



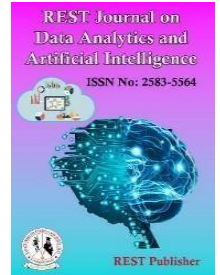
# REST Journal on Data Analytics and Artificial Intelligence

Vol: 4(4), December 2025

REST Publisher; ISSN: 2583-5564

Website: <http://restpublisher.com/journals/jdaai/>

DOI: <https://doi.org/10.46632/jdaai/4/4/2>



## Advancing Scalar-on-Function Regression: A Functional Data Analysis Approach to Basis Selection and Truncation Optimization

Harish Kasireddy

Virginia International University, USA.

\*Corresponding author Email: [Harishkasireddy87@gmail.com](mailto:Harishkasireddy87@gmail.com)

**Abstract:** The study Functional Data Analysis (FDA) has emerged as a critical framework for analyzing data with inherent smoothness, enabling insights beyond classical statistical methods. This paper focuses on the scalar-on-function regression paradigm, where the predictor is a functional variable, necessitating specialized methodologies due to the infinite-dimensional nature of the coefficient function. We introduce and compare two key approaches for transforming this infinite-dimensional problem into a finite-dimensional one: basis expansion and Functional Principal Component Regression (FPCR). Both techniques rely on a truncation parameter to approximate the coefficient function, a crucial element that influences model performance. Through theoretical exploration and Monte Carlo simulations, we examine strategies for optimal truncation parameter selection using cross-validation. Finally, we apply these insights to predict octane numbers of gasoline samples from Near-Infrared (NIR) absorption spectra, showcasing the practical implications of functional regression in real-world applications. Limitations and potential extensions are discussed, paving the way for future research into robust and adaptive FDA methodologies.

**Keywords:** Monte-Carlo Simulation, Functional Data Analysis, Near-Infrared absorption curves, Functional Linear Regression.

### 1. INTRODUCTION

Functional Data Analysis (FDA) has its roots in the work of Ulf Grenander and Kari Karhunen and is gaining traction as researchers from different fields collect data that is functional in nature. Although classical statistical methods can often process this data, FDA has advantages allowing it to extract information given by properties such as the smoothness of the underlying process or its derivatives [1]. As Kokoszka and Reimherr [2] describe, using methods from FDA should be considered when one variable of a given data set can be seen as smooth curves. Examples of such curves are the absorption curves of electromagnetic radiation in the Near-infrared (NIR) spectrum by chemical samples. This paper introduces Functional Linear Regression in a scalar-on-function setting. The distinct feature of this framework is that the regressor is a function, which makes a different approach to estimation necessary because the problem of estimating an unknown coefficient function is inherently infinite-dimensional. We introduce two ways of translating this infinite-dimensional problem into a finite-dimensional problem that can be addressed using theory from multivariate regression: First, a so-called basis expansion of the coefficient function, and second, Functional Principal Component Regression (FPCR). Both methods depend on a parameter called a truncation parameter for a functional basis, and this paper focuses on exploring the selection of these parameters using cross-validation.

In Section II, we introduce the necessary theoretical concepts, describe the estimation procedures, and address the theoretical importance of the truncation parameter. Section III contains a description of our Monte-Carlo Simulation, which aims to provide information on how to choose an appropriate functional basis and truncation parameter for the aforementioned methods. The application in Section IV uses the insights from theory and simulation to choose an

appropriate basis for the estimation of octane numbers of gasoline samples based on Near-Infrared absorption curves. In Section V, we describe the limitations of our approach and give an outlook on possible extensions for this paper.

## 2. THEORY

To introduce scalar-on-function regression, it is necessary to extend some concepts from multivariate regression to the realm of infinite-dimensional objects, as the statistics derived from infinite-dimensional random functions cannot be defined on a finite-dimensional space. One integral concept that must be defined are random functions as a special case of random variables. Paraphrasing a definition by Bauer [3], a random variable  $X: \Omega \rightarrow \Omega'$  is an  $A$ - $A'$ -measurable function, where  $(\Omega, A, P)$  is a probability space and  $(\Omega', A')$  is a measure space. The typical case for a random variable realizing in  $\mathbb{R}$  is  $(\Omega', A') = (\mathbb{R}, B)$ , where  $B$  is the canonical  $\sigma$ -algebra on the real numbers. As a first intuition, it is possible to imagine a similar concept where a random variable does not realize as an element of the real numbers but as a function in a function space. A formalization of this idea makes some more in-depth considerations necessary. The following theoretical introduction closely follows chapters 2.3 and 2.4 from Hsing and Eubank [4] and chapters 4.4 and 4.6 from Kokoszka and Reimherr [2].

### 1. Inner Products and Hilbert Spaces

The first concept we will introduce is the concept of Hilbert spaces. We start from inner product spaces but restrict our analysis to vector spaces over  $\mathbb{R}$  for clarity. Let  $V$  be a vector space over  $\mathbb{R}$ . Then, a function  $\langle \bullet, \bullet \rangle: V \times V \rightarrow \mathbb{R}$  is called an inner product, if  $\forall v, v_1, v_2 \in V$  and  $a_1, a_2 \in \mathbb{R}$  the following properties hold.

- 1)  $\langle v, v \rangle \geq 0$
- 2)  $\langle v, v \rangle = 0$  if  $v = 0$
- 3)  $\langle a_1 v_1 + a_2 v_2, v \rangle = a_1 \langle v_1, v \rangle + a_2 \langle v_2, v \rangle$
- 4)  $\langle v_1, v_2 \rangle = \langle v_2, v_1 \rangle$

A vector space with an associated inner product is called an inner product space. The inner product defines a norm and an associated distance on the vector space.

$$\|v\| = \langle v, v \rangle^{\frac{1}{2}} \text{ and } d(v_1, v_2) = \langle v_2 - v_1, v_2 - v_1 \rangle^{\frac{1}{2}} \quad (1)$$

If the inner product space is complete with respect to the induced distance, it is called a Hilbert space, denoted  $H$  in the following. To extend the known concept of a basis in a finite dimensional space to potentially infinite Hilbert spaces, it is necessary to define the closed span of a sequence of elements of  $H$ . Recall that the span of a set of vectors  $S \subseteq \mathbb{R}^p$  is given by

$$\text{span}(S) = \left\{ \sum_{i=1}^k \lambda_i v_i \mid k \in \mathbb{N}, v_i \in S, \lambda_i \in \mathbb{R} \right\} \quad (2)$$

The closed span  $\overline{\text{span}}(S)$   $\text{span}(S)$  of a sequence  $S$  in  $H$  is defined as the closure of the span with respect to the distance induced by the norm and  $S$  is called a basis of  $H$  if  $\overline{\text{span}}(S) = H$ . It is called an orthonormal basis if, in addition, the following properties hold.

- 1)  $\langle v_i, v_j \rangle = 0 \quad \forall v_i, v_j \in S \text{ for } i \neq j$
- 2)  $\|v\| = 1 \quad \forall v \in S$

Each element of a Hilbert space can be expressed in terms of a corresponding basis. Using a Fourier expansion of an element  $x \in H$  with respect to a basis  $S = \{s_n\}$  leads to the following representation.

$$x = \sum_{j=1}^{\infty} \langle x, s_j \rangle s_j \quad (3)$$

Differing from the finite-dimensional case, these representations can be limits of series. As using an infinite number of basis functions is infeasible in applied contexts, an intuitive way to approximate elements in a Hilbert space is to use a truncated series.

$$x \approx \sum_{j=1}^K \langle x, s_j \rangle s_j \quad (4)$$

## 2. Random Functions in the Hilbert Space of Square-Integrable Functions

In functional data analysis, one Hilbert space of particular importance is the space of square-integrable functions on  $[0,1]$ , but analogous constructions can be made for different domains. Denoted by  $L^2[0,1]$ , this space consists of all Lebesgue-measurable functions  $f(t)$  on  $[0,1]$  that fulfill the following condition.

$$\|f\|_2 = \int_0^1 |f|^2 d\mu < \infty \quad (5)$$

This ensures that a random function has a finite second moment so that the variance and covariance functions can be defined. The inner product on  $L^2[0,1]$  is defined by Equation 6.

$$\langle f_1, f_2 \rangle = \int_0^1 f_1 f_2 d\mu. \quad (6)$$

A random function on  $L^2[0,1]$  can now be defined formally as a function  $X : \Omega \rightarrow L^2[0,1]$  defined on a probability space  $(\Omega, \mathcal{A}, P)$  where  $\Omega$  is a sample space with  $\sigma$ - algebra  $\mathcal{A}$  and a probability space  $P$ .

## 3. FUNCTIONAL DATA SETS

If we take a random function  $X(\omega)$  as defined in the previous section, then its realizations  $x(t)$  are called sample curves of the random function. This presence of functional observations  $x_i(t)$  in a data set defines functional data sets. However, realizations of random functions are not typically observed in their functional form. Instead, each curve is observed at a set of discrete measurement points. Consider the case of a data set containing observations  $x_i(t)$  of a random function  $X(\omega)$ .

$$x_i(t_{i,j}) \in \mathbb{R}, i = 1, \dots, N, j = 1, \dots, J_i, t_{i,j} \in [T_1, T_2] \quad (7)$$

Each curve  $x_i(t)$  exists  $\forall t \in [T_1, T_2]$ , but is only observed at measurement points  $t_{i,j}$ . These measurement points can be different for each sample curve. In this paper, we only consider the case where curves share their measuring points. To use the unique capabilities of functional data analysis with functional data obtained in this form, it is necessary to restore its functional structure. Therefore, we introduce methods such as basis representations in the following parts of the Theory Section.

As in the finite-dimensional setting, the concept of independent and identically distributed (i.i.d.) data is important for many aspects of functional data analysis. One example of i.i.d. curves could be Near-Infrared absorption spectra of gasoline samples where each sample is produced by the same production process and can therefore be interpreted as a realization of and i.i.d. random process itself. More information about this example can be found in Appendix VI-A.

## 4. REPRESENTING A FUNCTION IN TERMS OF A BASIS

As previously described, a basis of a Hilbert space can be used to express its elements using a Fourier expansion. Let  $\{\phi_i(t) \mid i \in \mathcal{I}\}$  be a basis used to express a realization  $x(t)$  of  $X(\omega)$ . The following equation shows how to use a basis to express a function as a weighted sum of its elements

$$x(t) = \sum_{j \in \mathcal{I}} a_j \phi_j(t) \quad (8)$$

One important question in this context is how the coefficients  $a_j$  for  $j \in \mathcal{I}$  are derived for a given function  $x(t)$ . In this paper, this process will remain a black box, but detailed information on the derivation of these coefficients can be found in chapter 4 of Ramsay and Silverman [5]. Three examples of bases used to approximate elements of  $L^2[0,1]$  in practice and in the later parts of this paper are explained in the following. Diagrams showing these bases are found in Appendix VI-B.

- a) *Monomial Basis:* A straightforward idea to approximate functions in  $L^2[0,1]$  is to take inspiration from the well-known Taylor expansion and to use the monomials as a basis. For entire functions  $f(t)$ , such as polynomials, the exponential function, or trigonometric functions, we can express the function as a potentially infinite but converging sum of weighted monomials.

$$f(t) = \sum_{i=1}^{\infty} a_i t^i \quad \text{where} \quad a_i = \frac{f^{(i)}(0)}{i!} \quad (9)$$

The monomials are only a basis for the space of entire functions and not for  $L^2[0,1]$ . However, even for functions that do not fall into this category, using a truncated Taylor expansion around a chosen point can lead to reasonable approximations around this specific point or even on  $R$  as a whole. As in the case of the Taylor expansion, it is not necessary to approximate a function around zero, as shown above. Instead, one can introduce a shift parameter  $\alpha$ . This leads to the following formalization of the Monomial basis.

$$\phi_i^M(t) = (t - \alpha)^i \quad i \in \mathbb{N} \quad (10)$$

Due to the implementation of our simulation, we limit our paper to the case of  $\alpha = 0$ . A different choice of  $\alpha$  could lead to performance improvements. As the monomials are not pairwise orthogonal, this basis is prone to collinearity problems, which can result in numerically unstable estimates. This restricts the number of basis functions that can be used in the estimation procedures limiting its ability to capture pronounced local peculiarities. The effects of this problem will be addressed in more detail in later parts of this paper. The limitation to a low number of monomials can additionally lead to undesirable behavior away from the point of evaluation [5].

- b) *Fourier Basis:* In the same way the Monomial basis is connected to the Taylor series, the Fourier basis corresponds to the Fourier series, which can be used to decompose a periodic function into trigonometric functions. Equation 11 shows an example for a function  $s(x)$  with a period of  $T = 1$ .

$$s(x) = \frac{A_0}{2} + \sum_{i=1}^{\infty} A_i \cos(2\pi i x - \phi_i) \quad (11)$$

$$= \frac{a_0}{2} + \sum_{i=1}^{\infty} [a_i \cos(2\pi i x) + b_i \sin(2\pi i x)] \quad (12)$$

Typically, the series is represented in the so-called amplitude phase form. This, however, is impractical for the estimation procedures shown in the later parts of this paper due to the phase shift parameter. Rewriting the series in its sine-cosine form, as shown above, is necessary. The Fourier basis for  $L^2[0,1]$  is thus given by the following sequence of functions defined on  $[0,1]$  directly corresponding to the terms of the sine-cosine form of the Fourier series.

$$\phi_i^F(x) = \begin{cases} 1 & \text{if } i = 1 \\ \sqrt{2} \cos(\pi i x) & \text{if } i \text{ is even} \\ \sqrt{2} \sin(\pi(i-1)x) & \text{otherwise} \end{cases} \quad (13)$$

To stay true to the original amplitude-phase form, it is reasonable to restrict the number of Fourier basis functions to odd-numbered values. The Fourier basis' elements are cyclical which is useful to expand functions that represent a periodic or seasonal underlying process. Due to the nature of the trigonometric functions, it is especially suitable to expand functions with a similar curvature across their domain, resulting in uniformly smooth expansions [4].

- c) *B-spline Basis*: Following chapter 3.5 from Ramsay and Silverman [5], splines are defined by first dividing an interval of interest  $[\tau_0, \tau_B]$  into  $B$  subintervals of non-negative length bounded by a non-decreasing sequence of points  $(\tau_b)_{b=0, \dots, B}$  called knots. On each subinterval, a spline is a polynomial of order  $m = n + 1$  where  $n$  is its degree. If there are no multiplicities in the set of knots, the polynomials on neighboring subintervals share derivatives up to order  $m - 2$  at the boundary knot  $\tau_b$ . A typical case that is often used in practice is an equidistant grid of knots. In some settings, however, it can be sensible to place multiple knots at the same value to replicate specific properties of the data structure, allowing for a reduced number of matching derivatives at the corresponding knots. For the purposes of this paper, we will focus on the case of equidistant knots without multiplicity at inner knots.

B-splines are a specific system of spline functions developed by Boor [6] and are defined by a recursive procedure. Let  $\phi_{b,m}^{BS}(x)$  for  $b \in \{1, \dots, B + m - 2\}$  be a B-spline of order  $m$  for an interval  $[\tau_0, \tau_B]$  and knots  $\{\tau_b \mid b = 0, \dots, B\}$ , then it is defined by the Cox-de Boor recursion formula as follows.

$$\begin{aligned} \phi_{b,0}^{BS}(x) &= \begin{cases} 1 & \text{if } x \in [\tau_b, \tau_{b+1}) \\ 0 & \text{otherwise} \end{cases} \\ \phi_{b,m}^{BS}(x) &= \frac{x - \tau_b}{\tau_{b+m} - \tau_b} \phi_{b,m-1}^{BS}(x) + \frac{\tau_{b+m+1} - x}{\tau_{b+m+1} - \tau_{b+1}} \phi_{b+1,m-1}^{BS}(x) \end{aligned} \quad (14)$$

As this equation references knots that are not defined by the original vector of knots, implementations typically repeat the knots at the boundaries of the interval,  $\tau_0$  and  $\tau_B$ , an additional  $m$  times. This padding of the knot vector allows calculating every object that is needed for the definition of the basis over the original set of knots. This does not lead to a basis of  $L^2[0,1]$  as the closed span of this finite sequence of functions is not equal to  $L^2[0,1]$ . However, to focus on specific approximation errors in the later parts of this paper, we will assume that a B-spline basis representation of a function in  $L^2[0,1]$  will serve as a sufficient approximation for an appropriately chosen number of B-spline basis functions. As the B-spline basis does not have infinitely many elements, it is slightly misleading to speak of truncating the B-spline basis at a truncation parameter  $L$ . For the sake of keeping the notation concise, we will still keep this notation. By convention, truncating a B-spline basis at truncation parameter  $L$  shall mean using a B-spline basis consisting of  $L$  functions from this point on.

## 5. APPROXIMATION AND SMOOTHING VIA BASIS TRUNCATION

Realized curves from a data set can be expressed in terms of a chosen functional basis. For this expansion, it is possible to use a complete basis of  $L^2[0,1]$ . In many cases, this is not a desirable approach as this expansion can introduce high amounts of variance or even lead to the approximation of noise in the sample curves, the latter being a typical case of overfitting. To combat this problem, smoothing methods such as acceleration penalties are employed to enforce a degree of smoothness in the analyzed curves. On the other hand, important information on the curves could be

missed by oversmoothing the data, giving too much weight to a chosen penalty term leading to oversmoothing and loss of valuable information. This is a typical occurrence of the Bias-Variance tradeoff.

A usual setup is described by Goldsmith et al. [7] in which an explicit smoothing term is used to tune the smoothness of the estimator  $\hat{\beta}(t)$  while setting the number of basis functions sufficiently high. To provide intuition for this approach, let

$$PSSE_{\lambda}(\alpha, \beta) = \sum_{i=1}^N \left[ Y_i - \alpha - \int_0^1 \beta(t) X_i(t) dt \right]^2 + \lambda \int [D^m \beta(t)]^2 dt \tag{15}$$

denote the penalized residual sum of squares for the derivative of order  $m$ . A typical choice is the second derivative, as highly variable functions are expected to exhibit large second derivatives and therefore a larger penalty. The smoothing parameter  $\lambda$  is set to minimize the  $PSSE_{\lambda}(\alpha, \beta)$ , which can be achieved by different criteria as shown in Lee [8]. A different approach is to enforce smoothing by limiting the number of basis functions in the approximation of the functional objects. Here, the parameter of choice is not a weighting term for the penalty but the number of basis functions. Exploring this alternative smoothing method in two different estimation procedures is the main focus of this paper. A truncated basis expansion as described above is given in Equation 16.

$$X(\omega_0) = x(t) = \sum_{j \in \mathcal{I}} a_j(\omega_0) \phi_j(t) = \sum_{j=1}^L a_j(\omega_0) \phi_j(t) + \delta(t) \tag{16}$$

$$\approx \sum_{j=1}^L a_j(\omega_0) \phi_j(t) \tag{17}$$

Here,  $\delta(t)$  is the truncation error and  $L \leq \max(j)$  for all  $L \in \mathcal{I}$ . In later parts, this approximation error is explicitly denoted in the derivation and then omitted for the final approximations.  $L$  can be chosen subjectively, but also through applying data-driven methods such as Cross-Validation (CV). Figure 1 shows the effect of choosing different numbers of basis functions for one NIR absorption curve from the gasoline data set, which exemplifies the tradeoffs at the core of the truncation parameter choice.

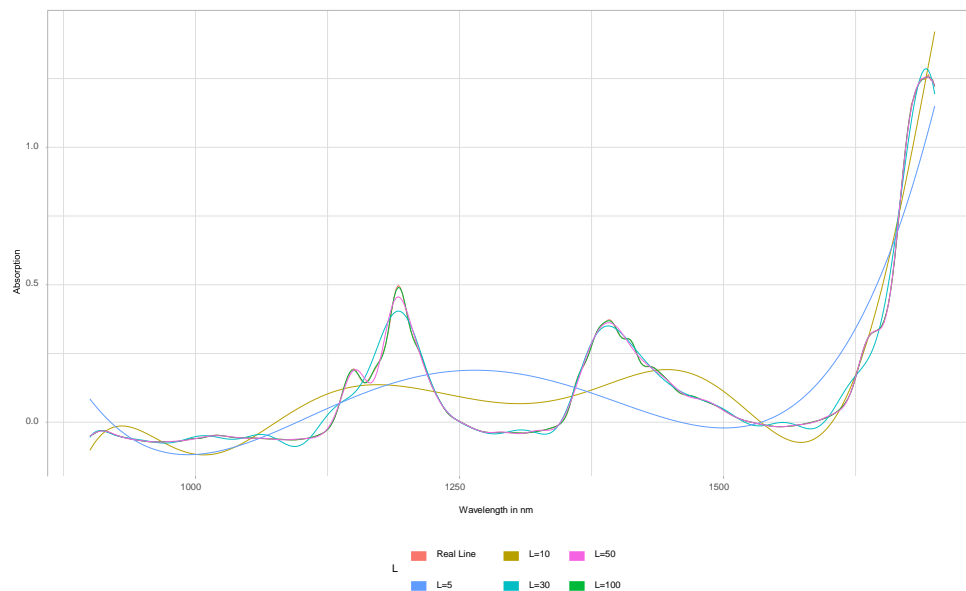


FIGURE 1. B-spline Approximations of NIR Absorption Spectra with different Basis Truncation Parameters



$$\int_0^1 \hat{c}(t, s) \hat{\nu}(s) ds = \hat{\lambda} \hat{\nu}(t) \quad (25)$$

In the following, we will call the  $\nu^m(t)$  Functional Principal Components (FPC's) to distinguish the sample analogs from the theoretical Eigenfunctions  $\nu^m(t)$ . As in the case of ordinary principal components, the number of FPC's corresponding to non-zero Eigenvalues is limited [5]. As each curve is infinite-dimensional, there is no upper limit to this number due to the dimensionality. However, the number of curves still imposes an upper limit of  $N - 1$  non-zero Eigenvalues, where  $N$  is the number of curves in the data set. A second upper limit is given by the number  $L$  of basis functions available in the derivation of the FPC's. Define therefore  $O := \min(N - 1, L)$ . In case the number of FPC's is limited by  $L$ , an exact representation of the functional observations is impossible, introducing an approximation error. For the purposes of this paper, we will not address this error and assume that an exact representation is possible. Using this assumption,  $O$  is implicitly assumed to always equal  $N - 1$ .

These sample analogs naturally lead to the following representation of each sample curve  $x_i(t)$ .

$$x_i(t) = \hat{\mu}(t) + \sum_{j=1}^O \hat{\xi}_i^m \hat{\nu}^m(t) \quad (26)$$

Here, the  $\hat{\xi}_j^m$  are derived as

$$\hat{\xi}_i^m(\omega) = \langle x_i - \hat{\mu}, \hat{\nu}^m \rangle = \int_0^1 (x_i(s) - \hat{\mu}(s)) \hat{\nu}^m(s) ds \quad (27)$$

In reality, these calculations are often implemented using basis representations of both the functional principal components  $\nu^m$  and the observations  $x_i(t)$  leading to the following representation. For clarity, we assume that the bases used for the expansion of both the observations and the coefficient function are proper bases of  $L^2[0,1]$  and can therefore be used to express the corresponding objects exactly. The  $\delta_i^J(s)$  denotes the approximation error of the demeaned observation due to the basis truncation and  $\delta_\nu^{m,K}(s)$  denotes the corresponding error for the FPC.

$$\hat{\xi}_i^m = \int_0^1 (x_i(s) - \hat{\mu}(s)) \hat{\nu}^m(s) ds \quad (28)$$

$$= \int_0^1 \left( \sum_{j \in \mathcal{I}} a_{i,j} \phi_j(s) \right) \left( \sum_{k \in \mathcal{L}} d_k^m \psi_k(s) \right) ds \quad (29)$$

$$= \int_0^1 \left( \sum_{j=1}^J a_{i,j} \phi_j(s) + \delta_i^J(s) \right) \quad (30)$$

$$= \left( \sum_{k=1}^K d_k^m \psi_k(s) + \delta_\nu^{m,K}(s) \right) ds \quad (31)$$

$$= \sum_{j=1}^J \left[ a_{i,j} \sum_{k=1}^K d_k^m \int_0^1 \phi_j(s) \psi_k(s) ds \right] \quad (32)$$

$$= + \sum_{k=1}^K d_k^m \int_0^1 \delta_i^J(s) \psi_k(s) ds \quad (33)$$

$$+ \sum_{j=1}^J a_{i,j} \int_0^1 \phi_j(s) \delta_\nu^{m,K}(s) ds + \int_0^1 \delta_i^J(s) \delta_\nu^{m,K}(s) \quad (34)$$

In practice, a typical choice is to use the same basis  $(\phi_j(t))_{j \in I}$  and the same truncation parameter  $L$  for the basis expansion of the demeaned observations  $(x_i(t) - \mu^{\wedge}(t))$  and the functional principal components  $\hat{v}^m$ . This leads to the following simplification of Equation 35.

$$\hat{\xi}_i^m = \sum_{j=1}^L \left[ a_{i,j} \sum_{k=1}^L d_k^m \int_0^1 \phi_j(s) \phi_k(s) ds \right] \quad (35)$$

$$+ \sum_{k=1}^L d_k^m \int_0^1 \delta_i^L(s) \phi_k(s) ds \quad (36)$$

$$+ \sum_{j=1}^L a_{i,j} \int_0^1 \phi_j(s) \delta_{\nu}^{m,L}(s) ds \quad (37)$$

$$+ \int_0^1 \delta_i^L(s) \delta_{\nu}^{m,L}(s) ds \quad (38)$$

We define the following objects

$$\tilde{\xi}_i^{m,L} := \sum_{j=1}^L \left[ a_{i,j} \sum_{k=1}^L d_k^m \int_0^1 \phi_j(s) \phi_k(s) ds \right] \quad (39)$$

$$\delta_{\xi,i}^L := \hat{\xi}_i^m - \tilde{\xi}_i^{m,L} \quad (40)$$

$$\tilde{v}^{m,L}(t) := \sum_{k=1}^L d_k^m \phi_k(t) \quad (41)$$

$$\delta_{\nu}^{m,L}(t) := \hat{v}^m(t) - \tilde{v}^{m,L}(t) \quad (42)$$

This method of deriving or approximating the Eigenfunctions and scores from a data set is introduced in chapter 8.4.2 of Ramsay and Silverman [5] and implemented in the R package *fda*. The following considerations and results of the simulation study might provide information about the performance of this method in a scenario where a limited number of basis functions is provided to the method instead of using more conventional smoothing approaches.

## 7. SCALAR-ON-FUNCTION REGRESSION

In the simple scalar setting, one of the essential tools in econometrics is linear regression. To motivate the jump from multivariate regression to scalar-on-function regression, assume a data generating process as follows.

$$Y = X\beta + \epsilon \quad (43)$$

Here,  $Y$  is the vector of response variables,  $X$  is the matrix containing the corresponding regressors in its columns, and  $\beta = (\beta_0, \beta_1, \dots, \beta_p)'$  is the vector containing the unknown coefficients. In this finite-dimensional setting, one important question is how to estimate the unknown coefficients  $\beta$ . The most famous estimator, the Ordinary Least Squares estimator, fulfills this purpose.

$$\hat{\beta}^{OLS} = (X'X)^{-1}X'Y \quad (44)$$

The concept of linear regression can be extended to the setting of functional data, where a scalar response variable is assumed to be dependent on a functional regressor. Integrating over the product of an observation with the coefficient function is not the only functional that can be used to create a data generating process involving functional

observations. However, it is the most typical as it naturally extends the intuition from multiple linear regression to the realm of infinite-dimensional objects. Therefore, we will always assume a data generating process as follows in this paper.

$$\delta_v^{m,L}(t) := \hat{v}^m(t) - \tilde{v}^{m,L}(t) \tag{45}$$

Similar to the finite-dimensional setting, a challenge is to estimate the unknown coefficient function  $\beta(t)$  given a data set containing realizations of a random function and associated scalar response variables. A simple extension of the OLS estimator to allow for infinite-dimensional objects is not possible. Therefore, other options have to be considered, two of which are explained in the following.

- 1) Estimation using Basis-Representation: The most common way to make this problem tractable is via a basis representation of  $\beta(t)$ . Let  $\{\phi_j(t) \mid j \in \mathcal{I}\}$  be a basis of  $L^2[0,1]$  and represent  $\beta(t)$  in terms of this basis.

$$\beta(t) = \sum_{j \in \mathcal{I}} b_j \phi_j(t) \tag{46}$$

This enables us to write Equation 45 using this representation to obtain a formulation as a sum of scalar random variables  $Z_j(\omega)$ .

$$Y(\omega) = \alpha + \int_0^1 \beta(s) X(\omega)(s) ds + \epsilon(\omega) \tag{47}$$

$$= \alpha + \int_0^1 \left[ \sum_{j \in \mathcal{I}} b_j \phi_j(s) \cdot X(\omega)(s) \right] ds + \epsilon(\omega) \tag{48}$$

$$= \alpha + \sum_{j \in \mathcal{I}} \left[ b_j \cdot \int_0^1 X(\omega)(s) \phi_j(s) ds \right] + \epsilon(\omega) \tag{49}$$

$$= \alpha + \sum_{j \in \mathcal{I}} b_j \cdot Z_j(\omega) + \epsilon(\omega) \tag{50}$$

This representation translates the original problem of regressing a scalar on a continuously observed function to a problem where a scalar is regressed on what is possibly a countably infinite sequence of regressors. Using a truncation of the basis at some parameter  $J$  can be used to make this problem tractable if we assume that the approximation error created by this truncation is small.

$$Y(\omega) = \alpha + \int_0^1 \left[ \sum_{j=1}^J b_j \phi_j(s) + \hat{\beta}^J(s) \right] X(\omega)(s) ds + \epsilon(\omega) \tag{51}$$

$$= \alpha + \sum_{j=1}^J b_j \int_0^1 \phi_j(s) X(\omega)(s) ds \tag{52}$$

$$+ \int_0^1 \hat{\beta}^J(s) X(\omega)(s) ds \tag{53}$$

$$+ \epsilon(\omega) \tag{54}$$

In practice, it is common to not only express the coefficient function in terms of a basis but also the observations. Therefore two bases,  $(\phi_j(t))_{j \in \mathcal{I}}$  and  $(\psi_k(t))_{k \in \mathcal{L}}$ , and two corresponding truncation parameters,  $J$  and  $K$ , can be chosen. This leads to the following representation.

$$Y(\omega) = \alpha + \int_0^1 \beta(s) X(\omega)(s) ds + \epsilon(\omega) \quad (55)$$

$$= \alpha + \int_0^1 \left[ \sum_{j \in \mathcal{I}} b_j \phi_j(s) \sum_{k \in \mathcal{L}} a_k(\omega) \psi_k(s) \right] ds + \epsilon(\omega) \quad (56)$$

$$= \alpha + \int_0^1 \left[ \sum_{j=1}^J b_j \phi_j(s) + \delta_\beta^J(s) \sum_{k=1}^K a_k(\omega) \psi_k(s) + \delta_X^K(\omega)(s) \right] ds + \epsilon(\omega) \quad (57)$$

$$ds + \epsilon(\omega) \quad (58)$$

$$= \alpha + \sum_{j=1}^J b_j \sum_{k=1}^K a_k(\omega) \int_0^1 \phi_j(s) \psi_k(s) ds \quad (59)$$

$$+ \sum_{j=1}^J b_j \int_0^1 \phi_j(s) \delta_X^K(\omega)(s) ds \quad (60)$$

$$+ \sum_{k=1}^K a_k(\omega) \int_0^1 \delta_\beta^J(s) \psi_k(s) ds \quad (61)$$

$$+ \int_0^1 \delta_\beta^J(s) \delta_X^K(\omega)(s) ds \quad (62)$$

$$+ \epsilon(\omega) \quad (63)$$

A typical choice in this scenario is to use the same functional basis  $(\phi_j(t))_{j \in \mathcal{I}}$  and the same truncation parameter  $L$  for both the coefficient function and the approximation of the observations. Using the following notation

$$\tilde{Z}_j(\omega) = \sum_{k=1}^L \left[ a_k(\omega) \int_0^1 \phi_j(s) \phi_k(s) ds \right] \quad j = 1, \dots, L \quad (64)$$

This leads to a considerable simplification of Equation?? and an approximation by omitting the terms containing truncation errors.

$$Y(\omega) = \alpha + \sum_{j=1}^L b_j \tilde{Z}_j(\omega) + \sum_{j=1}^L b_j \int_0^1 \phi_j(s) \delta_X^L(\omega)(s) ds \quad (65)$$

$$+ \sum_{k=1}^L a_k \int_0^1 \delta_\beta^L(s) \phi_k(s) ds + \epsilon(\omega) \quad (66)$$

$$+ \int_0^1 \delta_\beta^L(s) \delta_X^L(\omega)(s) ds + \epsilon(\omega) \quad (67)$$

$$\approx \alpha + \sum_{j=1}^J b_j \tilde{Z}_j(\omega) + \epsilon(\omega) \quad (68)$$

A model in the form of Equation ?? naturally lends itself to be estimated using theory from multivariate linear regression.

Define therefore the following objects.

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad Z = \begin{pmatrix} 1 & \tilde{Z}_{1,1} & \dots & \tilde{Z}_{1,J} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \tilde{Z}_{N,1} & \dots & \tilde{Z}_{N,J} \end{pmatrix} \quad (69)$$

Then an OLS estimator can be calculated in the usual way to obtain an estimate for the values of  $\alpha$  and  $b_j$ , and an estimate of the coefficient function can be derived accordingly.

The performance of this estimation procedure depends in part on the quality of the approximation in Equation ?. Due to the nature of basis representations, the overall approximation error  $\delta(t)$  can only decrease with the number of basis functions. However, in specific points  $t_0$  which might coincide with high predictive power for the response, the error  $\delta(t_0)$  could increase, potentially increasing the prediction error. In addition, we have the potential benefits of smoothing such as reducing the potential of overfitting functional noise and reducing the degrees of freedom in the estimation. Because of this tradeoff between a reduction in the overall approximation error and beneficial smoothing properties, increasing the number of basis function and thereby decreasing the approximation error is not necessarily beneficial for the prediction accuracy. Therefore, the behavior of the estimator when changing the truncation parameter  $L$  is not intuitively clear even if we stay in the same basis system.

If we additionally derive and approximate the principal components and corresponding scores using a truncated basis representation as in Equation ??, we obtain the following. To not complicate things unnecessarily, the following equation assumes that the same basis  $(\phi_j(t))_{j \in I}$  was used in the derivation of the principal components and the expansion of the coefficient function. Additionally, the following approximation truncates the basis for the expansion of the coefficient function at the same parameter  $L$  that was used for the approximation of the principal components and scores.

Defining the following notation

$$\tilde{\alpha}^L = \alpha + \sum_{j=1}^L b_j \int_0^1 \phi_j(s) \hat{\mu}(s) ds + \int_0^1 \delta^L(s) \hat{\mu}(s) ds \quad (70)$$

we can express Equation ?? as follows.

$$y_i = \tilde{\alpha}^L + \sum_{m=1}^O \tilde{\xi}^{m,L}_i \cdot \sum_{j=1}^L b_j C_j^{m,L} + \epsilon_i$$

Higher-order terms

$$(71)$$

where:

$$C_j^{m,L} = \int_0^1 \tilde{\nu}^{m,L}(s) \phi_j(s) ds \quad (72)$$

The parameter  $M \in \{1, \dots, O\}$  corresponds to the chosen number of principal components and constitutes another choice in the approximation. Using  $M$  functional principal components leads to the following approximation

$$\begin{aligned} y_i &= \tilde{\alpha}^L + \sum \tilde{\xi}_i \mathbf{c} + \epsilon, \quad \mathbf{c} \\ &= \sum b_j \langle \phi_j, \tilde{\mathbf{v}} \rangle. \end{aligned} \quad (73)$$

As in the previous section, this equation lends itself for estimation with OLS and we can define the following objects.

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad Z = \begin{pmatrix} 1 & \tilde{\xi}_1^{1,L} & \dots & \tilde{\xi}_1^{M,L} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \tilde{\xi}_N^{1,L} & \dots & \tilde{\xi}_N^{M,L} \end{pmatrix} \quad (74)$$

We can then derive the following estimator for  $\tilde{\alpha}^L$  and  $\tilde{\mathbf{b}}^{m,L}$   $m = 1, \dots, M$ .

$$\begin{aligned} \tilde{\mathbf{b}} &= (Z'Z)^{-1} Z'Y \in \mathbb{R}^{M+1}, \quad \hat{\alpha} \\ &= \tilde{b}_1, \quad \hat{\beta}(t) = \sum_{m=1}^M \tilde{b}_{m+1} \tilde{\mathbf{v}}^m(t) \end{aligned} \quad (75)$$

As in the previous case, the performance of this estimation depends in parts on the quality of the approximations in the derivation of this estimator. The interactions are even more complex. Even though the approximations exhibit smaller errors when a larger number of basis functions is used, the interplay of the chosen way of smoothing in the construction of the FPC's and the choice of a number of FPC's used in the linear regression makes the behavior of this estimator difficult to predict.

### 3. SIMULATION STUDY

#### 1. Motivation

Instead of generating data from scratch, we use the gasoline data set from *R* package *refund*, which consists of 60 samples of Near-infrared absorption curves measured in increments of 2 nm from 900 to 1,700 nm, and a response variable, the octane number. We chose this setup to improve the approach towards the application in which we predict the octane numbers from the gasoline data set. We introduced different bases in Section II-D and demonstrated the importance of the truncation parameter  $L$  for the estimation in Section II-G. For the simulation study, we use Basis Expansion Regression and Functional Principal Component Regression (FPCR) with the introduced bases and focus on choosing the truncation parameter  $L$  as well as the number of FPC's by ten-fold CV using the Mean-Squared Prediction Error (MSPE). In practice, the number of FPC's is often chosen by using the lowest number that explains a specified proportion of variance of the smoothed curves [2]. This might not be optimal since FPC's with smaller Eigenvalues may have a more significant influence on the prediction [10]. In this simulation certain Eigenfunctions could correspond to certain chemical compounds and vibration overtones in the absorption bands of the spectrum that could have high predictive power, but explain only little variability in the NIR curves shown in Appendix VI-A.

#### 2. Generating Similar Curves

To avoid small sample problems, we generated 200 similar curves denoted by  $NIR_{sim}$ , from the absorption curves of the gasoline data set denoted  $NIR_{orig}$ . Our approach is motivated by Karhunen-Loeve Expansion. First, the initial curves are mapped to the interval  $[0,1]$  as described in Section II-B and expressed in terms of a cubic B-spline basis which is created using 50 knots. In the implementation of the *fda* package, these 50 knots account for 52 basis functions  $(50 + 4 - 2)$ . These smoothed curves are centered before applying the Karhunen-Loeve Expansion. For the purposes of data generation, we assume that the scores follow a multivariate normal distribution  $\tilde{\xi} = (\tilde{\xi}^1, \dots, \tilde{\xi}^M)' \sim \mathcal{N}(0_M, \text{diag}(\hat{\lambda}^1, \dots, \hat{\lambda}^M))$ . Finally, we obtain the generated curves  $NIR_{sim}$  as realizations of

$$\hat{X}(\omega)(t) = \hat{\mu}(t) + \sum_{m=1}^M \hat{\xi}^m(\omega) \tilde{v}^{m,L}(t) \quad (76)$$

where  $\hat{X}(\omega)(t)$ ,  $\hat{\mu}(t)$  and  $\tilde{v}^{m,L}(t)$  are approximated as vectors in  $\mathbb{R}^{401}$  for  $M = 30$  FPC's.

### 3. Simulation Setup

The simulation study follows Reiss and Ogden [11] as a guideline. Two different true coefficient functions,  $f_1(t)$  and  $f_2(t)$ , that differ in their smoothness, are created to compare our methods with differing true coefficient functions.

$$f_1(t) = 401 [2 \sin(0.5\pi t) + 4 \sin(1.5\pi t) + 5 \sin(2.5\pi t)] \quad (77)$$

The function  $f_2(t)$  was generated referring to Cardot [12] while function  $f_1(t)$  directly follows Reiss and Ogden [11].

Figure 3:2

f12(t),

bumpysmooth

function

We created two different error-terms by generating an i.i.d. standard normal error term  $Z \sim N(0,1)$  and multiplying it by two error variations  $\sigma_e$ . The error variations represent different signal-to-noise ratios of the responses to test the methods with low and high amounts of noise in the responses. They are created such that the squared multiple correlation coefficient  $R^2 = \text{var}(\langle X, f \rangle) / (\text{var}(\langle X, f \rangle) + \sigma_e^2)$  is equal to 0.9 and 0.6. The two error terms are then used to generate two sets of responses for  $f \in \{f_1(t), f_2(t)\}$ .

In total, we created four combinations, using the two coefficient functions and the two error terms. These four combinations are then used with a different number of Monomial basis functions, cubic B-spline basis functions and Fourier basis functions to predict the generated responses using the basis expansion approach and the FPCR approach.

To obtain valid out of sample properties for the FPCR, within each of the ten ten-fold cross-validation splits, we calculated the first  $n_{FPC} \in \{2,3,4\}$  FPC's for the training set  $T$ , which was smoothed with the respective basis specification. The approximated Eigenfunctions  $\tilde{v}_j^{m,L,T}$  are then used to estimate the scores of the holdout set, by Equation ???. In total, we conducted 5000 repetitions for each specified model, basis system and basis function. The models are denoted as a combination of the function  $f \in \{f_1(t), f_2(t)\}$  and created responses  $Y \in \{Y1, Y2\}$ .

### 4. Results

The discussed results and figures of  $\hat{\beta}(t)$  for the simulation can be found in Appendix VI-C and VI-E.

- 1) **Basis Expansion Regression:** These results follow from the model introduced in Section II-G1, in which we transform the smoothed curves to perform scalar-on-function regression.

Examining the results, it appears that the cross-validated MSPE exhibits a convex behavior in the number of basis functions for all bases.

- a) **Monomial Basis:** Due to the increasing collinearity problems for higher numbers of Monomial basis functions, simulations were conducted up until the sixth monomial, which already shows signs of numerical instability. For reasons outlined in Section II-D0a, they seem suited for  $f_1$ , where it shows a better performance than B-splines. A hypothesis for this could be that  $f_1$  is an entire function, which can be well approximated with a power series. In the case of  $f_2$ , this basis shows the weakest performance

- out of all bases, for which Figures 9 and ?? provide visual evidence. It seems like  $\hat{\beta}(t)$  is not changing in the amount of noise and shows exaggerated behavior at the boundaries. This weakness is especially pronounced in the MSPE for  $f_2, Y_1$ . For  $f_1$ , the simulation selects 5(3) and for  $f_2$  5(5) Monomial basis functions for the high (low) signal-to-noise ratio.
- b) **B-spline Basis:** Simulations with B-spline basis functions were possible from 4 to 18 basis functions. From 18 onward, the simulations ran into problems of numerical instability. This might be caused by the fact, that the B-splines are not pairwise orthogonal. Function  $f_1$  chooses 5(4) B-spline basis functions for the high (low) signal-to-noise ratio to obtain the best fit and shows the worst performance of the three bases. An explanation might be its extreme behavior at the boundaries and the exaggeration of the peculiarities of  $f_1$  (Figures ?? and ??). This is especially pronounced for the lower signal-to-noise ratio. For  $f_2$ , 11(6) B-spline basis functions are chosen for the high (low) signal-to-noise ratio. The B-spline basis outperforms the Monomial basis in  $f_2$  but comes second to the Fourier basis. While the basis seems to recognize the peculiarities in  $f_2, Y_1$ , it struggles for the noisy responses in  $f_2, Y_2$  (Figures 9 and ??). With the low signal-to-noise ratio in the responses, the simulation chooses a smaller number of basis functions, which could prevent overfitting the scalar noise using  $\hat{\beta}(t)$ .
- c) **Fourier Basis:** For  $f_1$ , the simulation chooses a smaller number of Fourier basis functions, 5(3) and a higher number for  $f_2$ , 9(7) for the setup with the high (low) signal-to-noise ratio. As for the B-splines basis, the simulation chooses a smaller number of basis functions for the noisy responses and a higher number for the high signal-to-noise responses. The Fourier basis performs the best for each setup for the Basis Expansion Regression. Several reasons could contribute to this. First, especially  $f_1$  shows similar curvature across the domain while the curvature of  $f_2$  does at least not display any erratic jumps. Second, both functions feature periodic behavior. Third,  $f_2$  does have the same value at the start and the end of the interval, lending itself to a periodic representation.
- 2) **Functional Principal Component Regression:** The model used for the FPCR is described in Section ?. The selection of the truncation parameter  $L$  is difficult since the approximated Eigenfunctions from the decomposition are influenced by the choice of  $L$ . In addition, the choice of the number of FPC's adds to the complexity of the model. Some eigenfunctions, which might contain important information, could correspond to small Eigenvalues and, therefore, be omitted as described in Section III-A. But since the FPCR is ultimately estimated with a linear model containing the corresponding scores as regressors, the relevant degrees of freedom in the estimation of the ultimate model are not affected by  $L$ , but only by  $n_{FPC} \in \{2, 3, 4\}$ . It seems that neither a convex behavior of the MSPE nor any clear relationship can be observed between the number of basis functions and the number of FPC's. This might be because this dependency is too complex to draw conclusions with this simulation study. Therefore, we will limit ourselves to a brief description. In this simulation, the cross-validated choice of basis functions indicates that the FPCR might not take the differing signal-to-noise ratios of the responses into account. This is indicated by the fact that the same number of basis functions is chosen for the different signal-to-noise ratios. A possible explanation might be that the FPC's, which affect the relevant degrees of freedom, are solely calculated from the smoothed curves, not considering the responses.
- a) **Monomial Basis:** For  $f_1, Y_1$ , the cross-validated MSPE seems to decrease in the number of FPC's and chosen basis functions for both the high and the low signal-to-noise ratio (4, 5, 6 basis functions for  $n_{FPC} = 2, 3, 4$ ). For the noisy responses of  $f_1, Y_2$  we find signs of overfitting for  $n_{FPC} = 4$ . In  $f_2$ , we also observe a decreasing MSPE in the number of FPC's, but no clear relationship for the chosen basis functions. The Monomial basis shows the weakest performance out of all three bases in each setting.
- b) **B-spline Basis:** For  $f_1, Y_1$ , the MSPE suggests that models with a higher number of FPC's perform better. While the number of basis functions stays at five for  $n_{FPC} = 2, 3$ , it increases to 6 basis functions for  $n_{FPC} = 4$ . In  $f_1, Y_2$ , the cross-validated MSPE is only slightly affected by  $n_{FPC}$ , but lowest for  $n_{FPC} = 3$ , indicating the possibility of overfitting for  $n_{FPC} = 4$ . Similar to the two setups with  $f_1$ , the chosen number of basis functions for  $f_2$  is increasing in the number of FPC's (4, 6, 23 for  $n_{FPC} = 2, 3, 4$ ).
- c) **Fourier Basis:** In  $f_1, Y_1$ , the MSPE decreases in  $n_{FPC}$  while the results in  $f_1, Y_2$  might indicate overfitting for  $n_{FPC} = 4$ . Both specifications using  $f_1$  select the lowest number of basis functions possible. This is interesting as the number of FPC's puts a binding lower bound on the number of basis functions used in

their construction. An exploration of the implications of this fact is out of the scope of this paper. Both configurations of  $f_2$  use 5, 15 and 7 basis functions for  $n_{FPC} = 2, 3, 4$ . The basis shows the lowest MSPE for all four settings.

### 5. Interpretation and Relevance for Application

A possible explanation applicable to the setups performing Basis Expansion Regression might be the effect of the Bias-Variance tradeoff and the following hypothesis: For  $f_1$ , only little bias seems to be introduced when choosing a small number of basis functions. For  $f_2$ , a higher number of basis functions seems appropriate. This could result from the inherent peculiarities of  $f_2$  that are more pronounced with higher  $L$ , therefore choosing higher numbers of basis functions since the bias is decreasing faster than the variance is increasing in the number of basis functions compared to  $f_1$ . The described convex behavior of the MSPE might also be partially attributable to the bias-variance tradeoff. This convex behavior was not observed for the FPCR where no clear relationship between the basis functions and the number of FPC's could be found. The MSPE of FPCR seems to be comparatively more stable when changing the number of basis functions than for Basis Expansion Regression. It seems like the chosen number of FPC's is more significant, especially for the two high signal-to-noise ratio settings which display a higher absolute and relative decrease of MSPE in  $n_{FPC}$ .

To examine this relationship further, we conducted additional simulations (Appendix VI-D) for B-spline basis specifications with a large truncation parameter  $L \in \{50, 70\}$  for  $n_{FPC} \in \{2, \dots, 7\}$ . We can find potential signs of overfitting in all setups for  $L = 70$ . For  $L = 50$ , signs of overfitting can be observed for the setups with the noisy responses. The additional simulations revealed evidence for a relationship from  $n_{FPC} = 3$  onwards for all setups. For  $n_{FPC} \in \{3, \dots, 7\}$ , the ten-fold cross-validated MSPE is always lower for 50 than for 70 basis functions. This might be the first sign of an  $L$  that is chosen too large and creates undersmoothing. It seems that once a sufficient number of basis functions is used to expand the curves, FPCR with B-splines performs better using a lower number of basis functions. While a higher number of  $n_{FPC}$  decreases the MSPE of  $f_1$ ,  $Y_1$  and  $f_2$ ,  $Y_1$  for  $L = 50$  compared to the main simulations, this does not hold true for the two setups with noisy responses. This strengthens our hypothesis of the high importance of  $n_{FPC}$  for the high signal-to-noise ratio setup. For  $n_{FPC} \in \{2, 3, 4\}$ , the additional B-splines simulations perform worse for these high number of basis functions compared to the main simulation. This might indicate beneficial attributes of a lower truncation parameter in settings with a lower number of  $n_{FPC}$ .

## 4. APPLICATION

Our application uses the methods and insights from the previous sections to predict the octane numbers of the gasoline data set. Although the simulation study granted valuable insights into the different methods in our four different settings, it is not enough to determine the optimal method and choice of basis for the application since there is too much uncertainty involved. To point out some sources of uncertainty: First, differing from the simulation setup, we do not know the true coefficient function. Visual inspection of the estimated coefficient functions shown in Appendix VI-G fuels the hypothesis that the real coefficient function might be more similar to  $f_2$  than to  $f_1$ , but the insights from the two functions are not sufficient to draw any conclusion. Second, we have no information about the signal-to-noise ratio of the measured octane numbers. Third, to generate similar curves, we made assumptions about the distribution of  $\xi$  that are not applicable in this application where we do not know their distribution.

Therefore, we rerun all specifications for the gasoline data set and use the results of the simulation study to improve our understanding of the results. The application is designed analogously to the simulation study: The 60 spectra of the gasoline data set are used to estimate a coefficient function, predict the reported octane numbers, and evaluate the results via MSPE using 12 fold CV with 5 elements per fold. In total, we conducted 1000 repetitions for each setting, choosing different fold partitions for each run. Over all specifications, the best-performing model is FPCR using 4 FPC's built on 7 Fourier basis functions (0.0435), followed by Basis Expansion Regression using 10 B-spline basis functions (0.04574).

### 1. Interpretation of Results

- 1) *Basis Expansion Regression*: The cross-validated MSPE selects 5 basis functions for the Monomial basis. For BSplines, the cross-validation selects 10 basis functions and 9 basis functions for the Fourier basis.

Comparing  $\hat{\beta}(t)$  for the different bases away from the boundaries (Figure ??), the B-spline and Fourier bases show similar behavior, differing from the Monomial estimate, which might be attributable to the lower number of basis functions. Especially at the lower boundary, the Monomial basis shows exaggerated behavior. However, we must exercise caution since from  $L = 6$  onwards, we could not calculate stable results for the Monomial basis. The MSPE for B-splines (0.04574) and Fourier (0.04808) are similar, while the Monomial basis (0.24181) shows distinctly worse properties. In contrast to the simulation study, the B-splines basis outperforms the Fourier basis. Driving factors for this improved performance could be that, first, we do not assume a periodic coefficient function with the same start- and end value and second, that the Fourier basis enforces identical start and end values on the curves of NIR. Note that the reported MSPE for the B-spline and Fourier basis are close to the errors reported in the simulation study for  $f_2$ ,  $Y_1$ , which might be caused by the hypothesized similarities between the actual coefficient function and  $f_2$ , but also by a similar high signal-to-noise ratio of the octane numbers.

- 2) *Functional Principal Component Regression*: The best performance for all numbers of FPC's was achieved with the Fourier Basis, as for the FPCR in the simulation study. In contrast to the simulation study, no evidence of overfitting was found for the best basis choices in any of the three bases. The MSPE for the optimal specification appears to be strictly decreasing in the number of FPC's. As for the simulation study, the interpretation of the results with respect to the chosen basis is difficult: Referring to the plotted estimates  $\hat{\beta}(t)$  (Figures ??, ?? and ??), it appears to be the case that the higher  $n_{FPC}$  is, the more similar the behavior of the corresponding estimates becomes. Differing from the simulation study, the number of basis functions steadily increases for the Monomial basis. The behavior for the B-spline basis is similar to the one reported in the simulation study (basis functions increasing in  $n_{FPC}$ ).

## 5. OUTLOOK

### 1. Limitations

- a) *Insights from Simulation cannot be Extended to More General Functions*: As described, the properties of the two used functions  $f \in \{f_1(t), f_2(t)\}$  influence the results. Although the resulting MSPE and visual inspection provided some evidence to support the hypothesis that the true coefficient function of the original gasoline data set might be similar to  $f_2$ , other factors that were not addressed could have caused this. Therefore, this simulation is not sufficient to base reasonable claims concerning more general coefficient functions on it.
- b) *Collinearity Problems in Basis Expansion Regression*: Except for the Fourier basis, the Basis Expansion Regression was in part limited by the numerical instability of the estimation procedure. This is mainly due to an increase in collinearity of the derived regressor matrix shown in Definition 69. This problem is inherent to basis systems whose functions are not pairwise orthogonal, such as the Monomial or B-splines bases, but gets more pronounced the more functions we add to the basis system and the higher the correlation between those functions. The numeric instability of the inversion of this matrix makes the estimates unreliable and therefore can make this approach infeasible for nonorthogonal bases in settings where the characteristics of the data set demand a higher number of basis functions than is feasible due to the properties of the estimation procedure.

### 2. Possible Extensions

- a) *Orthogonal Polynomials to Solve Collinearity Problems of Monomial Basis*: To address the collinearity problems described earlier, one possible idea would be to use a system of pair-wise orthogonal polynomials as a basis instead of the Monomial basis. One example is the system of Legendre polynomials which are orthogonal by construction and have the same closed span as the monomials [13]. Due to their orthogonality, the problem of collinearity in the regressor matrix is greatly reduced, which could allow for larger numbers of basis functions in the Basis Expansion Regression approach. The first eight Legendre polynomials are shown in Appendix VI-B.
- b) *Comparison to Penalty Based Smoothing Procedures*: In contrast to the more typical approach of using a arbitrary, large number of basis functions and smoothing using a penalty term involving, e.g. the second derivative, this paper focuses on smoothing by using a smaller number of basis functions. As the next step to

the analysis of this paper, we could compare both methods to see in which settings different approaches to smoothing perform better and if a possible combination of both approaches could be advantageous.

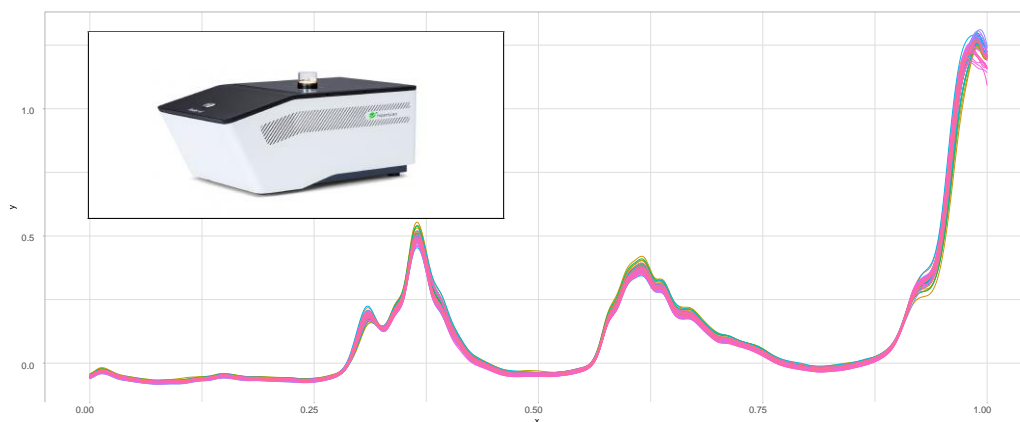
- c. *Larger Range for the Number of Specifications for FPCR*: The chosen specifications for the number of principal components and the number of basis functions used for FPCR might be insufficient to obtain a complete picture of the performance of this procedure when combined with smoothing using basis truncation. It would therefore be interesting to focus more attention on this limitation of this paper in a future extension to explore a broader array of possible specifications which might allow the derivation of more detailed insights.
- d. *Input from Physics*: It could be exciting to combine the findings of this paper with theoretical expertise from the field of physics. This could inform the choice of models and be conducive to a more meaningful interpretation of the estimates. For example, specific parts of the NIR spectrum could possibly be linked to specific molecules associated with the octane number. This could guide, for example, the construction of a B-spline basis that focuses on the relevant parts of the spectrum by using a vector of knots founded by field expertise. Alternatively, it could be used to link specific principal components to chemical compounds, which would be useful for the model selection and interesting for the interpretation of estimates.

## 6. APPENDIX

### 1. Near-infrared (NIR) Spectroscopy

NIR spectroscopy is a spectroscopic method that uses the near-infrared region of the electromagnetic spectrum (From 780 nm to 2500 nm). It measures the absorption and interaction of this spectrum of radiation with the sample. NIR spectroscopy is not only faster and cheaper than the standard test procedure – another significant advantage is that it does not need a reagent and thus does not destroy the sample. It is used for analysis in different sectors and fields, like the agrochemical industry and healthcare. Its non-invasive nature makes it also an asset for medical applications like the monitoring of diabetes in which NIR spectroscopy can detect the worsening of the blood glucose metabolic dysfunction [14].

In the context of this paper, the gasoline data set which is used for the simulation and the application is constructed using NIR spectroscopy. According to Gy et al. [15], NIR spectroscopy is a feasible method for the analysis of gasoline since most of the absorption that is observed within the described interval of wavelengths is due to overtones and interactions of the radiation with chemical combinations (carbon–hydrogen, carbon–carbon, carbon–oxygen, carbonyl associated groups, aromatic stretching, and deformation vibration of the hydrocarbon molecules). While this paper focuses on the prediction of the octane number of gasoline, other research focuses on different properties of gasoline such as the olefin, naphtaenic and aromatic content (Parisi et al. 1990, as cited in Gy et al. [15]) or the distillation characteristics (Pauls 1985, as cited in Gy et al. [15]).



**FIGURE 4.** NIR absorption spectra from the Gasoline Data Set and Finder SD (a Near-Infrared-Spectroscopy built by HiperScan GmbH)

## 2. Basic Plots



FIGURE 5. Fourier basis functions for  $i = 1, \dots, 7$

## 3. Simulation Study Results

The following results were rounded to 5 decimal places.

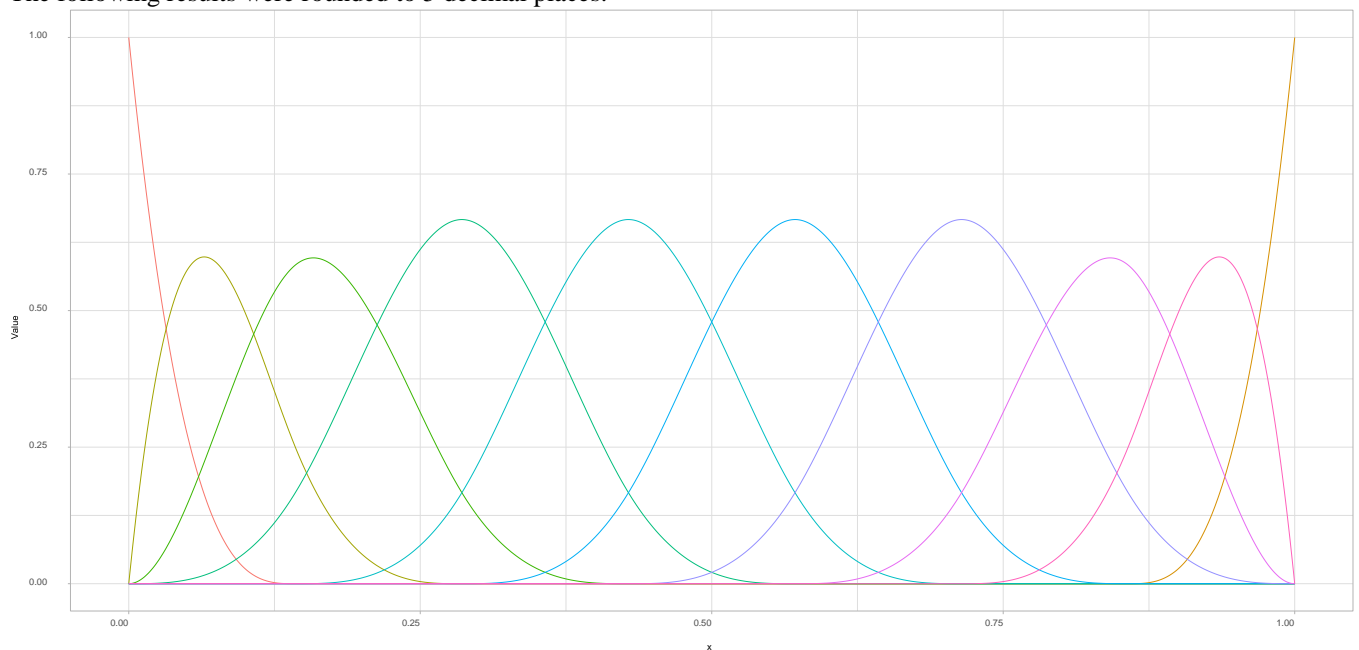
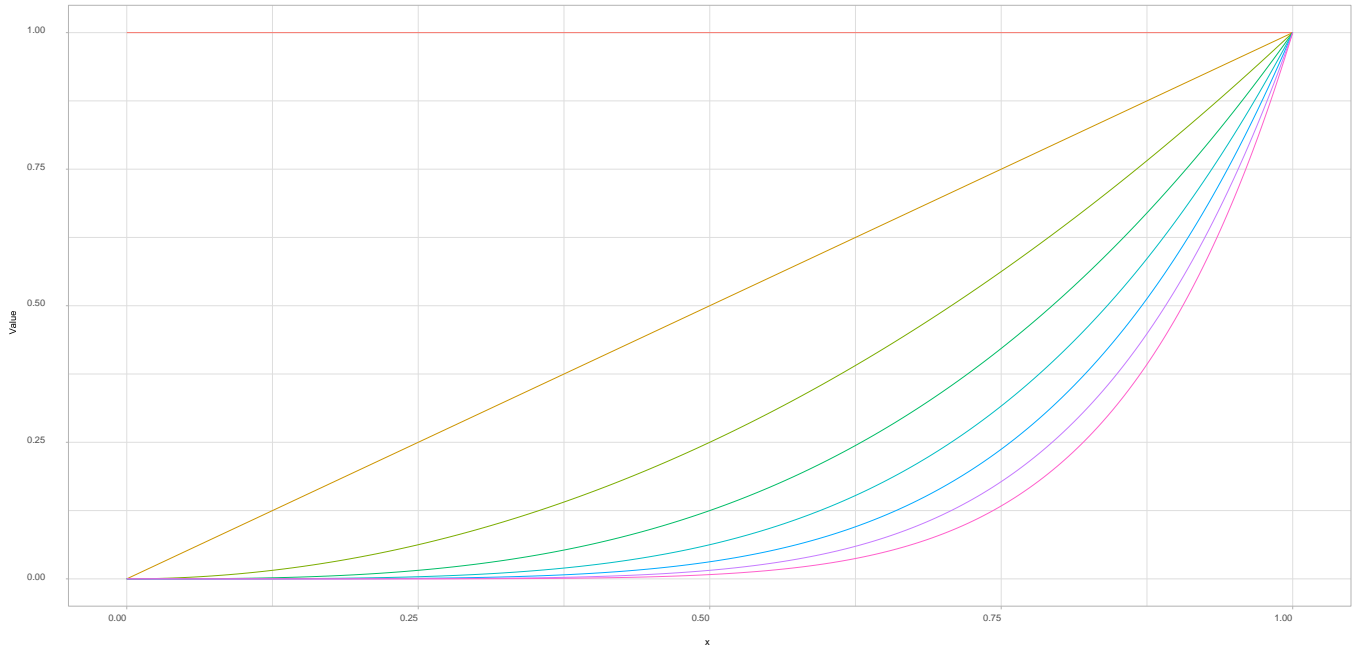


FIGURE 6. B-spline basis functions of order 4 for 8 equidistant knots on  $[0, 1]$



**FIGURE 7.** Monomial basis functions of degree 0 to 7

**4. Additional Simulation**

For each of these specifications, we ran 1000 repetitions of the simulation. The following results were rounded to 5 decimal places.

**5. Simulation - Coefficient Function Estimates**

**6. Application Results**

The following results were rounded to 5 decimal places.

**Table XIV: Application Results Basis Expansion Regression**

Monomial	B-Spline	Fourier	n_basis	fold_size	n_folds
2.29641			2	5	12
2.11258		2.07767	3	5	12
0.73444	0.73444		4	5	12
0.24181	0.2544	0.07430	5	5	12
8.88013	0.08621		6	5	12
	0.11177	0.05021	7	5	12
	0.05012		8	5	12
	0.07465	0.04808	9	5	12
	0.04574		10	5	12
	0.05629	0.05456	11	5	12
	0.05291		12	5	12
	0.06083	0.05558	13	5	12
	0.06926		14	5	12
	0.09058	0.07920	15	5	12
			16	5	12
		0.06161	17	5	12
			18	5	12
		0.10976	19	5	12

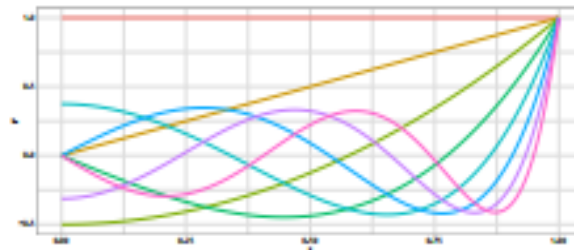


Figure 8: Legendre Polynomials of degree 0 to 7

Table I: Monomial Basis Expansion Regression

$f_1, Y_1$	$f_1, Y_2$	$f_2, Y_1$	$f_2, Y_2$	n_basis
7.17723	131.92717	0.89367	2.58395	2
4.17039	129.64259	0.81224	2.51177	3
3.91275	130.11708	0.38079	2.09047	4
3.6385	130.59817	0.09217	1.81343	5
6.01644	201.15535	0.76208	3.39373	6

Table XV: Application Results Monomial FPCR

2 FPC	3 FPC	4 FPC	n_basis	fold_size	n_folds
2.29642			2	5	12
2.17648	2.11259		3	5	12
2.21978	2.17702	0.73432	4	5	12
2.21854	2.23061	2.35328	5	5	12
2.21896	1.98803	0.84386	6	5	12
2.23156	2.07872	0.85358	7	5	12
2.19982	2.25104	0.11363	8	5	12
2.24343	2.21964	0.11476	9	5	12
2.18792	2.11694	0.12084	10	5	12
2.20919	2.15677	0.08986	11	5	12
2.27197	2.21142	0.14176	12	5	12

Table XVI: Application Results B-spline FPCR

2 FPC	3 FPC	4 FPC	n_basis	fold_size	n_folds
2.24364	2.1	0.73432	4	5	12
2.21916	2.21024	2.35745	5	5	12
2.2088	1.92703	0.55305	6	5	12
2.22042	2.10322	1.03165	7	5	12
2.17938	2.24933	0.076	8	5	12
2.20726	2.20003	0.18801	9	5	12
2.23642	2.08957	0.05669	10	5	12
2.23069	2.13131	0.09795	11	5	12
2.22878	2.14135	0.05523	12	5	12
2.1994	2.13177	0.06089	13	5	12
2.20168	2.14603	0.07197	14	5	12
2.20372	2.14908	0.05159	15	5	12
2.16851	2.13255	0.05724	16	5	12
2.17341	2.12780	0.06545	17	5	12
2.18263	2.00428	0.05399	18	5	12
2.15647	2.05871	0.06774	19	5	12
2.19268	1.96472	0.05542	20	5	12
2.17718	1.95345	0.0595	21	5	12
2.1798	1.99118	0.05519	22	5	12
2.1966	1.90692	0.05318	23	5	12
2.1786	1.97124	0.05725	24	5	12
2.18768	1.95758	0.0508	25	5	12

Table II: B-Spline Basis Expansion Regression

$f_1, Y_1$	$f_1, Y_2$	$f_2, Y_1$	$f_2, Y_2$	n_basis
3.91275	130.11708	0.38079	2.09047	4
3.64305	130.61095	0.09512	1.81643	5
3.65426	131.35355	0.07775	1.80913	6
3.67705	132.14205	0.07518	1.8168	7
3.71074	133.36537	0.0581	1.8165	8
3.71921	133.68149	0.0564	1.81931	9
3.74201	134.51217	0.05218	1.82576	10
3.7644	135.29727	0.0519	1.8361	11
3.80109	136.57581	0.05204	1.85315	12
3.83436	137.78407	0.05271	1.86997	13
3.86217	138.73431	0.05307	1.88269	14
3.86856	138.97427	0.0526	1.88482	15
3.88506	139.57419	0.05283	1.89335	16
3.91267	140.5149	0.05322	1.90619	17
3.94813	141.80885	0.05358	1.92312	18

Table III: Fourier Basis Expansion Regression

$f_1, Y_1$	$f_1, Y_2$	$f_2, Y_1$	$f_2, Y_2$	n_basis
3.69752	129.19134	0.69524	2.3944	3
3.6347	130.59282	0.07418	1.79582	5
3.67623	132.08248	0.05147	1.79343	7
3.71885	133.67575	0.05105	1.81291	9
3.76451	135.26219	0.05146	1.83463	11
3.81095	136.90282	0.05197	1.85724	13
3.85674	138.57258	0.05252	1.88021	15
3.90619	140.29178	0.05304	1.90283	17
3.95517	142.05727	0.05365	1.92718	19

Table XVII: Application Results Fourier FPCR

2 FPC	3 FPC	4 FPC	n_basis	fold_size	n_folds
2.13234	2.07788		3	5	12
2.26267	0.21409	0.19585	5	5	12
2.15572	0.06414	0.0435	7	5	12
2.16344	0.05448	0.05256	9	5	12
2.16087	0.06795	0.05276	11	5	12
2.13984	0.05848	0.05154	13	5	12
2.17941	0.14014	0.06486	15	5	12
2.16465	0.28983	0.05143	17	5	12
2.18426	0.43575	0.05255	19	5	12

7. Application - Coefficient Function Estimates

## 7 DEFINITIONS AND PROOFS

The following proofs are adapted from Alexanderian [9].

### 1. Definition (Hilbert-Schmidt Integral Operator)

Given a bounded domain  $A \subset \mathbb{R}^n$ , we call a function  $c$ :

$A \times A \rightarrow \mathbb{R}$  a Hilbert-Schmidt kernel if

### 2. Lemma (Three Traits of Scores)

The random function  $X(t)$  realizing in  $L^2[0, 1]$  is expanded by its Eigenfunctions  $\{\nu^m\}$  as shown in Equation ???. The coefficients  $\xi^m$  corresponding to Eigenfunctions  $\nu^m$  satisfy the following properties:

- 1)  $E[\xi^m(\omega)] = 0$   $\delta_{m,n} \lambda^m$
- 2)  $Cov(\xi^m(\omega), \xi^n(\omega)) =$  3)  $Var(\xi^m(\omega)) = \lambda^m$

where  $\delta^{m,n} = 0$  if  $m \neq n$ , otherwise 1.

*Proof.* Assume that  $F(t)$  is the centered process of  $X(t)$ , namely,  $F(t) = X(t) - \int_{\Omega} X(t) dP(\omega)$ . To obtain the first result, we can show that

$$\begin{aligned}
 E[\xi^m] &= E\left[\int_0^1 F(t)\nu_j(t)dt\right] \\
 &= \int_{\Omega} \int_0^1 F(t)\nu^m(t)dt dP(\omega) \\
 &= \int_0^1 \int_{\Omega} F(t)\nu^m(t) dP(\omega) dt && \text{(Fubini)} && (78) \\
 &= \int_0^1 \int_{\Omega} F(t) dP(\omega) \nu^m(t) dt \\
 &= \int_0^1 E[F(t)] \nu^m(t) dt = 0
 \end{aligned}$$

where  $E[F(t)]$  is 0 since  $F(t)$  is a centered process.

**TABLE IV.** Monomial FPCR, nharm = 2

$f1, Y1$	$f1, Y2$	$f2, Y1$	$f2, Y2$	n basis
7.17723	131.92717	0.89367	2.58395	2
5.89626	130.64918	0.81105	2.50111	3
<b>5.807</b>	<b>130.55792</b>	0.77154	2.46164	4
6.07681	130.82169	0.77836	2.4684	5
6.55003	131.28414	0.77506	2.465	6
7.55111	132.26672	0.73771	2.42761	7
11.62846	136.28281	0.74403	2.43402	8
17.29836	141.88633	0.69837	2.38887	9
18.84999	143.43309	<b>0.64904</b>	<b>2.33978</b>	10
18.88329	143.46571	0.69403	2.38461	11

18.99159    143.63082    0.81692  
 2.50713        12

Table V: Monomial FPCR,  $nharm = 3$

$f_1, Y_1$	$f_1, Y_2$	$f_2, Y_1$	$f_2, Y_2$	n_basis
4.17039	129.64259	0.81224	2.51177	3
4.08093	129.55048	0.77174	2.47128	4
<b>4.0038</b>	<b>129.49674</b>	0.75004	2.44986	5
4.17563	129.68591	<b>0.45954</b>	<b>2.1605</b>	6
4.23226	129.74146	0.52512	2.226	7
4.4198	129.91964	0.74368	2.4439	8
4.42156	129.92039	0.69479	2.39518	9
4.4125	129.91256	0.64087	2.34145	10
4.44546	129.94476	0.6792	2.37965	11
11.82311	137.32413	0.74329	2.44296	12

Table VI: Monomial FPCR,  $nharm = 4$

$f_1, Y_1$	$f_1, Y_2$	$f_2, Y_1$	$f_2, Y_2$	n_basis
3.91274	130.1171	0.38077	2.09045	4
3.95242	130.16289	0.73944	2.44922	5
<b>3.8366</b>	<b>130.05179</b>	0.27201	1.98372	6
3.96405	130.17512	0.30729	2.01904	7
4.38618	130.61211	<b>0.11108</b>	<b>1.82326</b>	8
4.44028	130.66082	0.13501	1.84706	9
4.4251	130.64453	0.1455	1.85748	10
4.44727	130.66669	0.14012	1.85223	11
7.88834	134.08085	0.17742	1.8879	12

Table VII: B-Spline FPCR,  $nharm = 2$

$f_1, Y_1$	$f_1, Y_2$	$f_2, Y_1$	$f_2, Y_2$	n_basis
10.93556	135.66335	<b>0.69614</b>	<b>2.38597</b>	4
<b>6.01366</b>	<b>130.75986</b>	0.77796	2.46801	5
6.36762	131.10562	0.79035	2.48029	6
6.76471	131.49507	0.7684	2.45834	7
7.21393	131.93622	0.7958	2.48568	8
7.77885	132.48984	0.75447	2.44437	9
8.48517	133.18046	0.71664	2.40654	10
8.95142	133.63714	0.70481	2.39475	11
9.21314	133.89421	0.71456	2.4045	12
9.29854	133.97794	0.73961	2.42952	13
9.31792	133.99732	0.74046	2.43037	14
9.32858	134.00809	0.74354	2.43344	15
9.27722	133.95811	0.77399	2.46383	16
9.23813	133.91976	0.77175	2.46159	17
9.33983	134.01705	0.76653	2.45632	18
9.24067	133.92097	0.78389	2.47367	19
9.34246	134.01837	0.75304	2.44285	20
9.38735	134.06288	0.76148	2.45127	21
9.35218	134.02839	0.75842	2.44823	22
9.50715	134.17875	0.74067	2.43049	23
9.51344	134.18666	0.7552	2.44502	24
9.52277	134.19463	0.75125	2.44107	25

The second claim is proved as

$$\begin{aligned}
 \mathbb{E}[\xi^m \xi^n] &= \mathbb{E}\left[\int_0^1 F(s)\nu^m(s)ds \int_0^1 F(t)\nu^n(t)dt\right] \\
 &= \mathbb{E}\left[\int_0^1 \int_0^1 F(s)\nu^m(s)F(t)\nu^n(t)dsdt\right] \text{(Fubini)} \quad (79) \\
 &= \int_0^1 \int_0^1 \mathbb{E}[F(s)F(t)]\nu^m(s)\nu^n(t)dsdt \\
 &= \int_0^1 \left(\int_0^1 c(s,t)\nu^m(s)ds\right)\nu^n(t)dt \\
 &= \int_0^1 [K\nu^m](t)\nu^n(t)dt \\
 &= \langle K\nu^m, \nu^n \rangle \\
 &= \langle \lambda^m \nu^m, \nu^n \rangle = \delta^{m,n} \lambda^m
 \end{aligned}$$

Where  $\delta^{m,n} = 1$  if  $m = n$ , otherwise 0. The result is

Table VIII: B-Spline FPCR,  $nharm = 3$

$f_1, Y_1$	$f_1, Y_2$	$f_2, Y_1$	$f_2, Y_2$	n_basis
4.26915	129.73814	0.65618	2.35517	4
3.99153	129.47951	0.75902	2.45877	5
4.13265	129.64227	0.42504	2.126	6
4.15027	129.66158	0.52603	2.22684	7
4.28266	129.7862	0.71107	2.41134	8
4.30464	129.8084	0.65968	2.36019	9
4.32579	129.82882	0.60062	2.3013	10
4.29797	129.80089	0.65658	2.35703	11
4.32782	129.82922	0.68072	2.38108	12
4.34042	129.84041	0.71583	2.41605	13
4.3526	129.8512	0.72262	2.42279	14
4.3577	129.85642	0.72621	2.42637	15
4.37694	129.87493	0.75746	2.45749	16
4.35973	129.85893	0.75291	2.45296	17
4.37223	129.87141	0.69875	2.39878	18
4.34433	129.84395	0.74346	2.44353	19
4.36545	129.86595	0.67405	2.37412	20
4.36942	129.86886	0.67941	2.37944	21
4.34272	129.843	0.6919	2.39205	22
4.38437	129.88403	0.64702	2.34708	23
4.35012	129.8495	0.68435	2.38446	24
4.36866	129.86795	0.67512	2.37514	25

Table IX: B-Spline FPCR,  $nharm = 4$

$f_1, Y_1$	$f_1, Y_2$	$f_2, Y_1$	$f_2, Y_2$	n_basis
3.91274	130.1171	0.38077	2.09045	4
3.97061	130.18108	0.75664	2.46627	5
3.83646	130.05682	0.19244	1.90433	6
3.8814	130.0934	0.3326	2.04419	7
4.2679	130.49905	0.09497	1.8071	8
4.31978	130.54252	0.15828	1.87039	9
4.32413	130.5527	0.10538	1.81745	10
4.31263	130.53896	0.13116	1.84322	11
4.33997	130.56608	0.10762	1.81971	12
4.34107	130.56642	0.11868	1.83076	13
4.35627	130.57893	0.12832	1.84037	14
4.35369	130.57747	0.11173	1.82379	15
4.35313	130.57695	0.11652	1.82861	16
4.3476	130.57119	0.11283	1.82497	17
4.31541	130.54208	0.08832	1.80044	18
4.32121	130.54666	0.10087	1.81304	19
4.2968	130.52361	0.08341	1.79552	20
4.30044	130.52787	0.08785	1.79994	21
4.29859	130.52514	0.08833	1.80045	22
4.28391	130.51071	0.08189	1.79394	23
4.29418	130.52054	0.08928	1.80134	24
4.28911	130.51512	0.0856	1.79766	25

Produced from orthonormality of the Eigenfunctions.

$$Cov(\xi^m, \xi^n) = E[\xi^m \xi^n] - E[\xi^m]E[\xi^n] = \delta^{m,n} \lambda^m \quad (80)$$

Where  $E[\xi^m] = E[\xi^n] = 0$  as shown in the first property. The

Table X: Fourier FPCR,  $nharm = 2$

$f_1, Y_1$	$f_1, Y_2$	$f_2, Y_1$	$f_2, Y_2$	n_basis
5.04859	129.81003	0.78889	2.4789	3
5.08647	129.84459	0.69756	2.38778	5
5.29235	130.04946	0.80393	2.49395	7
5.32414	130.07859	0.80074	2.49072	9
5.34403	130.09777	0.79714	2.48713	11
5.43601	130.18816	0.8141	2.50404	13
5.51333	130.26036	0.78908	2.47902	15
5.70153	130.44275	0.79659	2.48646	17
5.87259	130.6101	0.77783	2.46771	19

Table XI: Fourier FPCR,  $nharm = 3$

$f_1, Y_1$	$f_1, Y_2$	$f_2, Y_1$	$f_2, Y_2$	n_basis
3.69752	129.19134	0.69524	2.3944	3
5.06895	130.54881	0.1376	1.83947	5
5.21758	130.69839	0.13994	1.84156	7
5.22636	130.70697	0.12715	1.82896	9
5.27531	130.74943	0.14282	1.84453	11
5.29036	130.76091	0.13007	1.83175	13
5.16722	130.64051	0.10365	1.80524	15
4.98841	130.46747	0.11563	1.81733	17
4.92246	130.40322	0.14766	1.84937	19

Last assertion is confirmed from the two properties above.

$$Var[\xi^m] = E[(\xi^m - E[\xi^m])^2] = E[(\xi^m)^2] = \lambda^m \quad (81)$$

The original process  $X(t)$  also has the same properties as the centered one since

$$X(t) = E[X(t)] + F(t) = \mu(t) + \sum_{m=1}^{\infty} \xi^m \nu^m(t) \quad (82)$$

### C. Theorem (Karhunen-Loeve Expansion)

Let  $X: [0, 1] \rightarrow \mathbb{R}$  be a mean-square continuous stochastic process, meaning

$$\lim_{\epsilon \rightarrow 0} E[(X(t + \epsilon) - X(t))^2] = 0 \quad (83)$$

Such that  $X \in L^2 [0, 1]$ . Then there exists a basis  $\nu^m$  of  $L^2 [0, 1]$  such that for all  $t \in [0, 1]$  we have the following representation

$$X(t) = \mu(t) + \sum_{m=1}^{\infty} \xi^m \nu^m(t), \quad (84)$$

Here,  $\mu(t)$  is the mean function of  $X(t)$  and coefficients  $\xi^m$  are given by  $\int_0^1 (X(t) - \mu(t)) \nu^m(t) dt$ . These coefficients satisfy the following conditions.

Table XII: Fourier FPCR,  $nharm = 4$

$f_1, Y_1$	$f_1, Y_2$	$f_2, Y_1$	$f_2, Y_2$	$n\_basis$
3.61833	129.83284	0.07736	1.78876	5
3.66209	129.88835	0.06311	1.77442	7
3.68812	129.91351	0.0717	1.78321	9
4.10715	130.34635	0.08317	1.79518	11
4.17569	130.4147	0.08837	1.80046	13
4.21112	130.44645	0.09965	1.81164	15
4.2129	130.45091	0.08775	1.79975	17
4.21884	130.45501	0.09032	1.80229	19

Table XIII: Additional Runs B-Splines for  $L \in \{50, 70\}$

$n\_FPC$	$f_1, Y_1$	$f_1, Y_2$	$f_2, Y_1$	$f_2, Y_2$	$n\_basis$
2	12.01215	136.37917	0.74396	2.45234	50
2	12.63307	136.79449	0.73227	2.43754	70
3	4.98379	130.27154	0.62033	2.33437	50
3	6.70572	131.58072	0.69647	2.40584	70
4	4.33014	130.23059	0.08985	1.81309	50
4	5.81452	131.05453	0.26875	1.9835	70
5	4.07164	130.68035	0.06323	1.79676	50
5	5.083	131.08668	0.22106	1.94707	70
6	3.70117	131.04066	0.06245	1.80571	50
6	5.62771	132.62722	0.23593	1.97606	70
7	3.69911	131.80298	0.05566	1.8068	50
7	5.37466	133.80644	0.27455	2.03417	70

- 1)  $\mathbb{E}[\xi^m(\omega)] = 0$   $\delta^{m,n} \lambda^m$
- 2)  $Cov(\xi^m(\omega), \xi^n(\omega)) = 0$  3)  $Var(\xi^m(\omega)) = \lambda^m$

*Proof.* Let  $K$  be a Hilbert-Schmidt operator as in Equation 20. We know that  $K$  has a complete set of Eigenfunctions  $\nu^m$  in  $L^2[0, 1]$  and non-negative Eigenvalues  $\lambda^m$  since  $K$  is a positive compact self-adjoint operator. With the reminder that  $\xi^m$  satisfy the three conclusions by Lemma VII-B, we prove this expansion by considering where  $F(t)$  is the centered process of  $X(t)$ . Once it is shown that  $\lim_{N \rightarrow \infty} \epsilon_N(t) = 0$  uniformly in  $[0, 1]$ , the proof is completed.

$$\begin{aligned}
 \epsilon_N(t) &= \mathbb{E} \left[ \left( F(t) - \sum_{m=1}^N \xi^m \nu^m(t) \right)^2 \right] \\
 &= \mathbb{E}[F(t)^2] - 2\mathbb{E} \left[ F(t) \sum_{m=1}^N \xi^m \nu^m(t) \right] + \mathbb{E} \left[ \sum_{m=1}^N \sum_{n=1}^N \xi^m \xi^n \nu^m(t) \nu^n(t) \right]
 \end{aligned}
 \tag{85}$$

Here,  $\mathbb{E}[F(t)^2] = c(t, t)$  as in Equation 19 since  $F(t)$  is the centered process. Now, take the second term

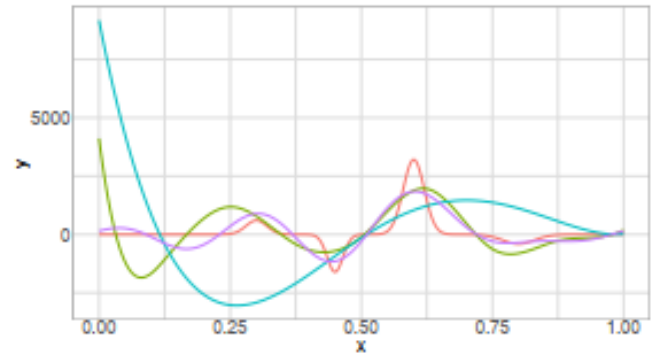


Figure 9: Basis Expansion Regression -  $f_2, Y_1$

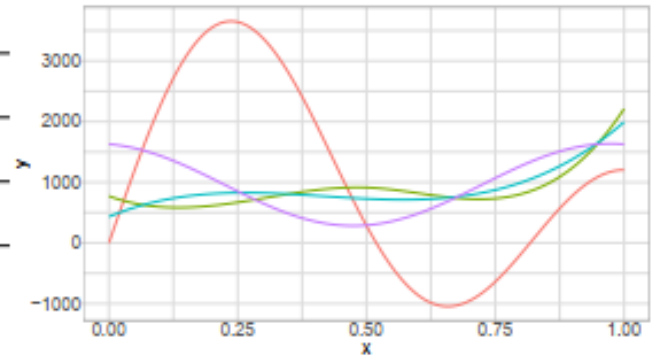
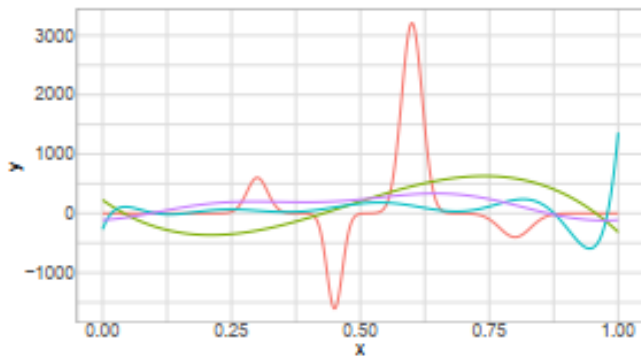


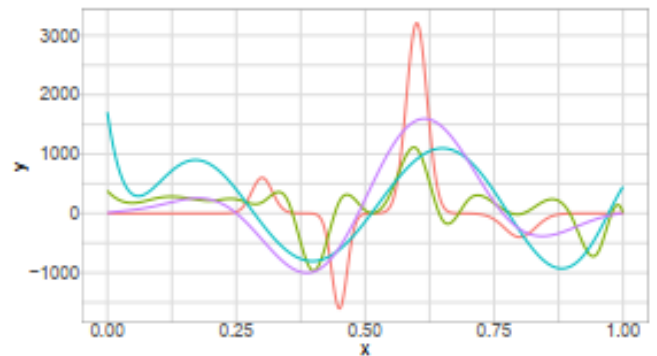
Figure 10: FPC Regression, 2 harmonics -  $f_1, Y_1$

$$\begin{aligned}
 \mathbb{E} \left[ F(t) \sum_{m=1}^N \xi^m \nu^m(t) \right] &= \mathbb{E} \left[ F(t) \sum_{m=1}^N \left( \int_0^1 F(s) \nu^m(s) ds \right) \nu^m(t) \right] \\
 &= \mathbb{E} \left[ \sum_{m=1}^N \left( \int_0^1 F(t) F(s) \nu^m(s) ds \right) \nu^m(t) \right] \\
 &= \sum_{m=1}^N \left( \int_0^1 \mathbb{E}[F(t) F(s)] \nu^m(s) ds \right) \nu^m(t) \\
 &= \sum_{m=1}^N \left( \int_0^1 c(t, s) \nu^m(s) ds \right) \nu^m(t) \\
 &= \sum_{m=1}^N [K \nu^m](t) \nu^m(t) \\
 &= \sum_{m=1}^N \lambda^m \nu^m(t) \nu^m(t) = \sum_{m=1}^N \lambda^m \nu^m(t)^2
 \end{aligned}
 \tag{86}$$

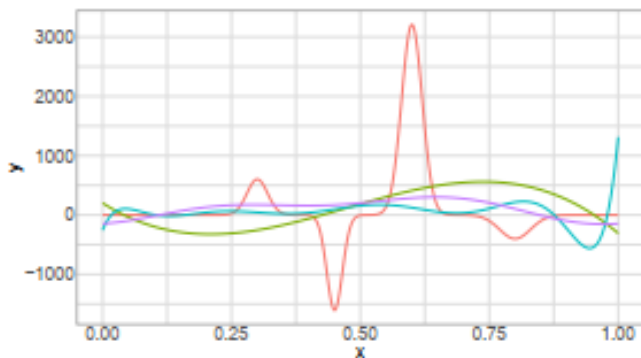
Where the covariance function  $c(t, s)$  has the Hilbert-Schmidt operator as in Equation 20. For the last term, we derive from Equation 79 that



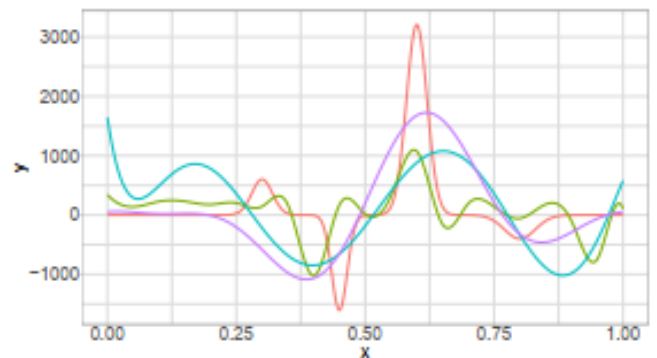
Curve ■ f2 ■ bspline\_nharm2 ■ monomial\_nharm2 ■ fourier\_nharm  
 Figure 11: FPC Regression, 2 harmonics -  $f_2, Y_1$



Curve ■ f2 ■ bspline\_nharm4 ■ monomial\_nharm4 ■ fourier\_nharm  
 Figure 13: FPC Regression, 4 harmonics -  $f_2, Y_1$



Curve ■ f2 ■ bspline\_nharm2 ■ monomial\_nharm2 ■ fourier\_nharm  
 Figure 12: FPC Regression, 2 harmonics -  $f_2, Y_2$



Curve ■ f2 ■ bspline\_nharm4 ■ monomial\_nharm4 ■ fourier\_nharm  
 Figure 14: FPC Regression, 4 harmonics -  $f_2, Y_2$

$$\begin{aligned}
 \mathbb{E} \left[ \sum_{m=1}^N \sum_{n=1}^N \xi^m \xi^n \nu^m(t) \nu^n(t) \right] &= \sum_{m=1}^N \sum_{n=1}^N \mathbb{E}[\xi^m \xi^n] \nu^m(t) \nu^n(t) \\
 &= \sum_{m=1}^N \sum_{n=1}^N \delta^{m,n} \lambda^m \nu^m(t) \nu^n(t) \\
 &= \sum_{m=1}^N \lambda^m \nu^m(t)^2
 \end{aligned}
 \tag{87}$$

where  $\delta_{m,n} = 1$  if  $m = n$ , otherwise 0.

Therefore, by Equations 85, 86, and 87 we obtain

$$\epsilon_N(t) = c(t, t) - \sum_{m=1}^N \lambda^m \nu^m(t) \nu^m(t)
 \tag{88}$$

Implementing Mercer’s Theorem this proof is concluded by

$$\lim_{N \rightarrow \infty} \epsilon_N(t) = \lim_{n \rightarrow \infty} \mathbb{E} \left[ \left( F(t) - \sum_{m=1}^N \xi^m \nu^m(t) \right)^2 \right] = 0
 \tag{89}$$

## REFERENCES

- [1]. D. J. Levitin, R. L. Nuzzo, B. W. Vines, and J. O. Ramsay, “Introduction to functional data analysis.,” *Canadian Psychology/Psychologie canadienne*, vol. 48, no. 3, pp. 135–155, 2007, doi: <https://doi.org/10.1037/cp2007014>.
- [2]. P. Kokoszka and M. Reimherr, *Introduction to functional data analysis*. Boca Raton, FL: Crc Press, 2017.
- [3]. H. Bauer, *Wahrscheinlichkeitstheorie und Grundzüge der Maßtheorie*. Walter de Gruyter GmbH & Co KG, 2020.
- [4]. T. Hsing and R. Eubank, *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. John Wiley & Sons, 2015.
- [5]. J. O. Ramsay and B. W. Silverman, *Functional Data Analysis*. Springer Science+Business Media, 2005. doi: <https://doi.org/10.1007/b98888>.
- [6]. Carl de Boor, “A Practical Guide to Splines.” *Mathematics of Computation*, vol. 34, no. 149, pp. 325–325, Jan. 1980, doi: <https://doi.org/10.2307/2006241>.
- [7]. J. Goldsmith, J. Bobb, C. M. Crainiceanu, B. Caffo, and D. Reich, “Penalized Functional Regression,” *Journal of Computational and Graphical Statistics*, vol. 20, no. 4, pp. 830–851, Jan. 2011, doi: <https://doi.org/10.1198/jcgs.2010.10007>.
- [8]. T. C. M. Lee, “Smoothing parameter selection for smoothing splines: a simulation study,” *Computational Statistics & Data Analysis*, vol. 42, no. 1–2, pp. 139–148, Oct. 2002, doi: [https://doi.org/10.1016/S0167-9473\(02\)00159-7](https://doi.org/10.1016/S0167-9473(02)00159-7).
- [9]. A. Alexanderian, “A brief note on the Karhunen-Loève expansion,” *arXiv.org*, 2015. <https://arxiv.org/abs/1509.07526>
- [10]. I. T. Jolliffe, “A Note on the Use of Principal Components in Regression,” *Applied Statistics*, vol. 31, no. 3, p. 300, 1982, doi: <https://doi.org/10.2307/2348005>.
- [11]. P. T. Reiss and R. T. Ogden, “Functional Principal Component Regression and Functional Partial Least Squares,” *Journal of the American Statistical Association*, vol. 102, no. 479, pp. 984–996, Sep. 2007, doi: <https://doi.org/10.1198/016214507000000527>.
- [12]. H. Cardot, “Spatially Adaptive Splines for Statistical Linear Inverse Problems,” *Journal of Multivariate Analysis*, vol. 81, no. 1, pp. 100–119, Apr. 2002, doi: <https://doi.org/10.1006/jmva.2001.1994>.
- [13]. G. Dattoli, Paolo Emilio Ricci, and C. Cesarano, “A Note on Legendre Polynomials,” *International Journal of Nonlinear Sciences and Numerical Simulation*, vol. 2, no. 4, Jan. 2001, doi: <https://doi.org/10.1515/ijnsns.2001.2.4.365>.
- [14]. Y. Li *et al.*, “Early Diagnosis of Type 2 Diabetes Based on Near-Infrared Spectroscopy Combined With Machine Learning and Aquaphotomics,” *Frontiers in Chemistry*, vol. 8, Dec. 2020, doi: <https://doi.org/10.3389/fchem.2020.580489>.
- [15]. Gy. Bohács, Z. Ovádi, and A. Salgó, “Prediction of Gasoline Properties with near Infrared Spectroscopy,” *Journal of Near Infrared Spectroscopy*, vol. 6, no. 1, pp. 341–348, Jan. 1998, doi: <https://doi.org/10.1255/jnirs.155>.