A mixture of experts regression model for functional response with functional covariates

Jean Steve TAMO TCHOMGUI

jean-steve.tamo-tchomgui@univ-lyon2.fr

Entrepôts, Représentation et Ingénierie des Connaissances

Julien JACQUES

Entrepôts, Représentation et Ingénierie des Connaissances

Guillaume FRAYSSE

Orange (France)

Vincent BARRIAC

Orange (France)

Stéphane CHRETIEN

Entrepôts, Représentation et Ingénierie des Connaissances

Research Article

Keywords: Mixture of Experts, Functional regression, EM algorithm, Ridge regularized estimation

Posted Date: March 26th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4142146/v1

Additional Declarations: No competing interests reported.

A mixture of experts regression model for functional response with functional covariates

Jean Steve TAMO TCHOMGUI[1,2], Julien JACQUES[2], Guillaume FRAYSSE[1], Vincent BARRIAC[1], Stéphane CHRETIEN[2]

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

Contributing authors: jean-steve.tamo-tchomgui@univ-lyon2.fr; julien.jacques@univ-lyon2.fr; guillaume.fraysse@orange.com; vincent.barriac@orange.com; stephane.chretien@univ-lyon2.fr;

Abstract

Due to the fast growth of data that are measured on a continuous scale, functional data analysis has undergone many developments in recent years. Regression models with a functional response involving functional covariates, also called "function-on-function", are thus becoming very common. Studying this type of model in the presence of heterogeneous data can be particularly useful in various practical situations. We mainly develop in this work a Function-on-Function Mixture of Experts (FFMoE) regression model. Like most of the inference approach for models on functional data, we use basis expansion (B-splines) both for covariates and parameters. A regularized inference approach is also proposed, it accurately smoothes functional parameters in order to provide interpretable estimators. Numerical studies on simulated data illustrate the good performance of FFMoE as compared with competitors. Usefullness of the proposed model is illustrated on two data sets: the reference Canadian weather data set, in which the precipitations are modeled according to the temperature, and a Cycling data set, in which the developed power is explained by the speed, the cyclist heart rate and the slope of the road.

Keywords: Mixture of Experts, Functional regression, EM algorithm, Ridge regularized estimation.

1 Introduction

During the past few decades, functional data have become a very popular type of measurement in a constantly growing number of industrial, societal and medical applications. A branch of statistics, Functional Data Analysis (FDA), was developed as a specific discipline for analysing such data. FDA's flexibility in handling complex, high-dimensional, and structured data makes it applicable to a broad range of scientific and practical problems, providing insights that traditional data analysis methods may not be able to unveil. Broadly speaking, this new paradigm concerns the statistical analysis of data where at least one of the variables of interest is treated as a curve, surface or volume (also called function for simplicity) observed over a domain set. Most notable recent applications encompass, in particular, Healthcare and Medicine (monitoring patient health over time, FMRI data), Environmental Science (temperature or precipitation trends over time), Economics and Finance (evolution of stocks or commodities, modelling consumer behaviour over time), Sports Science (Analyzing athletes' performance data over time or during an event to optimize training and performance), Meteorology (analyzing weather patterns and trends to improve forecasting models), Chemometrics (analyzing spectroscopy data to identify and quantify chemical substances), Genomics and Bioinformatics (analyzing gene expression data over time), Traffic Analysis and Urban Planning.

Our main focus in the present work is the extension of linear regression to the functional data setting, a model which has naturally become a major area of research in the field of FDA. Standard references for FDA are [1–4]. A broad overview of functional linear regression is given in [5] and [6]. Using the convention that first term denotes the type of the response and second term denotes the type of the covariate three different setups have been analyzed in the literature: Function-on-Scalar, Scalar-on-Function and Function-on-Function. In the present work, we will focus on the most challenging setup from both statistical and computational perspectives, i.e.

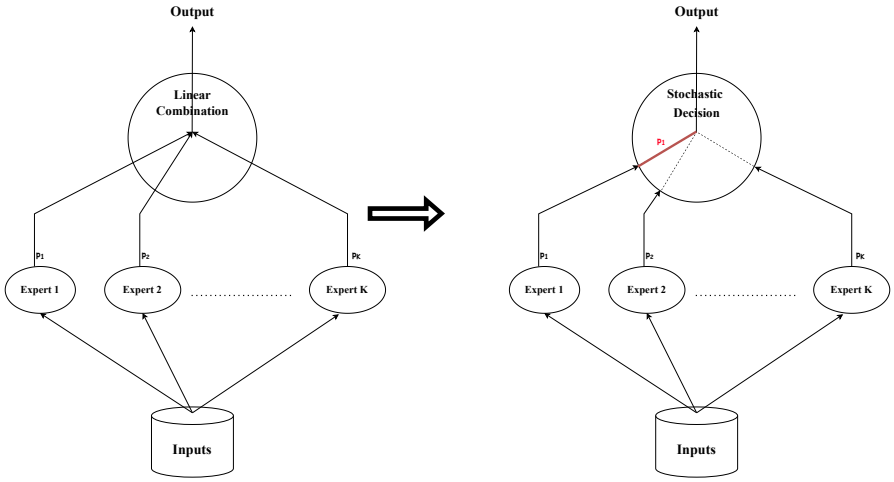
Function-on-Function regression problems. Function-on-Function regression problems have indeed been much less studied than the two other types of functional regression despite their relevance in many important applications. Recently, [7] proposed to estimate a Function-on-Function regression model using a penalized mixed model. A signal compression approach was also recently devised in [8], based on preprocessing the functional covariates using their wavelet transform and on proposing a method to estimate the functional parameter by characterizing them as solutions to a generalized functional eigenvalue problem. In a vast majority of current works in this area, one of the main issues is how to accurately select the most statistically relevant number of basis functions, and the location of the knots for spline models [9]. Another important issue is the interpretability of the obtained estimators [10]. In [11] proposed a Ridge-type penalization on second derivative of parameters using B-splines expansions for both functional covariates and parameters which is a first attempt at resolving the interpretability and model selection problems using convex sparsity-enforcing penalties.

Often in practice, the available data carry some heterogeneity, and the assumption that a unique relationship between the response variable and covariates holds for the full data set may not be valid. To circumvent this problem, a mixture of regression model can be proposed [12, 13]. As we know (see [14] and [15]), mixture models are very powerful at capturing subpopulation behaviour, a crucial capability in most applications. Mixture models have been studied in many different setups and specific algorithms, such as EM-type unpenalized and penalized models have been devised for the estimation of its parameters [16, 17]. Accelerated versions using space alternating schemes [18] and proximal interpretations [19, 20]. Sparsity-enforcing penalized versions were studied in [21].

Unfortunately, standard mixture models do not permit to parameterize the individual probability of each data to belong to a specific cluster. As this usually hampers

the predictive capabilities of mixture models, the framework of Mixture of Experts (MoE) models was first suggested in [22] as a powerful supervised learning procedure that can efficiently handle the potential heterogeneity often present in the data. The MoE model is based on a divide-and-conquer principle, which can be simply understood by realizing that each expert can specialize in smaller problems, and their predictive power can be combined together via a gating function in order to solve the full problem. The MoE model can also be viewed as a version of a multilayer supervised network in the sense that it is composed of K separate networks, each of which learning on a subset of the whole data data, as illustrated by Figure 1. From a more statistical learning perspective, the MoE model consists in a mixture model where both the mixture weights, a.k.a. Gating Functions, and component densities, a.k.a. Experts, depend on each data's covariate. The mixture model and its extension to MoE model has been investigated in the contexts of regression, clustering and discriminant analysis. A useful overview was proposed in [23], in which provides conditions for consistency and asymptotically normal properties are studied. Nevertheless, most MoE models only handle the scalar case. In the functional case, it would be relevant to implement efficient extensions of MoE model as well. This problem has already been tackled in [24], but for scalar response. Our contribution is to extend the MoE model to the Function-on-Function setup and provide an efficient inference algorithm.

The paper is organised as follows: Section 2 briefly presents the framework and the inference of Function-on-Function linear regression models. Section 3 presents the Function-on-Function MoE model we proposed and its inference. Section 4 describes how to implement a penalized version of the estimation scheme. Section 5 proposes extensive simulation experiments that explore the various aspects of the performance of the method. Section 6 finally presents an illustration of the method on two real-world data sets and shows the advantage, in terms of predictive quality, of considering MoE as compared with non-mixture-based approaches.



(a) (b)

Fig. 1: System of Experts and gating networks: The case of weighted linear combination (a) and the case of stochastic decision (b) to produce output.

2 The concurrent model

2.1 The functional model

The problem under study consists in modelling the relationship between functional covariates $\mathrm{X}^1(t), \ldots, \mathrm{X}^p(t)$ and a functional response $\mathrm{Y}(t)$ based on a n-sample $\{\mathrm{Y}_i(t), \mathrm{X}_i^1(t), \ldots, \mathrm{X}_i^p(t), t \in [0, \mathrm{T}]\}, i = 1, \ldots, n$. The functional response and covariates are assumed to belong to the separable Hilbert space $\mathrm{L}^2([0; \mathrm{T}])$ endowed with the Lebesgue measure. In the present work, we focus on the concurrent model [1] which assumes a linear relationship between the response and covariates, where the value of the response at a particular time stamp is modelled as a linear combination of the covariates at that specific time stamp, and the coefficients of the functional covariates

are univariate smooth functions of time:

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

with $\mathrm{X}_i(t) = (1 \;\; \mathrm{X}_i^1(t) \;\; \cdots \;\; \mathrm{X}_i^p(t))^\top$ and $\beta(t) = (\beta_0(t) \;\; \beta_1(t) \;\; \cdots \;\; \beta_p(t))^\top$.
$\beta_\ell(t)$ are the unknown functional parameters, assumed to be square integrable. The residuals $\varepsilon_i(t)$ are centered random variables with variance σ_i^2, specific to the i^{th} individual ([1], Chapter 13). Finally, $\varepsilon_i(t)$ and $\mathrm{X}_i^\ell(t)$ are assumed to be uncorrelated. The noise functions $\varepsilon_i(t)$ can also be rigorously defined using white noise theory as presented in [25]. In the framework of the present project, we will only use the property that when sampled at various times from a finite time set $\mathcal{T}$, the vector $(\varepsilon_i(t))_{t\in\mathcal{T}}$ can be expressed as a sum of a vector with independent and identically distributed (i.i.d.) components and a vector with prescribed covariance matrix, which can be a prescribed to a vector with constant components in the simplest case. Considering the concurrent model is of great interest because, as mentioned in [26], any functional linear model can be reduced to this form.

2.2 From functional to multivariate models

The parameters $\beta_\ell(t)$ of Model (1) can be estimated using the method discussed in [11], where the functional problem is rewritten as a classical multivariate regression problem by expanding the functional covariates and parameters into B-spline series, i.e.:

$$\mathrm{X}_i^\ell(t) = \sum_{j=1}^{q_{\mathrm{x}^\ell}} x_{ij}^\ell\, \mathrm{B}_j^\ell(t) = \mathrm{B}^\ell(t)^\top x_i^\ell \quad \text{and} \quad \beta_\ell(t) = \sum_{j=1}^{q_{\beta^\ell}} b_j^\ell\, \phi_j^\ell(t) = \phi^\ell(t)^\top b^\ell, \tag{2}$$

where $\mathrm{B}^\ell(t) = (\mathrm{B}_j^1(t), \ldots, \mathrm{B}_j^{q_{\mathrm{x}^\ell}}(t))^\top$ is the q_{x^ℓ} -dimensional vector of basis functions for the covariate $\mathrm{X}^\ell(t)$ and $x_i^\ell = (x_{i1}^\ell, \ldots, x_{iq_{\mathrm{x}^\ell}}^\ell)$ the corresponding basis expansion

coefficients. Analogously, $\{\phi^\ell(t),\ b^\ell\}$ are the basis functions and basis coefficients for $\beta_\ell(t)$. Then, using the following notations :

- $\Phi(t) = (\phi^0(t)^\top \ \ \phi^1(t)^\top \ \ \cdots \ \ \phi^p(t)^\top)$, a vector of length $\sum_\ell q_{\beta^\ell}$,
- $b = (b^{0^\top} \ \ b^{1^\top} \ \ \cdots \ \ b^{p^\top})^\top$, a vector of length $\sum_\ell q_{\beta^\ell}$,
- $B(t) = (1 \ \ B^1(t)^\top \ \ \cdots \ \ B^p(t)^\top)$, a vector of length $\sum_\ell q_{X^\ell}$,
- $x_i = (x_i^0(t)^\top \ \ x_i^1(t)^\top \ \ \cdots \ \ x_i^p(t)^\top)^\top$, a vector of length $\sum_\ell q_{X^\ell}$,

Model (1) can be written:

$$\mathrm{Y}_i(t) \ = \ x_i^\top \mathrm{B}(t)^\top \Phi(t)\ b + \varepsilon_i(t) = \mathrm{R}_i(t)^\top b + \varepsilon_i(t). \tag{3}$$

From this viewpoint, the concurrent model can be recast as a classical linear regression model with design matrix $\mathrm{R}_i(t) = \Phi(t)^\top \mathrm{B}(t)\, x_i$ and regression parameters b. When restricted to the observation grid consisting of the m successive timestamps $\{t_1, \ldots, t_m\}$, the problem reduces to:

$$\mathrm{Y}_i(t_j) = \mathrm{R}_i(t_j)^\top b + \varepsilon_i(t_j) \quad \text{with} \ \ 1 \leq i \leq n \text{ and } 1 \leq j \leq m. \tag{4}$$

There is nevertheless one peculiarity with this approach to underline. Indeed, in Model (4) the random variables $\varepsilon_i(t_1), \ldots, \varepsilon_i(t_m)$ representing the noise can not be assumed independent. In order to circumvent this issue, one possible approach is to use a linear mixed model (LMM) as advocated in [27]. For this purpose, we will assume that the model error can be decomposed as $\varepsilon_i(t_j) = \mathrm{U}_i + \eta_{ij}$, with η_{ij} a Gaussian white noise and U_i a random variable which takes into account the random effect in each individual $i \in \{1, \ldots, n\}$. In this framework, the estimation procedure proposed in [11] consists in maximizing the ridge-type penalised likelihood, with an ℓ_2-squared penalty on the second derivatives of $\beta_\ell(t)$. Such a penalty is recommended when smooth estimates

are sought for and provides sufficient flexibility that can still capture a substantial variety of complex shapes.

3 Mixture of experts of linear models for functional response with functional covariates

Mixture Regression (MR) models form a subset of the broad class of statistical models known as finite mixture models [13], which are designed to account for the statistical heterogeneity in a population through a finite set of empirical latent classes. MR models focus on identifying systematic differences between underlying latent groups in the population by the effect of covariates on the response. These models have to be distinguished from other mixture models that estimate the differences in levels and variance of the response variable between the groups (see [12]). MR models assumes that there are $\mathrm{K} \in \mathbb{N}^{\star}$ mixture components in the population. Component membership is indicated by a latent categorical variable (one-hot encoding as) $\mathrm{Z} = (\mathrm{z}_1, \ldots, \mathrm{z}_\mathrm{K})$ where z_k takes the value 1 if the observation belongs to the component k and 0 otherwise. The MR model can written

$$\mathrm{MR}(\mathrm{Y}|\mathrm{X}) = \sum_{k=1}^{\mathrm{K}} \pi_k \mathbb{E}_k[\mathrm{Y}|\mathrm{X}, \mathrm{z}_k = 1] \tag{5}$$

where π_k is the mixture proportion of group k associated with the k-th expert $\mathbb{E}_k[\mathrm{Y}|\mathrm{X}]$. In the present functional case, this expert is defined by

$$\mathbb{E}_k[\mathrm{Y}(t)|\mathrm{X}(t), \mathrm{z}_k = 1] = \mathrm{X}(t)^\top \beta_k(t) \tag{6}$$

where $\beta_k(t) = (\beta_{k,0}(t), \beta_{k,1}(t), \ldots, \beta_{k,p}(t))$ the functional parameters of the k^{th} expert.

Within the proposed model, there are two possible options for designing the probabilities π_k, $k = 1, \ldots, K$. The first one assumes that the covariates X are not related

to latent classes Z: $\pi_k = \mathbb{P}(z_k = 1)$. The second, and more general, assumes that Z depends on X: $\pi_k = \pi_k(\mathrm{X}) = \mathbb{P}(z_k = 1 \mid \mathrm{X})$.

The conditional density of $\mathrm{Y}(t)$ according to the Function-on-Function Mixture of Expert (FFMoE) model is

$$f(\mathrm{Y}(t)|\mathrm{X}(t), \Psi(t)) = \sum_{k=1}^{\mathrm{K}} \pi_k(\mathrm{X}(t), \alpha_k(t))\Phi(\mathrm{Y}(t); \mathrm{X}(t)\beta_k(t), \sigma_k^2), \tag{7}$$

with

- $\pi_k(\mathrm{X}(t),\ \alpha_k(t))$ the mixture proportion of group k, also called the k^{th} gated network function;
- $\Psi_k(t) = (\beta_k(t), \alpha_k(t))$ are the functional parameters;
- $\Phi(\mathrm{Y}(t); \mathrm{X}_i(t)\beta_k(t), \sigma_k^2)$ is the Gaussian density probability function of mean $\mathrm{X}(t)\beta_k(t)$ and variance σ_k^2.

3.1 Modelling the gated network function

The MoE model can be seen as a submodel of the Latent class model proposed by [28] named concomitant-variable latent class model. Various models for gated network have been proposed in the past "non-functional" related literature. One instance is the version of [22] where a multinomial logistic model is introduced. Another approach presented in [29] considers non parametric models. Turning to the functional setup, various authors have proposed extensions of the logistic regression model of [22]. Most of them assume in particular that the functional terms all belong to the space of real square integrable functions $\mathrm{L}^2([0, 1])$. See for instance [30] for an overview. In [31], it is shown that the functional nature of covariates raises important technical issues, some of them inherited from the non-functional setup but with higher complexities. Some of the more noticeable issues include the non-existence of maximum likelihood estimators

under general conditions, a remedy being working in a tailored Reproducing Kernel Hilbert Space (RKHS).

In the present work, under the realisation $\mathrm{x}_i(t)$ of $\mathrm{X}(t)$, we consider the following gating softmax function:

$$\pi_k(\mathrm{x}_i(t), \alpha_k(t)) = \frac{\exp(h_k(\mathrm{x}_i(t),\ \alpha_k(t)))}{1 + \sum_{k'=1}^{\mathrm{K}-1} \exp(h_{k'}(\mathrm{x}_i(t),\ \alpha_{k'}(t)))}, \tag{8}$$

where

$$h_k(\mathrm{x}_i(t),\ \alpha_k(t)) = \int_{\mathrm{T}} \alpha_k^\top(s)\mathrm{x}_i(s)ds \tag{9}$$

with $\alpha_k(t) = (\alpha_{k,0}(t), \alpha_{k,1}(t), \ldots, \alpha_{k,p}(t))^\top$. Notice that, in this model, the mixture proportion is constant over time.

As for the other functional parameters, $\alpha_k(t)$ is assumed to have an expansion into a basis of functions of the form:

$$\alpha_{k,\ell}(t) = \sum_{j=1}^{\mathrm{L}_{\alpha^\ell}} a_{k,j}^\ell \varrho_j^\ell(t) = \varrho^\ell(t)^\top a_k^\ell.$$

Similarly as for $\beta(t)$ in (2.2), we can write $\alpha_k(t) = \varrho(t)a_k$ and Equation ((9)) becomes:

$$h_k(\mathrm{x}_i(t), \alpha_k(t)) = \int_{\mathrm{T}} a_k^\top \varrho(s)^\top \mathrm{B}(s)\ x_i ds = a_k^\top \underbrace{\int_{\mathrm{T}} \varrho(s)^\top \mathrm{B}(s) dt x_i}_{r_i} = a_k^\top r_i,$$

Thus Model (8) can be written:

$$\pi_k(\mathrm{x}_i(t), \alpha_k(t)) = \frac{\exp(a_k^\top r_i)}{1 + \sum_{k'=1}^{\mathrm{K}-1} \exp(a_{k'}^\top r_i)}. \tag{10}$$

To guarantee the identifiability of $\alpha_k(t) \in \mathrm{L}^2(\mathbb{R}^{p+1})$, $k = 1, \ldots, \mathrm{K}$, $\alpha_{\mathrm{K}}(t)$ is set to the null function (and hence a_{K} is set to null vector) [32].

3.2 Estimation of the functional MoE via the EM algorithm

In practice, as expected, we only have access to a set of (noisy) observations at the timestamps in the set $\{t_1, \ldots, t_m\}$. For an observation i belonging to component k, the k^{th} expert model is given by

$$\mathrm{y}_i(t_j) = \beta_{k,0}(t_j) + \sum_{\ell=1}^{p} \beta_{k,\ell}(t)\mathrm{x}_i^{\ell}(t_j) + \varepsilon_i(t_j) = \beta_k(t_j)^{\top}\mathrm{x}_i(t_j) + \varepsilon_i(t_j), \tag{11}$$

where $\beta_k(t) = (\beta_{0,k}(t), \beta_{1,k}(t), \ldots, \beta_{p,k}(t))^{\top}$ for $k = 1 \ldots \mathrm{K}$, are the unknown functional experts parameters and are assumed to be square integrable.

As in the simple regression case, the successive observed values of a realisation i can not be assumed statistically independent. The mixed model approach of [27] can again be put to work after decomposing the observation error as $\varepsilon_i(t_j) = \mathrm{U}_i + \eta_{ij}$, with η_{ij} a Gaussian white noise and U_i a random variable which accounts for the random effect in each individual observation $i = 1, \ldots, n$. To sum up, model (11) consists of a LMM with fixed effects b_k and random effect U_i. In matrix form, this yields:

$$\mathbf{Y} = \mathrm{R}^{\top} b_k + \mathrm{WU} + \eta, \tag{12}$$

where $\mathbf{Y} = (\mathrm{y}_1(t_1), \ldots, \mathrm{y}_1(t_m), \mathrm{y}_2(t_1), \ldots, \mathrm{y}_n(t_m))^{\top}$, $\mathrm{R} = (\mathrm{R}_i(t_j))_{i,j}$ the design matrix of dimension $q_\beta \times nm$ with $q_\beta = \sum_\ell q_{\beta^\ell}$, $\mathrm{U} = (\mathrm{U}_1, \mathrm{U}_2, \cdots, \mathrm{U}_n)^{\top} \sim \mathcal{N}(\mathbf{0}, \Gamma)$, $\eta =$

$(\eta_{ij})_{i,j} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbb{I}_{nm})$ and

$$\mathrm{W} = \underbrace{\begin{pmatrix} 1_{m\times 1} & 0_{m\times 1} & \dots & 0_{m\times 1} \\ 0_{m\times 1} & 1_{m\times 1} & \dots & 0_{m\times 1} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{m\times 1} & 0_{m\times 1} & \dots & 1_{m\times 1} \end{pmatrix}}_{(nm\times n)\text{ - matrix}}.$$

We will make use of the notations $0_{k\times l}$ (resp. $1_{k\times l}$) of size $k \times l$ for the matrices of zeros (resp. ones) and the notation $\mathbf{0}$ for the null vector. We will also denote by Γ the unknown covariance matrix of the random effects. $\mathbb{I}_{nm}$ refers to the $nm \times nm$ identity matrix.

The conditional density of $\mathbf{Y}$, given the observations is a mixture of K Gaussian distributions of mean $b_k^\top \mathrm{R}$ and variance $\mathrm{V}_k = \mathrm{W}\Gamma\mathrm{W}^\top + \sigma_k^2 \mathbb{I}_{nm}$. So we have:

$$f(\mathbf{Y}|\mathbf{X}, \Psi) = \sum_{k=1}^{\mathrm{K}} \pi_k(\mathrm{x}_i(t), \alpha_k(t)) \Phi_{nm}(\mathbf{Y}; b_k^\top \mathrm{R}, \mathrm{V}_k), \tag{13}$$

where $\mathbf{X}$ is defined in the same way as $\mathbf{Y}$. $\Phi_\ell(x; \mu, \Sigma)$ denotes the probability density function of the L-dimensional Gaussian distribution with mean vector μ and covariance matrix Σ. $\Psi = ((a_1, b_1, \sigma_1^2), \dots, (a_{\mathrm{K}}, b_{\mathrm{K}}, \sigma_{\mathrm{K}}^2), \mathrm{U}, \Gamma)$ are the vector of parameters of the model to be estimated.

Inference of finite mixture model has been studied by various authors in the literature. We can mention for e.g. [22, 33] that compute Maximum Likelihood Estimators (MLE) via EM algorithm; Bayesian approaches have also been proposed as for instance in [34]; [29] present a parameter estimation approach in a semiparametric setting.

Now the FFMoE model can be defined using finite representation of functional terms. In this setting, we can easily write the observed data log-likelihood given by:

$$\mathcal{L}(\Psi; \{y_i(t_j), \mathrm{x}_i(t_j)\}_{i,j}) = \sum_{i=1}^{n} \log(\sum_{k=1}^{\mathrm{K}} \pi_k(\mathrm{x}_i(t), \alpha_k(t)) \Phi_m(\mathbf{y}_i; b_k^\top \mathrm{R}_i, \mathrm{V}_{k,i})) \quad (14)$$

where $\mathbf{y}_i$ is the vector of size m that contains all the measurements for observation i, R_i and $\mathrm{V}_{k,i}$ are respectively the design matrix and block covariance matrix of V_k associated with i. Then, the log-likelihood of Equation (14) becomes:

$$\sum_{i=1}^{n} \log(\sum_{k=1}^{\mathrm{K}} \frac{\exp(a_k^\top r_i)}{1 + \sum_{u=1}^{\mathrm{K}-1} \exp(a_u^\top r_i)} \cdot \frac{1}{\sqrt{(2\pi)^m |\mathrm{V}_{k,i}|}} \exp(-\frac{1}{2}(\mathbf{y}_i - b_k^\top \mathrm{R}_i)^\top \mathrm{V}_{k,i}^{-1}(\mathbf{y}_i - b_k^\top \mathrm{R}_i))).$$

As is well known in Finite Mixture Models, the log-likelihood maximisation problem is cumbersome to address without introducing clever intermediate steps that form the philosophy of EM-type algorithms, as extensively discussed in the landmark paper [16]. A basic requirement for the method is to complete the data by imputing latent group membership variables z_i for each observation $i = 1 \dots n$. These latent variables are represented by K binary variables $(z_{i1}, z_{i2}, \dots, z_{i\mathrm{K}})$. This model is called a complete model and leads to the complete data log-likelihood given by:

$$\begin{aligned}\mathcal{L}_c(\Psi; \{y_i(t_j), \mathrm{x}_i(t_j)\}_{i,j}) = \sum_{i=1}^{n} \sum_{k=1}^{\mathrm{K}} z_{ik} \log(&\frac{\exp(a_k^\top r_i)}{1 + \sum_{u=1}^{\mathrm{K}-1} \exp(a_u^\top r_i)} \cdot \frac{1}{\sqrt{(2\pi)^m |\mathrm{V}_{k,i}|}} \\ &\exp(-\frac{1}{2}(\mathbf{y}_i - b_k^\top \mathrm{R}_i)^\top \mathrm{V}_{k,i}^{-1}(\mathbf{y}_i - b_k^\top \mathrm{R}_i))). \quad (15)\end{aligned}$$

Let $\Psi^{(0)} = ((a_1^{(0)}, b_1^{(0)}, \sigma_1^{2^{(0)}}), \dots, (a_{\mathrm{K}}^{(0)}, b_{\mathrm{K}}^{(0)}, \sigma_{\mathrm{K}}^{2^{(0)}}), \mathrm{U}^{(0)}, \Gamma^{(0)})$ be an initial estimate of Ψ. The EM algorithm is a generic process consisting of repeating two steps to updates parameters such that the log-likelihood value monotonically increases:

- **E-step:** At this step, we compute the conditional expectation of the log-likelihood given the observed data and the current parameter (at iteration l) estimation $\Psi^{(l)}$. So we define the Q function for the EM algorithm defined by:

$$\mathrm{Q}(\Psi^{(l+1)}|\Psi^{(l)}) = \mathbb{E}(\mathcal{L}_c(\Psi^{(l+1)}; \{y_i(t_j), \mathrm{x}_i(t_j)\}_{i,j})|\{y_i(t_j), \mathrm{x}_i(t_j)\}_{i,j};\ \Psi^{(l)}). \quad (16)$$

This consists of computing the posterior probabilities $p_{ik}^{(l)}$ that the curves i-th sample $(y_i(t), \mathrm{x}_i(t))$ belongs to the k^{th} component of the mixture under the current model:

$$p_{ik}^{(l)} = \mathbb{E}(z_{ik}|\{y_i(t_j), \mathrm{x}_i(t_j)\}_{i,j}; \Psi^{(l)}) = \mathbb{P}(z_{ik} = 1|\{y_i(t_j), \mathrm{x}_i(t_j)\}_{i,j}; \Psi^{(l)}).$$

Using Bayes' theorem, this conditional probability $p_{ik}^{(l)}$ can be expressed as:

$$p_{ik}^{(l)} = \frac{\pi_k(\mathrm{x}_i(t), \alpha_k^{(l)}(t))\Phi_m(\mathbf{y}_i; b_k^{\top(l)}\mathrm{R}_i, \mathrm{V}_{k,i}^{(l)}, t \in \mathrm{T})}{\sum_{u=1}^{\mathrm{K}} \pi_u(\mathrm{x}_i(t),\ \alpha_u^{(l)}(t))\Phi_m(\mathbf{y}_i; b_u^{\top(l)}\mathrm{R}_i, \mathrm{V}_{u,i}^{(l)})}. \quad (17)$$

- **M-step:** Given the previous conditional probability and the observed data, this step updates the current parameters $\Psi^{(l)}$ by maximizing the conditional expectation of the complete data log-likelihood, that is $\Psi^{(l+1)}$:

$$\begin{aligned}\mathrm{Q}(\Psi^{(l+1)}|\Psi^{(l)}) &= \mathbb{E}(\mathcal{L}_c(\Psi^{(l+1)}; \{y_i(t_j), \mathrm{x}_i(t_j)\}_{i,j})|\{y_i(t_j), \mathrm{x}_i(t_j)\}_{i,j};\ \Psi^{(l)}) \\ &= \mathrm{Q}_1(a_k^{(l+1)}|\Psi^{(l)}) + \mathrm{Q}_2(b_k^{(l+1)},\ \mathrm{V}_k^{(l+1)}|\Psi^{(l)}). \quad (18)\end{aligned}$$

The EM algorithm was shown to be a particular case of the celebrated Proximal Point algorithm in [19, 20] using a Kullbak-Leibler type divergence for the proximity term. Another interesting interpretation in terms of alternating minimisation in given in [35]. Space alternating version of the EM algorithms where proposed in

[18, 36] and [21] for the nonsmoothly penalised case. In this paper, the maximisation of (Q) will be performed using a modified version of the **R** package [37]: in particular, the function `initFlexmix` which allows repeating the EM algorithm with different starting values and choosing the solution with the highest value of the likelihood while allowing concomitant variables, as developed in [38]. The global maximisation problem is split onto two separate maximisation problems (see Appendix A for details):

- the updating of gated network parameters via the maximisation of the function $\mathrm{Q}_1(a_k^{(l+1)}|\Psi^{(l)})$ and
- the updating of the expert's parameters via the maximisation of the function $\mathrm{Q}_2(b_k^{(l+1)},\ \mathrm{V}_k^{(l+1)}|\Psi^{(l)})$.

One will easily recognise in each of these two expressions, the likelihood of the multinomial logistic model $\mathrm{Q}_1(.)$ and of the linear Gaussian model $\mathrm{Q}_2(.)$ for which we know how to compute (at least numerically using e.g. Newton-Raphson iterations) the MLEs.

- The E and M steps are alternated repeatedly until numerical convergence i.e. the difference $\mathcal{L}(\Psi^{(l+1)};\{y_i(t_j),\mathrm{x}_i(t_j)\}_{i,j}) - \mathcal{L}(\Psi^{(l)};\{y_i(t_j);\mathrm{x}_i(t_j)\}_{i,j})$ changes by no more than an arbitrarily small value.

Stability and convergence properties of the method are established in the literature (see [17] for an overview and [20] for the proximal viewpoint).

With the estimates of gated network and experts parameters obtained, a hard-clustering of the link between $\mathrm{X}(t)$ and $\mathrm{Y}(t)$ is reach using Bayes' rule so that

$$\widehat{z}_{ik} = \begin{cases} 1 & \text{if } k = \mathrm{Arg}\max\limits_{1\leq k\leq \mathrm{K}} p_{ik}, \\ 0 & \text{otherwise,} \end{cases} \qquad i = 1\ldots n.$$

where p_{ik} is the value of Equation (17) at convergence.

3.3 Model selection

One important challenge in statistical estimation with potentially several possible models depending on hyperparameters is the selection of the most statistically relevant one. In the present model, choosing the correct number of components K is one crucial step of the estimation problem. In the regression setting, the selection can be done using information criteria such as AIC [39] or BIC [40], or using cross validation methods. The latter being time-consuming, we will use information criterion based approaches and more specifically the BIC criterion usually defined using log-likelihood (14) as:

$$\text{BIC} = \text{-}2\mathcal{L}(\Psi; \{y_i(t_j), \mathrm{x}_i(t_j)\}_{i,j}) - d\log(n) \tag{19}$$

where $d = \mathrm{K} \times (1 + \sum_{\ell=0}^{p} \mathrm{L}_{\beta^\ell} + \sum_{\ell=1}^{p} \mathrm{L}_{\alpha^\ell})$ is the number of free parameters of the model and n the number of observations.

3.4 Prediction

As we have already mentioned, one of the major limitations of simple mixture models is predictive modelling. Since for a new individual, its prediction will be given by the weighted sum of the predictions of each class. This is so far not ideal as this prediction is entirely driven by the prediction of the most probably class and these class probabilities will not change whatever the characteristics of the new individual. With the MoE model, we have seen that we can make this latent class probability depending on the covariates (concomitant variables). In this case, the prediction is given by expert prediction of the most probable class. To build such predictions we first need the conditional probabilities that any individual i belongs to a component

k given by:

$$\pi_k(\mathrm{X}_i(t),\, \widehat{\alpha}_k) = \frac{\exp(\widehat{a}_k^\top r_i)}{1 + \sum_{v=1}^{\mathrm{K}-1} \exp(\widehat{a}_v^\top r_i)}$$

where $\widehat{a}_k$ for $1 \leq k \leq \mathrm{K}-1$ are the gated parameters estimators.

We deduce, where component k_m is the most probable class for the i curve, the predictive curve by:

$$\widehat{\mathrm{Y}}_i(t) = b_{k_m}^\top \mathrm{R}_i(t).$$

As a result, estimating the group membership from covariates is essential to predict the response well.

4 Regularizing the Function-on-Function mixture of experts regression

In the FFMoE model (7) presented in Section 3, it is assumed that the functional covariates and parameters can be decomposed into a finite dimensional functional basis. This assumption allows to get the finite representation (13). The numbers of basis functions of each parameters and covariates should be correctly selected in order to avoid over- or under-fitting. Nevertheless, precise adjustment of these values often induces a high computational effort. In the case of the B-spline basis, even more parameters have to be properly tuned such as the choice of the spline order and the location of the knots. In order to reduce the expected cost of such a computationally demanding procedure, we made the choice of choosing a sufficiently large a priori value for L_β (or L_α) and then apply a penalty. This approach brings the benefit of tuning a single hyperparameter, which is the number of basis functions and improving the smoothness and then interpretability of the estimated functional coefficients. This last point is very interesting in the case of the linear model because as we already know,

the interpretation of the predictors-response relationship becomes more difficult as the shape of the functional parameter $\beta(t)$ (or $\alpha(t)$) does not have any simple structure.

Various approaches to regularize the parameter shape have been proposed in the literature. In our setting of interest, the main goal is to enhance the shape of parameters and then interpretability. [41] are among the first to explore the functional penalization and show that the obtained estimator are less sensitive to the rather subjective choice of the number of basis functions. [10] proposed a method called Functional Linear Regression That is Interpretable (FLiRTI) which address the issue of choosing relevant penalties. Based on variable selection ideas such as the Lasso penalty, FLiRTI produces accurate, flexible and highly interpretable estimates of the functional parameters. The main idea of FLiRTI method is, instead of enforcing sparsity on the function themselves, to enforce sparsity of the derivatives. Using the notation $\beta^{(l)}(t)$ for the l^{th} derivative of $\beta(t)$, we may deduce that $\beta^{(0)}(t) = 0$ guarantees $\mathrm{X}(t)$ has no effect on $\mathrm{Y}(t)$ at t; $\beta^{(1)}(t) = 0$ implies that $\beta(t)$ is constant at t; $\beta^{(2)}(t) = 0$ means that $\beta(t)$ is linear at t and so on.

4.1 Ridge-type penalty on second derivatives

Instead of the Lasso penalty, we proposed to estimated the functional MoE model (13) by maximizing a Ridge-type penalized log-likelihood. The penalty is based on the second derivative of the functional parameters (both gated and experts). This choice is mainly motivated by the desire to obtain a possibly locally constant relationship if needed. Moreover, the use the ridge penalty is motivated by the lack of exact sparsity observed in real problems and the clear benefits of getting a closed form formula for the estimators.

The corresponding penalized (data) log-likelihood function for the observed data is defined using (14) by:

$$\mathcal{L}_{pen}(\Psi; \{y_i(t_j), \mathrm{x}_i(t_j)\}_{i,j}) = \mathcal{L}(\Psi; \{y_i(t_j), \mathrm{x}_i(t_j)\}_{i,j}) + \mathrm{Pen}(\Psi), \tag{20}$$

in which the Ridge regularization term is given by

$$\mathrm{Pen}(\Psi) = \sum_{k=1}^{\mathrm{K}} \sum_{\ell=0}^{p} \lambda_{k,\ell} \int \beta_{k,\ell}^{''}(t)^2 dt + \sum_{k=1}^{\mathrm{K}-1} \sum_{\ell=1}^{p} \gamma_{k,\ell} \int \alpha_{k,\ell}^{''}(t)^2 dt$$

where

$$\int \beta_{k,\ell}^{''}(t)^2 dt = \int \left[\sum_{j=1}^{\mathrm{L}_{\beta^\ell}} b_{k,j}^{\ell} \varphi_j^{\ell^{''}}(t) \right]^2 dt = \sum_{s,u=1}^{\mathrm{L}_{\beta^\ell}} b_{k,s}^{\ell} b_{k,u}^{\ell} \Gamma_{su}^{\ell}$$

with $\Gamma_{su}^{\ell} = \int \varphi_s^{\ell^{''}}(t) \varphi_u^{\ell^{''}}(t) dt$, and

$$\int \alpha_{k,\ell}^{''}(t)^2 dt = \int \left[\sum_{j=1}^{\mathrm{L}_{\alpha^\ell}} a_{k,j}^{\ell} \varrho_j^{\ell^{''}}(t) \right]^2 dt = \sum_{s,u=1}^{\mathrm{L}_{\beta^\ell}} a_{k,s}^{\ell} a_{k,u}^{\ell} \Upsilon_{su}^{\ell}$$

with $\Upsilon_{su}^{\ell} = \int \varrho_s^{\ell^{''}}(t) \varrho_u^{\ell^{''}}(t) dt$.

So,

$$\mathrm{Pen}(\Psi) = \sum_{k=1}^{\mathrm{K}} \sum_{\ell=0}^{p} \lambda_{k,\ell} \sum_{s,u=1}^{\mathrm{L}_{\beta^\ell}} b_{k,s}^{\ell} b_{k,u}^{\ell} \Gamma_{su}^{\ell} + \sum_{k=1}^{\mathrm{K}-1} \sum_{\ell=1}^{p} \gamma_{k,\ell} \sum_{s,u=1}^{\mathrm{L}_{\beta^\ell}} a_{k,s}^{\ell} a_{k,u}^{\ell} \Upsilon_{su}^{\ell} \tag{21}$$

where $\lambda_{k,\ell}$ and $\gamma_{k,\ell}$ are the usual tuning regularization parameters which control the importance we want to place on the smoothness of estimators. As we know, selecting a good value of $\lambda_k = (\lambda_{k,\ell})_\ell$ (resp. $\gamma_k = (\gamma_{k,\ell})_\ell$) is very important to reduce the noise that less influential covariates create.

By using matrix terms, we get:

$$\mathcal{L}_{pen}(\Psi; \{y_i(t_j), \mathrm{x}_i(t_j)\}_{i,j}) = \mathcal{L}(\Psi; \{y_i(t_j), \mathrm{x}_i(t_j)\}_{i,j}) - \sum_{k=1}^{\mathrm{K}} b_k^\top (\lambda_k \mathrm{P})\ b_k - \sum_{k=1}^{\mathrm{K}-1} a_k^\top (\gamma_k \mathrm{Q})\ a_k$$

where $(\lambda_k \mathrm{P}) \in \mathbb{R}^{\mathrm{L}_\beta \times \mathrm{L}_\beta}$ is given by:

$$(\lambda_k \mathrm{P}) = \begin{pmatrix} \lambda_{k,0}\Gamma^0 & 0_{\mathrm{L}_{\beta^0} \times \mathrm{L}_{\beta^1}} & \dots & 0_{\mathrm{L}_{\beta^0} \times \mathrm{L}_{\beta^p}} \\ 0_{\mathrm{L}_{\beta^1} \times \mathrm{L}_{\beta^0}} & \lambda_{k,1}\Gamma^1 & \dots & 0_{\mathrm{L}_{\beta^1} \times \mathrm{L}_{\beta^p}} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{\mathrm{L}_{\beta^p} \times \mathrm{L}_{\beta^0}} & 0_{\mathrm{L}_{\beta^p} \times \mathrm{L}_{\beta^1}} & \dots & \lambda_{k,p}\Gamma^p \end{pmatrix} \quad \text{with } \Gamma^\ell = \begin{pmatrix} \Gamma^\ell_{11} & \Gamma^\ell_{12} & \dots & \Gamma^\ell_{1\mathrm{L}_{\beta^\ell}} \\ \Gamma^\ell_{21} & \Gamma^\ell_{22} & \dots & \Gamma^\ell_{2\mathrm{L}_{\beta^\ell}} \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma^\ell_{\mathrm{L}_{\beta^\ell}1} & \Gamma^\ell_{\mathrm{L}_{\beta^\ell}2} & \dots & \Gamma^\ell_{\mathrm{L}_{\beta^\ell}\mathrm{L}_{\beta^\ell}} \end{pmatrix};$$

and $(\gamma_k \mathrm{Q}) \in \mathbb{R}^{\mathrm{L}_\alpha \times \mathrm{L}_\alpha}$ by:

$$(\gamma_k \mathrm{Q}) = \begin{pmatrix} \gamma_{k,0}\Upsilon^0 & 0_{\mathrm{L}_{\alpha^0} \times \mathrm{L}_{\alpha^1}} & \dots & 0_{\mathrm{L}_{\alpha^0} \times \mathrm{L}_{\alpha^p}} \\ 0_{\mathrm{L}_{\alpha^1} \times \mathrm{L}_{\alpha^0}} & \gamma_{k,1}\Upsilon^1 & \dots & 0_{\mathrm{L}_{\alpha^1} \times \mathrm{L}_{\alpha^p}} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{\mathrm{L}_{\alpha^p} \times \mathrm{L}_{\alpha^0}} & 0_{\mathrm{L}_{\alpha^p} \times \mathrm{L}_{\alpha^1}} & \dots & \gamma_{k,p}\Upsilon^p \end{pmatrix} \quad \text{with } \Upsilon^\ell = \begin{pmatrix} \Upsilon^\ell_{11} & \Upsilon^\ell_{12} & \dots & \Upsilon^\ell_{1q_{\alpha^\ell}} \\ \Upsilon^\ell_{21} & \Upsilon^\ell_{22} & \dots & \Upsilon^\ell_{2\mathrm{L}_{\alpha^\ell}} \\ \vdots & \vdots & \ddots & \vdots \\ \Upsilon^\ell_{\mathrm{L}_{\alpha^\ell}1} & \Upsilon^\ell_{\mathrm{L}_{\alpha^\ell}2} & \dots & \Upsilon^\ell_{\mathrm{L}_{\alpha^\ell}\mathrm{L}_{\alpha^\ell}} \end{pmatrix}.$$

Here, $0_{\mathrm{L}_1 \times \mathrm{L}_2}$ is the standard notation for the null matrix of size $\mathrm{L}_1 \times \mathrm{L}_2$. As Γ^ℓ (resp. Υ^ℓ) is a symmetric positive-definite matrix for any $0 \leq \ell \leq p$, we can easily find its Cholesky decomposition, which can be efficiently leveraged in the implementation.

And for the penalized complete (data) log-likelihood, we made the same process and by using (15) we get:

$$\begin{aligned} \mathcal{L}^c_{pen}(\Psi; \{y_i(t_j), \mathrm{x}_i(t_j)\}_{i,j}) = \mathcal{L}_c(\Psi; \{y_i(t_j), \mathrm{x}_i(t_j)\}_{i,j}) - \\ \sum_{k=1}^{\mathrm{K}} b_k^\top (\lambda_k \mathrm{P})\ b_k - \sum_{k=1}^{\mathrm{K}-1} a_k^\top (\gamma_k \mathrm{Q})\ a_k. \end{aligned} \tag{22}$$

4.2 Maximum Likelihood estimation via the EM algorithm

The EM algorithm for the regularized FFMoE is developed for maximizing the penalized (data) log-likelihood (22). The algorithm is simply the same as in non penalized version with small changes. The E-step is exactly the same and the M-step is done by splitting the problem into two maximize problems as (see Appendix B for details):

$$\begin{aligned} \mathrm{Q}_{pen}(\Psi^{(l+1)}|\Psi^{(l)}) &= \mathbb{E}(\mathcal{L}^{c}_{pen}(\Psi^{(l+1)})|y(t),\ \mathrm{x}(t);\ \Psi^{(l)}) \\ &= \mathrm{Q}_{1,pen}(a_k^{(l+1)}|\Psi^{(l)}) \ +\ \mathrm{Q}_{2,pen}(b_k^{(l+1)},\ \sigma_k^{2\ (l+1)}|\Psi^{(l)}). \end{aligned} \tag{23}$$

5 Simulation study of mixture of experts functional models

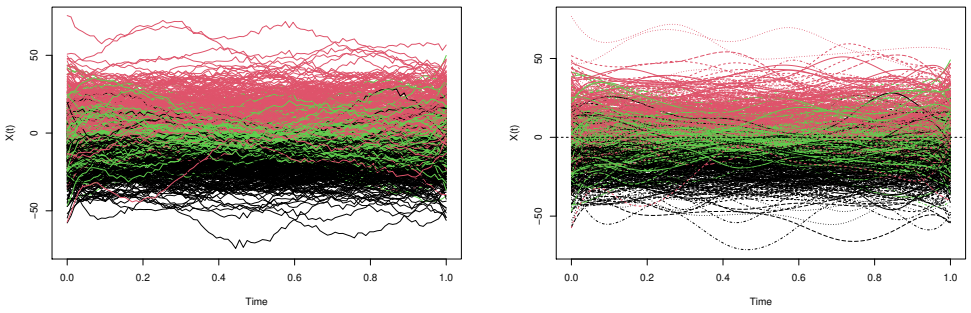
The goal of this section if to evaluate, on the basis of simulated data, the proposed model in the case of Function-on-Function regression model. The data simulation process if derived from [24].

5.1 Data simulation process

100 data sets are simulated according to the FFMoE model with $K = 3$ components and $p = 1$ covariate, on a time domain $[0, 1]$. The covariate is simulated with $\mathrm{X}_i(t) = x_i^\top \mathrm{B}(t)$, where $x_i = W.v_i$ with W a $10 \times 10-$matrix of $\mathcal{U}(0, 1)$, v_i a $10-$vector of $\mathcal{N}(0, 10)$ and $\mathrm{B}(t)$ is a 10-dimensional B-splines basis. The functional parameters are $\beta_{1,0}(t) = -5t$, $\beta_{2,0}(t) = 0$ and $\beta_{3,0}(t) = 5t$, $\beta_{2,1}(t) = -\beta_{1,1}(t)$, $\beta_{3,1}(t) = 100(t-0.5)^2 -$

Scenarios	number of sampling points: m	number of observations: n
S1	20	300
S2	20	800
S3	100	300
S4	100	800

Table 1: The four scenarios of the simulation study



(a) (b)

Fig. 2: Discrete observations (left) and cubic B-splines smoothing (right) of the functional covariate. Color depends on the component membership.

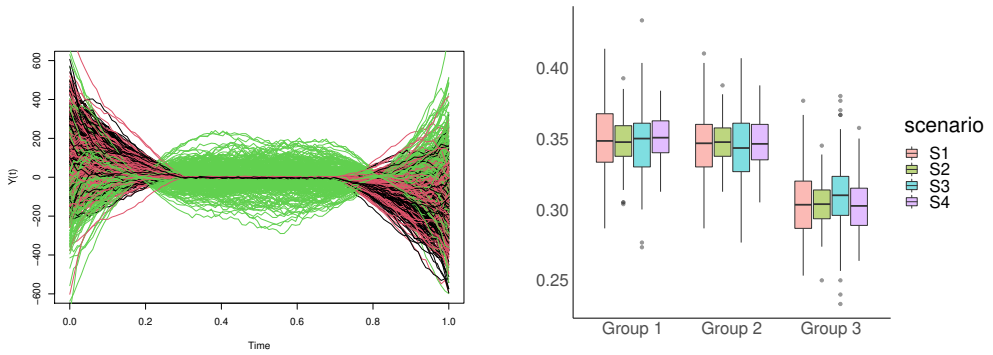
10 and

$$\beta_{1,1}(t) = \begin{cases} -50(t-0.5)^2 + 2 & \text{if } 0 \leq t < 0.3 \\ 0 & \text{if } 0.3 \leq t < 0.7 \\ 50(t-0.5)^2 - 2 & \text{if } 0.7 \leq t < 1 \end{cases}$$

The functional parameters of the gated network are $\alpha_{1,0} = \alpha_{2,0} = -10$, $\alpha_{3,0} = 0$, $\alpha_{1,1}(t) = 80(t-0.5)^2 - 8$, $\alpha_{2,1}(t) = -\alpha_{1,1}(t)$ and $\alpha_{3,1}(t) = 0$. Finally, the residuals are simulated with $\varepsilon_i(t) \sim \mathcal{N}(0, 4)$.

The number n of observations and the number m of sampling points are given in Table 1, defining thus four scenarios S1, S2, S3, S4.

Figure 2 plots the discrete covariate observations (left panel) and their corresponding B-splines smoothing (right panel) for Scenario S3. Figure 3 displays the discrete



(a) (b)

Fig. 3: Discrete observations of the functional output (left) and proportions of observations of each component on the mixture (right).

time sampled observations of the response $Y(t)$ (left) for Scenario S3, and the proportions of observations of each component on the mixture (right) for the four scenarios.

5.2 Assessment criteria of goodness of fit

The assessment of the proposed FFMoE model is performed using two specific indicators: first the estimation quality and second, the prediction quality. In addition, the efficiency of BIC for selecting the number of components is also investigated.

The quality of parameter estimation is evaluated with the Mean Square Error (MSE)

$$\text{MSE}(\beta_\ell(.)) = \left[\frac{1}{m}\sum_{j=1}^{m}(\beta_\ell(t_j) - \widehat{\beta}_\ell(t_j))^2\right]^{1/2}. \tag{24}$$

Knowing that the label-switching problem sometimes occurs, we will take care of relabelling the clusters using the estimated confusion matrix, a strategy which is relevant when the true number of mixture components has been guessed.

The quality of prediction is assessed using the Mean Relative Prediction Error (MRPE) on a generate test sample of length $n_{test} = 2000$ for each scenario:

$$\text{MRPE} = \frac{1}{m}\sum_{j=1}^{m}\left(\frac{\sum_{i=1}^{n_{test}}(\text{Y}_i(t_j) - \widehat{\text{Y}}_i(t_j))^2}{\sum_{i=1}^{n_{test}}\text{Y}_i(t_j)^2}\right). \tag{25}$$

Notice that this criterion can be highly irrelevant if the observation is associated with the wrong expert. Subsequently, two additional criteria will be defined: MRPE.good, computed only of the observations associated with the correct expert, and MRPE.bad for those associated with a wrong expert.

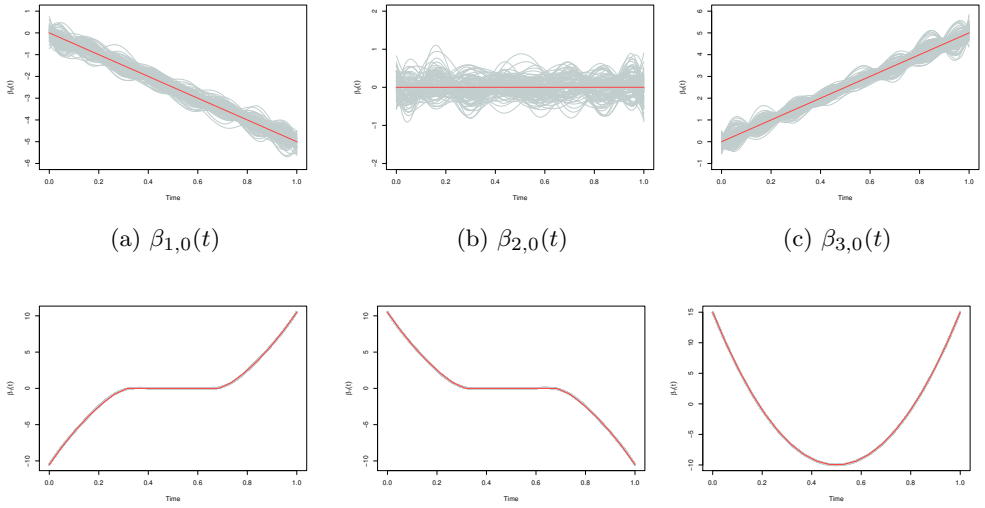
5.3 Competitors

The competitors are the non-mixture penalized Function-on-Function regression models PenFFR [11] and pffr [7].

The PenFFR estimation process uses basis expansion of functional covariates and parameters to transform a functional model to multivariate. Estimation scheme is achieved by maximising the penalised log-likelihood using a ridge-type penalty on the second derivatives. Cubic B-splines basis functions were employed for both for functional covariates and the functional parameters. The number of basis functions was set to 10 for both the functional parameters and covariates.

The pffr estimation process uses observed values of functional covariates. An approach that matches with densely or sparsely sampled functions. The functional parameters is estimated using restricted maximum likelyhood (REML) in an associated mixed model. For the implementations of the method, we used default settings of the `pffr` function available in the `R` package `refund`. We only set the number of basis functions to 10 both for functional covariates and parameters.

Finally, for FFMoE and PenFFMoE, we also set the number of basis functions to 10 both for both for functional covariates and parameters.



(a) $\beta_{1,0}(t)$ (b) $\beta_{2,0}(t)$ (c) $\beta_{3,0}(t)$

(d) $\beta_{1,1}(t)$ (e) $\beta_{2,1}(t)$ (f) $\beta_{3,1}(t)$

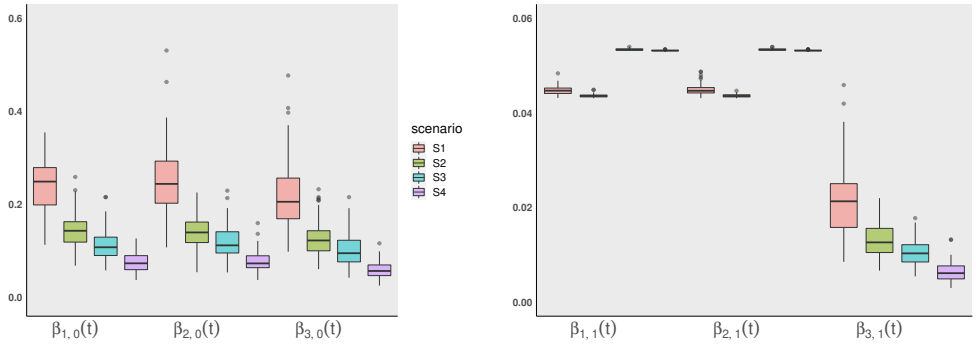
Fig. 4: Estimation of the regression coefficients for Scenario S3 with FFMoE. The red curves are the actual parameters, the gray curves are the estimation.

5.4 Simulation results

Concerning the ability of BIC to select the right number of components, BIC selects indeed the correct number $K = 3$ in 100% of the case for the four scenarios, and thus for FFMoE and PenFFMoE.

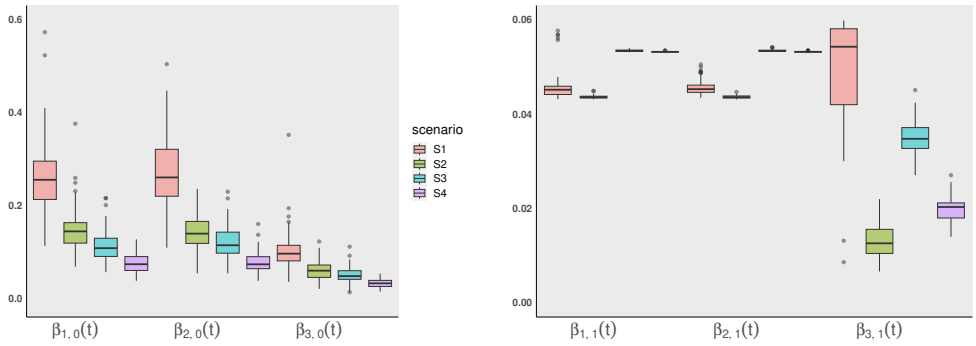
5.4.1 Parameter estimation

The relevance of our model is reflected by the parameter estimation. Figure 4 for FFMoE and Figure C1 for PenFFMoE in appendix show the estimated versus actual parameters from Scenario S3. The estimation of the covariate effect is remarkably accurate. This observation is also supported by the MSE values reported in boxplot given in Figure 6 for PenFFMoE and Figure 5 for FFMoE in all scenarios.



(a) Intercept (b) Effect of X

Fig. 5: Boxplot of MSE between actual and estimated parameters for FFMoE. Functional intercept $\beta_0(t)$ (left) and functional effect $\beta_1(t)$ of X(t) (right) in each of the 3 components mixture for our 4 simulated scenarios.



(a) Intercept (b) Effect of X

Fig. 6: Boxplot of MSE between actual and estimated parameters for PenFFMoE. Functional intercept $\beta_0(t)$ (left) and functional effect $\beta_1(t)$ of X(t) (right) in each of the 3 components mixture for our 4 simulated scenarios.

5.4.2 Prediction accuracy

In Table 2, we present the predictive accuracy through MRPE of the proposed method (FFMoE and PenFFMoE) and of its competitors (PenFFR and pffr) on the test sample. The results are clearly better for FFMoE and PenFFMoE. Let's remark that the

		Expert affectation accuracy	MRPE.good	MRPE.bad	MRPE
S1	FFMoE	91.3% (0.014)	0.006 ($<10^{-4}$)	2.560 (0.30)	0.230 (0.07)
	PenFFMoE	92.3% (0.018)	0.006 ($<10^{-4}$)	2.568 (0.29)	0.205 (0.07)
	PenFFR	-	-	-	1.213 (0.07)
	pffr	-	-	-	1.266 (0.08)
S2	FFMoE	93.0% (0.005)	0.006 ($<10^{-4}$)	2.500 (0.18)	0.180 (0.02)
	PenFFMoE	93.3% (0.003)	0.006 ($<10^{-4}$)	2.501 (0.18)	0.174 (0.01)
	PenFFR	-	-	-	1.192 (0.04)
	pffr	-	-	-	1.252 (0.05)
S3	FFMoE	92.0% (0.026)	0.016 ($<10^{-3}$)	2.715 (0.49)	0.280 (0.38)
	PenFFMoE	92.8% (0.040)	0.016 ($<10^{-3}$)	2.730 (0.45)	0.219 (0.16)
	PenFFR	-	-	-	1.227 (0.07)
	pffr	-	-	-	1.290 (0.09)
S4	FFMoE	93.9% (0.004)	0.015 ($<10^{-4}$)	2.628 (0.21)	0.174 (0.02)
	PenFFMoE	94.3% (0.003)	0.016 ($<10^{-4}$)	2.614 (0.20)	0.165 (0.01)
	PenFFR	-	-	-	1.237 (0.05)
	pffr	-	-	-	1.314 (0.06)

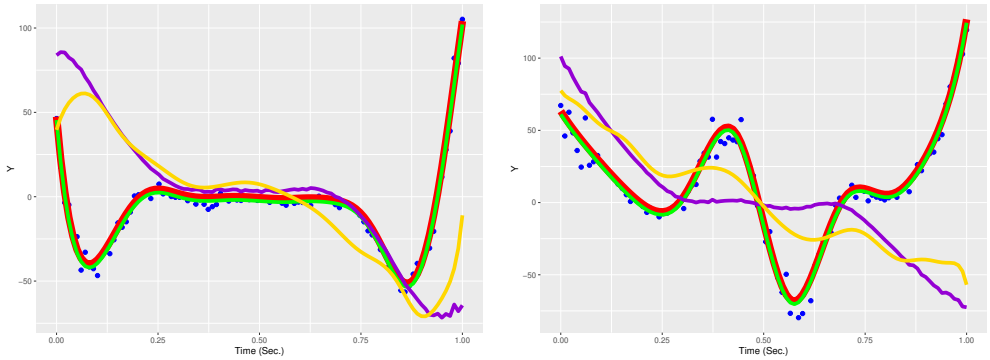
Table 2: Expert affectation accuracy and average (standard deviation) of MRPE on a test sample.

difference between MRPE.good and MRPE.bad show that it is important to correctly affect the observations to the correct expert.

Finally, Figure 7 gives the prediction for four randomly chosen observations compared to actual values. Figure (7a) and Figure (7b) correspond to situations where the observations are assigned to the correct clusters; Figure (7c) corresponds to a case where the data is assigned to the correct clusters by the penalized method but not for the non penalized method and Figure (7d) corresponds to cases where the data is assigned to a wrong cluster, for both methods.

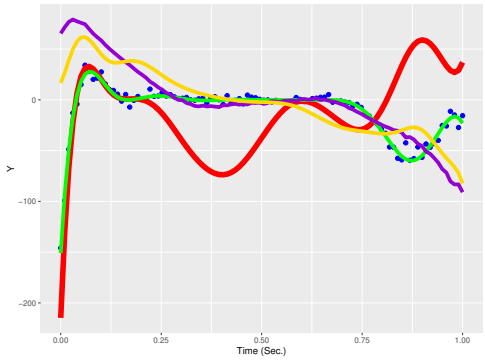
6 Application to real-world data

In this section, we perform our proposed methodology FFMoE and PenFFMoE on two real-world data sets: Canadian Weather (CW, available in the R package [42]) and Cycling (available in the R package [43]).

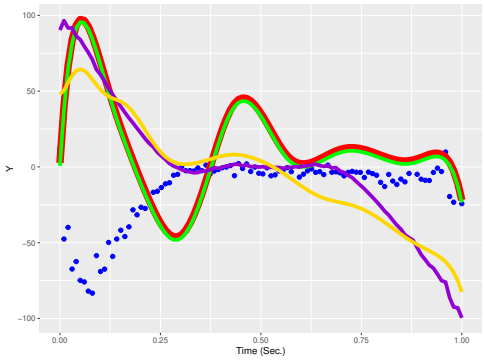


(a) Prediction 1

(b) Prediction 2



(c) Prediction 3



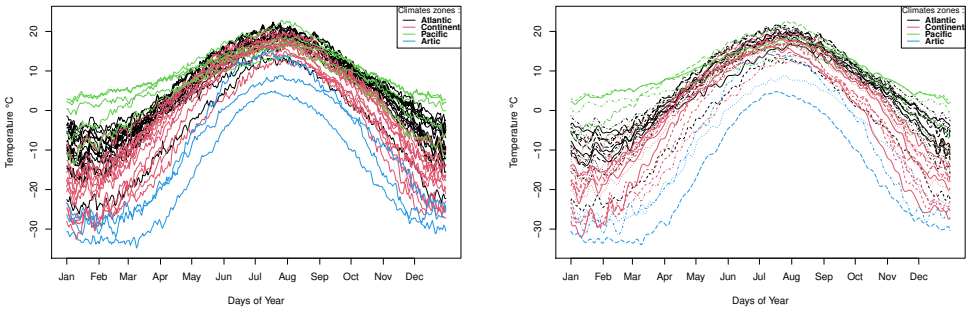
(d) Prediction 4

Fig. 7: observed data vs Fitted response functions for four chosen individuals on the test sample. Red and green lines match to FFMoE and PenFFMoE resp.; gold and violet lines are the prediction pffr and PenFFR resp.; the actual data isthe blue dots.

In each of these data sets, the prediction accuracy of FFMoE and PenFFMoE is compared with the competitors PenFFR [11] and pffr [7]. Let us remark that PenFFR and pffr consider a single model and not a mixture as compared with FFMoE and PenFFMoE. Comparison is done by the leave-one-out cross-validation integrated square error (ISE):

$$\mathrm{ISE}_i = \int_0^T (\mathrm{Y}_i(t) - \widehat{\mathrm{Y}}_i^{(-i)}(t))^2 \, dt \,,$$

where $\widehat{\mathrm{Y}}_i^{(-i)}(t)$ is the prediction of the i^{th} observation given by the model trained on a dataset of all the observations without the i^{th} one. Computationally, this criterion is



(a) (b)

Fig. 8: 35 daily mean raw (a) and processed (b) temperature measurement curves.

approximated by the L^2-norm between the actual and prediction values on a grid of values t is used as a surrogate. It is given by:

$$\widehat{\mathrm{ISE}}_i = \sum_{j=1}^{m} (\mathrm{Y}_i(t_j) - \widehat{\mathrm{Y}}_i^{(-i)}(t_j))^2. \tag{26}$$

6.1 Canadian Weather data

The Canadian Weather data set consists of $m = 365$ daily temperature measurements (average over the year 1961 to 1994) at $n = 35$ weather stations in Canada, and their corresponding daily precipitation (in log scale). Our goal is to predict the (log) daily precipitations functions $\mathrm{Y}_i(t)$ using its corresponding temperatures $\mathrm{X}_i(t)$, for $t \in [0, 365]$.

Figure 8 displays the raw temperatures and their cubic B-splines smoothing with $\mathrm{L_X} = 100$ basis functions and equispaced knots. Figure 9 shows the raw log precipitations profiles to predict.

Following the target to obtain smooth estimates of parameter curves (or surfaces) and accurate predictions, we must correctly choose the number of basis functions of functional parameters without forgetting that the number of parameters of the

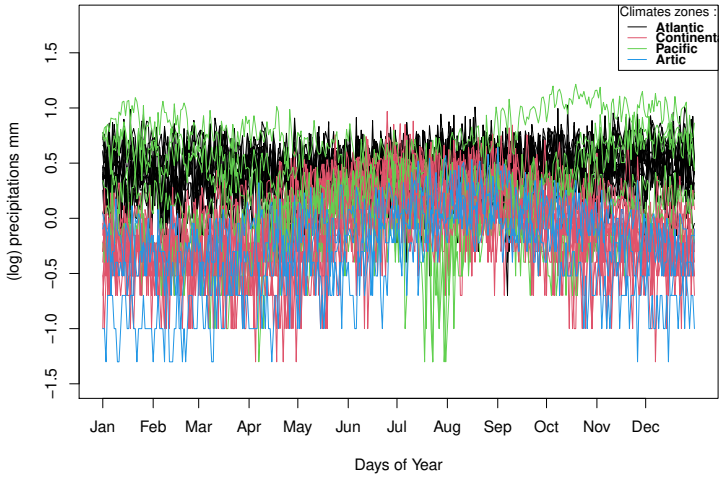


Fig. 9: raw log precipitations profiles of the 35 Canadian weather stations.

simple model is nearly multiplied by the number of components to get the number of parameters of MoE models. So we set L_β the number of basis functions to 8 both for FFMoE and PenFFMoE. And for the non-mixture models (PenFFR and pffr), we set L_β to 40. The penalty parameters λ_0 and λ_1 for the intercept and temperature effect are selected using cross-validation on a predefined grid of values (3 equispaced values between 0 and 0.5). Model selection is made using the BIC criterion for each LOO model with the number of expert components K in the set $\{1, 2, 3, 4, 5\}$. We observe in Table 3 that both for FFMoE and the PenFFMoE, the number of experts component the most often selected is $K = 4$. The same situation is observed between the two methods due to the fact that the cross-validation procedure leads to selecting mostly a null value of λ.

Table 4 shows the average value of $\widehat{\text{ISE}}_i$, the standard deviation and the median over the $n = 35$ weather stations. It is important to recall that the statistics are computed over LOO cross-validation, so on different model estimations (including the

Number of components	1	2	3	4	5
FFMoE	0%	0%	20.00%	51.43%	28.57%
PenFFMoE	0%	0%	11.43%	65.71%	22.86%

Table 3: Proportion of number of experts per model obtained by BIC selection.

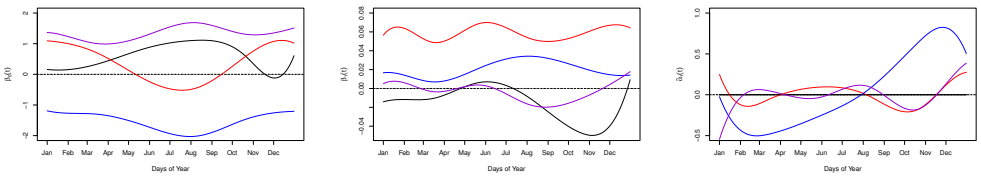
Methods	average $\widehat{\text{ISE}}$	sd $\widehat{\text{ISE}}$	median $\widehat{\text{ISE}}$
PenFFMoE	**29.91**	**21.07**	22.83
FFMoE	30.00	30.37	21.17
PenFFR	36.40	40.42	**21.04**
pffr	89.51	52.06	71.22

Table 4: Average, standard deviation and median of $\widehat{\text{ISE}}_i$ for the Canadian Weather data set. The best result is in boldface.

choice of K). We note a little enhancement in the predictive quality of the mixture models (PenFFMoE, FFMoE) compared to the models without mixture (PenFFR, pffr), and also a smaller inter-individual variance.

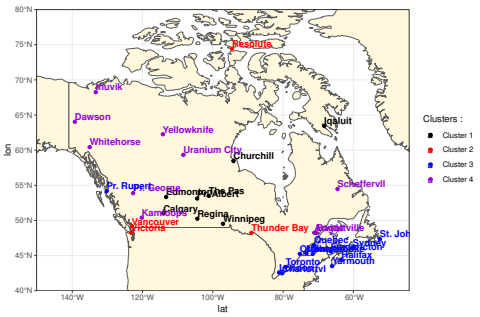
Another advantage of mixture models is the interpretation of the mixture component belonging. For this, new estimations of PenFFMoE and FFMoE are performed on the whole data set. The BIC criterion selects K = 3 components for FFMoE and K = 4 for PenFFMoE. Figure 10 shows the regression coefficient $\hat{\beta}_k(t)$ and gated network parameters $\hat{\alpha}_k(t)$ for the PenFFMoE version. Note that for the gated network parameters, we only have $K-1$ curves due to the identifiability condition, which imposes that $\alpha_1(t) = 0$. We also observed that PenFFMoE parameters are slightly smoother than for FFMoE parameters (see Appendix D). This led to a better highlighting of all components of the impact of temperatures on precipitations at different times of the year.

Figure 11a plots the geographical positions of the stations. We note a high correlation with the four climate zones of Canada, which is confirmed by the confusion matrices given in Table (11b). Finally, Figure 12 gives predictions for two randomly chosen weather stations (Churchill and Edmonton) and are compared with the actual precipitation.



(a) Intercepts (b) Expert parameters (c) Gated parameters

Fig. 10: Functional coefficients and gated network parameters obtained by PenFFMoE on Canadian Weather data. Color depends on group membership.

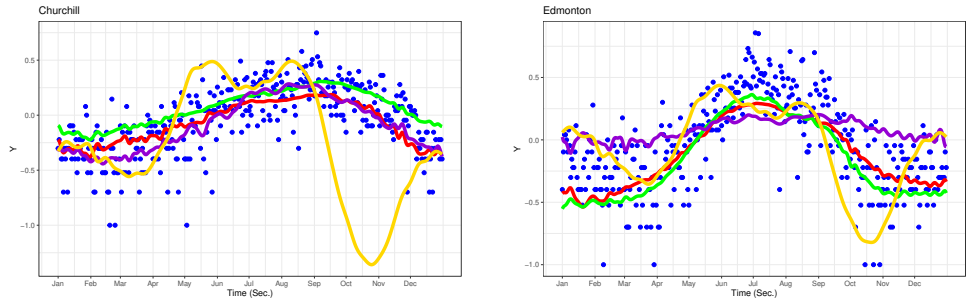


(a)

PenFFMoE	**Clusters**			
	1	2	3	4
Atlantic	0	0	12	3
Continental	7	1	0	4
Pacific	0	2	1	2
Artic	1	1	0	1

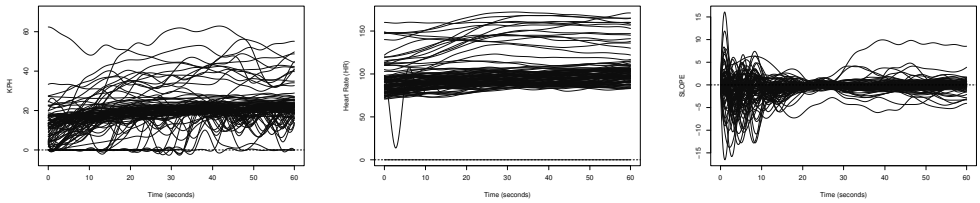
(b)

Fig. 11: Geographic visualization of the 35 weather stations clustering by PenFFMoE and confusion matrix between clusters and climates zones.



(a) Churchill station (b) Edmonton station

Fig. 12: Prediction for two randomly chosen stations. Blue points are the actual data, red and green lines are the predictions for FFMoE and PenFFMoE resp. The violet and gold lines are the predictions given by PenFFR and pffr resp.



(a) FD KPH (b) FD HR (c) FD SLOPE

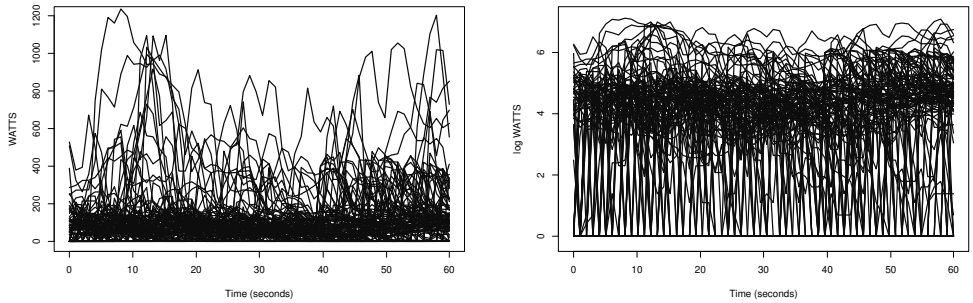
Fig. 13: Raw and functional expansion curves of speed (a,d), heart rate (b,e) and slope (c,f) for 100 cyclists.

6.2 Cycling Data

The Cycling data set, initially studied in [44], contains the measurements of several parameters during 216 cycling sessions of 30 minutes. The parameters are the power developed by the cyclist (in watts), its heart rate (in beats per minute), the pedalling cadence (in rotation per minute), the speed (in km/h), the slope (in percentage), the outdoor temperature (in Celsius degree) the altitude (in meters). The sampling rate is one measure per second. Our goal in this study is to predict the developed power according to the three parameters known to have an impact [44]: speed (KPH), heart rate (HR) and slope (SLOPE). Due to the high variability of these parameters during a period of 30 minutes, we restrict our analysis to a small portion of the curve, corresponding to the 20^{th}-minute (chosen arbitrarily).

Figure 13 shows the functional expansion in cubic B-splines with $\mathrm{L_X} = 50$ basis functions and equispaced knots for the three covariates. Figure 14a plots the developed power. Due to its dispersion, a logarithmic transformation is applied (Figure 14b).

We evaluate on this data set FFMoE, PenFFMoE, PenFFR and pffr. Predictive performances are evaluated through the ISE. The data set is split into train and test subsets with proportions 80% and 20%. The number of components for FFMoE, PenFFMoE is made using BIC with K in the set $\{1, 2, \ldots, 15\}$. The number of basis functions of both expert parameters L_β and gated parameters L_α are set to 10. For the



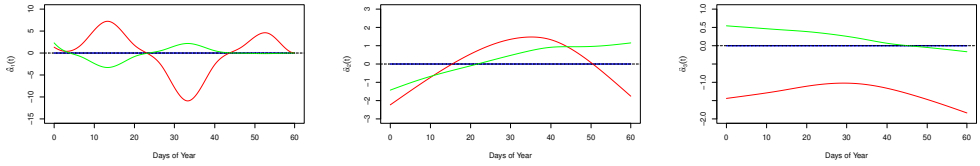
(a) raw WATTS (b) logarithmic transformation

Fig. 14: Power developed by 100 cyclists and the corresponding logarithmic transformation.

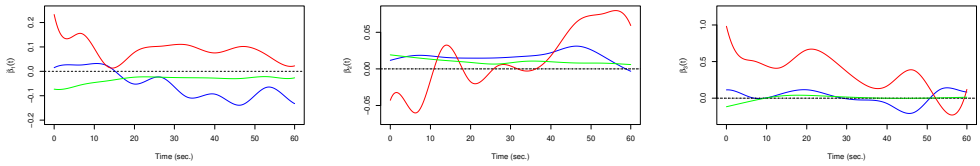
non mixture models PenFFR and pffr, the number of basis functions for parameters are set to 15.

We obtained K = 4 for FFMoE and K = 3 for PenFFMoE, with a better BIC for FFMoE. Figure 15 shows the gated and expert parameters for PenFFMoE, which allows interesting interpretation. For instance, for the green cluster, the effect of the three features is almost constant, which means that the cyclist has a regular effort, with regular speed, heart rate and slope. On the contrary, for the blue cluster, the effect of KPH goes from positive to negative, whereas the effect of HR remains positive: probably that this session corresponds to an end of a climb: during the climb, the cyclist goes slowly whereas developing a high power and high HR, and then, after the summit of the climb, keeping a high power allows him to go fastly with a decreasing HR. Figure E3 in Appendix E shows the same results for FFMoE method.

Table 5 presents the average and standard deviation of ISE (over the test set) for the different models. If we consider the ISE averaged over the individuals of the test set, the best results are obtained with pffr. But looking at the median ISE, we conclude that most individuals are better predicted with the PenFFMoE method. This is in particular confirmed by Figure 16, which plots the predictions on two randomly



(a) KPH gated parameters (b) HR gated parameters (c) SLOPE gated parameters



(d) KPH's expert parameters (e) HR's expert parameters (f) SLOPE's exp. parameters

Fig. 15: Functional gated (first row) and expert (second row) parameters obtained by PenFFMoE on Cycling data. corresponding colors matched for the same cluster.

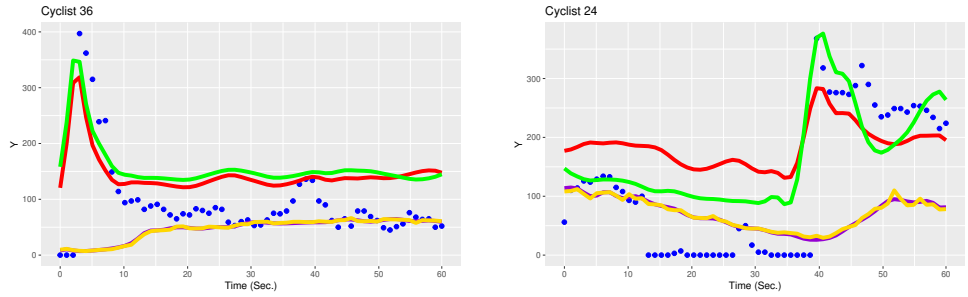
Methods	BIC	Nb clusters	Average ISE	sd $\widehat{ISE}$	median $\widehat{ISE}$
PenFFMoE	26729.9	3	160.72	225.64	**34.63**
FFMoE	26275.2	4	155.20	202.94	47.66
PenFFR		/	155.07	181.85	47.31
pffr		/	**154.78**	**181.61**	46.82

Table 5: The average and standard deviation of $\widehat{\text{ISE}}$ for Cycling data set. The best result is in boldface.

chosen cycling sessions, on which we can see that the prediction with FFMoE and PenFFMoE better follow the general shape of the curves.

7 Conclusion

Functional data analysis has now reached a high level of maturity and its manifold applications span a wide range of scientific fields. In the present paper we developed a novel estimation scheme for MoE in the framework where both covariates and response are of functional type using the concurrent linear model with Gaussian error. Preliminary investigations based on plain maximum likelihood estimation, and using functional expansions in standard bases, lead us to the observation of a



(a) (b)

Fig. 16: Prediction on two randomly chosen cycling sessions. Blue points are the actual data, and red and green lines are the predictions for FFMoE and PenFFMoE resp. The violet and gold lines are the predictions given by PenFFR and pffr resp.

lack of smoothness of the estimators in various experiments with real-world datasets. In order to circumvent these issues, we introduced a ridge-type penalisation on the second derivatives and obtained a more stable estimator, still capable of handling substantial variability of first-order behaviours. Numerical experiments showed that the FFMoE (also PenFFMoE) has satisfactory behaviour in terms of parameter estimation (interpretability) and predictive accuracy on simulated datasets.

We then illustrate this performance on two real-world datasets. On Canadian weather dataset, PenFFMoE and FFMoE cluster the weather stations in $K = 4$ clusters that match the various climate zones. The predictive accuracy shows a definite advantage of mixture of experts over non-mixture based models. On Cycling data, the predictive quality is certainly not as good as non-mixture models, but it gives predictions that detect regime changes more easily.

Extensions of this work are potentially manifold. One possible avenue is to explore the more general exponential family on the functional response side. A second possible direction would be to investigate possible solutions for producing relevant prediction bounds, using for instance conformal prediction [45] which has attracted great interest lately in the machine learning community.

Appendix A EM for the FFMoE

Given the complete data log-likelihood and the parameters at current iteration l, we define the Q function for the EM algorithm defined by:

$$\mathrm{Q}(\Psi^{(l+1)} \,|\, \Psi^{(l)}) \,=\, \mathbb{E}(\mathcal{L}_c(\Psi^{(l+1)}; \{y_i(t_j), \mathrm{x}_i(t_j)\}_{i,j}) \,|\, \{y_i(t_j), \mathrm{x}_i(t_j)\}_{i,j};\ \Psi^{(l)})$$

Now we are going to describe the EM algorithm for maximizing (15):

- **E-step:**

  At this step, we compute the conditional expectation of the log-likelihood given the observed data and the current parameter (at iteration l) estimation $\Psi^{(l)}$. This is equivalent to update the posterior probabilities $p_{ik}^{(l)}$ that the curves $\mathrm{x}_i(t)$ belongs to the k^{th} component of the mixture under the current model:

$$p_{ik}^{(l)} \,=\, \mathbb{E}(z_{ik} \,|\, \{y_i(t_j), \mathrm{x}_i(t_j)\}_{i,j}; \Psi^{(l)}) = \mathbb{P}(z_{ik} = 1 \,|\, \{y_i(t_j), \mathrm{x}_i(t_j)\}_{i,j}; \Psi^{(l)}).$$

  Using Bayes' theorem, the conditional probability $p_{ik}^{(l)}$ can be expressed as:

$$\begin{aligned} p_{ik}^{(l)} \,&=\, \frac{\mathbb{P}(z_{ik} = 1)\, \mathbb{P}(\{y_i(t_j), \mathrm{x}_i(t_j)\}_{i,j};\ \Psi^{(l)} \,|\, z_{ik} = 1)}{\mathbb{P}(\{y_i(t_j), \mathrm{x}_i(t_j)\}_{i,j};\ \Psi^{(l)})} \\ &= \frac{\mathbb{P}(z_{ik} = 1)\, \mathbb{P}(\{y_i(t_j), \mathrm{x}_i(t_j)\}_{i,j};\ \Psi^{(l)} \,|\, z_{ik} = 1)}{\sum_{u=1}^{\mathrm{K}} \mathbb{P}(z_{iu} = 1) \mathbb{P}(\{y_i(t_j), \mathrm{x}_i(t_j)\}_{i,j};\ \Psi^{(l)} \,|\, z_{iu} = 1)} \\ p_{ik}^{(l)} \,&=\, \frac{\pi_k(\mathrm{x}_i(t), \alpha_k^{(l)}(t))\, \Phi_m(\mathbf{y}_i\,;\, b_k^{\top(l)}\mathrm{R}_i, \mathrm{V}_{k,i}^{(l)})}{\sum_{u=1}^{\mathrm{K}} \pi_u(\mathrm{x}_i(t), \alpha_u^{(l)}(t))\, \Phi_m(\mathbf{y}_i\,;\, b_u^{\top(l)}\mathrm{R}_i, \mathrm{V}_{u,i}^{(l)})} \end{aligned} \tag{A1}$$

- **M-step:**

  Given the previous posterior probability and the observed data, this step updates the current parameters $\Psi^{(l)}$ by maximizing the complete (data) log-likelihood, that

is $\Psi^{(l+1)}$:

$$
\begin{aligned}
\mathrm{Q}(\Psi^{(l+1)} \,|\, \Psi^{(l)}) &= \mathbb{E}(\mathcal{L}_c(\Psi^{(l+1)}; \{y_i(t_j), \mathrm{x}_i(t_j)\}_{i,j}) \,|\, \{y_i(t_j), \mathrm{x}_i(t_j)\}_{i,j};\ \Psi^{(l)}) \\
&= \mathbb{E}(\sum_{i=1}^{n}\sum_{k=1}^{\mathrm{K}} z_{ik}\ \log(\frac{\exp({a_k^{(l+1)}}^{\top} r_i)}{1+\sum_{u=1}^{\mathrm{K}-1}\exp({a_u^{(l+1)}}^{\top} r_i)} \cdot \frac{1}{\sqrt{(2\pi)^m |\mathrm{V}_{k,i}^{(l+1)}|}} \\
&\quad \exp(-\frac{1}{2}(\mathbf{y}_i - {b_k^{(l+1)}}^{\top}\mathrm{R}_i)^{\top}\mathrm{V}_{k,i}^{-1^{(l+1)}}(\mathbf{y}_i - {b_k^{(l+1)}}^{\top}\mathrm{R}_i))) \\
&\qquad\qquad \Big| \{y_i(t_j), \mathrm{x}_i(t_j)\}_{i,j};\ \Psi^{(l)}) \\
&= \sum_{i=1}^{n}\sum_{k=1}^{\mathrm{K}} \mathbb{E}(z_{ik} \,|\, \{y_i(t_j), \mathrm{x}_i(t_j)\}_{i,j}; \Psi^{(l)})\ \log(\frac{\exp({a_k^{(l+1)}}^{\top} r_i)}{1+\sum_{u=1}^{\mathrm{K}-1}\exp({a_u^{(l+1)}}^{\top} r_i)} \\
&\quad \frac{1}{\sqrt{(2\pi)^m |\mathrm{V}_{k,i}^{(l+1)}|}} \exp(-\frac{1}{2}(\mathbf{y}_i - {b_k^{(l+1)}}^{\top}\mathrm{R}_i)^{\top}\mathrm{V}_{k,i}^{-1^{(l+1)}}(\mathbf{y}_i - {b_k^{(l+1)}}^{\top}\mathrm{R}_i))))
\end{aligned}
$$

$$
\begin{aligned}
= \underbrace{\sum_{i=1}^{n}\sum_{k=1}^{\mathrm{K}} p_{ik}^{(l)}\ \log(\frac{\exp({a_k^{(l+1)}}^{\top} r_i)}{1+\sum_{u=1}^{\mathrm{K}-1}\exp({a_u^{(l+1)}}^{\top} r_i)})}_{\mathrm{Q}_1(a_k^{(l+1)} \,|\, \Psi^{(l)})} \quad + \\
\underbrace{\sum_{i=1}^{n}\sum_{k=1}^{\mathrm{K}} p_{ik}^{(l)}\ \log(\frac{1}{\sqrt{(2\pi)^m |\mathrm{V}_{k,i}^{(l+1)}|}} \exp(-\frac{1}{2}(\mathbf{y}_i - {b_k^{(l+1)}}^{\top}\mathrm{R}_i)^{\top}\mathrm{V}_{k,i}^{-1^{(l+1)}}(\mathbf{y}_i - {b_k^{(l+1)}}^{\top}\mathrm{R}_i)))}_{\mathrm{Q}_2(b_k^{(l+1)},\ \mathrm{V}_k^{(l+1)} \,|\, \Psi^{(l)})}
\end{aligned}
$$

$$
\mathrm{Q}(\Psi^{(l+1)} \,|\, \Psi^{(l)}) = \mathrm{Q}_1(a_k^{(l+1)} \,|\, \Psi^{(l)}) + \mathrm{Q}_2(b_k^{(l+1)},\ \mathrm{V}_k^{(l+1)} \,|\, \Psi^{(l)}).
$$

The global maximization problem is split onto two separate maximization problems: the updating of gated network parameters via the maximization of the function $\mathrm{Q}_1(a_k^{(l+1)} \,|\, \Psi^{(l)})$ and the updating of experts parameters via the maximization of the function $\mathrm{Q}_2(b_k^{(l+1)},\ \sigma_k^{2\ (l+1)} \,|\, \Psi^{(l)})$. It obsvious to recognise in each of these two expressions the likelihood of the multinomial logistic model $\mathrm{Q}_1(.)$ and the linear

gaussian model $\mathrm{Q}_2(.)$ for which we know how to calculate (at least numerically through Newton-Raphson method for e.g) MLEs.

Appendix B EM for PenFFMoE

- **E-step:**

  Same as in non penalize case

- **M-step:**

  Given the previous posterior probability and the observed data, this step updates the current parameters $\Psi^{(l)}$ by maximizing the penalized complete (data) log-likelihood, that is $\Psi^{(l+1)}$. We define:

$$
\begin{aligned}
\mathrm{Q}_{pen}(\Psi^{(l+1)} \mid \Psi^{(l)}) &= \mathbb{E}(\mathcal{L}^c_{pen}(\Psi^{(l+1)}; \{y_i(t_j), \mathrm{x}_i(t_j)\}_{i,j}) \Big| \{y_i(t_j), \mathrm{x}_i(t_j)\}_{i,j};\ \Psi^{(l)}) \\
&= \mathbb{E}(\sum_{i=1}^{n}\sum_{k=1}^{\mathrm{K}} z_{ik}\ \log(\frac{\exp({a_k^{(l+1)}}^\top r_i)}{1+\sum_{u=1}^{\mathrm{K}-1}\exp({a_u^{(l+1)}}^\top r_i)} \cdot \frac{1}{\sqrt{(2\pi)^m|\mathrm{V}_{k,i}^{(l+1)}|}} \\
&\quad \exp(-\frac{1}{2}(\mathbf{y}_i - {b_k^{(l+1)}}^\top \mathrm{R}_i)^\top \mathrm{V}_{k,i}^{-1^{(l+1)}}(\mathbf{y}_i - {b_k^{(l+1)}}^\top \mathrm{R}_i))) - \\
&\quad \sum_{k=1}^{\mathrm{K}} {b_k^{(l+1)}}^\top(\lambda_k \mathrm{P})\ b_k^{(l+1)} - \sum_{k=1}^{\mathrm{K}-1} {a_k^{(l+1)}}^\top(\gamma_k \mathrm{Q})\ a_k^{(l+1)} \\
&\qquad\qquad \Big| \{y_i(t_j), \mathrm{x}_i(t_j)\}_{i,j};\ \Psi^{(l)}) \\
&= \sum_{i=1}^{n}\sum_{k=1}^{\mathrm{K}} \mathbb{E}(z_{ik} \mid \{y_i(t_j), \mathrm{x}_i(t_j)\}_{i,j}; \Psi^{(l)})\ \log(\frac{\exp({a_k^{(l+1)}}^\top r_i)}{1+\sum_{u=1}^{\mathrm{K}-1}\exp({a_u^{(l+1)}}^\top r_i)} \\
&\quad \frac{1}{\sqrt{(2\pi)^m|\mathrm{V}_{k,i}^{(l+1)}|}} \exp(-\frac{1}{2}(\mathbf{y}_i - {b_k^{(l+1)}}^\top \mathrm{R}_i)^\top \mathrm{V}_{k,i}^{-1^{(l+1)}}(\mathbf{y}_i - {b_k^{(l+1)}}^\top \mathrm{R}_i))))) \\
&\quad -\sum_{k=1}^{\mathrm{K}} {b_k^{(l+1)}}^\top(\lambda_k \mathrm{P})\ b_k^{(l+1)} - \sum_{k=1}^{\mathrm{K}-1} {a_k^{(l+1)}}^\top(\gamma_k \mathrm{Q})\ a_k^{(l+1)}
\end{aligned}
$$

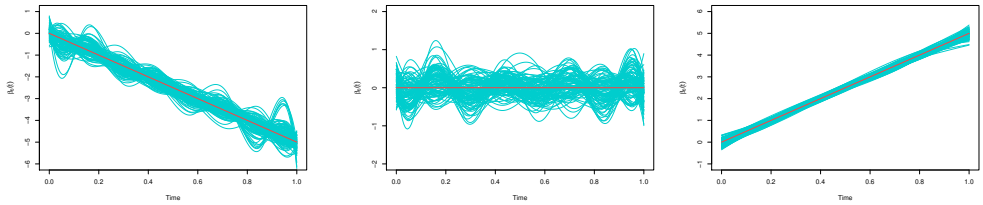
$$= \underbrace{\sum_{i=1}^{n}\sum_{k=1}^{\mathrm{K}} p_{ik}^{(l)}\ \log(\frac{\exp(a_k^{(l+1)^\top} r_i)}{1+\sum_{u=1}^{\mathrm{K}-1}\exp(a_u^{(l+1)^\top} r_i)})}_{\mathrm{Q}_1(a_k^{(l+1)}\,|\,\Psi^{(l)})} - \sum_{k=1}^{\mathrm{K}-1} a_k^{(l+1)^\top}(\gamma_k \mathrm{Q})\ a_k^{(l+1)} -$$

$$\sum_{k=1}^{\mathrm{K}} b_k^{(l+1)^\top}(\lambda_k \mathrm{P})\ b_k^{(l+1)} +$$

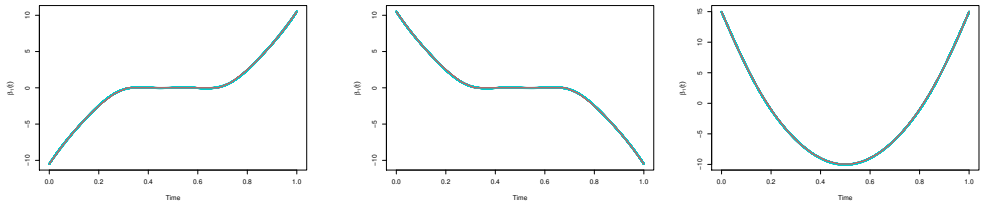
$$\underbrace{\sum_{i=1}^{n}\sum_{k=1}^{\mathrm{K}} p_{ik}^{(l)}\ \log(\frac{1}{\sqrt{(2\pi)^m|\mathrm{V}_{k,i}^{(l+1)}|}}\exp(-\frac{1}{2}(\mathbf{y}_i - b_k^{(l+1)^\top}\mathrm{R}_i)^\top \mathrm{V}_{k,i}^{-1^{(l+1)}}(\mathbf{y}_i - b_k^{(l+1)^\top}\mathrm{R}_i)))}_{\mathrm{Q}_2(b_k^{(l+1)},\ \mathrm{V}_k^{(l+1)}\,|\,\Psi^{(l)})}$$

$$\begin{aligned}\mathrm{Q}(\Psi^{(l+1)}\,|\,\Psi^{(l)}) &= \underbrace{\mathrm{Q}_1(a_k^{(l+1)}\,|\,\Psi^{(l)}) - \sum_{k=1}^{\mathrm{K}-1} a_k^{(l+1)^\top}(\gamma_k \mathrm{Q})\ a_k^{(l+1)}}\ + \\ &\qquad \underbrace{\mathrm{Q}_2(b_k^{(l+1)},\ \mathrm{V}_k^{(l+1)}\,|\,\Psi^{(l)}) - \sum_{k=1}^{\mathrm{K}} b_k^{(l+1)^\top}(\lambda_k \mathrm{P})\ b_k^{(l+1)}} \\ &= \mathrm{Q}_{1,\,pen}(a_k^{(l+1)}\,|\,\Psi^{(l)})\ +\ \mathrm{Q}_{2,\,pen}(b_k^{(l+1)},\ \sigma_k^{2\ (l+1)}\,|\,\Psi^{(l)}).\end{aligned}$$

Appendix C Parameters in simulation study



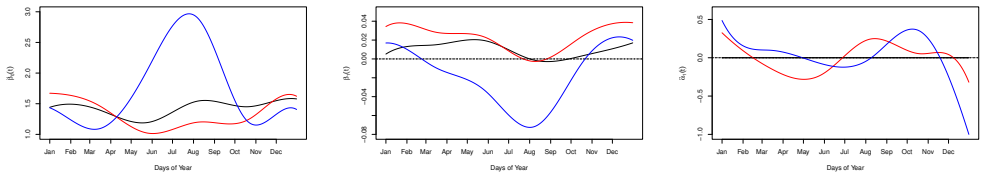
(a) $\beta_{1,0}(t)$ (b) $\beta_{2,0}(t)$ (c) $\beta_{3,0}(t)$



(d) $\beta_{1,1}(t)$ (e) $\beta_{2,1}(t)$ (f) $\beta_{3,1}(t)$

Fig. C1: Estimation of the regression coefficients for Scenario S3 with PenFFMoE. The red curves are the actual parameters, the cyan curves are the estimation.

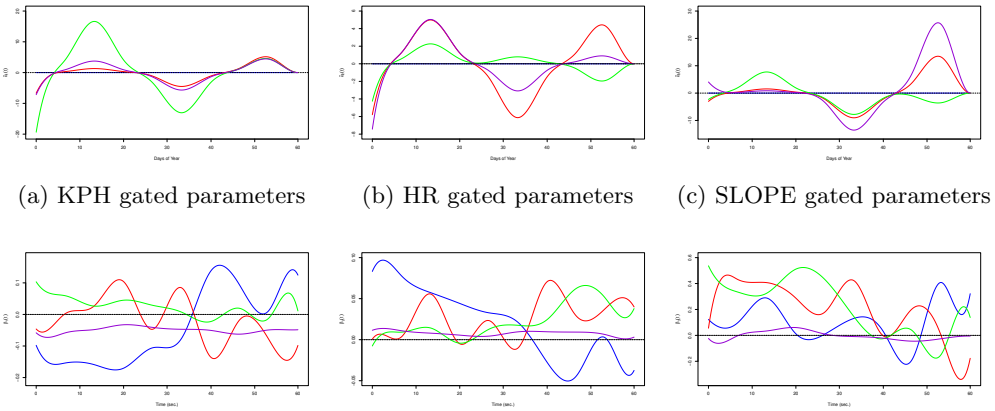
Appendix D Estimators for Canadian weather data



(a) Intercepts (b) Expert parameter (c) gated parameters

Fig. D2: Functional coefficients and gated network parameters obtained by FFMoE on Canadian Weather data. Color depends on group membership.

Appendix E Estimators for Cycling data



(a) KPH gated parameters (b) HR gated parameters (c) SLOPE gated parameters

(d) KPH's expert parameters (e) HR's expert parameters (f) SLOPE's exp. parameters

Fig. E3: Functional gated (first row) and expert (second row) parameters obtained by FFMoE on Cycling data. corresponding colors matched for the same cluster.

References

[1] Ramsay, J., Silverman, B.W.: Functional Data Analysis, 2nd edn. Springer Series in Statistics. Springer, New York (2005). https://doi.org/10.1007/b98888

[2] Ramsay, J.O., Hooker, G., Graves, S.: Functional Data Analysis with R and MATLAB, 1st edn. Springer, New York (2009)

[3] Horváth, L., Kokoszka, P.: Inference for Functional Data with Applications. Springer Series in Statistics. Springer, New York (2012). https://doi.org/10.1007/978-1-4614-3655-3 . https://www.springer.com/gp/book/9781461436546 Accessed 2021-05-27

[4] Kokoszka, P., Reimherr, M.: Introduction to Functional Data Analysis, (2017). https://doi.org/10.1201/9781315117416

[5] Goldsmith, J., Bobb, J., Crainiceanu, C.M., Caffo, B., Reich, D.: Penalized functional regression. Journal of Computational and Graphical Statistics **20**(4), 830–851 (2011) https://doi.org/10.1198/jcgs.2010.10007 https://doi.org/10.1198/jcgs.2010.10007

[6] Morris, J.: Functional regression. Annual Review of Statistics and Its Application **2** (2014) https://doi.org/10.1146/annurev-statistics-010814-020413

[7] Ivanescu, A., Staicu, A.-M., Scheipl, F., Greven, S.: Penalized function-on-function regression. Computational Statistics **30**(2), 539–568 (2015) https://doi.org/10.1007/s00180-014-0548-4

[8] Luo, R., Qi, X., Wang, Y.: Functional wavelet regression for linear function-on-function models. Electronic Journal of Statistics **10**(2), 3179–3216 (2016) https://doi.org/10.1214/16-EJS1204

[9] Li, Y., Ruppert, D.: On the asymptotics of penalized splines. Biometrika **95**(2), 415–436 (2008)

[10] James, G.M., Wang, J., Zhu, J.: Functional linear regression that's interpretable. Annals of Statistics **37**(5A), 2083–2108 (2009)

[11] Tamo Tchomgui, J.S., Jacques, J., Barriac, V., Fraysse, G., Chrétien, S.: A Penalized Spline Estimator for Functional Linear Regression with Functional Response. working paper or preprint (2023). https://hal.science/hal-04120709

[12] DeSarbo, W.S., Cron, W.L.: A maximum likelihood methodology for clusterwise linear regression. Journal of Classification **5**(2), 249–282 (1988)

[13] McLachlan, G.J., Peel, D.: Finite Mixture Models. Probability and Statistics – Applied Probability and Statistics Section, vol. 299. Wiley, New York (2000)

[14] Makov, E., Titterington, D.M., Smith, A.F.M.: Statistical Analysis of Finite Mixture Distributions, 1st edn. wiley, New-York (1985)

[15] Lindsay, B.G.: Mixture Models: Theory, Geometry, and Applications. NSF-CBMS regional conference series in probability and statistics, Institute of Mathematical Statistics (1995). https://books.google.fr/books?id=VFDzNhikFbQC

[16] Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society: Series B (Methodological) **39**(1), 1–22 (1977)

[17] McLachlan, G.J., Krishnan, T.: The EM Algorithm and Extensions. Wiley, Series in Probability and Statistics (2007). https://books.google.fr/books?id=NBawzaWoWa8C

[18] Celeux, G., Chrétien, S., Forbes, F., Mkhadri, A.: A component-wise em algorithm

for mixtures. Journal of Computational and Graphical Statistics **10**(4), 697–712 (2001)

[19] Chrétien, S., Hero, A.: Acceleration of the em algorithm via proximal point iterations. In: Proceedings. 1998 IEEE International Symposium on Information Theory (Cat. No. 98CH36252), p. 444 (1998). IEEE

[20] Chrétien, S., Hero, A.O.: Kullback proximal algorithms for maximum-likelihood estimation. IEEE transactions on information theory **46**(5), 1800–1810 (2000)

[21] Chrétien, S., Hero, A., Perdry, H.: Space alternating penalized kullback proximal point algorithms for maximizing likelihood with nondifferentiable penalty. Annals of the Institute of Statistical Mathematics **64**, 791–809 (2012)

[22] Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive Mixtures of Local Experts. Neural Computation **3**(1), 79–87 (1991) https://doi.org/10.1162/neco.1991.3.1.79

[23] Nguyen, H.D., Chamroukhi, F.: Practical and theoretical aspects of mixture-of-experts modeling: An overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1246 (2018) https://doi.org/10.1002/widm.1246

[24] Chamroukhi, F., Pham, N.T., Hoang, V.H., McLachlan, G.J.: Functional mixtures-of-experts. Statistics and Computing **34**(98) (2024)

[25] Hida, T., Hui-Hsiung, K., Potthoff, J., Streit, L.: White Noise: An Infinite Dimensional Calculus. Mathematics and its applications, Kluwer Academic Publishers (1993). https://books.google.fr/books?id=-XitQgAACAAJ

[26] Hastie, T., Tibshirani, R.: Varying-coefficient models. Journal of the Royal Statistical Society. Series B (Methodological) **55**(4), 757–796 (1993). Accessed 2022-12-13

[27] Wood, S.N.: On confidence intervals for generalized additive models based on penalized regression splines. Australian & New Zealand Journal of Statistics **48**(4), 445–464 (2006)

[28] Dayton, C.M., Macready, G.B.: Concomitant-variable latent-class models. Journal of the American Statistical Association **83**(401), 173–178 (1988)

[29] Young, D.S., Hunter, D.R.: Mixtures of regressions with predictor-dependent mixing proportions. Computational Statistics & Data Analysis **54**(10), 2253–2266 (2010) https://doi.org/10.1016/j.csda.2010.04.002

[30] Mousavi, S., Sørensen, H.: Functional logistic regression: a comparison of three methods. Journal of Statistical Computation and Simulation **88**, 1–19 (2018) https://doi.org/10.1080/00949655.2017.1386664

[31] Berrendero, J.R., Bueno-Larraz, B., Cuevas, A.: On functional logistic regression: some conceptual issues. TEST: An Official Journal of the Spanish Society of Statistics and Operations Research **32**(1), 321–349 (2023) https://doi.org/10.1007/s11749-022-00836-

[32] Jiang, W., Tanner, M.A.: Hierarchical mixtures-of-experts for exponential family regression models: Approximation and maximum likelihood estimation. Annals of Statistics **27**(3), 987–1011 (1999) https://doi.org/10.1214/aos/1018031265

[33] Jordan, M., Jacobs, R.: Hierarchical mixtures of experts and the. Neural computation **6**, 181 (1994)

[34] Peng, F., Jacobs, R.A., Tanner, M.A.: Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition. Journal of the American Statistical Association **91**(435), 953–960 (1996). Accessed 2023-10-03

[35] Neal, R.M., Hinton, G.E.: A view of the em algorithm that justifies incremental, sparse, and other variants. In: Learning in Graphical Models, Springer, pp. 355–368 (1998)

[36] Fessler, J.A., Hero, A.O.: Space-alternating generalized expectation-maximization algorithm. IEEE Transactions on signal processing **42**(10), 2664–2677 (1994)

[37] Grün, B., Leisch, F.: Flexmix version 2: Finite mixtures with concomitant variables and varying and constant parameters. Journal of Statistical Software **28**(4), 1–35 (2008) https://doi.org/10.18637/jss.v028.i04

[38] Grün, B., Leisch, F.: Identifiability of Finite Mixtures of Multinomial Logit Models with Varying and Fixed Effects. Journal of Classification **25**(2), 225–247 (2008) https://doi.org/10.1007/s00357-008-9022-8

[39] Akaike, H.: A new look at the statistical model identification. IEEE Transactions on Automatic Control **19**(6), 716–723 (1974) https://doi.org/10.1109/TAC.1974.1100705

[40] Schwarz, G.: Estimating the Dimension of a Model. Annals of Statistics **6**(2), 461–464 (1978) https://doi.org/10.1214/aos/1176344136

[41] Leurgans, S.E., Moyeed, R.A., Silverman, B.W.: Canonical correlation analysis when the data are curves. Journal of the Royal Statistical Society **55**(3), 725–740 (1993) https://doi.org/10.1111/j.2517-6161.1993.tb01936.x

[42] Ramsay, J.O., Graves, S., Hooker, G.: Fda: Functional Data Analysis. (2022). R package version 6.0.5. https://CRAN.R-project.org/package=fda

[43] Samardzic, S.: FREG: Functional Regression Models. (2022). R package version 1.1

[44] Jacques, J., Samardzic, S.: Analyzing cycling sensors data through ordinal logistic regression with functional covariates. Journal of the Royal Statistical Society **71**(4), 969–986 (2022)

[45] Angelopoulos, A.N., Bates, S.: A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification (2022)