we were able to express this σ in terms of residuals from the un-augmented model. That is not easy to do in this case; instead, we separate the formula for $\hat{m}(x)$ into terms depending on Y_{n+1} and terms not depending on Y_{n+1}. Then we can solve directly for Y_{n+1}. We can write

$$
\begin{aligned}
r_i &= Y_i - e_1(Z_{x_i}' W_{x_i} Z_{x_i})^{-1}(Z_{x_i}^{n'} W_{x_i}^{n} Y^n + Z_{x_i,n+1} W_{x_i,n+1} Y_{n+1}) \\
&= Y_i - e_1(Z_{x_i}' W_{x_i} Z_{x_i})^{-1} Z_{x_i}^{n'} W_{x_i}^{n} Y^n - e_1(Z_{x_i}' W_{x_i} Z_{x_i})^{-1} Z_{x_i,n+1} W_{x_i,n+1} Y_{n+1} \\
&= Y_i - A_i - B_i Y_{n+1}
\end{aligned}
$$

where neither A_i nor B_i depend on Y_{n+1}.

Again, as per Equation (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 conformal prediction intervals based on the local linear regression conformity score:

Algorithm 2

1. Calculate the K eigenfunctions $\hat{\boldsymbol{\phi}}_1, \ldots, \hat{\boldsymbol{\phi}}_K$ of $\boldsymbol{X}^T\boldsymbol{X}$ corresponding to the K largest eigenvalues where $1 \leq K \leq rank(X)$. Let $\boldsymbol{\phi} = (\hat{\boldsymbol{\phi}}_1, \ldots, \hat{\boldsymbol{\phi}}_K)^T$.
2. Calculate $\xi = (\hat{\xi}_1, \ldots, \hat{\xi}_K)^T = X\phi^T$, the scores of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n+1}$ with respect to the basis $\hat{\boldsymbol{\phi}}_1, \ldots, \hat{\boldsymbol{\phi}}_K$. Any covariates may be appended to this matrix.
3. Using the above notation, calculate A_i and B_i for all i.
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 $Y_{n+1} = (L_{(\lfloor (n+1)\alpha \rfloor)}, U_{(\lceil (1-\alpha)(n+1) \rceil)})$

Similarly to Algorithm 1, this algorithm makes the assumption that $1 - B_{n+1} \pm B_i > 0$, $\forall i$ (**Contiguity Condition 2**) in order to guarantee contiguous prediction intervals. Again, check supplemental section 7.1 for instructions on modifying this algorithm if this condition is not satisfied.

3.2 Partial Linear Conformity Score based on Signature

The major issue facing the previous algorithm is the so-called "curse of dimensionality". This term refers to the poor performance of nonparametric methods caused by the sparsity of data in more than 2 or 3 dimensions. As a result, applying Algorithm 2 to more principal component scores can often result in large prediction sets with inconsistent conditional coverage that often are not even contiguous. This practical restriction to only a few principal components limits the ability of FPCA to capture the true nature of the relationship, since mFPCA tends to require slightly more principal components to capture most of the information contained in the multiple functional predictors. We want to find another regression model to describe nonlinear relationships without encountering this same issue.

While functional PCA is a natural choice, other methods of extracting feature sets exist. The signature method, one such approach first described by Chen (1958), has recently been applied to rough path theory in areas of machine learning by Hambly and Lyons (2010). However, very little work has been done to apply this method to functional data analysis despite the obvious similarities.

In this area of research, a d-dimensional path is defined as a continuous mapping from $[a, b]$ to $\mathbb{R}^d$. A function $X(t), t \in [0, 1]$ can be considered a specific case of a 2-dimensional path $X_t = \{X_t^1, X_t^2\} = \{t, X(t)\}$. For example, the function $X(t) = 2t^2$, $t \in [0, 1]$ is equivalent to the 2-dimensional path $X_t = \{X_t^1, X_t^2\} = \{t, 2t^2\}$, $t \in [0, 1]$. Note that a 1-dimensional path alone is insufficient to capture the information contained within $X(t)$, as it will be clear later that the signature of the 1-path dimensional path only depends on the image of the mapping. Similarly, the p functions $X_1(t), X_2(t), \ldots X_p(t)$ 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 as described by Chevyrev and Kormilitzin (2016). The path integral of a 1-dimensional path Y_t against another 1-dimensional path X_t is defined as the integral

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

Now we can define the signature of a d-dimensional path. For all $i \in \{1, 2, \ldots, d\}$:

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

For $i, j \in \{1, 2, \ldots, d\}$, the double-iterated integral is defined as:

$$S(X_t)^{i,j}_{a,t} = \int_{a<s<t} S(X_t)^i_{a,s} dX^j_s = \int_{a<r<s<t} dX^i_r dX^j_s$$

In general, for $i_1, i_2, \ldots, i_k \in \{1, 2, \ldots, d\}$, 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 sequence

$$S(X_t)_{a,b} = (1, S(X_t)^1_{a,b}, \ldots, S(X_t)^d_{a,b}, S(X_t)^{1,1}_{a,b}, S(X_t)^{1,2}_{a,b}, \ldots)$$

where all d^k k-fold iterated integrals with unique superscripts are included for $k = 1, 2, \ldots$. Chen (1958) provides a derivation and analysis of iterated integrals, while Hambly and Lyons (2010) shows how this definition of the signature naturally arises from the study of controlled differential equations.

Furthermore, define the k-th order signature $[S(X_t)]^k_{a,b}$ to be the sequence $\{S(X_t)^{1,\ldots,1}_{a,b}, \ldots, S(X_t)^{d,\ldots,d}_{a,b}\}$ of all d^k k-fold iterated integrals with unique superscripts. Then the signature $S(X_t)_{a,b} = (1, [S(X_t)]^1_{a,b}, [S(X_t)]^2_{a,b}, \ldots)$.

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 displacement $X^i_b - X^i_a$ 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 $A = (S(X_t)^{1,2}_{a,b} - S(X_t)^{2,1}_{a,b})/2$, where A is the *Lévy Area* of the path. The *Lévy Area* of a path is the signed area enclosed by the path and the chord connecting the endpoints. In the example in Figure 1, the *Lévy Area* is calculated $A = A_+ - A_-$

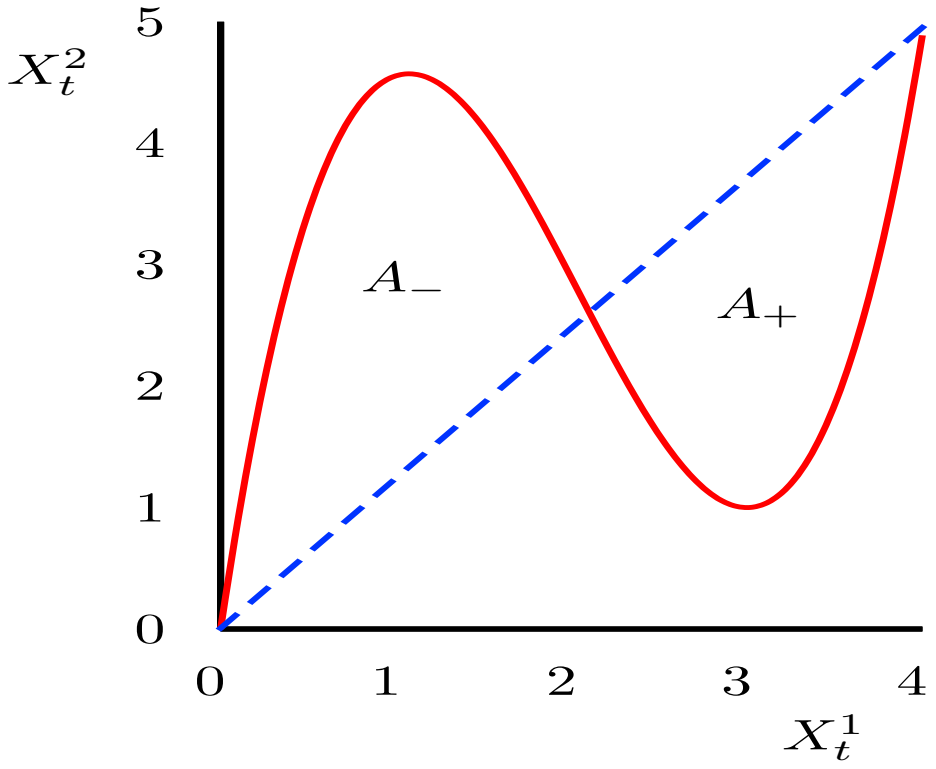

Figure 1: Geometric intuition of second order signature

3.2.1 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 $S(X_t)_{a,b}^{i_1,\ldots,i_k}$ as $S^{i_1,\ldots,i_k}$ when there is no confusion. Also, we use $[S(X_t)]_{a,b}^k$ 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 Chen (1958) for all continuously differentiable functions and later expanded to paths of bounded variance by Hambly and Lyons (2010).

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 $X(t)$ 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 the domain of each predictor is the same for all observations, then vertical location is the only information lost.

The *shuffle property* is another reason why the signature method is a promising approach for functional nonparametric regression. First proved by Ree (1958), the shuffle property states that any product of two terms $S(X_t)_{a,b}^{i_1,i_2,\ldots,i_k}$ and $S(X_t)_{a,b}^{j_1,j_2,\ldots,j_n}$ can be expressed as the sum of terms of $S(X_t)_{a,b}$ determined entirely by 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 $J = (j_1, j_2, \ldots, j_n)$,

$$S(X_t)_{a,b}^I S(X_t)_{a,b}^J = \sum_{K \in I \sqcup J} S(X_t)_{a,b}^K$$

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:

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

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. In the context of (non)linear controlled differential equations (see section 3.3 for more details), Lyons (2014) shows that the truncated signature uniformly converges to the signature, but does not generalize this result to other settings. Nevertheless, we 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. If the response indeed only depends on the a finite order of signature terms, the produced prediction interval will have conditional coverage and efficiency in addition to the finite sample marginal coverage guarantee.

In functional data regression settings in which observations are recorded on the same domain, we can further make use of the *shuffle property* to reduce the number of terms needed for regression. Since the i-th sample predictor $\boldsymbol{X}_i(t)$ is represented

as a path $X_{it} = \{t, X_{1i}(t), \ldots, X_{pi}(t)\}$, the signature terms $S_i^1, S_i^{11}, S_i^{111}, \ldots$, which only depend on the first 1-dimensional path will take the same value for all sample functions. Based on the *shuffle property*, we can express all signature terms of order k as linear combinations of signature terms of order $k+1$. 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 have the following (approximate) equalities

$$
\begin{aligned}
& f(X_1(t), X_2(t), \ldots, X_p(t)) \\
=& \tilde{f}(S(X_t)), \text{ by uniqueness of the signature} \\
=& L(S(X_t)), \text{ by the shuffle property} \\
\approx& L([S(X_t)]^1, [S(X_t)]^2, \ldots, [S(X_t)]^k), \text{ by the finite order approximation} \\
=& L([S(X_t)]^k \setminus S(X_t)^{1,1,\ldots,1}), \text{ by the shuffle property in the functional data context}
\end{aligned}
$$

Of course, the first equality above ignores the issue of vertical translation. Any dependence the response has on vertical location of the predictors will not be captured by this feature set. As such, additional variables representing the vertical location of each predictors should be added to the signature expansion. However, there is no reason to assume that the response linearly depends on these terms. Therefore, it is natural to consider a semiparametric model.

One possibility for the semiparametric model would be a multiple index model. In this model, we have $Y = f([S]^k \beta, Z_1, \ldots, Z_p) + \epsilon$, where f is an unknown function, $[S_i]^k$ is the k-th order signature expansion for X_i, and Z_j, $1 \leq j \leq p$, are scalar

predictors which capture the vertical location of each predictor function $X_j(t)$. For example, Z_j can be the average value of $X_j(t)$ over all t. 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 Lei et al. (2018) to obtain an approximate conformal prediction set based on this model.

We propose using the signature-based multiple partial linear model

$$Y_i = [S_i]^k \beta + \sum_{j=1}^{p} g_j(Z_{ji}) + \epsilon_i, \tag{6}$$

where g_j, $1 \leq j \leq p$, 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 a balance between model flexibility and simplicity.

To derive the conformal prediction interval, we need to fit the model Equation (6) 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 $[S]^k = ([S]_1^k, \ldots, [S]_{n+1}^k)^T$. Extending the work of Robinson (1988), 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 Opsomer (2000). 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 Opsomer (2000), the equations for the Y-residuals and S-residuals are:

$\hat{r}_{yi} = Y_i - \hat{g}_y(\boldsymbol{Z}_i) = Y_i - \boldsymbol{W}_{M_i,} Y = Y_i - \sum_{j=1}^{n} \boldsymbol{W}_{M_i,j} Y_j - \boldsymbol{W}_{M_i,n+1} Y_{n+1}$

$\hat{r}_{si} = [S_i]^k - \hat{g}_s(\boldsymbol{Z}_i) = [S_i]^k - \boldsymbol{W}_{M_i,} [S]^k,$

As defined by Opsomer (2000), $\boldsymbol{W}_M$ 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 Equation (6) is

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

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

$$\begin{aligned}
r_i &= Y_i - \hat{Y}_i \\
&= Y_i - \hat{r}_{si}\hat{\beta} - \hat{g}_y(\boldsymbol{Z}_i) \\
&= Y_i - \hat{r}_{si}(c_1 + c_2 Y_{n+1}) - (\sum_{j=1}^{n} \boldsymbol{W}_{M_{ij}} Y_j + \boldsymbol{W}_{M_{i,n+1}} Y_{n+1}) \\
&= Y_i - a_i - b_i Y_{n+1} \qquad (7)
\end{aligned}$$

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, $|r_{n+1}| > |r_i| \Leftrightarrow |Y_{n+1} - a_{n+1} - b_{n+1}Y_{n+1}| > |Y_i - a_i - b_iY_{n+1}|$. Solving this inequality leads to the following four inequalities:

Inequalities to determine upper bound:

$$\begin{cases} (1 - b_{n+1} + b_i)Y_{n+1} < Y_i - a_i + a_{n+1} \\ (1 - b_{n+1} - b_i)Y_{n+1} < -Y_i + a_i + a_{n+1} \end{cases} \tag{8}$$

Inequalities to determine lower bound:

$$\begin{cases} (1 - b_{n+1} + b_i)Y_{n+1} > Y_i - a_i + a_{n+1} \\ (1 - b_{n+1} - b_i)Y_{n+1} > -Y_i + a_i + a_{n+1} \end{cases} \tag{9}$$

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 3)

1. Calculate the k-th order signature $[S_i]^k$ of each $\boldsymbol{X}_i$, $i = 1, \ldots, n+1$. Construct the matrix $[\boldsymbol{S}]^k = ([S_1]^k, \ldots, [S_{n+1}]^k)^T$. Any linear covariates may be appended to this matrix.
2. Calculate Z_{ij}, for each $X_{ij}(t)$, $i = 1, \ldots, n+1, j = 1, \ldots, p$. Construct the matrix $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{n+1})^T$. Any additive nonlinear covariates may be appended to this matrix.
3. Calculate $\hat{r}_{yi}$, $\hat{r}_{Si}$, $\hat{\beta}$, and a_i, b_i as described in Equation (7).
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 $Y_{n+1} = (L_{(\lfloor \alpha(n+1) \rfloor)}, U_{(\lceil (1-\alpha)(n+1) \rceil)})$

To ensure the resulting set is a contiguous interval, this algorithm assumes a **Contiguity Condition 3**

$$1 - b_{n+1} \pm b_i > 0, \forall i. \tag{10}$$

It is obvious from Equation (7) that b_i is the degree to which the value of Y_{n+1} affects r_i and therefore $\hat{y}_i$. Thus, this condition (along with the previous two) is a restriction on the *perturb one sensitivity*, similar to Assumption 3 in Section 3.3. The restriction roughly translates into a requirement that $\boldsymbol{X}_{n+1}$ not be "too far" away from the rest of the $\boldsymbol{X}_i$.

3.3 Theoretical Results

In this section, we first list Properties 1 - 4 for conformal prediction intervals. These properties have been proved in previous papers under a list of assumptions (Lei et al., 2018). and we include that here for completeness. We then discuss some of assumptions in the functional data context, and with respect to the specific conformity scores we proposed. Then we discuss the contiguity property of the constructed prediction intervals.

Property 1 (Finite Sample Coverage): If (X_i, Y_i), $i = 1, 2, \ldots, n+1$ are i.i.d., then $P(Y_{n+1} \in C(X_{n+1})) \geq 1 - \alpha$, where $C(X_{n+1})$ is the conformal prediction interval produced by algorithm 1, 2, or 3.

Property 2 (Finite Sample Accuracy): Under the same conditions as Theorem 1, with the additional assumption that for all $y \in \mathbb{R}$, the fitted absolute residuals $R_{y,i}$ have a continuous joint distribution, then $P(Y_{n+1} \in C(X_{n+1})) \leq 1 - \alpha + \frac{1}{n+1}$.

The following four assumptions are necessary to establish properties 3 and 4 regarding conditional coverage and efficiency:

Assumption 1: (X_i, Y_i) are sampled i.i.d. from a distribution P on $\mathbb{R}^{\mathbb{R}} \times \mathbb{R}$, with mean function $m(x) = \mathbb{E}(Y|X = x)$, $x \in \mathbb{R}^{\mathbb{R}}$.

Assumption 2: The noise $\epsilon = Y - m(X)$ is independent of X, and the density of ϵ is symmetric about 0 and nonincreasing on $[0, \infty)$.

As noted in the source paper, this assumption is weak, and can even be dropped, but is included for convenience.

Assumption 3: Let $\hat{m}_n$ be the estimated regression function from the original n observations, $\hat{m}_{n,(X,y)}$ be the estimated regression function from the augmented data with $(n+1)^{st}$ observation (X, y), and $\mathcal{Y}$ be compact interval of fixed length. Then for large enough n,

$$\mathbb{P}\left(\sup_{y \in \mathcal{Y}} ||\hat{m}_n - \hat{m}_{n,(X,y)}||_\infty \geq \eta_n\right) \leq \rho_n$$

for some sequences satisfying $\eta_n = o(1)$, $\rho_n = o(1)$ as $n \to \infty$.

This assumption requires the fitted regression function does not change too much as we vary the y value of the $(n+1)^{st}$ observation.

Assumption 4: For n large enough,

$$\mathbb{P}\left(\mathbb{E}_X\left[(\hat{m}_n(X) - m(X))^2 | \hat{m}_n\right] \geq \eta_n\right) \leq \rho_n$$

for some sequences satisfying $\eta_n = o(1)$, $\rho_n = o(1)$ as $n \to \infty$.

This is a condition on the consistency of the estimator $\hat{m}$.

Property 3 (Conditional Coverage of the Conformal Prediction Set): Under assumptions 1-4, then the sequence C_n of (possibly) random prediction sets have asymptotic conditional coverage at the level $(1-\alpha)$. Specifically, there exists a sequence of (possibly) random sets $\Lambda_n \subseteq \mathbb{R}^{\mathbb{R}}$ such that $\mathbb{P}(X \in \Lambda_n | \Lambda_n) = 1 - o_{\mathbb{P}}(1)$ and

$$\inf_{x \in \Lambda_n} |\mathbb{P}(Y \in C_n(x) | X = x) - (1-\alpha)| = o_{\mathbb{P}}(1).$$

Property 4 (Efficiency of the Conformal Prediction Set): Under the same assumptions as in Theorem 4, the Lebesgue measure of the symmetric difference between $C_{n,conf}(X)$ and C_s^* is $o_{\mathbb{P}}(1)$, where C_s^* is the "super oracle" interval $[m(x) - q_\alpha, m(x) + q_\alpha]$, and q_α is the α upper quantile of $\mathcal{L}(|\epsilon|)$. This super oracle interval has the shortest length among all bands with conditional coverage, and the shortest average length among all bands with marginal coverage.

Assumption 4 is an important assumption to verify. While conformal sets have guaranteed marginal coverage regardless, we would like to claim that the conditional coverage and length of the prediction interval approaches optimality in some respect.

Of initial interest is the FPCA based local linear conformity score. Lei notes that the standard multivariate local polynomial regression satisfies this assumption, but that doesn't have an obvious extension to the functional case. Similarly, Ferraty and Vieu (2006) was able to prove pointwise almost complete convergence for his Nadaraya-Watson type estimator, but did not extend this to a functional local linear model.

Baíllo and Grané (2009) was able to determine the convergence rate of the functional local linear regression when the basis $\{\phi_j\}_{j \geq 1}$ is the trigonometric basis under the following conditions:

- There exist two real constants $0 < c_I < C_I < \infty$ such that the kernel K satisfies

$c_I 1_{[0,1]} \leq K \leq C_I 1_{[0,1]}$

- For any $\epsilon > 0$, the small ball probability $\varphi_x(\epsilon) \coloneqq P(||X - x|| < \epsilon) = O(\epsilon^\tau)$ for some $\tau > 0$ is strictly positive.
- With probability one, any trajectory $X(\cdot, \omega)$ of X has derivative of ν-th order which is uniformly bounded on $[0, 1]$ by a constant independent of ω.
- The element x has derivative of ν-th order which is uniformly bounded on $[0, 1]$.
- The bandwidth $h = O(n^{-1/(4+\tau)})$, and the number of principal component scores $K = O(h^{-2/\nu})$

If these conditions are satisfied, then $E[(\hat{m}(x) - m(x))^2|X] = O_P(n^{-4/(4+\tau)})$. In particular, the requirement that the basis $\{\phi_j\}_{j\geq 1}$ be the trigonometric basis was used to show that $\max_{i=1,\ldots,n} |m'_x(X_i - x) - \sum_{j=1}^K m'_{x,j} c_{ij}| = O(K^{-\nu})$, a bound on the basis approximation of the derivative of $m(X)$. To extend this result for a generic FPCA basis, a similar result will need to be shown.

As mentioned earlier, there are not many results concerning the consistency of the signature as a feature set in the functional regression setting. The few results which do appear are within the context of Controlled Differential Equations. Expanding on earlier work in Levin et al. (2013), focusing only on Linear Controlled Differential Equations, Boedihardjo et al. (2015) studied Controlled Differential Equations of the form

$$dY_t = f(Y_t)dX_t$$

$$Y_0 = y_0$$

In other words, the response Y is also a function of t and the change in the value of Y at time t is equal to the product of some function of Y_t and the change in

the predictor X at time t. In this context, Boedihardjo showed that there exists a constant C_p depending only on p such that

$$\left| Y_t - Y_s - \sum_{k=1}^{\lfloor \gamma \rfloor} f^{\circ k}(Y_s) X_{s,t}^k \right| \leq \frac{1}{\left(\frac{\lfloor \gamma \rfloor}{p}\right)!} \beta^{\lfloor \gamma \rfloor} C_p M_{p,\gamma} ||f||_{\circ\gamma} ||X||_{p-var,[s,t]}^{\gamma}$$

where $X = (1, X^1, \ldots, X^{\lfloor p \rfloor})$ is a p-weak geometric rough path for $p \geq 1$, f is a $Lip(\gamma - 1)$ vector field with $\gamma > p$, and $M_{p,\gamma}$, $||f||_{\circ\gamma}$, and β are complex functions of p and γ (see reference for details).

In the functional regression setting, this result implies that the difference $Y_1 - \langle [a_i]^{1:K}, [S]^{1:K} \rangle$ for some coefficients $[a_i]^{1:K}$ is approximately proportional to $\frac{1}{K!}$ for large K. Therefore, if we allow K to grow slowly with n, the resulting approximation should converge.

Unfortunately, this model is not always appropriate, including situations in which Y cannot meaningfully be thought of as a function of t. In these situations, there are no results regarding the consistency of the truncated signature.

Property 5 (Contiguity of Conformal Prediction Sets): The prediction sets $C(X_{n+1})$ generated by algorithms 1, 2, and 3 will be contiguous if Contiguity Conditions 1, 2, or 3 are satisfied respectively.

Proof: For convenience, we will prove this result only for algorithm 1. The proofs for algorithms 2 and 3 follow analogously. Assume Contiguity Condition 1 holds. Thus the n regions from step 5 of algorithm 1.1b (see Section 7.1) will all take the form $[\min(a_i, b_i), \max(a_i, b_i)]$. Every union of $\lceil (n+1)(1-\alpha) \rceil$ of these sets will take the form $[\min_{i \in p}(\min(a_i, b_i)), \max_{i \in p}(\max(a_i, b_i))]$, where p is any selection of $\lceil (n+1)(1-\alpha) \rceil$ unique elements from $\{1, 2, \ldots, n\}$ (Note that in this case, every region will necessarily contain 0, meaning that no two sets may be disjoint). Finally, the intersection of these sets will take the form

$[\max_{p\in P}(\min_{i\in p}(\min(a_i, b_i))), \min_{p\in P}(\max_{i\in p}(\max(a_i, b_i)))]$, where P is the set of all unique p. This set is a contiguous interval.

Let us consider the actual meaning of these conditions. Contiguity Condition 1 states that $1 - h_{n+1,n+1} \pm h_{i,n+1} > 0 \ \forall i$. In the linear regression model, $\hat{Y} = HY$. Thus, $h_{n+1,n+1}$ is the amount that $\hat{Y}_{n+1}$ will change given a one unit increase in the value of Y_{n+1}. In other words, $h_{n+1,n+1}$ represents the sensitivity of $\hat{Y}_{n+1}$ to a perturbation of Y_{n+1}. Similarly, $h_{i,n+1}$ represents the sensitivity of $\hat{Y}_i$ to a perturbation of Y_{n+1}. This condition therefore plays a similar role to assumption 3 by bounding the change in fitted regression function after a change in Y_{n+1}.

Now consider Contiguity Conditions 2 and 3. Through our notation, we have made it obvious that the B_i in condition 2 and the b_i in condition 3 fully capture the degree to which the value of Y_{n+1} affects r_i and therefore $\hat{y}_i$. Thus, these conditions are also limits on the perturb one sensitivity.

To understand what these conditions entail in practice, start with the simple linear model: $Y = \beta_0 + \beta_1 X + \epsilon$, $X, Y \in \mathbb{R}$. To further simplify the math, assume X_i are centered such that $\sum_{i=1}^{n} X_i = 0$. In this case, Contiguity Condition 1 simplifies to the inequality for all $i = 1, 2, \ldots, n$:

$$max(-\frac{n+1}{nX_i}\sum_{i=1}^{n} X_i^2, \frac{n-1}{nX_i}\sum_{i=1}^{n} X_i^2) < X_{n+1} < min(-\frac{n+1}{nX_i}\sum_{i=1}^{n} X_i^2, \frac{n-1}{nX_i}\sum_{i=1}^{n} X_i^2)$$

If we make the approximations that $\sum_{i=1}^{n} X_i^2 \approx n\text{Var}(X)$, and $n-1 \approx n \approx n+1$, then this makes the the system of inequalities approximately equivalent to:

$$-\frac{n\text{Var}(X)}{\max(|X_i|)} < X_{n+1} < \frac{n\text{Var}(X)}{\max(|X_i|)}$$

This condition boils down to a weak restriction on the values X_{n+1} may take. In the case where X are i.i.d. $N(0,1)$, the expected value of $\max(|X_i|)$ is $\mathcal{O}(\sqrt{\log n})$, so the bounds on X_{n+1} for the resulting prediction set to be a contiguous interval are on the order of $\pm\frac{n}{\sqrt{\log n}}$.

Extending this to the case where $X_i \in \mathbb{R}^d$ is relatively straightforward conceptually, even if it is a bit more complicated mathematically. The restriction $1 - h_{n+1,n+1} \pm h_{i,n+1} > 0 \ \forall i$ translates into a requirement that X_{n+1} not be "too far" away from the rest of the X_i, where the precise meaning depends on the spread of X_i and n.

In local linear regression, this condition requires X_{n+1} be sufficiently close to other points relative to the bandwidth. If this condition cannot be satisfied, lowering the dimension and increasing the bandwidth will both help meet this condition. In the partial linear regression, this condition requires that the $(n+1)$-th residual of the regression of X on Z be near the other residuals.

Of course, we ultimately wish to use these algorithms in the functional regression problem. Therefore, it must be the feature set of X that satisfies the conditions, as opposed to X itself. It is difficult to translate these restrictions into restrictions on the actual functions X_{n+1}.

4.0 Simulations

4.1 Univariate Simulations

To test the performance and robustness of the three algorithms, we performed multiple simulations under a variety of conditions. There were a couple main goals with these simulations. 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.

The final goal was to compare the performance of the proposed algorithms 1-3 under different settings. To accomplish this, we performed simulations using different bases to generate X, using linear and quadratic functions for the relationship between Y and X, using different sample sizes, and using different error distributions. 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". We also compared the results to intervals generated via the sample splitting method shown by Lei et al. (2015) to have valid finite sample coverage. Specifically, we split the original sample into equally-sized training and testing sets, estimated the local linear mFPC model on the training set, then used the fitted model to obtain residuals for each observation in the testing set. These residuals were then used to construct the conformal prediction set for the new observation.

We generated X according to $X_i(t) = \sum_{k=1}^{K} \xi_{ik}\phi_k(t)$. In the first simulation setting, $K = 22$, and the basis functions $\{10, 10t, 10\sin(2\pi it), 10\cos(2\pi it)\}$ (orthogonalized by Gram-Schmidt algorithm). The scores ξ_{ik} are i.i.d $N(0, k^{-1.2})$. In the second setting, we used the cubic B-Spline basis with 18 knots. Scores were generated from the standard normal distribution, and the resulting functions were scaled by a factor of 4 to produce a similar range of values as the modified Trigonometric basis. Figure 2 displays example X functions from each of the bases.

Our response Y was calculated as $Y_i = \int X_i(t)\beta(t)dt + \epsilon_i,\ i = 1, \ldots, n$, in the linear settings and $Y_i = \frac{1}{100}(\int X_i(t)\beta(t)dt)^2 + \epsilon_i$ in the quadratic settings, where $\epsilon_i \sim N(0, 0.5)$ or $\epsilon_i \sim t_3/\sqrt{2}$ in the heavy-tail setting. The function β, as shown in Figure 3, was generated as $\sum_{k=1}^{4} b_k\phi_k(t)$, where $\phi_k(t)$ are the four modified Trigonometric basis mentioned above, and $b_k = (-1)^{\lfloor k/2 \rfloor} i^{-1},\ k = 1, 2, 3, 4$.

We note that when $X(t)$ is generated from the modified Trigonometric basis, the relationship between Y and X is completely determined through the first four functional principal component components of X, and indeed mostly captured by the first two or three components. So this is a low dimensional case for function principal component based nonparametric regression. When $X(t)$ are generated from the cubic B-Spline basis, the relationship between Y and X can not be well captured if we only represent X using the first few functional principal components, and the

nonparametric regression based on FPCA may not work.

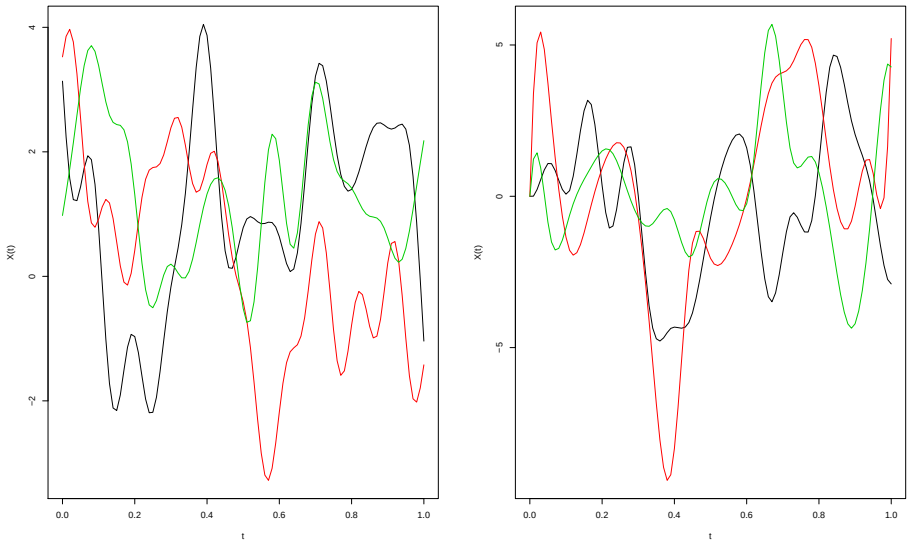

Figure 2: Examples of $X(t)$ functions generated from modified trigonometric basis (left) and B-spline basis (right)

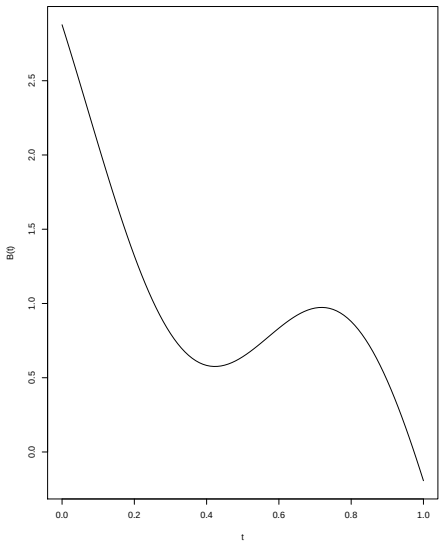

Figure 3: $\beta(t)$

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.

The results of the simulations are displayed in the following tables. Tables 1-4 display the mean coverage of the intervals produced by the specified algorithm in the given setting. Tables 5-8 display the median length of these interval. The naive "oracle" predictions are constructed as follows. In the linear relationship settings, the residuals are derived from the linear FPCA regression model, while in the quadratic relationship settings, they are derived from the local linear FPCA regression model. In the settings with normal error, the critical values are calculated using the residual standard deviation and normal quantile, while the empirical percentiles are used in the t_3 error settings. While these oracle intervals have asymptotic coverage at the

95% level, they may tend to undercover in practice. However, they still perform well as a baseline for comparison.

Table 1: Linear Relationship with Modified Trig Basis: we report the mean coverage from 1000 simulations for linear mFPCA conformity score, local linear FPCA conformity score, partial linear signature conformity score, naive "oracle", and split sample conformity score.

Error	N	Lin-FPCA	LL-FPCA	PL-Sig	"Oracle"	SS Interval
Normal	200	0.950	0.950	0.951	0.944	0.947
	500	0.951	0.950	0.951	0.947	0.949
t_3	200	0.950	0.951	0.950	0.952	0.947
	500	0.950	0.951	0.951	0.951	0.953

Table 2: Linear Relationship with B-Spline Basis: we report the mean coverage from 1000 simulations for linear mFPCA conformity score, local linear FPCA conformity score, partial linear signature conformity score, naive "oracle", and split sample conformity score.

Error	N	Lin-FPCA	LL-FPCA	PL-Sig	"Oracle"	SS Interval
Normal	200	0.950	0.955	0.949	0.940	0.950
	500	0.949	0.949	0.950	0.947	0.956
t_3	200	0.949	0.949	0.951	0.953	0.950
	500	0.951	0.949	0.950	0.950	0.951

Table 3: Quadratic Relationship with Modified Trig Basis: we report the mean coverage from 1000 simulations for linear mFPCA conformity score, local linear FPCA conformity score, partial linear signature conformity score, naive "oracle", and split sample conformity score.

Error	N	Lin-FPCA	LL-FPCA	PL-Sig	"Oracle"	SS Interval
Normal	200	0.953	0.951	0.954	0.922	0.948
	500	0.956	0.950	0.956	0.930	0.944
t_3	200	0.947	0.951	0.950	0.937	0.948
	500	0.947	0.948	0.949	0.946	0.948

Table 4: Quadratic Relationship with B-Spline Basis: we report the mean coverage from 1000 simulations for linear mFPCA conformity score, local linear FPCA conformity score, partial linear signature conformity score, naive "oracle", and split sample conformity score.

Error	N	Lin-FPCA	LL-FPCA	PL-Sig	"Oracle"	SS Interval
Normal	200	0.955	0.957	0.952	0.943	0.952
	500	0.957	0.947	0.951	0.944	0.951
t_3	200	0.951	0.950	0.956	0.958	0.949
	500	0.953	0.955	0.950	0.946	0.955

From tables 1-4, we find that the mean coverage of all 3 algorithms is approximately 0.95 in every setting, even when the model on which the algorithm is based

is not correct (such as the linear model in the quadratic relationship). This is one of the main benefits to using a conformal prediction method. The mean coverage of the naive "oracle" interval is a bit worse in smaller sample sizes as they rely on asymptotic properties.

Table 5: Linear Relationship with Modified Trig Basis: we report the median length of the constructed intervals from 1000 simulations for linear mFPCA conformity score, local linear FPCA conformity score, partial linear signature conformity score, naive "oracle", and split sample conformity score.

Error	N	Lin-FPCA	LL-FPCA	PL-Sig	"Oracle"	SS Interval
Normal	200	2.8	2.9	3.1	2.8	5.1
	500	2.9	2.9	2.9	2.8	5.1
t_3	200	4.8	4.8	5.2	4.9	6.3
	500	4.7	4.8	5.0	4.6	6.1

Table 6: Linear Relationship with B-Spline Basis: we report the median length of the constructed intervals from 1000 simulations for linear mFPCA conformity score, local linear FPCA conformity score, partial linear signature conformity score, naive "oracle", and split sample conformity score.

Error	N	Lin-FPCA	LL-FPCA	PL-Sig	"Oracle"	SS Interval
Normal	200	3.6	4.3	3.1	3.3	4.8
	500	3.1	3.9	3.0	3.1	4.8
t_3	200	5.6	5.3	5.2	5.6	5.9
	500	4.8	5.5	4.8	4.7	5.9

Table 7: Quadratic Relationship with Modified Trig Basis: we report the median length of the constructed intervals from 1000 simulations for linear mFPCA conformity score, local linear FPCA conformity score, partial linear signature conformity score, naive "oracle", and split sample conformity score.

Error	N	Lin-FPCA	LL-FPCA	PL-Sig	"Oracle"	SS Interval
Normal	200	6.9	3.3	4.5	3.2	7.0
	500	6.8	3.4	4.4	3.2	7.0
t_3	200	8.1	5.7	6.3	4.9	8.2
	500	8.0	5.0	6.0	4.8	8.2

Table 8: Quadratic Relationship with B-Spline Basis: we report the median length of the constructed intervals from 1000 simulations for linear mFPCA conformity score, local linear FPCA conformity score, partial linear signature conformity score, naive "oracle", and split sample conformity score.

Error	N	Lin-FPCA	LL-FPCA	PL-Sig	"Oracle"	SS Interval
Normal	200	6.0	5.6	4.2	5.8	6.1
	500	5.9	4.5	3.6	4.5	6.1
t_3	200	7.4	7.0	5.7	7.3	7.5
	500	7.2	7.2	5.1	5.8	7.5

From tables 5-8, we find that Algorithm 1, based on a linear conformity score, produces nice results when the true relationship is linear, but is inefficient (larger intervals than other algorithms) in cases the true relationship is non-linear. This result is as expected.

The other two algorithms based on nonparametric functional regressions produce intervals with reasonable length. A closer look reveals that Algorithm 2, based a on local linear FPCA conformity score, performs worse in the cases that the predictor functions X are generated from B-Spline basis, especially in the quadratic case. This is also as expected, since in these cases the regression relationship need to be characterized using a relatively large number of principal components, which leads to unstable fitting in a local linear approach. The naive "oracle" has a similar problem as the fitting of the mean function is based on functional principal components regression. Meanwhile, Algorithm 3, based on a partial linear signature conformity score, produces reasonably good results in all settings. Finally, we see that the

split sample method tends to produce overly-large intervals, regardless of simulation setting. This is likely related to the large variance associated with out-of-sample evaluation.

The basis used to generate X and the type of relationship obviously had a large and varying impact on performance of the three algorithms. The effects of sample size and error distribution were smaller and more consistent. Overall we saw a small reduction in interval size when increasing n, which likely corresponds to more accurately estimating the mean function $\hat{m}(x)$. Some results for $\alpha = 0.1$ are in the supplementary chapter, which show no difference in performance based on significance level.

4.2 Multivariate Simulations

To example the performance of the MPL-Sig and LL-mFPCA algorithms in a multivariate predictors setting, we extended the simulation setting from Wong et al. (2019) 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 Wong et al. (2019), 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.

1000 data sets were generated, with either 200 or 800 observations in each. A new observation was generated in the same manner for each trial, and prediction intervals constructed for this observation using 4 unique methods. First, we use our algorithm based on the multiple partial linear signature conformity score. We compare this with the local linear mFPCA based algorithm. Next, we compare these intervals to a "naive oracle": 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 by the empirical quantiles. In the linear setting, we simply used a linear regression of our response Y on the mFPCA scores to obtain our $\hat{m}(x)$. We used the PLFAM

model to obtain $\hat{m}(x)$ in the additive model. Finally, we also included the interval obtained from the split sample algorithm using the MPL-Sig model for approximate conformal intervals.

Table 9: 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, naive "oracle", and split sample interval.

Error	N	MPL-Sig	LL-mFPCA	Naive "oracle"	SS Interval
Normal	200	0.949 (0.063)	0.946 (0.116)	0.939 (0.022)	0.956 (0.093)
	800	0.947 (0.058)	0.955 (0.089)	0.947 (0.009)	0.950 (0.108)
t_3	200	0.949 (0.050)	0.947 (0.091)	0.943 (0.021)	0.948 (0.093)
	800	0.953 (0.026)	0.951 (0.085)	0.949 (0.008)	0.950 (0.109)

Table 10: 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, naive "oracle", and split sample interval.

Error	N	MPL-Sig	LL-mFPCA	Naive "oracle"	SS Interval
Normal	200	0.951 (0.086)	0.949 (0.100)	0.945 (0.048)	0.942 (0.141)
	800	0.950 (0.070)	0.951 (0.087)	0.946 (0.017)	0.952 (0.118)
t_3	200	0.951 (0.062)	0.950 (0.069)	0.941 (0.043)	0.952 (0.113)
	800	0.948 (0.055)	0.950 (0.074)	0.948 (0.012)	0.953 (0.114)

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 "oracle's" prediction intervals do tend to undercover, especially in smaller samples, but still serves as a good lower bound for interval length. The split sample intervals have accurate coverage, but have more variation than the exact algorithms.

Table 11: 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, naive "oracle", and split sample interval.

Error	N	MPL-Sig	LL-mFPCA	Naive "oracle"	SS Interval
Normal	200	4.85 (0.48)	6.44 (0.49)	3.90 (0.27)	7.40 (0.63)
	800	4.60 (0.44)	6.36 (0.21)	3.93 (0.14)	7.40 (0.36)
t_3	200	5.37 (0.80)	6.80 (0.60)	4.53 (0.59)	7.78 (0.85)
	800	5.16 (0.57)	6.71 (0.29)	4.54 (0.30)	7.95 (0.54)

Table 12: 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, naive "oracle", and split sample interval.

Error	N	MPL-Sig	LL-mFPCA	Naive "oracle"	SS Interval
Normal	200	5.68 (0.58)	5.95 (0.44)	4.55 (0.32)	8.21 (0.76)
	800	5.21 (0.47)	5.77 (0.21)	4.03 (0.13)	8.44 (0.45)
t_3	200	6.08 (0.73)	6.31 (0.57)	4.92 (0.56)	8.63 (0.97)
	800	5.66 (0.60)	6.27 (0.27)	4.59 (0.29)	8.92 (0.56)

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.

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.

We see that the approximate split sample intervals are far larger than the intervals from either of the exact methods, justifying our focus on deriving closed form solutions for these prediction sets.

5.0 Public Datasets

5.1 PM 2.5 Dataset

Photos of smog-covered Chinese cities have captured the attention of international media and prompted organizations worldwide to search for solutions to this growing issue. Several studies have identified fine particulate matter as especially harmful to health. Specifically, particulate matter with diameter less than 2.5 micrometers (PM2.5) has been linked to an increased risk of morbidity and mortality from cardiovascular and respiratory diseases. Therefore, it is important to be able to model the PM2.5 level and predict its level in the future. This allows people to take steps to minimize exposure at times with high pollution.

A group lead by Songxi Chen gathered meteorological and PM data collected hourly at various posts in 5 Chinese cities. Liang et al. (2016) identified several variables affecting PM2.5 levels: air pressure, dew point, temperature, cumulative wind power, cumulative precipitation, and wind direction. Furthermore, PM2.5 levels were highly elevated between November and March due to the use of fossil fuels for heating during winter. With these findings, we would like to apply the proposed method to construct prediction intervals for PM2.5 levels that provide actionable information.

To prepare the data for analysis, we averaged the PM2.5 measurements from 6 to 8 AM to estimate air quality during morning rush hour. As the PM2.5 readings were highly right skewed, the data was logged. Air humidity was treated as a functional predictor, spanning from 9 AM the previous day to 5 AM the day we wish to predict. The daily mean and signature expansion of the humidity data was obtained. All other

meteorological variables were treated as scalar predictors, using the value they have at 5 AM. The partial linear algorithm was then used to obtain prediction intervals for log(PM2.5) with the humidity signature and other weather variables as predictors. It is worth mentioning that the previous day PM2.5 value adds only a small amount of predictive power for today's value. We performed analysis with/without PM2.5 from previous day as a predictor and the results were similar.

The predicted intervals (leave-one-out prediction) as well as the observed PM2.5 values are plotted in Figure 4 for 95% coverage and in Figure 5 for 50% coverage. We can see that the 95% intervals are big, which is partially due to the fact that PM2.5 prediction is a difficult task, and the predictors only explain a moderate amount of the variation.

To further check the predictive power, we referenced the PM2.5 concentration categories described in Liang et al. (2015): Low ($<35\mu g/m^3$), High (35-$150\mu g/m^3$), and Severe ($>150\mu g/m^3$). We first look at the classification results based on the point prediction as shown in Table 13. The point prediction is from the signature based partial linear model. The performance is reasonable, making correct category predictions 70% of the time. We then look at the predicted intervals. We classify it as "low to high" if the predicted interval spans two categories "low" and "high", etc.

In Table 14 to Table 16, we report the classification results for $\alpha = 0.05$, 0.2, and 0.5. When the coverage level increases, the predicted intervals get larger and unlikely to fall into one single category entirely, and the classification outcome becomes less specific, but with a guaranteed coverage, the misclassification rate is always controlled under α. This mimics the trade off between "specificity" and "sensitivity".

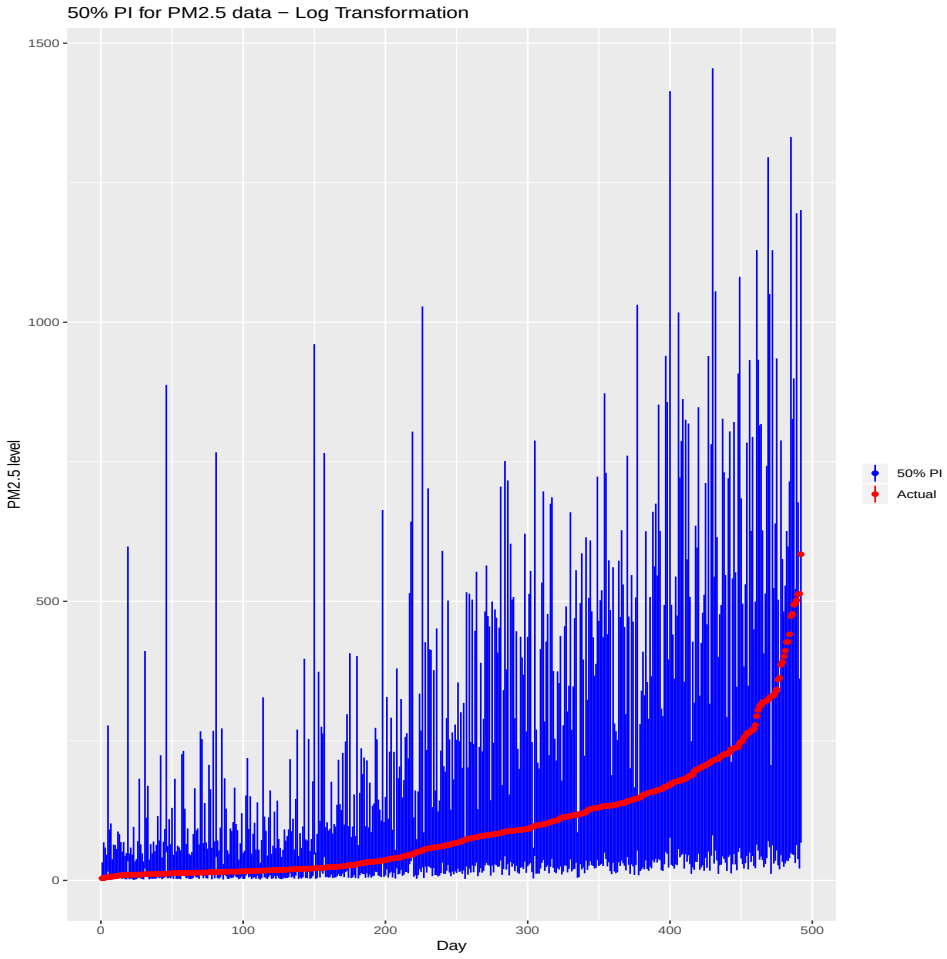

Figure 4: Plot of 95% prediction intervals vs. actual PM2.5 values

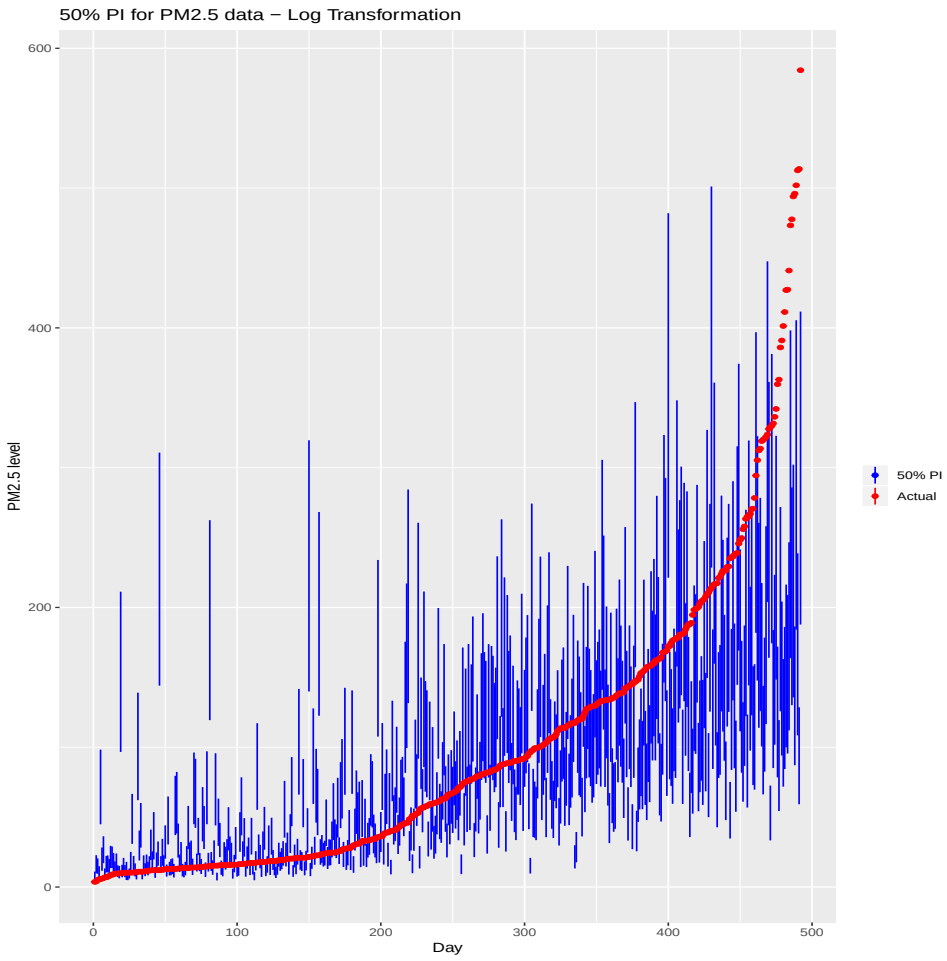

Figure 5: Plot of 50% prediction intervals vs. actual PM2.5 values

Table 13: Classification Accuracy of Point Prediction

Actual Category	Predicted Category		
	Low	High	Severe
Low	145	47	4
High	18	151	14
Severe	0	65	48

Table 14: Classification Accuracy of 95% Conformal Interval

Actual Category	Predicted Category/Categories					
	Low	Low to High	High	High to Severe	Severe	Low to Severe
Low	5	140	0	4	0	47
High	0	20	0	15	0	148
Severe	0	0	0	52	0	61

Table 15: Classification Accuracy of 80% Conformal Interval

Actual	Predicted Category/Categories					
Category	Low	Low to High	High	High to Severe	Severe	Low to Severe
Low	62	122	0	11	0	1
High	2	69	0	98	1	14
Severe	0	3	0	103	12	7

Table 16: Classification Accuracy of 50% Conformal Interval

Actual	Predicted Category/Categories					
Category	Low	Low to High	High	High to Severe	Severe	Low to Severe
Low	118	47	26	5	0	0
High	7	33	72	70	1	0
Severe	0	2	26	73	12	0

5.2 Crop Yield Dataset

To illustrate our method in a multiple predictor setting, we analyzed the crop yield dataset described in Wong et al. (2019). 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 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 6 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 Wong et al. (2019) 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.

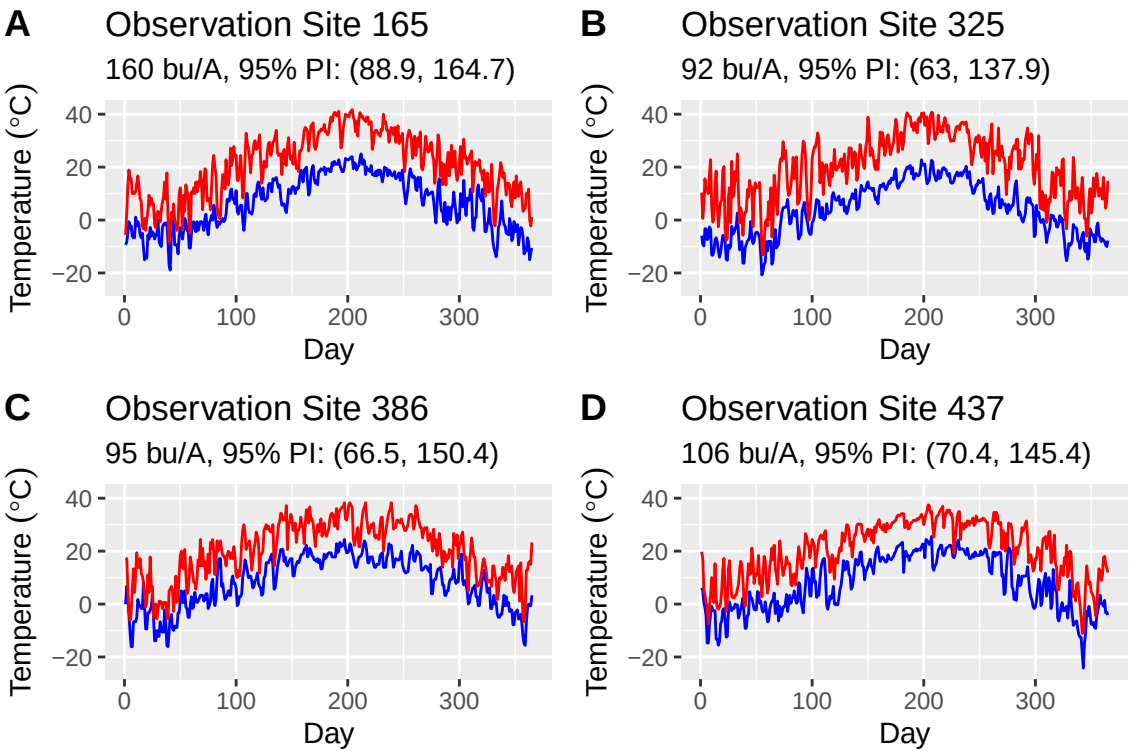

Figure 6: Daily minimum (blue) and maximum (red) temperatures for 4 randomly selected observation sites. Actual corn yield in bushels per acre reported alongside 95% out of sample prediction interval.

6.0 Conclusions

In this paper, we have derived several algorithms for constructing prediction intervals for multiple functional regression, while making minimal assumptions. The foundation for these algorithms is the conformal prediction method, which guarantees finite sample coverage while only requiring exchangeability of the data and choice of conformity score. Choosing an appropriate conformity score not only allows for construction of exact prediction intervals, but also efficiency of the constructed interval. Furthermore, we have developed conditions for each of the algorithms which guarantee that the resulting prediction set is a contiguous interval.

In simulations and analysis of public data sets, these algorithms showed encouraging results. Across simulations with different sample sizes, error distributions, and mean relationships, we found that using an appropriate algorithm for the setting resulted in intervals not much larger than the "oracle" intervals. In public data sets, we found that the algorithms produced useful prediction intervals, although they can obviously be large when the relationship is sufficiently complex.

This paper also explored the use of a relatively novel feature set for representing functional data. Thanks to the shuffle property, the signature of a set of functions can be an extremely useful tool for modeling non-linear transformations of those functions. The ability to express any transformation of the functions as a linear combination of the feature set allows us to avoid some of the issues associated with nonparametric methods in high dimensions.

Because the methods outlined in this paper are so generalizable, it is easy to think of possible extensions of this work to related problems. Most obvious would be extending the methods to conformity scores based on other multiple functional

regression estimators. There may be strong theoretical or empirical reasons for a researcher to hold specific beliefs about their data (perhaps it has a certain sparsity structure, or fits a certain form of relationship), and that knowledge can be used to select a conformity score best suited for the problem at hand. Even if this conformal score does not allow for construction of an exact prediction set, the split sample conformal interval is often better than nothing.

Additionally, we can consider regressions with a broader class of predictors. Beyond the "standard" functional predictor, researchers may wish to use various forms of data such as images and surfaces to predict their response. As these types of data are even more complex to analyze, the value of a distribution-free interval with finite sample coverage increases greatly.

We can also consider types of analysis without scalar responses. In the PM 2.5 example, pollution levels can be classified as "Low", "High", or "Severe". Although we approached this problem by converting a numerical prediction interval into a prediction set of categories, a more sensible way to do this is to transform the continuous response to a variable with values 0, 1, and 2 and directly train a functional classification model with a functional conformal method. The conformal classification method has been studied in Lei (2014) and Sadinle et al. (2019), and the extension to functional predictors is of great interest. It seems likely that there exists a conformity score which would allow for construction of exact classification sets. Alternatively, the response might also be functional. Choosing a conformity score that allows for construction of exact intervals seems difficult in this setting, but could be possible with careful analysis.

Ultimately, this paper contributes to an area which has received relatively little attention. Prediction intervals are a key tool for researchers across all fields when regressing on multivariate scalar data, but few methods exist for functional predic-

tors. Those that do exist often require strict modeling assumptions, in contrast to this method. As such, the conformal prediction approach can play an important role in better understanding regression models across a wide range of fields.

7.0 Supplemental Material

7.1 Additional Results

The following algorithm details the process to produce conformal prediction sets when the Contiguity Condition 1 is not satisfied.

Algorithm 1b

1. Calculate the residuals r_i^n for each point (X_i, Y_i) using the linear regression estimates from the data $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$.
2. Calculate the hat matrix $H = X(X'X)^{-1}X'$, where $X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_{n+1} \end{bmatrix}$.
3. Calculate $a_i = \frac{r_i^n}{1-h_{n+1,n+1}+h_{i,n+1}}$ and $b_i = \frac{-r_i^n}{1-h_{n+1,n+1}-h_{i,n+1}}$ for $i = 1, 2, \ldots, n$.
4. Construct n regions in the following manner:

   If $\text{sgn}\,(1 - h_{n+1,n+1} + h_{i,n+1}) = +$ and $\text{sgn}\,(1 - h_{n+1,n+1} - h_{i,n+1}) = +$, then your region is $[\min(a_i, b_i), \max(a_i, b_i)]$.

   Else, your region is $(-\infty, \min(a_i, b_i)] \cup [\max(a_i, b_i), \infty)$.
5. Take the intersection of unions of all combinations of $\lceil (n+1)(1-\alpha) \rceil$ of these sets. This resulting set R is the set of all residuals in the $100(1-\alpha)\%$ prediction interval for Y_{n+1}.
6. Construct the $100(1-\alpha)\%$ prediction interval for Y_{n+1} as $\hat{y}_{n+1}^n + R$.

Steps 4 to 6 may be used to produce sets in Algorithms 2 and 3 when the respective Contiguity Conditions are not satisfied.

7.2 Additional Simulations

These are the results of simulations performed with beta generated using all 22 sin/cos basis functions.

Table 17: Linear FPCA based Algorithm

Linear Relationship	Mean Coverage	Median Length
$n = 200$, $\alpha = 0.05$, Normal Error	0.951	3.0
$n = 500$, $\alpha = 0.05$, Normal Error	0.950	2.9
$n = 200$, $\alpha = 0.10$, Normal Error	0.902	2.4
$n = 200$, $\alpha = 0.05$, t_3 Error	0.950	4.7

Table 18: Linear FPCA based Algorithm

Quadratic Relationship	Mean Coverage	Median Length
$n = 200$, $\alpha = 0.05$, Normal Error	0.947	7.0
$n = 500$, $\alpha = 0.05$, Normal Error	0.952	6.9
$n = 200$, $\alpha = 0.10$, Normal Error	0.906	4.8
$n = 200$, $\alpha = 0.05$, t_3 Error	0.954	8.2

Table 19: Local Linear FPCA based Algorithm

Linear Relationship	Mean Coverage	Median Length
$n = 200$, $\alpha = 0.05$, Normal Error	0.950	3.3
$n = 500$, $\alpha = 0.05$, Normal Error	0.950	3.2
$n = 200$, $\alpha = 0.10$, Normal Error	0.900	2.7
$n = 200$, $\alpha = 0.05$, t_3 Error	0.950	4.8

Table 20: Local Linear FPCA based Algorithm

Quadratic Relationship	Mean Coverage	Median Length
$n = 200$, $\alpha = 0.05$, Normal Error	0.950	3.5
$n = 500$, $\alpha = 0.05$, Normal Error	0.952	3.5
$n = 200$, $\alpha = 0.10$, Normal Error	0.899	2.9
$n = 200$, $\alpha = 0.05$, t_3 Error	0.951	5.7

Table 21: Partial Linear Signature based Algorithm

Linear Relationship	Mean Coverage	Median Length
$n = 200$, $\alpha = 0.05$, Normal Error	0.951	3.2
$n = 500$, $\alpha = 0.05$, Normal Error	0.949	3.1
$n = 200$, $\alpha = 0.10$, Normal Error	0.904	2.8
$n = 200$, $\alpha = 0.05$, t_3 Error	0.949	5.2

Table 22: Partial Linear Signature based Algorithm

Quadratic Relationship	Mean Coverage	Median Length
$n = 200$, $\alpha = 0.05$, Normal Error	0.958	4.9
$n = 500$, $\alpha = 0.05$, Normal Error	0.953	4.5
$n = 200$, $\alpha = 0.10$, Normal Error	0.897	3.8
$n = 200$, $\alpha = 0.05$, t_3 Error	0.954	6.4

Bibliography

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

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

Cardot, H., C. Crambes, and P. Sarda
2005. Quantile regression when the covariates are functions. *Nonparametric Statistics*, 17(7):841–856.

Chen, K. and H.-G. Müller
2012. Conditional quantile analysis when covariates are functions, with application to growth data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):67–89.

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(2):395–407.

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

Fan, J., Q. Yao, and H. Tong
1996. Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, 83(1):189–206.

Ferraty, F. and P. Vieu
2006. *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media.

Geenens, G. et al.

2011. Curse of dimensionality and related issues in nonparametric functional regression. *Statistics Surveys*, 5:30–43.

Hall, P., J. L. Horowitz, et al.
2007. Methodology and convergence rates for functional linear regression. *The Annals of Statistics*, 35(1):70–91.

Hall, P., H.-G. Müller, J.-L. Wang, et al.
2006. Properties of principal component methods for functional and longitudinal data analysis. *The annals of statistics*, 34(3):1493–1517.

Hall, P., R. C. Wolff, and Q. Yao
1999. Methods for estimating a conditional distribution function. *Journal of the American Statistical association*, 94(445):154–163.

Hambly, B. and T. Lyons
2010. Uniqueness for the signature of a path of bounded variation and the reduced path group. *Annals of Mathematics*, Pp. 109–167.

Jiang, F., S. Baek, J. Cao, and Y. Ma
2018. A functional single index model. *Statistica Sinica*.

Koenker, R., A. Chesher, and M. Jackson
2005. *Quantile Regression*, Econometric Society Monographs. Cambridge University Press.

Lei, J.
2014. Classification with confidence. *Biometrika*, 101(4):755–769.

Lei, J., M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman
2018. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111.

Lei, J., A. Rinaldo, and L. Wasserman
2015. A conformal prediction approach to explore functional data. *Annals of Mathematics and Artificial Intelligence*, 74(1-2):29–43.

Lei, J. and L. Wasserman
2014. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):71–96.

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

Liang, X., S. Li, S. Zhang, H. Huang, and S. X. Chen
2016. Pm2. 5 data reliability, consistency, and air quality assessment in five chinese cities. *Journal of Geophysical Research: Atmospheres*, 121(17).

Liang, X., T. Zou, B. Guo, S. Li, H. Zhang, S. Zhang, H. Huang, and S. X. Chen
2015. Assessing beijing's pm2. 5 pollution: severity, weather impact, apec and winter heating. *Proc. R. Soc. A*, 471(2182):20150257.

Lyons, T.
2014. Rough paths, signatures and the modelling of functions on streams. *arXiv preprint arXiv:1405.4537*.

Müller, H.-G. and F. Yao
2008. Functional additive models. *Journal of the American Statistical Association*, 103(484):1534–1544.

Opsomer, J. D.
2000. Asymptotic properties of backfitting estimators. *Journal of Multivariate Analysis*, 73(2):166–179.

Politis, D. N.
2015. Model-free prediction in regression. In *Model-Free Prediction and Regression*, Pp. 57–80. Springer.

Ree, R.
1958. Lie elements and an algebra associated with shuffles. *Annals of Mathematics*, Pp. 210–220.

Robinson, P. M.
1988. Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, Pp. 931–954.

Sadinle, M., J. Lei, and L. Wasserman
2019. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234.

Silverman, B. W. et al.
1996. Smoothed functional principal components analysis by choice of norm. *The Annals of Statistics*, 24(1):1–24.

Vovk, V., A. Gammerman, and G. Shafer
2005. *Algorithmic learning in a random world*. Springer Science & Business Media.

Wong, R. K., Y. Li, and Z. Zhu
2019. Partially linear functional additive models for multivariate functional data. *Journal of the American Statistical Association*, 114(525):406–418.