

# Scalar on Function Regression

Jonathan Willnow, Jakob Juergens, Jonghun Baek

Presentation Day

# Introduction

Jona

Introductory Example  $\rightarrow$  Octane/NIR-spectrum

# Theory

Jona

Motivation from multivariate regression (multivariate dgp).

# 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

# Theory

Jakob

- Random function represented as linear combination of basis functions ✓
- Just transform to multiple linear regression setting ✓
- You already know that from the beginning ✓

# Estimation via Basis Representation

Assume the following data generating process

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

- $Y$  and  $\epsilon$  realize in  $\mathbb{R}$  and  $F$  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  $Y$
- $f_i(t)$ : a realization of  $F$

# 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:

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



# Estimation via Basis Representation

$$Z_j(\omega) = \int_0^1 F(\omega)(s)b_j(s)ds$$

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.

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

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

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

# Estimation via Basis Representation

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

# Theory - FPCA

Jakob

- Let's assume you know the theory of PCA (pc from varcov matrix) ✓
- Introduce mean and covariance functions of random functions ✓
- There is another cool basis  $\rightarrow$  Eigenbasis (Karhunen-Loeve Expansion) ✓
- Sample Analog! (create a basis from observations and use for basis regression) ✓
- Plot fpcs and approximation of function realization

# Spectral Representation of Random Vectors

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

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

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

# Karhunen-Loève Expansion

**Mean Function:**

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

**Autocovariance Function:**

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

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

$$\int_0^1 c(t, s) \nu(s) ds = \lambda \nu(t)$$

# Karhunen-Loève Expansion

A random function  $F$  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  $\xi_j$  are scalar-valued random variables with the following properties.

1  $\mathbb{E}[\xi_i(\omega)] = 0$

2  $\text{Var}(\xi_i(\omega)) = \lambda_i$

3  $\text{Cov}(\xi_i(\omega), \xi_j(\omega)) = 0$  for  $i \neq j$

This representation is called the **Karhunen-Loève 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).

$\Sigma_X$  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  $\{\hat{\gamma}_i \mid i = 1, \dots, p\}$  be the orthonormal **Eigenvectors** of  $\hat{\Sigma}_X$
- Let  $\{\hat{\lambda}_i \mid i = 1, \dots, p\}$  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 $\mathbb{E}[Z_i(\omega)] = 0$               | 3 $\text{Cov}(Z_i(\omega), Z_j(\omega)) = 0$ for |
| 2 $\text{Var}(Z_i(\omega)) = \hat{\lambda}_i$ | $i \neq j$                                       |



# 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:**

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

# 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:

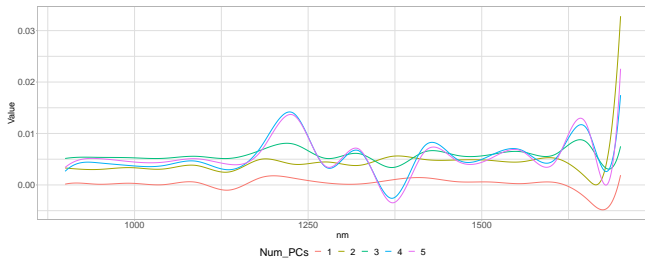
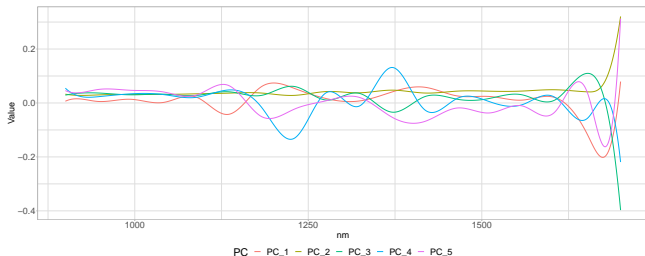
$$\int_0^1 \hat{c}(t, s) \hat{\nu}(s) ds = \hat{\lambda} \hat{\nu}(t)$$

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

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

# FPCA - Plots



Lorem Ipsum  
Dolor Sit  
Amet

# Simulation Setup & Application

Jona

- Compare b-spline / fourier regression chosen via criterion (cv/aic/...)
- Similar for fpca
- generate new curves from observed curves motivated by Karhunen-Loeve expansion
- Compare optimal variants with test and training sets
- Connect to Application

# Summary

Jona

Just summarize what we have done...

# further reading

Put footnotes here!