Model-based (co-)clustering in some high-dimensional scenarios

Faicel Chamroukhi

Working Group on Model-Based Clustering summer session

Ann Arbor, July 15-21, 2018

Outline

1. Model-Based Co-Clustering of Multivariate Functional Data

   Joint work with Christophe Biernacki, INRIA-Lille

2. Regularized Mixture-of-Experts for high-dimensional data

   Joint work with Bao Tuyen Huynh, Unicaen, LMNO

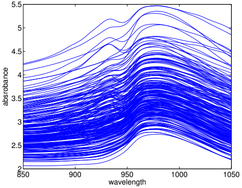
Outline

1. Model-Based Co-Clustering of Multivariate Functional Data
   - Motivation
   - Model-based co-clustering
   - Temporal curve segmentation (RHLP)
   - Model-based co-clustering embedding RHLP
   - Conclusion and perspectives

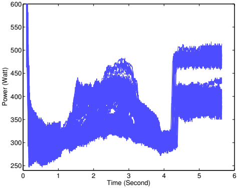
2. Regularized Mixture-of-Experts for high-dimensional data

Functional data are increasingly frequent

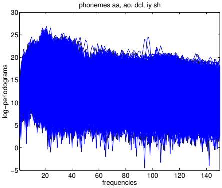
[James and Hastie, 2001; James and Sugar, 2003]
[Ramsay and Silverman, 2005]
[Chamroukhi et al., 2010]
[Bouveyron and Jacques, 2011]
[Samé et al., 2011]
[Jacques and Preda, 2014]
[Bouveyron et al., 2018]
[Chamroukhi and Nguyen, 2018]



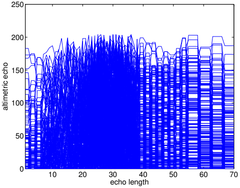
Tecator data



Railway switch curves



Phonemes curves



Satellite waveforms

Clustering of functional data

$\hookrightarrow$ a growing investigation of Model-Based Clustering (MBC) for functional data

Some Reviews on MBC for functional data: [Jacques and Preda, 2014; Chamroukhi and Nguyen, 2018]

Clustering of functional data

$\hookrightarrow$ a growing investigation of Model-Based Clustering (MBC) for functional data

Some Reviews on MBC for functional data: [Jacques and Preda, 2014; Chamroukhi and Nguyen, 2018]

Tecator data set[1]: $n = 240$ spectra with $m = 100$

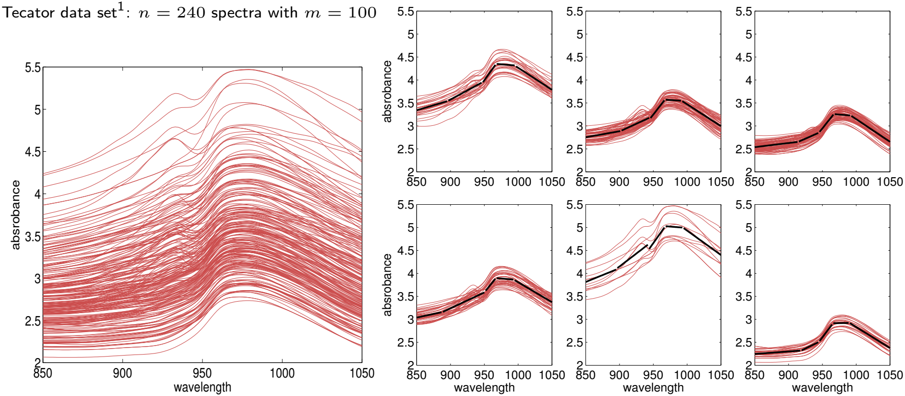


Figure: Original data and clustering results from Chamroukhi [2016b] for the data considered in the same setting as in Hébrail et al. [2010] (six clusters, each cluster is approximated by five linear segments ($R = 5, p = 1$))

Clustering of functional data

Topex/Poseidon satellite data[2]: $n = 472$ waveforms of $m = 70$ measured echoes

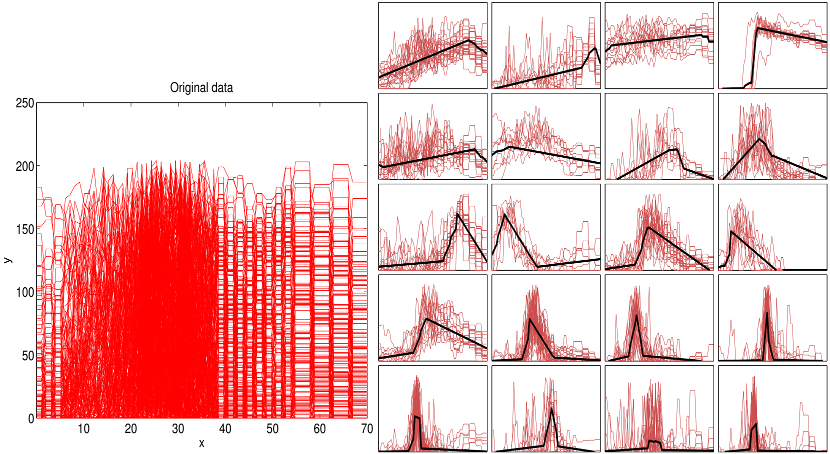


Figure: Original data and clustering results from Chamroukhi [2016b] with the same setting as in Hébrail et al. [2010]: twenty clusters and a piecewise linear approximation of four segments.

[2] Satellite data are available at `http://www.lsp.ups-tlse.fr/staph/npfda/npfda-datasets.html`.

Clustering of functional data

Phonemes data set[3]: $n = 1000$ log-periodograms for $m = 150$ frequencies

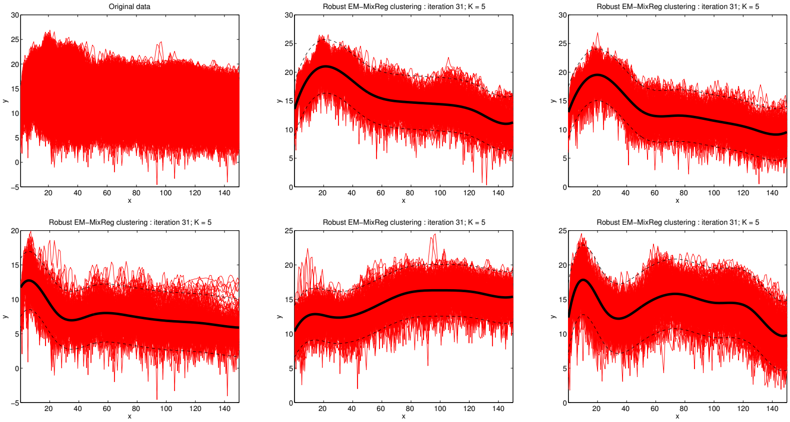


Figure: Original data and clustering results from Chamroukhi [2016b]

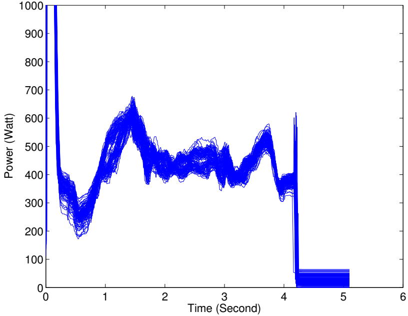
[3] Data from `http://www.math.univ-toulouse.fr/staph/npfda/`, used in Ferraty and Vieu [2003]

Clustering of functional data

Clustering real curves of high-speed railway-switch operations

Data: $n = 115$ curves of $m \simeq 510$ observations

$K = 2$ clusters: operating state without/with possible defect

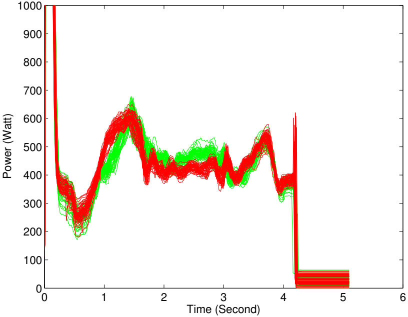


Clustering switch operations

Clustering real curves of high-speed railway-switch operations

Data: $n = 115$ curves of $m \simeq 510$ observations

$K = 2$ clusters: operating state without/with possible defect



Outline

1. Model-Based Co-Clustering of Multivariate Functional Data
   - Motivation
   - Model-based co-clustering
   - Temporal curve segmentation (RHLP)
   - Model-based co-clustering embedding RHLP
   - Conclusion and perspectives
2. Regularized Mixture-of-Experts for high-dimensional data

This talk: Multivariate functional data clustering

- Multivariate functional data are increasingly present
- e.g: Data continuously recorded for different subjects from multiple subject' sensors

↪ Measurements collected from different network elements (transceivers, cells, sites. . .):

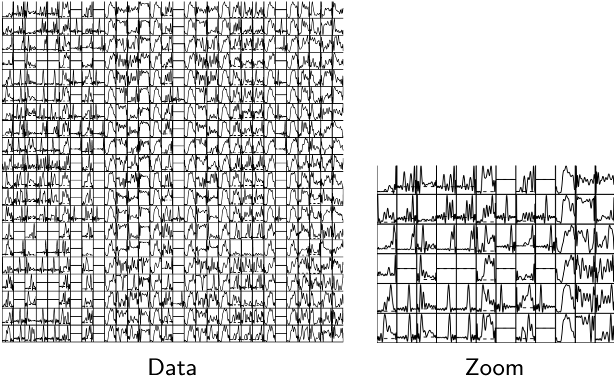


Figure: An example with $d = 30$ and $n = 20$ daily observations [Ben Slimen et al., 2016].

This talk

Questioning

Clustering of highly multivariate functional data with two guidelines:

- (1) Mathematical guideline: warranty for estimation and selection
- (2) User guideline: keep a user-friendly meaning of the process

Both are important because clustering is a highly risky task...

Proposed answering

(1) Model-based co-clustering with (2) temporal curve segmentation

Novelty corresponds to combining both (1) and (2)

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. $\rightarrow \boldsymbol{W}$)

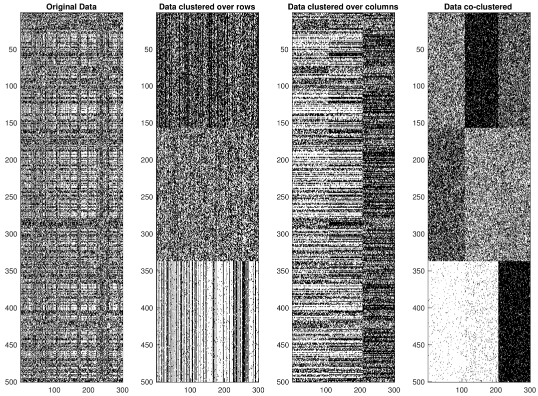


Figure: Binary data set with $n = 500$, $d = 300$, $K = M = 3$

Latent block model for co-clustering

The Latent Block Model [Govaert and Nadif, 2013]

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

Hypotheses

- The latent variables $\boldsymbol{Z}$ and $\boldsymbol{W}$ are independent: $\mathbb{P}(\boldsymbol{Z},\boldsymbol{W}) = \mathbb{P}(\boldsymbol{Z})\mathbb{P}(\boldsymbol{W})$ and iid:
  $\mathbb{P}(\boldsymbol{Z}) = \prod_i \mathbb{P}(z_i)$ with $z_i \sim$ Multinomial$(\pi_1,\ldots,\pi_K)$ where $\pi_k = \mathbb{P}(z_k = k)$
  $\mathbb{P}(\boldsymbol{W}) = \prod_j \mathbb{P}(w_j)$ with $w_j \sim$ Multinomial$(\rho_1,\ldots,\rho_M)$ where $\rho_\ell = \mathbb{P}(w_j = \ell)$
- Conditional independence: $x_{ij}|(z_i,w_j) \perp x_{i'j'}|(z_{i'},w_{j'})$

Latent block model for co-clustering

The Latent Block Model [Govaert and Nadif, 2013]

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

Hypotheses

- The latent variables $\boldsymbol{Z}$ and $\boldsymbol{W}$ are independent: $\mathbb{P}(\boldsymbol{Z},\boldsymbol{W}) = \mathbb{P}(\boldsymbol{Z})\mathbb{P}(\boldsymbol{W})$ and iid:

  $\mathbb{P}(\boldsymbol{Z}) = \prod_i \mathbb{P}(z_i)$ with $z_i \sim$ Multinomial$(\pi_1,\ldots,\pi_K)$ where $\pi_k = \mathbb{P}(z_k = k)$

  $\mathbb{P}(\boldsymbol{W}) = \prod_j \mathbb{P}(w_j)$ with $w_j \sim$ Multinomial$(\rho_1,\ldots,\rho_M)$ where $\rho_\ell = \mathbb{P}(w_j = \ell)$
- Conditional independence: $x_{ij}|(z_i,w_j) \perp x_{i'j'}|(z_{i'},w_{j'})$

$\hookrightarrow$ binary data: binary [Govaert and Nadif, 2003, 2008; Keribin et al., 2012],

$\hookrightarrow$ categorical data: multinomial [Keribin et al., 2014]

$\hookrightarrow$ contingency table: Poisson [Govaert and Nadif, 2003, 2006, 2008]

$\hookrightarrow$ continuous data: Gaussian [Lomet, 2012; Govaert and Nadif, 2013]

$\hookrightarrow$ functional data: functional PCA + Gaussian, see further [Ben Slimen et al., 2016]

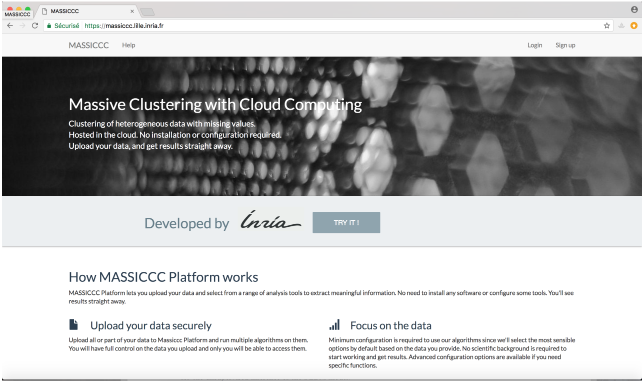
Inference for the latent block model

Inference of the latent block model

- variational block EM (VBEM) for maximum likelihood estimation and fuzzy co-clustering [Govaert and Nadif, 2006, 2008].
- block classification EM (CEM) algorithm for maximum classification likelihood and hard co-clustering [Govaert and Nadif, 2003, 2006, 2008]
- Bayesian inference [Keribin et al., 2012, 2014]: Bayesian latent block mixtures for binary data and categorical data & a variational Bayesian inference and Gibbs sampling.
- Number of blocks estimation: ICL criterion [Lomet, 2012; Keribin et al., 2014]

Package blockcluster on the cloud

massiccc.lille.inria.fr



Functional data notation

- Data: (discretized) values of underlying smooth functions, not just vectors
- Data: A sample of n heterogeneous univariate curves $(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n)$
- $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ consists of m_i observations $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{im_i})$ observed at the independent covariates, (e.g., time t in time series), $(x_{i1}, \ldots, x_{im_i})$

Functional data modeling: "classical" approach

[Ramsay and Silverman, 2005] and many others

- Step 1: $(\boldsymbol{x}, \boldsymbol{y})$ decomposed into a finite basis of function (B-spline...) : $Y_i(t) \approx \sum_{r=1}^{d} c_{ir}\phi_r(x_i(t))$ with $\mathbf{c}$ estimated by OLS
- Step 2: functional principal components analysis (PCA) which is performed as a usual PCA of the basis expansion coefficients $\mathbf{c}$ using a metric defined by the inner products between the basis functions
- Step 3: set a probability distribution on $\mathbf{c}$, typically Gaussian

It defines a distribution on $\mathbf{c}$ instead of $\boldsymbol{y}$...

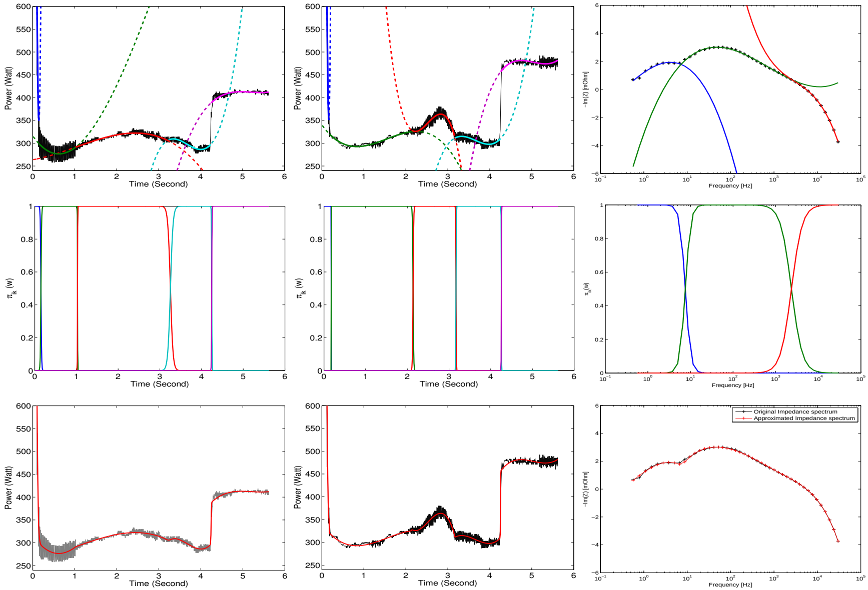
Functional data modeling: regression RHLP

Alternatively, use a segmentation via generative piecewise polynomial regression modeling of $f(\boldsymbol{y}|\boldsymbol{x})$ [Chamroukhi et al.])

$\hookrightarrow$ Regression with Hidden Logistic Process (RHLP)
$\hookrightarrow$ See formula later

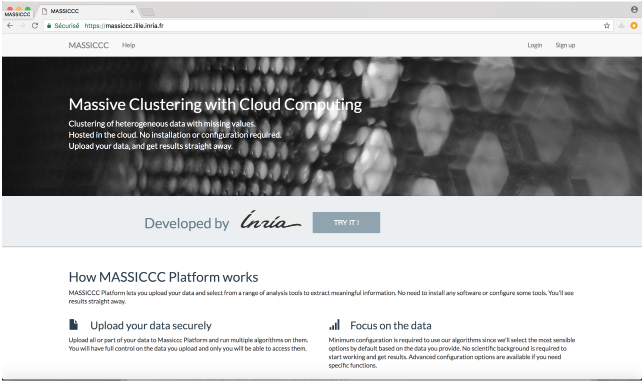
It gives a distribution on $\boldsymbol{y}$ and also a meaningful segmentation of the curve

RHLP for modeling different types of functions



Package mixtcomp on the cloud

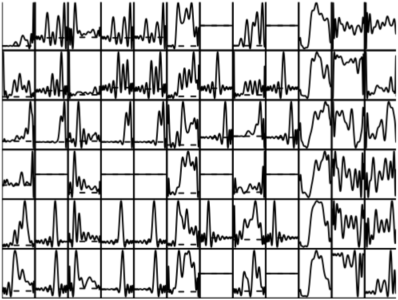
massiccc.lille.inria.fr



Multivariate functional data co-clustering

[Chamroukhi and Biernacki, 2017]

- Data: $\boldsymbol{Y} = (\boldsymbol{y}_{ij})$ a data sample matrix of n individuals defined on a set $\mathcal{I}$ and d continuous functional variables defined on a set $\mathcal{J}$.
- Each variable $\boldsymbol{y}_{ij}$ is an univariate curve $\boldsymbol{y}_{ij} = \big(y_{ij}(t_1), \ldots, y_{ij}(t_{T_{ij}})\big)$ of T_{ij} observations $y(t) \in \mathbb{R}$ linked to covariates $\boldsymbol{x}_{ij} = \big(x_{ij}(t_1), \ldots, x_{ij}(t_{T_{ij}})\big)$ at the points $(t_1, \ldots, t_{T_{ij}})$, typically a sampling time



Embedding RHLP in co-clustering

[Chamroukhi and Biernacki, 2017]

- Functional Latent Block Model for Co-clustering:

$$\begin{aligned} f(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\Psi}) &= \sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} \mathbb{P}(\boldsymbol{Z};\boldsymbol{\pi})\mathbb{P}(\boldsymbol{W};\boldsymbol{\rho}) f(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{W};\boldsymbol{\theta}) \\ &= \sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} \prod_{i,k} \pi_k^{z_{ik}} \prod_{j,\ell} \rho_\ell^{w_{j\ell}} \prod_{i,j,k,\ell} \underbrace{f(\boldsymbol{y}_{ij}|\boldsymbol{x}_{ij};\boldsymbol{\theta}_{k\ell})}_{\text{RHLP}}{}^{z_{ik}w_{j\ell}}. \end{aligned}$$

with parameter vector $\boldsymbol{\Psi} = (\boldsymbol{\pi}^T, \boldsymbol{\rho}^T, \boldsymbol{\theta}^T)^T$, where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)^T$, $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_M)^T$, and $\boldsymbol{\theta} = (\boldsymbol{\theta}_{11}^T, \ldots, \boldsymbol{\theta}_{k\ell}^T, \ldots, \boldsymbol{\theta}_{KM}^T)^T$.

Embedding RHLP in co-clustering

- RHLP [Chamroukhi et al., 2009]: model the conditional data distribution for each block kl, assuming that each functional variable $\boldsymbol{y}_{ij}$ is governed by an $S_{k\ell}$-state hidden process of y_{ij}:

$$f(\boldsymbol{y}_{ij}|\boldsymbol{x}_{ij};\boldsymbol{\theta}_{k\ell}) = \prod_{t=1}^{T_{ij}} \sum_{r=1}^{S_{k\ell}} \alpha_{k\ell r}(t;\boldsymbol{\xi}_{k\ell})\mathcal{N}\left(y_{ij}(t);\boldsymbol{\beta}_{k\ell r}^T \boldsymbol{x}_{ij}(t),\sigma_{k\ell r}^2\right)$$

where the dynamical weights α's are given by the multinomial logistic:

$$\alpha_{k\ell r}(t;\boldsymbol{\xi}_{k\ell}) = \frac{\exp\left(\xi_{k\ell r0}+\xi_{kr\ell 1}t\right)}{1+\sum_{r'=1}^{S_{k\ell}-1}\exp\left(\xi_{k\ell r'0}+\xi_{k\ell r'1}t\right)}.$$

$\hookrightarrow$ Can be seen as a generative piecewise polynomial regression model where the transition points are smoothly controlled by logistic weights

$\hookrightarrow$ a particular mixture-of-experts model [Jacobs et al., 1991; Jordan and Jacobs, 1994]/(parametric) mixture of regressions with predictor-dependent mixing proportions [Young and Hunter, 2010]

Block mean curve approximation and segmentation

- Approximation: a prototype mean curve

$$y_t|(z_i, w_j) \approx \widehat{y}_t = \mathbb{E}[Y(t)|z_i, w_j, x(t); \widehat{\boldsymbol{\Psi}}] = \sum_{s=1}^{S_{kl}} \alpha_{k\ell r}(t; \widehat{\boldsymbol{\xi}}_{k\ell}) \widehat{\boldsymbol{\beta}}_{k\ell r}^T \boldsymbol{x}_i(t)$$

$\hookrightarrow$ A smooth and flexible approximation thanks to the the logistic weights

- Curve segmentation:

$$\widehat{h}_t|(z_i, w_j) = \arg \max_{1 \leq s \leq S_{kl}} \mathbb{E}[H_t|z_i, w_j, x_{ij}(t); \widehat{\boldsymbol{\xi}}] = \arg \max_{1 \leq k \leq K} \alpha_{k\ell r}(t; \widehat{\boldsymbol{\xi}}_{k\ell})$$

Parameter estimation: EM not feasible

- The complete-data log-likelihood:

$$\begin{aligned}\log L_c(\boldsymbol{\Psi}) &= \log f(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{H} | \boldsymbol{X}; \boldsymbol{\Psi}) \\ &= \sum_{i,k} z_{ik} \log \pi_k + \sum_{j,\ell} w_{j\ell} \log \rho_\ell \\ &\quad + \sum_{i,j,k,\ell,t,r} z_{ik} w_{j\ell} h_{tr} \log \left[\alpha_{k\ell r}(t; \boldsymbol{\xi}_{k\ell}) \mathcal{N}\left(y_{ij}(t); \boldsymbol{\beta}_{k\ell r}^T \boldsymbol{x}_{ij}(t), \sigma_{k\ell r}^2 \right) \right]\end{aligned}$$

where $(h_{tr}; t = 1, \ldots, T_{ij}, r = 1, \ldots, S_{k\ell})$ is a binary variable indicating from which state the observation $y_{ij}(t)$ within the block cluster $k\ell$ is originated

Parameter estimation: EM not feasible

- The E-Step computes the expected complete-data log-likelihood, given the observed curves $(\boldsymbol{X},\boldsymbol{Y})$, and the current parameter estimation $\boldsymbol{\Psi}^{(q)}$

$$
\begin{aligned}
Q(\boldsymbol{\Psi},\boldsymbol{\Psi}^{(q)}) =& \mathbb{E}\left[\log L_c(\boldsymbol{\Psi})\middle|\boldsymbol{X},\boldsymbol{Y};\boldsymbol{\Psi}^{(q)}\right]\\
=&\sum_{i,k}\mathbb{P}(z_{ik}=1|\boldsymbol{y}_{ij},\boldsymbol{x}_{ij})\log\pi_k+\sum_{j,\ell}\mathbb{P}(w_{j\ell}=1|\boldsymbol{y}_{ij},\boldsymbol{x}_{ij})\log\rho_\ell\\
&+\sum_{i,j,k,\ell,t,r}\mathbb{P}(z_{ik}w_{j\ell}=1|\boldsymbol{y}_{ij},\boldsymbol{x}_{ij})\mathbb{P}(h_{tr}=1|z_{ik},w_{j\ell},y_{ij}(t),x_{ij}(t))\times\\
&\log\left[\alpha_{k\ell r}(t;\boldsymbol{\xi}_{k\ell})\mathcal{N}\left(y_{ij}(t);\boldsymbol{\beta}_{k\ell r}^T\boldsymbol{x}_{ij}(t),\sigma_{k\ell r}^2\right)\right]
\end{aligned}
$$

$\hookrightarrow$ Requires the calculation of the posterior joint distribution $\mathbb{P}(z_{ik}w_{j\ell}=1|\boldsymbol{y}_{ij},\boldsymbol{x}_{ij})$

$\hookrightarrow$ does not factorize due to the conditional dependence on the observed curves of the row and the column labels

$\Rightarrow$ [Govaert and Nadif, 2008, 2013] proposed a variational approximation by relying on the Neal and Hinton's interpretation of the EM algorithm [Neal and Hinton, 1998].

$\hookrightarrow$ We adopt this variational approximation in our context

Variational block EM algorithm

$$\mathbb{P}(z_{ik}w_{j\ell} = 1|\boldsymbol{y}_{ij}, \boldsymbol{x}_{ij}) \approx \mathbb{P}(z_{ik} = 1|\boldsymbol{y}_{ij}, \boldsymbol{x}_{ij}) \times \mathbb{P}(w_{j\ell} = 1|\boldsymbol{y}_{ij}, \boldsymbol{x}_{ij})$$

Variational block EM algorithm

$$\mathbb{P}(z_{ik}w_{j\ell} = 1|\boldsymbol{y}_{ij}, \boldsymbol{x}_{ij}) \approx \mathbb{P}(z_{ik} = 1|\boldsymbol{y}_{ij}, \boldsymbol{x}_{ij}) \times \mathbb{P}(w_{j\ell} = 1|\boldsymbol{y}_{ij}, \boldsymbol{x}_{ij})$$

Initialization: start from an initial solution at iteration $q = 0$, and then alternate at the $(q+1)$th iteration between the following variational E- and M- steps until convergence:

VE Step Estimate the variational approximated posterior memberships:

1. $\tilde{z}_{ik}^{(q+1)} \propto$
$\pi_k^{(q)} \exp\Big(\sum_{j,\ell,t,r} \tilde{w}_{j\ell}^{(q)}\ \tilde{h}_{tr}^{(q)} \log\Big[\alpha_{k\ell r}(t; \boldsymbol{\xi}_{k\ell}^{(q)})\mathcal{N}\left(y_{ij}(t); \boldsymbol{\beta}_{k\ell r}^{T^{(q)}} \boldsymbol{x}_{ij}(t), \sigma_{k\ell r}^{(q)^2}\right)\Big]\Big)$
2. $\tilde{w}_{j\ell}^{(q+1)} \propto$
$\rho_\ell^{(q)} \exp\Big(\sum_{i,k,t,r} \tilde{z}_{ik}^{(q)}\ \tilde{h}_{tr}^{(q)} \log\Big[\alpha_{k\ell r}(t; \boldsymbol{\xi}_{k\ell}^{(q)})\mathcal{N}\left(y_{ij}(t); \boldsymbol{\beta}_{k\ell r}^{T^{(q)}} \boldsymbol{x}_{ij}(t), \sigma_{k\ell r}^{(q)^2}\right)\Big]\Big)$
3. $\tilde{h}_{tr}^{(q+1)} \propto \alpha_{k\ell r}^{(q)}(t; \boldsymbol{\xi}_{k\ell}^{(q)})\mathcal{N}\left(y_{ij}(t); \boldsymbol{\beta}_{k\ell r}^{(q)^T} \boldsymbol{x}_{ij}(t), \sigma_{k\ell r}^{(q)^2}\right)$

where:

- $\tilde{z}_{ik} = \mathbb{P}(z_{ik} = 1|\boldsymbol{y}_{ij}, \boldsymbol{x}_{ij})$,
- $\tilde{w}_{j\ell} = \mathbb{P}(w_{j\ell} = 1|\boldsymbol{y}_{ij}, \boldsymbol{x}_{ij})$,
- $\tilde{h}_{tr} = \mathbb{P}(h_{tr} = 1|z_i, w_j, y_{ij}(t), x_{ij}(t))$

Variational block EM algorithm

M Step update the parameters estimates $\boldsymbol{\theta}^{(q+1)}$ given the estimated posterior memberships at the current iteration $q+1$:

1. $\pi_k^{(q+1)} = \frac{\sum_i \tilde{z}_{ik}^{(q+1)}}{n}$
2. $\rho_\ell^{(q+1)} = \frac{\sum_j \tilde{w}_{j\ell}^{(q+1)}}{d}$

Variational block EM algorithm

M Step update the parameters estimates $\boldsymbol{\theta}^{(q+1)}$ given the estimated posterior memberships at the current iteration $q+1$:

1. $\pi_k^{(q+1)} = \frac{\sum_i \tilde{z}_{ik}^{(q+1)}}{n}$
2. $\rho_\ell^{(q+1)} = \frac{\sum_j \tilde{w}_{j\ell}^{(q+1)}}{d}$

   The update of each block parameters $\boldsymbol{\theta}_{k\ell}$ consists in a weighted version of the RHLP updating rules:

3. $\boldsymbol{\xi}_{k\ell}^{(new)} = \boldsymbol{\xi}_{k\ell}^{(old)} - \left[\frac{\partial^2 F(\boldsymbol{\xi}_{k\ell})}{\partial \boldsymbol{\xi}_{k\ell} \partial \boldsymbol{\xi}_{k\ell}^T}\right]^{-1}_{\boldsymbol{\xi}_{k\ell}=\boldsymbol{\xi}_{k\ell}^{(old)}} \frac{\partial F(\boldsymbol{\xi}_{k\ell})}{\partial \boldsymbol{\xi}_{k\ell}}\Big|_{\boldsymbol{\xi}_{k\ell}=\boldsymbol{\xi}_{k\ell}^{(old)}}$ which is the IRLS maximisation of $F(\boldsymbol{\xi}_{k\ell}) = \sum_{i,j,t} \tilde{z}_{ik}^{(q)} \tilde{w}_{j\ell}^{(q)} \tilde{h}_{tr}^{(q)} \log \alpha_{k\ell r}(t; \boldsymbol{\xi}_{k\ell})$ w.r.t $\boldsymbol{\xi}_{k\ell}$.

Variational block EM algorithm

M Step update the parameters estimates $\boldsymbol{\theta}^{(q+1)}$ given the estimated posterior memberships at the current iteration $q+1$:

1. $\pi_k^{(q+1)} = \frac{\sum_i \tilde{z}_{ik}^{(q+1)}}{n}$
2. $\rho_\ell^{(q+1)} = \frac{\sum_j \tilde{w}_{j\ell}^{(q+1)}}{d}$

   The update of each block parameters $\boldsymbol{\theta}_{k\ell}$ consists in a weighted version of the RHLP updating rules:
3. $\boldsymbol{\xi}_{k\ell}^{(new)} = \boldsymbol{\xi}_{k\ell}^{(old)} - \left[\frac{\partial^2 F(\boldsymbol{\xi}_{k\ell})}{\partial \boldsymbol{\xi}_{k\ell} \partial \boldsymbol{\xi}_{k\ell}^T}\right]^{-1}_{\boldsymbol{\xi}_{k\ell}=\boldsymbol{\xi}_{k\ell}^{(old)}} \frac{\partial F(\boldsymbol{\xi}_{k\ell})}{\partial \boldsymbol{\xi}_{k\ell}}\Big|_{\boldsymbol{\xi}_{k\ell}=\boldsymbol{\xi}_{k\ell}^{(old)}}$ which is the IRLS maximisation of $F(\boldsymbol{\xi}_{k\ell}) = \sum_{i,j,t} \tilde{z}_{ik}^{(q)} \tilde{w}_{j\ell}^{(q)} \tilde{h}_{tr}^{(q)} \log \alpha_{k\ell r}(t; \boldsymbol{\xi}_{k\ell})$ w.r.t $\boldsymbol{\xi}_{k\ell}$.

   The regression parameters updates consist in analytic WLS problems:
4. $\boldsymbol{\beta}_{k\ell r}^{(q+1)} = \left[\sum_{i,j} \tilde{z}_{ik}^{(q)} \tilde{w}_{j\ell}^{(q)} \mathbf{X}_{ij}^T \boldsymbol{\Lambda}_{ijkr}^{(q)} \mathbf{X}_{ij}\right]^{-1} \sum_{i,j} \tilde{z}_{ik}^{(q)} \tilde{w}_{j\ell}^{(q)} \mathbf{X}_{ij}^T \boldsymbol{\Lambda}_{ijkr}^{(q)} \boldsymbol{y}_{ij}$
5. $\sigma_{k\ell r}^{2(q+1)} = \frac{\sum_{i,j} \tilde{z}_{ik}^{(q)} \tilde{w}_{j\ell}^{(q)} \|\sqrt{\boldsymbol{\Lambda}_{ijkr}^{(q)}}(\boldsymbol{y}_{ij} - \mathbf{X}_{ij}\boldsymbol{\beta}_{kr}^{(q+1)})\|^2}{\sum_{i,j} \tilde{z}_{ik}^{(q)} \tilde{w}_{j\ell}^{(q)} \ \mathrm{trace}(\boldsymbol{\Lambda}_{ijkr}^{(q)})}$ where $\mathbf{X}_{ij}$ is the design matrix for the ith curve, $\boldsymbol{\Lambda}_{ijkr}^{(q)}$ is the diagonal matrix whose diagonal elements are the posterior segment memberships $\{\tilde{h}_{ijtr}^{(q)}; t = 1, \ldots, T_{ij}\}$.

$\hookrightarrow$ It is also possible to use the Classification EM (CEM) approximation of EM [Celeux and Govaert, 1992].

Parameter estimation by an SEM algorithm: SEM-FLBM

- $\hookrightarrow$ The SEM algorithm [Celeux and Diebolt, 1985] allows to overcome some drawbacks of the variational-EM algorithm, including its sensitivity to starting values; SEM does not use an approximation.
- Eg. SEM for latent block models for categorical data [Keribin et al., 2012, 2014]
- The formulas of VEM-FLBM and SEM-FLBM are essentially the same, except that we incorporate a stochastic step consisting of sampling binary indicator variables z_{ik}, $w_{j\ell}$ and h_{tr} according to $\tilde{z}_{ik}$, $\tilde{w}_{j\ell}$ and $\tilde{h}_{tr}$.

Conclusion and perspectives

Conclusion

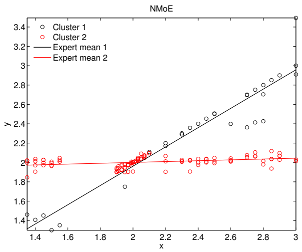
- A full generative framework for the cluster analysis and segmentation of high-dimensional non-stationary functional data
- The model inference can be performed by a variational EM algorithm or SEM

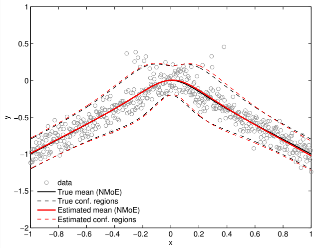
Perspectives

- Numerical experiments
- Package

Outline

Context





- Heterogeneous regression data $(x, y) \hookrightarrow$ underlying unknown partition $\mathbf{z}$
- Data issued from non-linear regression function $f(y|x)$

Modeling framework

Mixture-of-experts/(parametric) mixture of regressions with predictor-dependent mixing proportions :

$$p(y_i|\boldsymbol{x}_i) \quad = \quad \sum_{z_i} \mathbb{P}(z_i|\boldsymbol{x}_i) p(y_i|\boldsymbol{x}_i, z_i),$$

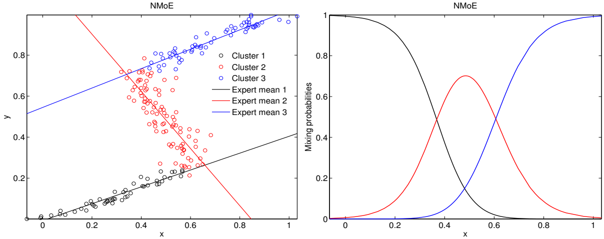
Mixture-of-Experts (MoE) modeling framework

- Observed pairs of data $(\boldsymbol{x}, y)$ where the response $y \in \mathbb{R}$ for the predictors $\boldsymbol{x} \in \mathbb{R}^p$ governed by a hidden categorical random variable Z
- Mixture of experts (MoE) [Jacobs et al., 1991; Jordan and Jacobs, 1994] :

$$f(y|\boldsymbol{x};\boldsymbol{\theta}) = \sum_{k=1}^{K} \underbrace{\pi_k(\boldsymbol{x};\mathbf{w})}_{\text{Gating network}} \underbrace{f_k(y|\boldsymbol{x};\boldsymbol{\theta}_k)}_{\text{Expert Network}}$$

- Gating network (e.g softmax): $\pi_k(\boldsymbol{x};\mathbf{w}) = \frac{\exp(w_{k0}+\boldsymbol{w}_k^T\boldsymbol{x})}{1+\sum_{\ell=1}^{K-1}\exp(w_{\ell 0}+\boldsymbol{w}_\ell^T\boldsymbol{x})}$
- Experts network (e.g Gaussian regressors): $f_k(y|\boldsymbol{x};\boldsymbol{\theta}_k) = \phi\big(y;\mu(\boldsymbol{x};\boldsymbol{\beta}_k),\sigma_k^2\big)$ with parametric (non-)linear regression functions $\mu(\boldsymbol{x};\boldsymbol{\beta}_k)$
- Non-normal MoE, for data with atypical observations, and with possible heavy tailed and asymmetric distributions: [Chamroukhi, 2016a, 2017; Nguyen and Chamroukhi, 2018]
- parameter vector $\boldsymbol{\theta} = (\mathbf{w}^T, \boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_K^T)^T$

Illustration



Standard MLE of the MoE model

- MLE: $\boldsymbol{\theta}$ is commonly estimated by maximizing the observed-data log-likelihood:

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

with

$L(\boldsymbol{\theta}) = \ln f((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_1); \boldsymbol{\theta}) = \sum_{i=1}^{n} \ln \sum_{k=1}^{K} \pi_k(\boldsymbol{x}_i; \boldsymbol{w}) f(\boldsymbol{y}_i|\boldsymbol{x}_i; \boldsymbol{\theta}_k)$.

$\hookrightarrow$ the EM algorithm (Dempster et al. [1977])

Standard MLE of the MoE model

- MLE: $\boldsymbol{\theta}$ is commonly estimated by maximizing the observed-data log-likelihood:

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

with

$L(\boldsymbol{\theta}) = \ln f((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_1); \boldsymbol{\theta}) = \sum_{i=1}^{n} \ln \sum_{k=1}^{K} \pi_k(\boldsymbol{x}_i; \boldsymbol{w}) f(\boldsymbol{y}_i|\boldsymbol{x}_i; \boldsymbol{\theta}_k)$.

$\hookrightarrow$ the EM algorithm (Dempster et al. [1977])

$\hookrightarrow$ Consider a high-dimensional setting

$\hookrightarrow$ Looking for a sparse models

Regularized MLE of the MoE

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

$$\widehat{\boldsymbol{\theta}}_n \in \arg\max_{\boldsymbol{\theta}\in\Theta} PL(\boldsymbol{\theta})$$

with $PL(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) - \text{Pen}(\boldsymbol{\theta})$

- $\hookrightarrow$ $\text{Pen}(\boldsymbol{\theta})$ should encourage sparsity
- parameter estimation and selection problem

Proposed Regularized Mixture of Experts model

$$\mathrm{Pen}(\boldsymbol{\theta}) = \underbrace{\sum_{k=1}^{K} \lambda_k \|\boldsymbol{\beta}_k\|_1}_{\text{Lasso-like pen.}} + \underbrace{\sum_{k=1}^{K-1} \gamma_k \|\boldsymbol{w}_k\|_1 + \frac{\rho}{2} \|\boldsymbol{w}_k\|_2^2}_{\text{Elastic-Net like pen.}}$$

- Lasso penalty for the experts $\hookrightarrow$ encourage a sparse solution
- The elastic net penalty (Zou and Hastie [2005]) for the gating network:
  $\hookrightarrow$ reduce the norm of the estimated values of the gating network parameters by using the L_2 penalties;
  $\hookrightarrow$ the Lasso penalty to recover a sparse solution
- The convexity of L_1 and L_2 penalties have also advantageous numerical properties.
- If the correlation between the features is high, one can add L_2 penalties for the expert network.

Regularized MLE via an EM algorithm

- The penalized log-likelihood function:

$$PL(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) - \sum_{k=1}^{K} \lambda_k \|\boldsymbol{\beta}_k\|_1 - \sum_{k=1}^{K-1} \gamma_k \|\boldsymbol{w}_k\|_1 - \frac{\rho}{2}\|\boldsymbol{w}_k\|_2^2$$

- The penalized complete-data log-likelihood function:

$$PL_c(\boldsymbol{\theta}) = L_c(\boldsymbol{\theta}) - \sum_{k=1}^{K} \lambda_k \|\boldsymbol{\beta}_k\|_1 - \sum_{k=1}^{K-1} \gamma_k \|\boldsymbol{w}_k\|_1 - \frac{\rho}{2}\|\boldsymbol{w}_k\|_2^2$$

with

$$L_c(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} \log\left[\pi_k(\boldsymbol{x}_i;\boldsymbol{w}) f(\boldsymbol{y}_i|\boldsymbol{x}_i;\boldsymbol{\theta}_k)\right]$$

such that $z_{ik} = 1$ iff $z_i = k$ (the data pair $(\boldsymbol{x}_i, y_i)$ originates from expert k

Parameter estimation for RMoE

Khalili's method [Khalili, 2010]:

- Approximates the L_1 penalty function in a some neighborhood by an ε -local quadratic function

$$\eta|t| \approx \eta|t_0| + \frac{\eta}{2(|t_0| + \varepsilon)}(t^2 - t_0^2).$$

  $\hookrightarrow$ Almost surely none of the components will be exactly zero.

- Needs using a threshold to recover the zero coefficients
  $\hookrightarrow$ The size of threshold affects the degree of sparsity of the solution.

- The Newton-Raphson algorithm is used to update the M-step of the EM algorithm.
  $\hookrightarrow$ This approach still require computing the inverse matrix.

In our proposal:

- A block EM algorithm with coordinate ascent algorithm to estimate the parameters:
  $\hookrightarrow$ Exact L_1 penalty regularization;
  $\hookrightarrow$ Avoids computing matrix inversion;
  $\hookrightarrow$ Avoids using a threshold to recover the zero coefficients.

Block EM algorithm with coordinate ascent

E-step

- Compute the conditional expectation of the penalized complete-data log-likelihood

$$\begin{aligned}
Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(q)}) &= \mathbb{E}\left[PL_c(\boldsymbol{\theta})|\mathcal{D};\boldsymbol{\theta}^{(q)}\right]\\
&= \sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}^{(q)}\log\left[\pi_k(\boldsymbol{x}_i;\boldsymbol{w})f_k(\boldsymbol{y}_i|\boldsymbol{x}_i;\boldsymbol{\theta}_k)\right]\\
&\quad -\sum_{k=1}^{K}\lambda_k\|\boldsymbol{\beta}_k\|_1-\sum_{k=1}^{K-1}(\gamma_k\|\boldsymbol{w}_k\|_1-\frac{\rho}{2}\|\boldsymbol{w}_k\|_2^2).
\end{aligned}$$

$\hookrightarrow$ Calculate the posterior component probabilities:

$$\tau_{ik}^{(q)}=\mathbb{P}(Z_i=k|\boldsymbol{y}_i,\boldsymbol{x}_i;\boldsymbol{\theta}^{(q)})=\frac{\pi_k(\boldsymbol{x}_i;\boldsymbol{w}^{(q)})\mathcal{N}(y_i;\beta_{k0}^{(q)}+\boldsymbol{x}_i^T\boldsymbol{\beta}_k^{(q)},\sigma_k^{(q)2})}{\sum_{l=1}^{K}\pi_l(\boldsymbol{x}_i;\boldsymbol{w}^{(q)})\mathcal{N}(y_i;\beta_{l0}^{(q)}+\boldsymbol{x}_i^T\boldsymbol{\beta}_l^{(q)},\sigma_l^{(q)2})}.$$

$\hookrightarrow$ As in standard MoE

Block EM algorithm with coordinate ascent (cont.)

M-step

- Maximizing the Q function: $\boldsymbol{\theta}^{(q+1)} \in \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(q)})$ with

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(q)}) = Q(\boldsymbol{w}; \boldsymbol{\theta}^{(q)}) + Q(\boldsymbol{\beta}, \sigma; \boldsymbol{\theta}^{(q)}),$$

where

$$Q(\boldsymbol{w}; \boldsymbol{\theta}^{(q)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(q)} \log \pi_k(\boldsymbol{x}_i; \boldsymbol{w}) - \sum_{k=1}^{K-1} (\gamma_k \|\boldsymbol{w}_k\|_1 - \frac{\rho}{2} \|\boldsymbol{w}_k\|_2^2), \quad (1)$$

$\hookrightarrow$ a weighted regularized multiclass logistic regression problem

and

$$Q(\boldsymbol{\beta}, \sigma; \boldsymbol{\theta}^{(q)}) = \sum_{k=1}^{K} \sum_{i=1}^{n} \tau_{ik}^{(q)} \log \mathcal{N}(y_i; \beta_{k0} + \boldsymbol{x}_i^T \boldsymbol{\beta}_k, \sigma_k^2) - \sum_{k=1}^{K} \lambda_k \|\boldsymbol{\beta}_k\|_1 \quad (2)$$

$\hookrightarrow$ K independent weighted LASSO problems

Updating the gating network parameters

- Coordinate ascent algorithm to update $\boldsymbol{w}$ Tseng [1988, 2001]
- w_{kj} is updated by maximizing the component (k, j) of (1) given by

$$Q(w_{kj}; \boldsymbol{\theta}^{(q)}) = \begin{cases} F(w_{kj}; \boldsymbol{\theta}^{(q)}) - \gamma_k w_{kj} & , \text{ if } w_{kj} > 0 \quad (F_1) \\ F(0; \boldsymbol{\theta}^{(q)}) & , \text{ if } w_{kj} = 0 \\ F(w_{kj}; \boldsymbol{\theta}^{(q)}) + \gamma_k w_{kj} & , \text{ if } w_{kj} < 0 \quad (F_2) \end{cases},$$

$$F(w_{kj}; \boldsymbol{\theta}^{(q)}) = \sum_{i=1}^{n} \tau_{ik}^{(q)} (w_{k0} + \boldsymbol{w}_k^T \boldsymbol{x}_i) - \sum_{i=1}^{n} \log\Big(1 + \sum_{l=1}^{K-1} e^{w_{l0} + \boldsymbol{w}_l^T \boldsymbol{x}_i}\Big) - \frac{\rho}{2} w_{kj}^2. \quad (3)$$

Univariate Newton-Raphson algorithm

- F_1 and F_2 are smooth univariate concave functions in w_{kj}. $\hookrightarrow$ Univariate Newton-Raphson algorithm can be used to update w_{kj}

$$w_{kj}^{(s+1)} = w_{kj}^{(s)} - \left(\frac{\partial^2 F(w_{kj}; \boldsymbol{\theta}^{(q)})}{\partial^2 w_{kj}}\right)^{-1}\Bigg|_{w_{kj}^{(s)}} \left(\frac{\partial F(w_{kj}; \boldsymbol{\theta}^{(q)})}{\partial w_{kj}} - \gamma_k \text{sign}(w_{kj})\right)\Bigg|_{w_{kj}^{(s)}},$$

where $\frac{\partial^2 F(w_{kj};\boldsymbol{\theta}^{(q)})}{\partial^2 w_{kj}}$ and $\frac{\partial F(w_{kj};\boldsymbol{\theta}^{(q)})}{\partial w_{kj}}$ have closed-form.

Updating the expert parameters

M-step (cont.)

- Update β_{kj} using coordinate ascent algorithm with soft-thresholding operator

$$\beta_{kj}^{[s+1]} = \mathcal{S}_{\lambda_k \sigma_k^{(q)2}}\Big(\sum_{i=1}^{n} \tau_{ik}^{(q)} r_{ikj}^{[s]} x_{ij}\Big) \Big/ \sum_{i=1}^{n} \tau_{ik}^{(q)} x_{ij}^2,$$

where $r_{ikj}^{[s]} = y_i - \beta_{k0}^{[s]} - \boldsymbol{\beta}_k^{[s]T} \boldsymbol{x}_i + \beta_{kj}^{[s]} x_{ij}$, $[\mathcal{S}_\gamma(u)]_j = \mathsf{sign}(u_j)(|u_j| - \gamma)_+$ and $(x)_+ = \max\{x, 0\}$ in the sth loop of the coordinate ascent algorithm.

$$\beta_{k0}^{[s+1]} = \sum_{i=1}^{n} \tau_{ik}^{(q)} (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_k^{[s+1]}) \Big/ \sum_{i=1}^{n} \tau_{ik}^{(q)}.$$

Updating the expert parameters

M-step (cont.)

- Update β_{kj} using coordinate ascent algorithm with soft-thresholding operator

$$\beta_{kj}^{[s+1]} = \mathcal{S}_{\lambda_k \sigma_k^{(q)2}} \Big(\sum_{i=1}^{n} \tau_{ik}^{(q)} r_{ikj}^{[s]} x_{ij}\Big) \Big/ \sum_{i=1}^{n} \tau_{ik}^{(q)} x_{ij}^2,$$

where $r_{ikj}^{[s]} = y_i - \beta_{k0}^{[s]} - \boldsymbol{\beta}_k^{[s]T} \boldsymbol{x}_i + \beta_{kj}^{[s]} x_{ij}$, $[\mathcal{S}_\gamma(u)]_j = \mathsf{sign}(u_j)(|u_j| - \gamma)_+$ and $(x)_+ = \max\{x, 0\}$ in the sth loop of the coordinate ascent algorithm.

$$\beta_{k0}^{[s+1]} = \sum_{i=1}^{n} \tau_{ik}^{(q)} (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_k^{[s+1]}) \Big/ \sum_{i=1}^{n} \tau_{ik}^{(q)}.$$

- Rerun the E-step, keep

$$(w_{k0}^{(q+2)}, \boldsymbol{w}_k^{(q+2)}) = (w_{k0}^{(q+1)}, \boldsymbol{w}_k^{(q+1)});\ (\beta_{k0}^{(q+2)}, \boldsymbol{\beta}_k^{(q+2)}) = (\beta_{k0}^{(q+1)}, \boldsymbol{\beta}_k^{(q+1)}),$$

and update $\sigma_k^{2(q+2)}$ as follows

$$\sigma_k^{2(q+2)} = \sum_{i=1}^{n} \tau_{ik}^{(q+1)} (y_i - \beta_{k0}^{(q+2)} - \boldsymbol{x}_i^\top \boldsymbol{\beta}_k^{(q+2)})^2 \Big/ \sum_{i=1}^{n} \tau_{ik}^{(q+1)}.$$

Simulation study

Simulation protocol

- $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}; \boldsymbol{\Sigma})$ with corr$(x_{ij}, x_{ij'}) = 0.5^{|j-j'|}$; $K = 2$
- Sample size: $n = 300$, 100 different data sets;
- The regression coefficients:

$$\begin{aligned}(\beta_{10}, \boldsymbol{\beta}_1)^T &= (0, 0, 1.5, 0, 0, 0, 1)^T; \sigma_1 = 1\\ (\beta_{20}, \boldsymbol{\beta}_2)^T &= (0, 1, -1.5, 0, 0, 2, 0)^T; \sigma_2 = 1\\ (w_{10}, \boldsymbol{w}_1)^T &= (1, 2, 0, 0, -1, 0, 0)^T; \sigma_3 = 1\end{aligned}$$

Considered approaches for comparison

- The standard MoE;
- MoE+L_2 (MoE with L_2 penalties in the gating network);
- MoE-BIC (MoE with model selection using BIC criterion - 100 submodels);
- MIXLASSO (MLR with Lasso penalties) (see Khalili and Chen [2007]);

Evaluation criteria

- The sensitivity/specificity (sparsity);
- The parameter estimation (density estimation);
- The misclassification error: Adjust rand index - ARI (clustering).

Sensitivity/specificity result

- *Sensitivity (S_1):* proportion of correctly estimated zero coefficients;
- *Specificity (S_2):* proportion of correctly estimated nonzero coefficients.

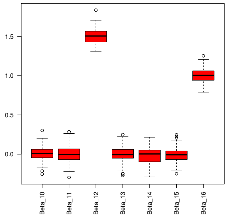
Method	Expert 1		Expert 2		Gate	
	S_1	S_2	S_1	S_2	S_1	S_2
MoE	0.000	1.000	0.000	1.000	0.000	1.000
MoE+L_2	0.000	1.000	0.000	1.000	0.000	1.000
MoE-BIC	0.920	1.000	0.930	1.000	0.850	1.000
MIXLASSO	0.775	1.000	0.693	1.000	N/A	N/A
Our MoE-Lasso+L_2	0.700	1.000	0.803	1.000	0.853	0.945

Table: Sensitivity (S_1) and specificity (S_2) results.

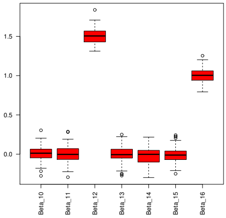
- MoE and MoE+L_2 could not be considered as model selection methods since their sensitivity equal zero.
- MIXLASSO can detect the zero coefficients in the experts. However, this model has a poor result when clustering the data.
- The MoE-Lasso+L_2 model can detect the zero coefficients in the experts and the gating network.

Parameter estimation for expert 1

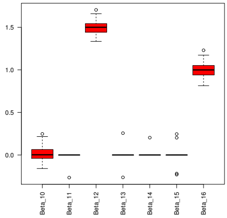
- $(\beta_{10}, \boldsymbol{\beta}_1)^T = (0, 0, 1.5, 0, 0, 0, 1)^T$.



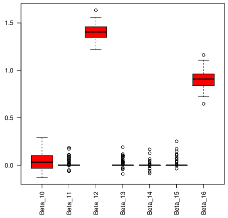
MoE



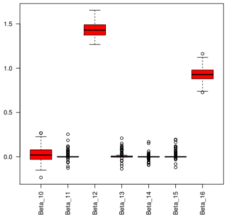
MoE-L_2



MoE-BIC



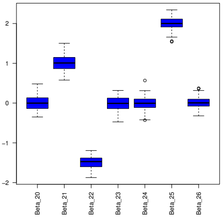
MIXLASSO



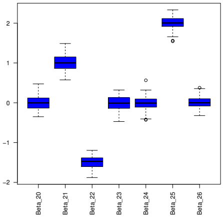
MoE-Lasso + L_2

Parameter estimation for expert 2

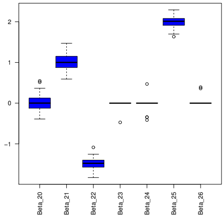
- $(\beta_{20}, \boldsymbol{\beta}_2)^T = (0, 1, -1.5, 0, 0, 2, 0)^T$.



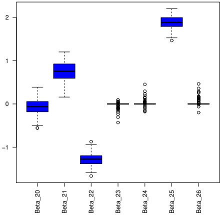
MoE



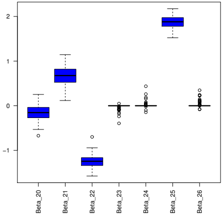
MoE-L_2



MoE-BIC



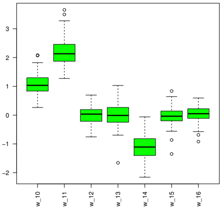
MIXLASSO



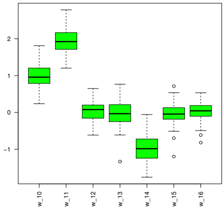
MoE-Lasso + L_2

Parameter estimation for gating network

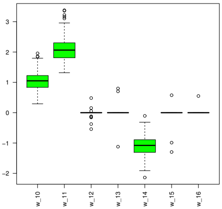
- $(w_{10}, \boldsymbol{w}_1)^T = (1, 2, 0, 0, -1, 0, 0)^T.$



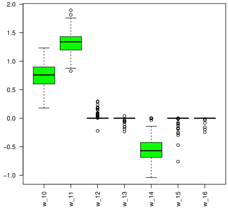
MoE



MoE-L_2



MoE-BIC



MoE-Lasso + L_2

Result for data clustering

Model	MoE	MoE+L_2	MoE-BIC	**MoE-Lasso + L_2**	MIXLASSO
C. rate	$89.57\%_{(1.65\%)}$	$89.62\%_{(1.63\%)}$	$90.05\%_{(1.65\%)}$	$89.46\%_{(1.76\%)}$	$82.89\%_{(1.92\%)}$

Table: clustering accuracy results (correct classification rate and Adjusted Rand Index).

Remarks

- MoE-BIC provides the best results. However, it is hard to apply BIC in reality especially for high dimensional data, since this involves a huge collection of model candidates.
- MIXLASSO can detect zero coefficients in the experts, but it provides a poor result when clustering data.
- MoE-Lasso+L_2 can detect zero coefficients in the model and provide a competitive result with MoE, MoE-L_2 in term of clustering, although it also causes bias to the non-zero coefficients.

Applications to real data sets

- For real data sets, we calculate the mean squared error between the response variable Y with its prediction $\widehat{Y}$, where

$$\widehat{Y} = \sum_{k=1}^{K} \pi_k(\boldsymbol{x}; \widehat{\boldsymbol{w}})(\widehat{\beta}_{k0} + \boldsymbol{x}^T \widehat{\boldsymbol{\beta}}_k).$$

- Housing data: 13 features, 506 observations, $K = 2$.

	MoE	MoE-Lasso+L_2 (Khalili)	**MoE-Lasso + L_2**
MSE	$0.1544_{(.577)}$	$0.2044_{(.709)}$	$0.1989_{(.619)}$

Table: Results for Housing data set.

- Baseball salary data: 32 features, 337 observations, $K = 2$.

	MoE	**MoE-Lasso + L_2**	MIXLASSO
MSE	$0.2625_{(.758)}$	$0.2821_{(.633)}$	$1.1858_{(2.792)}$

Table: Results for Baseball salaries data set.

The proximal Newton method

- We recently improve the proposed algorithm by using the proximal Newton method (Lee et al. [2006], Lee et al. [2014] and Friedman et al. [2010]) for updating the gating network parameters.
- The idea of the proximal Newton method:

  • Approximate the smooth part of $Q(\boldsymbol{w};\boldsymbol{\theta}^{(q)})$ with its local quadratic form;

  • Use coordinate ascent with soft-thresholding operator to solve the resulting approximated convex optimization problem;

  • Combine with backtracking line search to update $\boldsymbol{w}$.

Extension result for proximal Newton method

- Coordinate ascent algorithm (CA) VS proximal Newton (PN) method:

Criteria	**MoE-Lasso** + L_2 (CA)	**MoE-Lasso** + L_2 (PN)
C.Rate	$89.46\%_{(1.76\%)}$	$89.53\%_{(1.65\%)}$
$PL(\boldsymbol{\theta})$ value	$-558.140_{(12.99)}$	$-558.410_{(13.03)}$

Table: Simulation results.

- Application of the proximal Newton algorithm to the residential building data set: 107 features, 372 observations, $K = 3$.

Proximal Newton	$0.0120_{(.879)}$

Table: Results for residential building data set.

Conclusion and perspectives

Conclusion

- We propose a regularized MoE which does not require using approximations as in standard MoE regularization
- A blockwise EM algorithm with coordinate ascent algorithm is proposed to monotonically maximize the RMoE objective function
- The updating of the gating network for some situations is time consuming since we don't have a closed-form
- The algorithm has been improved by using proximal Newton method to update the gating network, which has a closed-form update for each parameter and improve the running time
- Future work: Estimation and feature selection for hierarchical MoE and MoE with discrete data, ...

References I

Y. Ben Slimen, S. Allio, and J. Jacques. Model-Based Co-clustering for Functional Data. HAL preprint hal-01422756, December 2016. URL `https://hal.inria.fr/hal-01422756`.

C. Bouveyron and J. Jacques. Model-based clustering of time series in group-specific functional subspaces. *Adv. Data Analysis and Classification*, 5(4):281–300, 2011.

C. Bouveyron, L. Bozzi, J. Jacques, and F.-X. Jollois. The functional latent block model for the co-clustering of electricity consumption curves. *Journal of the Royal Statistical Society, Series C*, 2018.

G. Celeux and J. Diebolt. The SEM algorithm a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly*, 2(1):73–82, 1985.

G. Celeux and G. Govaert. A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, 14:315–332, 1992.

F. Chamroukhi. Robust mixture of experts modeling using the t-distribution. *Neural Networks - Elsevier*, 79:20–36, 2016a. URL `https://chamroukhi.users.lmno.cnrs.fr/papers/TMoE.pdf`.

F. Chamroukhi. Skew t mixture of experts. *Neurocomputing - Elsevier*, 266:390–408, 2017. URL `https://chamroukhi.users.lmno.cnrs.fr/papers/STMoE.pdf`.

F. Chamroukhi and C. Biernacki. Model-Based Co-Clustering of Multivariate Functional Data. In *ISI 2017 - 61st World Statistics Congress*, Marrakech, Morocco, Jul 2017. URL `https://hal.archives-ouvertes.fr/hal-01653782`.

F. Chamroukhi, A. Samé, G. Govaert, and P. Aknin. Time series modeling by a regression approach based on a latent process. *Neural Networks*, 22(5-6):593–602, 2009. PDF.

F. Chamroukhi, A. Samé, G. Govaert, and P. Aknin. A hidden process regression model for functional data description. application to curve discrimination. *Neurocomputing*, 73(7-9):1210–1221, 2010. URL `https://chamroukhi.users.lmno.cnrs.fr/papers/chamroukhi_neucomp_2010.pdf`.

Faicel Chamroukhi. Piecewise regression mixture for simultaneous functional data clustering and optimal segmentation. *Journal of Classification*, 33(3):374–411, 2016b. URL `https://chamroukhi.users.lmno.cnrs.fr/papers/Chamroukhi-PWRM-JournalClassif-2016.pdf`.

References II

Faicel Chamroukhi and Hien D. Nguyen. Model-based clustering and classification of functional data. 2018. URL https://chamroukhi.users.lmno.cnrs.fr/papers/MBCC-FDA.pdf. arXiv:1803.00276v2.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of The Royal Statistical Society, B*, 39(1):1–38, 1977.

F. Ferraty and P. Vieu. Curves discrimination: a nonparametric functional approach. *Computational Statistics & Data Analysis*, 44(1-2):161–173, 2003.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

G. Govaert and M. Nadif. Clustering with block mixture models. *Pattern Recognition*, 36(2):463 – 473, 2003. Biometrics.

G. Govaert and M. Nadif. Fuzzy clustering to estimate the parameters of block mixture models. *Soft Computing*, 10(5): 415–422, 2006.

G. Govaert and M. Nadif. Block clustering with Bernoulli mixture models: Comparison of different approaches. *Computational Statistics and Data Analysis*, 52(6):3233 –3245, 2008.

G. Govaert and M. Nadif. *Co-Clustering*. Computer engineering series. Wiley-ISTE, November 2013. 256 pages.

G. Hébrail, B. Hugueney, Y. Lechevallier, and F. Rossi. Exploratory analysis of functional data via clustering and optimal segmentation. *Neurocomputing*, 73(7-9):1125–1141, March 2010.

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

Julien Jacques and Cristian Preda. Functional data clustering: A survey. *Adv. Data Anal. Classif.*, 8(3):231–255, September 2014. ISSN 1862-5347. doi: 10.1007/s11634-013-0158-y. URL http://dx.doi.org/10.1007/s11634-013-0158-y.

G. M. James and T. J. Hastie. Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society Series B*, 63:533–550, 2001.

G. M. James and C. Sugar. Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, 98 (462), 2003.

References III

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

C. Keribin, V. Brault, G. Celeux, and G. Govaert. Model selection for the binary latent block model. In *Proceedings of COMPSTAT*, 2012.

C. Keribin, V. Brault, G. Celeux, and G. Govaert. Estimation and selection for the latent block model on categorical data. *Statistics and Computing*, pages 1–16, 2014. ISSN 0960-3174. doi: 10.1007/s11222-014-9472-2. URL `http://dx.doi.org/10.1007/s11222-014-9472-2`.

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

A. Khalili and J. Chen. Variable selection in finite mixture of regression models. *Journal of the American Statistical association*, 102(479):1025–1038, 2007.

Jason D Lee, Yuekai Sun, and Michael A Saunders. Proximal newton-type methods for minimizing composite functions. *SIAM Journal on Optimization*, 24(3):1420–1443, 2014.

Su-In Lee, Honglak Lee, Pieter Abbeel, and Andrew Y Ng. Efficient l_1 regularized logistic regression. In *AAAI*, volume 6, pages 401–408, 2006.

A. Lomet. *Sélection de modèle pour la classification croisée de données continues*. Ph.D. thesis, Université de Technologie de Compiègne, 2012.

R. Neal and G. E. Hinton. *A view of the EM algorithm that justifies incremental, sparse, and other variants*, pages 355–368. Dordrecht: Kluwer Academic Publishers, 1998.

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

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

A. Samé, F. Chamroukhi, G. Govaert, and P. Aknin. Model-based clustering and segmentation of time series with changes in regime. *Advances in Data Analysis and Classification*, pages 1–21, 2011. ISSN 1862-5347.

References IV

P. Tseng. Coordinate ascent for maximizing nondifferentiable concave functions. 1988.

P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494, 2001.

DS Young and DR Hunter. Mixtures of regressions with predictor-dependent mixing proportions. *Computational Statistics and Data Analysis*, 55(10):2253–2266, 2010.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

Thank you for your attention!