M7777 Applied Functional Data Analysis

8. Functional Data Simulation

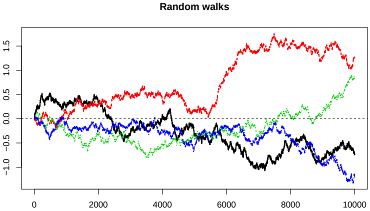
Jan Koláček (kolacek@math.muni.cz)

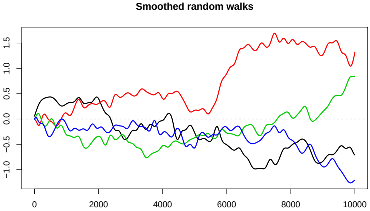
Dept. of Mathematics and Statistics, Faculty of Science, Masaryk University, Brno

Functional Data Simulation

1. Wiener process (limit of a Random Walk)

$$x_i(t_k) = S_k = \frac{1}{\sqrt{N}} \sum_{j=1}^{k} U_j, \qquad iid \; U_j \sim N(0,1), \; j = 1, \ldots, N$$





Combinations with Wiener process

$$x_i(t_k) = m(t_k) + S_k$$

Functional Data Simulation

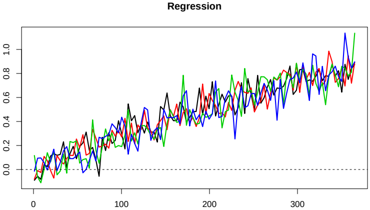
2. Regression model

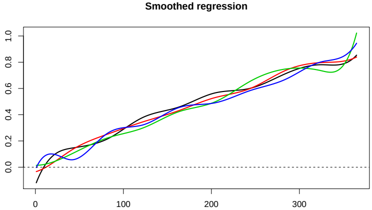
- Simulate $N \times M$ measurements

$$x_i(t_k) = m(t_k) + \varepsilon_{ik}, \qquad iid\ \varepsilon_{ik} \sim N(0, \sigma^2),$$

$m(t) \ldots$ any regression function, $i = 1, \ldots, N,\ k = 1, \ldots, M$

- Smooth the data by FDA $\Rightarrow x_i(t), i = 1, \ldots, N$



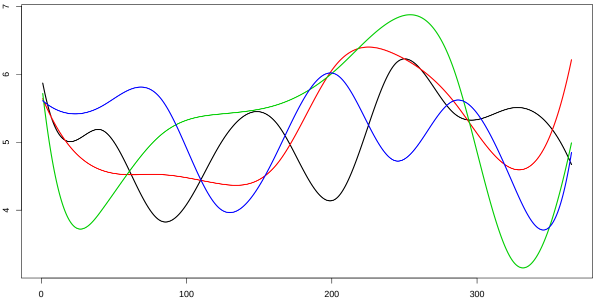


Functional Data Simulation

3. Basis expansion

$$x_i(t_k) = \sum_{j=1}^{K} C_{ij}\Phi_j(t_k).$$

- $\mathbf{\Phi}^*(t) = (\Phi_1(t), \ldots, \Phi_K(t))$ ... a given **basis system**
- C_{ij} ... iid **random** basis coefficients for i-th curve, $j = 1, \ldots, K$



Functional Data Simulation

4. Gaussian process

Let us consider a regression model

$$x_i(t_k) = m(t_k) + \varepsilon_{ik}$$

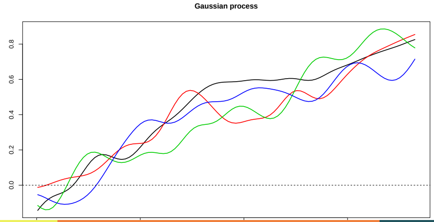
with a covariance function $c(r, s)$, i.e. $\mathrm{Cov}(\varepsilon_{ir}, \varepsilon_{is}) = c(t_r, t_s)$, usually

$$c(u, v) = \sigma^2 \exp\left(-\frac{1}{2l^2}(u - v)^2\right).$$

Set

$\mathbf{m} = (m(t_1), \ldots, m(t_M))', \boldsymbol{\Sigma} = (c(t_i, t_j))_{i,j=1}^{M}, \mathbf{x}_i = (x_i(t_1), \ldots, x_i(t_M))'$

Thus $\mathbf{x}_i \sim N_M(\mathbf{m}, \boldsymbol{\Sigma})$ and $x_i(t) = \lim_{M\to\infty} \mathbf{x}_i$.



Functional Data Simulation

5. Random Gaussian process

Let be given $(t_1^*, y_1^*) \dots, (t_L^*, y_L^*), L < M$, and suppose

$$x_i(t_k^*) = y_k^* + \varepsilon_{ik}^*, \ \boldsymbol{\varepsilon}_i^* \sim N_L(\mathbf{0}, \sigma_*^2 \mathbf{I}_L).$$

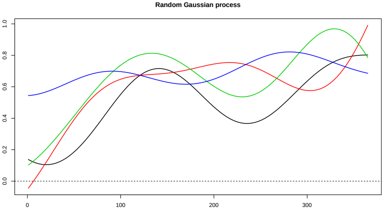
Then

$$\mathbf{x}_i | \mathbf{y}^* \sim N_M(\mathbf{m}^*, \boldsymbol{\Sigma}^*)$$

with

$$\mathbf{m}^* = \boldsymbol{\Sigma}_{\mathbf{t}\mathbf{t}^*} \left(\boldsymbol{\Sigma}_{\mathbf{t}^*\mathbf{t}^*} + \sigma_*^2 \mathbf{I}_L \right)^{-1} \mathbf{y}^*,$$

$$\boldsymbol{\Sigma}^* = \boldsymbol{\Sigma}_{\mathbf{t}\mathbf{t}} - \boldsymbol{\Sigma}_{\mathbf{t}\mathbf{t}^*} \left(\boldsymbol{\Sigma}_{\mathbf{t}^*\mathbf{t}^*} + \sigma_*^2 \mathbf{I}_L \right)^{-1} \boldsymbol{\Sigma}_{\mathbf{t}^*\mathbf{t}}, \text{ where } \boldsymbol{\Sigma}_{\mathbf{ab}} = \mathrm{Cov}(\mathbf{a}, \mathbf{b}).$$



Functional Data Simulation

Regression Simulation

1. Generate y_i with known $\alpha, \beta(t), x_i(t)$ and ε_i, $i = 1, \dots, 30$
2. Get estimates $\hat{\beta}(t), \hat{y}$ by considered methods
   a) Estimation through a basis expansion $\dots \hat{\beta}_{BE}(t), \hat{y}_{BE}$
   b) Estimation with a roughness penalty $\dots \hat{\beta}_{RP}(t), \hat{y}_{RP}$
   c) Regression on functional principal components $\dots \hat{\beta}_{PC}(t), \hat{y}_{PC}$
   d) Nonparametric regression $\dots \hat{\beta}_{NR}(t), \hat{y}_{NR}$
3. Compare $\hat{\beta}_{BE}, \hat{\beta}_{RP}, \hat{\beta}_{PC}, \hat{\beta}_{NR}$ with known β
4. Compare estimates $\hat{y}_{BE}, \hat{y}_{RP}, \hat{y}_{PC}, \hat{y}_{NR}$ with known model fits.

Functional Data Simulation

Regression Simulation 1

Let $(t_1, \ldots, t_M) = (1, 2, \ldots 365)$, we will simulate 30 regression curves $x_i(t)$ as the Gaussian process with

$$m(t) = \sin(t/365), \quad c(u, v) = 0.01 \exp\left(-\frac{1}{1\,000}(u - v)^2\right).$$

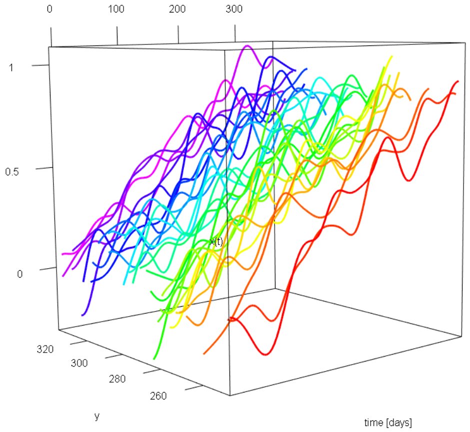
The regression model takes the form

$$y_i = -10 + \int_1^{365} \beta(t) x_i(t) dt + \varepsilon_i$$

with $\beta(t) = 1 + 2t/365 - (t/365)^2$ and $\varepsilon_i \sim N(0, 5)$.

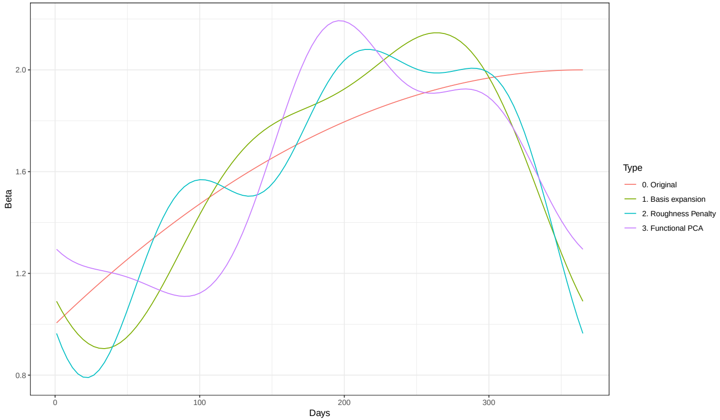
Functional Data Simulation

Simulated Data

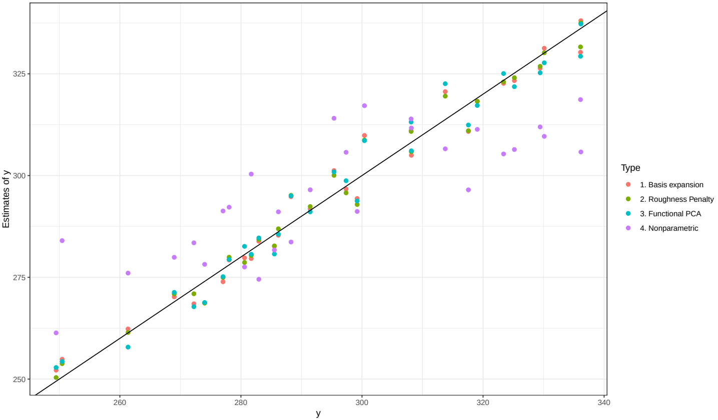


Functional Data Simulation

Comparison of β



Comparison of fits



Functional Data Simulation

Regression Simulation 2

Let $(t_1, \ldots, t_M) = (0, 0.01, \ldots 1)$, we will simulate 30 regression curves $x_i(t)$ as the Gaussian process with

$$m(t) = \exp(t/2\pi), \quad c(u, v) = 0.5 \exp\left(-10(u - v)^2\right).$$

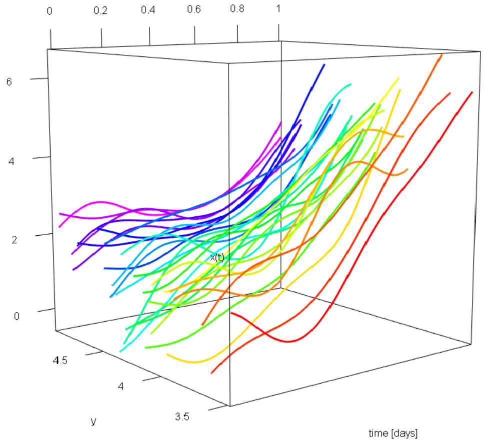
The regression model takes the form

$$y_i = 5 + \int_0^1 \beta(t) x_i(t) dt + \varepsilon_i$$

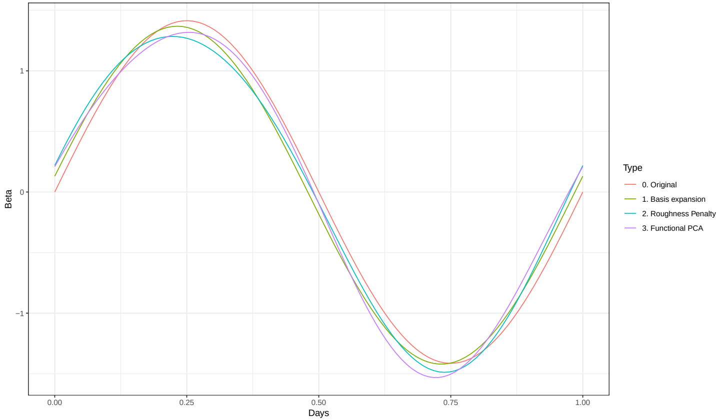
with $\beta(t) = \sin(2\pi t)$ and $\varepsilon_i \sim N(0, 0.1)$.

Functional Data Simulation

Simulated Data

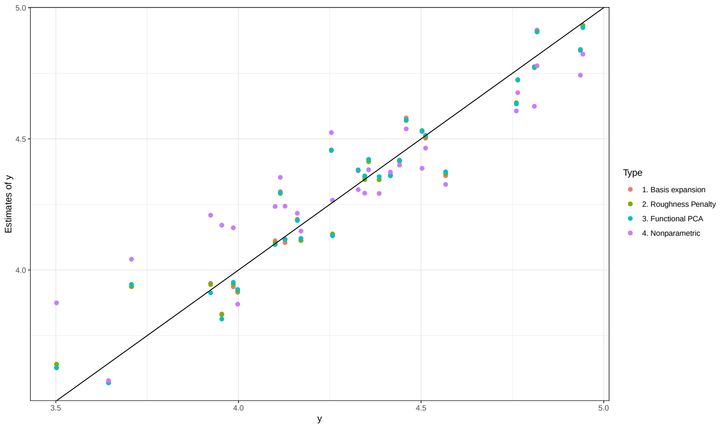


Comparison of β



Functional Data Simulation

Comparison of fits



Problems to solve

1. Conduct the following simulation (Kokoszka and Reimherr, 2017).
   - Generate 1 000 random functions

     $$X(t_j) = Zt_j + U + \eta(t_j) + \epsilon(t_j),$$

     where t_j are 101 equidistantly distributed points on $[0, 1]$, $\eta(t_j) \sim N(0, 1)$, $Z \sim N(1, 0.2^2)$, $U \sim UNIF(0, 5)$ and the random curves $\epsilon(t)$ will be generated as

     $$\epsilon(t) = \sum_{k=1}^{10} \frac{1}{k} \left\{ Z_{1k} \sin(2\pi tk) + Z_{2k} \cos(2\pi tk) \right\}$$

     with independent standard normal Z_{1k}, Z_{2k}.
   - Consider a regression model of the form

     $$y_i = 0.01 \int_0^1 \beta(t) X_i(t) dt + \varepsilon_i$$

     with $\beta(t) = -f_1(t) + 3f_2(t) + f_3(t)$ and $\varepsilon_i \sim N(0, 0.4)$, where f_1, f_2, f_3 are normal densities $N(0.2, 0.03^2)$, $N(0.5, 0.04^2)$, $N(0.75, 0.05^2)$, respectively.

Problems to solve

- Try all regression approaches studied in the previous lesson, i.e.
  - estimation through a basis expansion,
  - estimation with a roughness penalty and
  - regression on FPCA.

  Plot the estimates $\hat{\beta}(t)$ and compare it with the original $\beta(t)$ (see Figure 1).
- Conduct the nonparametric regression. Plot estimated values $\hat{y}_i$ against the simulated y_i for all cases (see Figure 2).

Problems to solve

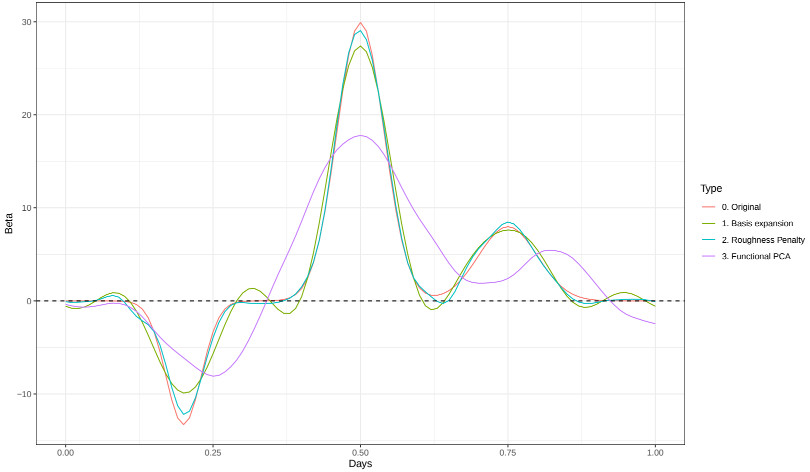


Figure 1.

Problems to solve

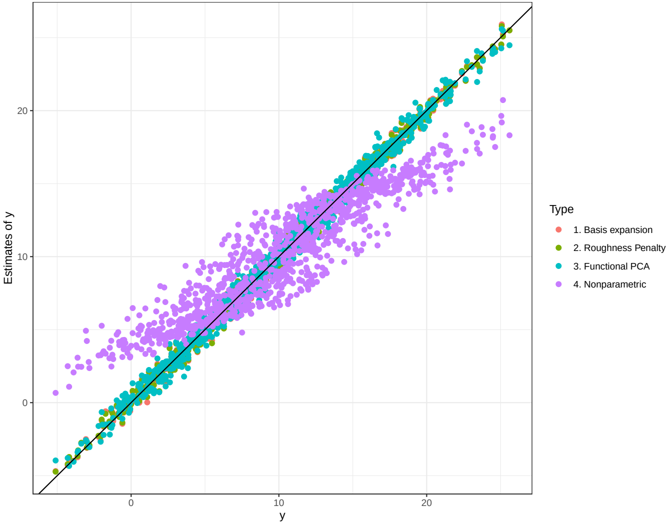


Figure 2.