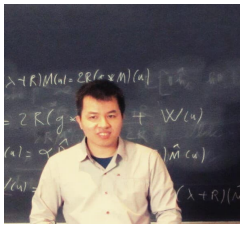
Mixtures-of-Exeperts with Functional Predictors

Faicel Chamroukhi

with NT Pham (Caen), VH Hoang (Vietnamese University), GJ McLachlan (UQ, Brisbane)


CMStatistics: The 14th International Conference of the ERCIM WG on Computational and Methodological Statistics, King's College London, 18-20 December 2021

PROJET FINANCÉ PAR L'ANR - PROJECT FUNDED BY THE ANR

Outline

Mixture-of-Experts [Jacobs et al., 1991, Jordan and Jacobs, 1994]

- Data : $\{X_i, Y_i\}_{i=1}^n$ a sample of n i.i.d random pairs with observed values $\{x_i, y_i\}_{i=1}^n$ where $Y_i \in \mathcal{Y}$ is the response for some covariates $X_i \in \mathcal{X}$.
- ME framework : Mixture-of-Experts explore the relationships of y given x via a fully conditional K-component mixture distribution of the form

$$\text{ME}(y|x) = \sum_{k=1}^{K} G_k(x) E_k(y|x)$$

$G_k(x)$: predictor-dependent mixing weights, referred to as gating network

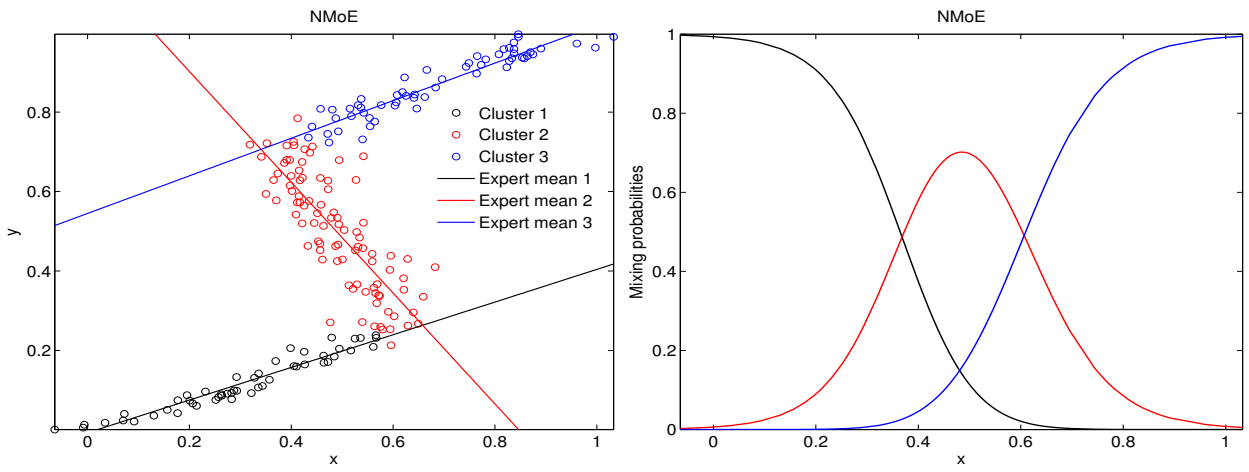
$E_k(y|x)$: conditional mixture components, referred to as experts network

Mixture-of-Experts for regression

- Data : $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$ an i.i.d sample of n observed values of a univariate response $Y \in \mathbb{R}$ for vector predictors $\boldsymbol{X} \in \mathbb{R}^p$
- ME model : $f(y|\boldsymbol{x}; \boldsymbol{\Psi}) = \sum_{k=1}^{K} \underbrace{\pi_k(\boldsymbol{x}; \mathbf{w})}_{\text{Gating network}} \underbrace{f_k(y|\boldsymbol{x}; \boldsymbol{\theta}_k)}_{\text{Expert Network}}$
- Softmax Gating network : $\pi_k(\boldsymbol{x}; \mathbf{w}) = \exp\left(w_{k0} + \boldsymbol{w}_k^T \boldsymbol{x}\right) \Big/ \sum_{\ell=1}^{K} \exp\left(w_{\ell 0} + \boldsymbol{w}_\ell^T \boldsymbol{x}\right)$
- Gaussian regression experts network : $f_k(y|\boldsymbol{x}; \boldsymbol{\theta}_k) = \phi\big(y; \mu(\boldsymbol{x}; \boldsymbol{\beta}_k), \sigma_k^2\big)$ with parametric (non-)linear regression functions $\mu(\boldsymbol{x}; \boldsymbol{\beta}_k)$

Maximum Likelihood Estimation via EM [Dempster et al., 1977, Jacobs et al., 1991]

- MLE : $\boldsymbol{\Psi}$ is commonly estimated by maximizing the observed-data log-likelihood : $\widehat{\boldsymbol{\Psi}}_n \in \arg\max_{\boldsymbol{\Psi}\in\Theta} L(\boldsymbol{\Psi})$ with $L(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k(\boldsymbol{x}_i;\boldsymbol{w}) f(\boldsymbol{y}_i|\boldsymbol{x}_i;\boldsymbol{\Psi}_k)$



Regularized MLE of the ME [Khalili, 2010] [Chamroukhi and Huynh, 2019]

$\boldsymbol{\Psi}$ is estimated by maximizing a penalized observed-data log-likelihood :

$$\widehat{\boldsymbol{\Psi}}_n \in \arg\max_{\boldsymbol{\Psi}\in\Theta} L(\boldsymbol{\Psi}) - \mathrm{Pen}_\lambda(\boldsymbol{\Psi})$$

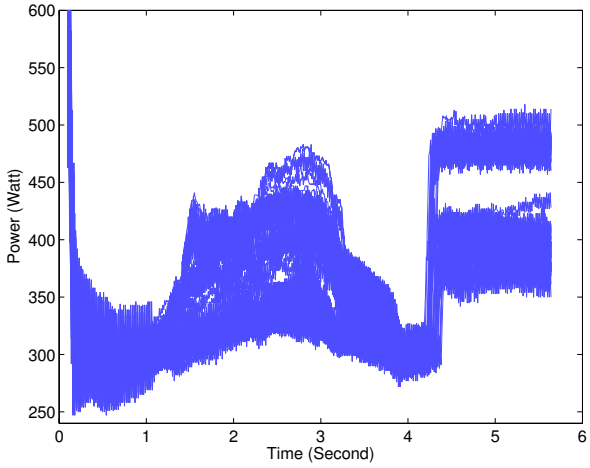
- $\hookrightarrow$ $\mathrm{Pen}_\lambda(\boldsymbol{\Psi})$ LASSO penalties for experts and the gating network
- encourages sparse solutions / performs parameter estimation and feature selection

$\hookrightarrow$ Doesn't apply to functional data (e.g functional inputs and/or outputs)

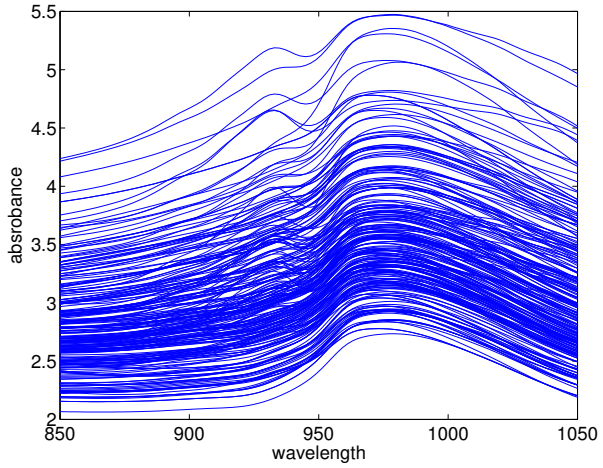
Functional data analysis context

Functional data are increasingly frequent

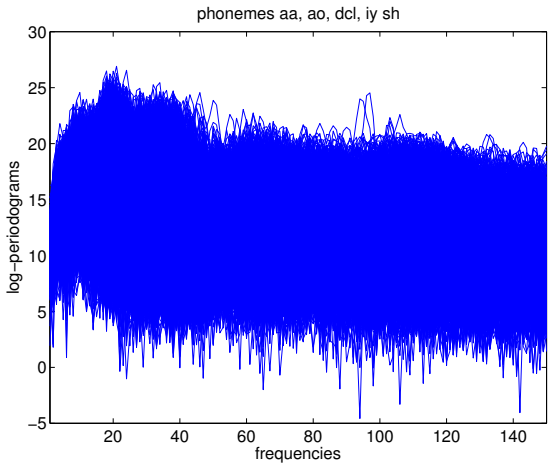
A broad literature : e.g. [Ramsay and Silverman, 2005, Chamroukhi and Nguyen, 2019]



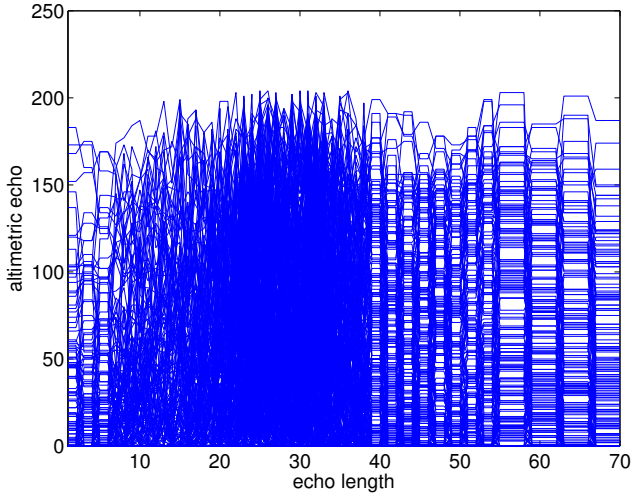
Railway time-series trajectories



Tecator data



Phonemes curves



Satellite waveforms

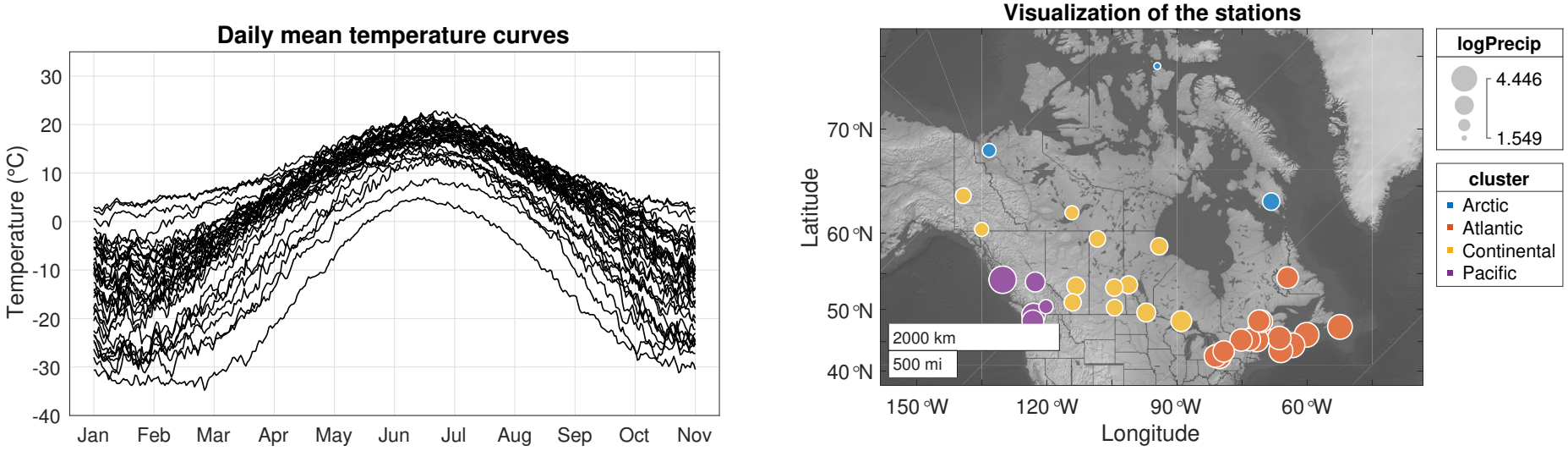


FIGURE – $n = 35$ daily mean temperature measurement curves (X_i's) in different stations (Left) and the $\log$ of precipitation values (Y_i's) visualized with the climate regions (Z_i's) (Right).

Mixtures-of-Experts with functional predictors

- ME to relate functional predictors $\{X(t) \in \mathbb{R}; t \in \mathcal{T} \subset \mathbb{R}\}$ to a scalar response $Y \in \mathcal{Y} \subset \mathbb{R}$
- Let $\{X_i(\cdot), Y_i\}_{i=1}^n$, be a random i.i.d sample from the pair $\{X(\cdot), Y)\}$

Stochastic representation of the FME model

Functional experts network

- The experts are formulated as functional regression models (see eg. James [2002])

$$Y_i = \beta_{z_i,0} + \int_{\mathcal{T}} X_i(t)\beta_{z_i}(t)dt + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

$z_i \in [K] = \{1, , \ldots, K\}$ is the unknown expert label for $(X_i(.), Y_i)$

$\beta_{z_i,0} \in \mathbb{R}$ is an unknown intercept coefficient of functional LR z_i

$\{\beta_{z_i}(t) \in \mathbb{R}; t \in \mathcal{T}\}$ is the unknown function of parameters of functional expert z_i

$\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2_{z_i})$ with $\sigma^2_{z_i} \in \mathbb{R}^+$ the variance of expert z_i

Stochastic representation of the FME model

Functional experts network

- The experts are formulated as functional regression models (see eg. James [2002])

$$Y_i = \beta_{z_i,0} + \int_{\mathcal{T}} X_i(t)\beta_{z_i}(t)dt + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

$z_i \in [K] = \{1, , \ldots, K\}$ is the unknown expert label for $(X_i(.), Y_i)$

$\beta_{z_i,0} \in \mathbb{R}$ is an unknown intercept coefficient of functional LR z_i

$\{\beta_{z_i}(t) \in \mathbb{R}; t \in \mathcal{T}\}$ is the unknown function of parameters of functional expert z_i

$\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2_{z_i})$ with $\sigma^2_{z_i} \in \mathbb{R}^+$ the variance of expert z_i

Functional gating network

- Multinomial logistic (softmax) functional gated network : For $z = 1, \cdots, K-1$:

$$\begin{aligned} h_z(x(\cdot)) &= \log\left(\frac{\mathbb{P}(Z = z|X(\cdot)}{\mathbb{P}(Z = K|X(\cdot)}\right) = \alpha_{z,0} + \int_{\mathcal{T}} X(t)\alpha_z(t)dt \\ \mathbb{P}(Z = z|\{X(t), t \in \mathcal{T}\}) &= \frac{\exp\left(\alpha_{z,0} + \int_{\mathcal{T}} X(t)\alpha_z(t)dt\right)}{1 + \sum_{z'=1}^{K-1} \exp\left(\alpha_{z',0} + \int_{\mathcal{T}} X(t)\alpha_{z'}(t)dt\right)}, \end{aligned} \tag{2}$$

- $\alpha_{z,0} \in \mathbb{R}$ is an unknown intercept parameter
- $\{\alpha_z(t) \in \mathbb{R}; t \in \mathcal{T}\}$ is the unknown function of parameters of gating network z

Representation of the functional predictors

$$
\begin{aligned}
Y_i &= \beta_{z_i,0} + \int_{\mathcal{T}} X_i(t)\beta_{z_i}(t)dt + \varepsilon_i, \quad i = 1, \ldots, n, \\
h_{z_i}(x_i(\cdot)) &= \alpha_{z_i,0} + \int_{\mathcal{T}} X_i(t)\alpha_{z_i}(t)dt.
\end{aligned}
$$

- Estimating the coefficient functions $\alpha(.)$ and $\beta(.)$ is a high-dimensional problem
  $\hookrightarrow$ needs approximation for dimensionality reduction

$\hookrightarrow$ Here we represent the functional data by using a basis expansion :

$$X_i(t) = \sum_{j=1}^{r} x_{ij} b_j(t) = \boldsymbol{x}_i^\top \mathbf{b}_r(t), \tag{3}$$

- $\mathbf{b}_r(t) = (b_1(t), b_2(t), \ldots, b_r(t))^\top$ is an r-dimensional basis ((B-)spline, Wavelet,..)
- $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ir})^\top$ can be seen as the vector representation of $X_i(.)$

Here the X's are directly observed (We later consider the case when they are not).
$\hookrightarrow$ The x_{ij}s' can be computed explicitly by $x_{ij} = \int_{\mathcal{T}} X_i(t) b_j(t) dt$ for $j = 1, \ldots, r$

Representation of the functional gating network

Functional linear predictor for the gating network defined as :

$$h_z(x(\cdot)) = \log\left\{\frac{\mathbb{P}(Z=z|\{x(t), t\in\mathcal{T}\}}{\mathbb{P}(Z=K|\{x(t), t\in\mathcal{T}\}}\right\} = \alpha_{z,0} + \int_{\mathcal{T}} X(t)\alpha_z(t)dt$$

$\hookrightarrow$ The function $\alpha_z(t)$ is represented similarly as for X function by

$$\alpha_z(t) = \sum_{j=1}^{q} \zeta_{z,j} b_j(t) = \boldsymbol{\zeta}_z^\top \mathbf{b}_q(t) \tag{4}$$

where

- $\mathbf{b}_q(t) = (b_1(t), \ldots, b_q(t))^\top$ is a q-dimensional basis (of the same type as X).
- $\boldsymbol{\zeta}_z = (\xi_{z,1}, \xi_{z,2}, \ldots, \xi_{z,q})^\top$ is the vector of logistic regression coefficients

Representation of the functional gating network

Then the functional linear predictor $h_z(x_i(\cdot))$ for $i = 1, \ldots, n$ is represented as

$$\begin{aligned} h_{z_i}(x_i(\cdot); \boldsymbol{\alpha}) &= \alpha_{z_i,0} + \int_{\mathcal{T}} X_i(t)\alpha_{z_i}(t)dt = \alpha_{z_i,0} + \int_{\mathcal{T}} \boldsymbol{x}_i^\top \mathbf{b}_r(t)\mathbf{b}_q^\top(t)\boldsymbol{\zeta}_{z_i} dt \\ &= \alpha_{z_i,0} + \boldsymbol{x}_i^\top \left(\int_{\mathcal{T}} \mathbf{b}_r(t)\mathbf{b}_q^\top(t)dt \right) \boldsymbol{\zeta}_{z_i} \\ &= \alpha_{z_i,0} + \boldsymbol{\zeta}_{z_i}^\top \mathbf{r}_i, \end{aligned}$$

where $\mathbf{r}_i = \left(\int_{\mathcal{T}} \mathbf{b}_r(t)\mathbf{b}_q(t)^\top dt\right)^\top \boldsymbol{x}_i$

The FME gating network (2) is then now phrased as

$$\pi_k(\mathbf{r}_i; \boldsymbol{\xi}) = \frac{\exp\{\alpha_{k,0} + \boldsymbol{\zeta}_k^\top \mathbf{r}_i\}}{1 + \sum_{k'=1}^{K-1} \exp\{\alpha_{k',0} + \boldsymbol{\zeta}_{k'}^\top \mathbf{r}_i\}} \tag{5}$$

where $\boldsymbol{\xi} = ((\alpha_{1,0}, \boldsymbol{\zeta}_1^\top), \ldots, (\alpha_{K-1,0}, \boldsymbol{\zeta}_{K-1}^\top))^\top \in \mathbb{R}^{(K-1)\times(q+1)}$ is the unknown parameter vector of the gating network, to be estimated.

Representation of the functional experts

$$Y_i = \beta_{z_i,0} + \int_{\mathcal{T}} X_i(t)\beta_{z_i}(t)dt + \varepsilon_i, \quad i = 1, \ldots, n\cdot$$

- The coeffecient function $\beta_z(\cdot)$ is represented by the following expansion :

$$\beta_z(t) = \sum_{j=1}^{p} \eta_{z,j} b_j(t) + e(t) = \boldsymbol{\eta}_z^\top \mathbf{b}_p(t) + e(t) \tag{6}$$

- $\mathbf{b}_p(t) = (b_1(t), b_2(t), \ldots, b_p(t))^\top$ is a p-dimensional basis ((B-)spline, Wavelet,..)
- $\boldsymbol{\eta}_z = (\eta_{z,1}, \eta_{z,2}, \ldots, \eta_{z,p})^\top$ is the vector of regression coefficients
- $e(t) \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_e^2)$, $e(\cdot) \perp X_i$'s and represents the approximation error of $\beta_z(t)$ by linear projection $\mathbf{b}_p(t)^\top \boldsymbol{\eta}_z$.

Representation of the functional experts

The functional linear expert regressor z is then represented as :

$$Y_i = \beta_{z_i,0} + \int_{\mathcal{T}} X_i(t)\beta_{z_i}(t)dt + \varepsilon_i = \beta_{z_i,0} + \int_{\mathcal{T}} \boldsymbol{x}_i^\top \mathbf{b}_r(t)\left(\mathbf{b}_p^\top(t)\boldsymbol{\eta}_{z_i} + e_i(t)\right)dt + \varepsilon_i$$

$$= \beta_{z_i,0} + \boldsymbol{x}_i^\top \left(\int_{\mathcal{T}} \mathbf{b}_r(t)\mathbf{b}_p^\top(t)dt\right)\boldsymbol{\eta}_{z_i} + \int_{\mathcal{T}} X_i(t)e(t)dt + \varepsilon_i$$

$$= \beta_{z_i,0} + \boldsymbol{\eta}_{z_i}^\top \mathbf{x}_i + \underbrace{\varepsilon_i + \int_{\mathcal{T}} X_i(t)e(t)dt}_{\varepsilon_i^\star}$$ where :

- $\mathbf{x}_i = \left(\int_{\mathcal{T}} \mathbf{b}_r(t)\mathbf{b}_p(t)^\top dt\right)^\top \boldsymbol{x}_i$
- $\varepsilon_i^\star = \varepsilon_i + \int_{\mathcal{T}} X_i(t)e(t)dt \sim \mathcal{N}(0, {\sigma_{z_i}^\star}^2)$.

The FME expert (1) can thus be expressed as

$$Y_i \quad = \quad \beta_{z_i,0} + \boldsymbol{\eta}_{z_i}^\top \mathbf{x}_i + \varepsilon_i^\star, \quad i = 1, \ldots, n, \tag{7}$$

and we have $f(y_i|x_i(.), z_i = k; \boldsymbol{\theta}_k) = \phi(y_i; \beta_{k,0} + \boldsymbol{\eta}_k^\top \mathbf{x}_i, {\sigma_k^\star}^2)$ where $\boldsymbol{\theta}_k = (\beta_{k,0}, \boldsymbol{\eta}_k^\top, {\sigma_k^\star}^2)^\top \in \mathbb{R}^{p+2}$ is the unknown parameter vector of expert density k

FME model

The Functional ME model

Combining (5) and (7), the resulting FME density is defined by

$$f(y_i|x_i(\cdot);\boldsymbol{\Psi}) = \sum_{k=1}^{K} \pi_k(\mathbf{r}_i;\boldsymbol{\xi})\phi(y_i;\beta_{k,0} + \boldsymbol{\eta}_k^\top \mathbf{x}_i, \sigma_k^{\star 2}) \tag{8}$$

where $\pi_k(\mathbf{r}_i;\boldsymbol{\xi}) = \exp\{\alpha_{k,0} + \boldsymbol{\zeta}_k^\top \mathbf{r}_i\} \Big/ 1 + \sum_{k'=1}^{K-1} \exp\{\alpha_{k',0} + \boldsymbol{\zeta}_{k'}^\top \mathbf{r}_i\}$ and $\boldsymbol{\Psi} = (\boldsymbol{\xi}^T, \boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_K^T)^T$ the unknown parameter vector of the model

Model fitting

- MLE via the EM algorithm [Jacobs et al., 1991, Dempster et al., 1977, McLachlan and Krishnan, 2008]
- Regularized ME to encourage sparsity (eg. lasso penalty [Tibshirani, 1996])
- Regularized MLE (lasso-type regularization) on the derivatives of the $\alpha(\cdot)$ and $\beta(\cdot)$ function, by relying on the methodology in [James et al., 2009]

1) FME and MLE via the EM algorithm

Maximum-Likelihood Estimation

$$\widehat{\boldsymbol{\Psi}} \in \arg\max_{\boldsymbol{\Psi}} L(\boldsymbol{\Psi})$$

log-likelihood : $L(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k(\mathbf{r}_i; \boldsymbol{\xi}) \phi(y_i; \beta_{k,0} + \boldsymbol{\eta}_k^\top \mathbf{x}_i, \sigma_k^{\star 2})$

The EM algorithm [Dempster et al., 1977]

$$\boldsymbol{\Psi}^{new} \in \arg\max_{\boldsymbol{\Psi} \in \Omega} \mathbb{E}_{\boldsymbol{\Psi}^{old}}[L_c(\boldsymbol{\Psi}) | \{X_i(\cdot), Y_i\}_{i=1}^n]$$

where $L_c(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{ik} \log \left[\pi_k(\mathbf{r}_i; \boldsymbol{\xi}) \phi(y_i; \beta_{k,0} + \boldsymbol{\eta}_k^\top \mathbf{x}_i, \sigma_k^{\star 2})\right]$, $Z_{ik} = \mathbb{1}_{\{z_i = k\}}$

Clustering, Regression

- Clustering : $\widehat{z}_i = \arg\max_{1 \leq k \leq K} \mathbb{E}[Z_{ik} | x_i(.); \widehat{\boldsymbol{\Psi}}], \quad (i = 1, \ldots, n)$
- Expert's mean function :
  $\widehat{y}_i | \{x_i(\cdot), \widehat{z}_i = k\} = \widehat{\beta}_{k,0} + \widehat{\boldsymbol{\eta}}_k^\top \mathbf{x}_i, \quad (i = 1, \ldots, n;\ k = 1, \ldots, K)$
- FME mean function : $\widehat{y}_i = \sum_{k=1}^{K} \pi_k(\mathbf{r}_i; \widehat{\boldsymbol{\xi}}) \{\widehat{\beta}_{k,0} + \widehat{\boldsymbol{\eta}}_k^\top \mathbf{x}_i\}, \quad (i = 1, \ldots, n)$

ML parameter estimation via EM (FME-EM)

The E-Step

Compute the expectation of the complete-data log-likelihood, given the observed data $\{x_i(\cdot), y_i\}_{i=1}^n$, using the current parameter vector $\boldsymbol{\Psi}^{(s)}$:

$$\begin{aligned} Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(s)}) &= \mathbb{E}\left[L_c(\boldsymbol{\Psi})|\{x_i(\cdot), y_i\}_{i=1}^n;\boldsymbol{\Psi}^{(s)}\right] \\ &= \sum_{i=1}^n\sum_{k=1}^K \tau_{ik}^{(s)} \log\left[\pi_k(\mathbf{r}_i;\boldsymbol{\xi})\phi(y_i;\beta_{k,0}+\boldsymbol{\eta}_k^\top\mathbf{x}_i,\sigma_k^{\star 2})\right], \end{aligned} \tag{9}$$

where $\tau_{ik}^{(s)} = \phi(y_i;\beta_{0,k}^{(s)}+\mathbf{x}_i^\top\boldsymbol{\eta}_k^{(s)},{\sigma^2}_k^{(s)})/f(y_i|x_i(.);\boldsymbol{\Psi}^{(s)})$ is the conditional probability that the pair $\{x_i(t), t\in\mathcal{T}; y_i\}$ is generated by the kth expert.

The M-Step

- Update the value of the parameter vector $\boldsymbol{\Psi}$ by $\boldsymbol{\Psi}^{(s+1)} = \arg\max_{\boldsymbol{\Psi}} Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(s)})$
- Separate maximizations w.r.t the gating network and the experts network

$$\boldsymbol{\xi}^{(s+1)} = \arg\max_{\boldsymbol{\Psi}}\{Q(\boldsymbol{\xi};\boldsymbol{\Psi}^{(s)}) = \sum_{i=1}^n\sum_{k=1}^K \tau_{ik}^{(s)}\log\pi_k(\mathbf{r}_i;\boldsymbol{\xi})\} \tag{10}$$

$$\boldsymbol{\theta}_k^{(s+1)} = \arg\max_{\boldsymbol{\Psi}}\{Q(\boldsymbol{\theta}_k;\boldsymbol{\Psi}^{(s)}) = \sum_{i=1}^n \tau_{ik}^{(s)}\log\phi(y_i;\beta_{k,0}+\boldsymbol{\eta}_k^\top\mathbf{x}_i,\sigma_k^{\star 2})\} \tag{11}$$

2) Regularized MLE via an EM-lasso algorithm

$\hookrightarrow p \gg n$ to ensure a good approximation of $\beta_z(t)$ by $\boldsymbol{\eta}_z^\top \mathbf{b}_p(t)$ (tradeoff between smoothness of the functional predictor and complexity of the model.)

Regularized Maximum-Likelihood Estimation

$$\widehat{\boldsymbol{\Psi}} \in \arg\max_{\boldsymbol{\Psi}} L(\boldsymbol{\Psi}) - \text{Pen}_{\lambda,\chi}(\boldsymbol{\Psi})$$

$$\text{log-likelihood} : L(\boldsymbol{\Psi}) = \sum_{i=1}^n \log \sum_{k=1}^K \pi_k(X_i;\boldsymbol{\xi})\phi(y_i;\beta_{k,0} + \boldsymbol{\eta}_k^\top \mathbf{x}_i, \sigma_k^{\star 2})$$

The EM-lasso algorithm

$$\boldsymbol{\Psi}^{new} \in \arg\max_{\boldsymbol{\Psi}\in\Omega} \mathbb{E}[L_{\lambda,\chi}^c(\boldsymbol{\Psi}) | \{X_i(\cdot), Y_i\}_{i=1}^n, \boldsymbol{\Psi}^{old}]$$

completed data log-likelihood :

$$L_{\lambda,\chi}^c(\boldsymbol{\Psi}) = \sum_{i=1}^n \sum_{k=1}^K Z_{ik} \log\left[\pi_k(\mathbf{r}_i;\boldsymbol{\xi})\phi(y_i;\beta_{k,0} + \boldsymbol{\eta}_k^\top \mathbf{x}_i, \sigma_k^{\star 2})\right] - \text{Pen}_{\lambda,\chi}(\boldsymbol{\Psi})$$

Lasso regularization

$$\text{Pen}_{\lambda,\chi}(\boldsymbol{\Psi}) = \lambda \sum_{k=1}^K \|\boldsymbol{\eta}_k\|_1 + \chi \sum_{k=1}^{K-1} \|\boldsymbol{\xi}_k\|_1 \tag{12}$$

where λ and χ are positive real values representing tuning parameters.

Regularized MLE via EM-lasso (FME-EMlasso)

The EM-lasso algorithm for FME

- E-Step : unchanged
- M-Step : $\boldsymbol{\Psi}^{(s+1)} = \arg\max_{\boldsymbol{\Psi}} \{Q_{\lambda,\chi}(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(s)}) = Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(s)}) - \text{Pen}_{\lambda,\chi}(\boldsymbol{\Psi})\}$

Updating the expert' network parameters

$\boldsymbol{\theta}_k^{(s+1)} \in \arg\max_{\boldsymbol{\theta}_k} Q\lambda(\boldsymbol{\theta}_k;\boldsymbol{\Psi}^{(s)})$ with

$$Q_\lambda(\boldsymbol{\theta}_k;\boldsymbol{\Psi}^{(s)}) = \sum_{i=1}^{n} \tau_{ik}^{(s)} \log \phi(y_i; \beta_{k,0} + \boldsymbol{\eta}_k^\top \mathbf{x}_i, \sigma_k^{\star 2}) - \lambda \sum_{j=1}^{p} |\eta_{kj}|,$$

- ↪ A weighted LASSO problem for the $\boldsymbol{\eta}_k$'s
- ↪ Apply the LASSO machinery
- ↪ the update of $\sigma_k^{\star 2}$ is a weighted variant of the standard univariate Gaussian regression

Updating the gating network parameters

Updating the gating network parameters

$\boldsymbol{\xi}^{(s+1)} \in \arg\max_{\boldsymbol{\xi}} Q\chi(\boldsymbol{\xi}; \boldsymbol{\Psi}^{(s)})$ with

$$Q_\chi(\boldsymbol{\xi}; \boldsymbol{\Psi}^{(s)}) = \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}^{(s)} \log \pi_k(\mathbf{r}_i; \boldsymbol{\xi}) - \chi \sum_{k=1}^{K-1}\sum_{j=1}^{q} |\xi_{kj}|$$

$$= \sum_{i=1}^{n}\left(\sum_{k=1}^{K-1} \tau_{ik}^{(s)} \left(\alpha_{k,0} + \boldsymbol{\zeta}_k^\top \mathbf{r}_i\right) - \log\left(1 + \sum_{k'=1}^{K-1} \exp\{\alpha_{k',0} + \boldsymbol{\zeta}_{k'}^\top \mathbf{r}_i\}\right)\right) - \chi \sum_{k=1}^{K-1}\sum_{j=1}^{q} |\xi_{kj}|,$$

↪ A weighted version of the regularized multinomial logistic problem (e.g [Mousavi and Sørensen, 2017])

- There is no closed-form solution
- we then use a Newton-Raphson with Coordinate Ascent updates of the gating network coefficients ξ_{kj}.

Coordinate Ascent for the gating network

For each expert k, for $j = 1, \ldots, p$:

$$\begin{aligned} \zeta_{k,j}^{(t+1)} &= \frac{\mathcal{S}\Big(\sum_{i=1}^{n} w_{ik}\mathbf{r}_{ij}(\tilde{h}_i^{(t)} - \tilde{z}_i^{(t)}); \chi\Big)}{\sum_{i=1}^{n} w_{ik}\mathbf{r}_{ij}^2} \\ &= \mathcal{S}\left(\mathbf{R}_j^T \mathbf{W}_k^{(t)}(\tilde{\boldsymbol{h}}_i^{(t)} - \tilde{\boldsymbol{z}}_i^{(t)}); \chi\right) / (\mathbf{R}_j^T \mathbf{W}_k^{(t)} \mathbf{R}_j) \end{aligned} \tag{13}$$

where

- $h_i^{(s)} = \alpha_{k,0}^{(s)} + \mathbf{r}_i^\top \boldsymbol{\zeta}_k + (\tau_{ik}^{(s)} - \pi_k(\mathbf{r}_i; \boldsymbol{\xi}^{(s)}))/w_{ik}$ is the working response
- $\tilde{z}_i^{(s)} = \alpha_{k,0}^{(s)} + \mathbf{r}_i^\top \boldsymbol{\zeta}_k - \mathbf{r}_{ij}\zeta_{k,j}^{(t+1)}$;fitted value excluding the contribution from $\zeta_{k,j}$
- $w_{ik} = \pi_k(\mathbf{r}_i; \boldsymbol{\xi}^{(t)})(1 - \pi_k(\mathbf{r}_i; \boldsymbol{\xi}^{(t)}))$
- $\mathbf{W}_k^{(t)} = \text{diag}(\boldsymbol{w}_k)$ with $\boldsymbol{w}_k = (w_{1k}, \ldots, w_{nk})^\top$ and $\mathbf{R}_j$ is the jth column of $\mathbf{R} = (\mathbf{r}_1, \ldots, \mathbf{r}_n)^\top$,
- $\mathcal{S}(\cdot)$ is a soft-thresholding operator defined by $\mathcal{S}(u, \chi) = \text{sign}(u)(|u| - \chi)_+$ and $(x)_+$ a shorthand for $\max\{x, 0\}$

For $\alpha_{k,0}$, the update is given by

$$\alpha_{k,0}^{(t+1)} = \frac{\sum_{i=1}^{n} w_{ik}(\tilde{h}_i^{(t)} - \mathbf{r}_i^\top \boldsymbol{\zeta}_k^{(t)})}{\sum_{i=1}^{n} w_{ik}} = \boldsymbol{w}_k^{(t)\top}(\tilde{\boldsymbol{h}}_i^{(t)} - \mathbf{R}\boldsymbol{\zeta}_i^{(t)})/\text{trace}(\mathbf{W}_k^{(t)})$$

Coordinate Ascent for the expert network

For each expert k, for $j = 1, \ldots, p$:

$$\begin{aligned}
\eta_{kj}^{(q+1)} &= \mathcal{S}\left(\sum_{i=1}^{n} \tau_{ik}^{(s)}(y_i - \beta_{k0}^{(s)} - \mathbf{x}_i^\top \boldsymbol{\beta}_k^{(s)} + \eta_{kj}^{(q)}\mathbf{x}_{ij}); \lambda {\sigma_k^{(s)}}^2\right) / \sum_{i=1}^{n} \tau_{ik}^{(s)} \mathbf{x}_{ij}^2 \\
&= \mathcal{S}\left(\mathbf{X}_j^T \mathbf{W}_k^{(q)} \boldsymbol{r}_{kj}^{(q)}; \lambda {\sigma_k^{(s)}}^2\right) / (\mathbf{X}_j^T \mathbf{W}_k^{(q)} \mathbf{X}_j) \qquad (14)
\end{aligned}$$

where $\mathbf{X}_j$ is the jth column of the design matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$,
$\mathbf{W}_k^{(q)} = \text{diag}(\tau_{1k}^{(q)}, \ldots, \tau_{nk}^{(q)}) = \text{diag}(\boldsymbol{\tau}_k^{(s)})$,
$\boldsymbol{r}_{kj}^{(q)} = \boldsymbol{y} - \beta_{k0}^{(q)} \mathbf{1}_n - \mathbf{X}\boldsymbol{\beta}_k^{(q)} + \beta_{kj}^{(q)} \mathbf{X}_j$ is the residual without the contribution of the jth coefficient
$\mathcal{S}(u, \eta) := \text{sign}(u)(|u| - \eta)_+$ is the soft-thresholding operator with $(.)_+ = \max\{., 0\}$.

$$\beta_{k,0}^{(s+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(s)}(y_i - \mathbf{x}_i^\top \boldsymbol{\eta}_k^{(s)})}{\sum_{i=1}^{n} \tau_{ik}^{(s)}} = {\boldsymbol{\tau}_k^{(s)}}^\top (\boldsymbol{y} - \mathbf{X}\boldsymbol{\eta}_k^{(s)}) / \text{trace}(\mathbf{W}_k^{(s)}), \qquad (15)$$

$$\begin{aligned}
{\sigma_k^{2(s+1)}} &= \frac{\sum_{i=1}^{n} \tau_{ik}^{(s)} \left(y_i - \beta_{k,0}^{(s+1)} - \mathbf{x}_i^\top \boldsymbol{\eta}_k^{(s+1)}\right)^2}{\sum_{i=1}^{n} \tau_{ik}^{(s)}} \\
&= \left\|\sqrt{\mathbf{W}_k^{(s+1)}} \left(\boldsymbol{y} - \beta_{k0}^{(s+1)} \mathbf{1}_n - \mathbf{X}\boldsymbol{\eta}_k^{(s+1)}\right)\right\|_2^2 / \text{trace}(\mathbf{W}_k^{(q)}) \qquad (16)
\end{aligned}$$

3) FME by regularizing functional derivatives

- For FME-LASSO regularization described previously, there is no actually reason that the functions $\beta(.)$ and $\alpha(.)$ be sparse.
- So regularizing the parameter vectors representing these functions has no obvious interpretability

↪ The methodology [James et al., 2009] offers an interpretable and sparse fit for functional linear regression

- Regularization is performed on the the derivatives of the coefficient function, rather than on the paramters of the function

↪ We rely on intrepretable regularization for the regression functions $\beta_{z_i}(t)$ (and $\alpha_{z_i}(t)$)

iFME : determine whether the dth derivative of $\beta_{z_i}(t)$ is zero or not at each point t_j.

↪ can produce a meaningful sparse estimates for $\beta_{z_i}(t)$ curves :
$\beta_{z_i}^{(0)}(t) = 0$ implies that $X(t)$ has no effect on Y at t
$\beta_{z_i}^{(1)}(t) = 0$ means that $\beta_{z_i}(t)$ is constant at t,
$\beta_{z_i}^{(0)}(t) = 1$ shows that $\beta_{z_i}(t)$ is a linear function of t, etc.

- Let D^d be the dth finite difference operator defined recursively as

$$D^1\mathbf{b}(t_j) = p[\mathbf{b}(t_j) - \mathbf{b}(t_{j-1})],$$
$$D^2\mathbf{b}(t_j) = D[D\mathbf{b}(t_j)] = p^2[\mathbf{b}(t_j) - 2\mathbf{b}(t_{j-1}) + \mathbf{b}(t_{j-2})],$$
$$D^d\mathbf{b}(t_j) = D[D^{d-1}\mathbf{b}(t_j)].$$

- $D^d\mathbf{b}(t_j)$ is an approximation for $\mathbf{b}^{(d)}(t_j) = [b_1^{(d)}(t_j), \ldots, b_p^{(d)}(t_j)]^\top$
- $\mathbf{A}_p = [D^d\mathbf{b}(t_1), D^d\mathbf{b}(t_2), \ldots, D^d\mathbf{b}(t_p)]^\top$ (the approximate derivative matrix)
- Let $\boldsymbol{\gamma}_{z_i} = \mathbf{A}_p\boldsymbol{\eta}_{z_i}$

$\hookrightarrow$ If $\beta_{z_i}^{(d)}(t) = 0$ over a large regions of t for some d, then $\boldsymbol{\gamma}_{z_i}$ is sparse.

$\hookrightarrow$ $\boldsymbol{\gamma}_{z_i} = [\gamma_{z_i,1}, \ldots, \gamma_{z_i,p}]^\top$ provides a sparse estimate for $[\beta_{z_i}^{(d)}(t_1), \ldots, \beta_{z_i}^{(d)}(t_p)]^\top$.

Functional expert' network of iFME

$$\begin{aligned} Y_i &= \beta_{z_i,0} + \boldsymbol{\eta}_{z_i}{}^\top\mathbf{x}_i + \varepsilon_i^\star = \beta_{z_i,0} + (\mathbf{A}_p^{-1}\boldsymbol{\gamma}_{z_i})^\top\mathbf{x}_i + \varepsilon_i^\star \\ &= \beta_{z_i,0} + (\mathbf{x}_i^\top\mathbf{A}_p^{-1})\boldsymbol{\gamma}_{z_i} + \varepsilon_i^\star \\ &= \beta_{z_i,0} + \mathbf{v}_i^\top\boldsymbol{\gamma}_{z_i} + \varepsilon_i^\star. \end{aligned}$$

and we now have $\boldsymbol{\theta}_k = (\beta_{k,0}, \boldsymbol{\gamma}_k^\top, \sigma_k^{\star 2})^\top$ parameter vector of expert density k

Gating network of interpretable FME

- Similarly, let $\boldsymbol{\omega}_k = \mathbf{A}_q \boldsymbol{\zeta}_k$ where $\mathbf{A}_q = [D^d\mathbf{b}(t_1), D^d\mathbf{b}(t_2), \ldots, D^d\mathbf{b}(t_q)]^\top$
  $\hookrightarrow$ we get $\boldsymbol{\zeta}_k = \mathbf{A}_q^{-1}\boldsymbol{\omega}_k$.

The gating network probabilities become

$$\pi_k(\boldsymbol{\nu}_i; \mathbf{w}) = \frac{\exp\{\alpha_{k,0} + {\boldsymbol{\zeta}_k}^\top \mathbf{r}_i\}}{1 + \sum_{k'=1}^{K-1} \exp\{\alpha_{k',0} + {\boldsymbol{\zeta}_{k'}}^\top \mathbf{r}_i\}} = \frac{\exp\{\alpha_{k,0} + \boldsymbol{\nu}_i^\top \boldsymbol{\omega}_k\}}{1 + \sum_{k'=1}^{K-1} \exp\{\alpha_{k',0} + \boldsymbol{\nu}_i^\top \boldsymbol{\omega}_{k'}\}} \tag{17}$$

with $\boldsymbol{\nu}_i = \mathbf{r}_i^\top \mathbf{A}_q^{-1}$ is the new predictor and the new gating network parameter vector $\mathbf{w} = ((\alpha_{1,0}, \boldsymbol{\omega}_1^\top), \ldots, (\alpha_{K-1,0}, \boldsymbol{\omega}_{K-1}^\top))^\top$ and $(\alpha_{K-1,0}, \boldsymbol{\omega}_K^\top)^\top$ is a null vector.

The resulting FME distribution and parameter estimation

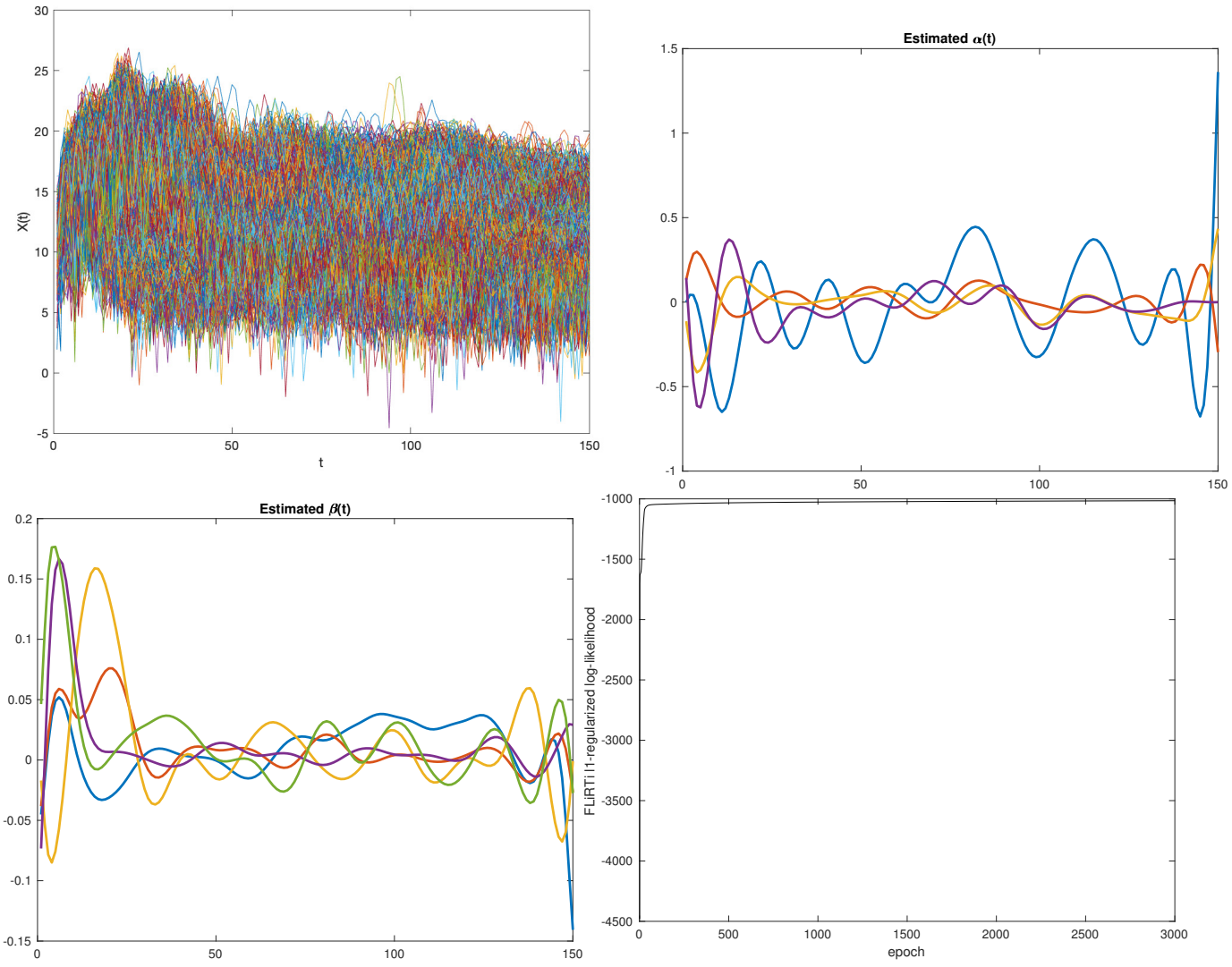
$$f(y_i|u_i(.); \boldsymbol{\Psi}) = \sum_{k=1}^{K} \pi_k(\boldsymbol{\nu}_i; \mathbf{w}) \phi(y_i; \beta_{k,0} + \boldsymbol{\gamma}_k^\top \mathbf{v}_i, \sigma_k^{\star 2}) \tag{18}$$

where $\boldsymbol{\Psi} = (\mathbf{w}^T, \boldsymbol{\Psi}_1^T, \ldots, \boldsymbol{\Psi}_K^T)^T$ the unknown parameter vector of the model

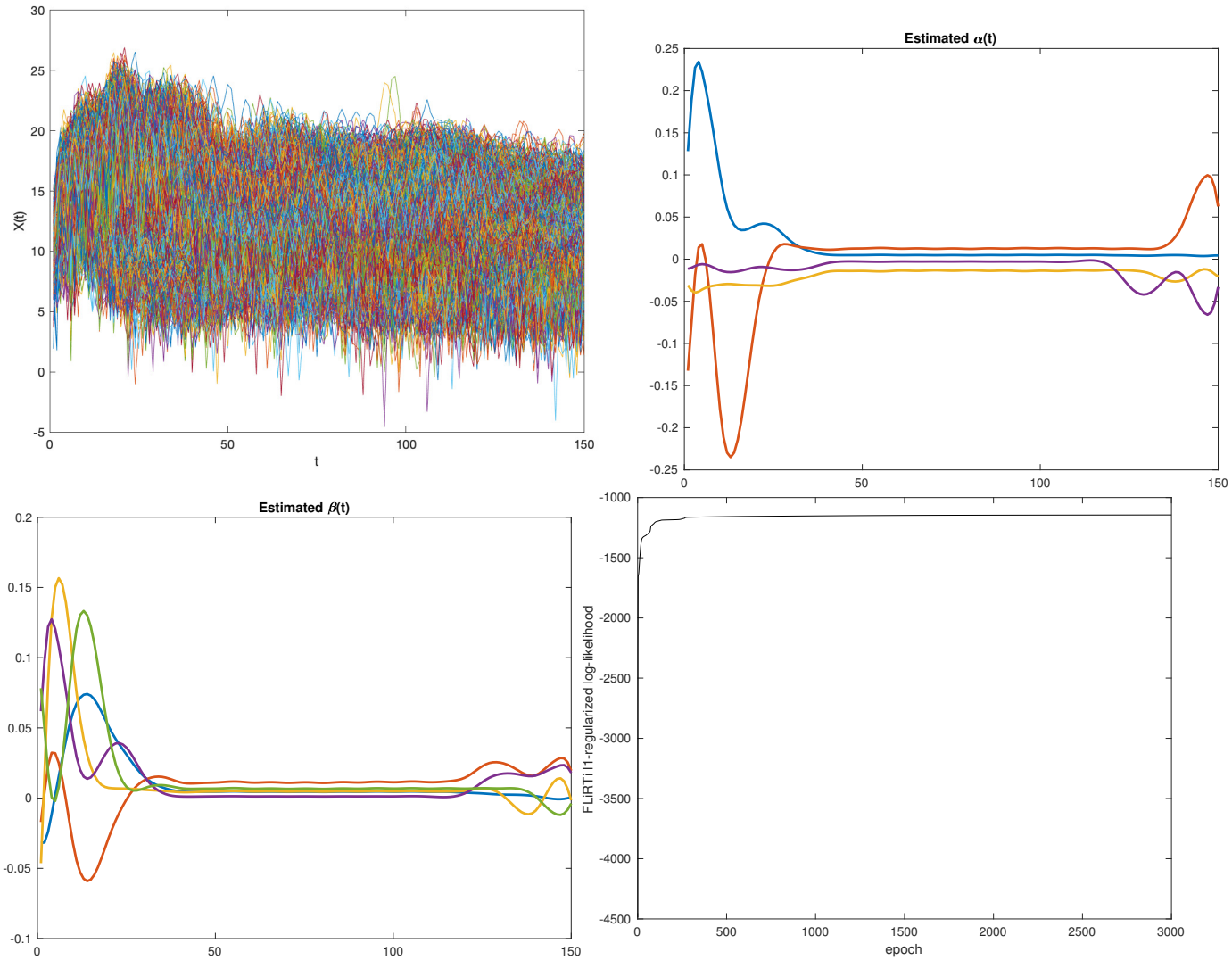
$\hookrightarrow$ Apply the EM-Lasso algorithm developed previously with :

- Predictors : $\mathbf{v}_i = \mathbf{x}_i^\top \mathbf{A}_p^{-1}$ and $\boldsymbol{\nu}_i = \mathbf{r}_i^\top \mathbf{A}_q^{-1}$
- Regularization : on $\boldsymbol{\omega}$'s and $\boldsymbol{\gamma}$'s : $\text{Pen}_{\lambda,\chi}(\boldsymbol{\Psi}) = \lambda \sum_{k=1}^{K} \|\boldsymbol{\gamma}_k\|_1 + \chi \sum_{k=1}^{K-1} \|\boldsymbol{\omega}_k\|_1$

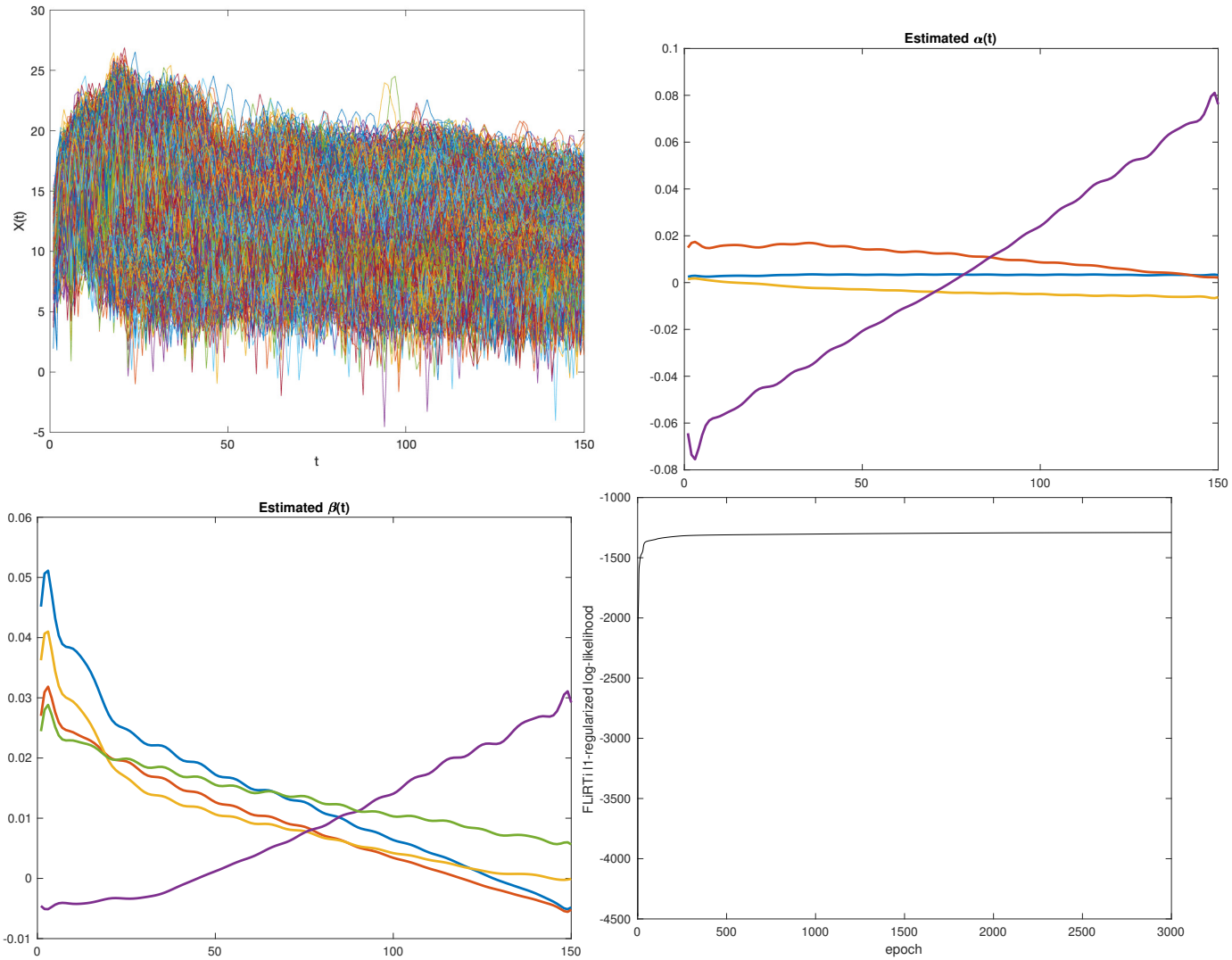
Example : Phonemes data (K=5), d=0



Example : Phonemes data (K=5), d=1



Example : Phonemes data (K=5), d=2



FME for noisy predictors

The functional predictors $X_i(t)$ are in general unobserved directly

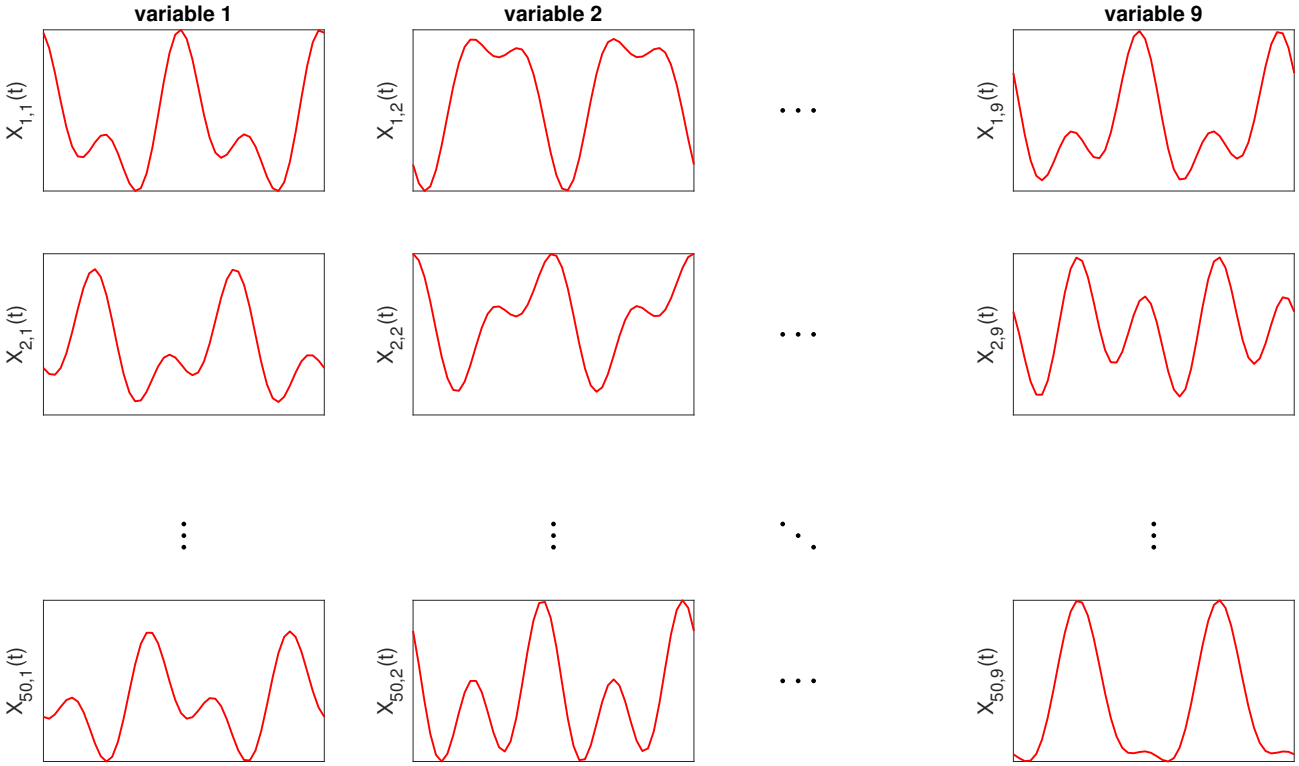


FIGURE – functional predictors $X_{ij}(t)$ $t \in \mathcal{T}$

FME for noisy predictors

We rather observe $U_i(t)$ a noisy version of $X_i(t)$

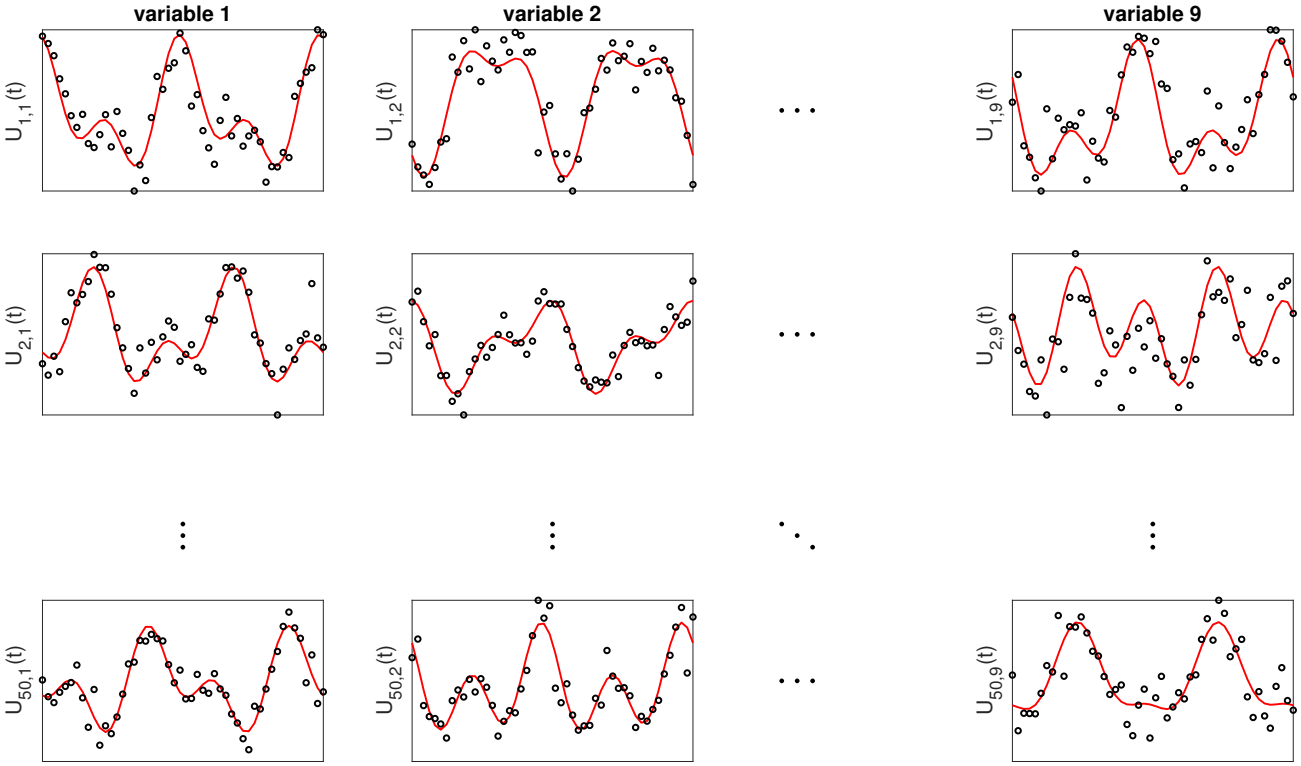


FIGURE – Noisy functional predictors $U_{ij}(t)$ $t \in \mathcal{T}$

Until now the functional predictors $X_i(t)$ are represented by basis expansion as

$$X_i(t) = \sum_{j=1}^{r} x_{ij} b_j(t) = \boldsymbol{x}_i^\top \mathbf{b}_r(t),$$

$\hookrightarrow$ the coefficients $x_{ij} = \int_{\mathcal{T}} X_i(t) b_j(t) dt$ are unknown since $X_i(t)$ is not observed
$\hookrightarrow$ We first model $U_i(t)$ (for a single variable) as

$U_i(t) = X_i(t) + \delta_i(t), \quad i = 1, \ldots, n, \quad \delta_i \sim \mathcal{N}(0, \sigma_\delta^2)$
We assume that the δ_i's are independent of the $X_i(\cdot)$'s and the Y_i's.

and propose an unbiased estimator of x_{ij} from $U_i(t)$ defined as

$$\widehat{x}_{ij} := \int_{\mathcal{T}} U_i(t) b_j(t) dt.$$

$\mathbb{E}[\widehat{x}_{ij}] = \int_{\mathcal{T}} \mathbb{E}(U_i(t)) b_j(t) dt = \int_{\mathcal{T}} X_i(t) b_j(t) dt = x_{ij}$.
$\hookrightarrow$ Thus, an estimate $\widehat{X}_i(t)$ of $X_i(t)$ can be given as

$$\widehat{X}_i(t) = \widehat{\boldsymbol{x}}_i^\top \mathbf{b}_r(t), \quad i = 1, \ldots, n, \tag{19}$$

with $\widehat{\boldsymbol{x}}_i = (\widehat{x}_{i1}, \ldots, \widehat{x}_{ir})^\top$.
$\hookrightarrow$ The previous models/algorithms apply by replacing $\boldsymbol{x}_i$ by its estimate $\widehat{\boldsymbol{x}}_i$

Simulation study

Simulations according to a functional mixture of Gaussian experts with $K = 3$ clusters with the time-domain $\mathcal{T} = [0, 1]$. Data generation process sampling density

$$\begin{aligned} Y_i|Z_i = z_i, U_i(.) &\sim \mathcal{N}\Big(\beta_{z_i,0} + \int_{\mathcal{T}} U_i(t)\beta_{z_i}(t)dt;\ \sigma^2_{z_i}\Big), \\ Z_i|U_i(.) &\sim \mathcal{M}\Big(1, \big(\pi_1(U_i(.);\boldsymbol{\alpha}), \ldots, \pi_K(U_i(.);\boldsymbol{\alpha})\big)\Big) \\ U_i(t) &= X_i(t) + \delta_i(t), \text{ with } \delta_i(t) \sim \mathcal{N}(0, \sigma^2_\delta) \\ X_i(t) &= \boldsymbol{x}_i^\top \mathbf{b}(t); \mathbf{b}(t) \text{ is a 10-dimensional B-spline basis }, \boldsymbol{x}_i = \boldsymbol{W}\boldsymbol{v}_i, \end{aligned} \tag{20}$$

where $\boldsymbol{W} \in \mathbb{R}^{10\times 10}$ is a matrix of i.i.d random values $\sim \mathcal{U}(0, 1)$ and $\boldsymbol{v}_i \in \mathbb{R}^{10}$ is a vector of i.i.d random values $\sim \mathcal{N}(0, 10)$, with the following expert and gating functions parameters :

$$\begin{aligned} \beta_1(t) &= \begin{cases} -50(t-0.5)^2 + 4 & \text{if } 0 \le t < 0.3, \\ 0 & \text{if } 0.3 \le t < 0.7, \\ 50(t-0.5)^2 - 4 & \text{if } 0.7 \le t \le 1, \end{cases} \\ \beta_2(t) &= -\beta_1(t), \\ \beta_3(t) &= 100(t-0.5)^2 - 10, \quad 0 \le t \le 1, \\ (\beta_{1,0}, \beta_{2,0}, \beta_{3,0})^\top &= (-5, 0, 5)^\top, \\ (\sigma_1^2, \sigma_2^2, \sigma_3^2)^\top &= (5, 5, 5)^\top, \\ \alpha_1(t) &= 80(t-0.5)^2 - 8, \\ \alpha_2(t) &= -\alpha_1(t), \quad \alpha_3(t) = \mathbf{0}, \quad 0 \le t \le 1, \\ (\alpha_{1,0}, \alpha_{2,0}, \alpha_{3,0})^\top &= (-10, -10, 0)^\top. \end{aligned}$$

The expert parameter functions $\beta_1(t)$ and $\beta_2(t)$ have a flat region in the interval $0.3 \le t < 0.7$, out of that they are quadratic, while $\beta_3(t)$ and the gating parameter functions $\alpha_1(t)$, $\alpha_2(t)$ are all quadratic on the whole domain.

Scenario	S1	S2	S3	S4
σ_δ^2	1	1	4	4
m	100	50	100	50

TABLE – Specifications of scenarios $S1, \ldots, S4$ with measurement noise σ_δ^2 and curve length m.

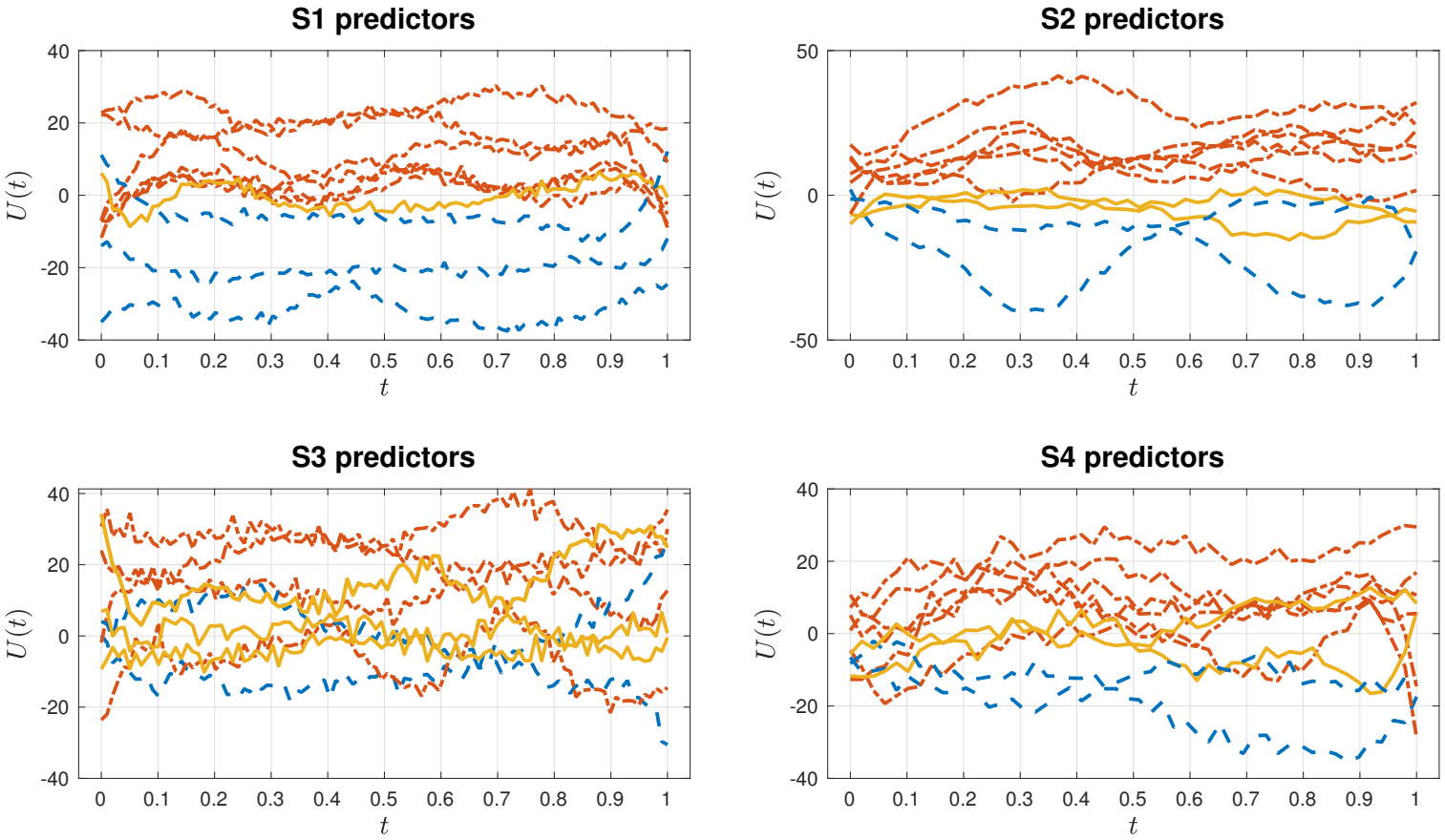


FIGURE – 10 randomly taken predictors in scenarios $S1, S2, S3$ and $S4$

Simulation results

	RPE	Corr	RI	ARI
S1 $(m = 100, \sigma_\delta^2 = 1)$				
FME	$.1135_{(.0319)}$	$.9374_{(.0191)}$	$.9407_{(.0169)}$	$.8675_{(.0379)}$
FME-Lasso	$.1067_{(.0284)}$	$.9411_{(.0170)}$	$.9421_{(.0165)}$	$.8706_{(.0371)}$
iFME	$\mathbf{.1058_{(.0290)}}$	$\mathbf{.9415_{(.0176)}}$	$\mathbf{.9441_{(.0173)}}$	$\mathbf{.8752_{(.0389)}}$
S2 $(m = 50, \sigma_\delta^2 = 1)$				
FME	$.1167_{(.0270)}$	$.9341_{(.0156)}$	$.9384_{(.0135)}$	$.8624_{(.0307)}$
FME-Lasso	$\mathbf{.1079_{(.0238)}}$	$\mathbf{.9391_{(.0141)}}$	$.9398_{(.0138)}$	$.8656_{(.0314)}$
iFME	$.1084_{(.0249)}$	$.9387_{(.0144)}$	$\mathbf{.9422_{(.0152)}}$	$\mathbf{.8709_{(.0343)}}$
S3 $(m = 100, \sigma_\delta^2 = 4)$				
FME	$.1245_{(.0228)}$	$.9312_{(.0131)}$	$.9361_{(.0157)}$	$.8576_{(.0356)}$
FME-Lasso	$.1181_{(.0199)}$	$.9346_{(.0117)}$	$.9378_{(.0157)}$	$.8613_{(.0356)}$
iFME	$\mathbf{.1158_{(.0204)}}$	$\mathbf{.9360_{(.0115)}}$	$\mathbf{.9392_{(.0144)}}$	$\mathbf{.8644_{(.0327)}}$
S4 $(m = 50, \sigma_\delta^2 = 4)$				
FME	$.1236_{(.0227)}$	$.9296_{(.0165)}$	$.9326_{(.0160)}$	$.8495_{(.0364)}$
FME-Lasso	$.1206_{(.0211)}$	$.9313_{(.0149)}$	$.9325_{(.0151)}$	$.8493_{(.0344)}$
iFME	$\mathbf{.1194}_{(.0217)}$	$\mathbf{.9320}_{(.0155)}$	$\mathbf{.9339}_{(.0148)}$	$\mathbf{.8525}_{(.0337)}$

TABLE – Evaluation of FME, FME-Lasso and iFME models on test data in scenarios $S1, \ldots, S4$. The reported values are averaged over 100 trials with standard errors in parentheses.

Simulation results

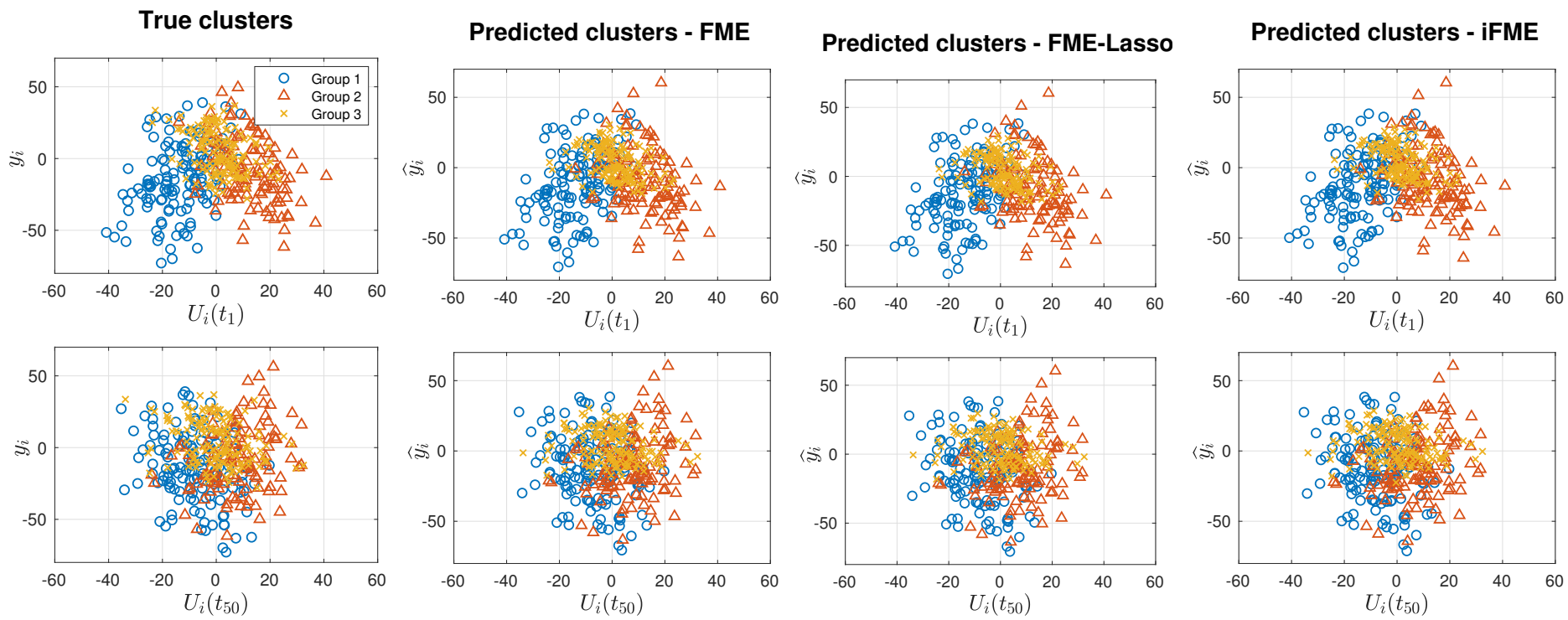


FIGURE – Scatter plots of $\hat{y}_i$ vs $U_i(t_1)$ (top panels) and $U_i(t_{50})$ (bottom panels) for a random dataset. Here, the clustering errors are 5.5%, 4.75% and 5.0% for FME, FME-Lasso and iFME models, respectively.

Simulation results

Model selection with BIC or modified BIC [Städler et al., 2010] defined as

$$\text{mBIC} = L(\widehat{\boldsymbol{\Psi}}) - \text{df}(\widehat{\boldsymbol{\Psi}})\frac{\log n}{2},$$

with $\text{df}(\widehat{\boldsymbol{\Psi}})$ is the effective number of (non-zero) parameters

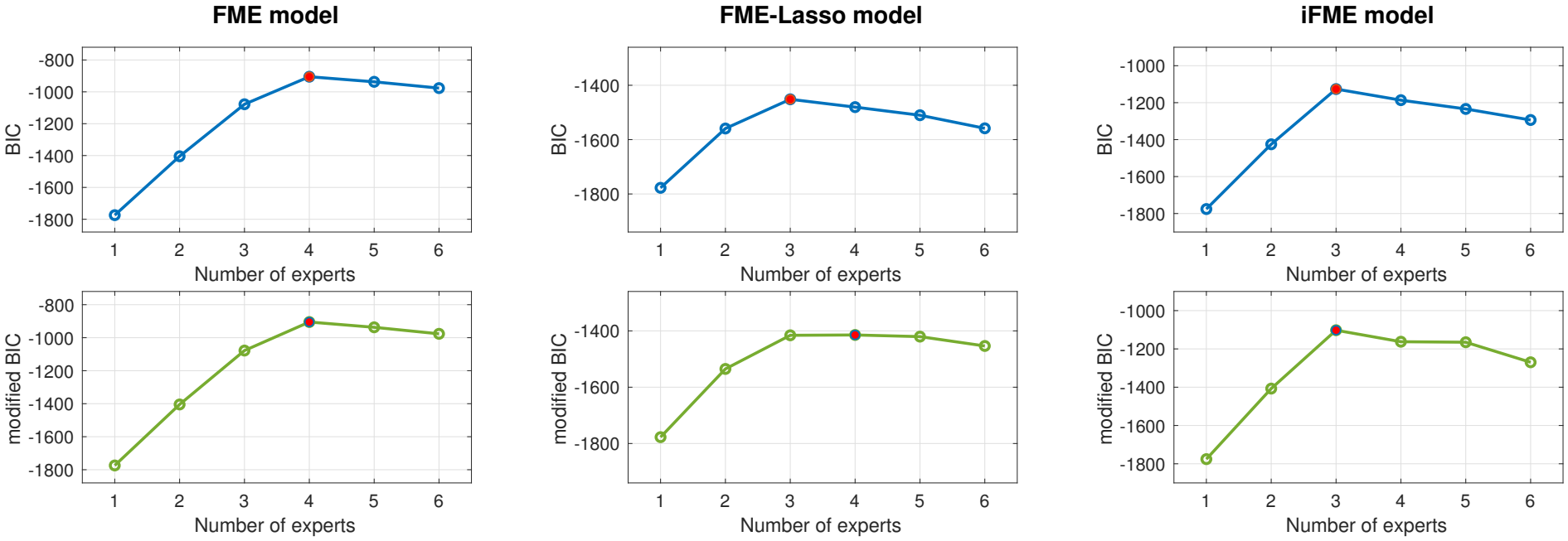
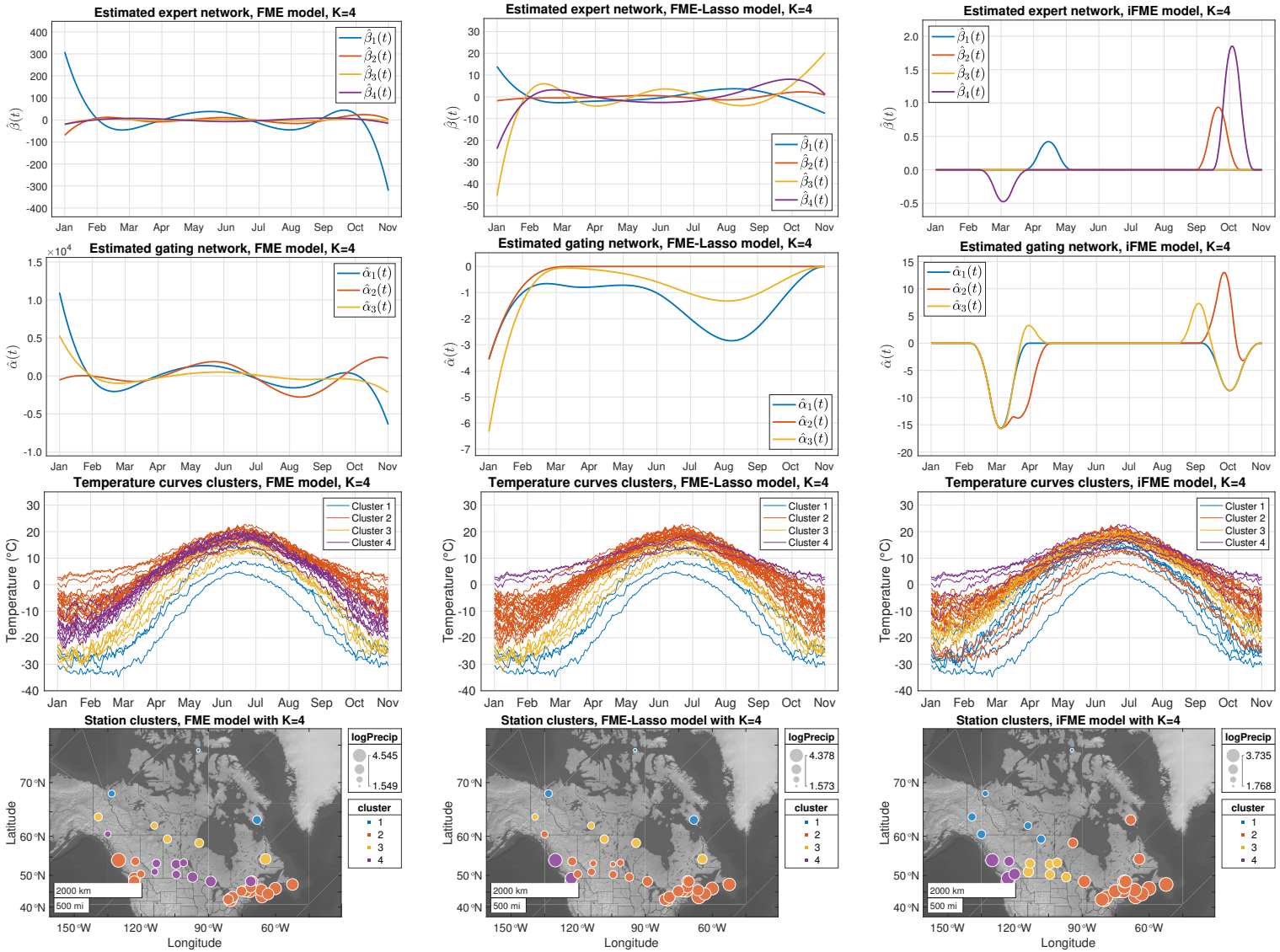


FIGURE – The BIC (top) and modified BIC (bottom) values of (a) FME, (b) FME-Lasso and (c) iFME model against the number of experts, fitted on a random taken dataset in scenario S1. Here, the red points are the ones with maximum BIC/modified BIC values.

Canadian weather data

FME models applied to Canadian weather data (K =4)



Canadian weather data

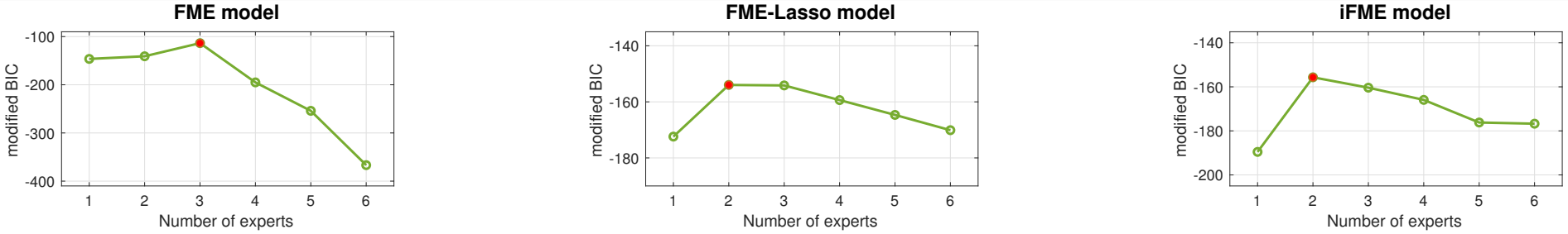
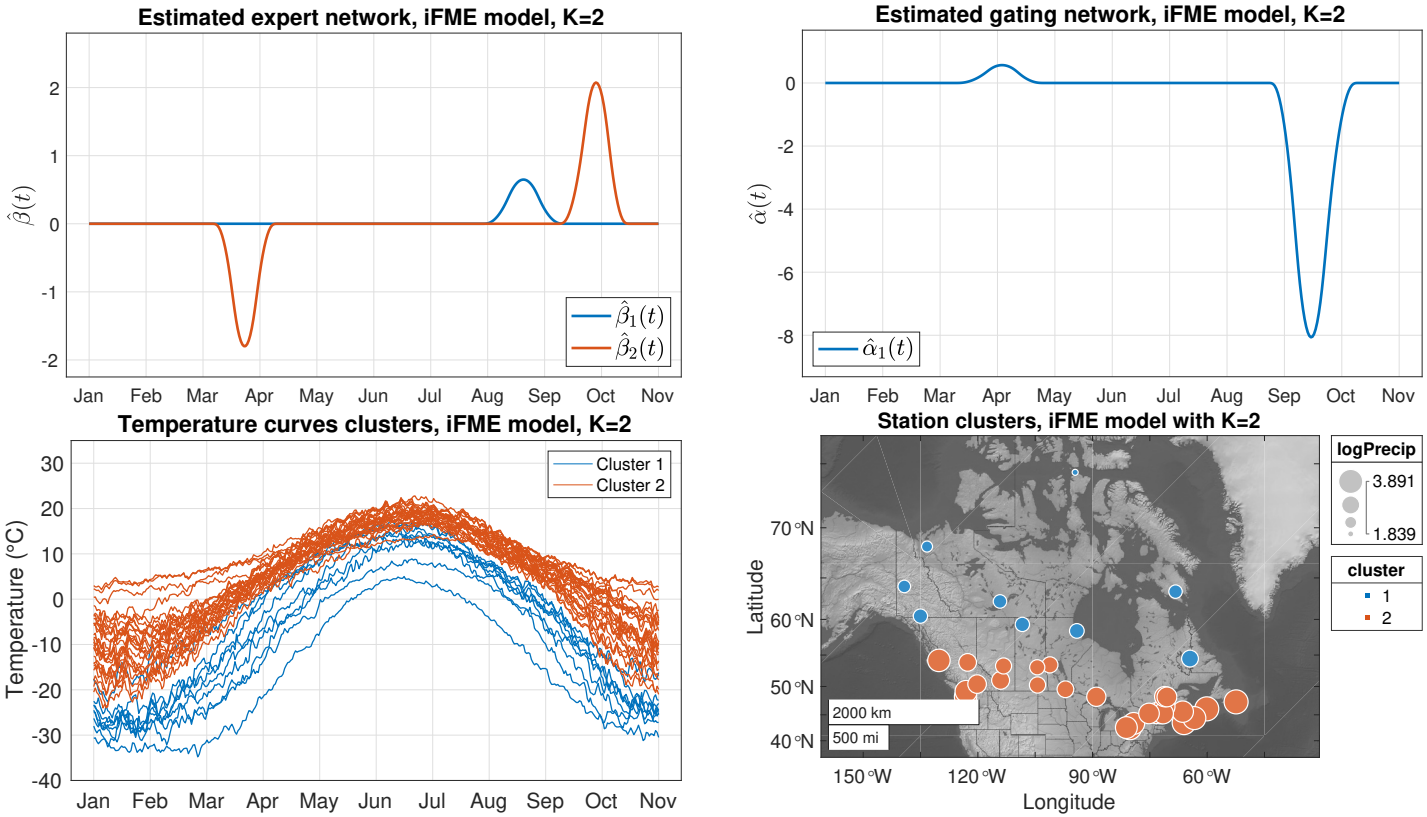


FIGURE – Selection of the number of functional mixture components K via modified BIC for Canadian weather data. Here, iFME is implemented with $d_1 = 0$, $d_2 = 3$ and $\rho = \varrho = 100$.



Concluding remarks

- A model for regression conditional density estimation with functional predictors
- The model inference can be performed by lasso-regularized EM algorithm
- Allows to perform feature selection and to keep the it interpretable

Ongoing :

- Soon on ArXiv
- Package (currently codes ares written in Matlab and will be made public soon)
- Extension to the multivariate setting

References

Faicel Chamroukhi and Bao T. Huynh. Regularized Maximum Likelihood Estimation and Feature Selection in Mixtures-of-Experts Models. *Journal de la Société Française de Statistique*, 160(1) :57–85, March 2019. URL `https://chamroukhi.com/papers/Chamroukhi_Huynh_jsfds-published.pdf`.

Faicel Chamroukhi and Hien D. Nguyen. Model-based clustering and classification of functional data. *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery*, Dec 2019. URL `https://chamroukhi.com/papers/MBCC-FDA.pdf`. DOI : 10.1002/widm.1298.

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, B*, 39(1) :1–38, 1977.

R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1) : 79–87, 1991.

Gareth M. James. Generalized linear models with functional predictors. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64(3) :411–432, 2002. ISSN 13697412, 14679868. URL `http://www.jstor.org/stable/3088780`.

Gareth M. James, Jing Wang, and Ji Zhu. Functional linear regression that's interpretable. *Annals of Statistics*, 37(5A) : 2083–2108, 2009.

M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6 :181–214, 1994.

A. Khalili. New estimation and feature selection methods in mixture-of-experts models. *Canadian Journal of Statistics*, 38(4) : 519–539, 2010.

G. J. McLachlan and T. Krishnan. *The EM algorithm and extensions*. New York : Wiley, second edition, 2008.

Seyed Nourollah Mousavi and Helle Sørensen. Multinomial functional regression with wavelets and lasso penalization. *Econometrics and Statistics*, 150–166, 2017. ISSN 2452-3062. doi: 10.1016/j.ecosta.2016.09.005.

Hans Müller and Ulrich Stadtmüller. Generalized functional linear models. *Ann. Statist*, pages 774–805, 2005.

Hien D. Nguyen and Faicel Chamroukhi. Practical and theoretical aspects of mixture-of-experts modeling : An overview. *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery*, pages e1246–n/a, Feb 2018. ISSN 1942-4795. doi: 10.1002/widm.1246. URL `http://dx.doi.org/10.1002/widm.1246`.

J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer, June 2005.

Nicolas Städler, Peter Bühlmann, and Sara Van De Geer. $\ell 1$-penalization for mixture regression models. *Test*, 19(2) :209–256, 2010.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1) : 267–288, 1996.

Thank you for your attention !