Function-on-Function Mixture of Experts Regression Models

Speaker: Jean Steve TAMO TCHOMGUI[1,2]

work supervised by: Vincent BARRIAC[1], Guillaume FRAYSSE[1], Stéphane CHRETIEN[2], Julien JACQUES[2]

December 18th, 2022

[1] Orange Innovation, Paris France. [2] Univ Lyon, Univ Lyon 2, ERIC, Lyon France.

CFE-CMStatistics 2022, London

Outline

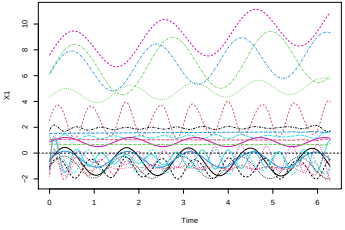
Motivations

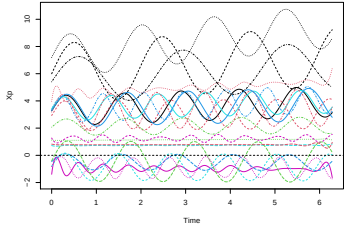
- My PhD is entitled ”Predicting the Quality of Experience of 5G verticals” from networks metrics;
- The shape of the available data led us to the Functional Data Analysis Framework;
- It further led us to the particular case of function-on-function regression models;
- With heterogeneity in the data, the latent class model is well adapted;
- For predictive modeling, the Mixture of Experts (MoE) is a more convenient architecture.

Data

Data ($1 \leq i \leq n$):

- Inputs: $\mathrm{X}_i(t) = \left(\mathrm{X}_i^1(t), \ldots, \mathrm{X}_i^p(t)\right)$
- Output: $\mathrm{Y}_i(t)$





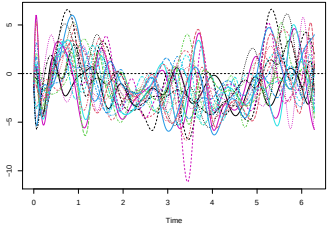


Table of Contents

Models for homogeneous data set

We consider the following models, Ramsay and Silverman (2005):

- **Concurrent model:**

$$\mathrm{Y}_i(t) \quad = \quad \beta_0(t) + \sum_{l=1}^{p} \beta_l(t)\, \mathrm{X}_i^l(t) + \varepsilon_i(t) \; = \beta(t)^\top \mathrm{X}_i(t) + \varepsilon_i(t), \tag{1}$$

Models for homogeneous data set

We consider the following models, Ramsay and Silverman (2005):

- **Concurrent model:**

$$\mathrm{Y}_i(t) \;=\; \beta_0(t) + \sum_{l=1}^{p} \beta_l(t)\,\mathrm{X}_i^l(t) + \varepsilon_i(t) \;= \beta(t)^\top \mathrm{X}_i(t) + \varepsilon_i(t), \tag{1}$$

- **(Feed forward) integral model:**

$$\mathrm{Y}_i(t) \;=\; \gamma_0(t) + \sum_{l=1}^{p} \int_0^t \gamma_l(s,t)\,\mathrm{X}_i^l(s)\,\mathrm{d}s + \varepsilon_i(t) \;=\; \gamma_0(t) + \int_0^t \gamma(s,t)^\top \mathrm{X}_i(s)\,\mathrm{d}s + \varepsilon_i(t). \tag{2}$$

Mixture of Experts (MoE), Jacobs et al. (1991)

- **MoE** is Flexible supervised learning architecture dedicated for complex relationship;

Mixture of Experts (MoE), Jacobs et al. (1991)

- **MoE** is Flexible supervised learning architecture dedicated for complex relationship;
- The MoE regression model with functional response involving functional covariates in a K-component mixture is defined by:

$$f\Big(\mathrm{Y}_i(t)\,|\,\mathrm{X}_i(t),\ t\in\mathrm{T},\ \Psi\Big) \quad = \quad \sum_{k=1}^{\mathrm{K}} \underbrace{\pi_k\big(\mathrm{X}_i(t),\ t\in\mathrm{T},\ \alpha_k\big)}_{\text{component mixture weight}} \underbrace{f_k\big(\mathrm{Y}_i(t)\,|\,\mathrm{X}_i(t),\ \theta_k\big)}_{\text{Local expert } k} \tag{3}$$

Component mixture weights

- Model for the gated network function based on a functional multinomial logistic model. Dayton and Macready (1988), Chamroukhi et al. (2022):

$$\begin{aligned}\pi_k\big(\mathrm{X}_i(t),\ t\in\mathrm{T},\ \alpha_k\big) &= \mathbb{P}\Big(\mathrm{Z}=k\,|\,\mathrm{X}_i(t),\ t\in\mathrm{T},\ \alpha_k\Big)\\ &= \frac{\exp\Big(h_k\big(\mathrm{X_i}(t)\ t\in\mathrm{T},\ \alpha_k\big)\Big)}{1+\sum_{k'=1}^{\mathrm{K}-1}\exp\Big(h_{k'}\big(\mathrm{X_i}(t)\ t\in\mathrm{T},\ \alpha_{k'}\big)\Big)}\end{aligned} \tag{4}$$

Where $h_k\big(\mathrm{X}_i(t),\ t\in\mathrm{T},\ \alpha_k\big)=\alpha_{k,0}+\int_{\mathrm{T}}\alpha_k(t)\,\mathrm{X}_i(t)\,dt$

Locals experts

We use for local models the functional linear model **(1)** defined in an expert k by:

$$\mathrm{Y}_i(t) = \beta_{0,k}(t) + \sum_{l=1}^{p} \beta_{l,k}(t)\, \mathrm{X}_i^l(t) + \varepsilon_i(t) = \beta_k(t)^\top \mathrm{X}_i(t) + \varepsilon_i(t) \quad \text{if } i \in \text{Group}_k$$

$$\text{with } \beta_k(t) = \begin{pmatrix} \beta_{0,k}(t) & \beta_{1,k}(t) & \dots & \beta_{p,k}(t) \end{pmatrix}^\top, \mathrm{X}_i(t) = \begin{pmatrix} 1 & \mathrm{X}_i^1(t) & \dots & \mathrm{X}_i^p(t) \end{pmatrix}^\top.$$

We expand in functional basis the functional covariates and parameters:

$$\mathrm{x}_i^l(t) = \sum_{j=1}^{\mathrm{L}_{\mathrm{x}^l}} x_{ij}^l\, \mathrm{B}_j^l(t) = \mathrm{B}^l(t)^\top x_i^l \text{ and } \beta_{l,k}(t) = \sum_{j=1}^{\mathrm{L}_{\beta^l}} b_{j,k}^l\, \phi_j^l(t) = \phi^l(t)^\top b_k^l$$

The model can be written with this expression as linear mixed model:

$$\mathrm{Y}_i(t) \quad = \quad b_k^\top \underbrace{\Phi(t)^\top \mathrm{B}(t)\, x_i}_{\mathrm{R}_i(t)} + \varepsilon_i(t) \tag{5}$$

Table of Contents

Estimation via the EM algorithm, Dempster et al. (1977)

We combine models (4) and (5) in K components mixture as:

$$f\Big(\mathrm{Y}_i(t)\,|\,\mathrm{X}_i(t),\ t\in\mathrm{T},\ \Psi\Big)=\sum_{k=1}^{\mathrm{K}}\frac{\exp\Big(\alpha_{k,0}+\delta_k^\top r_i\Big)}{1+\sum_{k'=1}^{\mathrm{K}-1}\exp\Big(\alpha_{k',0}+\delta_{k'}^\top r_i\Big)}\,\Phi\Big(\mathrm{Y}_i(t)\,;\,b_k^\top \mathrm{R}_i(t),\sigma_k^2\Big);\qquad(6)$$

with $\Psi=\Big((\alpha_{1,0},\delta_1,b_1,\sigma_1^2),\ \ldots,\ (\alpha_{\mathrm{K},0},\delta_{\mathrm{K}},b_{\mathrm{K}},\sigma_{\mathrm{K}}^2)\Big)$ the parameters vector to estimate.

EM algorithm

- **E-step:** Given the current parameter Ψ^{old}, compute here the posterior probability that the individual i belongs to the class k:

$$\widehat{p}_{ik} = \frac{\pi_k \left(\mathrm{X}_i(t),\ t \in \mathrm{T},\ \alpha_k^{\text{old}}\right) \Phi\left(\mathrm{Y}_i(t)\,;\, b_k^{\text{old}^\top} \mathrm{R}_i(t), \sigma_k^{2^{\text{old}}}\right)}{\sum_{k'=1}^{\mathrm{K}} \pi_{k'} \left(\mathrm{X}_i(t),\ t \in \mathrm{T},\ \alpha_{k'}^{\text{old}}\right) \Phi\left(\mathrm{Y}_i(t)\,;\, b_{k'}^{\text{old}^\top} \mathrm{R}_i(t), \sigma_{k'}^{\text{old}^2}\right)}$$

EM algorithm

- **E-step:** Given the current parameter Ψ^{old}, compute here the posterior probability that the individual i belongs to the class k:

$$\widehat{p}_{ik} = \frac{\pi_k\left(\mathrm{X}_i(t),\ t \in \mathrm{T},\ \alpha_k^{\text{old}}\right) \Phi\left(\mathrm{Y}_i(t)\,;\, b_k^{\text{old}^\top} \mathrm{R}_i(t), \sigma_k^{2^{\text{old}}}\right)}{\sum_{k'=1}^{\mathrm{K}} \pi_{k'}\left(\mathrm{X}_i(t),\ t \in \mathrm{T},\ \alpha_{k'}^{\text{old}}\right) \Phi\left(\mathrm{Y}_i(t)\,;\, b_{k'}^{\text{old}^\top} \mathrm{R}_i(t), \sigma_{k'}^{2^{\text{old}}}\right)}$$

- **M-step:** Given the posterior probabilities, we update the parameters by maximizing:

$$\mathrm{Q}(\Psi^{\text{new}} \mid \Psi^{\text{old}}) = \mathrm{Q}_1(b^{\text{new}}, \sigma^{2^{\text{new}}} \mid \Psi^{\text{old}}) + \mathrm{Q}_2(\delta^{\text{new}} \mid \Psi^{\text{old}})$$

Where:

$$\mathrm{Q}_1(b^{\text{new}}, \sigma^{2^{\text{new}}} \mid \Psi^{\text{old}}) = \sum_{i=1}^{N} \sum_{k=1}^{\mathrm{K}} \widehat{p}_{ik} \log\left(\Phi\left(\mathrm{Y}_i(t)\,;\, b_k^{\text{new}^\top} \mathrm{R}_i(t), \sigma_k^{2^{\text{new}}}\right)\right)$$

$$\mathrm{Q}_2(\delta^{\text{new}} \mid \Psi^{\text{old}}) = \sum_{i=1}^{N} \sum_{k=1}^{\mathrm{K}} \widehat{p}_{ik} \log\left(\pi_k\left(\mathrm{X}_i(t),\ t \in \mathrm{T},\ \alpha_k^{\text{new}}\right)\right)$$

EM algorithm

- Q_1 and Q_2 can be maximized separately using weighted ML estimation of GLMs and multinomial logistic model respectively;
- The algorithm is stopped if the relative change in the log-likelihood is smaller than a pre-specified ϵ;
- We repeat the EM algorithm with different initializations and choose the one with the maximum likelihood;
- Implementations details can be seen in Grün and Leisch (2008).

Model selection:

- The number K of mixture components can be chosen using BIC.

Table of Contents

Simulation study

- We simulate two data sets with $n = 500$ and $n = 1000$ each with $p = 5$ functional predictors defined on $[0, 2\pi]$ given by the equation:

$$\begin{cases} \mathrm{U}_i^l(t_j) = & \xi_{i,1}^l + \Big(\log(10 + t_j)\Big)^{\xi_{i,2}^l} + \xi_{i,3}^l \, \sin\left(\frac{2\pi t_j}{\xi_{i,4}^l}\right) \\ & \xi_{i,r}^l \qquad \text{constants drawn from } \mathcal{U}\big([\text{-}1, 1]\big), \; 1 \leq r \leq 4 \end{cases}$$

- And the functional response by the concurrent model with the formula:

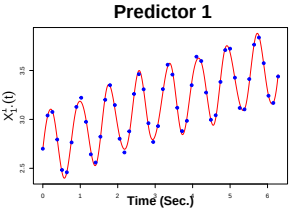
$$\mathrm{Y}_i(t_j) \quad = \quad \beta_{0,k}(t_j) + \sum_{l=1}^{p} \beta_{l,k}(t_j) \, \mathrm{X}_i^l(t_j) + \varepsilon_{ij} \qquad \text{if } i \in \mathrm{Group}_k,$$

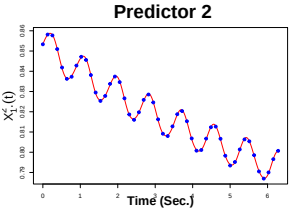
with $\begin{cases} \beta_{0,k}(t_j) & = \Big(\log(10 + t_j)\Big)^{\rho_{0,k}} \\ \beta_{l,k}(t_j) & = \rho_1^{l,k} \, \sin\left(\frac{2\pi t_j}{\rho_2^{l,k}}\right) \end{cases}$ where $\begin{cases} \varepsilon_{ij} & \sim \mathcal{N}(0, 4)\,, \\ \rho_{0,k},\ \rho_1^{l,k},\ \rho_2^{l,k} & \text{are chosen constants.} \end{cases}$

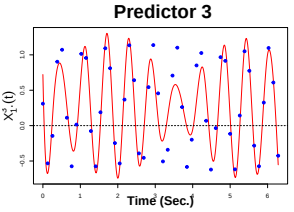
- $\mathrm{K} = 4$ groups, so we have $4 \times 6 = 24$ parameters ; In each scenario, we run $N = 50$ experiences.

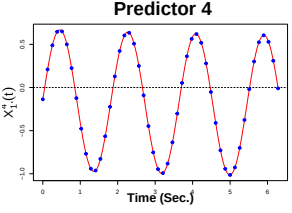
Functional predictors and functional concurrent response for a randomly chosen individual:

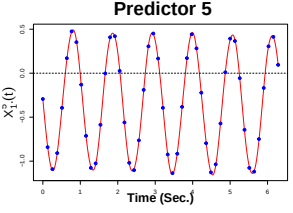
The blue dots are the observed discrete data and the red curves is the functional reconstructions of predictors

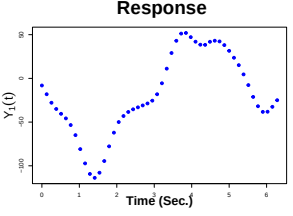




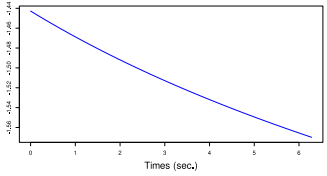




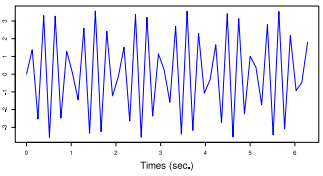




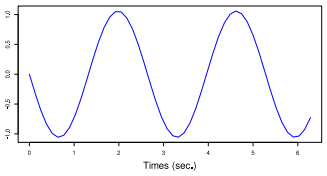
Functional parameters for the component (expert) 1:



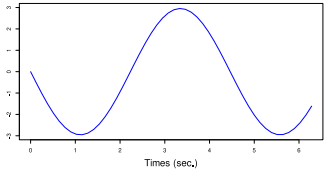
$\beta_{0,1}(t)$



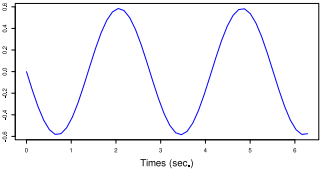
$\beta_{1,1}(t)$



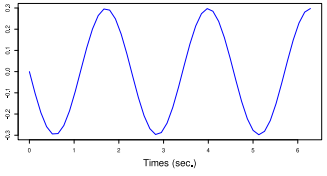
$\beta_{2,1}(t)$



$\beta_{3,1}(t)$



$\beta_{4,1}(t)$



$\beta_{5,1}(t)$

Assessment of quality of estimation

- **Choice of** K: Ratio of models with the right number of components ;
- **Estimation of the** $\beta_l(t)$: The accuracy of the estimated parameters of models that have the good number of components via the Root Mean Square Error (RMSE):

$$\text{RMSE}\Big(\beta_l(.)\Big) = \left[\frac{1}{m}\sum_{j=1}^{m}\Big(\beta_l(t_j) - \widehat{\beta}_l(t_j)\Big)^2\right]^{1/2}.$$

- **Prediction quality:** via the Mean Relative Prediction Error (MRPE) given by:

$$\text{MRPE} = \frac{1}{m}\sum_{j=1}^{m}\left(\frac{\sum_{i=1}^{n}\Big(\text{Y}_i(t_j) - \widehat{\text{Y}}_i(t_j)\Big)^2}{\sum_{i=1}^{n}\text{Y}_i(t_j)^2}\right).$$

Choice of K

We evaluate the efficiency of BIC for choosing K:

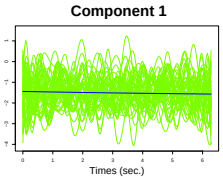
- For each sample, different values for $K \in \{2, \ldots, 6\}$ has been tested, and BIC is used to select the best one;
- **Ratio of models with the right number of components:**

Number of classes		3	**4**	5
% of models	n.train = 500	6%	**86%**	8%
	n.train = 1000	0%	**92%**	8%

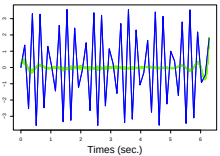
Table: Results of our $N = 50$ simulations in terms of number of components.

Estimation of the of the $\beta_l(t)$

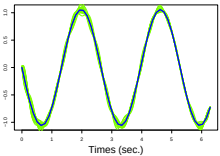
Figure: Estimated (green) and actual (blue) parameters for our simulations with $n_{train} = 500$.



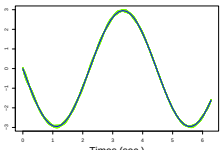
$\beta_{0,1}(t)$

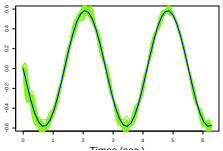


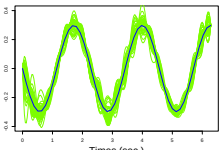
$\beta_{1,1}(t)$



$\beta_{2,1}(t)$

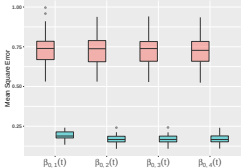




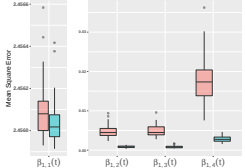


Estimation of the $\beta_l(t)$

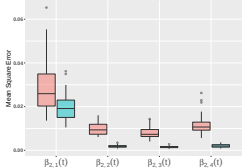
Figure: Boxplot of RMSE in all the estimated parameters for our two scenarios : $n = 500$ (red) and $n = 1000$ (blue).



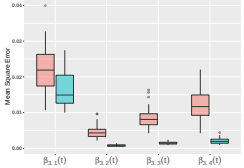
$\beta_0(t)$

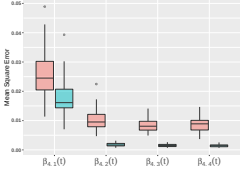


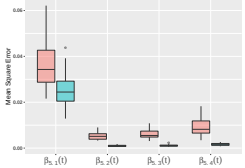
$\beta_1(t)$



$\beta_2(t)$

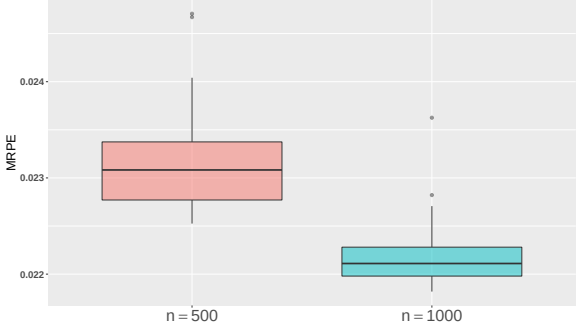




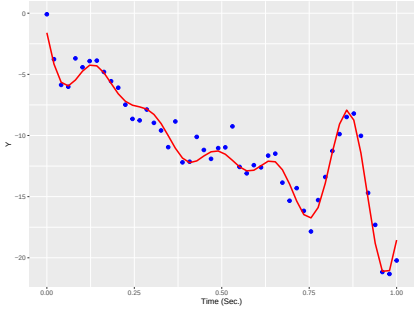


Prediction quality : MRPE

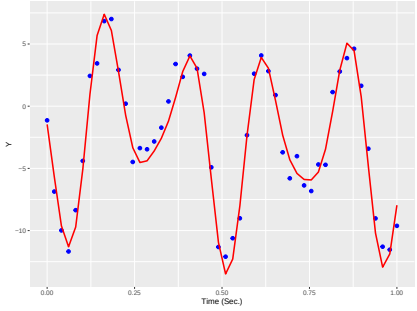
We evaluate here the prediction error on a test sample of length $n_{test} = 2000$ through the MRPE :



Prediction Quality : some plots



(a) prediction 1



(b) prediction 2

Figure: Prediction and actual values of the functional response for two randomly chosen individuals.

Conclusion and Future works

- MoE is a flexible architecture that allows to efficiently represents complex relationship ;
- EM algorithm designed here for functional data can be seen as extension vectorized version of MoE ;

Conclusion and Future works

- MoE is a flexible architecture that allows to efficiently represents complex relationship ;
- EM algorithm designed here for functional data can be seen as extension vectorized version of MoE ;
- With the functional basis expansion of our parameters, we must enhance both the local expert and gated network estimated parameters by applied a roughness penalty to get them interpretable :

$$\mathrm{Pen}(\beta_l) = \lambda_l \int \beta_l^{''}(t)^2 \, dt = \lambda_l \int \left[\sum_{j=1}^{\mathrm{L}_{\beta^l}} b_j^l \phi_j^{l"}(t) \right]^2 dt = \lambda_l \sum_{s,k=1}^{\mathrm{L}_{\beta^l}} b_s^l b_k^l \Phi_{sk}^l$$

$$\text{with } \Phi_{sk}^l = \int \phi_s^{l"}(t) \, \phi_k^{l"}(t) \, dt.$$

Some references

F. Chamroukhi, N. T. Pham, V. H. Hoang, and G. J. McLachlan, "Functional mixtures-of-experts," arXiv preprint arXiv:2202.02249, 2022.

C. M. Dayton and G. B. Macready, "Concomitant-variable latent-class models," Journal of the american statistical association, vol. 83, no. 401, pp. 173–178, 1988.

A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," Journal of the Royal Statistical Society: Series B (Methodological), vol. 39, no. 1, pp. 1–22, 1977.

B. Grün and F. Leisch, "Flexmix version 2: Finite mixtures with concomitant variables and varying and constant parameters," Journal of Statistical Software, vol. 28, no. 4, p. 1–35, 2008. [Online]. Available: https://www.jstatsoft.org/index.php/jss/article/view/v028i04

R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive Mixtures of Local Experts," Neural Computation, vol. 3, no. 1, pp. 79–87, 03 1991. [Online]. Available: https://doi.org/10.1162/neco.1991.3.1.79

J. Ramsay and B. W. Silverman, Functional Data Analysis, 2nd ed., ser. Springer Series in Statistics. New York: Springer-Verlag, 2005. [Online]. Available: https://www.springer.com/gp/book/9780387400808

Questions?

Appendix - Recover the functional nature of the predictors

- In the particular case of the cubic spline model, the expression in functional basis $\{\mathrm{B}_l(t)\}_{l\geq 1}$ truncated at q becomes $\left\{1,\, t,\, t^2,\, t^3,\, (t-\tau_1)_+^3,\, \ldots,\, (t-\tau_{q-4})_+^3\right\}$ and then

$$\mathrm{x}_i^l(t) \quad = \quad x_{i1}^l + x_{i2}^l\, t + x_{i3}^l\, t^2 + x_{i4}^l\, t^3 + \sum_{l=1}^{q-4} x_{i(4+l)}^l\, (t-\tau_l)_+^3.$$

$$\begin{pmatrix} \mathrm{x}_i(t_{i1}) \\ \mathrm{x}_i(t_{i2}) \\ \vdots \\ \mathrm{x}_i(t_{im_i}) \end{pmatrix} = \begin{pmatrix} x_{i1} + x_{i2}\, t_{i1} + x_{i3}\, t_{i1}^2 + x_{i4}\, t_{i1}^3 + \sum_{l=1}^{q-4} x_{i(4+l)}\, (t_{i1}-\tau_l)_+^3 \\ x_{i1} + x_{i2}\, t_{i2} + x_{i3}\, t_{i2}^2 + x_{i4}\, t_{i2}^3 + \sum_{l=1}^{q-4} x_{i(4+l)}\, (t_{i2}-\tau_l)_+^3 \\ \vdots \\ x_{i1} + x_{i2}\, t_{im_i} + x_{i3}\, t_{im_i}^2 + x_{i4}\, t_{im_i}^3 + \sum_{l=1}^{q-4} x_{i(4+l)}\, (t_{im_i}-\tau_l)_+^3 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & t_{i1} & t_{i1}^2 & t_{i1}^3 & (t_{i1}-\tau_1)_+^3 & (t_{i1}-\tau_2)_+^3 & \ldots & (t_{i1}-\tau_{q-4})_+^3 \\ 1 & t_{i2} & t_{i2}^2 & t_{i2}^3 & (t_{i2}-\tau_1)_+^3 & (t_{i2}-\tau_2)_+^3 & \ldots & (t_{i2}-\tau_{q-4})_+^3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_{im_i} & t_{im_i}^2 & t_{im_i}^3 & (t_{im_i}-\tau_1)_+^3 & (t_{im_i}-\tau_2)_+^3 & \ldots & (t_{im_i}-\tau_{q-4})_+^3 \end{pmatrix} \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iq} \end{pmatrix}$$

$$\mathrm{x}_i = \mathcal{B}_i\, x_i.$$

It is recommended (Ruppert, 2002) to choose q large enough to capture the complexity of the predictor and then apply a roughness penalty to limit overfitting of the form:

$$\textbf{(C1)} \max_{5\le l\le q} |x_{il}| < \mathrm{C} \qquad \textbf{(C2)} \sum_{l=5}^{q} |x_{il}| < \mathrm{C} \qquad \textbf{(C3)} \sum_{l=5}^{q} x_{il}^2 < \mathrm{C}.$$

$$\begin{cases} \min\limits_{x_i} \|\mathrm{x}_i - \mathcal{B}_i\, x_i\|^2 \\ \text{s.c. } x_i^\top\, \mathrm{D}\, x_i \le \mathrm{C} \end{cases} \iff \min_{x_i} \|\mathrm{x}_i - \mathcal{B}_i\, x_i\|^2 + \lambda\, x_i^\top\, \mathrm{D}\, x_i \qquad (\lambda \ge 0)$$

with $\mathrm{D} = \begin{bmatrix} 0_{4\times 4} & 0_{4\times(q-4)} \\ 0_{(q-4)\times 4} & \mathrm{I}_{(q-4)\times(q-4)} \end{bmatrix}$

Appendix - Penalizing the model

- The estimation of functional parameters $\beta(t)$ expressed in functional basis depends on the number of basis functions L_β;
- Unfortunately, the functional basis based fitting procedure lead to a non regular estimate curve $\hat{\beta}(t)$ and then hard to interpret;
- To correct this non-regularity effect, we enhance the model by introducing a Ridge roughness penalty in the model.

In this case, the penalized least squares risk to be minimized is given by:

$$\mathcal{R}(b) \quad = \quad \sum_{i=1}^{n} \sum_{j=1}^{m} \left(\mathrm{Y}_i(t_j) - b^\top \mathrm{R}_i(t_j) \right)^2 + \sum_{l=0}^{p} \mathrm{Pen}(\beta_l), \tag{7}$$

where $\mathrm{Pen}(\beta_l) = \lambda_l \int \beta_l^{''}(t)^2 \, dt = \lambda_l \int \left[\sum_{j=1}^{\mathrm{L}_{\beta^l}} b_j^l \phi_j^{l"}(t) \right]^2 dt = \lambda_l \sum_{s,k=1}^{\mathrm{L}_{\beta^l}} b_s^l b_k^l \Phi_{sk}^l$ with $\Phi_{sk}^l = \int \phi_s^{l"}(t) \, \phi_k^{l"}(t) \, dt.$

Appendix - Penalizing the model

Return to presentation

Then for a fixed value λ_l, the estimation of $\left(\beta_l(t)\right)_{0\leq l\leq p}$ is calculated by solving the problem:

$$\begin{aligned}\min_b \mathcal{R}(b) &= \min_b \sum_{i=1}^{n}\sum_{j=1}^{m}\left(\mathrm{Y}_i(t_j) - b^\top \mathrm{R}_i(t_j)\right)^2 + \sum_{l=0}^{p}\lambda_l \sum_{s,k=1}^{\mathrm{L}_{\beta^l}} b_s^l b_k^l \Phi_{sk}^l \\ &= \min_b \left(\mathrm{Y}-\mathrm{R}b\right)^\top\left(\mathrm{Y}-\mathrm{R}b\right) + b^\top(\lambda\mathrm{P})\, b\,,\end{aligned}$$

where $\lambda\mathrm{P}$ the matrix of dimension $\mathrm{L}_\beta \times \mathrm{L}_\beta$ given by:

$$\lambda\mathrm{P} = \begin{pmatrix} \lambda_0\Psi^0 & \mathbb{0} & \dots & \mathbb{0} \\ \mathbb{0} & \lambda_1\Psi^1 & \dots & \mathbb{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{0} & \mathbb{0} & \dots & \lambda_p\Psi^p \end{pmatrix} \quad \text{with } \Psi^l = \begin{pmatrix} \Phi_{11}^l & \Phi_{12}^l & \dots & \Phi_{1\mathrm{L}_{\beta^l}}^l \\ \Phi_{21}^l & \Phi_{22}^l & \dots & \Phi_{2\mathrm{L}_{\beta^l}}^l \\ \vdots & \vdots & \ddots & \vdots \\ \Phi_{\mathrm{L}_{\beta^l}1}^l & \Phi_{\mathrm{L}_{\beta^l}2}^l & \dots & \Phi_{\mathrm{L}_{\beta^l}\mathrm{L}_{\beta^l}}^l \end{pmatrix}.$$

With $\mathbb{0}$ a notation to refer to the corresponded block of null matrix. As Ψ^l for any $0 \leq l \leq p$ is a symmetric positive-definite matrix, we can easily find its Cholesky decomposition which will be used for the practical implementation.