Learning mixtures-of-experts from heterogenous and high-dimensional data

Faïcel Chamroukhi

CMStatistics, December 17, 2023

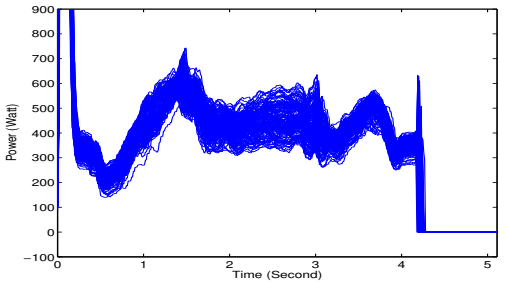
16th International Conference of the ERCIM WG on
Computational and Methodological Statistics (CMStatistics 2023)
16-18 December 2023, HTW Berlin, University of Applied Sciences, Berlin, Germany

Chamroukhi.com ChamroukhiF FChamroukhi ChamroukhiF GoogleScholar

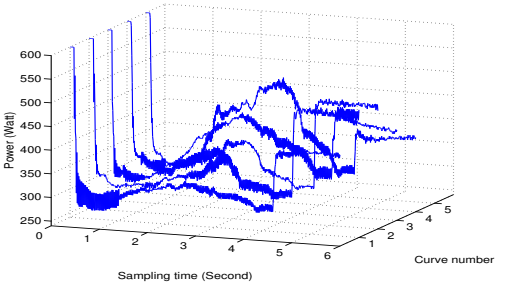
Real-world data are complex

- Heterogenous, Multimodal, High-Dimensional, Unlabeled, Possibly Massive ...
- Need for adapted analysis tools

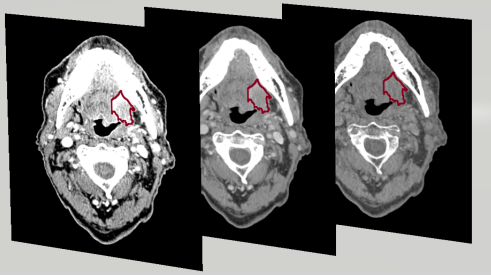
Transport : Railway switch curves diagnostic



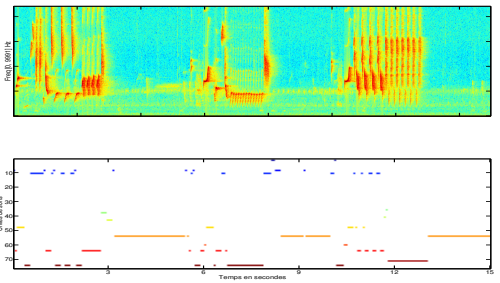
Predictive Maintenance

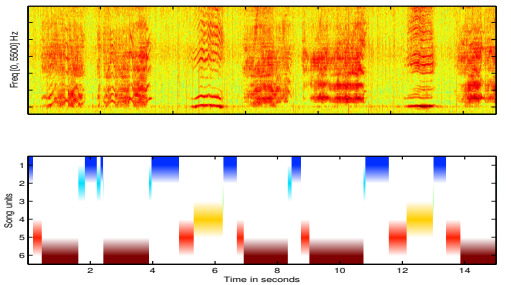


Health : Medical images

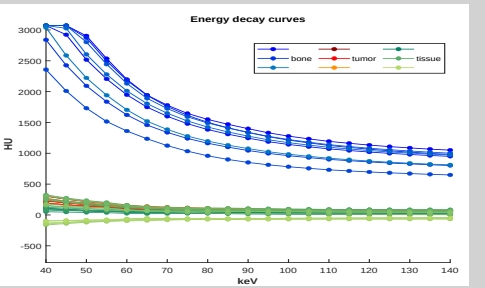


Acoustics : scene listening (marine, terrestrial)

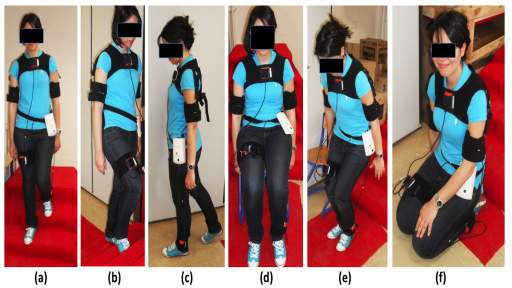




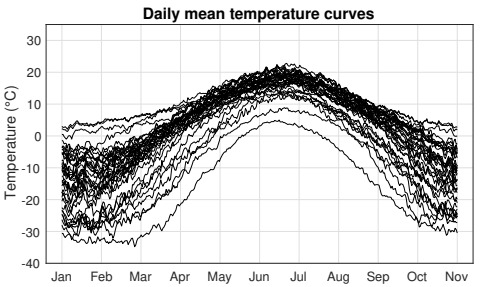
Dual-energy computed tomography

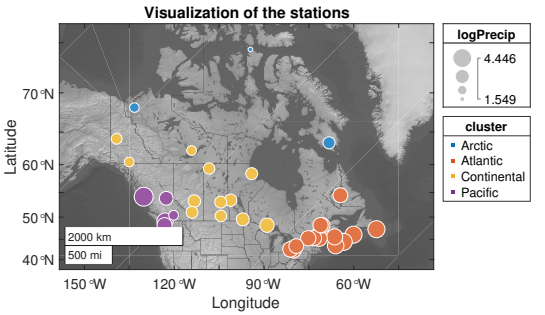


Health & Well Being : Activity recog.



Climate/Environment : meteorological data





Scientific Challenges

- Establish well-principled (with statistical guarantees) predictions in heterogeneous and high-dimensional situations,
- Construct efficient algorithms that operate in unsupervised way and provide interpretable solutions with computational guarantees.

Scientific framework

↪ Latent variable **models** : $f(x|\boldsymbol{\theta}) = \int_{\mathcal{Z}} f(x, z|\boldsymbol{\theta})\mathrm{d}z$

↪ **Inference**, **unsupervised representation** and **Selection** in high-dimension

Approximation capabilities of finite mixture distributions

Density approximation in Unsupervised Learning

- **Data** : observations $\{\boldsymbol{x}_i\}$ from $\boldsymbol{X} \in \mathbb{X} \subset \mathbb{R}^d$ of density (multimodal) $f \in \mathcal{F}$
- **Objective** : approximate the density f (and represent the data, e.g. *clustering*)
- **Solution :** Approximate f within the class $\mathcal{H}^{\varphi} = \bigcup_{K \in \mathbb{N}^{\star}} \mathcal{H}_K^{\varphi}$ of finite location-scale mixture h_K^{φ} (of K-components) of density φ (e.g., Gaussian), where

$$\mathcal{H}_K^{\varphi} = \left\{ \boxed{h_K^{\varphi}(\boldsymbol{x}) := \sum_{k=1}^{K} \pi_k \frac{1}{\sigma_k^d} \varphi\left(\frac{\boldsymbol{x} - \boldsymbol{\mu}_k}{\sigma_k}\right)}, \boldsymbol{\mu}_k \in \mathbb{R}^d, \sigma_k \in \mathbb{R}_+, \pi_k > 0 \, \forall k \in [K], \sum_{k=1}^{K} \pi_k = 1 \right\}$$

Theorem : Universal approximation of finite location-scale mixtures

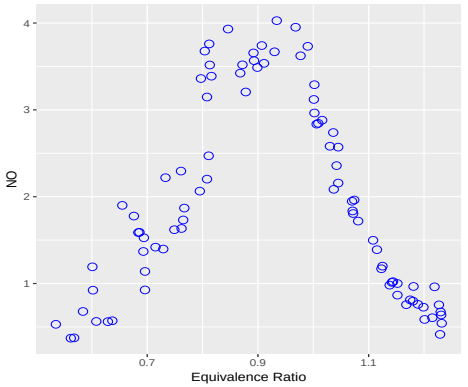
(a) Given any p.d.f $f, \varphi \in \mathcal{C}$ and a compact set $\mathbb{X} \subset \mathbb{R}^d$, there exists a sequence $(h_K^{\varphi}) \subset \mathcal{H}^{\varphi}$, such that $\lim_{K \to \infty} \sup_{\boldsymbol{x} \in \mathcal{X}} |f(\boldsymbol{x}) - h_K^{\varphi}(\boldsymbol{x})| = 0$.

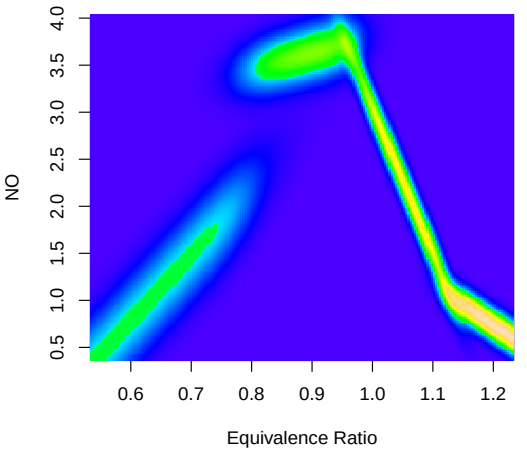
(b) For $p \in [1, \infty)$, if $f \in \mathcal{L}_p$ (Lebesgue p.d.f) and $\varphi \in \mathcal{L}_{\infty}$ (essentially bounded p.d.f), there exists a sequence $(h_K^{\varphi}) \subset \mathcal{H}^{\varphi}$, such that $\lim_{K \to \infty} \|f - h_K^{\varphi}\|_{\mathcal{L}_p} = 0$.

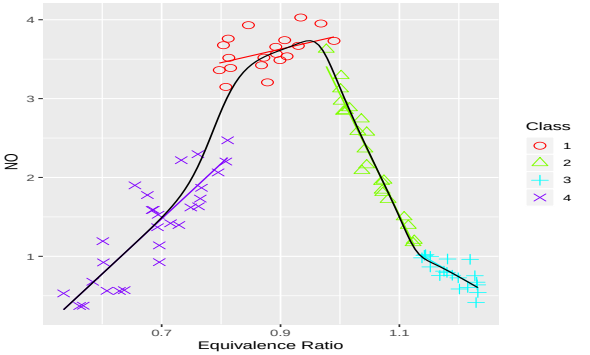
[J. Communications in Statistics - Theory and Methods, 2022] [PhD, TT Nguyen 2021]

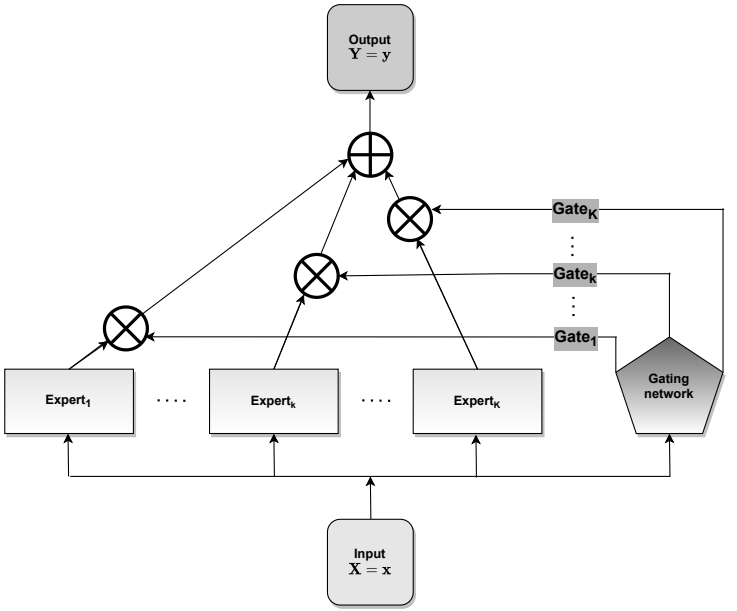
Heterogeneous regression-type data

Mixtures-of-Experts as good candidates to model a response Y given covariarte.s X when governed by a hidden structure accounting for heterogeneity









Apprentissage par Modèles de Mélanges d'Experts (ME)

- **Context :** n observations $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}$ from a pair $(\boldsymbol{X}, \boldsymbol{Y}) \in \mathbb{X} \times \mathbb{Y}$ with unknown conditional p.d.f $f \in \mathcal{F} = \left\{ f : \mathbb{X} \times \mathbb{Y} \to \mathbb{R}_+ \,|\, \int_{\mathbb{Y}} f(\boldsymbol{y}|\boldsymbol{x}) \, \mathrm{d}\lambda(\boldsymbol{y}) = 1, \forall \boldsymbol{x} \in \mathbb{X} \right\}$
- **High-dimensional setting** : $\mathbb{X} \subseteq \mathbb{R}^d$, $\mathbb{Y} \subseteq \mathbb{R}^q$, with $d, q \gg n$ and **heterogeneous**.
- **Objectifs :** Regression ; Clustering ; Model selection
- **Solution :** Approximate f within the class of **mixtures-of-experts** :

Let φ be a p.d.f (compactly supported on $\mathbb{Y} \subseteq \mathbb{R}^q$), we define the functional classes :

- Location-scale family : $\mathcal{E}_\varphi = \left\{ \phi_q(\boldsymbol{y}; \boldsymbol{\mu}, \sigma) := \frac{1}{\sigma^q} \varphi\left(\frac{\boldsymbol{y} - \boldsymbol{\mu}}{\sigma} \right) ; \boldsymbol{\mu} \in \mathbb{Y}, \sigma \in \mathbb{R}_+ \right\}$.
- Mixture of location-scale experts with softmax activation network : SGaME :

$$\mathcal{H}_S^\varphi = \left\{ \boxed{h_K^\varphi(\boldsymbol{y}|\boldsymbol{x}) := \sum_{k=1}^{K} g_k(\boldsymbol{x}; \boldsymbol{\gamma}) \phi_q(\boldsymbol{y}; \boldsymbol{\mu}_k, \sigma_k)} ; \quad \phi_q \in \mathcal{E}_\varphi \cap \mathcal{L}_\infty, g_k(\cdot; \boldsymbol{\gamma}) \in \{\text{softmax}\} \right\}$$

Theorem : Approximation capabilities of isotropic mixtures-of-experts SGaME

(a) For $p \in [1, \infty)$, $f \in \mathcal{F}_p \cap \mathcal{C}$, $\varphi \in \mathcal{F} \cap \mathcal{C}$, $\mathbb{X} = [0,1]^d$, there exists a sequence $\left(h_K^\varphi\right) \subset \mathcal{H}_S^\varphi$ such that $\lim_{K \to \infty} \left\| f - h_K^\varphi \right\|_{\mathcal{L}_p} = 0$.

(b) For $f \in \mathcal{F} \cap \mathcal{C}$, if $\varphi \in \mathcal{F} \cap \mathcal{C}$, $d = 1$, there exists a sequence $\left(h_K^\varphi\right) \subset \mathcal{H}_S^\varphi$ such that $\lim_{K \to \infty} h_K^\varphi = f$ almost uniformly.

[PhD TT. Nguyen, 2021] [Journal of Stat. Distributions and Applicat., 2021] [Neurocomputing, 2019] [WIREs DMKD 2018]

Open-Source Toolkit

Learning via the EM algorithm

SaMUraiS : open source software for statistical time-series analysis

SaMUraiS : **S**t**A**tistical **M**odels for the **U**nsupe**R**vised segment**A**t**I**on of time-**S**eries

▸ Github ▸ CRAN ▸ Matlab software

Available algorithms and Packages

`RHLP` : Regression with Hidden Logistic Process	▸ R software	▸ Matlab software
`HMMR` : Hidden Markov Model Regression	▸ R software	▸ Matlab software
`PWR` : Piece-Wise Regression	▸ R software	▸ Matlab software
`MRHLP` : Multivariate RHLP	▸ R software	▸ Matlab software
`MHMMR` : Multivariate HMMR	▸ R software	▸ Matlab software
`MPWR` : Multivariate PWR	▸ R software	▸ Matlab software

Include estimation, segmentation, approximation, model selection, and sampling

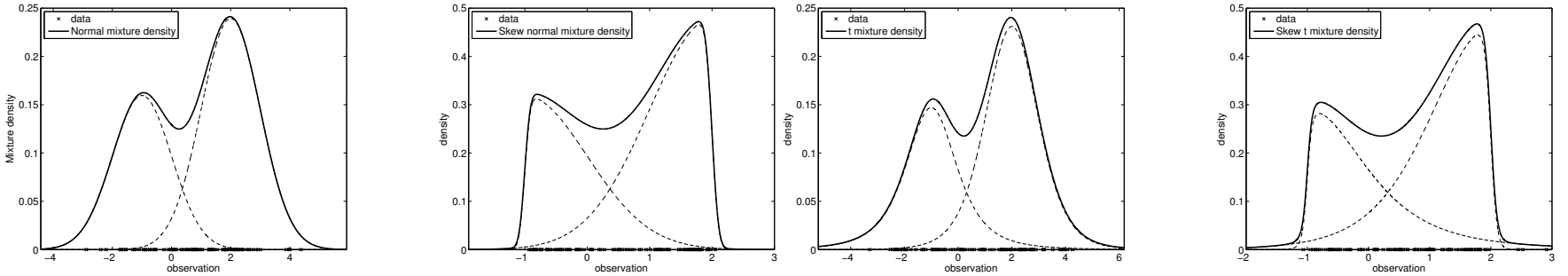
Principled robustness in learning with MoE

Principled robustness in regression and clustering

- **Questionings :** Prediction (non-linear regr., classification) & clustering in presence of **Outliers**, with potentially **skewed**, **heavy-tailed** distributions
- **Answering :** Robust MoE that accommodate asymmetry, heavy tails, and outliers

$$m(y|\boldsymbol{r},\boldsymbol{x};\boldsymbol{\theta}) = \sum_{k=1}^{K} \underbrace{g_k(\boldsymbol{r};\boldsymbol{\alpha})}_{\text{Softmax Gating Network}} \underbrace{\mathcal{ST}(y;\mu(\boldsymbol{x};\boldsymbol{\beta}_k),\sigma_k,\lambda_k,\nu_k)}_{\text{Skew-}t\text{ Expert Network}}$$

kth expert : has a skew t distribution [Azzalini and Capitanio 2003]



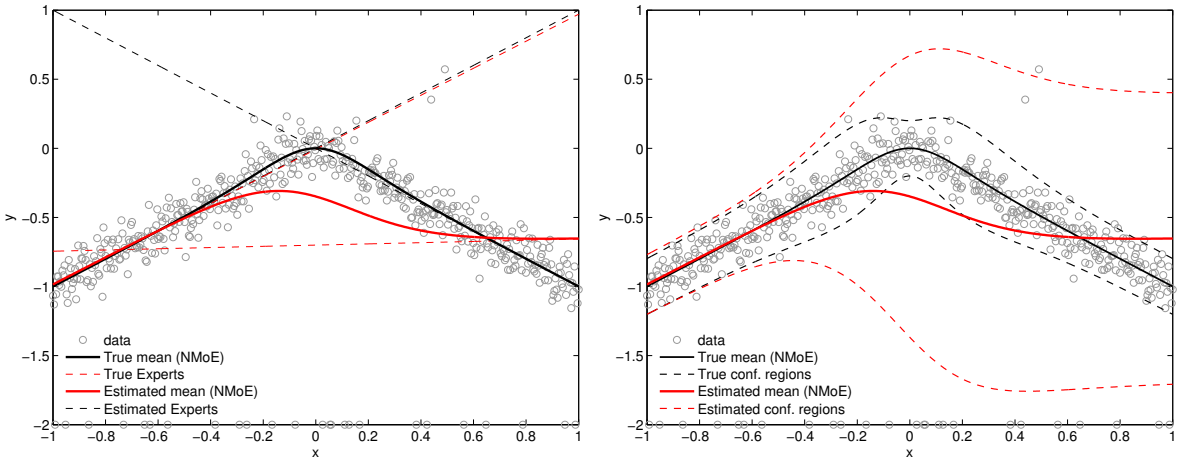
$\pi_k = [0.4, 0.6]$, $\mu_k = [-1, 2]$; $\sigma_k = [1, 1]$; $\nu_k = [3, 7]$; $\lambda_k = [14, -12]$;

Flexible and robust generalization of the standard MoE models

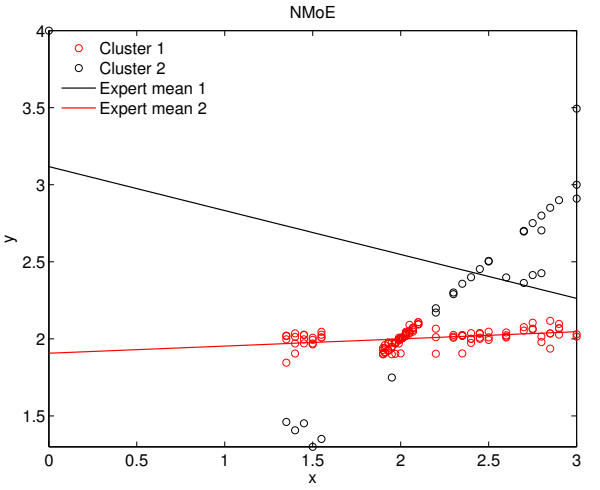
For $\{\nu_k\} \to \infty$, STMoE reduces to SNMoE ; For $\{\lambda_k\} \to 0$, STMoE reduces to TMoE.
For $\{\nu_k\} \to \infty$ and $\{\lambda_k\} \to 0$, StMoE approaches the NMoE.

[*Neurocomputing*, 2017] [*Neural Networks*, 2016c] [IJCNN 2016]

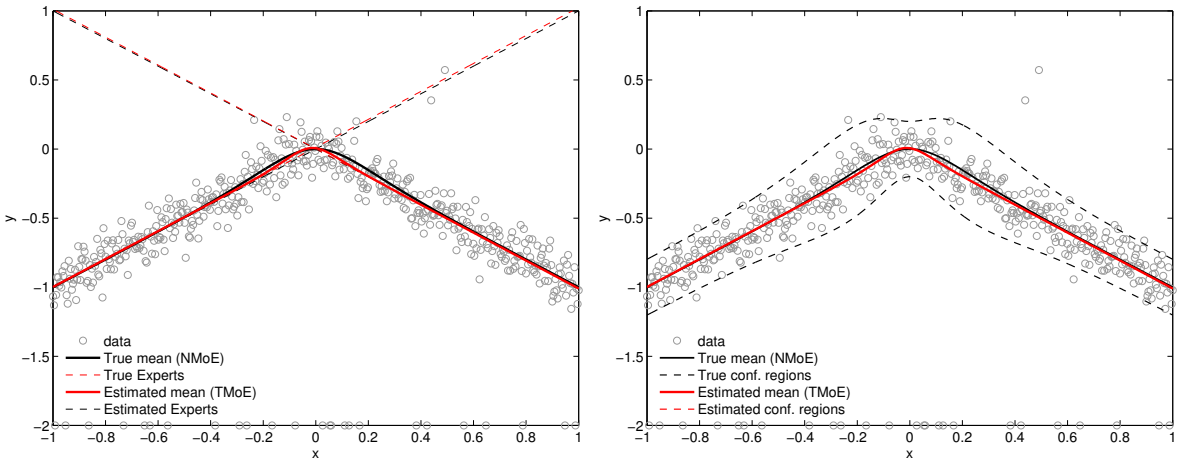
Robust learning with mixtures-of-experts models



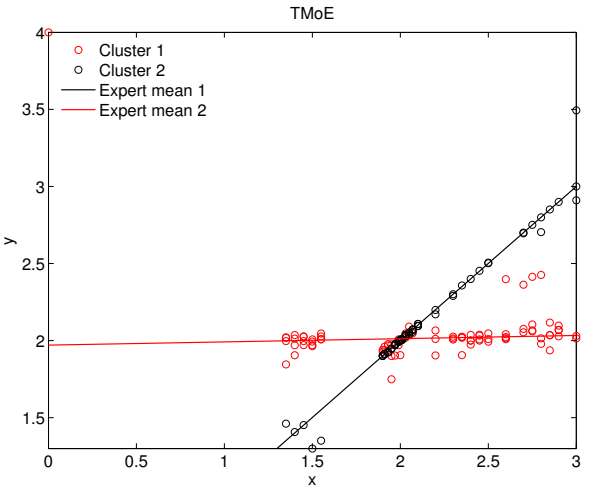
$n = 500$ observations with 5% of outliers $(x; y = -2)$: Normal fit



Tone data with 10 outliers $(0, 4)$: Normal fit



$n = 500$ observations with 5% of outliers $(x; y = -2)$: **Robust** fit

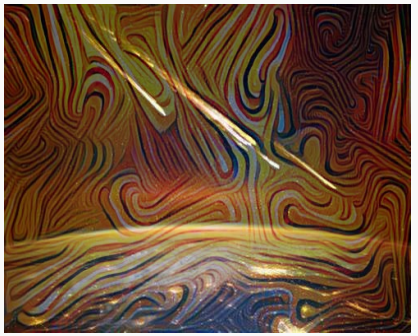


Tone data with 10 outliers $(0, 4)$: **Robust** fit

[*Neurocomputing*, 2017] [*Neural Networks*, 2016c] [IJCNN 2016]

Open-Source Toolkit

MEteorits : open-source soft. Robust learning with mixtures-of-experts models



MEteorits : **M**ixtures-of-**E**xper**T**s mod**E**ling for c**O**mplex and non-no**R**mal d**I**s**T**ributionS

▸ Github ▸ CRAN ▸ Matlab software

Available algorithms and Packages

`NMoE` : Normal Mixture-of-Experts	▸ R software	▸ Matlab software
`SNMoE` : Skew-Normal Mixture-of-Experts	▸ R software	▸ Matlab software
`tMoE` : Robust MoE using the t-distribution	▸ R software	▸ Matlab software
`StMoE` : Skew-t Mixture-of-Experts	▸ R software	▸ Matlab software

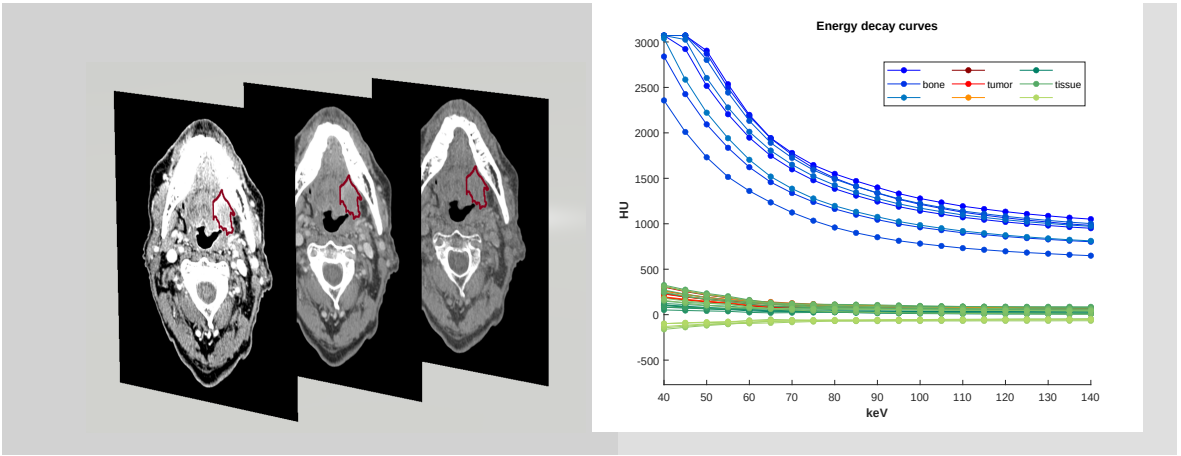
- Meteorits include sampling, fitting, prediction, clustering with each MoE model
- Non-normal mixtures (and MoE) is a very recent topic in the field

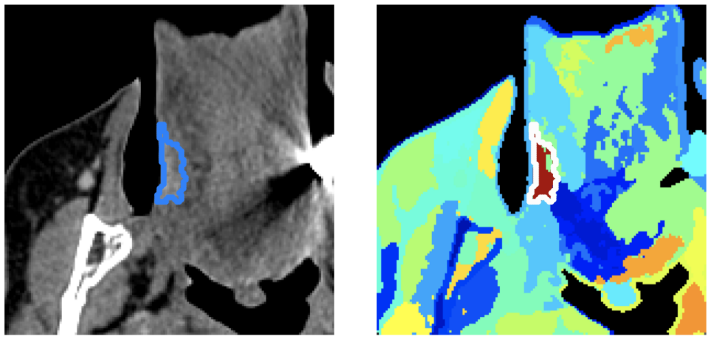
Dual-energy computed tomography (DECT) image Clustering

- Learning from Multimodal information in Healthcare/Radiology
- Cancer detection in Radiology : DECT clustering [Diagnostics (AI in medicine), 2022]

  Spatial mixture of functional regressions for dual-energy CT images

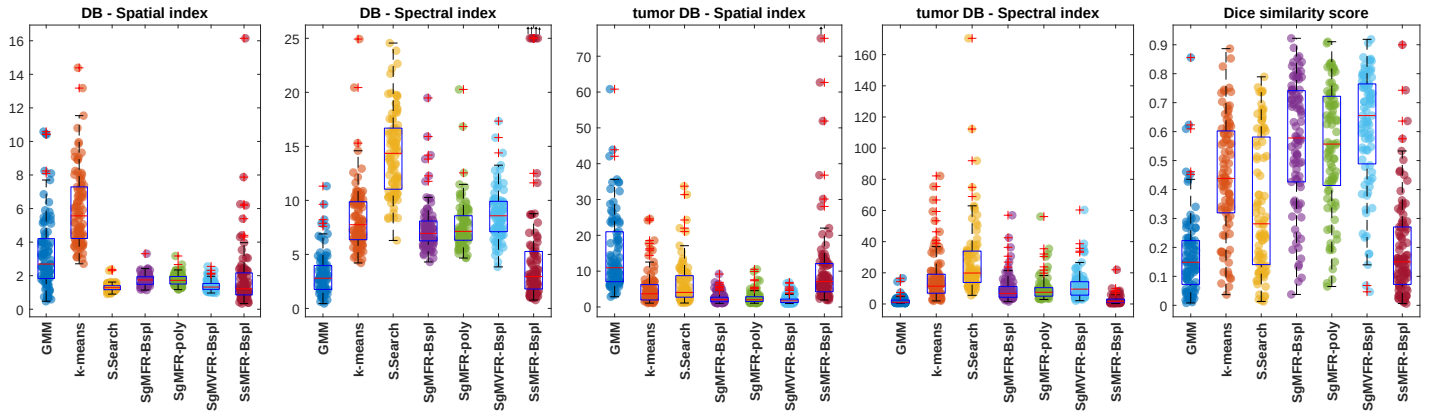
  $m(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{v};\boldsymbol{\theta}) = \sum_{k=1}^{K} \alpha_k(\boldsymbol{v};\boldsymbol{\alpha}) f_k(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta}_k)$ where $\alpha_k(\boldsymbol{v};\boldsymbol{\alpha}) = \frac{w_k \phi_3(\boldsymbol{v};\boldsymbol{\mu}_k,\mathbf{R}_k)}{\sum_{\ell=1}^{K} w_\ell \phi_3(\boldsymbol{v};\boldsymbol{\mu}_\ell,\mathbf{R}_\ell)}$





(a) Original slice (b) Dice=0.84, DB=1.64/6.92

DECT multimodal Data : 3D voxels & energy levels Expert Annotation Automatic Annotation



Questioning : Prediction (non-linear regr., classification) & clustering in presence of

[1.] **High-dimensional** predictors : $\boldsymbol{X}_i \in \mathbb{R}^p$ with $p \gg n$

[2.] **Functional** predictors : $X_i(t)$, $t \in \mathcal{T} \subseteq \mathbb{R}$ {eg. continuously recorded variables}

$\hookrightarrow$ Look for parsimonious and interpretable methods

[1.] HDME : High-Dimensional Mixtures-of-Experts

- Learning : PMLE $\widehat{\boldsymbol{\theta}}_{\mathsf{n}} \in \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log h_K^{\varphi}(\boldsymbol{y}_i|\boldsymbol{x}_i;\boldsymbol{\theta}) - \mathsf{pen}(\boldsymbol{\theta})$
- $\hookrightarrow$ LASSO penalty : $\mathsf{Pen}_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \underbrace{\sum_{k=1}^{K} \lambda_k \|\boldsymbol{\beta}_k\|_1}_{\text{Experts Net.}} + \underbrace{\sum_{k=1}^{K-1} \gamma_k \|\boldsymbol{w}_k\|_1}_{\text{Gating Net.}}$

$\hookrightarrow$ encourages sparse solutions & performs estimation and feature selection

$\hookrightarrow$ computationally attractive (Avoid matrix inversion ; univariate updates)

▸ Software Toolbox HDME on Github (GaussRMoE, LogisticRMoE, PoissonRMoE)

[PhD] Bao Tuyen Huynh. *Estimation and Feature Selection in High-Dimensional Mixtures-of-Experts Modesls* . PhD Thesis, Normandie Université, 2019.

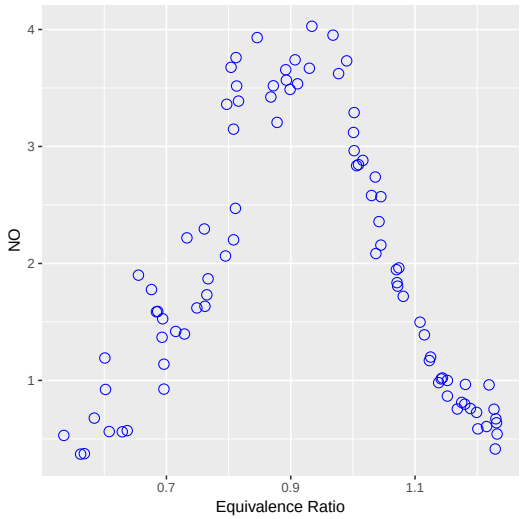
[J] Chamroukhi &Huynh. Regularized Maximum Likelihood Estimation and Feature Selection in Mixtures-of-Experts Models. Journal de la Société Francaise de Statistique, Vol. 160(1), pp :57–85, 2019

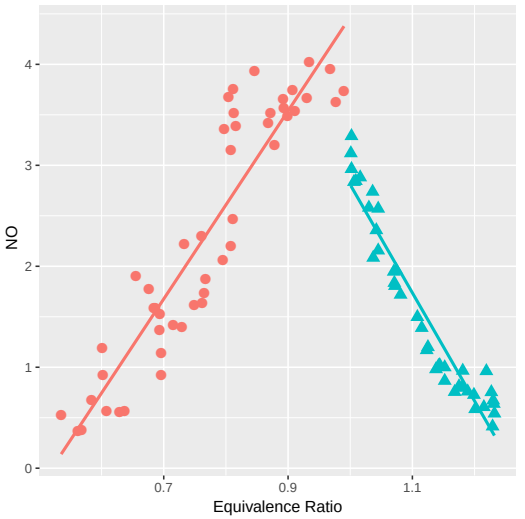
[J] Huynh & C. Estimation and Feature Selection in Mixtures of Generalized Linear Experts Models. arXiv :1810.12161, 2019

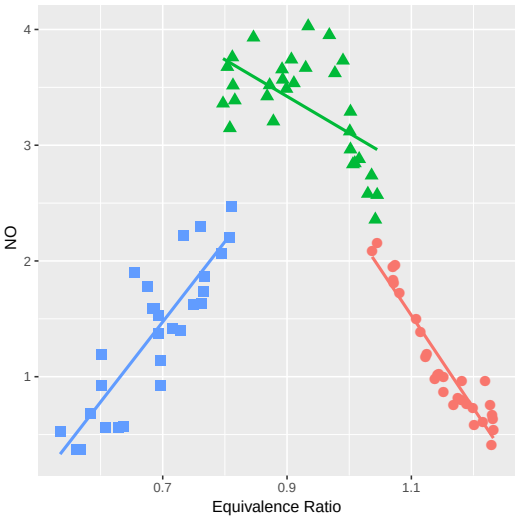
Model selection

(a) Raw Ethanol data set

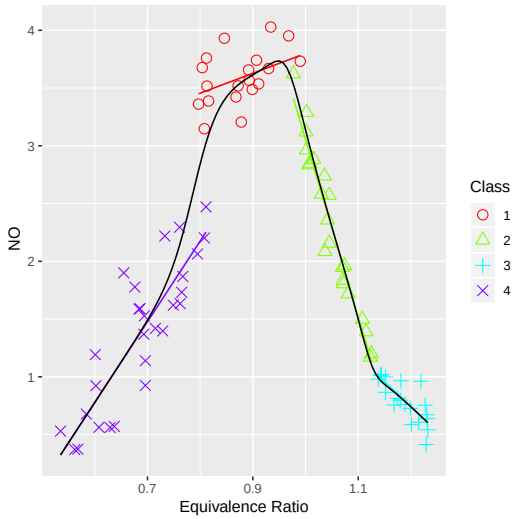
Collection of MoE models with linear mean functions characterized by 2-5 clusters

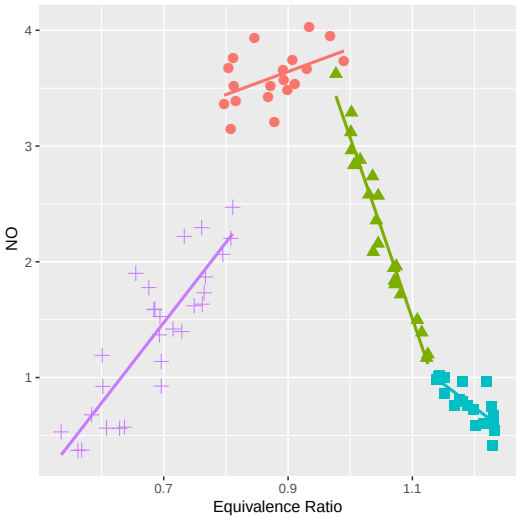


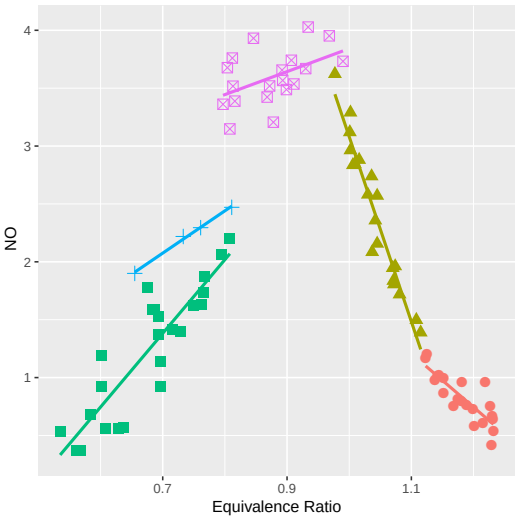




(b) Our best data-driven MoE model







Measuring uncertainty in high-dimensional learning

Questioning : Prediction (non-linear regr., classification) & clustering in presence of **High-dimensional** predictors : Data $\mathcal{D}_n = (\boldsymbol{X}_i, Y_i)_{i=1}^n$ where $\boldsymbol{X}_i \in \mathbb{R}^p$ with $p \gg n$

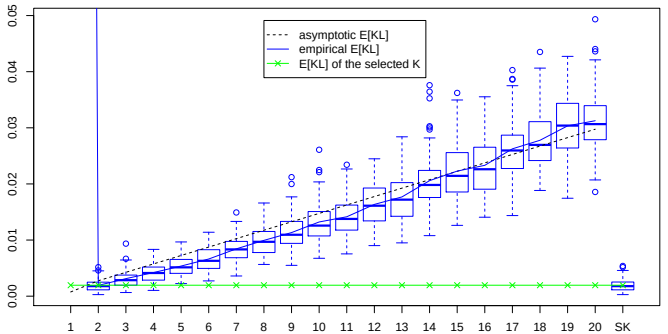
`HDME` : High-Dimensional MoE : PMLE $\widehat{\boldsymbol{\theta}}_{\mathsf{n}} \in \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^n \log h_K^{\varphi}(\boldsymbol{y}_i|\boldsymbol{x}_i;\boldsymbol{\theta}) - \mathrm{pen}(\boldsymbol{\theta})$

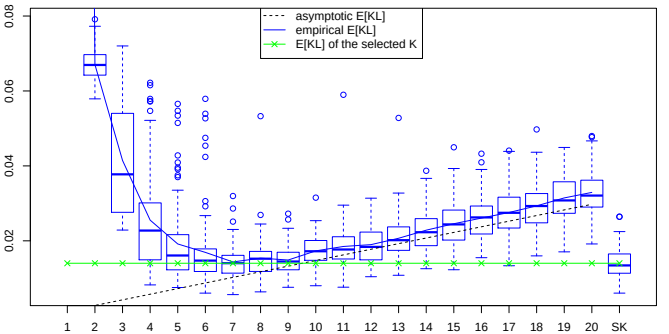
Theorem : Non-asymptotic oracle inequality for collection of MoE models

Result : $\exists$ constants C et $\kappa(\rho, C_1) > 0$ $(C_1 > 1)$ s.that whenever for $\mathbf{m} \in \mathcal{M}$, $\mathrm{pen}(\mathbf{m}) \geq \kappa(\rho, C_1)\left[(C + \ln n)\dim(\mathcal{H}_{\mathbf{m}}) + z_{\mathbf{m}}\right]$, the estimator PMLE $\widehat{h}_{\widehat{\mathbf{m}}}$ satisfies

$$\mathbb{E}\left[\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}\left(f, \widehat{h}_{\widehat{\mathbf{m}}}\right)\right] \leq C_1 \inf_{\mathbf{m}\in\mathcal{M}}\left(\inf_{h_{\mathbf{m}}\in\mathcal{H}_{\mathbf{m}}} \mathrm{KL}^{\otimes \mathrm{n}}(f, h_{\mathbf{m}}) + \frac{\mathrm{pen}(\mathbf{m})}{n}\right) + \frac{\kappa(\rho, C_1) C_1 \xi}{n} + \frac{\eta + \eta'}{n}.$$

- A non-asymptotic result. If $\mathrm{pen}(\mathbf{m})$ is well chosen, then our PMLE behaves in a comparable manner compared to **the best (oracle) model** $\mathcal{H}_{\mathbf{m}^\star}$ in the collection, minimizing the risk : $\inf_{\mathbf{m}\in\mathcal{M}}\left(\inf_{h_{\mathbf{m}}\in\mathcal{H}_{\mathbf{m}}} \mathrm{KL}^{\otimes \mathrm{n}}(f, h_{\mathbf{m}}) + \frac{\mathrm{pen}(\mathbf{m})}{n}\right)$ (f is unknown).





[PhD, Trung-Tin Nguyen, 2021.] [*Electronic Journal of Statistics*, 2022] [In revision, *JMVA*. arXiv :2104.08959. 2021b]

Functional Data Analysis (Open-Source Toolkit)

FLaMingoS : open source software for learning from functions

FLaMingoS : **F**unctional **L**atent dat**A** **M**odels for cluster**ING** heterogene**O**us time-**S**eries

▸ Github ▸ CRAN ▸ Matlab software

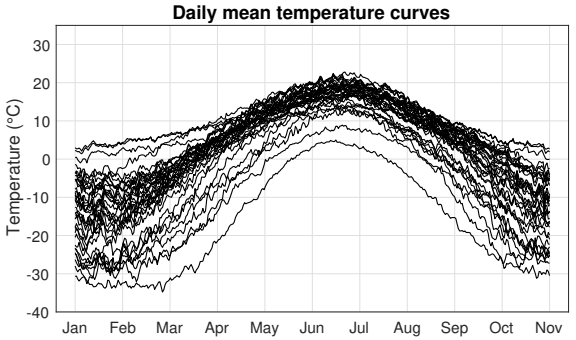
Available algorithms and Packages

`mixRHLP` : Mixture of Regressions with HLPs	▸ R	▸ Matlab
`mixHMM` : Mixture of Hidden Markov Models (HMMs)	▸ R	▸ Matlab
`mixHMMR` : Mixture of HMM Regressions	▸ R	▸ Matlab
`PWRM` : Piece-Wise Regression Mixture	▸ R	▸ Matlab
`uReMix` : Unsupervised Regression Mixtures	▸ R	▸ Matlab

↪ A flexible full generative modeling for FDA

↪ Could be extended to the multivariate case without a major effort

[2.] Learning with functional predictors



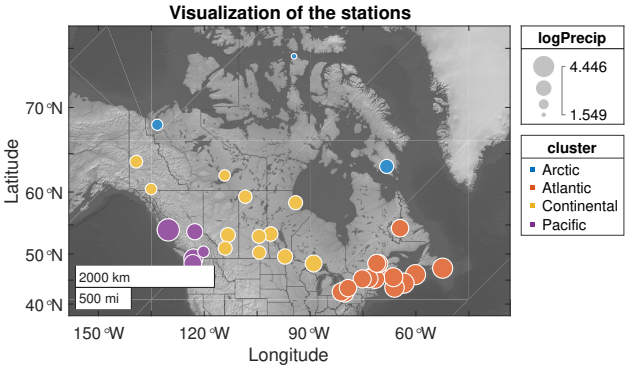


FIGURE – $n = 35$ daily mean temperature measurement curves (X_i's) in different stations (Left) and the log of precipitation values (Y_i's) visualized with the climate regions (Z_i's) (Right).

- Relate functional predictors $\{X(t) \in \mathbb{R}; t \in \mathcal{T} \subset \mathbb{R}\}$ to a scalar response $Y \in \mathcal{Y} \subset \mathbb{R}$
- Regression and classification of heterogeneous responses given functional predictors

  (1) generative functional modeling, sparsity and feature selection (high-dimension)

  (2) User guideline : keep an interpretable fit

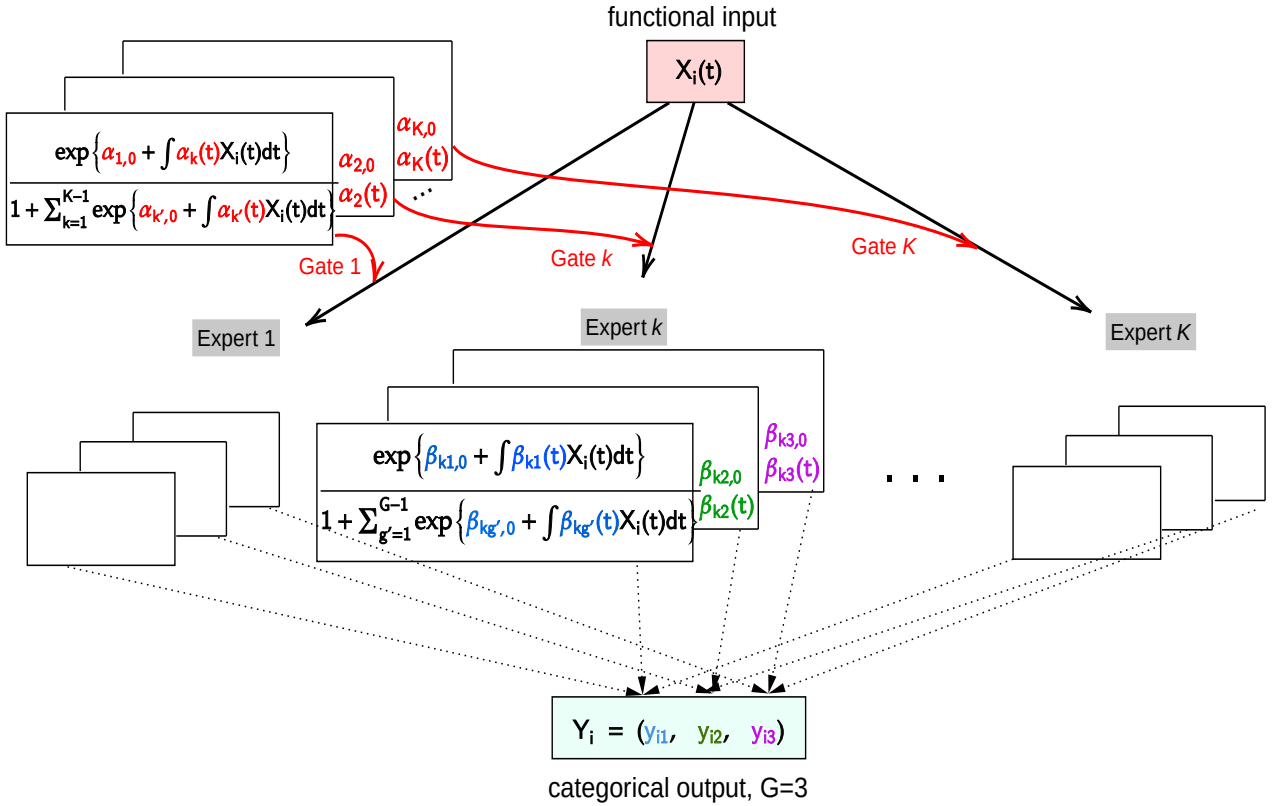
[2.] Functional Mixtures-of-Experts (and Different Learning strategies, in particular)

- $Y_i = \beta_{z_i,0} + \int_{\mathcal{T}} X_i(t)\beta_{z_i}(t)dt + \varepsilon_i$ avec $h_z(X_i(.)) = \alpha_{z_i,0} + \int_{\mathcal{T}} X_i(t)\alpha_{z_i}(t)dt$
- Lasso-type Regularized MLE w.r.t the derivatives of the $\alpha(\cdot)$ and $\beta(\cdot)$ functions

[Preprint] Chamroukhi, Pham, Hoang, McLachlan. Functional Mixtures-of-Experts. arXiv :2202.02249, Feb, 2022 (Under Review, Statistics and Computing)

Interpretable learning with time-series inputs

Mixture-of-Experts Architecture



$Y_i = \beta_{z_i,0} + \int_{\mathcal{T}} X_i(t)\beta_{z_i}(t)dt + \varepsilon_i$ with $h_z(X_i) = \alpha_{z_i,0} + \int_{\mathcal{T}} X_i(t)\alpha_{z_i}(t)dt$

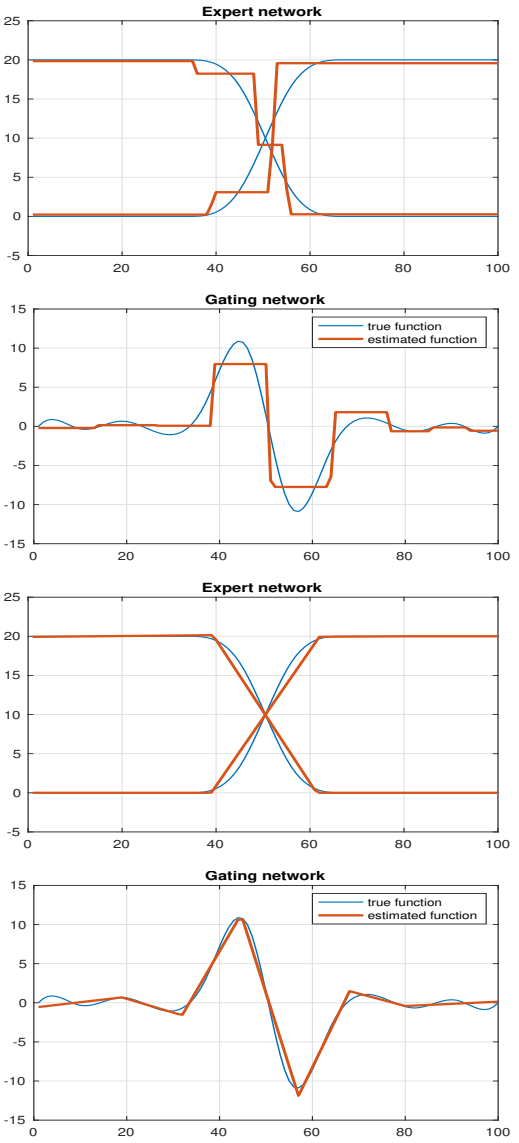
l_1-Regularized MLE w.r.t the derivatives of the $\alpha(\cdot)$ and $\beta(\cdot)$ functions

↪ produces a meaningful sparse estimates for $\beta_{z_i}(t)$ curves :

- $\beta^{(0)}_{z_i}(t) = 0$ implies that $X(t)$ has no effect on Y at t
- $\beta^{(1)}_{z_i}(t) = 0$ means that $\beta_{z_i}(t)$ is constant at t,
- $\beta^{(0)}_{z_i}(t) = 1$ shows that $\beta_{z_i}(t)$ is a linear function of t, etc.

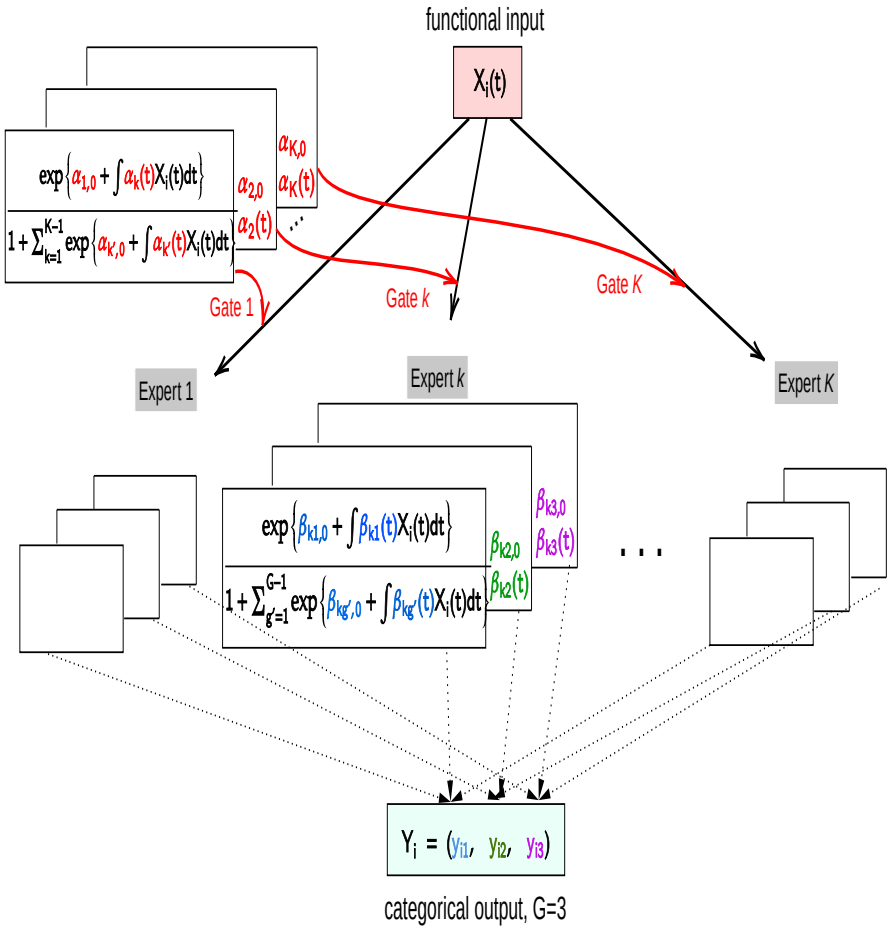
[PhD TN. Pham, 2022] [arXiv :2202.13934] [Submitted, 2023]

Interpretable fits

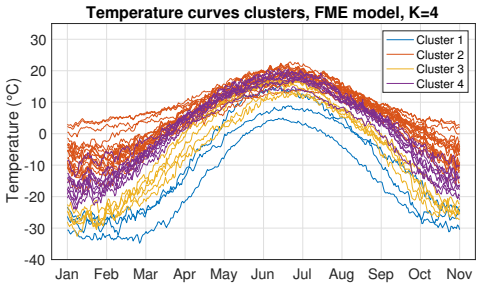


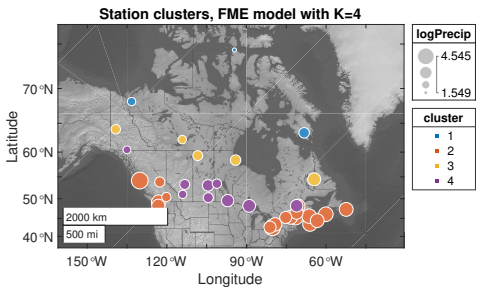
Interpretable learning with time-series inputs

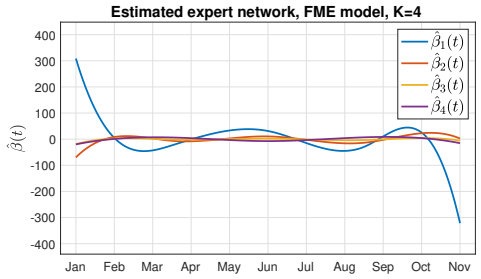
Mixture-of-Experts Architecture

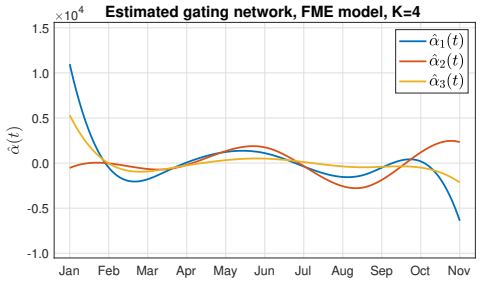


No regularization

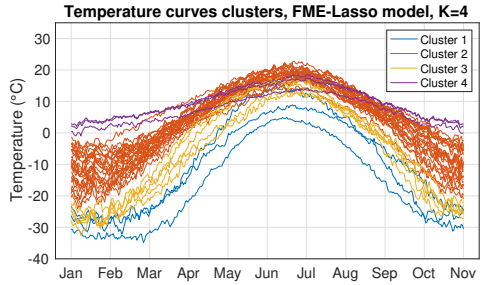


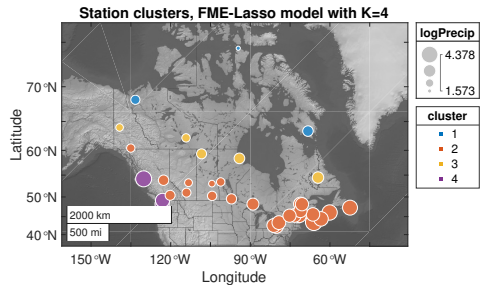


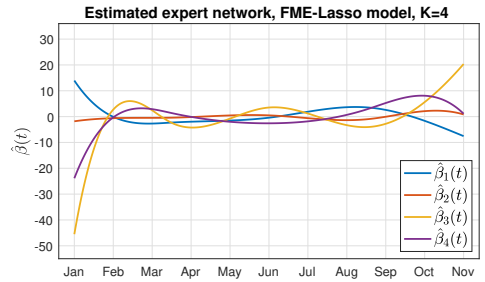


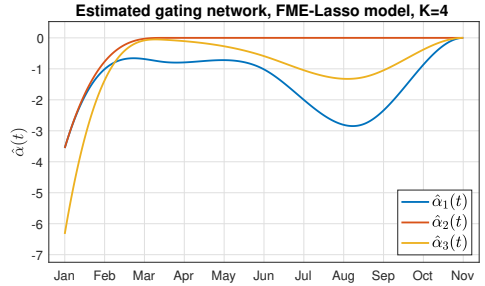


LASSO regularization





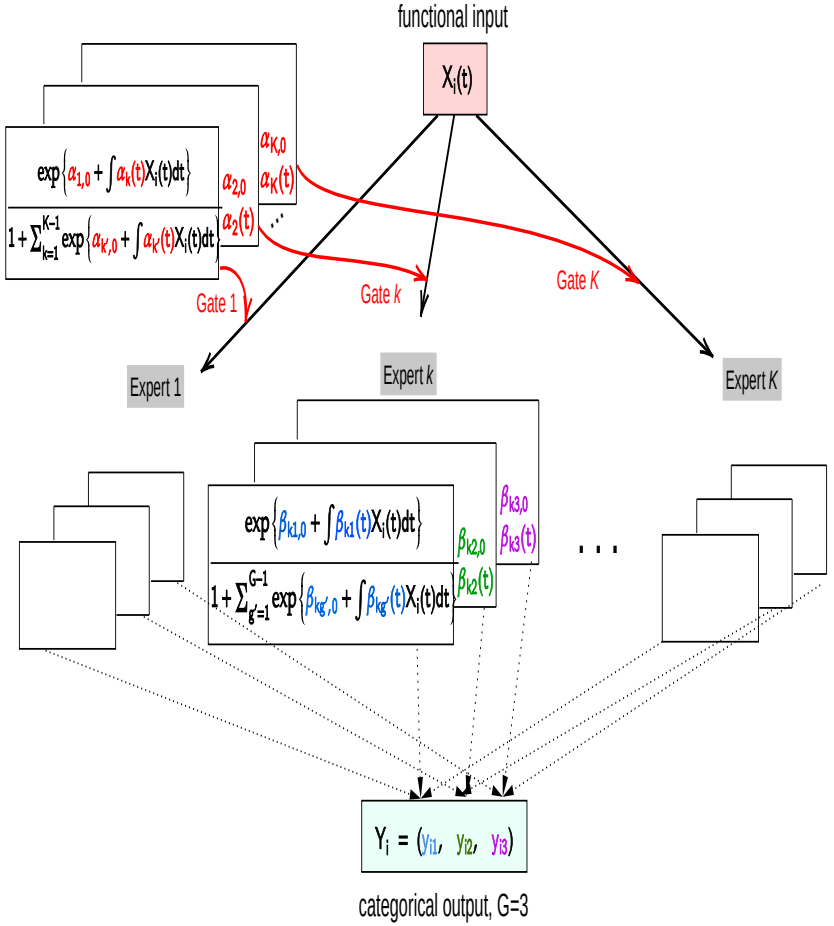




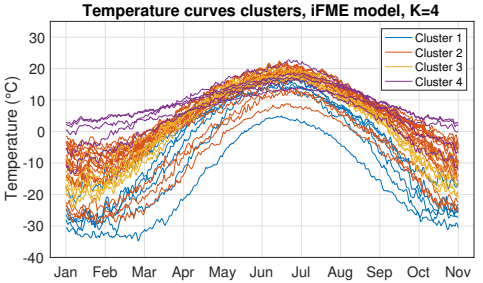
[PhD TN. Pham, 2022] [arXiv :2202.13934] [Submitted, 2023] [Contract, ANR SMILES]

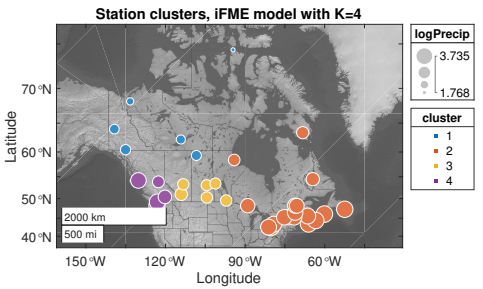
Interpretable learning with time-series inputs

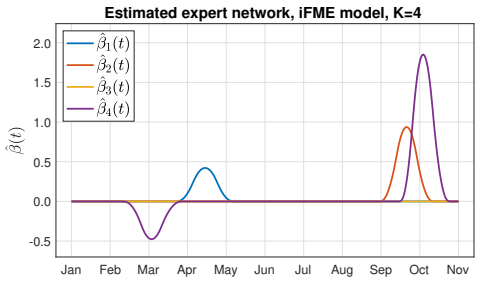
Mixture-of-Experts Architecture

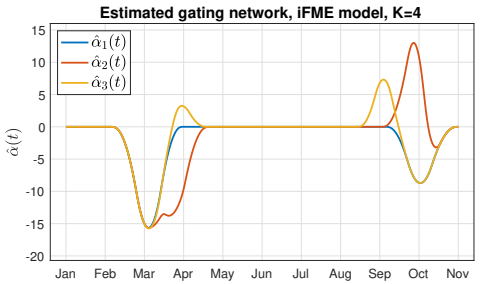


OUR regularization

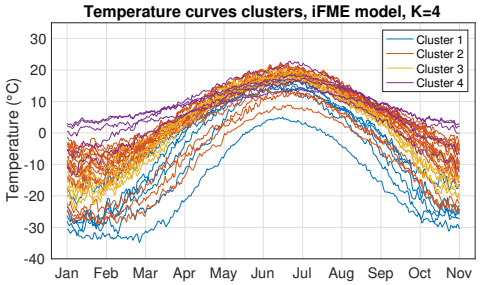


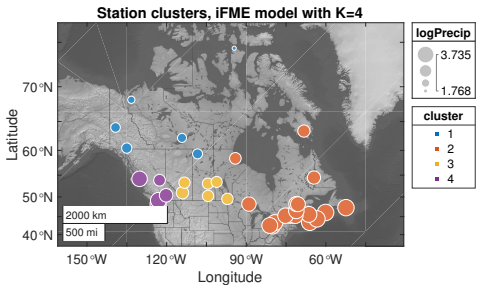


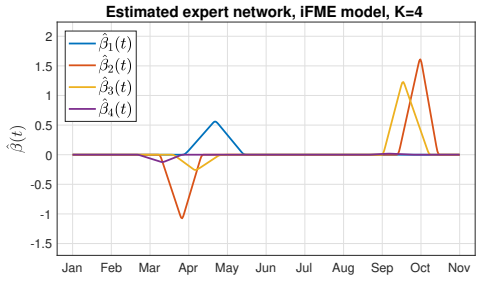


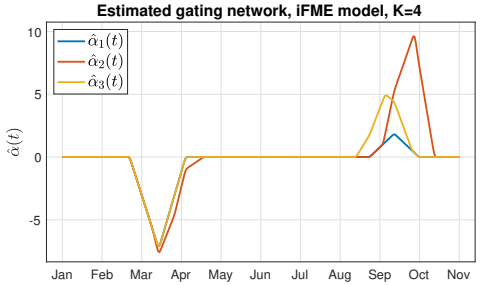


OUR regularization









produces a meaningful sparse estimates for $\beta_{z_i}(t)$ curves :

$\beta^{(0)}_{z_i}(t) = 0$ implies that $X(t)$ has no effect on Y at t

$\beta^{(1)}_{z_i}(t) = 0$ means that $\beta_{z_i}(t)$ is constant at t,

$\beta^{(0)}_{z_i}(t) = 1$ shows that $\beta_{z_i}(t)$ is a linear function of t, etc.

[PhD TN. Pham, 2022] [arXiv :2202.13934] [Submitted, 2023] [Contract, ANR SMILES]

Interpretable learning with functional inputs

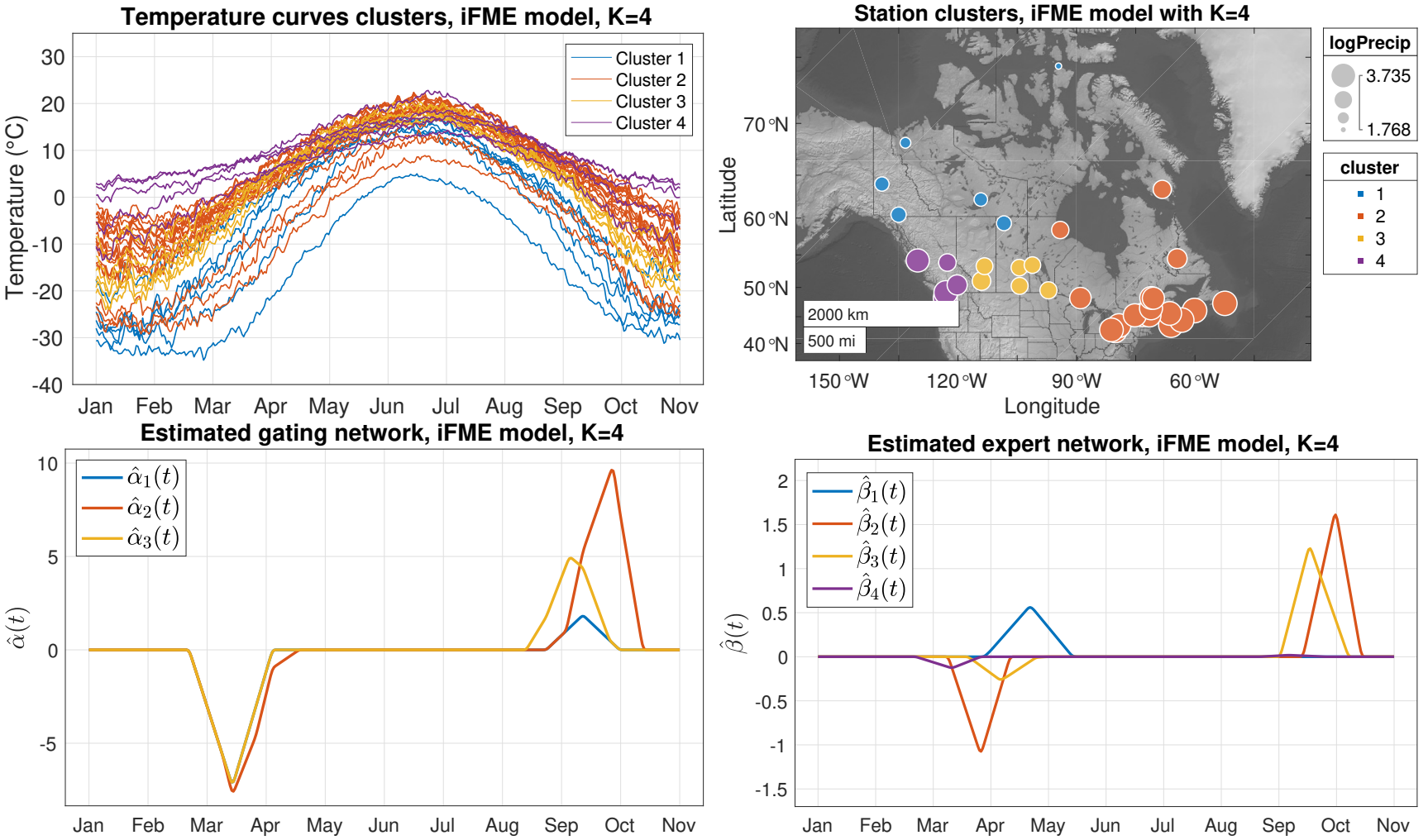


FIGURE – Clustering of temperatures (g.) predicted (log)-precipitations (d.) (Atlantic, Pacific, Continental and Arctic), and parameter functions (bg - bd.)

[PhD TN. Pham, 2022] [arXiv :2202.13934] [Submitted, 2023] [Contract, ANR SMILES]

Some references

- Faicel Chamroukhi, Thien Nhat Pham, Van Ha Hoang and Geoffrey J McLachlan. Functional Mixtures-of-Experts. arXiv preprint arXiv :2202.02249, 2022
- TrungTin Nguyen, Faicel Chamroukhi, Hien D. Nguyen and Geoffrey J. McLachlan. Approximation of probability density functions via location-scale finite mixtures in Lebesgue spaces. Communications in Statistics - Theory and Methods Taylor & Francis., Vol. 52(14), pages :5048-5059, 2023
- Nguyen, TrungTin, Nguyen, Hien Duy, Chamroukhi, Faicel and Forbes, Florence. A non-asymptotic approach for model selection via penalization in high-dimensional mixture of experts models. Electronic Journal of Statistics Institute of Mathematical Statistics and Bernoulli Society., Vol. 16(2), pages :4742-4822, 2022
- Chamroukhi, Faicel, Brivet, Segolene, Savadjiev, Peter, Coates, Mark and Forghani, Reza. DECT-CLUST : Dual-Energy CT Image Clustering and Application to Head and Neck Squamous Cell Carcinoma Segmentation. Diagnostics, Vol. 12(12), pages :3072, 2022
- Nguyen Hien Duy, Nguyen TrungTin, Chamroukhi Faicel and McLachlan Geoffrey John. Approximations of conditional probability density functions in Lebesgue spaces via mixture of experts models. Journal of Statistical Distributions and Applications SpringerOpen., Vol. 8(1), pages :1-15, 2021
- Nguyen, T Tin, Nguyen, Hien D, Chamroukhi, Faicel and McLachlan, Geoffrey J. Approximation by finite mixtures of continuous density functions that vanish at infinity. Cogent Mathematics & Statistics Cogent OA., Vol. 7(1), pages :1750861, 2020
- Faicel Chamroukhi and Bao T. Huynh. Regularized Maximum Likelihood Estimation and Feature Selection in Mixtures-of-Experts Models. Journal de la Soci Franse de Statistique, Vol. 160(1), pages :57-85, March, 2019

Thank you for your attention !