Statistical Learning from heterogenous, high-dimensional and distributed data

Faïcel Chamroukhi

Seminar, 04 april 2023, Versailles

name@irt-systemx.fr Chamroukhi.com ChamroukhiF FChamroukhi ChamroukhiF GoogleScholar

The Research and Technology Organisation (RTO) of Paris-Saclay

Research and Technology Organisation (RTO)
Non-profit Scientific Cooperation Foundation

Paris-Saclay

65
Economic partners
of which **1/3** are large groups
and **2/3** are SMEs

50
Academic
partners

Leads market-driven and applied research projects for the digital transformation of industry, services and territories:

1. Expertise: analysis, modeling, simulation and decision management
2. Own skills
3. Own assets: software, cyber-physical and tool-based platforms

5 main application domains

- Autonomous transport and Mobility
- Future industry
- Defense and Security
- Environment and Sustainable development
- Digital and Health

8 Scientific and technological fields

- Data science and AI
- Interaction and uses
- Scientific computing
- Optimisation
- Systems engineering
- Safety
- Digital security and blockchain
- IoT and networks

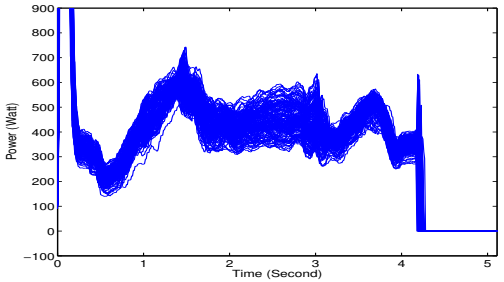
Founding members: ALSTOM, Bull atos technologies, COSMOTECH, IDnomic, OVHcloud, Renault Group, SAFRAN, SHERPA, Institut Polytechnique de Paris, Inria, Institut Mines-Télécom, université Paris-Saclay, Systematic Paris Region Digital Ecosystem

03/04/2023 An overview of SystemX 1

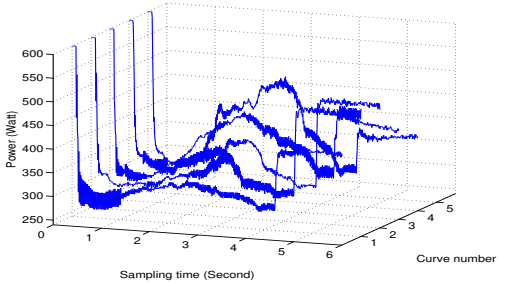
Real-world data are complex

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

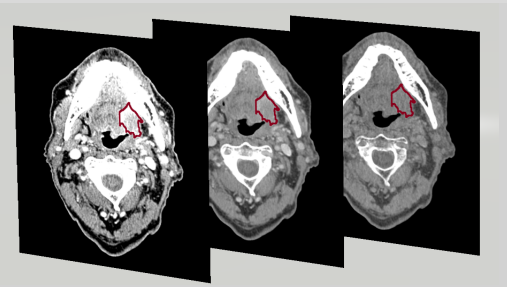
Transport : Railway switch curves diagnostic



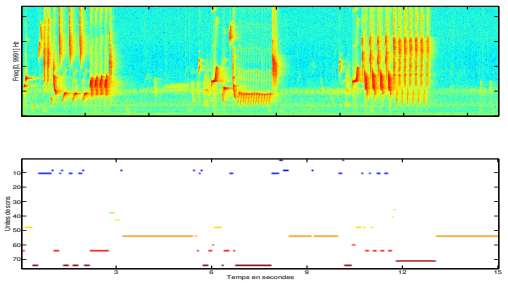
Predictive Maintenance

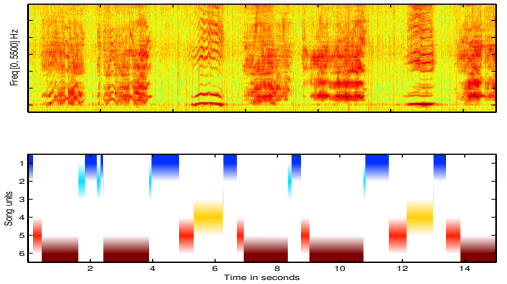


Health : Medical images

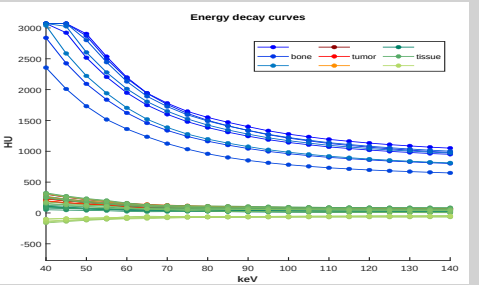


Acoustics : scene listening (marine, terrestrial)

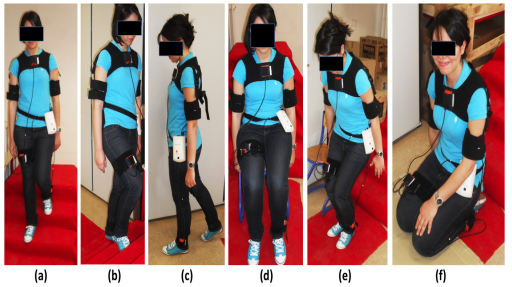




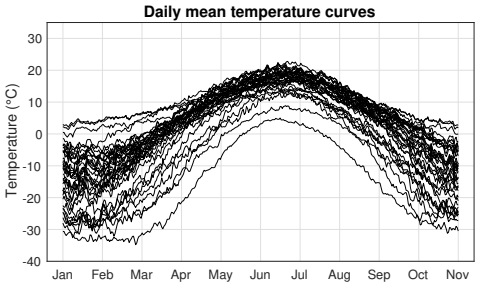
Dual-energy computed tomography

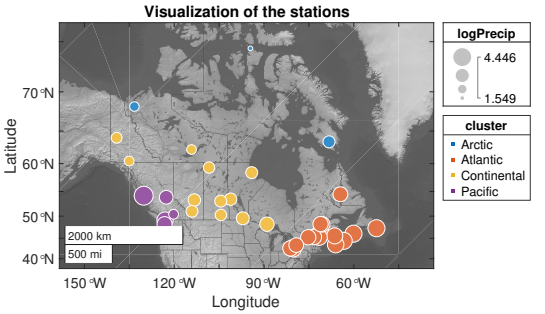


Health & Well Being : Activity recog.



Climate/Environment : meteorological data





Scientific Challenges

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

1. Deal with heterogeneous and potentially unlabeled data in different applications (e.g., time series, images,) $\hookrightarrow$ Need for models that explicitly accommodate heterogeneity and unsupervised analysis
2. Deal with high-dimensional data in complex (heterogeneous) situations, user-friendly $\hookrightarrow$ Need for models that encourage sparse solutions, while being interpretable
3. Learning form distributed (decentralized) data $\hookrightarrow$ Federating learning, with statistical and computational guarantees
4. Measure the precision quality in estimation & prediction $\hookrightarrow$ How far we are from optimal solutions, eg. Oracle inequalities

Outline

Cadre scientifique général

- ↪ **Modèles** à variables latente : $f(x|\boldsymbol{\theta}) = \int_{\mathcal{Z}} f(x, z|\boldsymbol{\theta})\mathrm{d}z$
- ↪ **Inférence**, **Sélection** et **représentation non supervisées** et à l'échelle

Approximation capabilities of finite mixture distributions

Density approximation in Unsupervised Learning

- **Data** : observations $\{\boldsymbol{x}_i\}$ d'une v.a $\boldsymbol{X} \in \mathbb{X} \subset \mathbb{R}^d$ à densité (multimodale) $f \in \mathcal{F}$
- **Objectif** : approcher la densité cible f (et représenter les données, e.g. *clustering*)
- **Solution :** Approcher f dans la classe $\mathcal{H}^\varphi = \bigcup_{K \in \mathbb{N}^\star} \mathcal{H}_K^\varphi$ des **mélanges finis** h_K^φ (à K-composants) **de translatées dilatées** d'une densité φ (e.g., gaussienne), où

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

Théorème : Universal approximation of finite mixtures models (FMM)

(a) Pour toute f.d.p $f, \varphi \in \mathcal{C}$ et un ensemble compact $\mathbb{X} \subset \mathbb{R}^d$, il existe une suite $(h_K^\varphi) \subset \mathcal{H}^\varphi$, telle que $\lim_{K \to \infty} \sup_{\boldsymbol{x} \in \mathcal{X}} |f(\boldsymbol{x}) - h_K^\varphi(\boldsymbol{x})| = 0$.

(b) Pour $p \in [1, \infty)$, si $f \in \mathcal{L}_p$ (f.d.p de Lebesgue) et $\varphi \in \mathcal{L}_\infty$ (f.d.p essentiellement bornée), il existe une suite $(h_K^\varphi) \subset \mathcal{H}^\varphi$, telle que $\lim_{K \to \infty} \|f - h_K^\varphi\|_{\mathcal{L}_p} = 0$.

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

Gaussian mixture models (GMMs)

The finite Gaussian mixture density is defined as :

$$h_K^{\mathcal{N}}(\boldsymbol{x}_i;\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_i;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$$

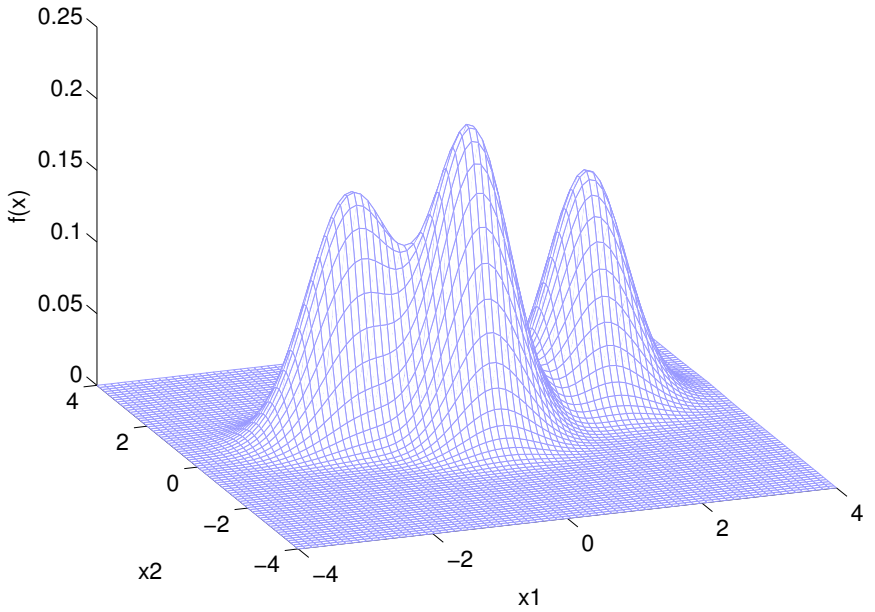


FIGURE – An example of a three-component Gaussian mixture density in $\mathbb{R}^2$.

Finite Mixture Models

$h_K^{\varphi}(\boldsymbol{x}_i;\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \varphi(\boldsymbol{x};\boldsymbol{\theta}_k)$ with $\pi_k > 0\ \forall k$ and $\sum_{k=1}^{K} \pi_k = 1$.

Maximum-Likelihood Estimation

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

log-likelihood : $\ln L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln \sum_{k=1}^{K} \pi_k \varphi(\boldsymbol{x}_i;\boldsymbol{\theta}_k)$.

The EM algorithm [DLR]

$$\boldsymbol{\theta}^{new} \in \arg\max_{\boldsymbol{\theta}\in\Omega} \mathbb{E}[\ln L_c(\boldsymbol{\theta})|\mathcal{D},\boldsymbol{\theta}^{old}]$$

complete log-likelihood : $\ln L_c(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{ik} \log\left[\pi_k \varphi(\boldsymbol{x}_i;\boldsymbol{\theta}_k)\right]$ where Z_{ik} is such that $Z_{ik} = 1$ if $Z_i = k$ and $Z_{ik} = 0$ otherwise.

Clustering

$$\widehat{z}_i = \arg\max_{1\leq k\leq K} \mathbb{P}(Z_i = k|\boldsymbol{x}_i;\widehat{\boldsymbol{\theta}}), \quad (i = 1,\ldots,n)$$

Max. de Vraisemblance : $\widehat{\boldsymbol{\theta}}_{\text{MLE}} \in \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta};\mathbf{x})$ with $L(\boldsymbol{\theta};\mathbf{x}) = \sum_{i=1}^{n} \ln h_K^{\varphi}(\boldsymbol{x}_i;\boldsymbol{\theta})$

↪ Algorithme.s EM {[DLR]} : $\boldsymbol{\theta}^{(\text{new})} \in \arg\max_{\boldsymbol{\theta}} \mathbb{E}\left[L_c(\boldsymbol{\theta};\mathbf{x},\mathbf{z})|\mathbf{x};\boldsymbol{\theta}^{(\text{old})}\right]$

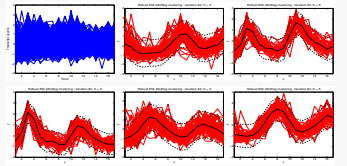
- Régularisation en apprentissage non-supervisé

↪ Inférence bayésienne de mélanges à effet mixtes

[J-13] *Journal of Statistical Computation and Simulation*, 2015.
[J-14] arXiv :1508.00635, 2015. et ESANN 2016

Images

Clustering en génomique



Apprentissage bayésien non-paramétrique

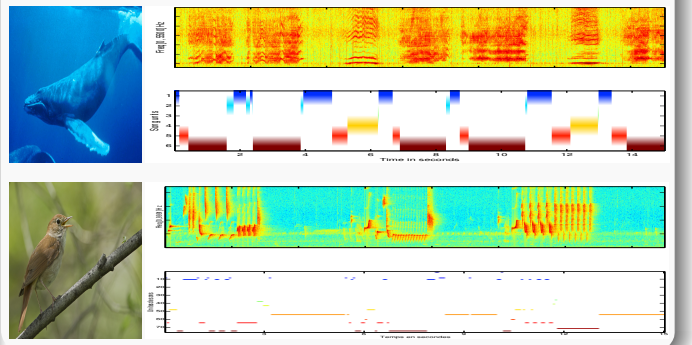
- Classification de séquences bio-acoustiques
- *Mélanges parcimonieux de Processus de Dirichlet*

↪ Cadre non-paramétrique : inférence et sélection

[J-15] FC et al. *Dirichlet Process Parsimonious Gaussian Mixture for clustering. arXiv :1501.03347v2, 2018*

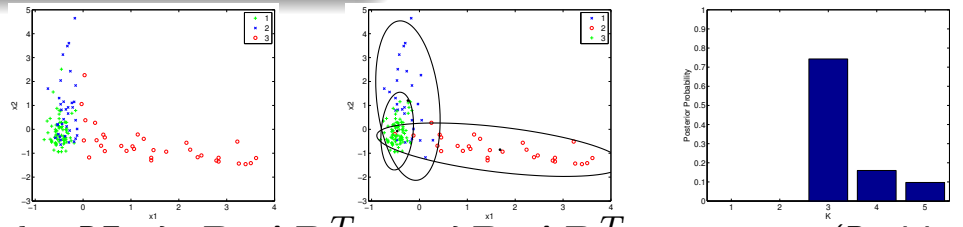
[Thèse] Marius BARTCUS, 2015, *UTLN*

Biologie - Découverte d'unités de chant



Diabetes data set

Clustering et choix de modèle



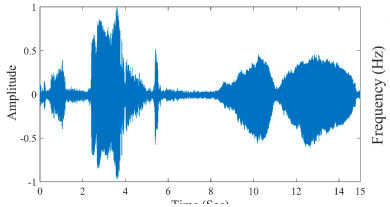
2 log BF : $\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T \;\; vs \;\; \lambda \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T \;=\; 199.58$ (Decisive)

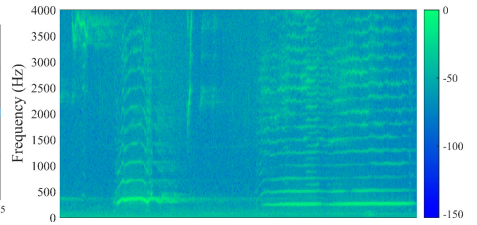
Automatic Scene Listening

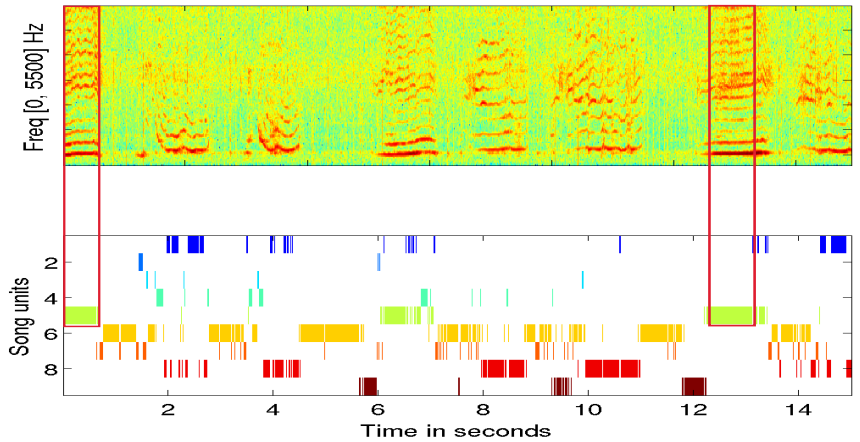
Unsupervised decomposition of whale song signals

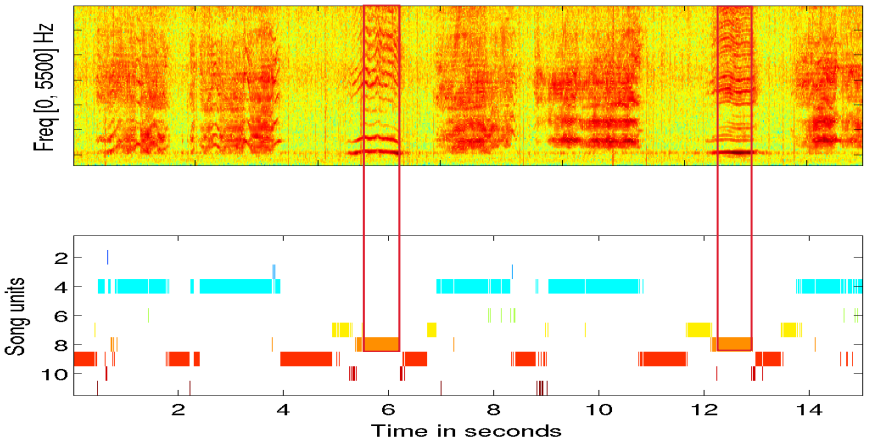
Biology/Bioacoustics/Environment

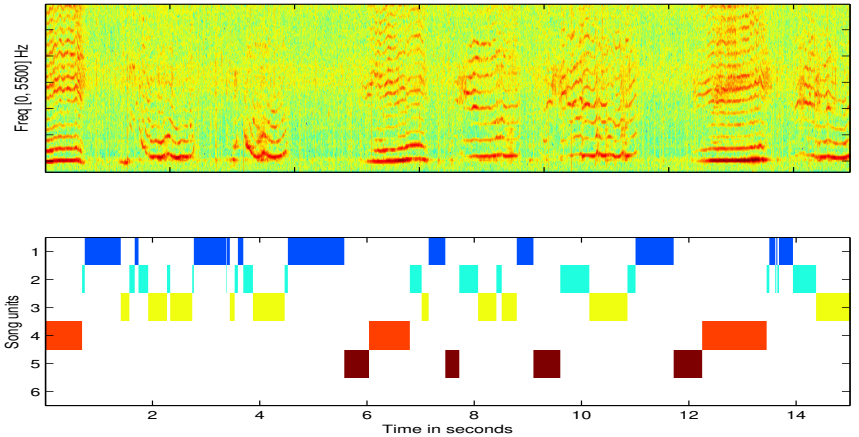
- Unsupervised decomposition (whale/bird song signals)

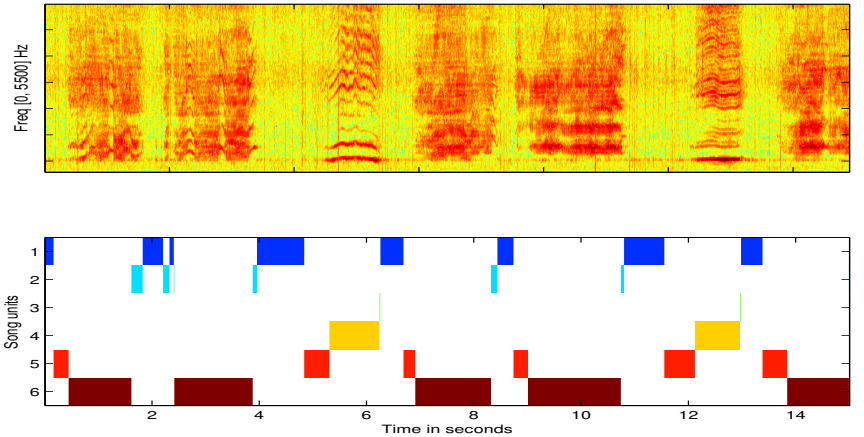








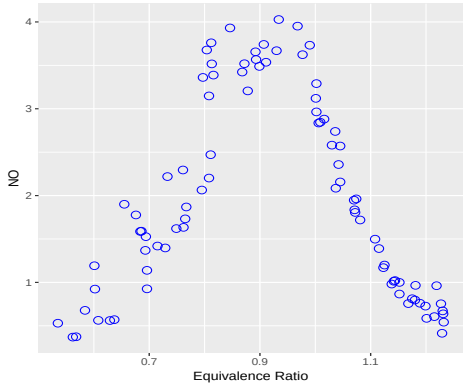


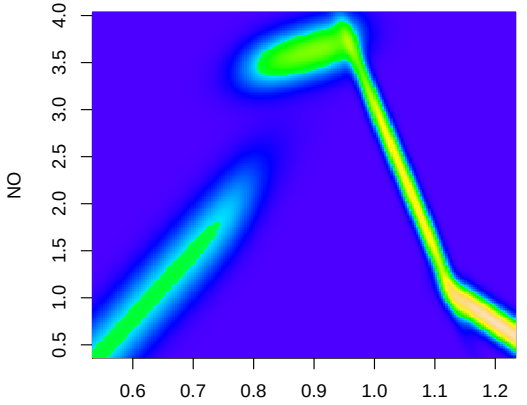


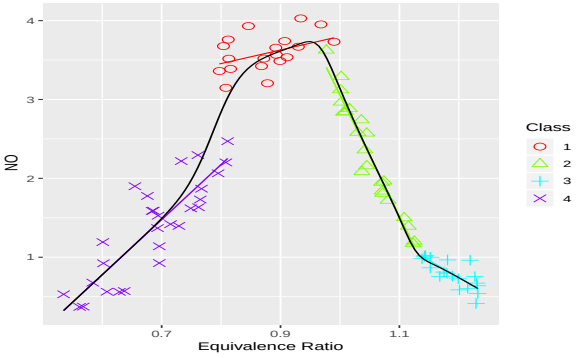
[PhD, M. Bartcus 2015] [Multimedia Tools and Applications, 2018] [wkp ICML 14, IPCR 14, IJCNN 15] [Sabiod.org]

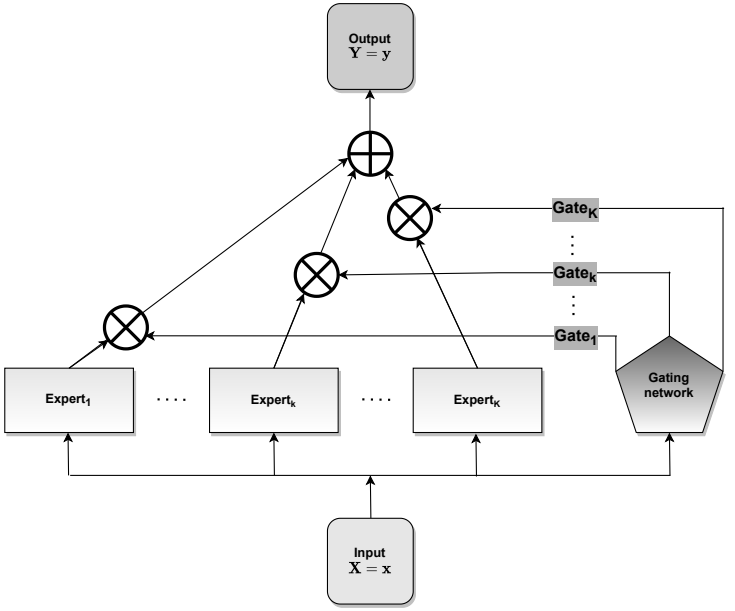
Heterogeneous regression-type data

Heterogeneous regression data : Pair of a response Y given covariarte.s X









Schematic diagram of the neural network architecture of a K-component MoE model.

Apprentissage par Modèles de Mélanges d'Experts (ME)

- **Contexte :** n observations $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}$ d'un couple $(\boldsymbol{X}, \boldsymbol{Y}) \in \mathbb{X} \times \mathbb{Y}$ lié via une f.d.p conditionnelle inconnue $f \in \mathcal{F} = \left\{f : \mathbb{X}\times\mathbb{Y} \to \mathbb{R}_+ \middle| \int_{\mathbb{Y}} f(\boldsymbol{y}|\boldsymbol{x})\, \mathrm{d}\lambda(\boldsymbol{y}) = 1, \forall \boldsymbol{x} \in \mathbb{X}\right\}$
- Scénario de **grande dimension** : $\mathbb{X} \subseteq \mathbb{R}^d$, $\mathbb{Y} \subseteq \mathbb{R}^q$, avec $d, q \gg n$ et **hétérogène**.
- **Objectifs :** Regression ; Clustering ; Sélection de modèle
- **Solution :** Approcher f dans la classe des **mélanges d'experts** :

Soit une f.d.p φ (et un support compact $\mathbb{Y} \subseteq \mathbb{R}^q$), on déifinit les classes suivantes :

- Transaltées-dilatées : $\mathcal{E}_\varphi = \left\{\phi_q(\boldsymbol{y}; \boldsymbol{\mu}, \sigma) := \frac{1}{\sigma^q}\varphi\left(\frac{\boldsymbol{y}-\boldsymbol{\mu}}{\sigma}\right); \boldsymbol{\mu} \in \mathbb{Y}, \sigma \in \mathbb{R}_+\right\}$.
- Mélanges d'experts transaltées-dilatés à réseau d'activation softmax : SGaME :

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

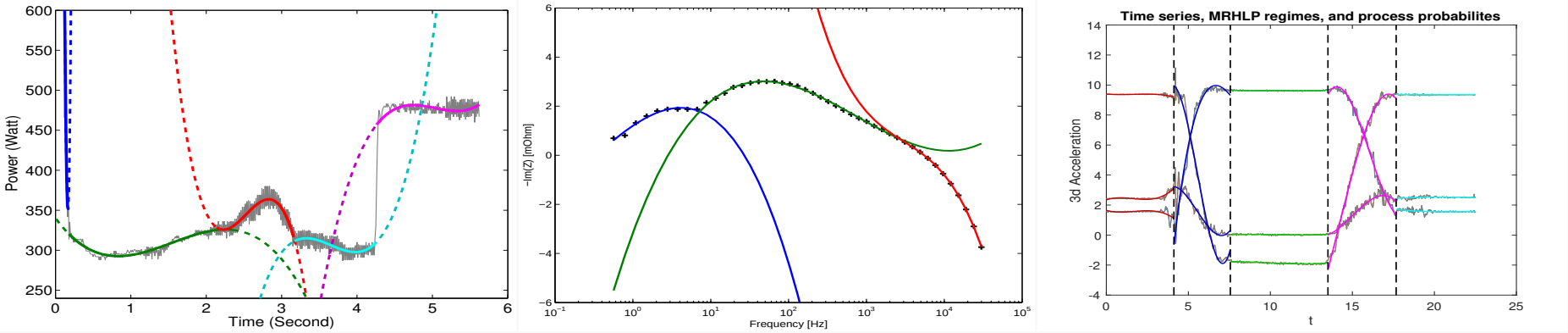
Theorem : Approximation capabilities of isotropic mixtures-of-experts SGaME

(a) Pour $p \in [1, \infty)$, $f \in \mathcal{F}_p \cap \mathcal{C}$, $\varphi \in \mathcal{F} \cap \mathcal{C}$, $\mathbb{X} = [0,1]^d$, il existe une suite $\left(h_K^\varphi\right) \subset \mathcal{H}_S^\varphi$ telle que $\lim_{K\to\infty} \left\|f - h_K^\varphi\right\|_{\mathcal{L}_p} = 0$.

(b) Pour toute $f \in \mathcal{F} \cap \mathcal{C}$, si $\varphi \in \mathcal{F} \cap \mathcal{C}$, $d = 1$, il existe une suite $\left(h_K^\varphi\right) \subset \mathcal{H}_S^\varphi$ telle que $\lim_{K\to\infty} h_K^\varphi = f$ presque uniformément.

[PhD TT. Nguyen, 2021] [Journal of Statistical Distributions and Applications, 2021] [Neurocomputing, 2019]

Time Series Modeling and Segmentation



Temporal data with unknown abrupt and/or smooth regime changes

Hidden Process Regression Models

$$y_i = \boldsymbol{\beta}_{z_i}^T \boldsymbol{x}_i + \sigma_{z_i}\epsilon_i \quad ; \epsilon_i \underset{\text{iid}}{\sim} \mathcal{N}(0,1), \qquad \mathbf{z} = (z_1, \ldots, z_n) \text{ : a hidden process}$$

$$h_K^{\mathcal{N}}(y_i|x_i;\boldsymbol{\theta}) = \sum_{k=1}^{K} g_k(x_i;\mathbf{w})\mathcal{N}\left(y_i;\boldsymbol{\beta}_k^T\boldsymbol{x}_i,\sigma_k^2\right); \qquad g_k(x_i;\mathbf{w}) = \mathbb{P}(Z_i = k|x_i;\mathbf{w})$$

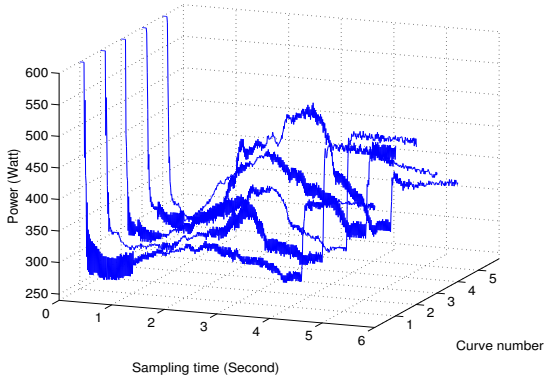
Optimization for Learning : $\widehat{\boldsymbol{\theta}}_{\text{MLE}} \in \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^n \log m(y_i|x_i;\boldsymbol{\theta})$

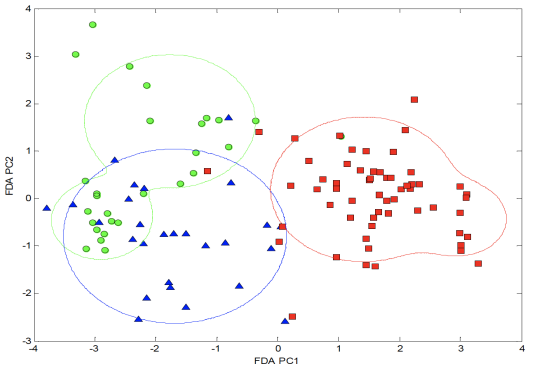
- MLE via the EM algorithm : $\boldsymbol{\theta}^{(q+1)} \in \arg\max_{\boldsymbol{\theta}} \mathbb{E}\left[L_c(\boldsymbol{\theta};\mathbf{x},\mathbf{y},\mathbf{z})|\mathbf{x},\mathbf{y};\boldsymbol{\theta}^{(q)}\right]$

[J-] WIREs DMKD 2018 — [J-] Neurocomputing 2010 & 2013 — [J-] ADAC 2011, [J-] IEEE TASE 2013, [J-] Sensors 2015

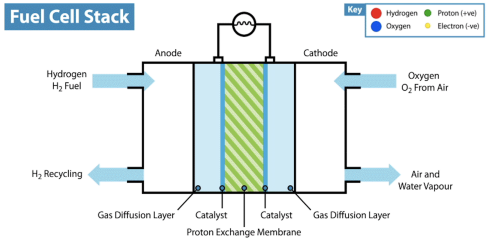
Time-Series Analysis Applications

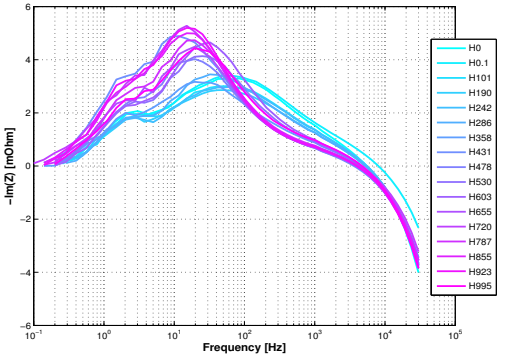
Transport : Railway switch operating state prediction {Collab. avec la SNCF ; Projet Switch-Rdf}

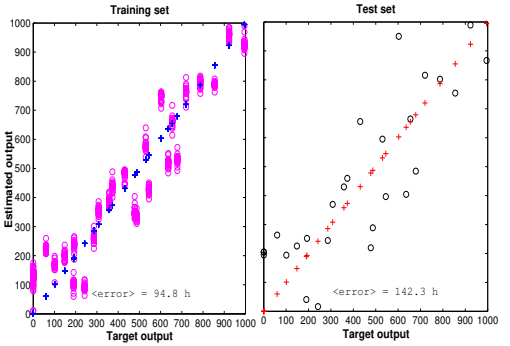




Energy : Fuel cell lifetime prediction {Collab. with Femto-ST, Phd of R. Onanena, 2012}

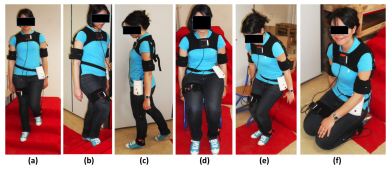


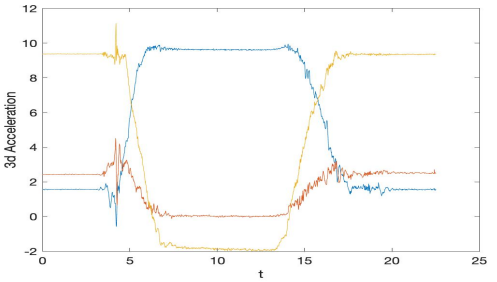


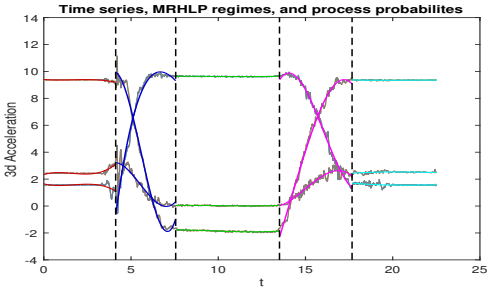


Health & well being : Human activity recognition {Collab. with Paris 12-LiSSi}

[PhD, D. Trabelsi, 2013][Sensors 2015, 748 citations]







Open-Source Toolkits

SaMUraiS : open source software for statistical time-series analysis

+13K téléchargements (à partir du canal R uniquement) depuis juillet 2019

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

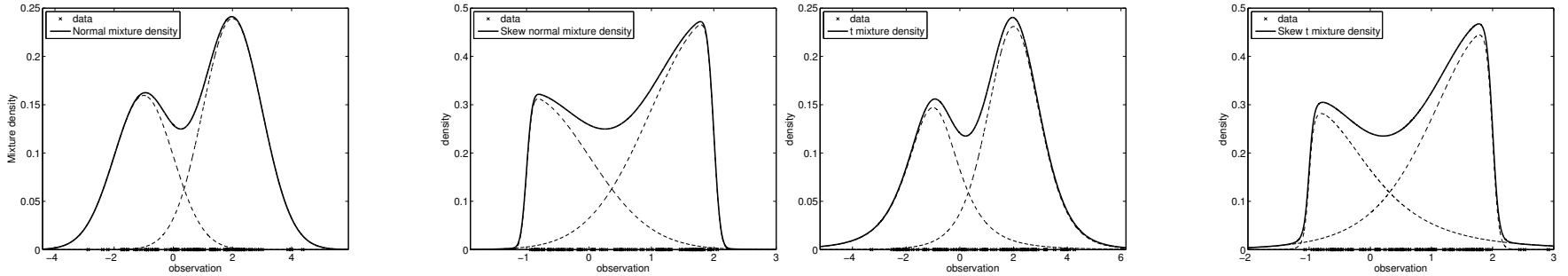
Principled robustness in learning with MoE

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]



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

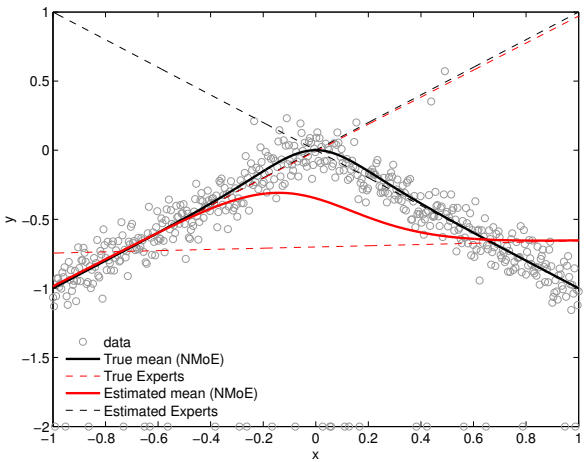
Flexible and robust generalization of the standard MoE models

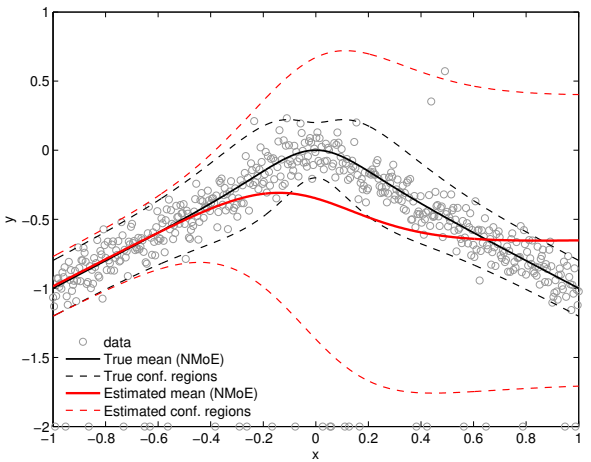
For $\{\nu_k\} \to \infty$, STMoE reduces to SNMoE ; For $\{\lambda_k\} \to 0$, STMoE reduces to TMoE.
For $\{\nu_k\} \to \infty$ and $\{\lambda_k\} \to 0$, StMoE approaches the NMoE.

[*Neurocomputing*, 2017] [*Neural Networks*, 2016c] [IJCNN 2016] [HDR 2015]

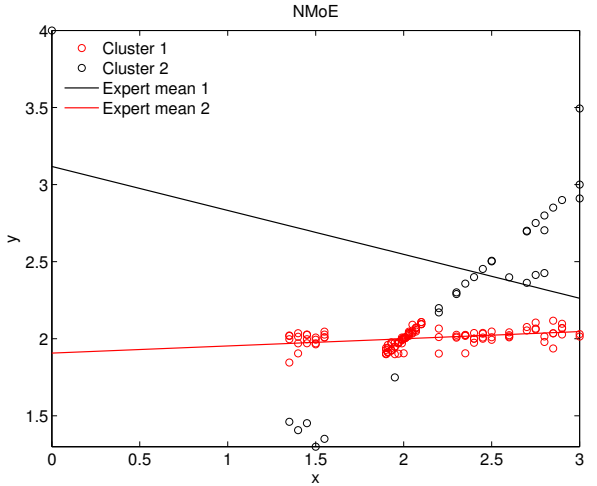
Robust learning with mixtures-of-experts models

Principled robustness in regression and clustering

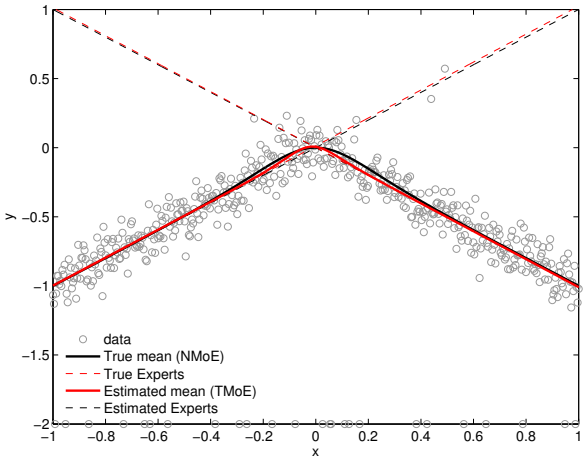


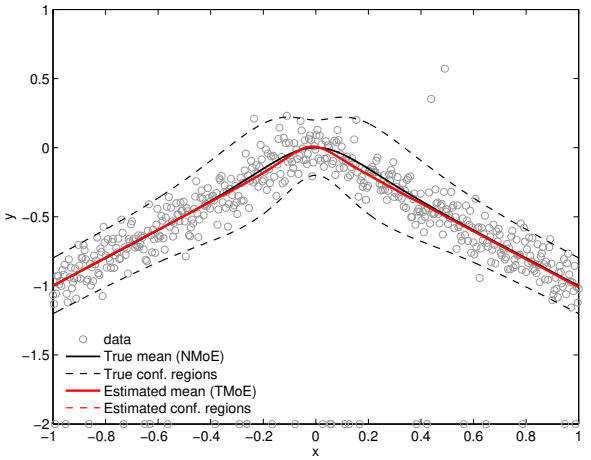


$n = 500$ observations with 5% of outliers $(x; y = -2)$: Normal fit

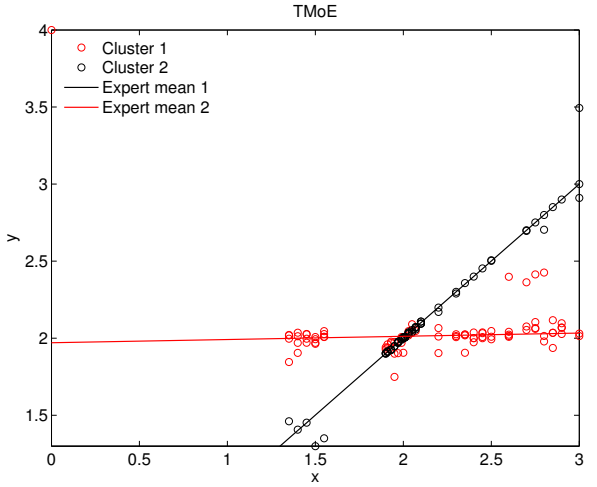


Tone data with 10 outliers $(0, 4)$: Normal fit





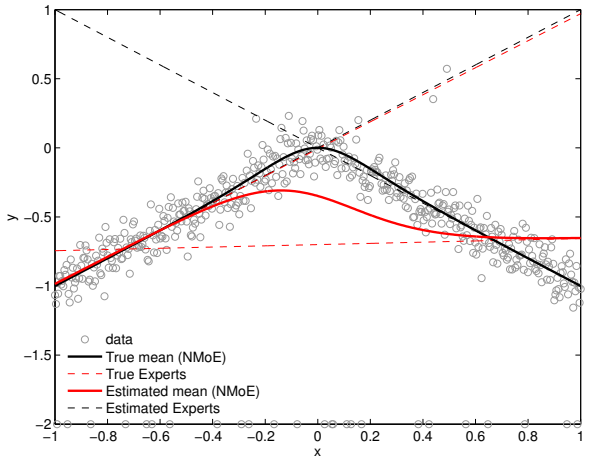
$n = 500$ observations with 5% of outliers $(x; y = -2)$: **Robust** fit

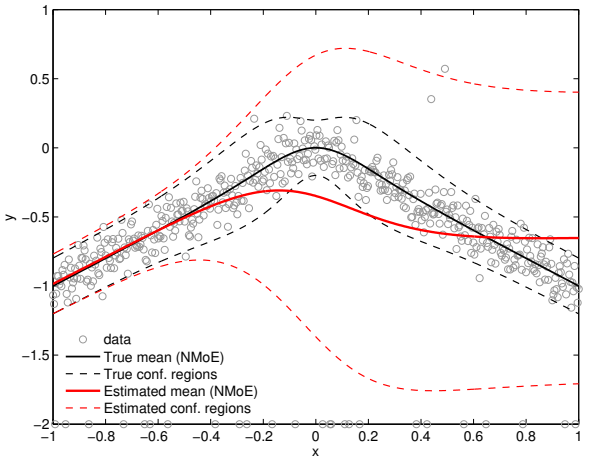


Tone data with 10 outliers $(0, 4)$: **Robust** fit

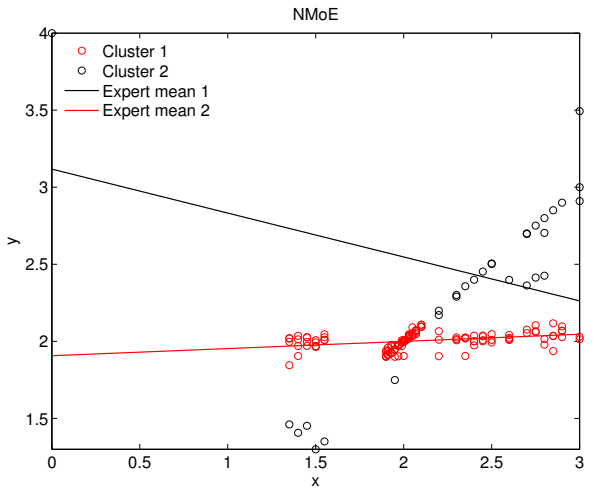
[*Neurocomputing*, 2017] [*Neural Networks*, 2016c] [IJCNN 2016] [HDR 2015]

Robustness in regression and clustering

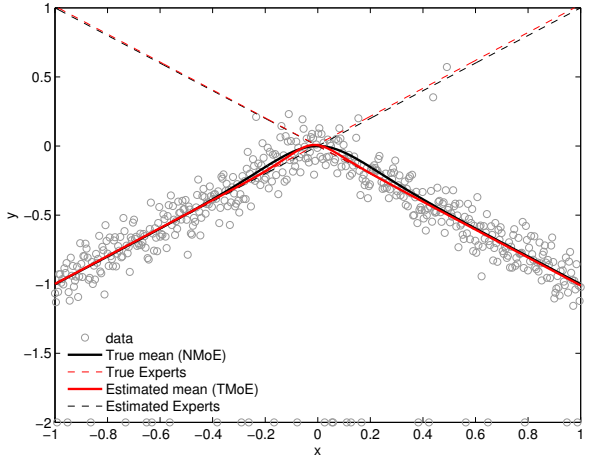


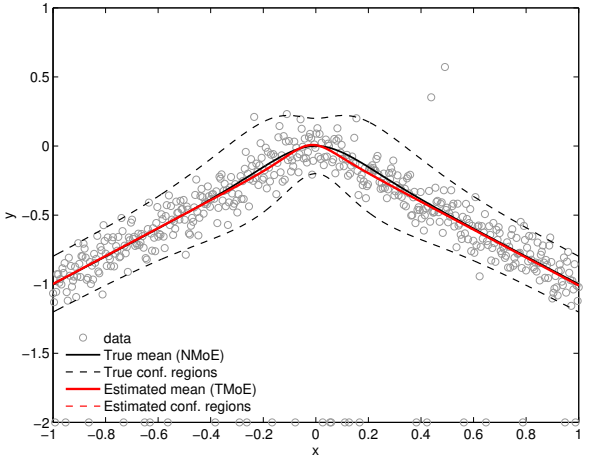


$n = 500$ observations with 5% of outliers $(x; y = -2)$: Normal fit

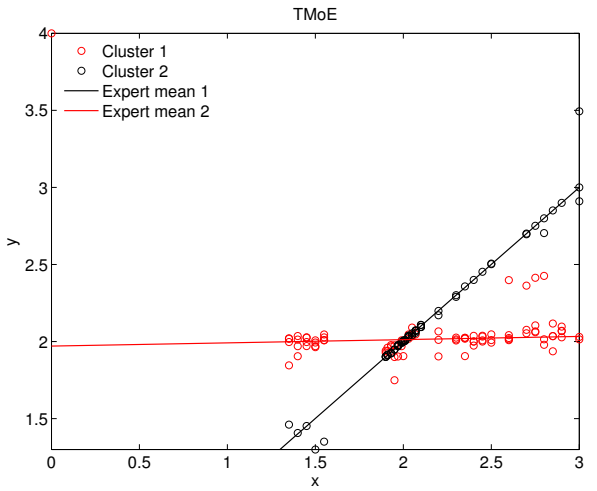


Tone data with 10 outliers $(0, 4)$: Normal fit





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

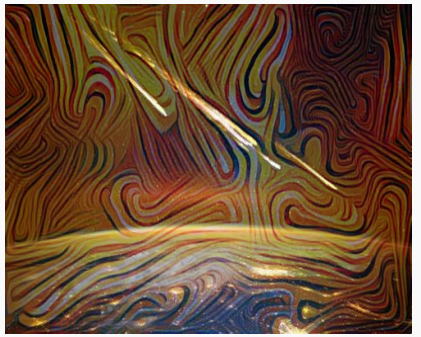


Tone data with 10 outliers $(0, 4)$: **Robust** fit

Open-Source Toolkits

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

+14K téléchargements (depuis janvier 2020) canal R uniquement



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

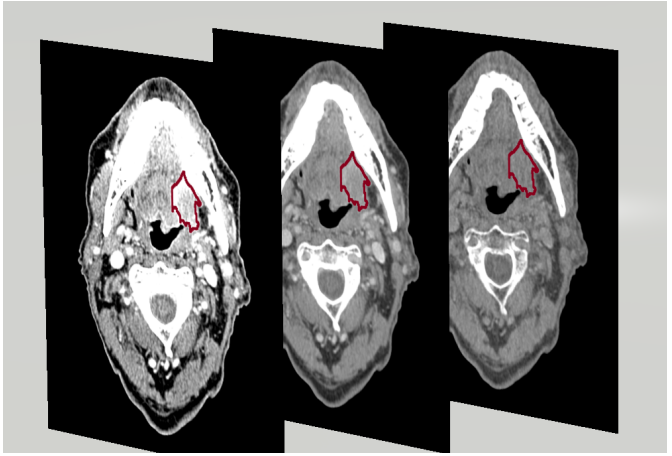
▸ Github ▸ CRAN ▸ Matlab software

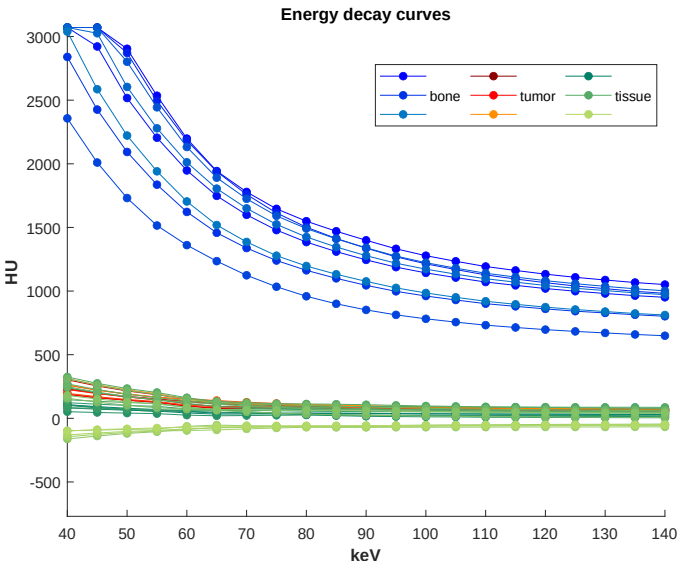
Available algorithms and Packages

`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

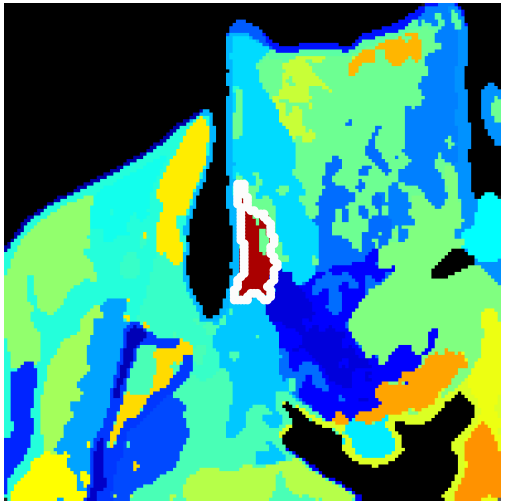
- Meteorits include sampling, fitting, prediction, clustering with each MoE model
- Non-normal mixtures (and MoE) is a very recent topic in the field

Application of ML in precision medicine (Radiology)





Expert Annotation

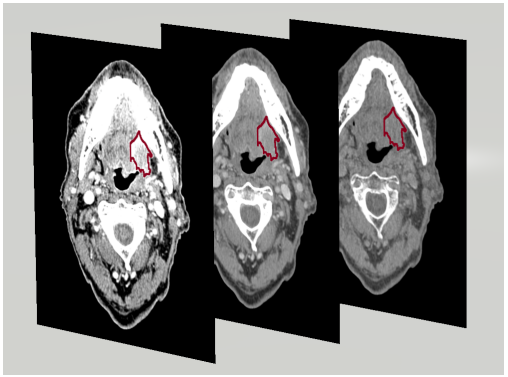


Automatic Annotation

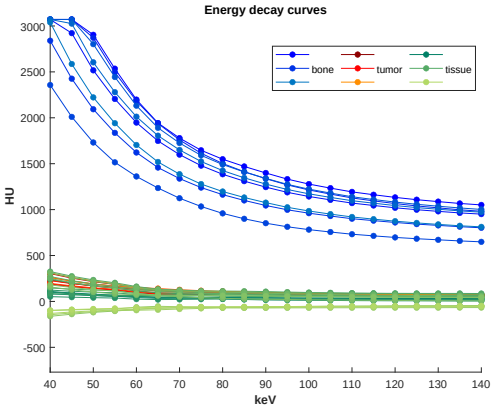
Spatial mixture of functional regressions {Diagnostics, 2022}

Spatial mixture of functional regressions for dual-energy CT images

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

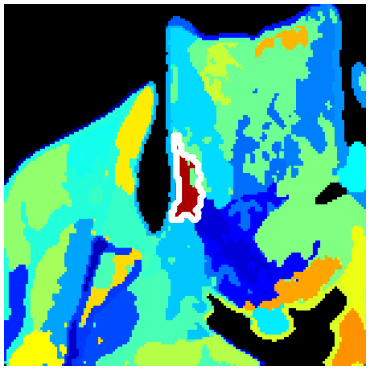


2D slices of VMIs at 40,65,140keV with tumor contour in red.



Examples of decay curves for different body locations.

Original slice



SgMFR Clustering. Dice score=0.84

Clustering with SgMFR. Note the robustness of the result in the presence of a metallic artifact in the RHS of the anatomical image.

[J] DECT-CLUST : Dual-Energy CT Image Clustering and Application to Head and Neck Squamous Cell Carcinoma Segmentation. Chamroukhi, Brivet, Savadjiev, Coates, Forghani. Diagnostics, 2022

— **[C]** CMStatistics. 2021

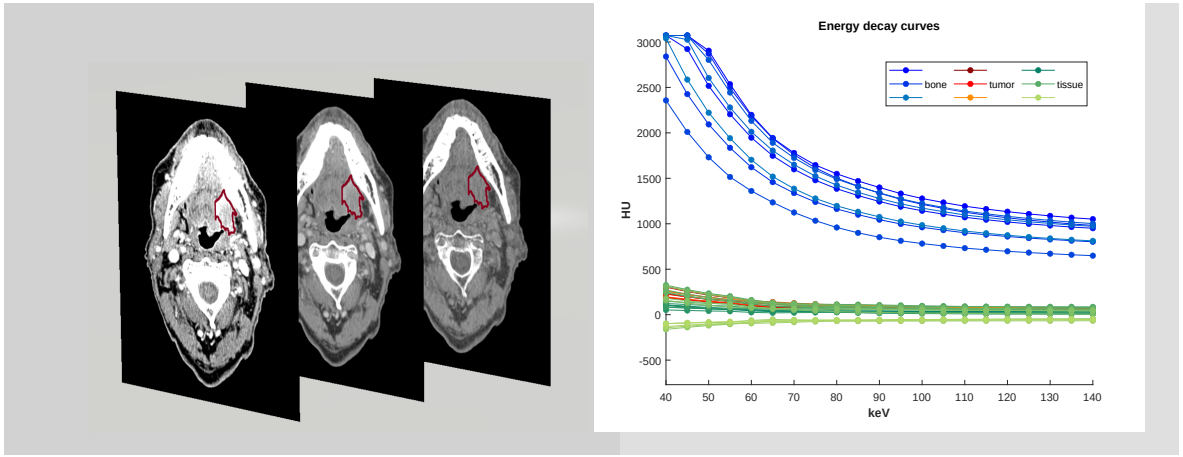
▸ **[S]** Source codes are publicly available on Github : `https://github.com/fchamroukhi/DECT-CLUST`

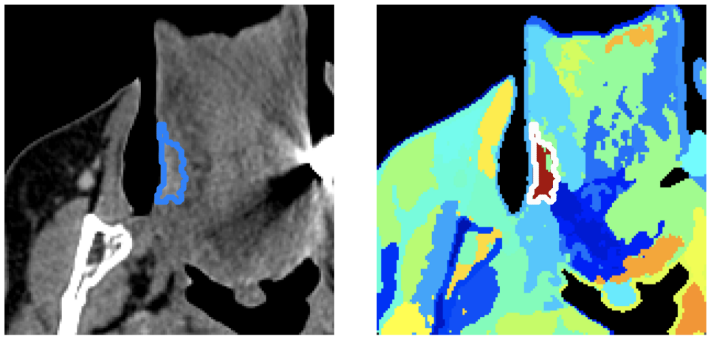
DECT image Clustering (Healthcare)

- Learning from Multimodal information in precision medicine
- HNSCC Cancer detection in Radiology : DECT clustering

{Collab. with McGill & Florida, College of Medicine} [Diagnostics (AI in medicine), 2022]

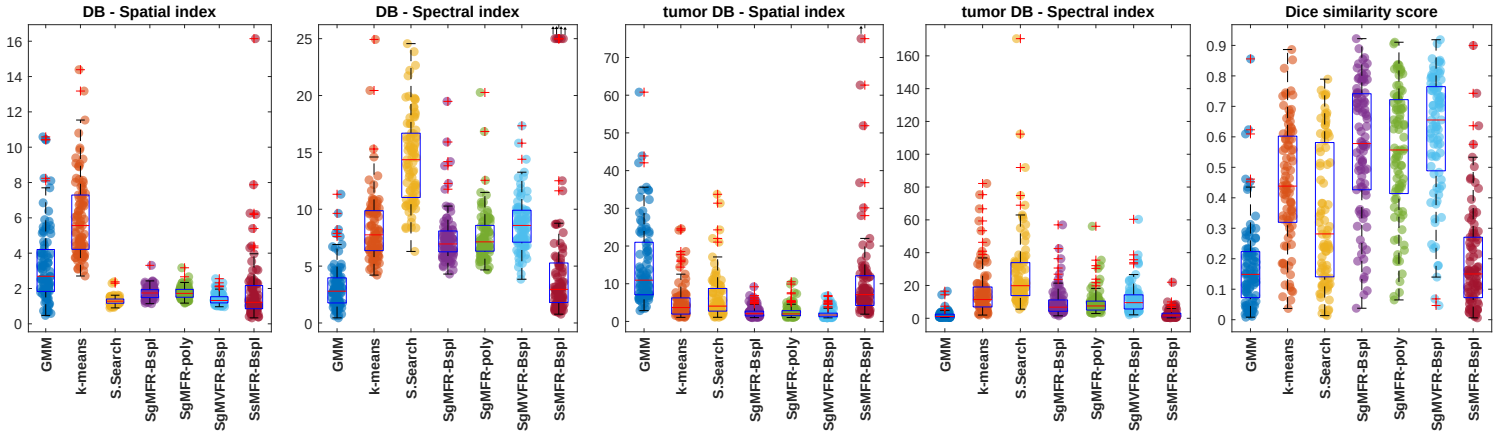
Emerging medical imaging system : need for models to analyse these data :





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

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



[A.] Inférence en grande dimension

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 : $\boldsymbol{X}_i(t)$, $t \in \mathcal{T} \subseteq \mathbb{R}$ {eg. continuous recorded variables}

$\hookrightarrow$ Méthodes d'**Inférence** et **Sélection parcimonieuses**, Soucis d'**interprétabilité**

[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 : $\mathsf{Pen}_{\lambda}(\boldsymbol{\theta}) = \underbrace{\sum_{k=1}^{K} \lambda_k \|\boldsymbol{\beta}_k\|_1}_{\text{Experts Net.}} + \underbrace{\sum_{k=1}^{K-1} \gamma_k \|\boldsymbol{w}_k\|_1}_{\text{Gating Net.}}$

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

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

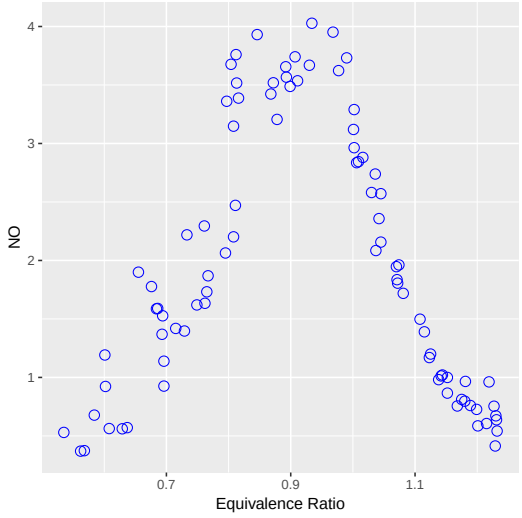
- High-Dimensional Clustering and Regression (with Gaussian and Poisson outputs)
- High-Dimensional Classification (Categorical outputs)
- EM-Lasso algorithms with proximal Newton and Coordinate Ascent for optimization

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

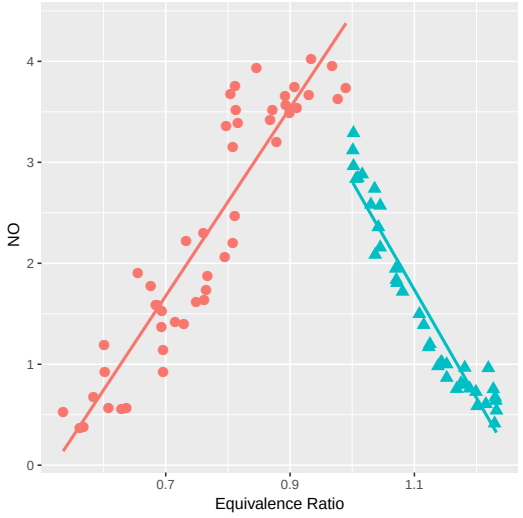
[Thèse] Bao Tuyen Huynh, *Estimation and Feature Selection in High-Dimensional Mixtures-of-Experts Modesls* . Thèse de

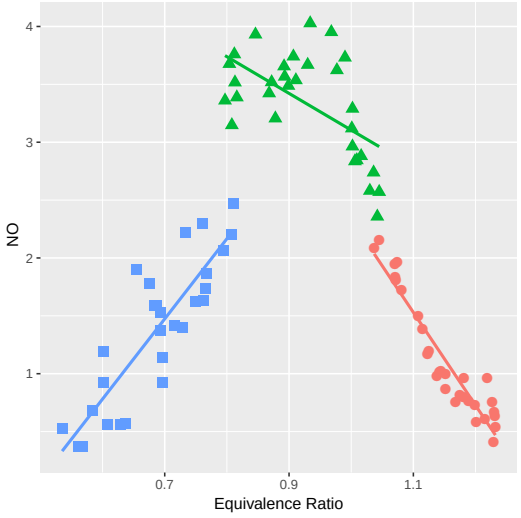
Model selection

(a) Raw Ethanol data set

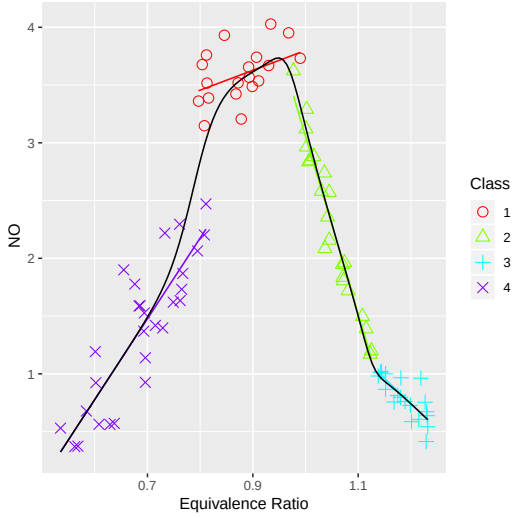


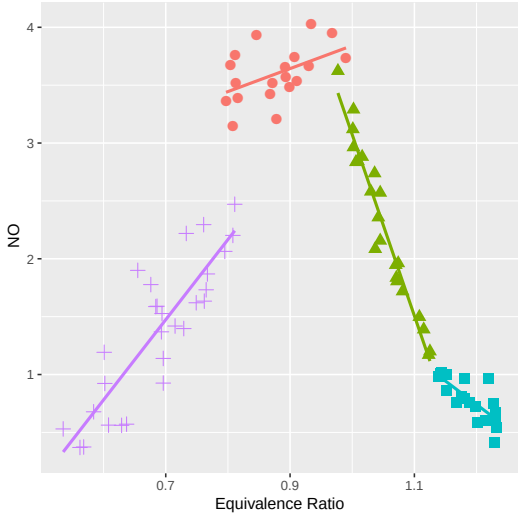
Collection of MoE models with linear mean functions characterized by 2-5 clusters

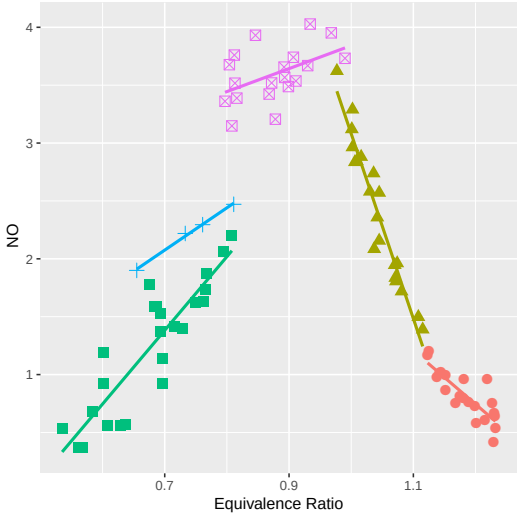




(b) Our best data-driven MoE model







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 $p \gg n$

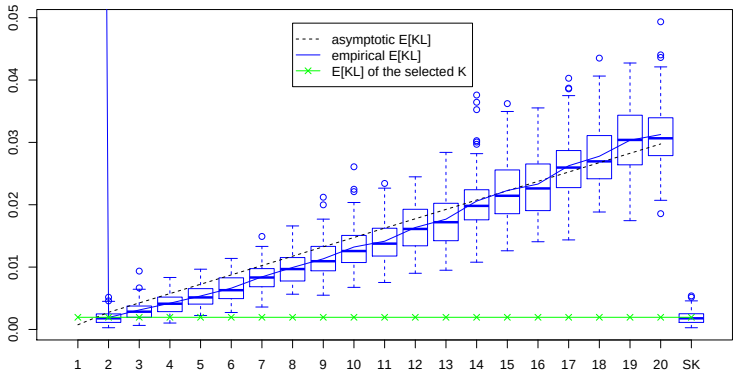
HDME : High-Dimensional MoE : PMLE $\widehat{\boldsymbol{\theta}}_{\mathrm{n}} \in \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^n \log h_K^{\varphi}(\boldsymbol{y}_i|\boldsymbol{x}_i;\boldsymbol{\theta}) - \mathrm{pen}(\boldsymbol{\theta})$

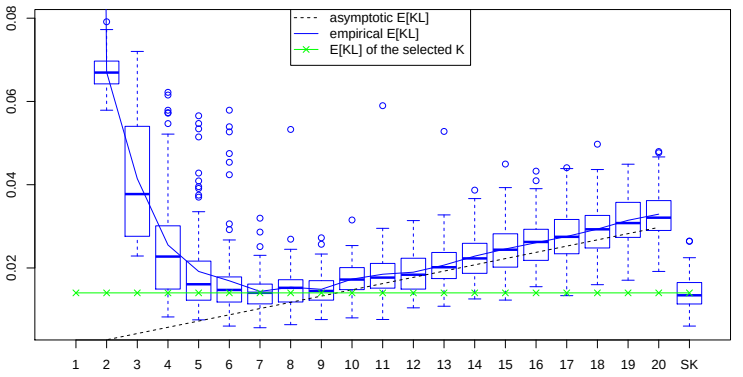
Theorem : Non-asymptotic oracle inequality for collection of MoE models

Résultat : $\exists$ des constantes C et $\kappa(\rho, C_1) > 0$ ($C_1 > 1$) t.q chaque fois que pour $\mathbf{m} \in \mathcal{M}$, $\mathrm{pen}(\mathbf{m}) \geq \kappa(\rho, C_1)\left[(C + \ln n)\dim(\mathcal{H}_{\mathbf{m}}) + z_{\mathbf{m}}\right]$, l'estimateur PMLE $\widehat{h}_{\widehat{\mathbf{m}}}$ satisfait

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

- Résultat non-asymptotique. Si $\mathrm{pen}(\mathbf{m})$ est bien choisie, alors notre PMLE se comporte de manière comparable au **meilleur modèle (oracle)** $h_{\mathbf{m}^\star}$ de la collection, minimisant le risque : $\inf_{\mathbf{m}\in\mathcal{M}}\left(\inf_{h_{\mathbf{m}}\in\mathcal{H}_{\mathbf{m}}} \mathrm{KL}^{\otimes \mathrm{n}}(f, h_{\mathbf{m}}) + \frac{\mathrm{pen}(\mathbf{m})}{n}\right)$ (f est inconnue).





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

Functional Data Analysis (Open-Source Toolkits)

FLaMingoS : open source software for learning from functions

+15K téléchargements R depuis août 2019

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

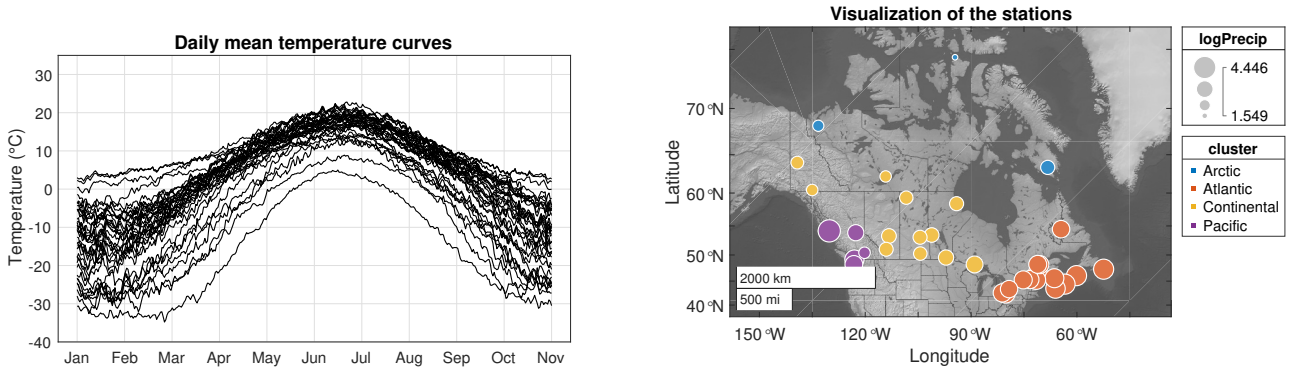


FIGURE – $n = 35$ daily mean temperature measurement curves (X_i's) in different stations (Left) and the $\log$ of precipitation values (Y_i's) visualized with the climate regions (Z_i's) (Right).

- 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)

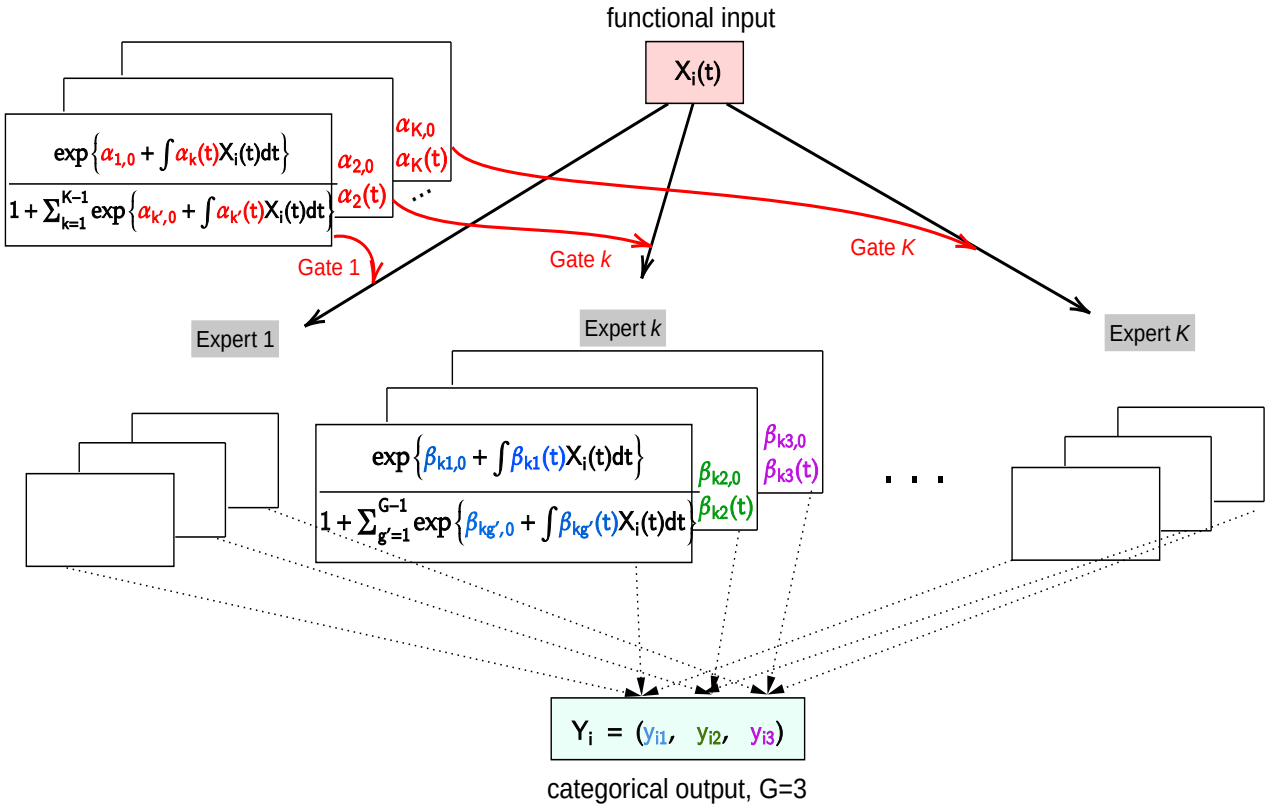
- $Y_i = \beta_{z_i,0} + \int_{\mathcal{T}} X_i(t)\beta_{z_i}(t)dt + \varepsilon_i$ avec $h_z(X_i(.)) = \alpha_{z_i,0} + \int_{\mathcal{T}} X_i(t)\alpha_{z_i}(t)dt$
- Lasso-type Regularized MLE w.r.t the derivatives of the $\alpha(\cdot)$ and $\beta(\cdot)$ functions

[Conf] Chamroukhi, Pham, Hoang, McLachlan. Mixtures-of-experts with functional predictors. CMStatistics 2021

[Preprint] —. 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 <u>derivatives</u> of the $\alpha(\cdot)$ and $\beta(\cdot)$ functions

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

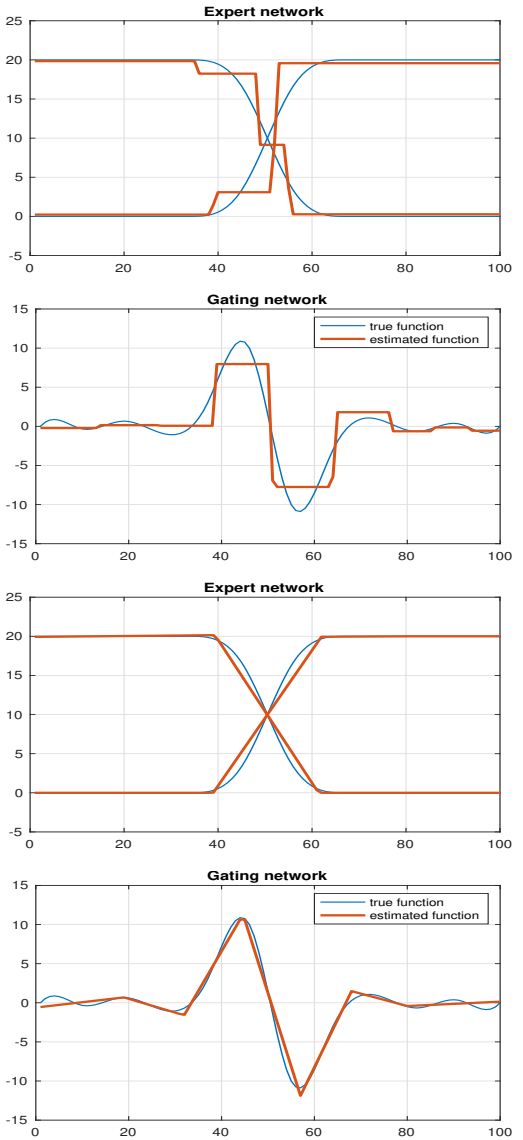
$\beta^{(0)}_{z_i}(t) = 0$ implies that $X(t)$ has no effect on Y at t

$\beta^{(1)}_{z_i}(t) = 0$ 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.

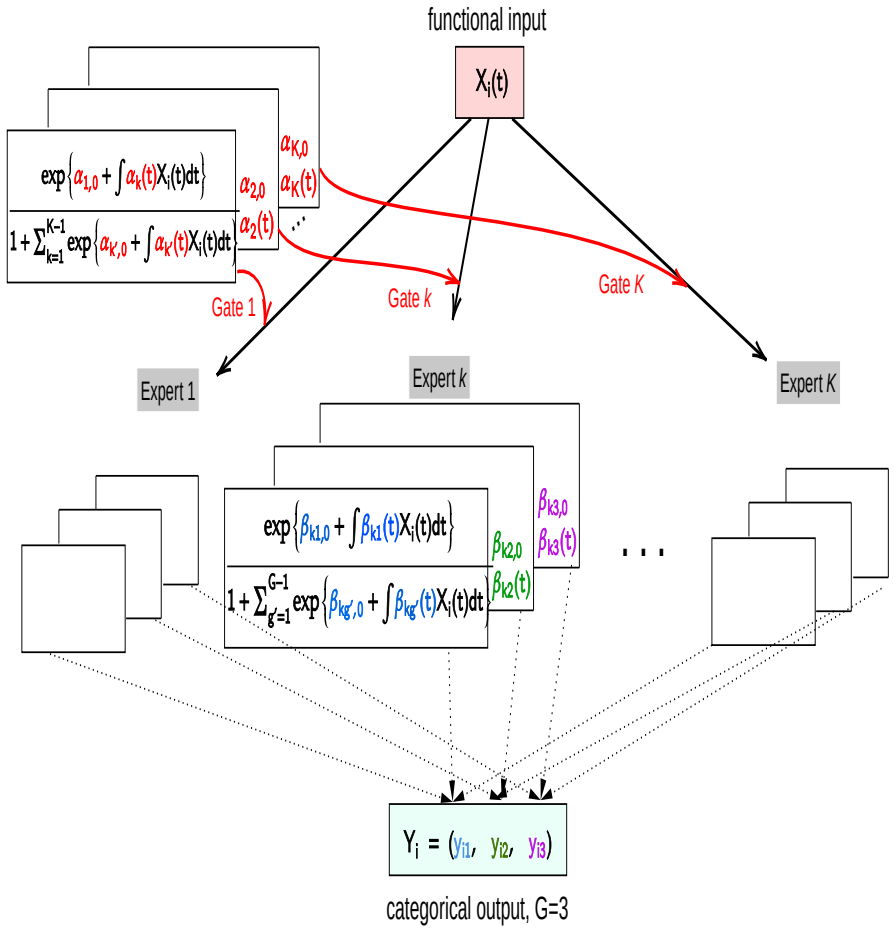
[PhD TN. Pham, 2022] [arXiv :2202.13934] [Submitted, 2023] [**Contract**, ANR SMILES]

Interpretable fits

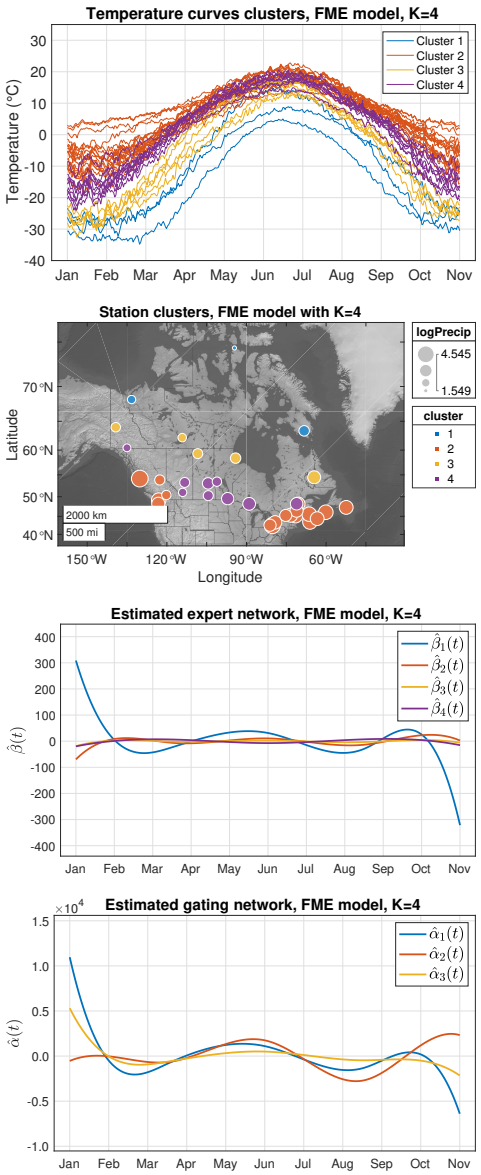


Interpretable learning with time-series inputs

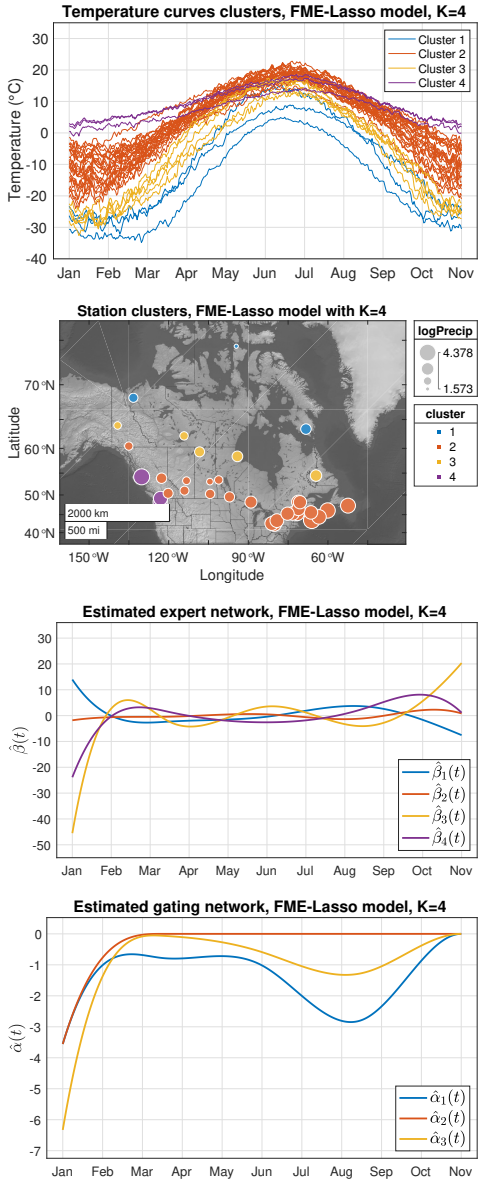
Mixture-of-Experts Architecture



No regularization



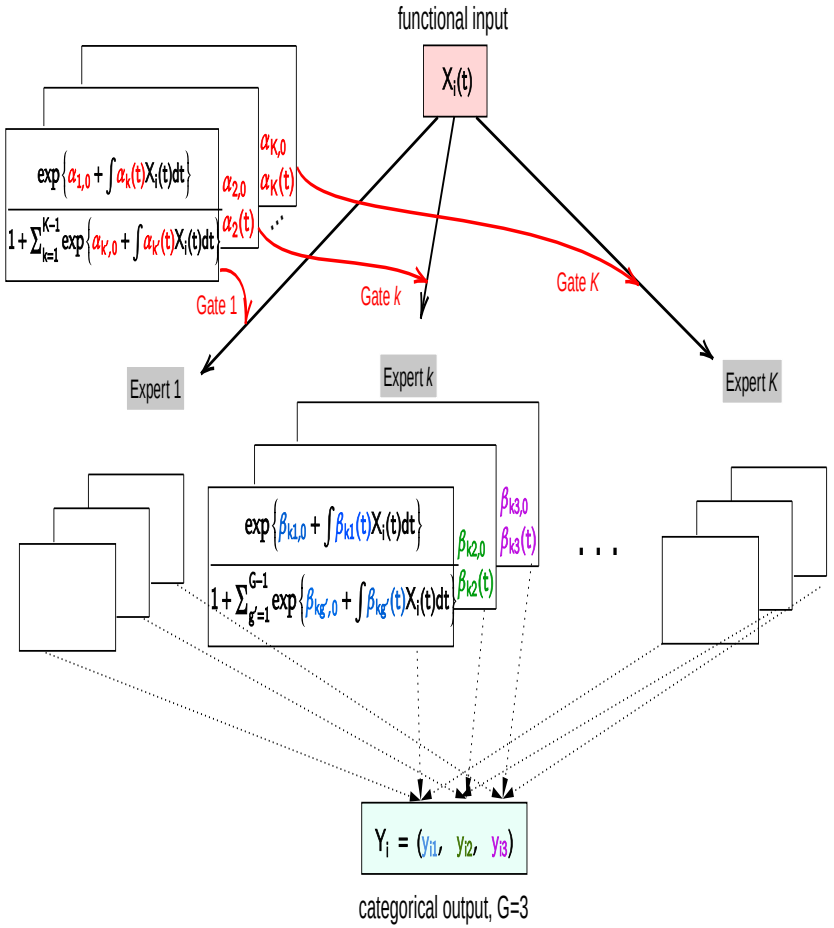
LASSO regularization



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

Interpretable learning with time-series inputs

Mixture-of-Experts Architecture



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

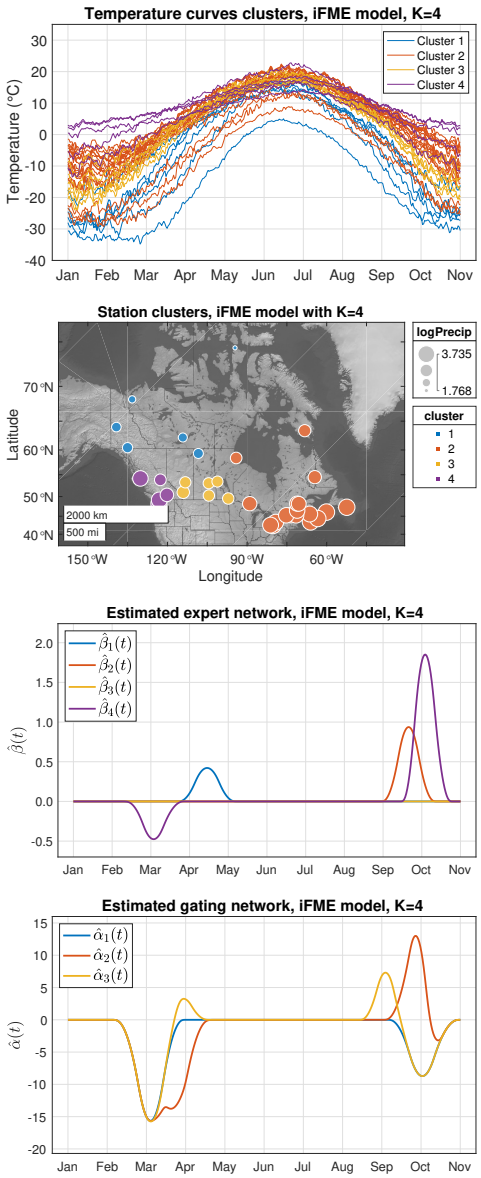
$\beta^{(0)}_{z_i}(t) = 0$ implies that $X(t)$ has no effect on Y at t

$\beta^{(1)}_{z_i}(t) = 0$ 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.

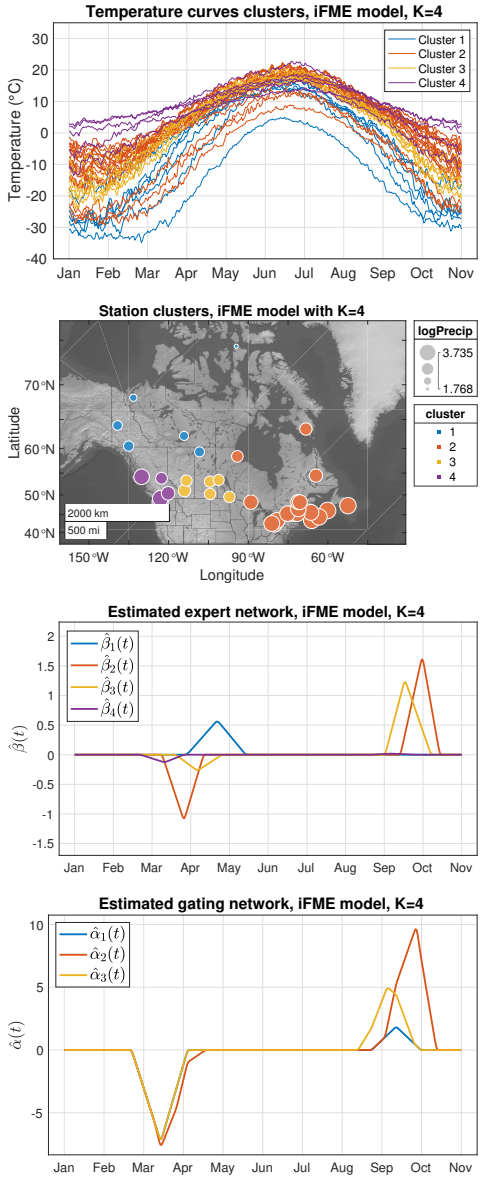
[PhD TN. Pham, 2022]

OUR regularization



[arXiv :2202.13934]

OUR regularization



[Submitted, 2023] [**Contract**, ANR SMILES]

Interpretable learning with functional inputs

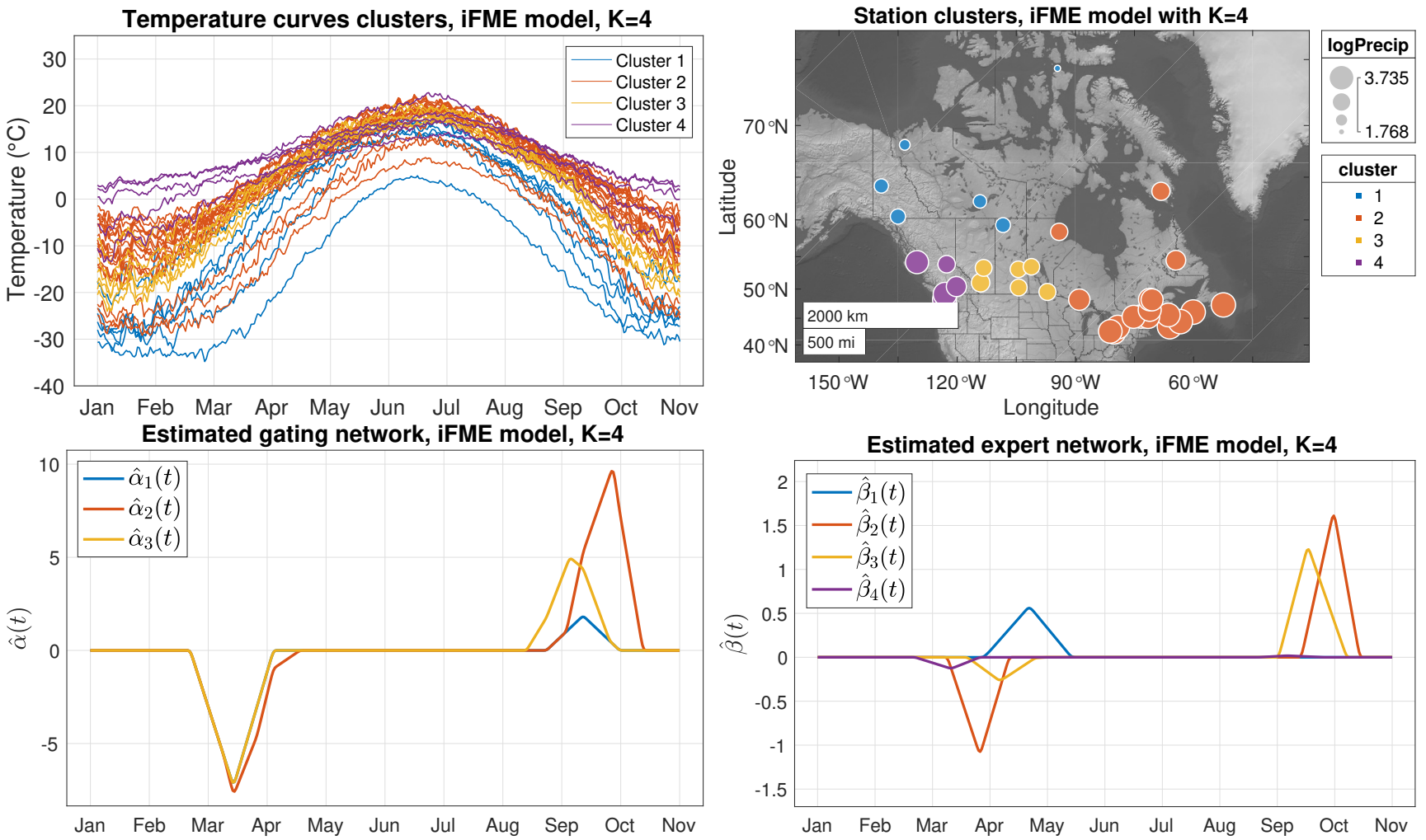


FIGURE – Clustering des températures (g.) et (log)-précipitations prédites (d.) (Atlantique, Pacifique, Continentale et Arctique), et fonctions paramètres (bg - bd.)

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

[B.] Données de gros volume : **Clustering** for an informative summary of the data

- la distribution des calculs est une façon naturelle de s'y prendre
- Stratégie : inférence (et échantillonnage) pour l'**agrégation de modèles DLVM**

↪ New statistical issues in estimation and model selection and computational issues

Statistical guidelines

- collaborative mixtures for large-scale model-based clustering
- aggregate local estimators to provide an overall proven aggregated estimator

↪ ↪ e.g minimize the KL divergence between mixtures

Computational guidelines

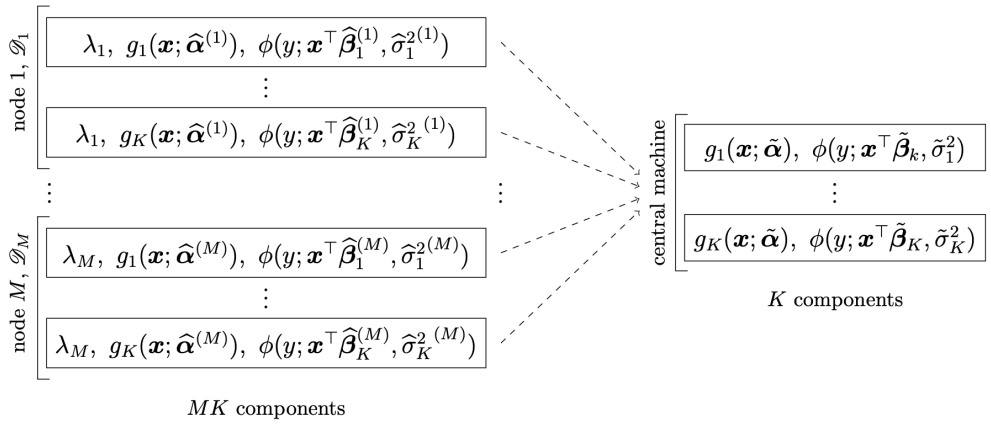
- Distributed and Parallel processing is a natural way to proceed
- ↪ key question : how to distribute data while controlling the quality of estimators
- ↪ ensemble methods (BLB effective in scaled supervised learning [a]

a. Kleiner et al. "A scalable bootstrap for massive data." *JRSS B(2014) 76 :795-816*

[Rep.] Chamroukhi & Pham **to be submitted 2023**, *Distributed Learning of Mixtures-of-Experts*
[Thèse] Pham. *Modeling and Learning with Mixtures of Experts for Functional Data and Distributed Data*. Thèse de Normandie Université, Nov. -2022.
[Contrat] ANR SMILES

[A.] Données de gros volume : Aggregating distributed mixtures-of-experts models

- **Clustering** for an informative summary of the data and **MoE** for better prediction
- collaborative mixtures-of-experts for large-scale data



- Local estimators : $\hat{f}_m = f(\cdot|\mathbf{x},\widehat{\boldsymbol{\theta}}_m) = \sum_{k=1}^K g_k(\mathbf{x},\widehat{\boldsymbol{\alpha}}^{(m)})\phi(\cdot;\mathbf{x}^\top\widehat{\boldsymbol{\beta}}_k^{(m)},\widehat{\sigma}_k^{2(m)})$,
- weighted average : $\bar{f} = f(y|\mathbf{x};\bar{\boldsymbol{\theta}}) = \sum_{m=1}^M \lambda_m \hat{f}_m$ where $\lambda_m = \frac{N_m}{N}$ the sample proportion. $\bar{f}$ is good but relates MK components so not our direct target.

$\hookrightarrow$ Reduced estimator : $\bar{f}^R = \underset{h_K\in\mathcal{M}_K}{\arg\inf}\ \rho\left(h_K, \sum_{m=1}^M \lambda_m \hat{f}_m\right)$: we seek for a K-component ME h that is closest to the MK-component ME $\bar{f} = \sum_{m=1}^M \lambda_m \hat{f}_m$ w.r.t a transportation divergence $\rho(\cdot,\cdot)$, e.g. KL.

{PhD, Pham. *Mixtures of Experts for Distributed Data, 2022*} [to be submitted 2023] [**Contrat**, ANR SMILES]

Federated Learning

Numerical results in Distributed clustering and Prediction

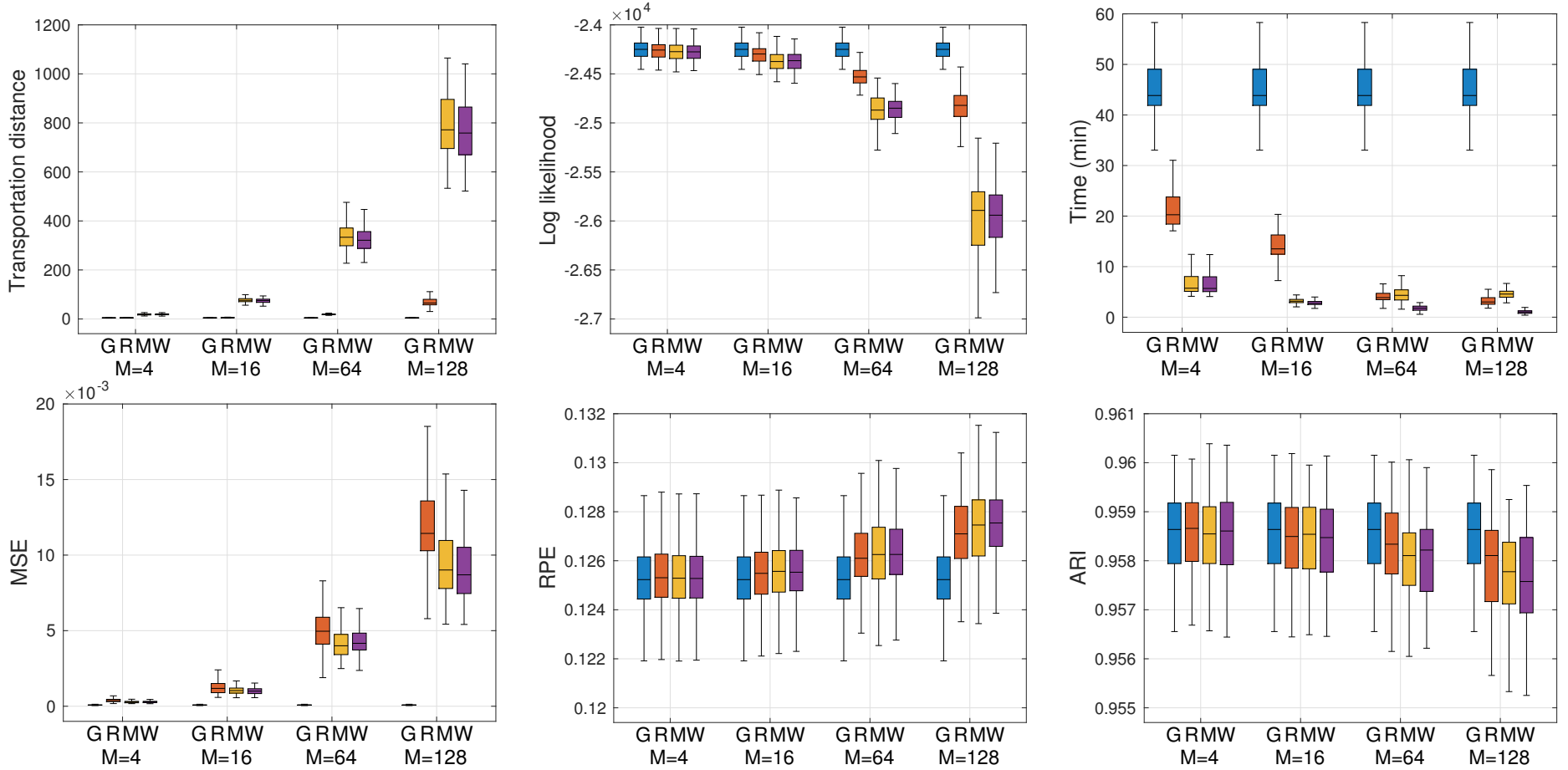


FIGURE – Performance of the Global ME (G), Reduction (R), Middle (M) and Weighted average (W) estimator for sample size $N = 10^6$ and M machines.

{PhD, Pham. *Mixtures of Experts for Distributed Data, 2022*} [to be submitted 2023] [**Contrat**, ANR SMILES]

Cadre scientifique général

↪ **Modèles** à variables latente : $f(x|\boldsymbol{\theta}) = \int_{\mathcal{Z}} f(x, z|\boldsymbol{\theta})\mathrm{d}z$

↪ **Inférence**, **Sélection** et **représentation non supervisées** et à l'échelle

Modélisation non supervisée à l'échelle par des MVL

- Apprentissage génératif de modèles à variables latentes ((non)-supervisé)

✔ Excellentes capacités de représentation

✔ Représenter explicitement la structure latente des données brutes et la révéler

✔ Cadre de choix en apprentissage non-supervisé (Clustering, Représentation)

↪ $\exists$ fondement théorique solide

↪ Outils afférents d'inférence et de choix de modèle

✖ Défis pour des traitements et analyses en grande dimension et en masse

Thank you for your attention !