Apprentissage statistique en vue du passage à l'echelle: représentation, inférence et sélection

Faïcel Chamroukhi

Séminaire, le 09 mai 2022

Recherche et contributions

- **Contexte :** Données complexes $\hookrightarrow$ *hétérogènes, temporelles dynamiques, fonctionnelles, peu ou non-annotées , de grande dimension, et disponibles en masse*
- **Objectif :** Transformation de telles données en connaissances inteprétables :
  $\hookrightarrow$ Reconstruction/révélation de structures cachées : groupes, hiérarchie de groupes, prototypes résumant bien les données, sélection de modèles, parcimonie
- **Actions :**
  - $\hookrightarrow$ Un cadre scientifique et technique pour les traiter, analyser et exploiter
  - $\hookrightarrow$ Avec une visibilité à l'international
  - $\hookrightarrow$ Applications et dans des activités contractuelles

Axe de recherche : Apprentissage statistique et analyse de données complexes

I. Apprentissage à partir de données hétérogènes, de grande dimension et massives ;

II. Apprentissage en présence de variables et co-variables fonctionnelles hétérogènes ;

III. Apprentissage non-supervisé de représentations.

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

Modèles de mélanges de lois

- **Data** : n observations $\{\boldsymbol{x}_i\}$ d'une v.a $\boldsymbol{X} \in \mathbb{X} \subset \mathbb{R}^d$ à population potentiellement **hétérogène** de densité inconnue $f \in \mathcal{F} = \left\{ f : \mathbb{X} \to \mathbb{R}_+ | \int_{\mathbb{X}} f(\boldsymbol{x}) \mathrm{d}\lambda(\boldsymbol{x}) = 1 \right\}$,
- **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 : Approximation universelle des mélanges finis de translatés dilatées

(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-] T Nguyen, F Chamroukhi, H Nguyen, and G McLachlan. **Approximation of probability density functions via location-scale finite mixtures in lebesgue spaces**. *Communications in Statistics - Theory and Methods*, 2022. doi :10.1080/03610926.2021.2002360.

[Thèse] T Nguyen. *Model Selection and Approximation in High-dimensional Mixtures of Experts Models : From Theory to Practice* . Thèse de Doctorat de Normandie Université, 2021. Directeur.

Apprentissage/Optimisation

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

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

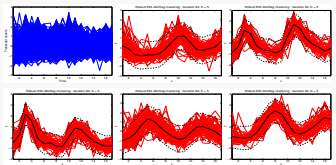
- Régularisation en apprentissage non-supervisé

↪ 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

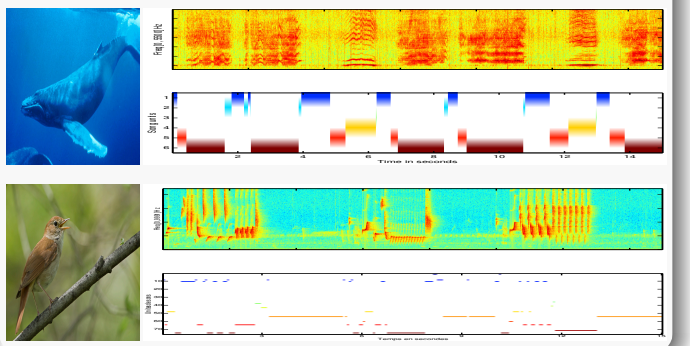
- Classification de séquences bio-acoustiques
- *Mélanges parcimonieux de Processus de Dirichlet*

↪ Cadre non-paramétrique : inférence et sélection

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

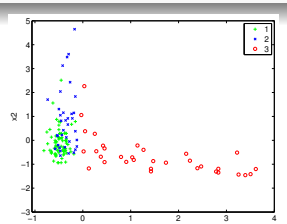
[Thèse] Marius BARTCUS, 2015, *UTLN*, Co-encadrant. 70%

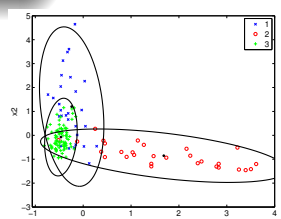
Biologie - Découverte d'unités de chant

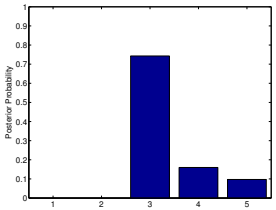


Diabetes data set

Clustering et choix de modèle

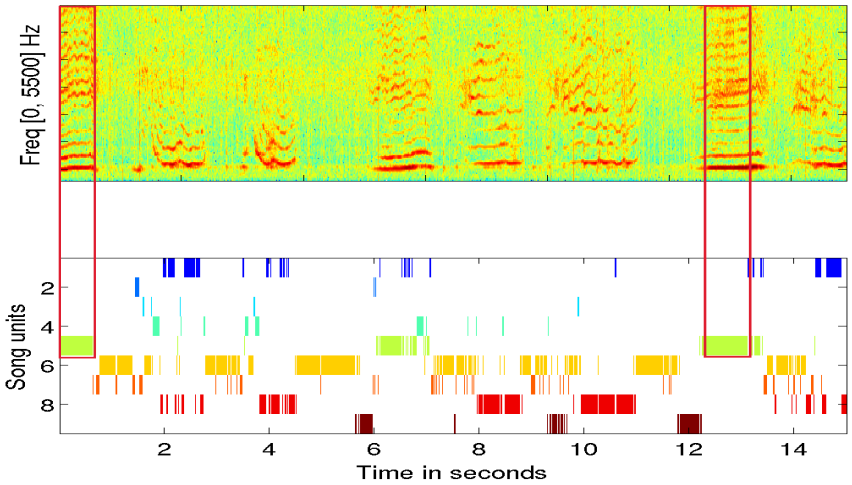


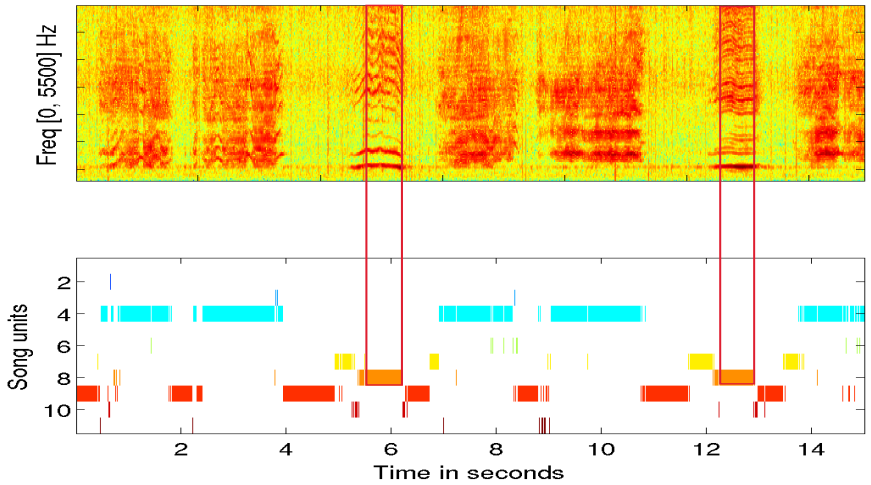




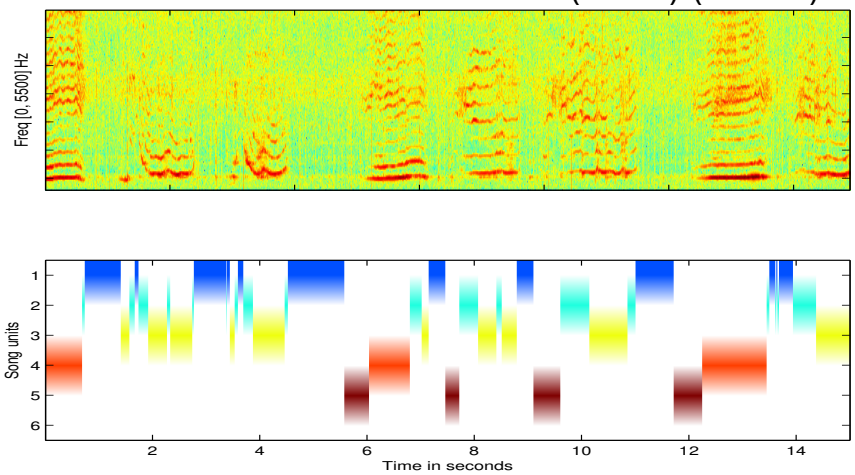
2 log BF : $\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T$ *vs* $\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T$ = 199.58 (Decisive)

Unsupervised decomposition of whale song signals

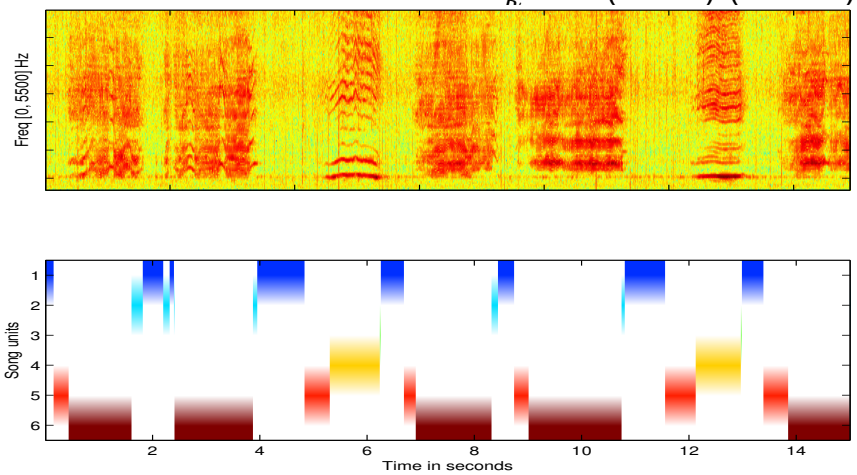




Sound demo of Unit 5 DPPM $\lambda \mathbf{I}$: (sec. 0) (sec. 12)



Sound demo of Unit 8 DPPM $\lambda_k \mathbf{A}$: (sec. 6) (sec. 12)



[-] Roger, Bartcus, Chamroukhi, Glotin. Unsupervised Bioacoustic Segmentation by Hierarchical Dirichlet Process Hidden Markov Model. Multimedia Tools and Applications for Environmental & Biodiversity Informatics, pp :113-130, 2018

[-] Bartcus, C., Glotin. Hierarchical Dirichlet Process HMM for Unsupervised Bioacoustic Analysis. IJCNN15. & wkp ICML14.

[-] Chamroukhi, Bartcus, Glotin. Bayesian non-parametric parsimonious Gaussian mixture for clustering. ICPR 2014.

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 $f \in \mathcal{F} = \left\{ f : \mathbb{X} \times \mathbb{Y} \to \mathbb{R}_+ | \int_{\mathbb{Y}} f(\boldsymbol{y}|\boldsymbol{x}) \, \mathrm{d}\lambda(\boldsymbol{y}) = 1, \forall \boldsymbol{x} \in \mathbb{X} \right\}$
- 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 :

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

Théorème : Capacités d'approximation des mélanges d'experts SGaME

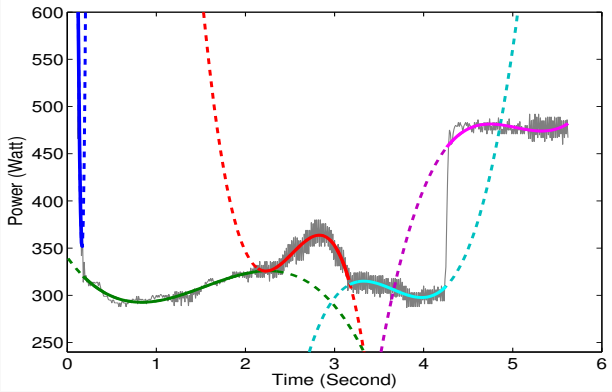
(a) Pour $p \in [1, \infty)$, $f \in \mathcal{F}_p \cap \mathcal{C}$, $\varphi \in \mathcal{F} \cap \mathcal{C}$, $\mathbb{X} = [0,1]^d$, il existe une suite $\left(h_K^\varphi\right) \subset \mathcal{H}_S^\varphi$ telle que $\lim_{K \to \infty} \left\| f - h_K^\varphi \right\|_{\mathcal{L}_p} = 0$.

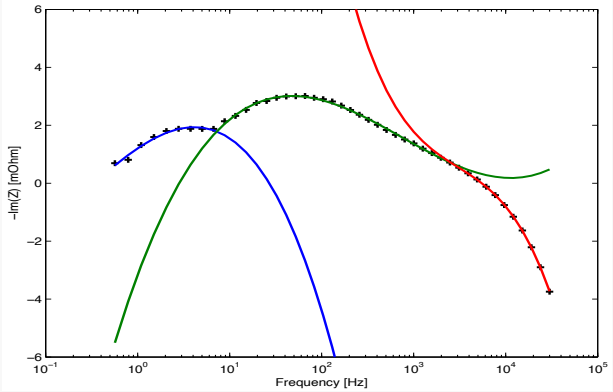
(b) Pour toute $f \in \mathcal{F} \cap \mathcal{C}$, si $\varphi \in \mathcal{F} \cap \mathcal{C}$, $d = 1$, il existe une suite $\left(h_K^\varphi\right) \subset \mathcal{H}_S^\varphi$ telle que $\lim_{K \to \infty} h_K^\varphi = f$ presque uniformément.

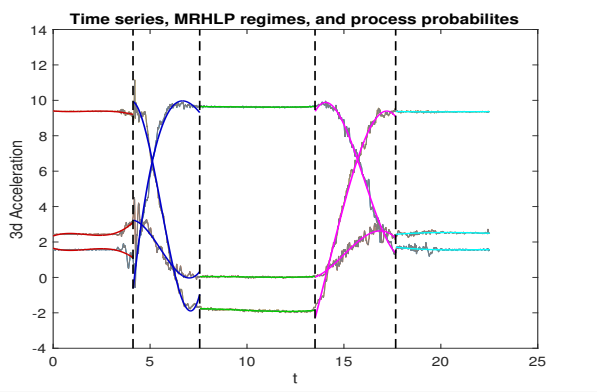
[J-] Nguyen, Chamroukhi, Forbes. *Neurocomputing*, 366, pp 208–214, 2019

[J-] NguyenH, NguyenT, Chamroukhi, McLachlan. Approximations of conditional probability density functions in Lebesgue spaces via mixture of experts models. *Journal of Statistical Distributions and Applications*, Springer, 8(1), pp :1–15, 2021

Time Series Modeling and Segmentation







Temporal data with unknown abrupt and/or smooth regime changes

Hidden Process Regression Models

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

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

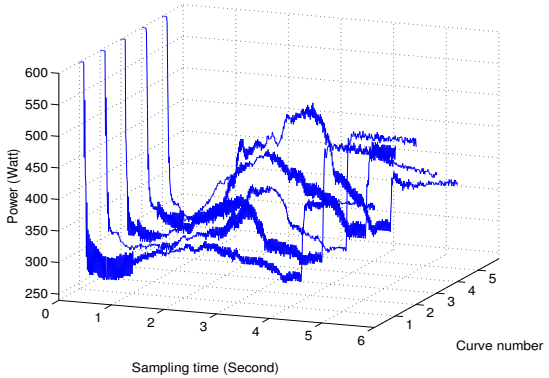
Optimization for Learning : $\widehat{\boldsymbol{\theta}}_{\text{MLE}} \in \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log m(y_i|x_i;\boldsymbol{\theta})$

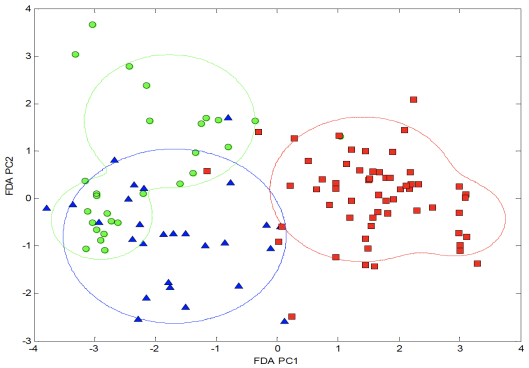
- MLE via the EM algorithm : $\boldsymbol{\theta}^{(q+1)} \in \arg\max_{\boldsymbol{\theta}} \mathbb{E}\left[L_c(\boldsymbol{\theta};\mathbf{x},\mathbf{y},\mathbf{z})|\mathbf{x},\mathbf{y};\boldsymbol{\theta}^{(q)}\right]$

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

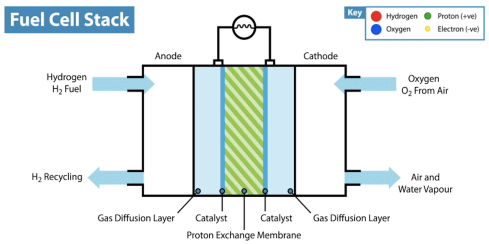
Time-Series Analysis Applications

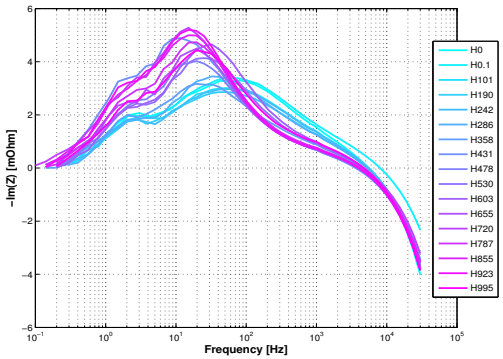
Transport : Railway switch operating state prediction {Collab. avec la SNCF ; Projet Switch-Rdf}

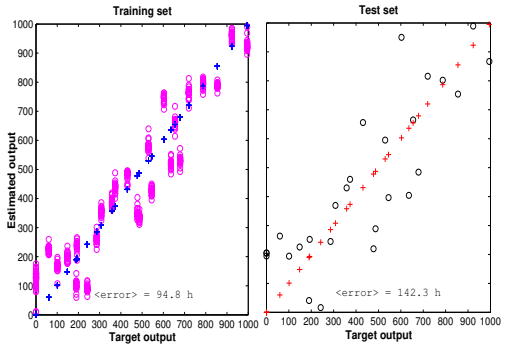




Energy : Fuel cell lifetime prediction {Collab. avec Femto-ST, Phd de R. Onanena, 2012}

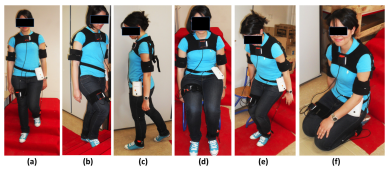


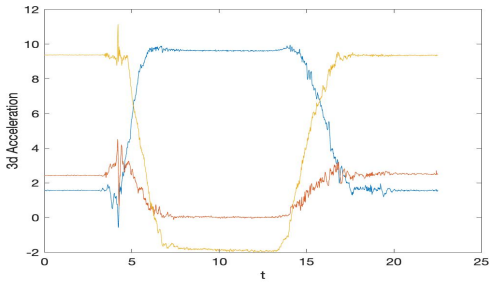


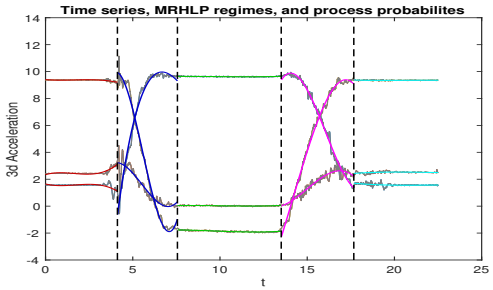


Health & well being : Human activity recognition {Collab. avec Paris 12-LiSSi}

{**[Thèse]** Thèse de Dorra TRABELSI, 2013, *Paris 12*, Co-encadrnt}







SaMUraiS : open source software for statistical time-series analysis

+11K téléchargements (à partir du canal R uniquement) **depuis juillet 2019**

`https://cranlogs.r-pkg.org/badges/grand-total/samurais`

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

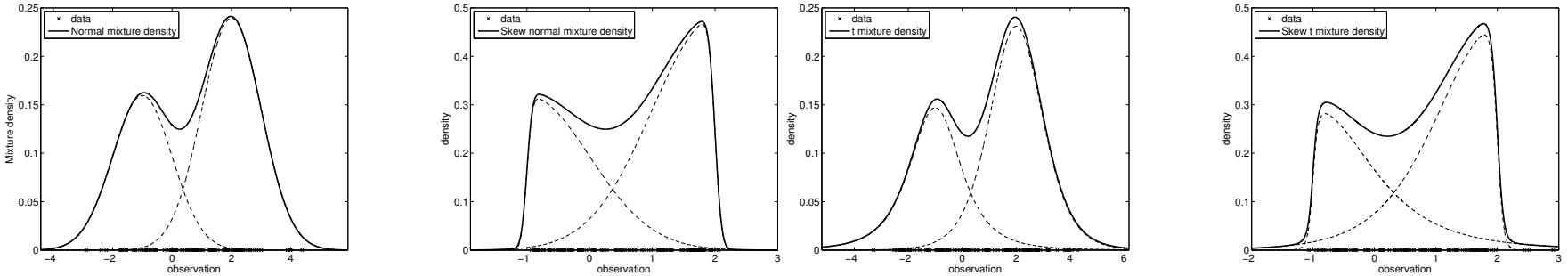
Include estimation, segmentation, approximation, model selection, and sampling

Robust learning with mixtures-of-experts models

- **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

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

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



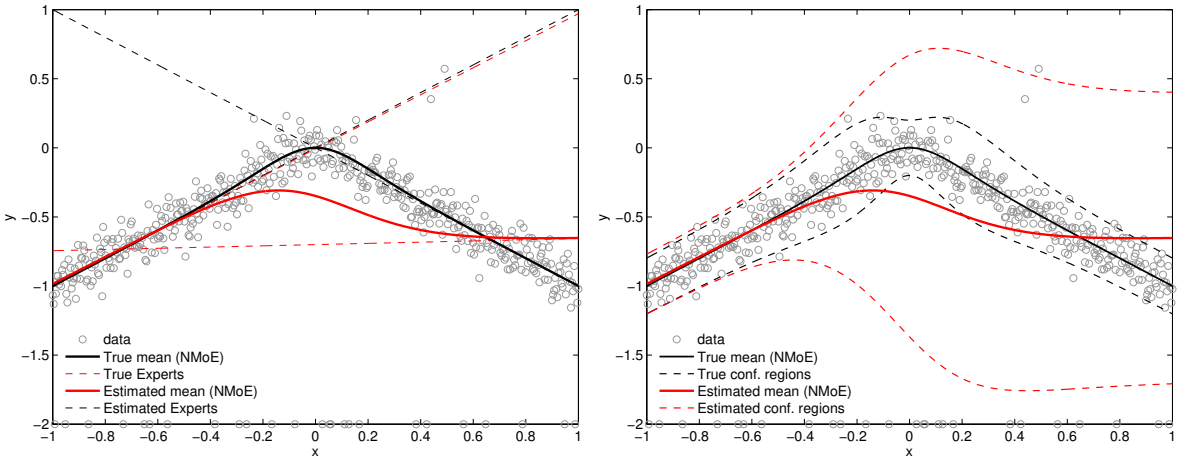
$\pi_k = [0.4, 0.6]$, $\mu_k = [-1, 2]$; $\sigma_k = [1, 1]$; $\nu_k = [3, 7]$; $\lambda_k = [14, -12]$;

Flexible and robust generalization of the standard MoE models

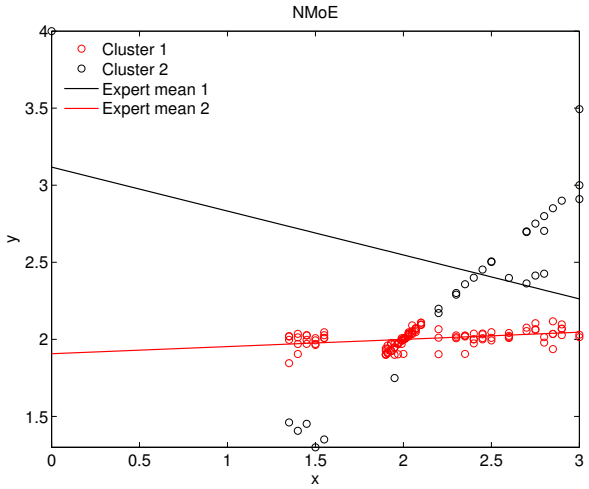
For $\{\nu_k\} \to \infty$, STMoE reduces to SNMoE ; For $\{\lambda_k\} \to 0$, STMoE reduces to TMoE.
For $\{\nu_k\} \to \infty$ and $\{\lambda_k\} \to 0$, StMoE approaches the NMoE.

[J-] F. Chamroukhi. Skew t mixture of experts. *Neurocomputing*, 266 :390–408, 2017.
[J-] F. Chamroukhi. Robust mixture of experts modeling using the t-distribution. *Neural Networks*, 79 :20–36, 2016c.
[C-] F. Chamroukhi. Skew-normal mixture of experts. IJCNN 2016b.

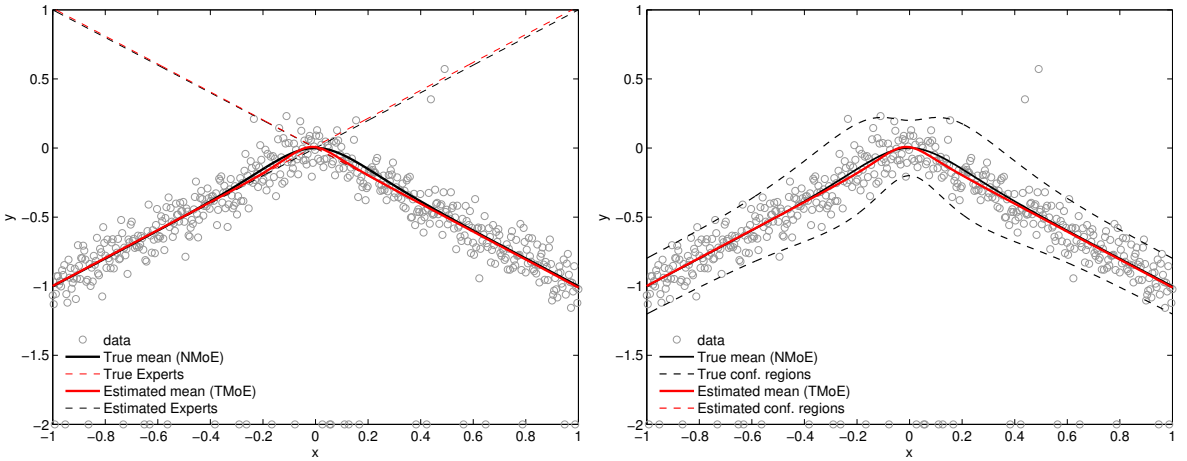
Robustness in regression and clustering



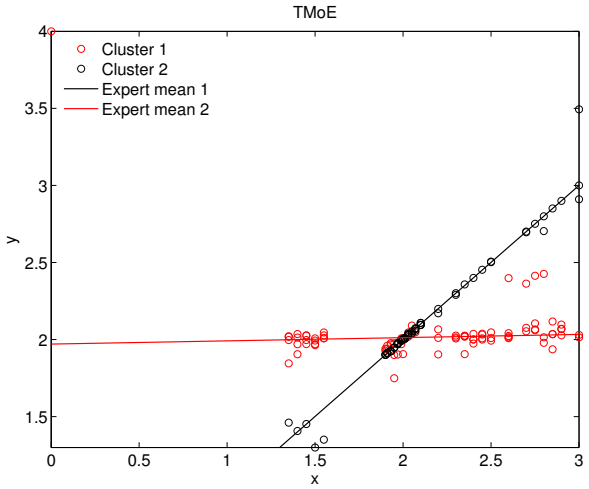
$n = 500$ observations with 5% of outliers $(x; y = -2)$: Normal fit



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



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

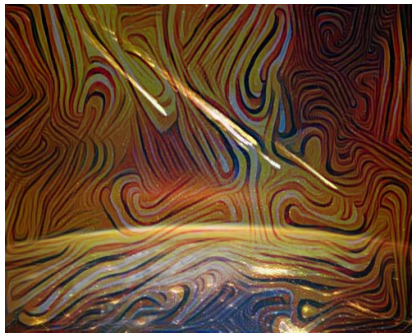


Tone data with 10 outliers $(0, 4)$: **Robust** fit

MEteorits : open-source soft. for mixtures-of-experts

+12K téléchargements R depuis janvier 2020

`https://cranlogs.r-pkg.org/badges/grand-total/meteorits`



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

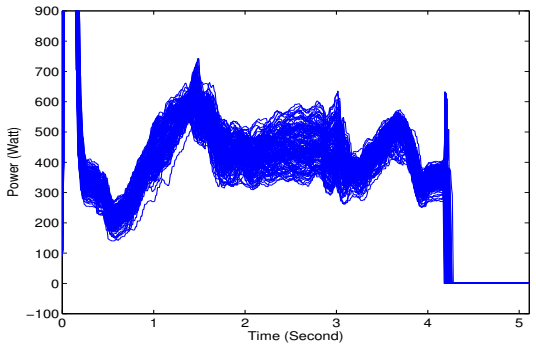
Available algorithms and Packages

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

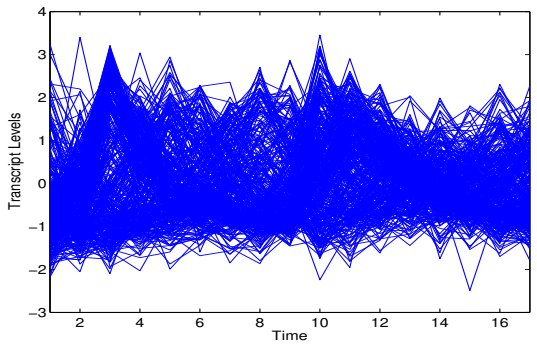
- Meteorits include sampling, fitting, prediction, clustering with each MoE model
- Non-normal mixtures (and MoE) is a very recent topic in the field

Apprentissage à partir de fonctions

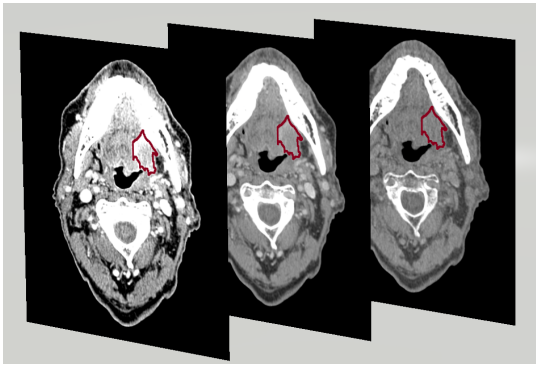
Transport : Railway switch curves



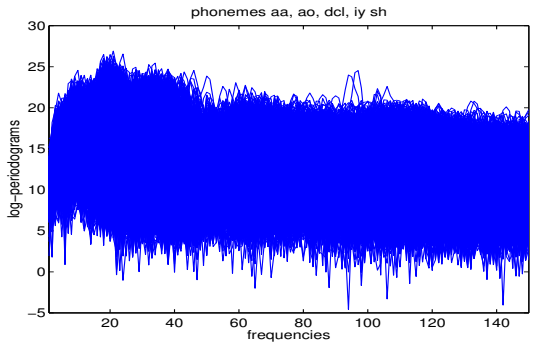
Genomics : Yeast cell cycle curves



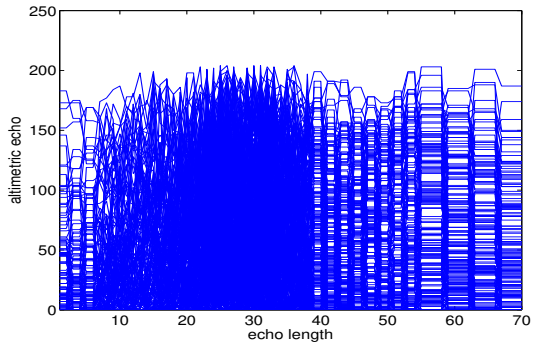
Health : Medical images



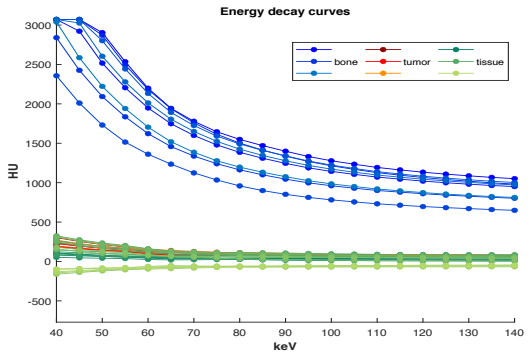
Speech recognition : Phonemes curves



Remote Sensing : Satellite waveforms



Dual-energy computed tomography



- Data : Entire functions (e.g., curves, surfaces) $((\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n)$
- $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{im_i})$ observed at covariates, (e.g., time), $(x_{i1}, \ldots, x_{im_i})$
- Functional regression, classification, clustering/segmentation

$\hookrightarrow$ (Un)supervised analysis of heterogeneous functions with hidden structures

Functional Mixture Models

$$m(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta}) = \sum_{k=1}^{K} \alpha_k f_k(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta}_k)$$

- standard $f_k(y|\boldsymbol{x})$ tailored to functional use (B-)spline regression, regression using wavelet bases etc, or Gaussian process regression, functional PCA
- more tailored to approximate smooth functions ; do not account for segmentation

↪ Here $f_k(y|\boldsymbol{x})$ itself exhibits a clustering property via hidden variables (regimes)

- Piecewise regression mixture model (PWRM)

$$m(\boldsymbol{y}_i|\boldsymbol{x}_i;\boldsymbol{\theta}) = \sum_{k=1}^{K} \alpha_k \underbrace{\prod_{r=1}^{R_k} \prod_{j\in I_{kr}} \mathcal{N}(y_{ij};\boldsymbol{\beta}_{kr}^T\boldsymbol{x}_{ij},\sigma_{kr}^2)}_{\text{Piecewise Regressors (PWR)}}$$

with $I_{kr} =]\xi_{kr}..\xi_{k,r+1}]$ index elements of segment r in cluster k

↪ Simultaneously accounts for curve clustering and segmentation

[J] C. & Nguyen. Model-Based Clustering and Classification of Functional Data. *WIREs Data Mining and Knowl. Discov.* 2019
[J] C. Piec. Reg. Mixture for Simultaneous Functional Data Clustering and Optimal Segmentation. *Journal of Classification*2016
[J] C. Unsupervised learn. of regression mixture models with unknown number of components. *J. of Stat. Comp. and Sim.* 2016

Functional Mixture Models

$$m(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta}) = \sum_{k=1}^{K} \alpha_k f_k(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta}_k)$$

- standard $f_k(y|\boldsymbol{x})$ tailored to functional use (B-)spline regression, regression using wavelet bases etc, or Gaussian process regression, functional PCA
- more tailored to approximate smooth functions ; do not account for segmentation

↪ Here $f_k(y|\boldsymbol{x})$ itself exhibits a clustering property via hidden variables (regimes)

- Mixture of regressions with hidden logistic processes (MixRHLP) :

$$m(\boldsymbol{y}_i|\boldsymbol{x}_i;\boldsymbol{\theta}) = \sum_{k=1}^{K} \alpha_k \underbrace{\prod_{j=1}^{m_i} \sum_{r=1}^{R_k} \pi_{kr}(x_j;\mathbf{w}_k)\mathcal{N}\left(y_{ij};\boldsymbol{\beta}_{kr}^T\boldsymbol{x}_j,\sigma_{kr}^2\right)}_{\text{RHLP}}$$

with $\pi_{kr}(x_j;\mathbf{w}_k) = \mathbb{P}(H_{ij} = r|Z_i = k, x_j;\mathbf{w}_k) \propto \exp\left(w_{kr0} + w_{kr1}x_j\right)$

↪ Simultaneously accounts for clustering and segmentation, in a probabilistic sense

[J] C. & Nguyen. Model-Based Clustering and Classification of Functional Data. *WIREs Data Mining and Knowl. Discov.* 2019
[J] C. Piec. Reg. Mixture for Simultaneous Functional Data Clustering and Optimal Segmentation. *Journal of Classification*2016
[J] C. Unsupervised learn. of regression mixture models with unknown number of components. *J. of Stat. Comp. and Sim.* 2016

Clustering of functional data : Topex/Poseidon satellite data

Topex/Poseidon satellite data : $n = 472$ waveforms of $m = 70$ measured echoes

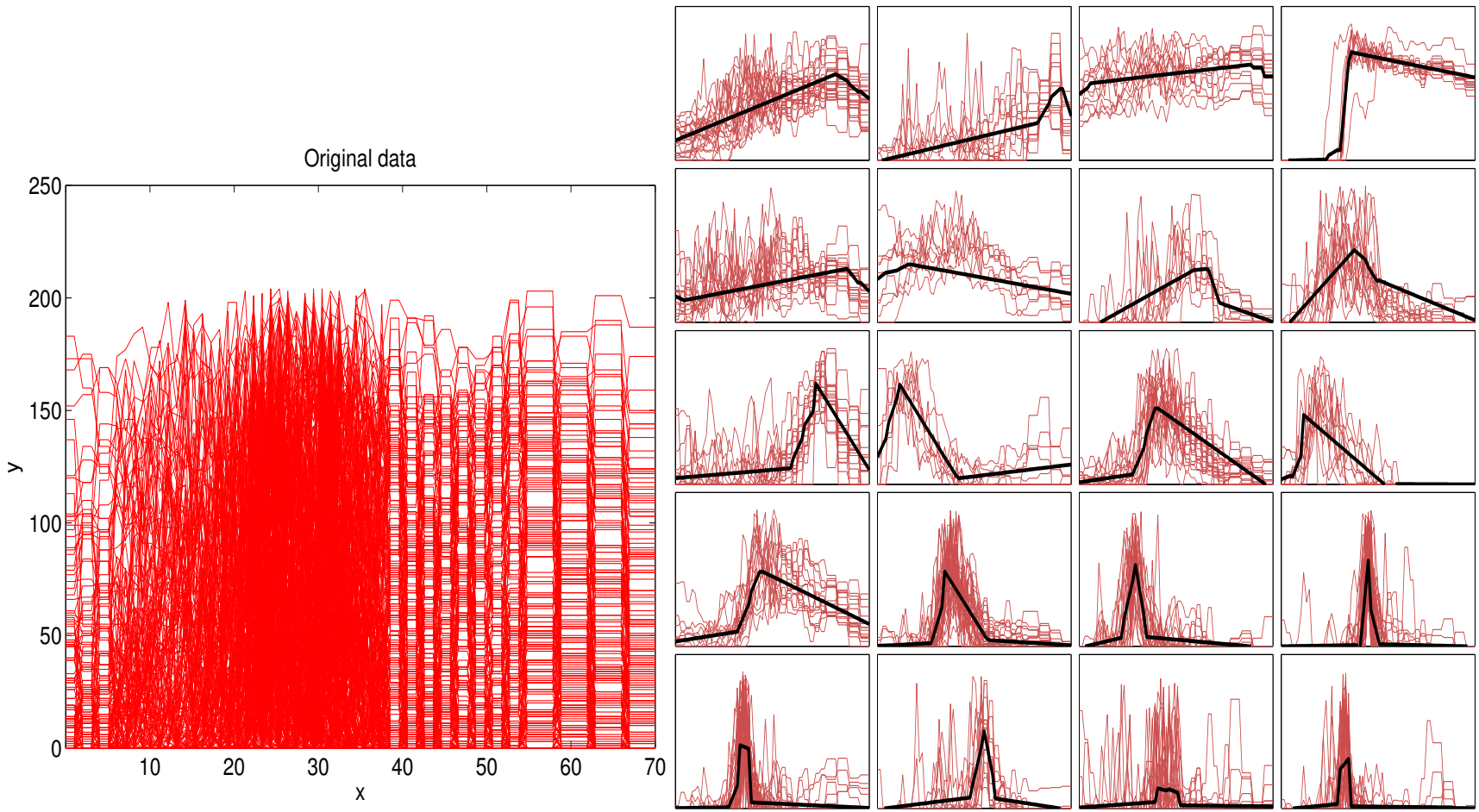
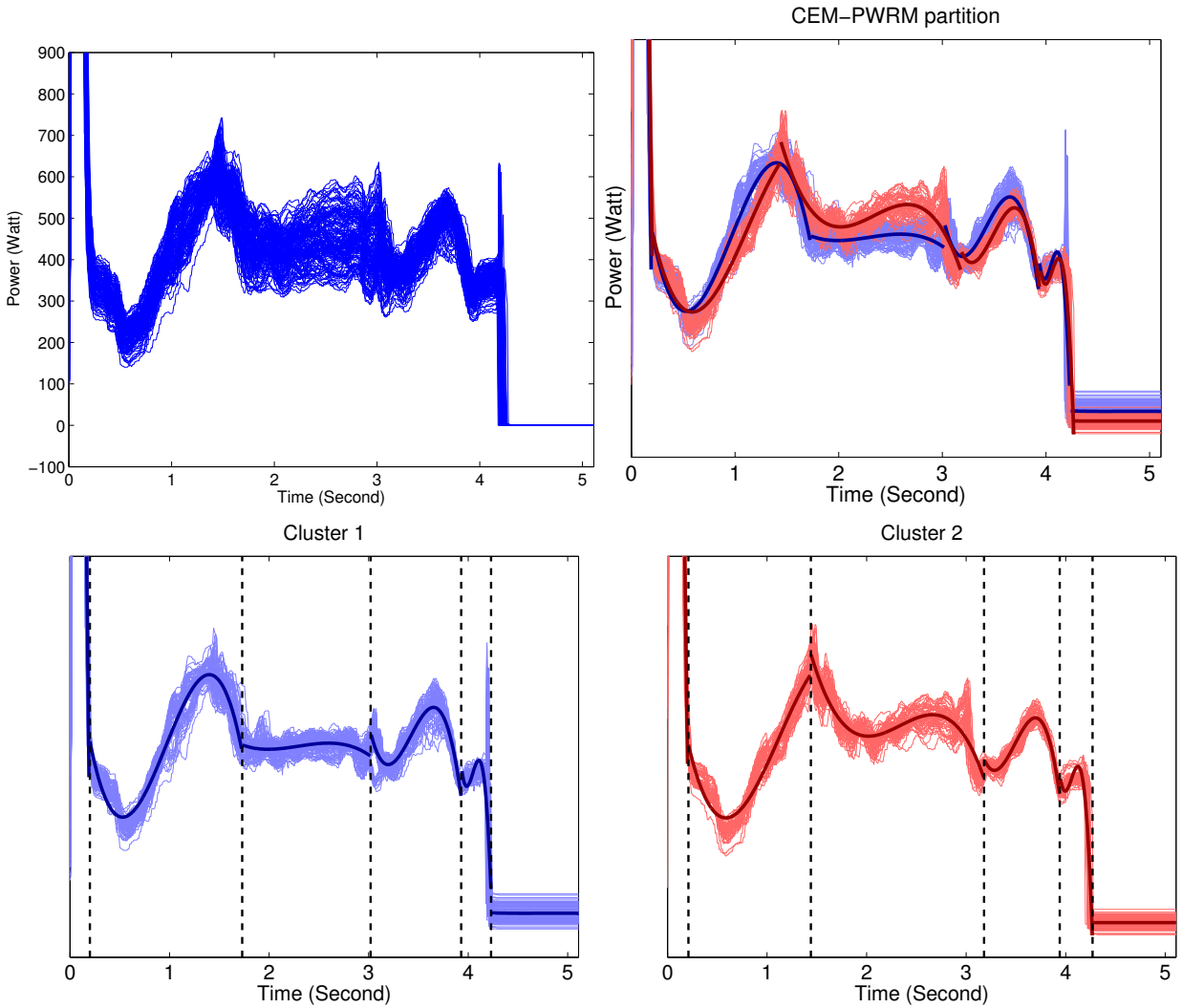


FIGURE – Original data and clustering results from C. 2016 with the same setting as in Hebrail et al. 2010 : twenty clusters and a piecewise linear approximation of four segments.

Clustering an segmentation of functional data : high-speed railway-switch curves

Data set : $n = 146$ real curves of $m = 511$ observations.

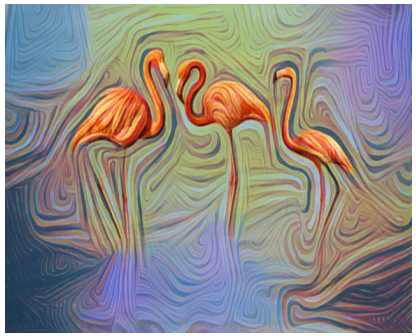
Each curve is composed of $R = 6$ electromechanical phases (regimes)



FLaMingoS : open source software for statistical functional data analysis

+12K téléchargements R depuis août 2019

`https://cranlogs.r-pkg.org/badges/grand-total/flamingos`



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

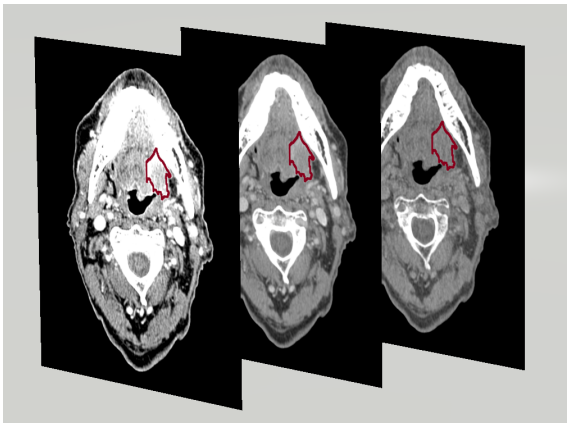
`mixRHLP` : Mixture of Regressions with HLPs	▸ R	▸ Matlab
`mixHMM` : Mixture of Hidden Markov Models (HMMs)	▸ R	▸ Matlab
`mixHMMR` : Mixture of HMM Regressions	▸ R	▸ Matlab
`PWRM` : Piece-Wise Regression Mixture	▸ R	▸ Matlab
`uReMix` : Unsupervised Regression Mixtures	▸ R	▸ Matlab

↪ A flexible full generative modeling for FDA

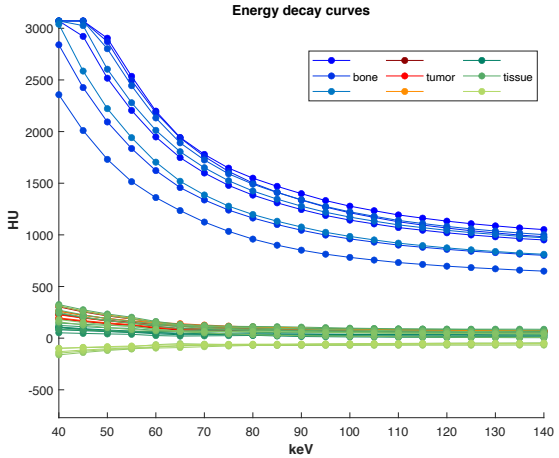
↪ Could be extended to the multivariate case without a major effort

Spatial mixture of functional regressions for dual-energy CT images {Collab avec McGill}

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

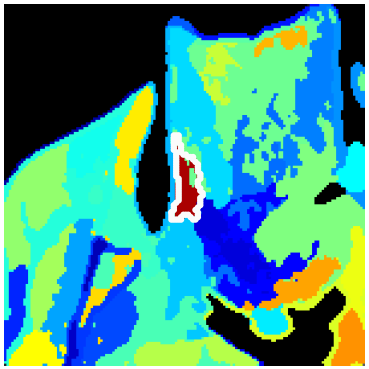


2D slices of VMIs at 40,65,140keV with tumor contour in red.



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] Brivet, Chamroukhi, Coates, Forghani and Savadjiev. Spectral image clustering on dual-energy CT scans using functional regression mixtures. **Under review** in *IEEE Trans. on Medical Imaging*. arXiv :2201.13398, **Jan, 2022** — **[C]** CMStatistics. 2021

▸ **[S]** Source codes are publicly available on Github : `https://github.com/segobrivet/DECT_clustering`

Difference between clustering and co-clustering

- Simultaneous clustering of lines/indiv. ($\boldsymbol{Z}$) and columns/var. ($\boldsymbol{W}$)
- Can be used as a way to reduce dimensionality (var. $\rightarrow \boldsymbol{W}$)

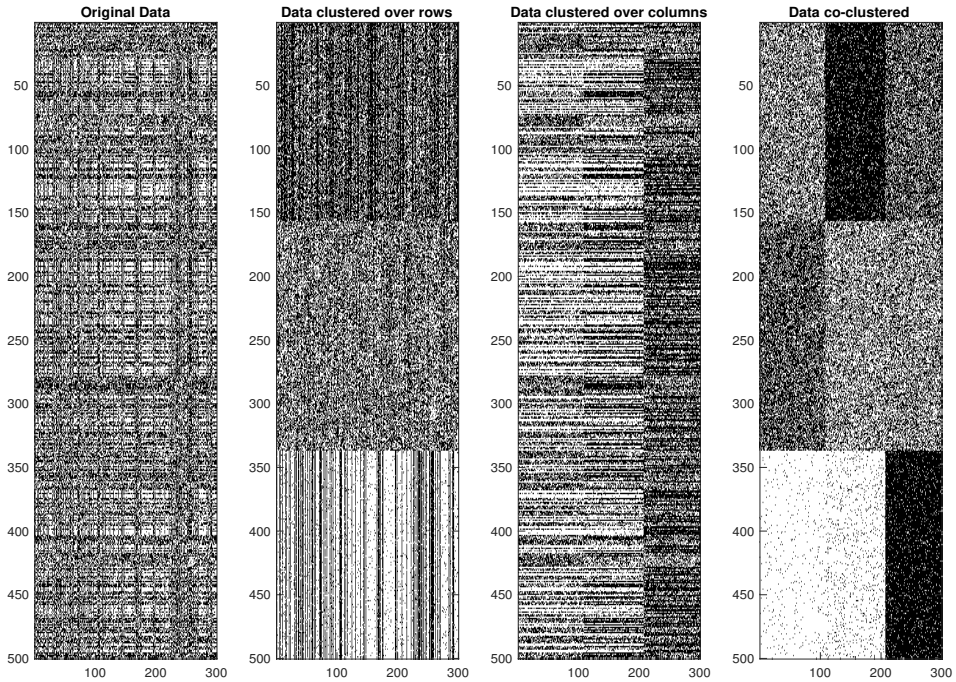


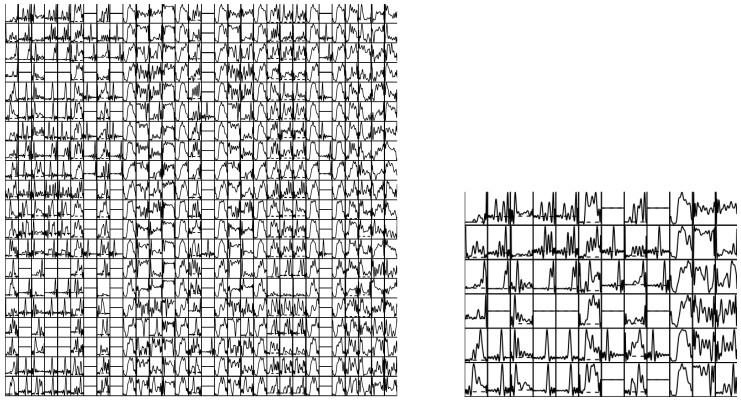
FIGURE – Binary data set with $n = 500$, $d = 300$, $K = M = 3$

The Latent Block Model [Govaert and Nadif, 2013]

$$f(\boldsymbol{X};\boldsymbol{\theta}) = \sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} \mathbb{P}(\boldsymbol{Z},\boldsymbol{W};\boldsymbol{\pi},\boldsymbol{\rho}) \underbrace{f(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{W};\boldsymbol{\theta})}_{\text{data kind dependent}}$$

Co-Clustering of highly multivariate functional data [Chamroukhi & Biernacki, ISI, 2017]

- Data continuously recorded for different subjects from multiple subject' sensors $\hookrightarrow$ e.g : KPI collected from different network elements (transceivers, cells, sites. . .) :



Data Zoom

FIGURE – An example with $d = 30$ and $n = 20$ daily observations [BenSliman 2016].

(1) Model-based co-clustering embedding (2) functional data modeling

- Simultaneous clustering of lines/indiv. ($\boldsymbol{Z}$) and columns/var. ($\boldsymbol{W}$)
- Functional Latent Block Model for Co-clustering :

$$f(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\theta}) = \sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} \mathbb{P}(\boldsymbol{Z};\boldsymbol{\pi})\mathbb{P}(\boldsymbol{W};\boldsymbol{\rho})\, \underbrace{f(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{W};\boldsymbol{\theta})}_{\text{functional data model}}$$

Apprentissage génératif non supervisé à l'échelle

[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}

↪ Méthodes d'**Inférence** et **Sélection parcimonieuses**, Soucis d'**interprétabilité**

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

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

↪ encourages sparse solutions & performs estimation and feature selection

↪ computationally attractive (Avoid matrix inversion ; univariate CF updates)

- High-Dimensional Clustering and Regression (with Gaussian and Poisson outputs)
- High-Dimensional Classification (Categorical outputs)
- EM 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 Doctorat de Normandie Université, 2019. Directeur.

[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

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

Hypothèses : Pour une collection $(\mathcal{H}_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$ de mélanges d'experts, $\rho \in (0,1)$, $C_1 > 1$, $\xi = \sum_{\mathbf{m}\in\mathcal{M}} e^{-z_{\mathbf{m}}} < \infty, z_{\mathbf{m}} \in \mathbb{R}^+, \forall \mathbf{m} \in \mathcal{M}$.

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

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

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

$\hookrightarrow$ (*i*) Choix de K étant donnée n ; $\hookrightarrow$ (*ii*) utiliser quelques experts complexes vs. combiner plusieurs experts simples, compte tenu du nombre total de paramètres.

- $\text{pen}(\mathbf{m}) \propto \dim(\mathcal{H}_{\mathbf{m}})$: pénalisant les modèles complexes.

[Thèse] Trung-Tin Nguyen. Thèse de Doctorat de Normandie Université, 2021. Directeur.

[J] NguyenT, NguyenH, Chamroukhi, McLachlan. An l_1-oracle inequality for the Lasso in mixture-of-experts regression models. **Under revision**, *ESAIM : Probability and Statistics* arXiv :2009.10622, 2020.

[J] Nguyen, Nguyen, Chamroukhi, Forbes. A non-asymptotic penalization criterion for model selection in mixture of experts models. **Under revision**, *Electronic Journal of Statistics*, arXiv :2104.02640. 2021c

[J] Nguyen, Chamroukhi, Nguyen, Forbes. Non-asymptotic model selection in block-diagonal mixture of polynomial experts models. **Under revision**, *Journal of Multivariate Analysis*. arXiv :2104.08959. 2021b

[2.] Learning with functional predictors

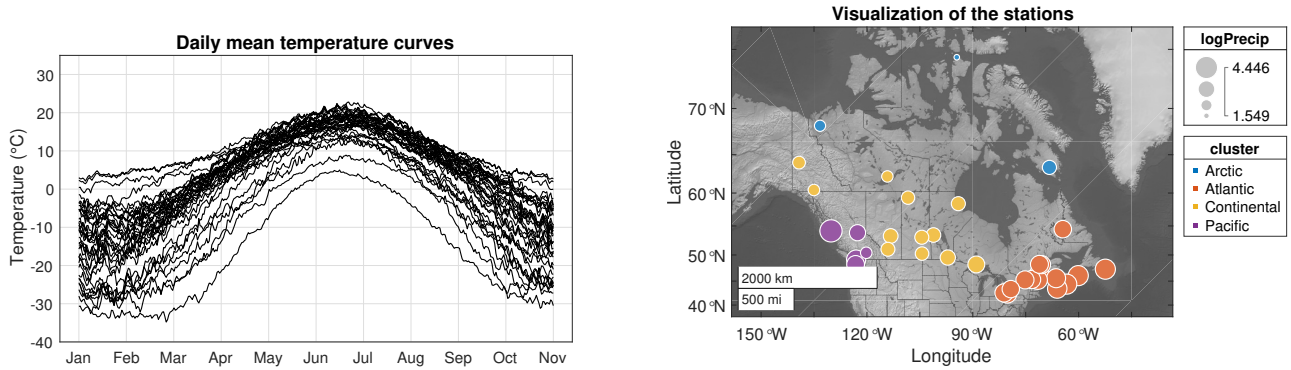


FIGURE – $n = 35$ daily mean temperature measurement curves (X_i's) in different stations (Left) and the log of precipitation values (Y_i's) visualized with the climate regions (Z_i's) (Right).

- 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**)

Representation of the functional ME (FME) model

Basis expansion of the expert regressor network and the gating network :

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

$$h_z(X_i(t), t \in \mathcal{T}; \boldsymbol{\alpha}) = \alpha_{z_i,0} + \int_{\mathcal{T}} X_i(t)\alpha_{z_i}(t)dt = \alpha_{z_i,0} + \underbrace{\boldsymbol{x}_i^\top \left(\int_{\mathcal{T}} \mathbf{b}_r(t)\mathbf{b}_q^\top(t)dt \right)}_{\mathbf{r}_i^\top} \boldsymbol{\zeta}_{z_i}$$

Different Learning strategies, in particular

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

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

$\hookrightarrow$ 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.

[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**)

- 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 $\boldsymbol{\gamma}_{z_i} = \mathbf{A}_p\boldsymbol{\eta}_{z_i}$

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

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

Functional expert' network of iFME

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

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

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

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

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

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 : $\text{Pen}_{\lambda,\chi}(\boldsymbol{\Psi}) = \lambda \sum_{k=1}^{K} \|\boldsymbol{\gamma}_k\|_1 + \chi \sum_{k=1}^{K-1} \|\boldsymbol{\omega}_k\|_1$

Illustratin : Interpretable fits using iFME

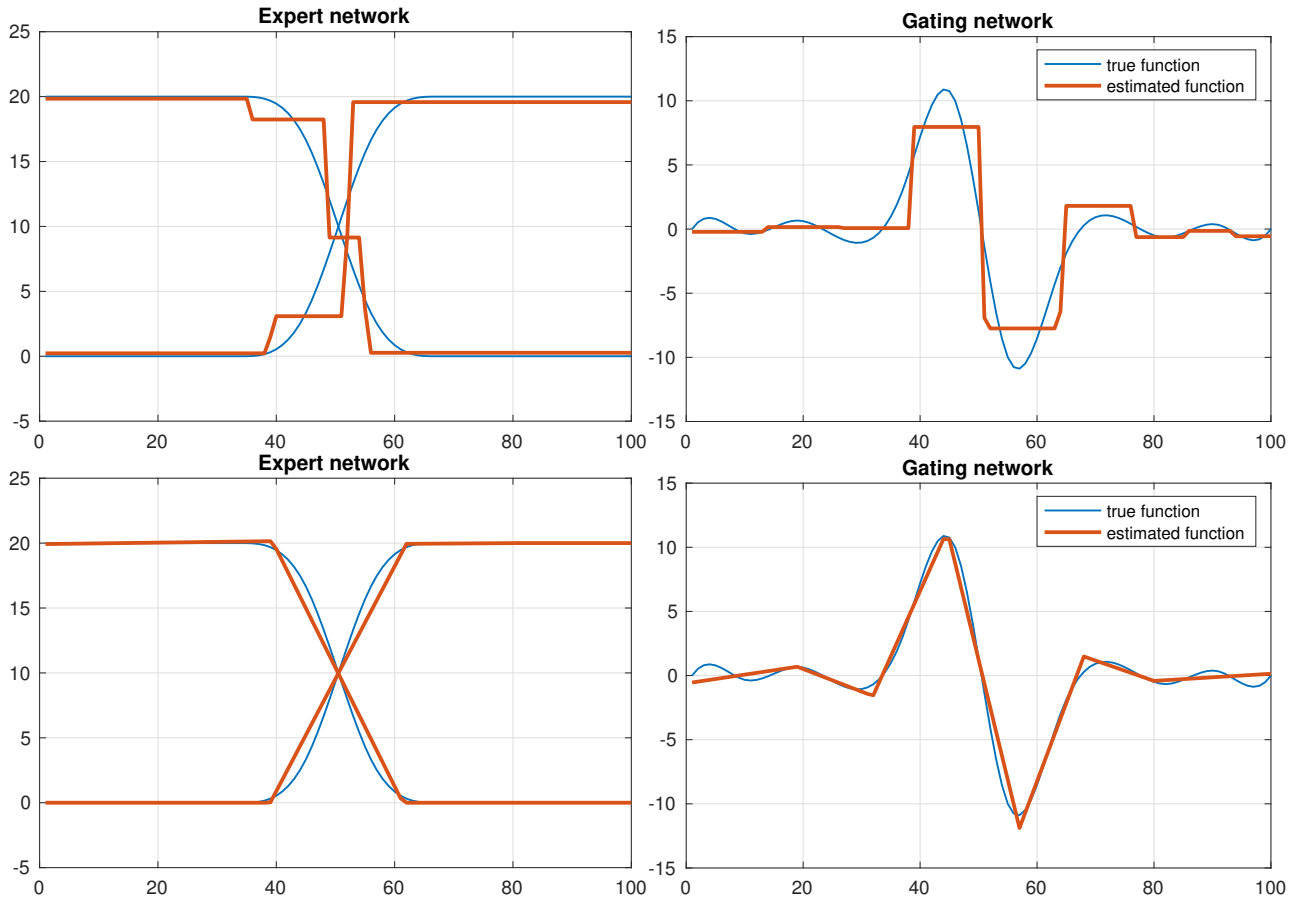


FIGURE – Expert and gating network estimated by FME-FLIRTI with penalization giving for zeroth derivative and first (top)/ second (bottom) derivative

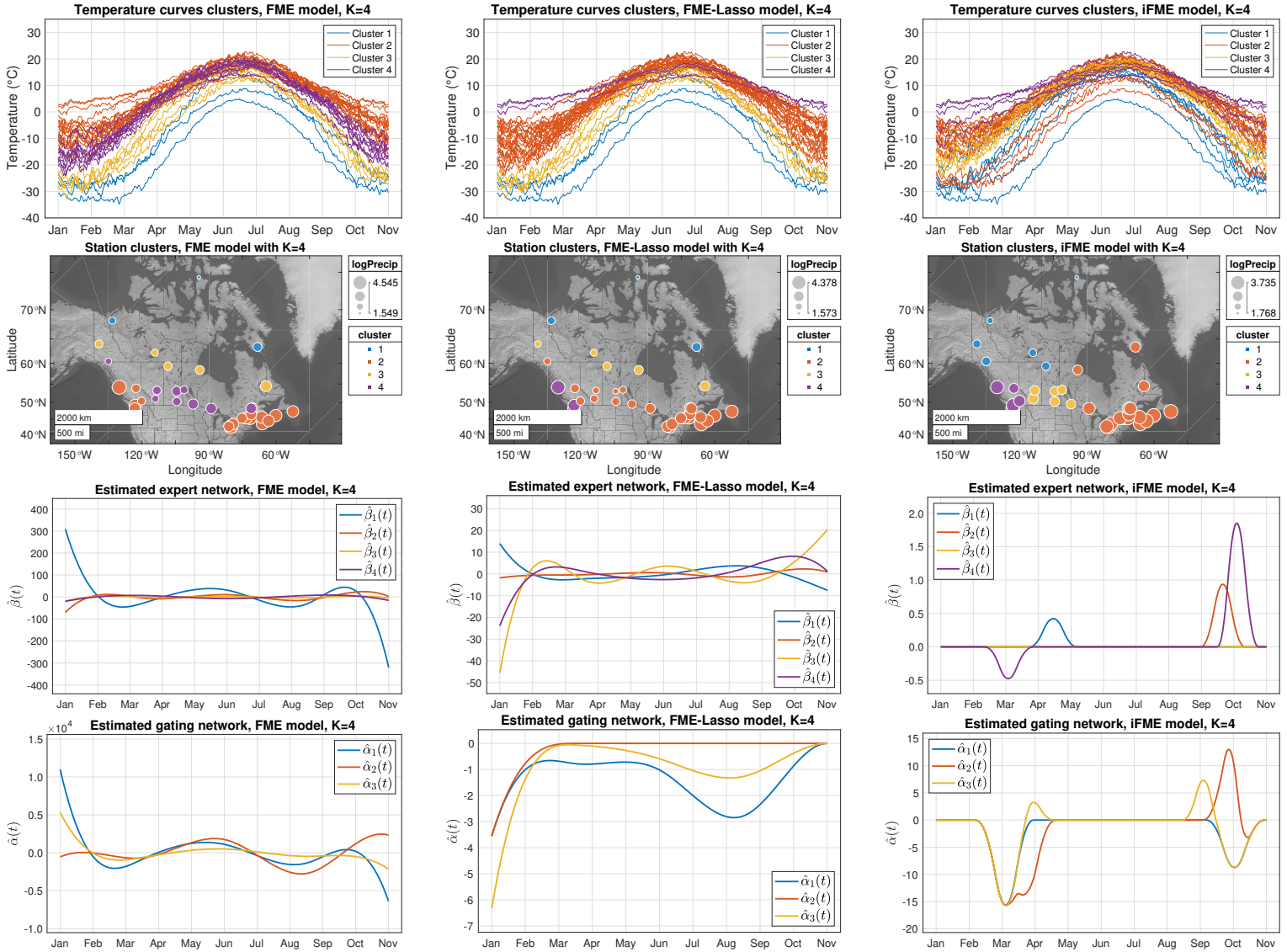


FIGURE – FME-MLE (left), FME-Lasso (middle) and iFME (right) on Canadian weather data

[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**)

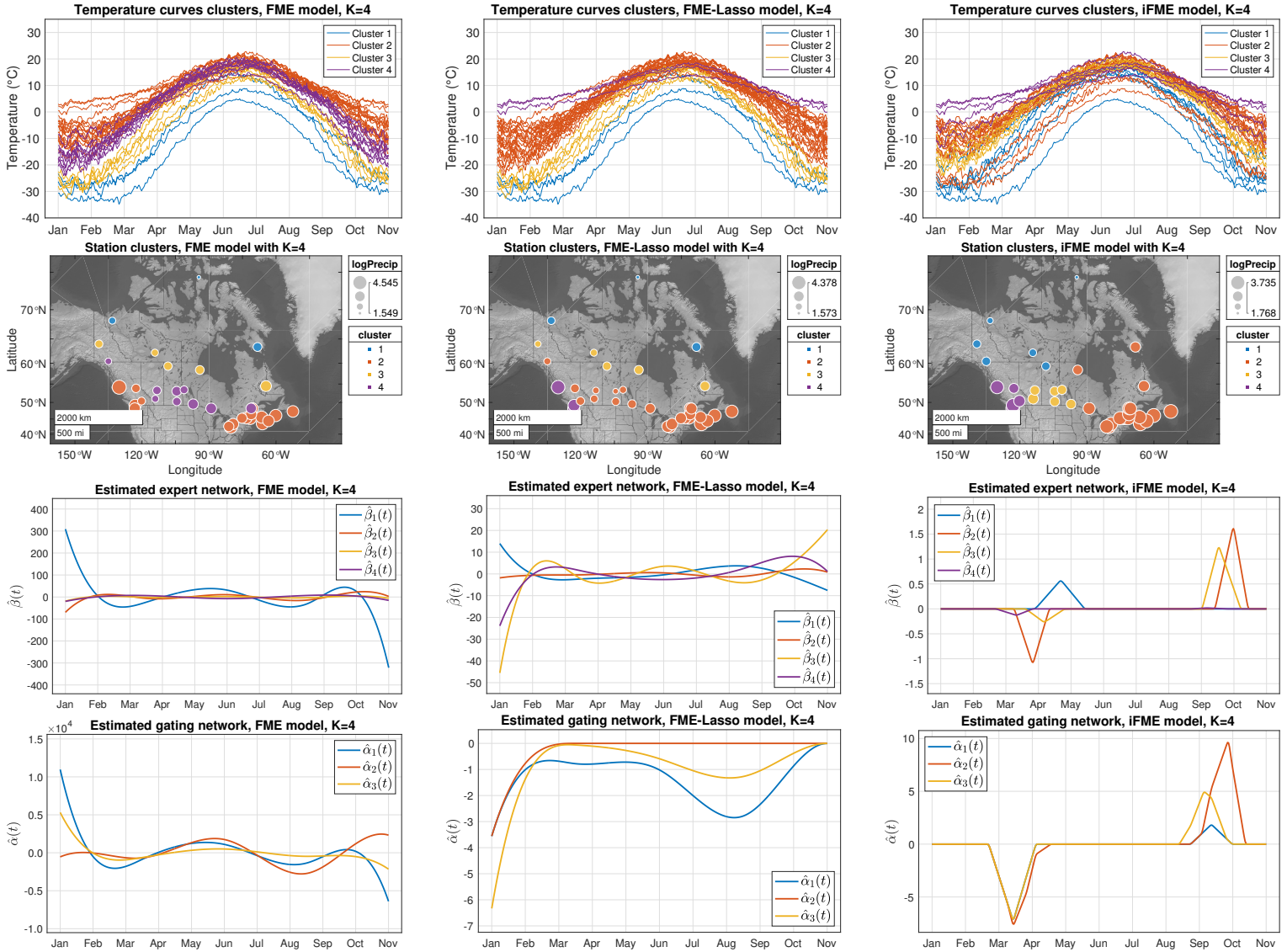


FIGURE – FME-MLE (left), FME-Lasso (middle) and iFME (right) on Canadian weather data

[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**)

[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

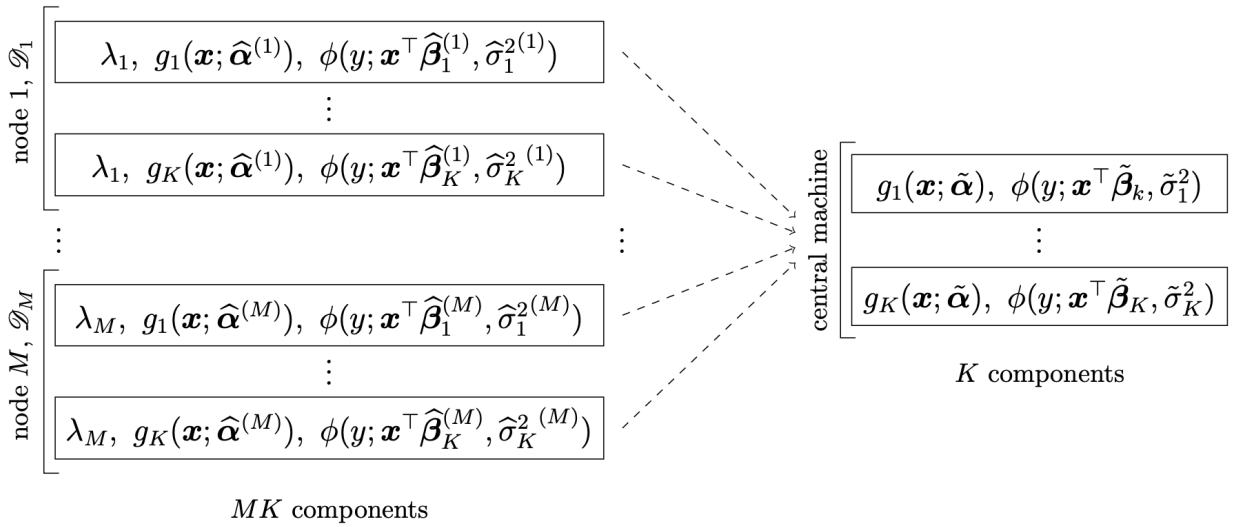
Computational guidelines

- 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*

[Prep.] Pham & Chamroukhi **In progress**, *Distributed Mixtures-of-Experts*, 2022
[Thèse] Pham. *Latent Data Models for Large-Scale Clustering*. Thèse en cours, apr.2018-2022. Directeur.
[Contrat] ANR SMILES

Aggregating distributed mixtures-of-experts models



- 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 related to MK componentsso not our direct target.

↪ 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.

[Prep.] Pham & Chamroukhi **In progress**, *Distributed Mixtures-of-Experts*, 2022
[Thèse] Pham. *Latent Data Models for Large-Scale Clustering*. Thèse en cours, apr.2018-2022. Directeur.
[Contrat] ANR SMILES

Numerical results in Distributed clustering and Prediction

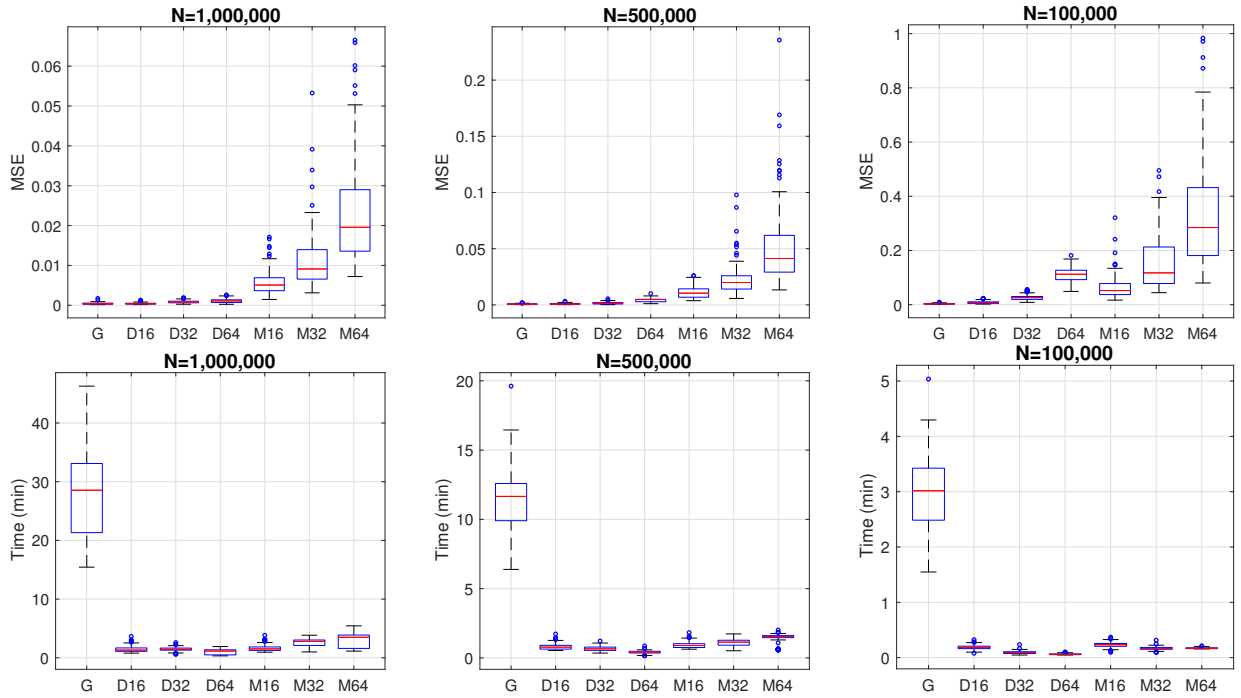


FIGURE – Global (G), Distributed (D) and Median (M) approaches (with different number of machines and sample sizes).

	Dataset size	1M	500k	100k
Clurstering Error	Global	$.0738_{(.0008)}$	$.0737_{(.0011)}$	$.0738_{(.0025)}$
	Distributed, 16 machines	$.0738_{(.0009)}$	$.0737_{(.0012)}$	$.0740_{(.0028)}$
	Distributed, 32 machines	$.0738_{(.0008)}$	$.0738_{(.0012)}$	$.0742_{(.0034)}$
	Distributed, 64 machines	$.0738_{(.0009)}$	$.0739_{(.0012)}$	$.0786_{(.0282)}$
	Dataset size	1M	500k	100k
Prediction Error	Global	$.1215_{(.0017)}$	$.1212_{(.0023)}$	$.1215_{(.0051)}$
	Distributed, 16 machines	$.1219_{(.0024)}$	$.1212_{(.0023)}$	$.1248_{(.0277)}$
	Distributed, 32 machines	$.1222_{(.0032)}$	$.1217_{(.0034)}$	$.1262_{(.0294)}$
	Distributed, 64 machines	$.1224_{(.0032)}$	$.1230_{(.0038)}$	$.1559_{(.1107)}$

- densité marginale des observations : $p_{\boldsymbol{\theta}}(\mathbf{x}) = \int_{\mathbb{R}^d} p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) d\mathbf{z} = \int_{\mathbb{R}^d} p(\mathbf{z}) p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) d\mathbf{z}$ intègre, dans ce contexte de représentation, des construction neuronales.

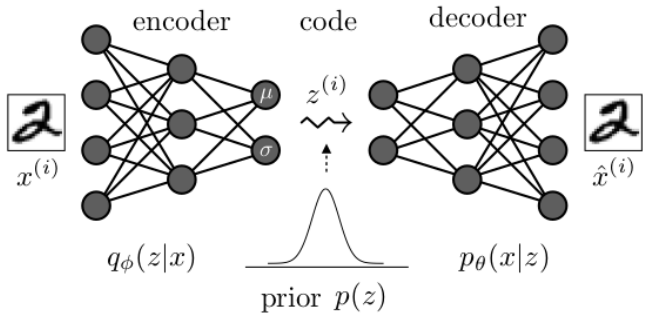


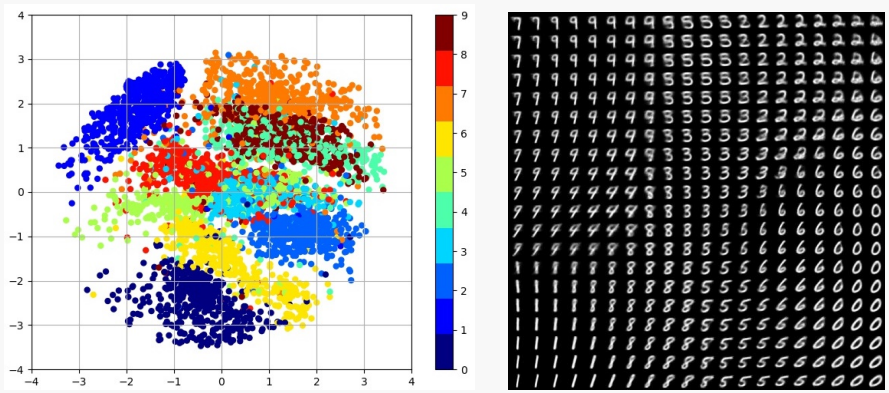
FIGURE – Variational Auto-Encoders (figure taken from Tschannen et al. 2018.)

- si $q^*(\mathbf{z}|\mathbf{x}) = \arg\max_{q \in \mathcal{Q}} \mathcal{L}\left[q(\mathbf{z}|\mathbf{x})\right]$ ($\mathcal{L}$ étant l'ELBO) la meilleure approximation variationnelle dans une famille de distributions $\mathcal{Q}$, l'écart d'inférence $\mathcal{G}$ se décompose :

$$\mathcal{G} = \underbrace{\mathrm{KL}\left(q^*_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})\right)}_{\text{Écart d'approximation}} + \underbrace{\mathrm{KL}\left(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})\right) - \mathrm{KL}\left(q^*_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})\right)}_{\text{Écart d'amortissement}}.$$

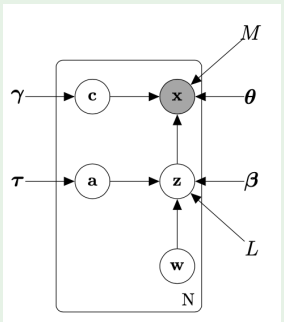
$\hookrightarrow$ une façon de réduire $\mathcal{G}$ est de réduire l'écart d'amortissement via une famille $\mathcal{Q}$ plus large d'une distribution variationnelle non gaussienne pour l'encodeur $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$.
$\hookrightarrow$ L'utilisation de mélanges (capacité d'approximation), contribuerait à réduire cet écart.

Motivation : Learn multimodal representation spaces and data spaces

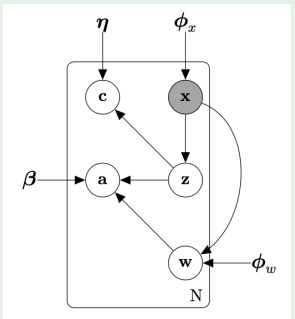


Clustering : Two-dimensional latent space of MNIST digits

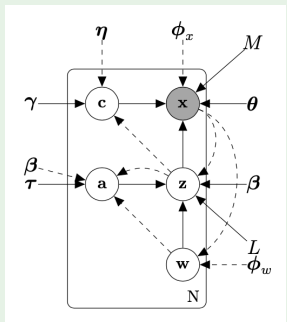
Proposal : Mixture of Prior-Decoder Variational Autoencoders (MPDVAE)



Mixture genrator $p_{\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\beta},\boldsymbol{\tau}}(\mathbf{x}, \mathbf{z}, \mathbf{c}, \mathbf{a}, \mathbf{w})$



Recognition $q_{\phi,\boldsymbol{\beta},\boldsymbol{\eta}}(\mathbf{z}, \mathbf{c}, \mathbf{a}, \mathbf{w} \mid \mathbf{x})$



complete model

↪ warranty for estimation via Stochastic GD or Stochastic EM

↪ Our MPDVAE would be the first ? model that captures the heterogeneous relationships in both data space and latent space

↪ Pb. Posterior collapse : L'effondrement de la postérieure $q_{\boldsymbol{\phi}}$ couramment observé dans les VAEs

↪ [Wang et al. Neurips 2021] : c'est un problème non spécifique à l'utilisation des réseaux de neurones ou de l'inférence variationnelle : le postérieur s'effondre si et seulement si les variables latentes sont non-identifiables dans le modèle génératif.

↪ Pistes de fonctions enforçant l'identifiabilité faisant appel au transport monotone de Brenier.

Merci pour votre attention !