Scalar on Function Regression

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
- Suited for field-monitoring / on-line analysis and diagnostics of e.g: prediction of octane ratings!
- Spectroscopy results in high-dimensional dataset.
- This set of measurements along a continuum can be viewed as set 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 $x_i(t_{j,i})$
- Growth curves, financial data, human perception (pitch), ...
- To abstract information from the curves, they must be interpretable!

Theory

Jonghun

- Random Functions (name square integrable functions)
- Motivate continuous stochastic processes (growth curves/electricity consumption/yield curves/stonks)
- Use curves to predict a scalar response (show typical dgp)

Theory

Jonghun

- Basis expansions (b-splines and fourier)
- Talk about purposes
- Plots and show bias variance tradeoff

Random Function

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

Random Function

Let E be the index set and this can be described as

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

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 $X(t)$ have been observed for every $t \in E$

Square Integrable Function

If a function f satisfies:

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

the function f is called square integrable function
and in the set $\mathbb{L}^2[0,1]$.

- Without loss of generality, the interval is defined in $[0,1]$
- $\mathbb{L}^2$ is the set of all square integrable functions
- We focus on $\mathbb{L}^2[0,1]$ since the domain of our function is on the real line.

Square Integrable Function

If $f, g \in \mathbb{L}^2[0,1]$,

$$(ab + bg)(t) = af(t) + bg(t), \quad t \in [0,1] and \forall a, b$$

, where a and b are scalars. (Maybe don't need)
We can additionally define the inner product as follows:

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

- Orthogonolality of two different functions
- Distance between functions

Stochastic Process Perspective

Functional data are the sample curves observed from continuous time stochastic precess.

- From this perspective, $X(t)$ is a random variable (?)
- $X(\cdot)$ is a collection of random variables by each time index(?)
- Realizations of a random function belong in large collection of functions

Plots

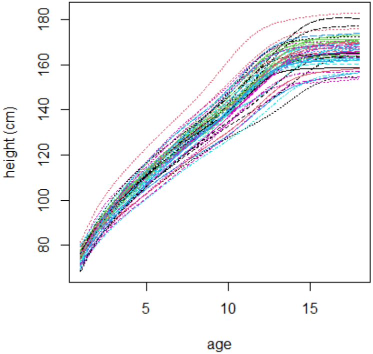


Figure: Growth curves of 54 girls age 1-18

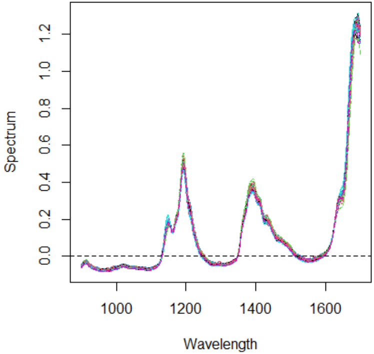


Figure: NIR spectrum of 60 gasoline samples

Data Generating Process

With $X_i(t_{ij}) \in \mathbb{L}^2[0,1]$, the measurement of X_{ij} is assumed to have noise with $\mathbb{E}(\epsilon)_{ij} = 0$ and **Var**(ϵ_{ij}^2), which are independent across i and j.
When the functional data has the same grid of domain, (SKIP)

Basis Expansion

Basis expansion is a linear combination of functions defining a function as described:

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

, where $\phi_k(t)$ is the k^{th} basis function of the expansion and c_{ik} is the corresponding coefficient.

- To make the function smoother
- To replace the original scalar data $X_n(t_{jn})$ by a smaller collection of c_{nm}

Two Typical Types of Basis Function

Fourier basis function is written as

$$f(x) = a_0 + \sum_{n=1}^{\infty} a_n cos(2\pi nx) + b_n sin(2\pi nx)$$

B-spline basis function is a flexible curve defined by degree and knots.
Each B-spline basis function, i-th B-spline basis function of degree p, $N_{i,p}(u)$ is defined on Cox-de Boor recursion formula.(Do you think I need to put the formula?)

Plots of Basis Functions

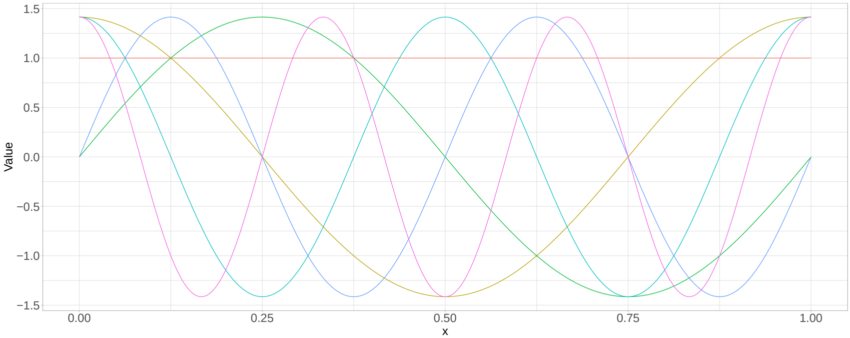


Figure: Fourier basis functions with order 9

Plots of Basis Functions

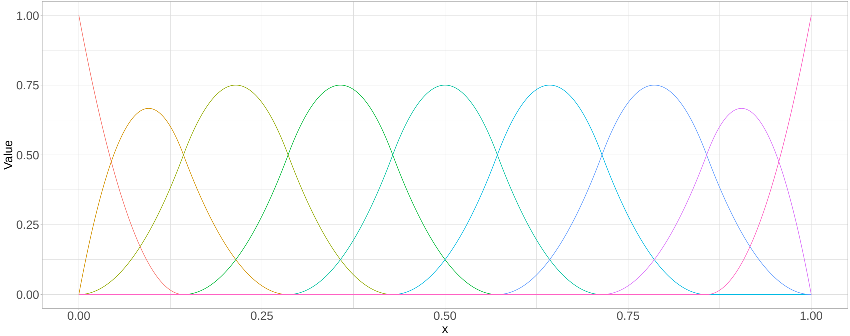


Figure: Bspline Basis with order 9

Trade Off between Bias and Variance

How do we choose the number K of basis functions?

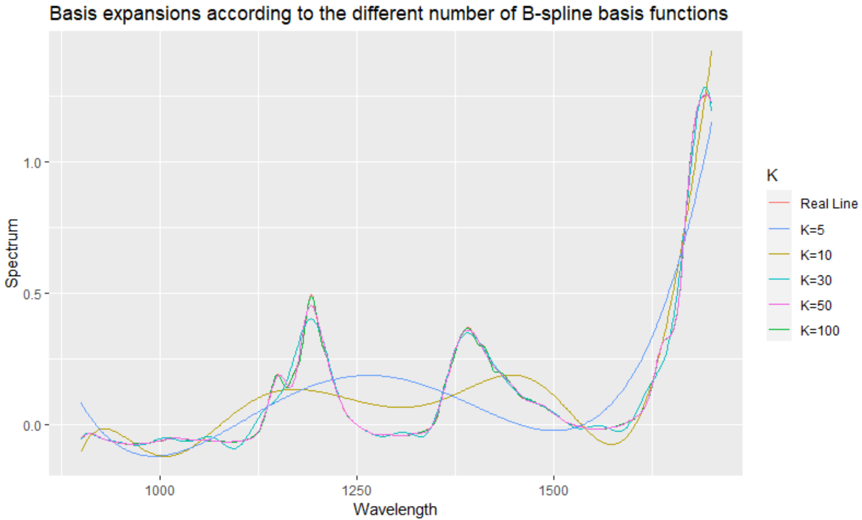
- The larger K, the better fit to the data with also fitting noise
- If K is too small, it would miss some significant information that we want to estimate

Trade Off between Bias and Variance

- $\mathbf{Bias}[\hat{x}(t)] = x(t) - E[\hat{x}(t)]$
- $\mathbf{Var}[\hat{x}(t)] = E[\{\hat{x}(t) - E[x(t)]\}^2]$
- $\mathbf{MSE}[\hat{x}(t)] = E[\{\hat{x}(t) - x(t)\}]$
- $\mathbf{MSE}[\hat{x}(t)] = \mathbf{Bias}^2[\hat{x}(t)] + \mathbf{Var}[\hat{x}(t)]$

In the sense of that, we need to concentrate on decreasing **MSE**.

Trade Off between Bias and Variance



Estimation via Basis Representation

Assume the following **data generating process**

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

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

Assume that we have a data set containing observations each of which is made up of:

- y_i: a scalar realization of $Y(\omega)$
- $f_i(t)$: a realization of $F(\omega)$

Estimation via Basis Representation

Let $\{b_i(t) \mid i = 1, \dots, \infty\}$ be a basis of $\mathbb{L}^2[0,1]$

Then we have the following representation of $\beta(t)$

$$\beta(t) = \sum_{j=1}^{\infty} \psi_j b_j(t) = \sum_{j=1}^{L} \psi_j b_j(t) + \delta(t) \approx \sum_{j=1}^{L} \psi_j b_j(t)$$

and we can transform the data generating process into:

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

Estimation via Basis Representation

$$Z_j(\omega) = \int_0^1 F(\omega)(s) b_j(s) \mathrm{d}s$$

This is a **scalar random variable** leading to the following transformation

$$Y(\omega) = \alpha + \sum_{j=1}^{\infty} \psi_j Z_j(\omega) + \epsilon(\omega)$$

Each combination of observation $f_i(t)$ and deterministic basis function $b_j(t)$ effectively gives us a realization of this random variable.

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

Estimation via Basis Representation

This allows us to write each observation in the data set as

- y_i: a scalar realization of Y
- $(Z_{i,j})_{j \in \mathbb{N}}$: a countably infinite sequence of scalars

Truncating the functional basis allows us to approximate the data set in the usual multivariate form.

- y_i: a scalar realization of Y
- $(Z_{i,1} \ \ldots \ Z_{i,L})'$: a vector of scalar regressors

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

Estimation via Basis Representation

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

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

This is dependent on...

- The functional basis $(b_j(t))_{j \in \mathcal{I}}$ for the estimation of $\beta(t)$
- The truncation parameter L
- (The functional basis used to approximate the observations)
- (The truncation parameter in the approximation of the observations)

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 $\{\gamma_i \mid i = 1, \dots, p\}$ be the orthonormal **Eigenvectors** of Σ_X
- Let $\{\lambda_i \mid i = 1, \dots, p\}$ 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 $i \neq j$

Karhunen-Loéve Expansion

Mean Function:

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

Autocovariance Function:

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

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

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

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 representation is called the **Karhunen-Loéve Expansion** of the random function F and the Eigenfunctions can serve as a basis to represent the function.

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 $i \neq j$

Functional Principal Component Analysis

This idea 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:

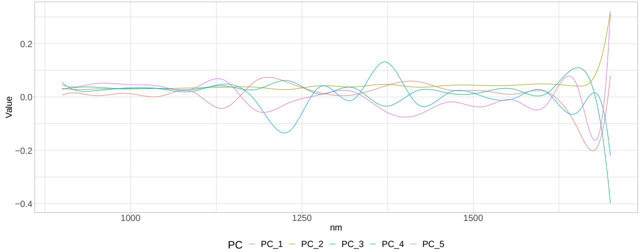
$$\int_0^1 \hat{c}(t,s)\hat{\nu}(s)\mathrm{d}s = \hat{\lambda}\hat{\nu}(t)$$

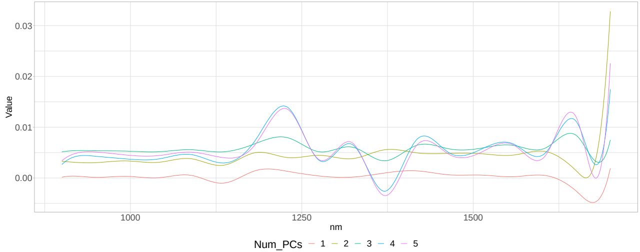
The $\{\hat{\nu}_i(s) \mid i \in \mathcal{I}\}$ are called **Functional Principal Components** and can serve as a basis for representing the original curves.

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 & Application

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

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

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

Simulation Setup & Application cont.

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

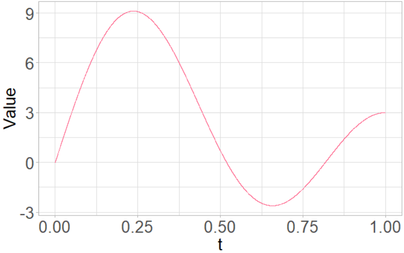


Figure: $f_1(t)$, smooth function

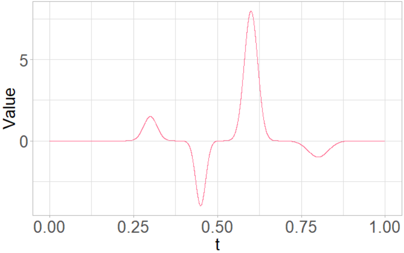


Figure: $f_2(t)$, bumpy function

Simulation Setup & Application cont.

Let

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

$$Y_{2,f} = \langle NIR, f \rangle + Z\frac{var(\langle NIR, f \rangle)}{0.9} - var(\langle NIR, f \rangle)$$

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 $n_{basis} \in (5, 6, ..., 25)$ to perform regression using basis expansion and the FPCR approach.
- Compare results via criteria (CV, Mallows CP,...)
- add results here!

Simulation Setup & Application cont.

- Use insights from the simulation study to uncover dependence.
- Similar setup, but using only bsplie basis expansion and initial 60 spectral curves.
- Validation set approach: Scores of testdata needs to be estimated by the trainingdata. explain in detail?
- Report results by MSE scaled by variance.

Summary

Jona
Just summarize what we have done...

further reading

Put footnotes here!