Statistical methods of data science

An introduction to Functional Data Analysis

Alessandra Menafoglio[1*]

[1]MOX, Department of Mathematics, Politecnico di Milano

*alessandra.menafoglio@polimi.it

4. Linear models

Agenda

4. Linear models

4.1. Functional Linear Models in Hilbert spaces

4.2. A case study on ground motion modelling

Agenda

4. Linear models

4.1. Functional Linear Models in Hilbert spaces

4.2. A case study on ground motion modelling

- Linear models in Hilbert spaces can be described by four types of models, depending on the type of regressors and response involved

	Multivariate Regressors	Functional Regressors
Multivariate Response	$\boldsymbol{x} = \mathbb{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ $x_i = \sum_{l=1}^{L} z_{il}\beta_l + \varepsilon_i$ $\boldsymbol{x}, \boldsymbol{\varepsilon} \in \mathbb{R}^n$ $\boldsymbol{\beta} \in \mathbb{R}^L$ $\mathbb{Z} \in \mathbb{R}^{n,L}$	$\boldsymbol{x} = \mathbb{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ $x_i = \sum_{l=1}^{L} \langle z_{il}, \beta_l \rangle + \varepsilon_i(t)$ $\boldsymbol{x}, \boldsymbol{\varepsilon} \in \mathbb{R}^n$ $\boldsymbol{\beta} \in G \subseteq H^L$ $\mathbb{Z} \in \mathcal{L}(G, \mathbb{R}^n)$
Functional Response	$\boldsymbol{x}(t) = \mathbb{Z}\boldsymbol{\beta}(t) + \boldsymbol{\varepsilon}(t)$ $x_i(t) = \sum_{l=1}^{L} z_{il}\beta_l(t) + \varepsilon_i(t)$ $\boldsymbol{x}, \boldsymbol{\varepsilon} \in H^n$ $\boldsymbol{\beta} \in G \subseteq H^L$ $\mathbb{Z} \in \mathbb{R}^{n,L}$	$\boldsymbol{x}(t) = (\mathbb{Z}\boldsymbol{\beta})(t) + \boldsymbol{\varepsilon}(t)$ $x_i(t) = \sum_{l=1}^{L} z_{il}(t)\beta_l(t) + \varepsilon_i(t)$ $x_i(t) = \sum_{l=1}^{L} \langle z_{il}, \beta_l(t, \cdot) \rangle + \varepsilon_i(t)$ $\boldsymbol{x}, \boldsymbol{\varepsilon} \in H^n$ $\boldsymbol{\beta} \in G \subseteq H^L$ $\mathbb{Z} \in \mathcal{L}(G, H^n)$

4.1. Functional Linear Models in Hilbert spaces

Introduction

- Linear models in Hilbert spaces can be described by four types of models, depending on the type of regressors and response involved
- In this part of the course, we will mainly focus on the space L^2, since the theory has been mainly developed and applied in L^2. Nevertheless, most of the results presented can be extended (or applied) in general Hilbert spaces
- Furthermore, we will focus on the case of functional responses

	Multivariate Regressors	**Functional Regressors**
Functional Response	$\boldsymbol{x}(t) = \mathbb{Z}\boldsymbol{\beta}(t) + \boldsymbol{\varepsilon}(t)$ $x_i(t) = \sum_{l=1}^{L} z_{il}\beta_l(t) + \varepsilon_i(t)$	$\boldsymbol{x}(t) = (\mathbb{Z}\boldsymbol{\beta})(t) + \boldsymbol{\varepsilon}(t)$ $x_i(t) = \sum_{l=1}^{L} z_{il}(t)\beta_l(t) + \varepsilon_i(t)$ $x_i(t) = \sum_{l=1}^{L} \langle z_{il}, \beta_l(t,\cdot)\rangle + \varepsilon_i(t)$
	$\boldsymbol{x}, \boldsymbol{\varepsilon} \in H^n$ $\boldsymbol{\beta} \in G \subseteq H^L$ $\mathbb{Z} \in \mathbb{R}^{n,L}$	$\boldsymbol{x}, \boldsymbol{\varepsilon} \in H^n$ $\boldsymbol{\beta} \in G \subseteq H^L$ $\mathbb{Z} \in \mathcal{L}(G, H^n)$

4.1. Functional Linear Models in Hilbert spaces

Premiss

- The theory on point-wise prediction for functional models is well-developed for parameter estimation; computational methods; arguments and proof of consistency of the estimators
- Developing inferential methods for, e.g., testing the significance of factors/regressors on a response is still a broad field of research
    - Solutions can be worked out theoretically for Gaussian processes
    - Recent techniques based on resampling (e.g., bootstrap) or permutation tests

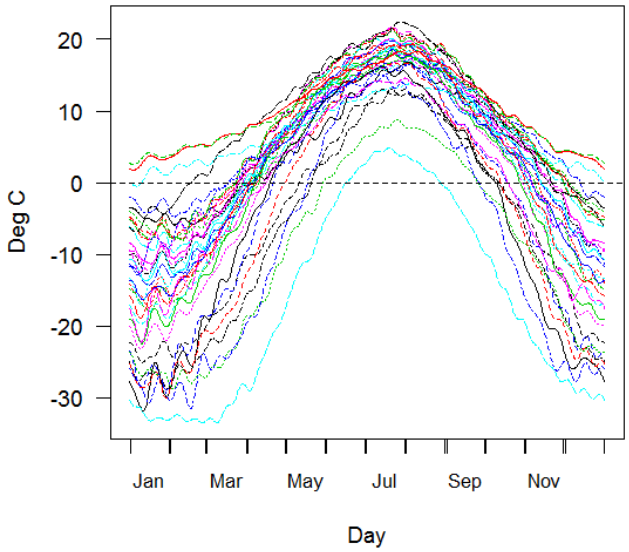
      Abramowicz et al. (2018) "Nonparametric inference for functional-on-scalar linear models applied to knee kinematic hop data after injury of the anterior cruciate ligament". Scand J Statist. 2018; 45: 1036–1061. https://doi.org/10.1111/sjos.12333
- Flexible regression methodologies have been developed within the context of GAM (Generalized Additive Models)

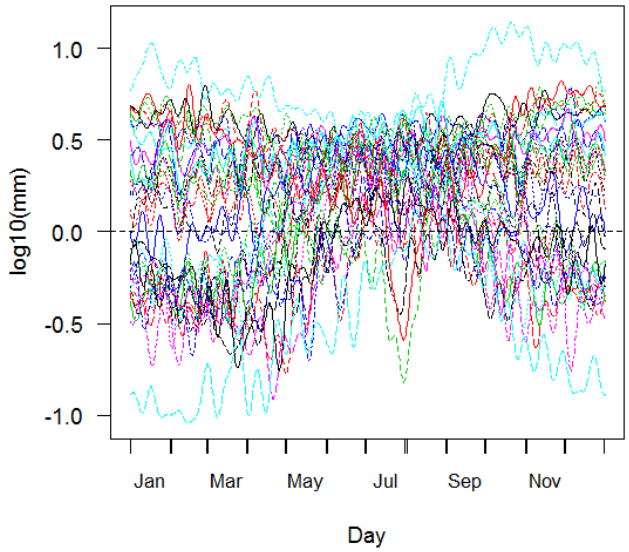
  McLean, Mathew W., Giles Hooker, Ana-Maria Staicu, Fabian Scheipl, and David Ruppert. "Functional Generalized Additive Models." Journal of Computational and Graphical Statistics 23, no. 1 (2014): 249–69. doi:10.1080/10618600.2012.729985.
  Greven, S. and Scheipl, F. (2017). A general framework for functional regression mod- elling (with discussion and rejoinder). Statistical Modelling, 17(1-2):1–35 and 100–115. 17, 21
- For the purpose of this course, we will mainly focus on parameter estimation and case studies in L^2, although theory and applications are also available in B^2

4.1. Functional Linear Models in Hilbert spaces

Motivating example

- We focus on Canadian weather data, that collect temperatures and precipitation for 35 weather stations across the country
- We consider the data smoothed with the methodologies devised before (smoothing through expression on a functional basis)
- We express precipitation with log, to respect the positivity constraint (note: precipitation are positive data, but not compositional)

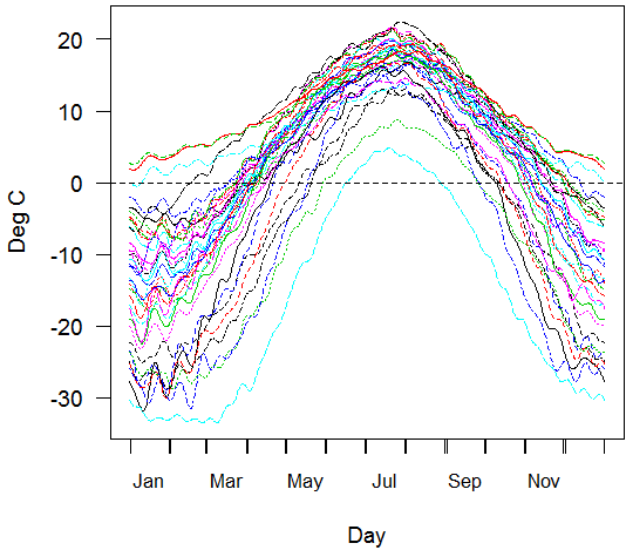




4.1. Functional Linear Models in Hilbert spaces

Motivating example

- Let's focus first on temperatures. We may ask ourselves, which portion of the data variability is explained, e.g., by the geographical area: we divide Canada into four meteorological zones, namely Atlantic, Continental, Pacific and Artic



- Let's focus first on temperatures. We may ask ourselves, which portion of the data variability is explained, e.g., by the geographical area: we divide Canada into four meteorological zones, namely Atlantic, Continental, Pacific and Artic
- We may thus want to build a functional ANOVA (FANOVA) model:

$$\texttt{Temp}_{mg}(t) = \mu(t) + \alpha_g(t) + \epsilon_{mg}(t)$$

  with $\sum \alpha_g(t) = 0$ for all *t*.
- The FANOVA is a particular kind of linear model with functional response and multivariate regressors. Indeed, let us define:
  - The design matrix of the model as the 35 x 5 matrix **Z** whose element *i,m* is 1 if the *i*-th station belongs to the *m*-th group, 0 otherwise.
  - The 5 functional coefficients as
- The model can be then formulated as $\boldsymbol{\beta} = (\mu, \alpha_1, \alpha_2, \alpha_3, \alpha_4)'$

$$\texttt{Temp}_{mg}(t) = \sum_{j=1}^{5} z_{(mg)j}\beta_j(t) + \epsilon_{mg}(t) \quad \text{or} \quad \texttt{Temp} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

$\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$ are now vectors of functions!

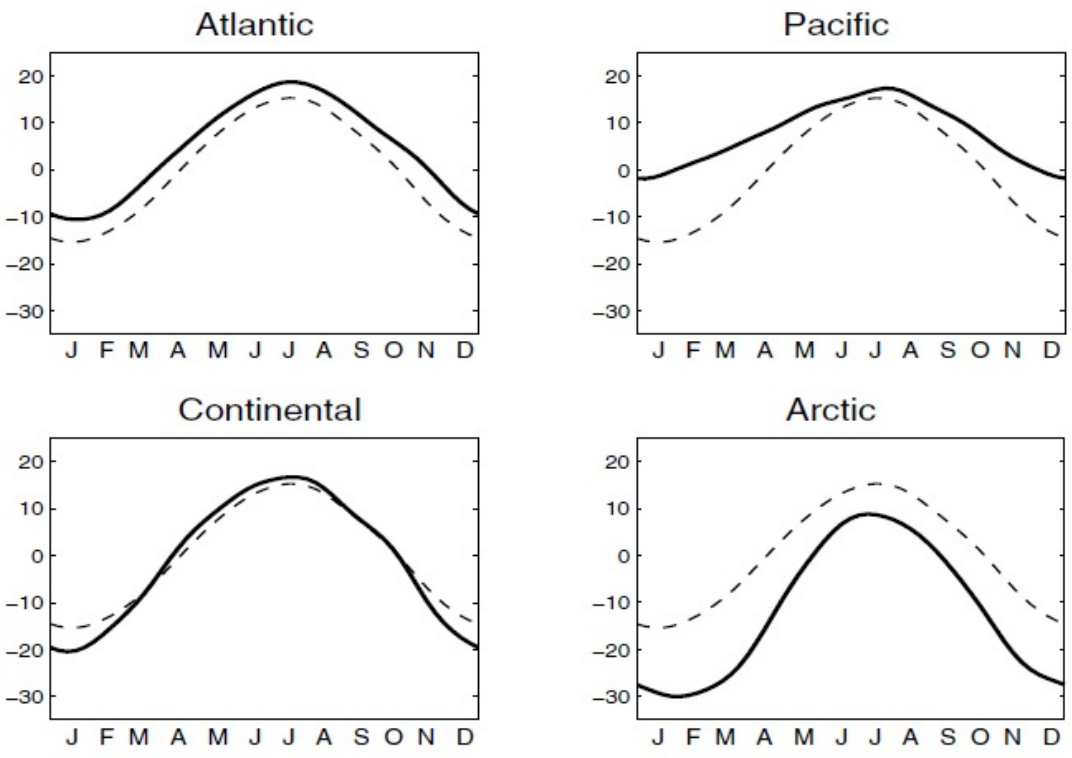
- If the model were a standard linear model, we would estimate the parameters as to minimize the residual sum of squares
- In the functional case, the residuals of the model are $\texttt{Temp}_i(t) - \mathbf{Z}_i\boldsymbol{\beta}(t)$, i.e., we have functional residuals.
- We can define the least square functional as

$$\texttt{LMSSE}(\boldsymbol{\beta}) = \sum_{g}^{4}\sum_{m}^{N_g}\int [\texttt{Temp}_{mg}(t) - \sum_{j}^{q} z_{(mg),j}\beta_j(t)]^2\, dt.$$

  and minimize this quantity under the constraint $\sum_2^5 \beta_j = 0$ (recall: the FANOVA coefficients were constrained).
- Pretending for the moment to be able to minimize LMSSE, we can obtain the coefficients β_j of the model, hence the mean within each group [$\boldsymbol{\beta} = (\mu, \alpha_1, \alpha_2, \alpha_3, \alpha_4)'$]

$$\texttt{Temp}_{mg}(t) = \mu(t) + \alpha_g(t) + \epsilon_{mg}(t)$$

- Results of the estimation (then we come back to how these parameters are estimated, in general)



- Atlantic stations: around 5 degrees C warmer than the Canadian average.
- Pacific stations: summer temperature close to the Canadian average, but much warmer winter.
- Continental stations: slightly warmer than average in the summer, but colder in the winter by about 5 degrees C.
- Arctic stations: colder than average (more in March than in January).

The figure shows the estimated means within the groups (solid line), and the overall mean (dashed line)

- The general model is

$$\mathbf{y}(t) = \mathbf{Z}\boldsymbol{\beta}(t) + \boldsymbol{\epsilon}(t)$$

with

  - $\mathbf{y}$ the vector of functional N observations
  - $\mathbf{Z}$ the design matrix N x q
  - $\boldsymbol{\beta}$ the vector of q functional parameters
  - $\boldsymbol{\epsilon}$ the vector of N functional errors

- Possible constraints on the parameters (e.g.. zero-sum as in FANOVA) are considered to be expressed as $\mathbf{L}\boldsymbol{\beta} = 0$ for a suitable matrix $\mathbf{L}$. [One can come back to the basic model by substitution of variables (e.g. $\boldsymbol{\beta} = \mathbf{C}\alpha$ with $\mathbf{C}$ appearing in the QR decomposition, and the model becomes $\mathbf{y}(t) = \mathbf{Z}\mathbf{C}\alpha(t) + \boldsymbol{\epsilon}(t)$).]

- We aim to minimize the residual sum of squares

$$\text{LMSSE}(\boldsymbol{\beta}) = \int [\mathbf{y}(t) - \mathbf{Z}\boldsymbol{\beta}(t)]'[\mathbf{y}(t) - \mathbf{Z}\boldsymbol{\beta}(t)]\, dt$$

Squared norm of the residuals; could be done in any *H*

4.1. Functional Linear Models in Hilbert spaces

Functional response on multivariate regressors: Estimating parameters

- Aim to minimize the residual sum of squares

$$\texttt{LMSSE}(\boldsymbol{\beta}) = \int [\mathbf{y}(t) - \mathbf{Z}\boldsymbol{\beta}(t)]'[\mathbf{y}(t) - \mathbf{Z}\boldsymbol{\beta}(t)]\, dt$$

  according to Ramsay & Silverman (2005) we can proceed in two ways:

  1. Minimize it **point-wise** on a grid of values of *t*
  2. Use a **regularized basis expansion**

- Let's focus on the **1st approach**. Given a point of evaluation *t*, the solution is given by

$$\widehat{\boldsymbol{\beta}}(t) = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}(t)$$
$$\widehat{\mathbf{y}}(t) = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}(t)$$

- It is easy to prove that point-wise minimization of LMSSE corresponds to formulating normal equations in L^2 similar as those used in the multivariate setting. That is, the optimal point-wise solution corresponds almost everywhere to the optimal solution in L^2 .

- This follows from the properties of the functional space L^2 . In particular, in L^2 we are allowed to perform **projections**, since it is a Hilbert space.

- More in general, let's consider the model formulated in any Hilbert space

$$\mathbf{y}(t) = \mathbf{Z}\boldsymbol{\beta}(t) + \boldsymbol{\epsilon}(t)$$

- We aim to minimize the residuals sum of squares (in *H*)

$$\|\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}\|_H^2$$

- The solution is given by

$$\widehat{\mathbf{y}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$$

  with the notation

$$[\mathbf{A}\mathbf{x}]_i = \sum_{ij} \mathbf{A}_{ij}\mathbf{x}_j$$

- *Note.* If the data are expressed through a basis representation, one can express the previous expressions in terms of the basis coefficients and then perform a multiple linear regression on the basis coefficients.

4.1. Functional Linear Models in Hilbert spaces

Functional response on multivariate regressors: Regularized basis expansion

- Let's now look at the **2nd approach**, that is using a regularized basis expansion. We thus assume that the functional observations and the parameters are expressed in a basis expansion form (e.g., through Fourier or B-spline basis functions).
- We focus on L^2, and assume that the observations are expressed on a basis $\boldsymbol{\phi}$ and the parameters are expressed on a basis $\boldsymbol{\theta}$ not necessarily coincident with $\boldsymbol{\phi}$

$$\mathbf{y} = \mathbf{C}\boldsymbol{\phi}$$
$$\hat{\boldsymbol{\beta}} = \mathbf{B}\boldsymbol{\theta}$$

- To impose a given degree of smoothness on the parameters we could
  a. Truncate their basis expansion (and go back to approach 1.), or
  b. Impose a roughness penalty on the parameters while estimating the model (i.e., approach 2.)
- We here pursue the choice b), and use a differential operator L to define the roughness penalty for (e.g., the curvature)

$$\text{PEN}_L(\beta) = \int [L\boldsymbol{\beta}(s)]'[L\boldsymbol{\beta}(s)]\, ds$$

- To estimate the parameters, we minimize a penalized least squares criterion

$$\texttt{PENSSE}(y|\boldsymbol{\beta}) = \int (\mathbf{C}\boldsymbol{\phi} - \mathbf{Z}\mathbf{B}\boldsymbol{\theta})'(\mathbf{C}\boldsymbol{\phi} - \mathbf{Z}\mathbf{B}\boldsymbol{\theta}) + \lambda \int (L\mathbf{B}\boldsymbol{\theta})'(L\mathbf{B}\boldsymbol{\theta}) \, .$$

- Using Kronecker products and linear algebra, one can find the minimizer of $\texttt{PENSSE}(y|\boldsymbol{\beta})$ as

$$\text{vec}\,(\mathbf{B}) = [\mathbf{J}_{\theta\theta} \otimes (\mathbf{Z}'\mathbf{Z}) + \mathbf{R} \otimes \lambda \mathbf{I}]^{-1} \text{vec}\,(\mathbf{Z}'\mathbf{C}\mathbf{J}_{\phi\theta})$$

with

$$\mathbf{J}_{\phi\phi} = \int \boldsymbol{\phi}\boldsymbol{\phi}' \, , \quad \mathbf{J}_{\theta\theta} = \int \boldsymbol{\theta}\boldsymbol{\theta}' \, , \quad \mathbf{J}_{\phi\theta} = \int \boldsymbol{\phi}\boldsymbol{\theta}' \quad \text{and} \quad \mathbf{R} = \int (L\boldsymbol{\theta})(L\boldsymbol{\theta})'$$

Note. The solution satisfies the normal equations

$$(\mathbf{Z}'\mathbf{Z}\mathbf{B}\mathbf{J}_{\theta\theta} + \lambda \mathbf{B}\mathbf{R}) = \mathbf{Z}'\mathbf{C}\mathbf{J}_{\phi\theta}$$

- Estimated the model, we can evaluate the goodness-of-fit. As in classical linear model, goodness-of-fit can be evaluated through appropriate measures of Sum of Squares / Variances.

- In particular we can consider:

$$\mathbf{Var}_{Tot}(t) = \sum_{i=1}^{N} [y_i(t) - \overline{y}]^2$$

$$\mathbf{Var}_{Reg}(t) = \sum_{i=1}^{N} [\mathbf{z}_i' \widehat{\boldsymbol{\beta}}(t) - \overline{y}]^2$$

$$\mathbf{Var}_{Res}(t) = \sum_{i=1}^{N} [y_i(t) - \mathbf{z}_i' \widehat{\boldsymbol{\beta}}(t)]^2$$

$$\mathbf{R}^2(t) = \frac{\mathbf{Var}_{Reg}(t)}{\mathbf{Var}_{Tot}(t)}$$

- We can imagine to consider functional regressors instead of scalar ones.

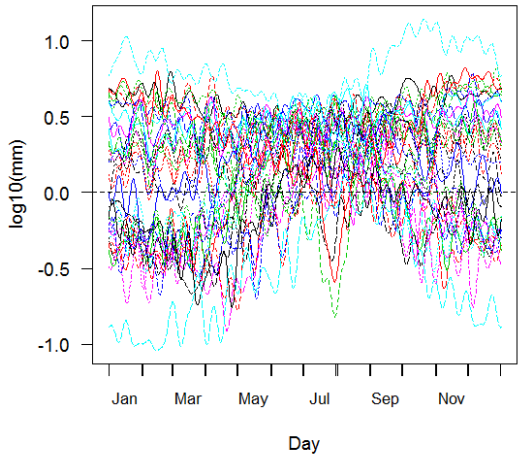
Example: Canadian weather dataset

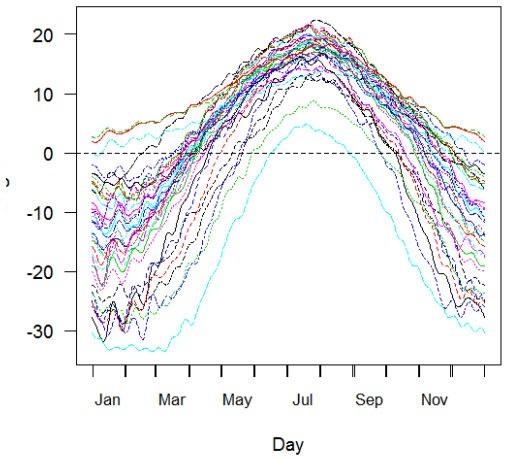
Let us consider the Canadian weather dataset. We aim to build a model for the (log)precipitation as a function of:

- The climatic region
- The temperature

We assume a point-wise dependence

$$\log[\texttt{Prec}_{mg}(t)] = \mu(t) + \alpha_g(t) + \texttt{TempRes}_{mg}(t)\beta(t) + \epsilon_{mg}(t).$$





- We can imagine to consider functional regressors instead of scalar ones.

Example: Canadian weather dataset

Let us consider the Canadian weather dataset. We aim to build a model for the (log)precipitation as a function of:

- The climatic region
- The temperature

We assume a point-wise dependence

$$\log[\texttt{Prec}_{mg}(t)] = \mu(t) + \alpha_g(t) + \texttt{TempRes}_{mg}(t)\beta(t) + \epsilon_{mg}(t).$$

Note. The variable $\texttt{TempRes}_{mg}$ is the residual temperature after removing the temperature effect of climate zone *g* by using the FANOVA model.
We remove temperature climate effects from the temperature profiles before using them in this model because we have already allowed for these effects in the model (we don't want climate zones in the equation twice).

- We can imagine to consider functional regressors instead of scalar ones.

Example: Canadian weather dataset

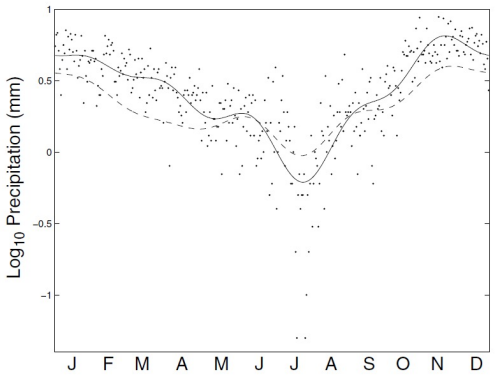
Let us consider the Canadian weather dataset. We aim to build a model for the (log)precipitation as a function of:

- The climatic region
- The temperature

We assume a point-wise dependence

$$\log[\texttt{Prec}_{mg}(t)] = \mu(t) + \alpha_g(t) + \texttt{TempRes}_{mg}(t)\beta(t) + \epsilon_{mg}(t).$$

How to solve the minimization problem in practice? As before, we may consider an approach based on a regularized basis expansion and obtain an estimate of the log-precipitation



Annual log-precipitation in Vancouver. Raw data (symbols); smoothed data (solid curve); predicted precipitation

- More in general, let us consider the model (in L^2)

$$y_i(t) = \sum_{j=1}^{q} z_{ij}(t)\beta_j(t) + \epsilon_i(t)$$

Concurrent model

or, in matrix notation

$$\mathbf{y}(t) = \mathbf{Z}(t)\boldsymbol{\beta}(t) + \boldsymbol{\epsilon}(t) \ ,$$

- We express both the data and the parameters on appropriate bases (not necessarily the same for data and parameters), and introduce a roughness penalty to control the smoothness of the β's, that is

$$\mathtt{PEN}_j(\beta_j) = \lambda_j \int [L_j\beta_j(t)]^2 \, dt$$

- We then minimize the weighted regularized fitting criterion, i.e.,

$$\mathtt{LMSSE}(\boldsymbol{\beta}) = \int \mathbf{r}(t)'\mathbf{r}(t) \, dt + \sum_{j}^{p} \lambda_j \int [L_j\beta_j(t)]^2 \, dt$$

with

$$\mathbf{r}(t) = \mathbf{y}(t) - \mathbf{Z}(t)\boldsymbol{\beta}(t).$$

- We introduce the following notation

$$\beta_j(t) = \sum_k^{K_j} b_{kj}\theta_{kj}(t) = \boldsymbol{\theta}_j(t)'\mathbf{b}_j(t)$$

$$\mathbf{b} = (\mathbf{b}_1', \mathbf{b}_2', \ldots, \mathbf{b}_q')'$$

$$\mathbf{R}_j = \lambda_j \int \boldsymbol{\theta}_j(t)\boldsymbol{\theta}_j'(t)\, dt$$

$$\boldsymbol{\Theta} = \begin{bmatrix} \boldsymbol{\theta}_1' & 0 & \cdots & 0 \\ 0 & \boldsymbol{\theta}_2' & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{\theta}_q' \end{bmatrix} \quad \mathbf{R} = \begin{bmatrix} \mathbf{R}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{R}_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \mathbf{R}_q \end{bmatrix}$$

- We express the model $\mathbf{y}(t) = \mathbf{Z}(t)\boldsymbol{\beta}(t) + \boldsymbol{\epsilon}(t)$, in terms of the basis expansions as:

$$\mathbf{y}(t) = \mathbf{Z}(t)\boldsymbol{\Theta}(t)\mathbf{b} + \boldsymbol{\epsilon}(t)$$

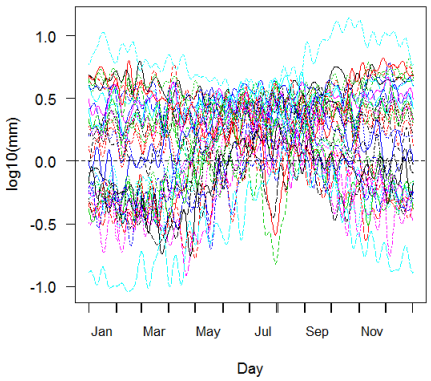
- Similar as in the previous case, the model is estimated by solving the normal equations

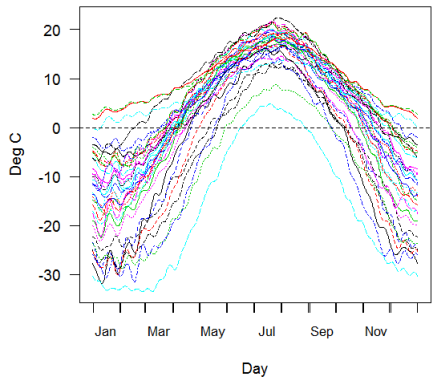
$$[\int \boldsymbol{\Theta}'(t)\mathbf{Z}'(t)\mathbf{Z}(t)\boldsymbol{\Theta}(t)\, dt + \mathbf{R}]\mathbf{b} = [\int \boldsymbol{\Theta}'(t)\mathbf{Z}'(t)\mathbf{y}(t)\, dt]$$

Example: Canadian weather dataset

- The functional linear model considered so far was based on modeling the dependence between the response and regressors as point-wise only

$$\log[\texttt{Prec}_{mg}(t)] = \mu(t) + \alpha_g(t) + \texttt{TempRes}_{mg}(t)\beta(t) + \epsilon_{mg}(t).$$





- We may think to predict the response not-only from concurrent information, but from the entire curve, e.g., through a model of the kind

$$\texttt{LogPrec}_i(t) = \alpha(t) + \int_0^{365} \texttt{Temp}_i(s)\beta(s,t)\,ds + \epsilon_i(t)$$

Total model

$\beta(s,t)$ quantifies the relative weight placed on the temperature at day *s* that is required to predict the log-precipitation of day *t*.

4.1. Functional Linear Models in Hilbert spaces

Functional response on functional regressors: fitting the model without regularization

- In general, the total model is, in matrix notation,

$$\mathbf{y}^*(t) = \int \mathbf{z}^*(s)\beta(s,t)\,ds + \boldsymbol{\epsilon}(t) \;=\; \int \mathbf{z}^*(s)\boldsymbol{\theta}'(s)\mathbf{B}\boldsymbol{\eta}(t)\,ds + \boldsymbol{\epsilon}(t) \;=\; \mathbf{Z}^*\mathbf{B}\boldsymbol{\eta}(t) + \epsilon(t),$$

with

$$\beta(s,t) = \sum_{k=1}^{K_1}\sum_{\ell=1}^{K_2} b_{k\ell}\,\eta_k(s)\theta_\ell(t) = \boldsymbol{\eta}(s)'\mathbf{B}\boldsymbol{\theta}(t) \quad \text{and} \quad \mathbf{Z}^* = \int \mathbf{z}^*(s)\boldsymbol{\theta}'(s)\,ds\ .$$

- One can follow previous arguments to minimize LMSSE, and express the normal equations as

$$\mathbf{Z}^{*'}\mathbf{Z}^*\mathbf{B}\int \boldsymbol{\eta}(t)\boldsymbol{\eta}'(t)\,dt = \mathbf{Z}^{*'}\int y(t)\boldsymbol{\eta}'(t)\,dt\ .$$

- Using Kronecker product one gets

$$[\mathbf{J}_{\eta\eta}\otimes(\mathbf{Z}^{*'}\mathbf{Z}^*)]\mathrm{vec}(\mathbf{B}) = \mathrm{vec}(\mathbf{Z}^{*'}\int \mathbf{y}(t)\boldsymbol{\eta}'(t)\,dt)\ ,$$

with $\mathbf{J}_{\eta\eta} = \int \boldsymbol{\eta}(t)\boldsymbol{\eta}'(t)\,dt$.

4.1. Functional Linear Models in Hilbert spaces

Functional response on functional regressors: fitting the model without regularization

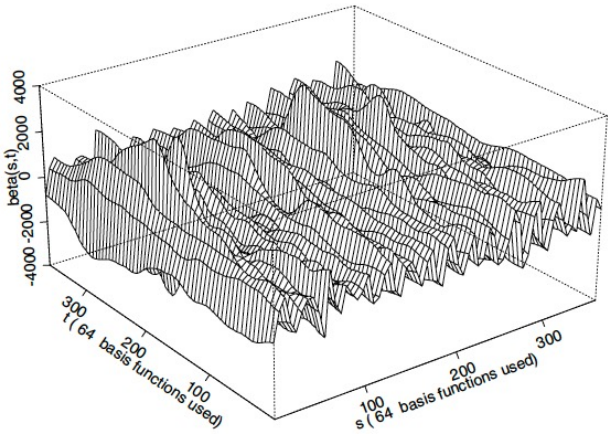


Figure 16.1. The functional parameter function β for the prediction of log precipitation from temperature, estimated direct from the data. The value $\beta(s,t)$ shows the influence of temperature at time s on log precipitation at time t.

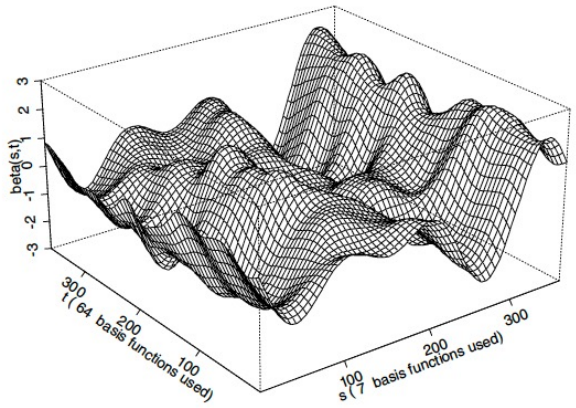


Figure 16.2. Perspective plot of estimated β function truncating the basis for the temperature covariates to 7 terms.

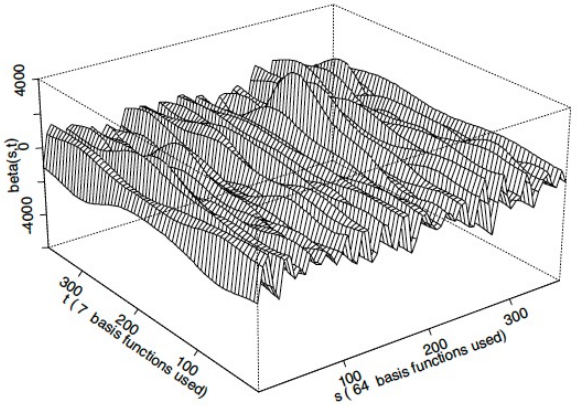


Figure 16.4. Perspective plot of estimated β function truncating the basis for the log precipitations to 7 terms.

- Alternatively, we can consider **penalizing an appropriate differential operator** associated with the parameters to control the smoothness of the estimate. For instance, we can penalize the curvature in *t* as

$$\begin{aligned} \texttt{PEN}_s(\beta) &= \int\int [L_s\beta(s,t)]^2\, ds\, dt \\ &= \text{trace}[\mathbf{B}'\mathbf{R}\mathbf{B}\mathbf{J}_{\eta\eta}] \ , \end{aligned}$$

with $\mathbf{R} = \int [L_s\boldsymbol{\theta}(s)][L_s\boldsymbol{\theta}'(s)]\, ds$.

- To penalize the curvature in *s*, analogously,

$$\begin{aligned} \texttt{PEN}_t(\beta) &= \int\int [L_t\beta(s,t)]^2\, ds\, dt \\ &= \text{trace}[\mathbf{B}'\mathbf{J}_{\theta\theta}\mathbf{S}\mathbf{B}] \ , \end{aligned}$$

$$\mathbf{S} = \int [L_t\boldsymbol{\eta}(t)][L_t\boldsymbol{\eta}'(t)]\, dt \ ,$$

$$\mathbf{J}_{\theta\theta} = \int \boldsymbol{\theta}(t)\boldsymbol{\theta}'(t)\, dt \ .$$

- Minimizing the residual sum of squares with the two penalizations, we can get the estimate as

$$\text{vec}(\hat{\mathbf{B}}) = [\mathbf{J}_{\eta\eta} \otimes (\mathbf{Z}^{*'}\mathbf{Z}^*) + \lambda_s\mathbf{J}_{\eta\eta} \otimes \mathbf{R} + \lambda_t\mathbf{S} \otimes \mathbf{J}_{\theta\theta}]^{-1}(\mathbf{J}_{\theta\eta} \otimes \mathbf{Z}^{*'})\text{vec}(\mathbf{C}) \ ,$$

$$\mathbf{J}_{\theta\eta} = \int \boldsymbol{\theta}(t)\boldsymbol{\eta}'(t)\, dt$$

Agenda

4. Linear models

4.1. Functional Linear Models in Hilbert spaces

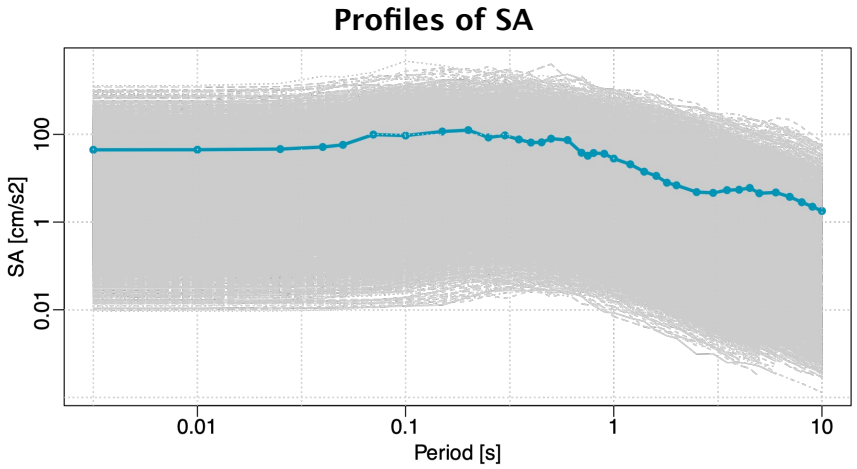
4.2. A case study on ground motion modelling

4.2. Examples

Toward a functional ground motion model for Italy

A **ground motion model** predicts the mean ground motion intensity measure at a site, conditionally on a given earthquake scenario.

Spectral acceleration (SA) is an intensity measure, defined over a range of vibration periods.



Benchmark: ITA18, scalar ground motion model for Italy (Lanzano et al. 2018).

$$\log_{10}(SA_j) = X_j\beta_j + \epsilon_j, j = 1, \dots, N.$$

Goal: extend ITA18 to a functional framework, i.e. formulate it as a **concurrent functional regression**, and estimate the functional coefficients.

4.2. Examples

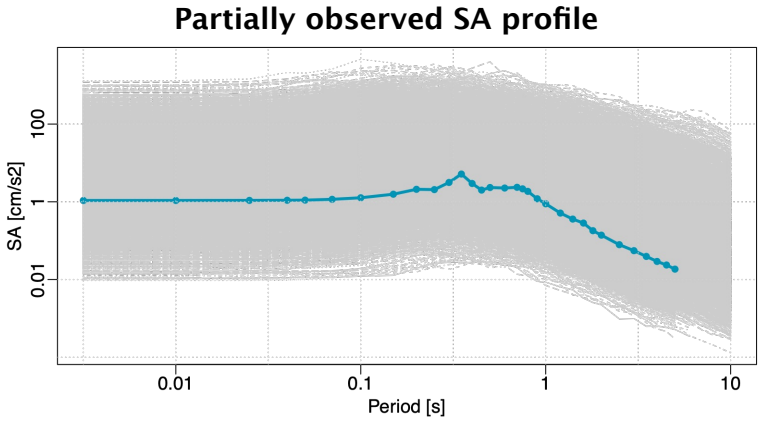
The peculiarity of spectral acceleration profiles

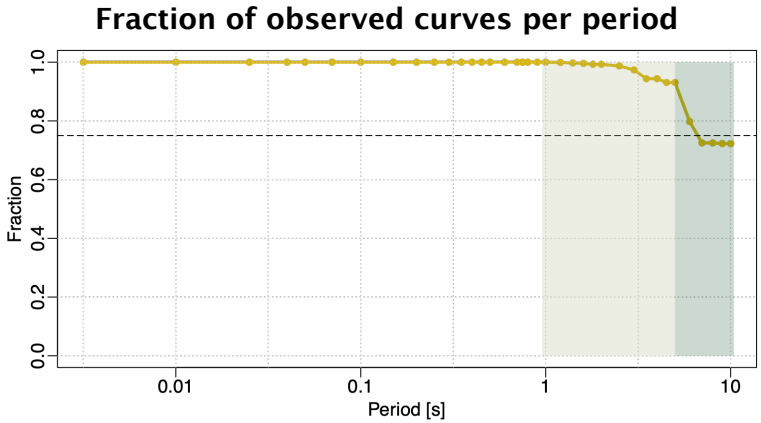
A non-negligible fraction of spectral acceleration profiles are only partially observed.

Methodological issue:
Handling partially observed response profiles.

Alternatives:

- Consider the restriction of the domain where the curves are observed
- Reconstruct the curves and consider the full domain (Kraus 2015; Kneip and Liebl 2020)
- Somewhere in between





4.2. Examples

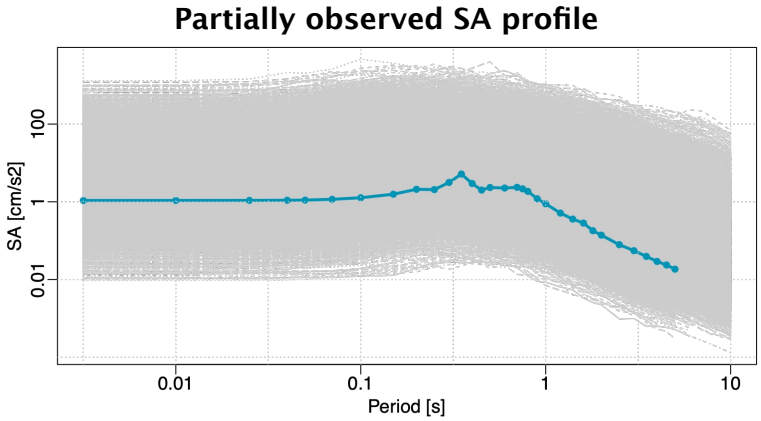
The peculiarity of spectral acceleration profiles

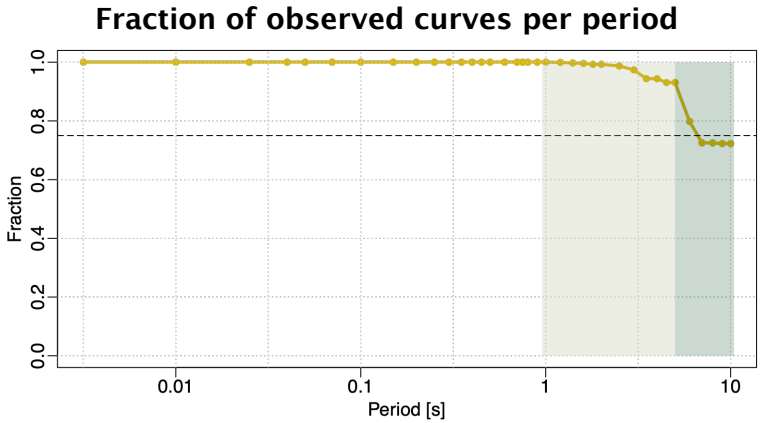
A non-negligible fraction of spectral acceleration profiles are only partially observed.

Methodological issue:
Handling partially observed response profiles.

Alternatives:

- Consider the restriction of the domain where the curves are observed
- Reconstruct the curves and consider the full domain (Kraus 2015; Kneip and Liebl 2020)
- Somewhere in between
  → Weighted functional data analysis





Let $y_1, \dots, y_n$ be reconstructed curves in $L^2(\mathcal{T})$. (Kraus 2015; Kneip and Liebl 2020)
Let $w: \mathcal{T} \to [0,1]$ be a weight.

Idea:
Couple each reconstructed curve y_i to a weight w_i, taking value 1 where the curve is observed and decreasing to zero the more the reconstruction becomes uncertain.

Analytical step:
Formulate new optimization criteria for functional regression, including the curve-specific functional weights, i.e. full weight will be given to errors made on the observed parts of the curves, and less weight to errors made on the reconstructed parts.

Concurrent linear regression

$$\boldsymbol{y}(t) = X(t)\boldsymbol{\beta}(t) + \boldsymbol{\epsilon}(t), t \in \mathcal{T}$$

Penalized weighted functional least-square criterion:
Minimize

$$\sum_{i=1}^{n} \int_{\mathcal{T}} w_i(s) \left(y_i(s) - \boldsymbol{x}_i(s)^T \boldsymbol{\beta}(s)\right)^2 ds + \sum_{j=1}^{q} \int_{\mathcal{T}} \lambda_j \left(D^2 \beta_j(s)\right)^2 ds.$$

Assumption: $\boldsymbol{\beta}(t) = \Theta(t)\boldsymbol{b}$.

Closed form solution for $\boldsymbol{b}$:

$$[J + R]\boldsymbol{b} = \int_{\mathcal{T}} \Theta(s)^T X(s)^T W(s)\boldsymbol{y}(s) ds,$$

where $J := \int_{\mathcal{T}} \Theta(s)^T X(s)^T W(s) X(s) \Theta(s) ds$, and R accounts for penalization.

In practice, only a finite number of longitudinal points of the curves are observed.
Assume to have $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n$ reconstructed: $\boldsymbol{y}_i = (y_i(t_1), \dots, y_i(t_N))$.

For each $\boldsymbol{y}_i$, we look for $f_i \in H^2(\mathcal{T})$ such that $y_i(t_j) = f_i(t_j) + e_{ij}$.

Penalized weighted least-square criterion

$$\hat{f}_i = \underset{f \in H^2(\mathcal{T})}{\operatorname{argmin}} \sum_{j=1}^{T} v_{ij}\left(y_i(t_j) - f(t_j)\right)^2 + \zeta ||D^2 f||^2_{L^2(\mathcal{T})}.$$

Solution in matrix form: $\hat{\boldsymbol{f}}(t) = C\boldsymbol{\phi}(t)$.

Closed form solution for $\boldsymbol{b}$:

$$[J + R]\boldsymbol{b} = \int_{\mathcal{T}} \Theta(s)^T X(s)^T W(s) C\boldsymbol{\phi}(s) ds.$$

Note: The weights for smoothing and regression could be different. In our work we set $v_{ij} = w_i(t_j)$.

We consider two possible systems of weights:

1. Logistic

Let y_i be observed up to t_i.

$$w_i(t) = \mathbf{1}_{(t \le t_i)} + \left(\frac{1}{1 + e^{(t-\mu_i)\alpha_i}} + c_i \right) \mathbf{1}_{(t > t_i)},$$

where $\mu_i = (t_i + t_N)/2$, and $\alpha_i = a\hat{\sigma}_{t_i}$.

2. Reconstruction driven (inspired by Kraus 2015)

Let y_i be observed on O_i and reconstructed on M_i.

$$w_i(t) = \mathbf{1}_{O_i}(t) + \left(1 - \sqrt{\frac{\hat{v}_i(t)}{\hat{c}(t)}} \right) \mathbf{1}_{M_i}(t),$$

- v_i is the diagonal of the kernel of $\mathcal{V}_i$, covariance operator of the reconstruction error,
- c is the diagonal of the kernel of $\mathcal{C}$, covariance operator of y_i,
- $\hat{v}_i$ and $\hat{c}$ empirical counterparts of v_i and c.

Note: The weights for smoothing and regression could be different. In our work we set $v_{ij} = w_i(t_j)$.

We consider two possible systems of weights:

1. Logistic SPECIFIC TO THE PROBLEM AT HAND

Let y_i be observed up to t_i.

$$w_i(t) = \mathbf{1}_{(t \le t_i)} + \left(\frac{1}{1 + e^{(t-\mu_i)\alpha_i}} + c_i \right) \mathbf{1}_{(t > t_i)},$$

where $\mu_i = (t_i + t_N)/2$, and $\alpha_i = a\hat{\sigma}_{t_i}$.

2. Reconstruction driven (inspired by Kraus 2015) GENERAL

Let y_i be observed on O_i and reconstructed on M_i.

$$w_i(t) = \mathbf{1}_{O_i}(t) + \left(1 - \sqrt{\frac{\hat{v}_i(t)}{\hat{c}(t)}} \right) \mathbf{1}_{M_i}(t),$$

- v_i is the diagonal of the kernel of $\mathcal{V}_i$, covariance operator of the reconstruction error,
- c is the diagonal of the kernel of $\mathcal{C}$, covariance operator of y_i,
- $\hat{v}_i$ and $\hat{c}$ empirical counterparts of v_i and c.

Simulation:

$$y_i(t_j) = \beta_0(t_j) + \beta_1(t_j)x_{1i} + \beta_2(t_j)x_{2i}(t_j) + \epsilon_{ij},$$

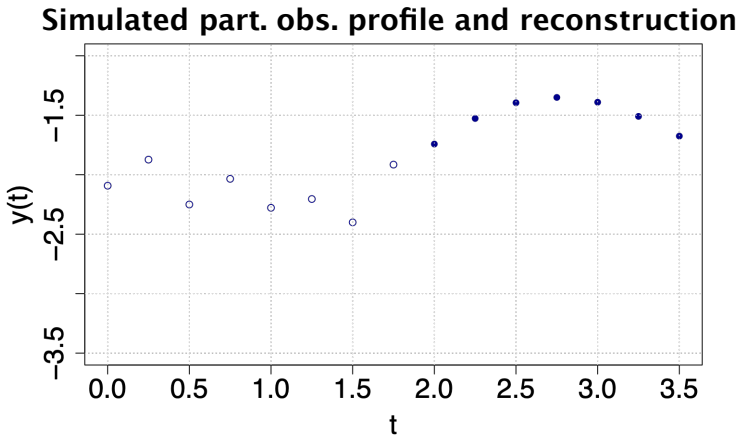
where ϵ_{ij} are zero-mean iid errors.

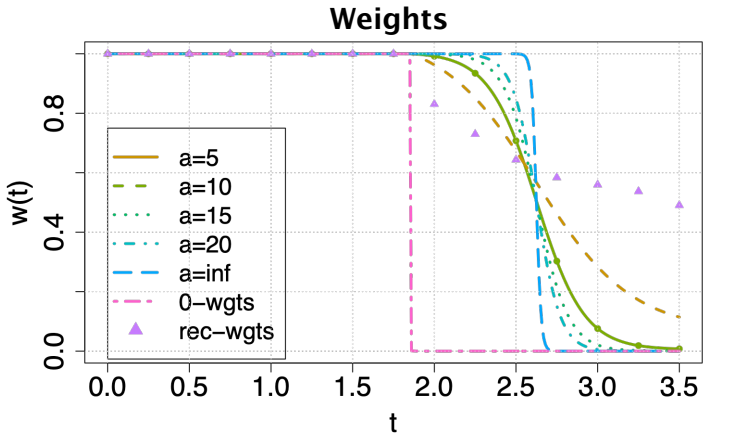
Partially observed longitudinal observations are generated by coupling a fraction of the curves with curve-specific observed domains.

Goal:

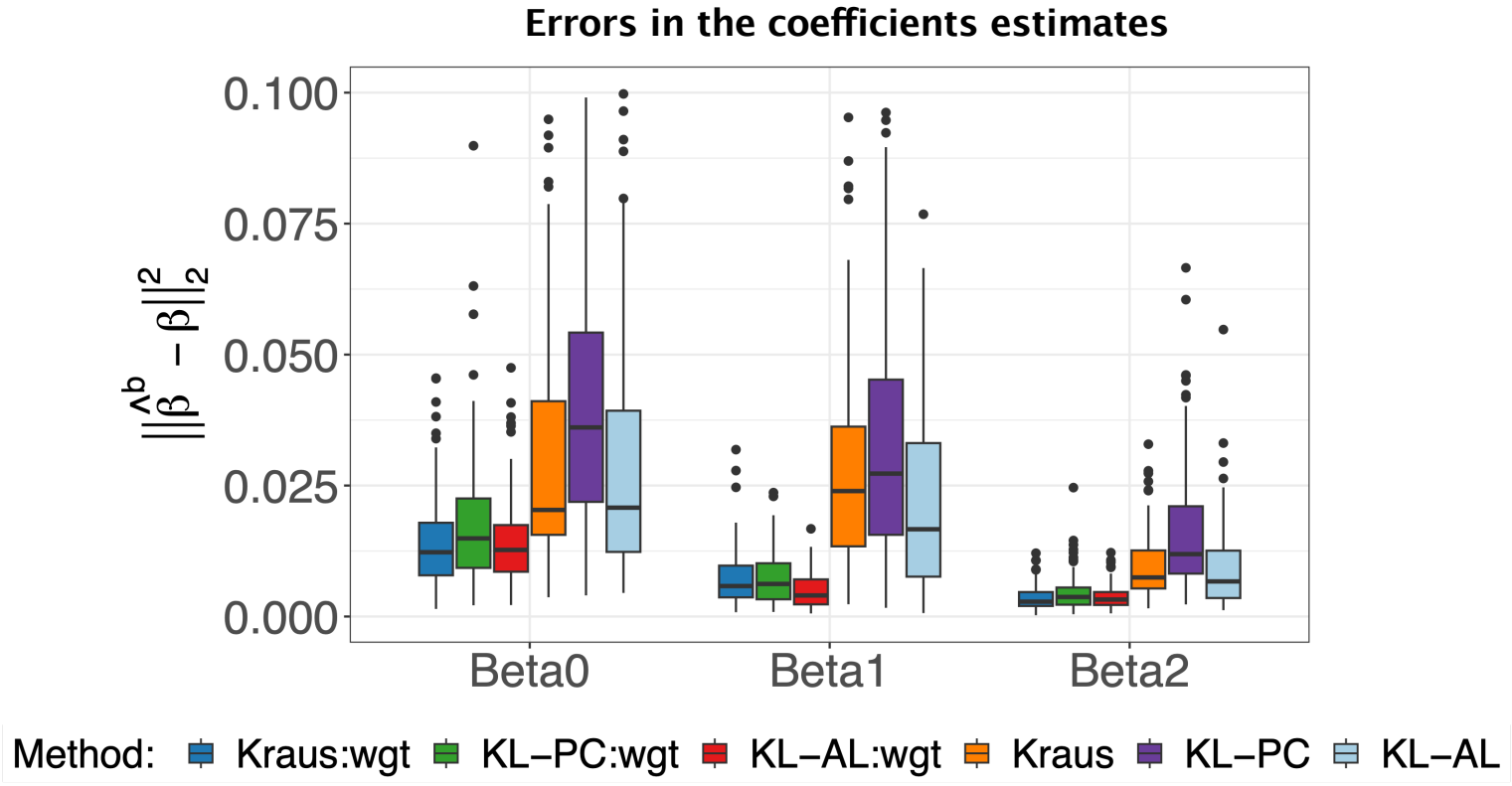
Validation of the weighted methodology on two main aspects:

1. Robustness of the **coefficients estimates** with respect to the reconstruction method,
2. Accuracy of the **coefficients estimates** and of the **predictions** as the definition of the weights changes.

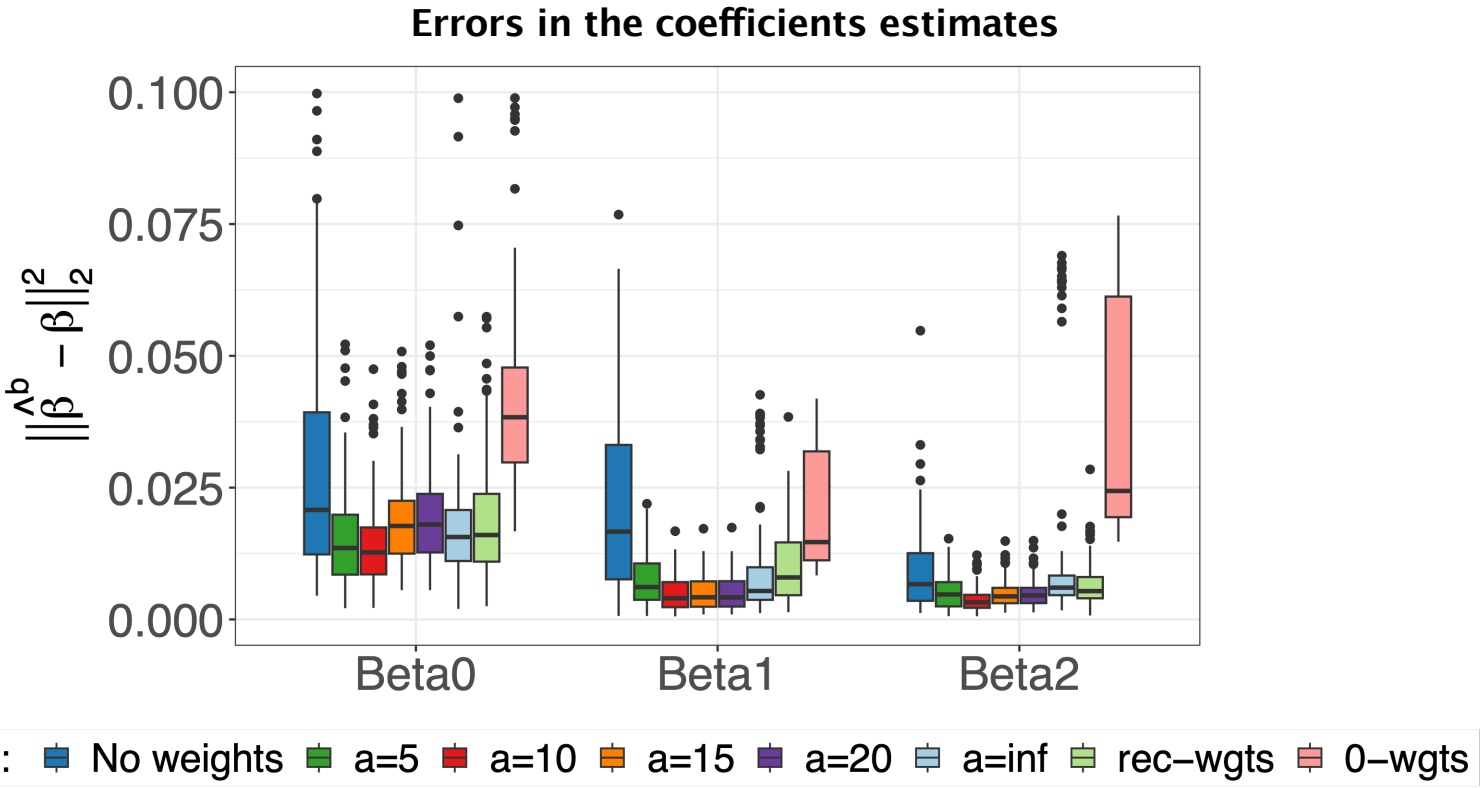




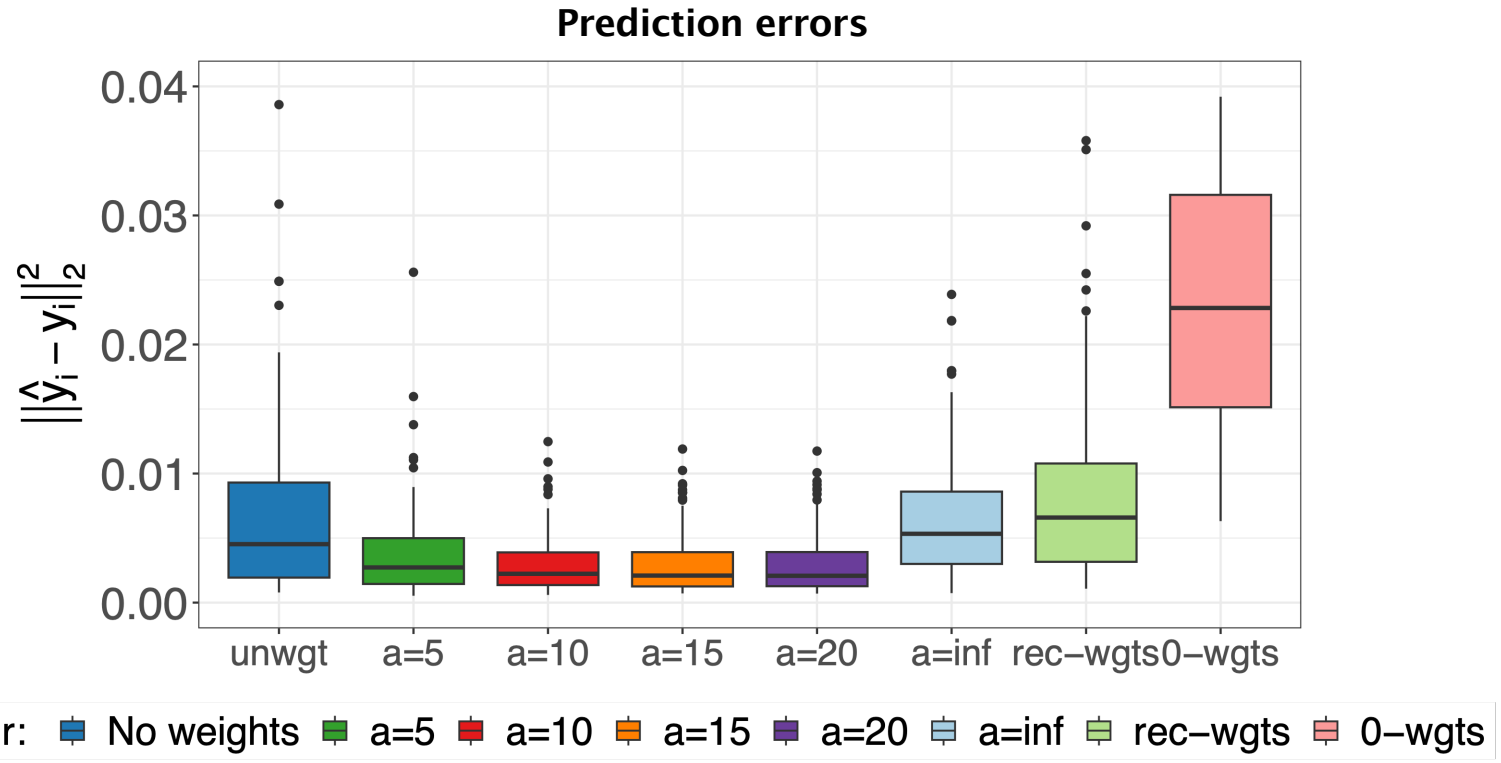
Monte Carlo simulation with B=100 repetitions.



Monte Carlo simulation with B=100 repetitions.



LOO cross-validation on a single simulated sample of 100 observations.



Data:
Parametric table of the ITA18 GMM for PGA, PGV and Spectral Acceleration ordinates. Istituto Nazionale di Geofisica e Vulcanologia. (Lanzano et al. 2022)

$$\log_{10} SA = \alpha + b_1 (M_w - M_h)\mathbf{1}_{(M_w \leq M_h)} + b_2 (M_w - M_h)\mathbf{1}_{(M_w > M_h)} + f_1 SoF_1 + f_2 SoF_2 + c_1 (M_w - M_{ref}) \log_{10} R + c_2 \log_{10} R + c_3 R + k \log_{10} \frac{V_0}{800} + \epsilon.$$

Critical steps:

1. Calibration of the ridge penalization parameters $\lambda_1, \dots, \lambda_9$ (Centofanti et al. 2023),
2. Choice of the weights,
3. Choice of the reconstruction method,
4. Estimate of the functional regression coefficients

4.2. Examples

Case study: Estimates of regression coefficients

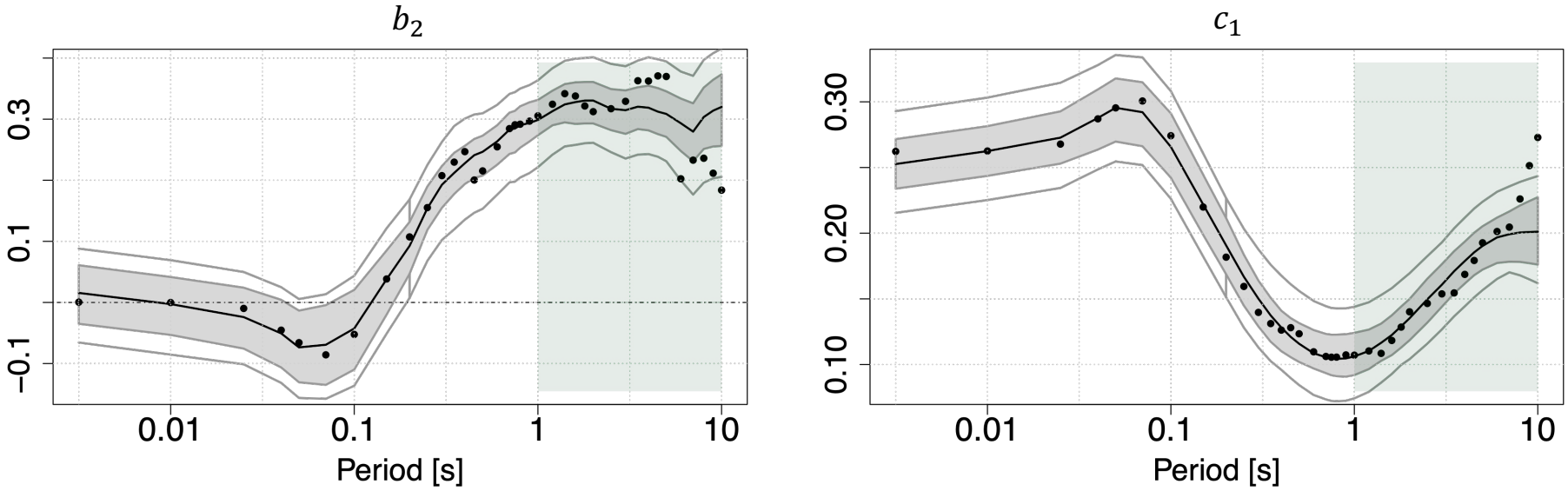


Figure: Functional boxplots of the bootstrap sample of the estimated functional regression coefficients and comparison with the ITA18 estimates (black dots). The black lines represent the point estimates of the coefficients. The bootstrap sample has size 1000.

Final remarks

Conclusions:

- The analysis of actual data often requires the development of ad hoc extensions. In this case, novel approach to the analysis of partially observed functional data
- The method proves effective in
  1. reducing the impact of the method adopted for reconstruction,
  2. improving the stability of the regression coefficients estimates,
- The optimal solution to the functional regression problem is found in the middle between considering solely the reduced domain or the entire domain where data are reconstructed (Stefanucci et al. 2018).

Analyses here presented can be replicated using the code available at
https://github.com/tbortolotti/WFDA.git

Full description and details in
Bortolotti, T., Peli, R., Lanzano, G., Sgobba, S., & Menafoglio, A. (2024). Weighted Functional Data Analysis for the Calibration of a Ground Motion Model in Italy. Journal of the American Statistical Association, 1–12. https://doi.org/10.1080/01621459.2023.2300506