Scalar on Function Regression with Applications to Near-Infrared Spectography

Jonathan Willnow, Jakob Juergens, Jonghun Baek

Presentation Day

Introduction

- **Near-Infrared** (NIR) **Spectroscopy** enables fast diagnostics by using the NIR region of the electromagnetic spectrum (from 780 nm to 2500 nm)
- Suited for field-monitoring / on-line analysis
- Spectroscopy results in high-dimensional dataset.
- This set of measurements serves as set of discretized approximations of smooth spectral curves
- Regression to determine relationship between octane rating and spectral curves

Theory

A simple functional dataset is given by

$$\{x_i(t_{j,i}) \in \mathbb{R} \mid i = 1, 2, ..., N,\ j = 1, 2, ..., J_i,\ t_{j,i} \in [T_1, T_2]\}$$

- Continuous underlying process, where $x_i(t)$ exists $\forall t \in [T_1, T_2]$
- Only observed at $x_i(t_{j,i})$
- Example: Gasoline dataset (60 x 400)
- Other Examples: Growth curves, financial data, human perception (pitch), ...

Random Function

A **Random Variable** is a function $X : \Omega \to \mathcal{S}$ which is defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where Ω is a probability space with a σ-algebra $\mathcal{F}$ and a probability measure $\mathbb{P}$.

- If $\mathcal{S} = \mathbb{R}$ then X is a random variable
- If $\mathcal{S} = \mathbb{R}^n$ then X is a random vector
- If $\mathcal{S}$ is a space of functions, X is called a **Random Function**

Random Function

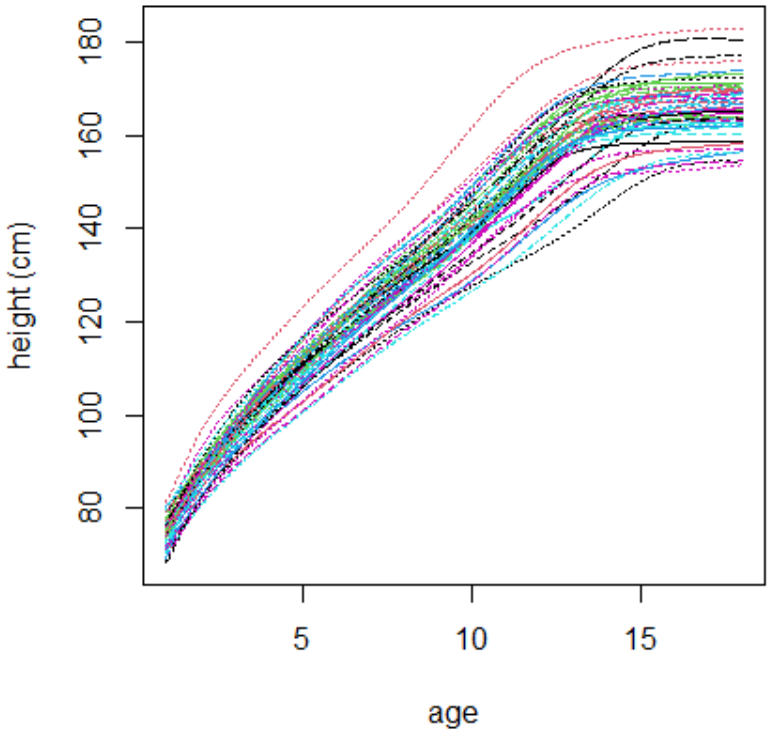
Let $\mathbb{E}$ be the index set and this can be described as

$$X = \{X(t, \omega) : t \in \mathbb{E}, \omega \in \Omega\},$$

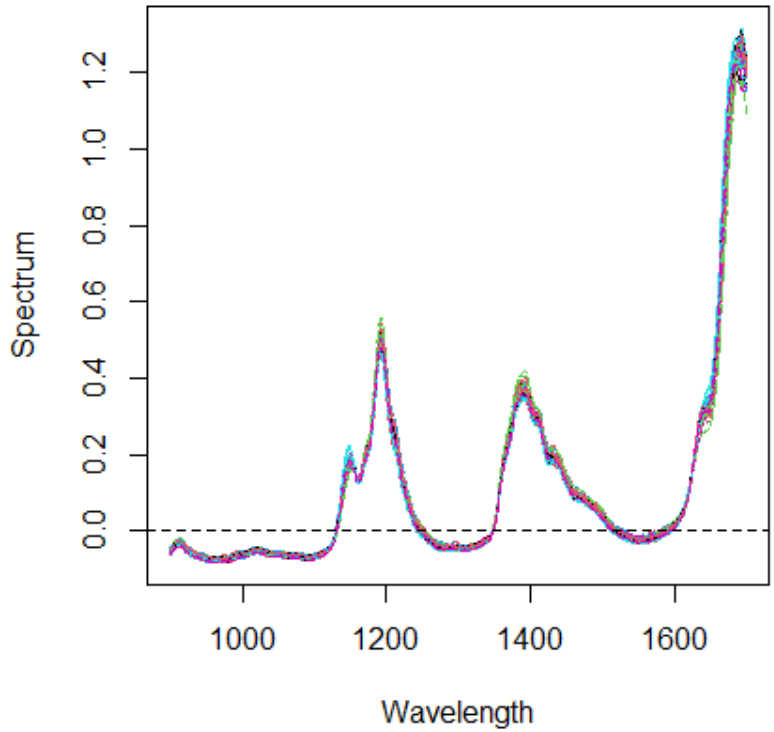
where $X(t, \cdot)$ is $\mathcal{F}$-measurable function on the sample space Ω.

- It can be shortened to $X(t)$ by omitting ω
- The function is realized when the $X(t)$ exists $\forall t \in \mathbb{E}$

Plots



Growth curves of
54 girls age 1-18



NIR spectrum of 60
gasoline samples

Square Integrable Function

If a function $f(t)$ satisfies

$$\int_0^1 (f(t))^2 \, dt < \infty$$

the function $f(t)$ is called **Square Integrable Function** written $f(t) \in \mathbb{L}^2[0,1]$.

- Without loss of generality, the interval is defined in $[0,1]$.
- $\mathbb{L}^2$ is the set of all square integrable functions.

Square Integrable Function

Let $f, g \in \mathbb{L}^2[0,1]$, then we can define inner product by

$$\langle f, g \rangle = \int_0^1 f(t)g(t)dt$$

- Orthogonality of two different functions with $\langle f, g \rangle = 0$
- Distance between functions

Basis Expansion

Basis Expansion is a linear combination of functions as described:

$$X_i(t) = \sum_{k=1}^{\infty} c_{ik}\phi_k(t) \approx \sum_{k=1}^{K} c_{ik}\phi_k(t), \quad i = 1, \ldots, n, \quad \forall t \in \mathbb{E}$$

where $\phi_k(t)$ is the k^{th} basis function of the expansion and c_{ik} is the corresponding coefficient. We truncate the basis at K to:

- make the function smoother
- replace the original curves $X_i(t)$ by a smaller collection of c_{nm}

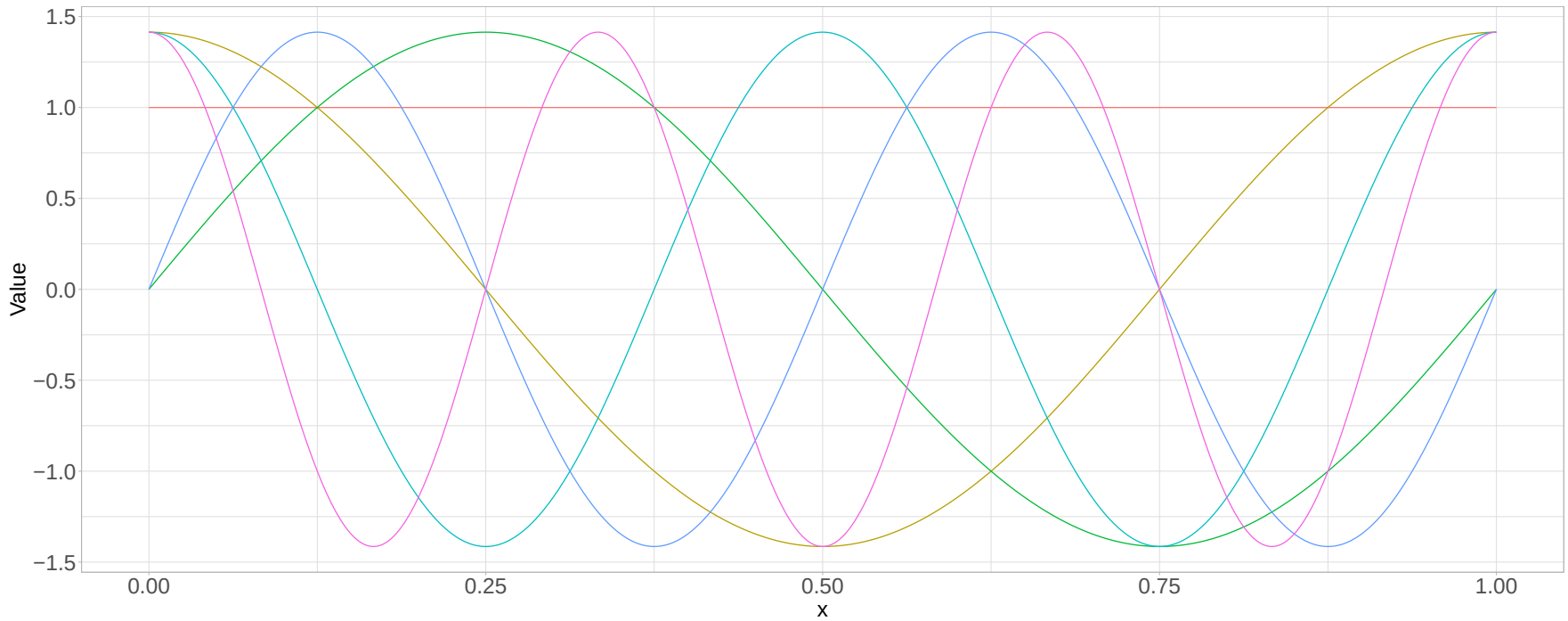
Two Typical Types of Basis Function

Fourier Basis Function is an element of the set:

$$\{\sqrt{2}sin(2\pi nx|n \in \mathbb{N})\} \cup \{\sqrt{2}cos(2\pi nx|n \in \mathbb{N})\} \cup \{1\}$$

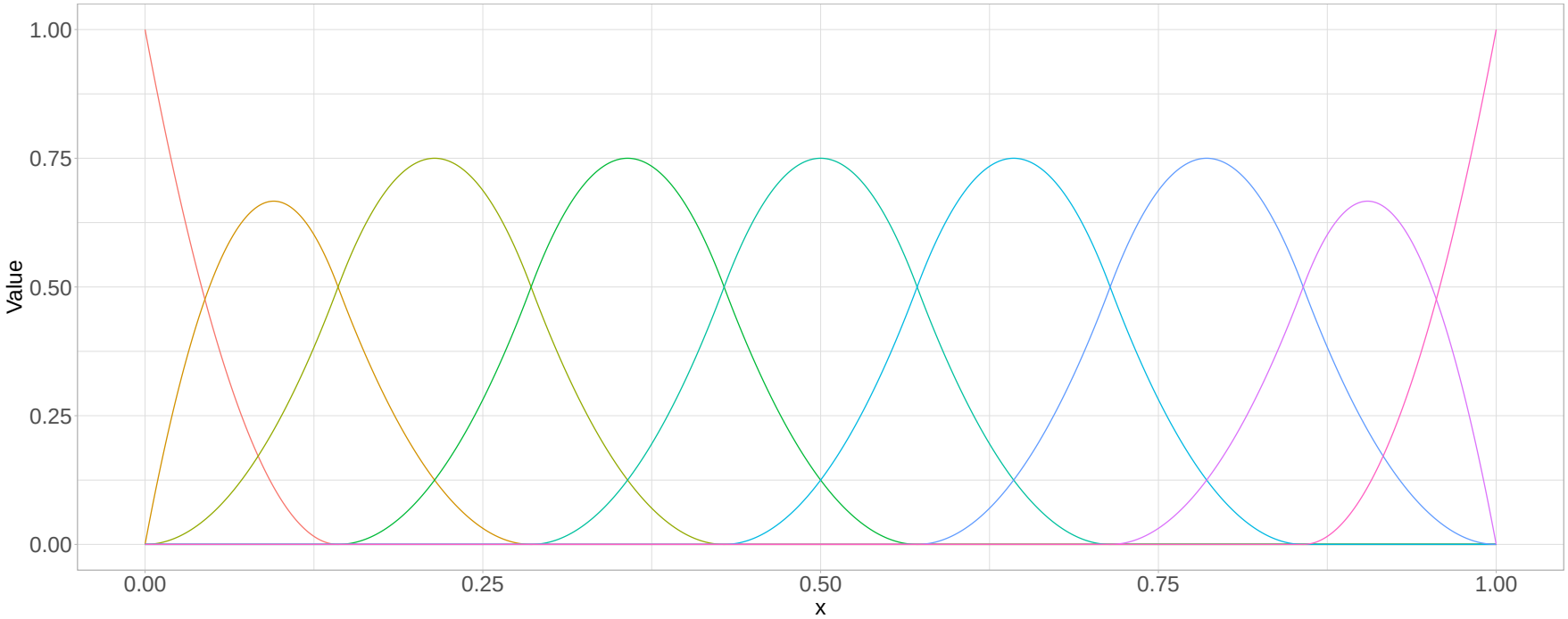
B-spline Basis Function is a polynomial function defined by order and knots.

Plots of Basis Functions



Fourier basis functions

Plots of Basis Functions



Bspline basis functions

Trade Off between Bias and Variance

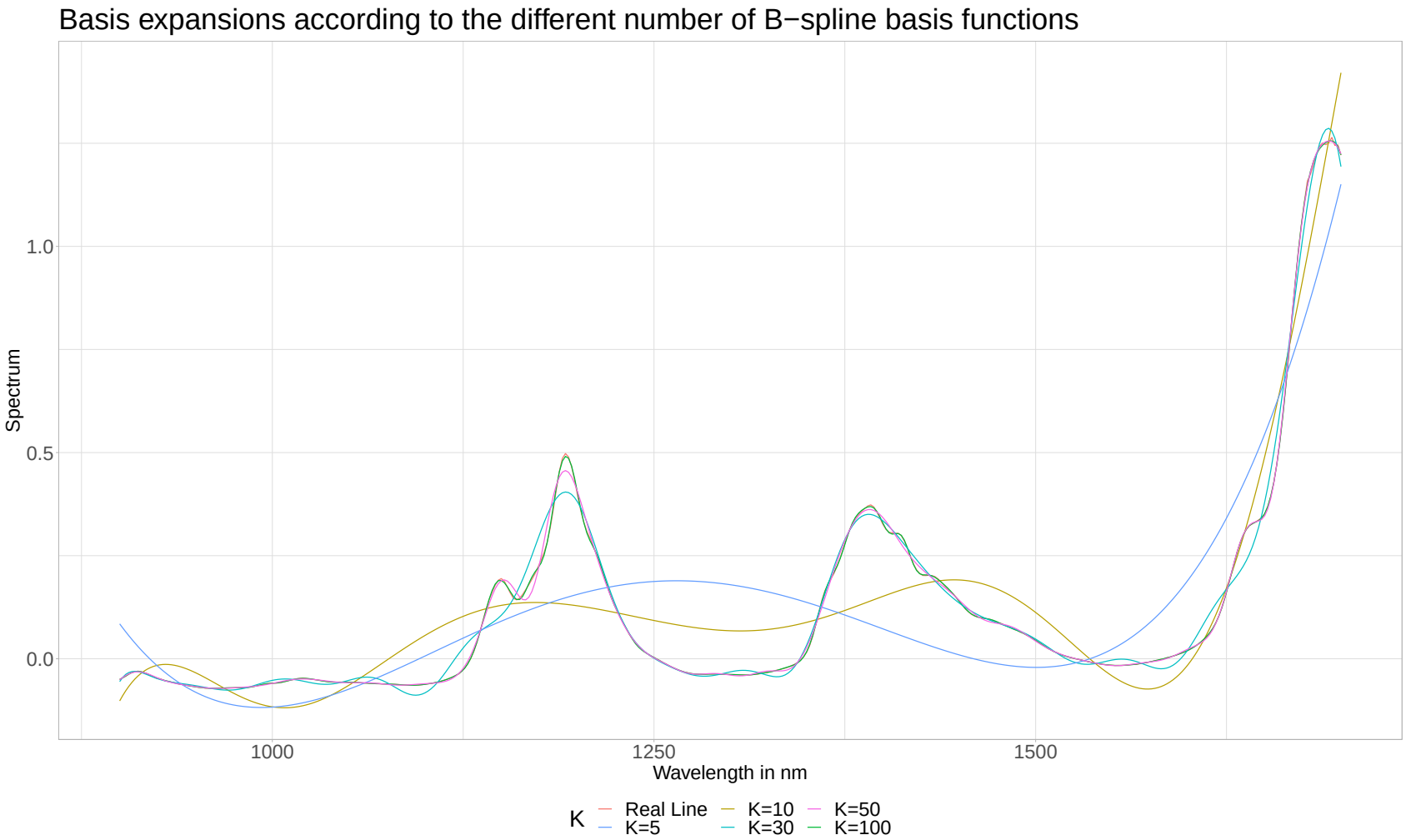
How do we choose the number K of basis functions?

$$\textbf{MSE}[\hat{X}(t)] = \textbf{Bias}^2[\hat{X}(t)] + \textbf{Var}[\hat{X}(t)]$$

$$\textbf{MISE}[\hat{X}] = \int_0^1 \textbf{MSE}[\hat{X}(t)]dt$$

- The larger K, the better fit to the data, but also more fitting noise
- If K is too small, the expansion would miss some significant information

Trade Off between Bias and Variance



Estimation via Basis Representation

Assume the following **Data Generating Process**

$$Y(\omega) = \alpha + \int_0^1 \beta(s)X(\omega)(s)\mathrm{d}s + \epsilon(\omega)$$

- $Y(\omega)$ and $\epsilon(\omega)$ realize in $\mathbb{R}$ and $X(\omega)$ realizes in $\mathbb{L}^2[0,1]$

Let $\{\phi_i(t) \mid i = 1, \ldots, \infty\}$ be a basis of $\mathbb{L}^2[0,1]$ leading to the following representation of $\beta(t)$

$$\beta(t) = \sum_{j=1}^{\infty} c_j \phi_j(t) \approx \sum_{j=1}^{L} c_j b_j(t)$$

Estimation via Basis Representation

We can transform the data generating process into:

$$
\begin{aligned}
Y(\omega) &= \alpha + \int_0^1 \left[\left(\sum_{j=1}^{\infty} c_j \phi_j(s) \right) X(\omega)(s) \right] \mathrm{d}s + \epsilon(\omega) \\
&= \alpha + \sum_{j=1}^{\infty} \left[c_j \int_0^1 X(\omega)(s) \phi_j(s) \mathrm{d}s \right] + \epsilon(\omega) \\
&= \alpha + \sum_{j=1}^{\infty} c_j Z_j(\omega) + \epsilon(\omega)
\end{aligned}
$$

Where a $Z_j(\omega)$ is a **scalar random variable**.

Estimation via Basis Representation

Each combination of $x_i(t)$ and $\phi_j(t)$ gives us

$$Z_{i,j} = \int_0^1 x_i(s) b_j(s) \mathrm{d}s$$

- This allows us to write each observation in the data set as $(y_i, Z_{i,1}, Z_{i,2}, \dots)$
- Truncating the functional basis yields an approximation $(y_i, Z_{i,1}, Z_{i,2}, \dots, Z_{i,L})$

Coefficients can then be estimated using theory from **multivariate regression** leading to an estimated coefficient vector $\hat{c} \in \mathbb{R}^L$.

Estimation via Basis Representation

This can be translated into an estimated coefficient function $\hat{\beta}(t)$:

$$\hat{\beta}_L(t) = \sum_{j=1}^{L} \hat{c}_{L,j}\phi_j(t)$$

This is dependent on...

- The basis $(\phi_j(t))_{j\in\mathcal{I}}$ for the estimation of $\beta(t)$
- The truncation parameter L
- The basis $(\psi_j(t))_{j\in\mathcal{I}}$ used for the expansion of the observations
- The truncation parameter in the approximation of the observations K

Karhunen-Loéve Expansion

Mean Function:

$$\mu(t) = \mathbb{E}\left[X(\omega)(t)\right]$$

Autocovariance Function:

$$c(t,s) = \mathbb{E}\Big[\left(X(\omega)(t) - \mu(t)\right)\left(X(\omega)(s) - \mu(s)\right)\Big]$$

The **Eigenvalues** and **Eigenfunctions**: $\{(\lambda_i, \nu_i) \mid i \in \mathcal{I}\}$ are solutions of the following equation:

$$\int_0^1 c(t,s)\nu(s)\mathrm{d}s = \lambda\nu(t)$$

Karhunen-Loéve Expansion

A random function $X(\omega)$ can be expressed in terms of its mean function and its Eigenfunctions:

$$X(\omega)(t) = \mu(t) + \sum_{j=1}^{\infty} \xi_j(\omega)\nu_j(t)$$

Where the ξ_j are scalar-valued random variables with the following properties.

1. $\mathbb{E}[\xi_i(\omega)] = 0$
2. $Var(\xi_i(\omega)) = \lambda_i$
3. $Cov(\xi_i(\omega), \xi_j(\omega)) = 0$ for $i \neq j$

This is called the **Karhunen-Loéve Expansion** of $X(\omega)$ and the Eigenfunctions can serve as a basis.

Functional Principal Component Analysis

Principal Component Analysis can be extended to functional regressors in the form of **Functional Principal Component Analysis** (FPCA).

Empirical Mean Function:

$$\hat{\mu}(t) = \frac{1}{n} \sum_{j=1}^{n} f_j(t)$$

Empirical Autocovariance Function:

$$\hat{c}(t,s) = \frac{1}{n} \sum_{j=1}^{n} \left(f_j(t) - \hat{\mu}(t)\right) \left(f_j(s) - \hat{\mu}(s)\right)$$

Functional Principal Component Analysis

The **Eigenvalues** and **Eigenfunctions**: $\{(\hat{\lambda}_i, \hat{\nu}_i) \mid i \in \mathcal{I}\}$ are solutions of the following equation:

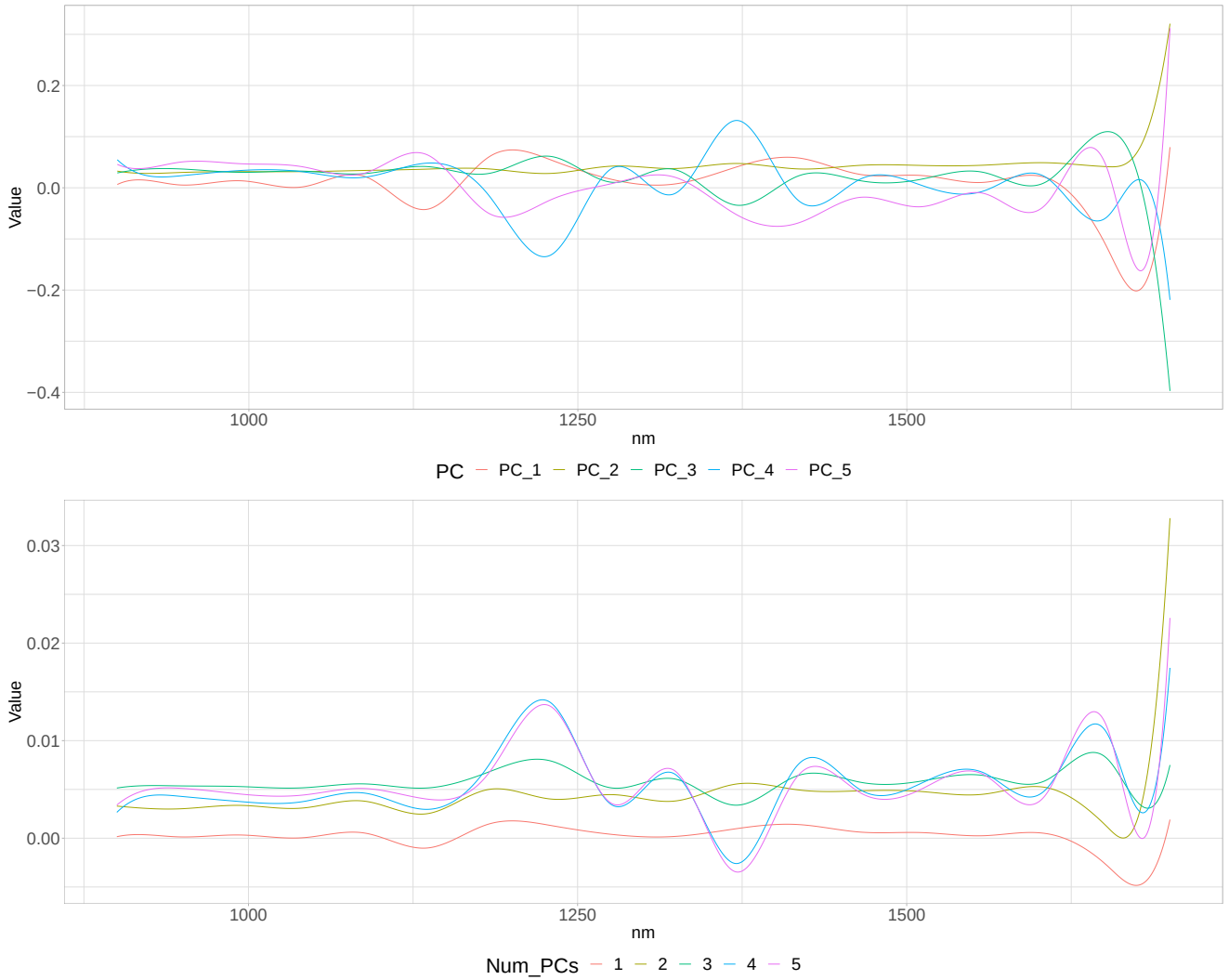
$$\int_0^1 \hat{c}(t,s)\hat{\nu}(s)\mathrm{d}s = \hat{\lambda}\hat{\nu}(t)$$

The $\{\hat{\nu}_i(s) \mid i \in \mathcal{I}\}$ are called **Functional Principal Components** and can serve as a basis for representing the original curves.

The corresponding scores $\hat{\xi}_i$ can be derived as

$$\hat{\xi}_j(\omega) = \int_0^1 (F(\omega)(s) - \hat{\mu}(s))\hat{\nu}_j(s)\mathrm{d}s$$

FPCA - Plots



The math is for intuition. In practice there are problems and the fpc's are derived differently.

Simulation Setup

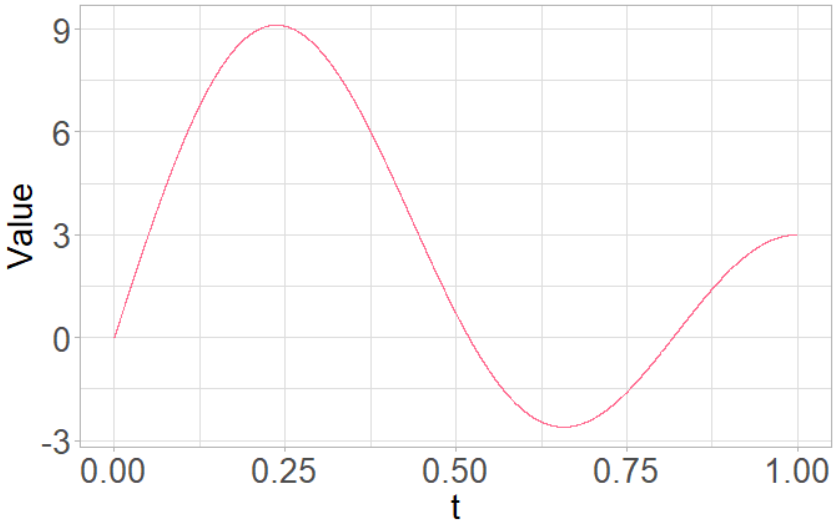
- Use the **Gasoline Dataset** (NIR-spectroscopy, 60×401) to predict octane ratings.
- Generate **similar curves** from gasoline dataset:

$$\tilde{X}(\omega)(t) = \hat{\mu}(t) + \sum_{j=1}^{J} \tilde{\xi}_j(\omega)\hat{\nu}_j(t)$$

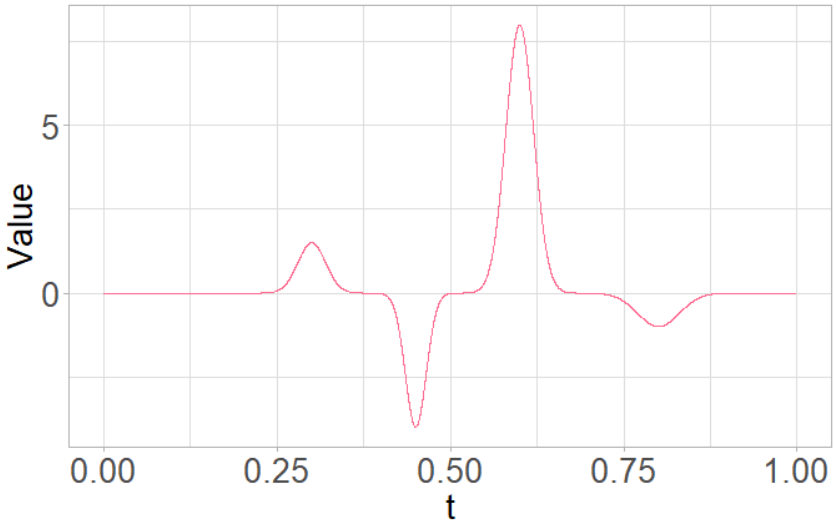
- $\tilde{\xi}_j \sim \mathcal{N}(0, \hat{\lambda}_j)$ and $\tilde{\xi}_j \perp\!\!\!\perp \tilde{\xi}_k$ for $j \neq k$
- Simplification: the ξ_j do not follow a normal
- $\tilde{X}(\omega)(t)$, $\hat{\mu}(t)$ and $\hat{\nu}_j(t)$ are approximated as vectors in $\mathbb{R}^{401}$.

Simulation Setup cont.

Following **Reiss and Ogden (2007)**, let $f_1(t)$ and $f_2(t)$ be two coefficient functions:



$f_1(t)$, smooth function



$f_2(t)$, bumpy function

Simulation Setup cont.

Let

$$Y_{1,f} = \langle NIR, f \rangle + Z \left(\frac{var(\langle NIR, f \rangle)}{0.9} - var(\langle NIR, f \rangle) \right)$$

$$Y_{2,f} = \langle NIR, f \rangle + Z \left(\frac{var(\langle NIR, f \rangle)}{0.6} - var(\langle NIR, f \rangle) \right)$$

where $Z \sim \mathcal{N}(0, 1)$ be two responses for $f \in \{f_1(t), f_2(t)\}$.

- Four combinations with different number of cubic basis-function $n_{basis} \in (4, 5, ..., 25)$ and fourier functions $(1, 3, ..., 25)$ to perform regression using basis expansion and the FPCR approach.
- Compare results via criteria (CV, Mallows CP,...)

Simulation Results

- add results here!

Application setup

- Use insights from the simulation study to uncover dependence.
- Similar setup, but using only bspline basis expansion and initial 60 spectral curves.
- Validation set approach: Scores of test data needs to be estimated by the training data.
- Report results by MSE scaled by variance.

Summary

- Concepts of functional data:
  - See dataset as smooth curve than as set of discrete measurements.
  - Theory of Random Functions, motivated from random variable.
  - Basis expansion and its Bias-Variance tradeoff.
  - Scalar on Function Regression via Basis Expansion.
  - Functional principal component Analysis and FPCR.
- Simulation study results: specification depends on function and signal-to-noise ratio
- Guided through application to predict octane ratings.

Thank you for your time!

Further Reading

Further Reading:

Gy. Bohács, Z. Ovádi, A. Salgó (1998). "Prediction of Gasoline Properties with near Infrared Spectroscopy". In: *Journal of near infrared spectroscopy.* 6, pp. 341–348.

Li, Yuanpeng et al. (2020). "Early Diagnosis of Type 2 Diabetes Based on Near-Infrared Spectroscopy Combined With Machine Learning and Aquaphotomics". In: *Frontiers in Chemistry* 8, p. 1133. ISSN: 2296-2646. DOI: 10.3389/fchem.2020.580489. URL: https://www.frontiersin.org/article/10.3389/fchem.2020.580489.

Reiss, Philip T. and R. Todd Ogden (2007). "Functional Principal Component Regression and Functional Partial Least Squares". In: *Journal of the American Statistical Association* 102.479, pp. 984–996. ISSN: 0162-1459. DOI: 10.1198/016214507000000527.

Spectral Representation of Random Vectors

Let $X(\omega)$ be a random vector realizing in $\mathbb{R}^p$.

- Let $\mu_x = \mathbb{E}(X)$ and $\Sigma_X = Cov(X)$
- Let $\{\gamma_i \mid i = 1, \ldots, p\}$ be the orthonormal **Eigenvectors** of Σ_X
- Let $\{\lambda_i \mid i = 1, \ldots, p\}$ be the corresponding **Eigenvalues** of Σ_X

Then X can also be represented as

$$X(\omega) = \mu_x + \sum_{i=1}^{p} \xi_i(\omega)\gamma_i$$

where the $\xi_i(\omega)$ have the following properties

1. $\mathbb{E}[\xi_i(\omega)] = 0$
2. $Var(\xi_i(\omega)) = \lambda_i$
3. $Cov(\xi_i(\omega), \xi_j(\omega)) = 0$ for $i \neq j$

Principal Component Analysis

A related concept is **Principal Component Analysis** (PCA).

Σ_X unknown $\rightarrow$ **sample analogues**

- Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ contain the standardized regressors
- Let $\hat{\Sigma}_X = \frac{\mathbf{X}'\mathbf{X}}{n}$
- Let $\{\hat{\gamma}_i \mid i = 1, \ldots, p\}$ be the orthonormal **Eigenvectors** of $\hat{\Sigma}_X$
- Let $\{\hat{\lambda}_i \mid i = 1, \ldots, p\}$ be the corresponding **Eigenvalues** of $\hat{\Sigma}_X$

Then $Z_i(\omega) = \hat{\gamma}_i' X(\omega)$ is called the i'th principal component and

1. $\mathbb{E}[Z_i(\omega)] = 0$
2. $Var(Z_i(\omega)) = \hat{\lambda}_i$
3. $Cov(Z_i(\omega), Z_j(\omega)) = 0$ for $i \neq j$