Statistical learning of latent data models for complex data analysis

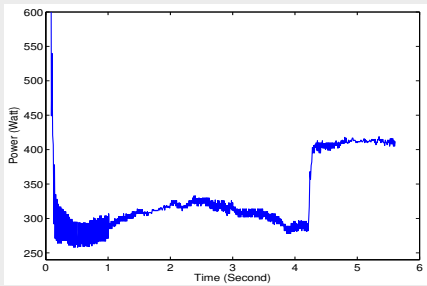
Faicel Chamroukhi

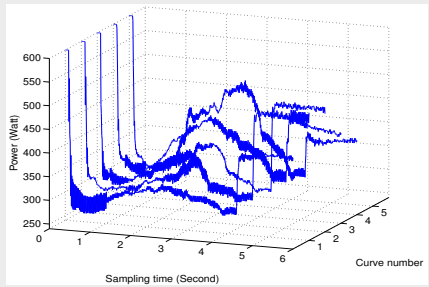
Laboratoire LSIS, UMR CNRS 7296
Université de Toulon

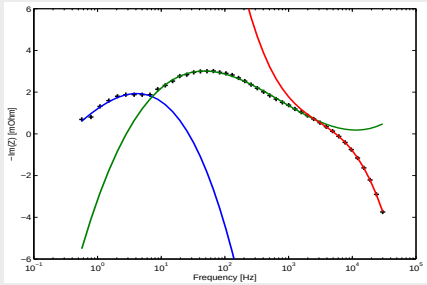
Sminaire du Laboratoire
Paul Painlevé, UMR CNRS 8524

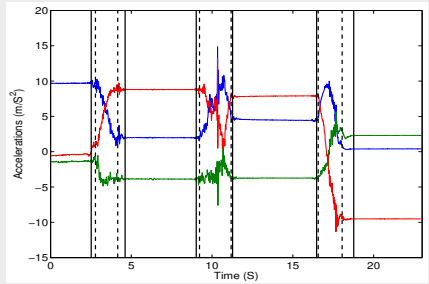
Mercredi 04 novembre 2015

- Temporal data with regime changes







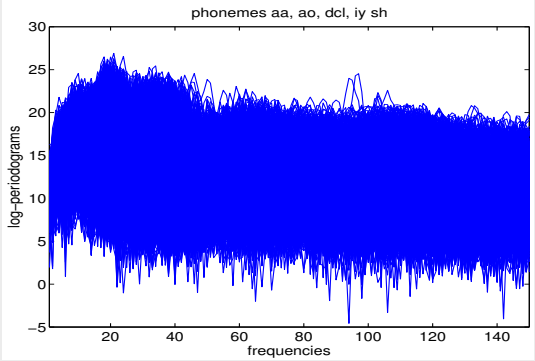


- Several regimes over time $\Rightarrow$ Abrupt and/or smooth regime changes
- Multidimensional temporal data

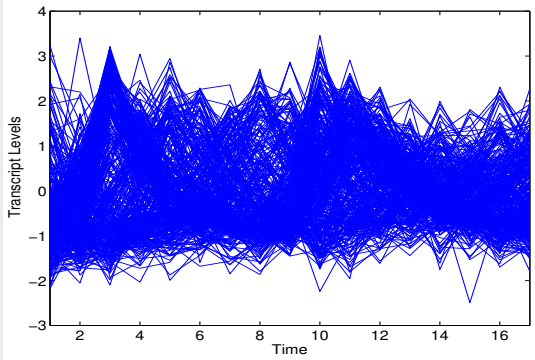
Objective

Temporal data modeling and segmentation

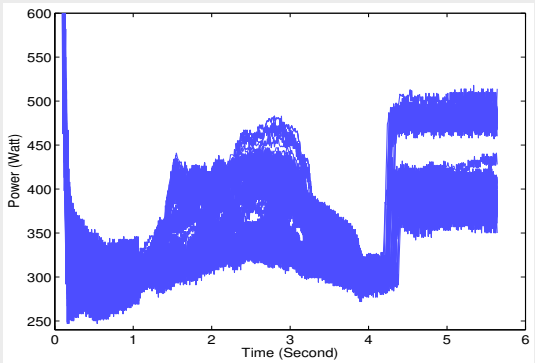
- Many curves to analyze



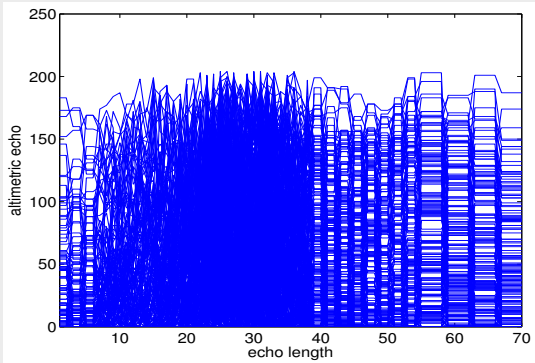
Phonemes curves



Yeast cell cycle curves



Railway switch waveforms



Satellite waveforms

Objectives

- Curve classification/clustering (functional data analysis framework)
- Deal with the problem of regime changes

- Data with possible atypical observations, skewed

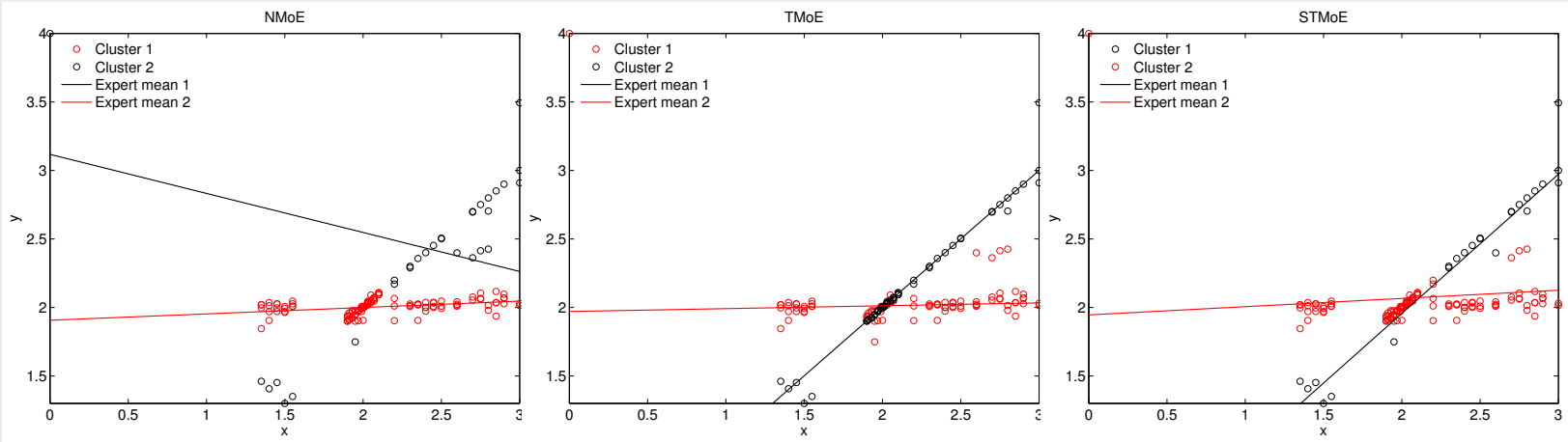


Figure: Fitting MoLE to the tone data set with ten outliers $(0, 4)$.

Objectives

- Derive robust models to fit at best the data and deal with possible features like skewness, heavy tails

Outline

Outline

Latent data models for temporal data segmentation

- This first research theme concerns the modeling and segmentation of complex temporal data, univariate and multivariate, and directly follows some of work I developed during my PhD thesis.
- This research axis can be organized into two sub-axes which are developed in what follows.
  1. Latent process regression models for univariate time series [J-1] [J,2] [J-3]
  2. Latent data models for dealing with the joint segmentation of multivariate time series [J-6][J-7][J-15].
     This main part initiated in 2010 was conducted in the framework of the PhD thesis of Dorra Trabelsi[1]

[1] D. Trabelsi. *Contribution à la reconnaissance non-intrusive d'activités humaines.* Ph.D. thesis, Université Paris-Est Créteil, Laboratoire Images, Signaux et Systèmes Intelligents (LiSSi), June 2013

Context

- In many application domains of data analysis, the data to be analyzed are presented as time series (also called signals, curves, etc).
- Time series analysis is a popular problem with a broad literature, and is studied by several scientific communities, including statistics, (statistical) signal processing, economics as well as statistical learning.
- In this study, we particularly focus on complex non-stationary time series that present non-linearities through various regime changes.
- Objectives: approximation, representation, summarizing model for prediction, segmentation for further categorization, etc.
- The problem of time series analysis is in reformulated into a time series segmentation problem
- The general problem of time series segmentation is common for different communities, including statistics, detection, signal processing, and machine learning.

- Here I mainly consider the framework of statistical learning of latent data models.
- Mixture models (Titterington et al., 1985; McLachlan and Peel., 2000; Frühwirth-Schnatter, 2006) and hidden Markov models (HMMs) (Rabiner and Juang, 1986; Rabiner, 1989; Frühwirth-Schnatter, 2006) are two well-known widely used examples of such models.
- In this framework of regime changing time series, it is natural to think that the observed time series is generated by an underlying stochastic process, with several possibly parametric states.
  $\Rightarrow$ The problem of time series modeling and segmentation therefore becomes the one of recovering the underling process and inferring the statistical parameters of each of its states.
- Classical approaches particularly concern abrupt change point detection $\Rightarrow$ Hence, if the regime changes may be smooth and/or abrupt, the piecewise regression model and the HMM based models, are not appropriate to provide smooth approximations.

Contribution

- Conventional solutions are subject to limitations in the control of the transitions between these states, leading to a non-smooth approximation.
- One can force the resulting approximation to be regular, but this leads to combinatorial optimization problems for the choice of the change points.
- Relaxing the regularity conditions leads to efficient dynamic programming algorithms, but also to non-smooth approximations.
- $\Rightarrow$ I relied on generative latent data modeling
- The regression model with a hidden logistic process (RHLP) [J-1], addresses these issues: accurate regular curve approximation and segmentation.
- The RHLP which represents a dynamical mixture model, allows for activating, simultaneously and preferentially, time-varying polynomial regression models with both smooth and abrupt regime changes.
- Also an alternative to solve the classical nonlinear regression problem [J-3].
- Two EM variants for efficient model inference
- Then, I studied the problem of modeling and joint segmentation of multivariate temporal data with hidden process regression models [J-6][J-7]

Regression with hidden logistic process

- The developed models are based on (multiple) regression with hidden processes.
- The aim of regression is to explore the relationship of an observed random variable Y given a covariate vector $\boldsymbol{X} \in \mathbb{R}^p$ via conditional density functions for $Y|\boldsymbol{X} = \boldsymbol{x}$ of the form $f(y|\boldsymbol{x})$, rather than only exploring the unconditional distribution of Y.
- For time series, the independent vector $\boldsymbol{x}$ in general relates the sampling time t, which we will consider hereafter.
- We are interested in parametric (non-)linear regression functions $f(y|\boldsymbol{x}) = \mu(\boldsymbol{x}; \boldsymbol{\beta})$.
- Let $\boldsymbol{y} = (y_1, \ldots, y_n)$ be a time series composed of n univariate observations $y_i \in \mathbb{R}$ $(i = 1, \ldots, n)$ observed at the time points $\mathbf{t} = (t_1, \ldots, t_n)$.

The regression with hidden logistic process (RHLP)

- The RHLP model assumes that the observed time series is governed by a K-"state" hidden process $\mathbf{z} = (z_1, \ldots, z_n)$ with the categorical random variable $z_i \in \{1, \ldots, K\}$ representing the unknown (hidden) label of the regime (Gaussian) generating the ith observation y_i

$$y_i = \boldsymbol{\beta}_{z_i}^T \boldsymbol{x}_i + \sigma_{z_i} \epsilon_i \quad ; \quad \epsilon_i \sim \mathcal{N}(0,1), \quad (i = 1, \ldots, n)$$

where $\boldsymbol{\beta}_{z_i} \in \mathbb{R}^{p+1}$ is the regression coefficient vector, $\boldsymbol{x}_i = (1, t_i, \ldots, t_i^p)^T$ is the time dependent predictor, $\sigma_{z_i}^2$ the associated noise variance.

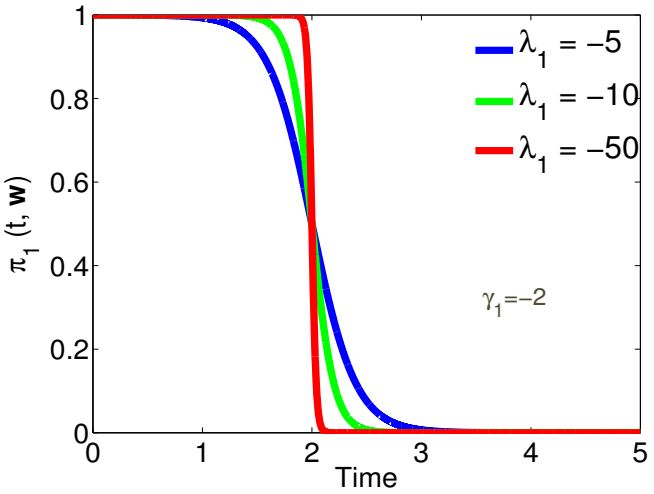
- The process $\boldsymbol{Z} = (Z_1, \ldots, Z_m)$ is assumed to be logistic: the hidden variable Z_i that allows to switch from one regression model to another during time follows the multinomial logistic distribution $\mathcal{M}(1, \pi_1(t_i; \mathbf{w}), \ldots, \pi_K(t_i; \mathbf{w}))$:

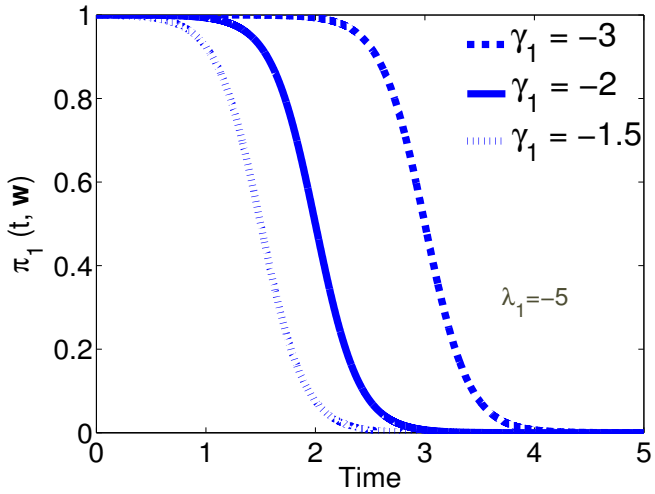
$$\pi_k(t_i; \mathbf{w}) = \mathbb{P}(Z_i = k | t_i; \mathbf{w}) = \frac{\exp\left(\boldsymbol{w}_k^T \boldsymbol{v}_i\right)}{\sum_{\ell=1}^{K} \exp\left(\boldsymbol{w}_\ell^T \boldsymbol{v}_i\right)},$$

where $\boldsymbol{v}_i = (1, t_i, , \ldots, t_i^u)^T \in \mathbb{R}^{u+1}$ is a time-dependent covariate vector, $\boldsymbol{w}_k \in \mathbb{R}^{u+1}$ is its associated coefficients $\boldsymbol{v}_i$ and $\mathbf{w} = (\boldsymbol{w}_1^T, \ldots, \boldsymbol{w}_{K-1}^T)^T \in \mathbb{R}^{(K-1)\times(u+1)}$ with $\boldsymbol{w}_K$ being the null vector.

The regression with hidden logistic process (RHLP)

- Modeling with the logistic distribution allows activating simultaneously and preferentially several regimes during time
- Flexibility of the logistic distribution: $\pi_r(t_i; \mathbf{w}) = \frac{\exp(\lambda_r(t_i+\gamma_r))}{\sum_{k=1}^{K}\exp(\lambda_\ell(t_i+\gamma_\ell))}$





$\Rightarrow$ The parameter λ_r controls the quality of transitions between regimes

$\Rightarrow$ The parameter γ_r is related to the transition time

If the goal is to segment the curves into contiguous segments, use linear logistic functions, that is by setting the value u of $\boldsymbol{w}_k$ to 1 (used hereafter)

The regression with hidden logistic process (RHLP)

- A K-component RHLP is defined by the following dynamical conditional mixture density:

$$f(y_i|t_i;\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k(t_i;\mathbf{w})\mathcal{N}\left(y_i;\boldsymbol{\beta}_k^T\boldsymbol{x}_i,\sigma_k^2\right),$$

- Parameter vector: $\boldsymbol{\theta} = (\mathbf{w}^T,\boldsymbol{\beta}_1^T,\ldots,\boldsymbol{\beta}_K^T,\sigma_1^2,\ldots,\sigma_K^2)^T$
- In the RHLP model, both the mixing proportions and the component parameters are time-varying, contrary to for example standard switching regression models or mixture of regression models
- It can be seen as a mixture of experts (ME) (Jordan and Jacobs, 1994) where the logistic weights are time-dependent, that is, the particular covariate variable used for the mixing proportions represents the time.

Illustration of the principle of the method

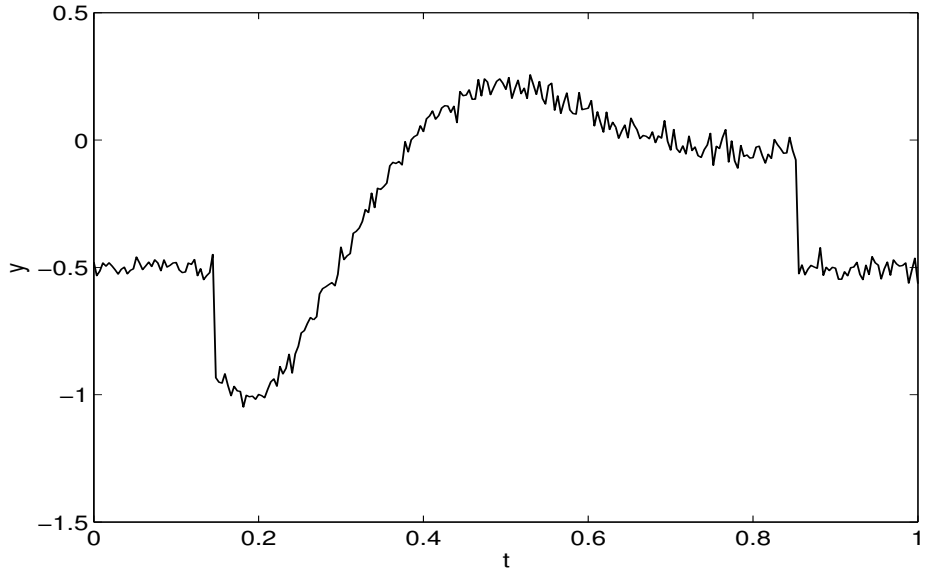
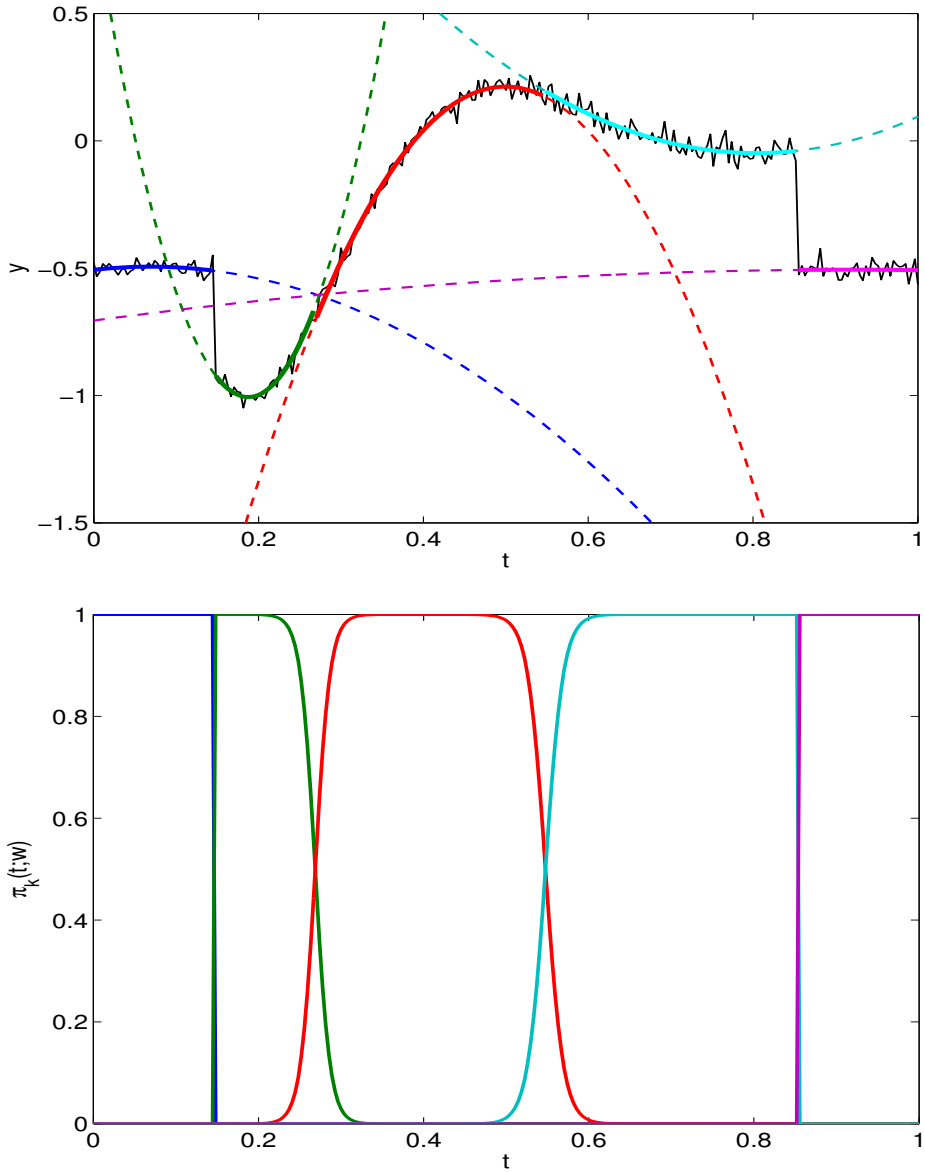


Illustration of the principle of the method



$K = 5$ polynomial components of degree $p = 2$

MLE of the RHLP via a dedicated EM

- The parameter vector $\boldsymbol{\theta}$ is estimated by monotonically maximizing the observed-data likelihood:

$$\log L(\boldsymbol{\theta}; \boldsymbol{y}, \mathbf{t}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k(t_i; \mathbf{w}) \mathcal{N}\left(y_i; \boldsymbol{\beta}_k^T \boldsymbol{x}_i, \sigma_k^2\right)$$

- Can not be performed in a closed form since the data are incomplete, that is, the labels $(z_1, \ldots, z_m)$ indicating from which component each observation of the time series is originated from, are unknown.
  $\Rightarrow$ The EM algorithm (Dempster et al., 1977; McLachlan and Krishnan, 2008) is particularly adapted to achieve this task [J-1].
- Complete-data log-likelihood

$$\log L_c(\boldsymbol{\theta}; \boldsymbol{y}, \mathbf{t}, \mathbf{z}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log[\pi_k(t_i; \mathbf{w}) \mathcal{N}\left(y_i; \boldsymbol{\beta}_k^T \boldsymbol{x}_i, \sigma_k^2\right)]$$

$z_{ik} = 1$ if $z_i = k$ (i.e., when y_i is belongs to the kth regime)

EM algorithm for the RHLP

- **The E-Step** computes the expected complete-data log-likelihood, given the observations $(\mathbf{t}, \boldsymbol{y})$ and a current parameter estimation $\boldsymbol{\theta}^{(q)}$

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)}) &= \mathbb{E}\left[\log L_c(\boldsymbol{\theta}; \boldsymbol{y}, \mathbf{t}, \mathbf{z}) | \boldsymbol{y}, \mathbf{t}; \boldsymbol{\theta}^{(q)}\right] \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}^{(q)} \left[\log \pi_k(t_i; \mathbf{w}) \mathcal{N}\left(y_i; \boldsymbol{\beta}_k^T \boldsymbol{x}_i, \sigma_k^2\right)\right], \end{aligned}$$

  $\Rightarrow$ simply requires calculation the posterior probability $\tau_{ik}^{(q)}$ that y_i $(i = 1, \ldots, m)$ originates from regime k $(k = 1, \ldots, K)$:

$$\tau_{ik}^{(q)} = \mathbb{P}(Z_i = k | y_i, t_i; \boldsymbol{\theta}^{(q)}) = \frac{\pi_k(t_i; \mathbf{w}^{(q)}) \mathcal{N}(y_i; \boldsymbol{\beta}_k^{T(q)} \boldsymbol{x}_i, \sigma_k^{2(q)})}{\sum_{\ell=1}^{K} \pi_\ell(t_i; \mathbf{w}^{(q)}) \mathcal{N}(y_i; \boldsymbol{\beta}_\ell^{T(q)} \boldsymbol{x}_i, \sigma_\ell^{2(q)})}.$$

- **The M-Step** computes the parameter vector update $\boldsymbol{\theta}^{(q+1)}$ by maximizing the expected complete-data log-likelihood, that is,

$$\boldsymbol{\theta}^{(q+1)} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)})$$

EM algorithm for the RHLP: M-Step

- The maximization of the Q-function w.r.t the regression coefficient vector $\boldsymbol{\beta}_k$ for each component k consists in analytically solving a weighted least-squares problem and the one w.r.t σ_k^2 is a weighted variant of the problem of estimating the variance of an univariate Gaussian density:

$$\begin{aligned}
\boldsymbol{\beta}_k^{(q+1)} &= \Big[\sum_{i=1}^{n} \tau_{ik}^{(q)} \boldsymbol{x}_i \boldsymbol{x}_i^T\Big]^{-1} \sum_{i=1}^{n} \tau_{ik}^{(q)} y_i \boldsymbol{x}_i, \\
\sigma_k^{2(q+1)} &= \frac{1}{\sum_{i=1}^{n} \tau_{ik}^{(q)}} \sum_{i=1}^{n} \tau_{ik}^{(q)} (y_i - \boldsymbol{\beta}_k^{T(q+1)} \boldsymbol{x}_i)^2.
\end{aligned}$$

EM algorithm for the RHLP: M-Step

- The maximization with respect to $\mathbf{w}$ is a multinomial logistic regression problem weighted by the $\tau_{ik}^{(q)}$'s; however cannot be solved in a closed form.

  $\Rightarrow$ It is solved with a multi-class Iteratively Reweighted Least Squares (IRLS) algorithm (Green, 1984; Chen et al., 1999)

$$\mathbf{w}^{(l+1)} = \mathbf{w}^{(l)} - \left[\frac{\partial^2 Q_{\mathbf{w}}(\mathbf{w}, \boldsymbol{\theta}^{(q)})}{\partial \mathbf{w} \partial \mathbf{w}^T}\right]^{-1}_{\mathbf{w}=\mathbf{w}^{(l)}} \frac{\partial Q_{\mathbf{w}}(\mathbf{w}, \boldsymbol{\theta}^{(q)})}{\partial \mathbf{w}}\Big|_{\mathbf{w}=\mathbf{w}^{(l)}}$$

- A convex optimization problem
- Analytic calculation of the Hessian and the gradient
- The EM-RHLP algorithm has a complexity of $\mathcal{O}(I_{\text{EM}} I_{\text{IRLS}} K^3 p^3 n)$, where I_{EM} is the number of iterations of the EM algorithm (more advantageous than dynamic programming).

Time series approximation and segmentation

1. Approximation: a prototype mean curve

$$\hat{y}_i = \mathbb{E}[y_i|t_i;\hat{\boldsymbol{\theta}}] = \sum_{k=1}^{K} \pi_k(t_i;\hat{\mathbf{w}})\hat{\boldsymbol{\beta}}_k^T \boldsymbol{x}_i$$

   - A smooth and flexible approximation thanks to the the logistic weights
   - The RHLP can be used to solve the nonlinear regression model $y_i = f(t_i;\boldsymbol{\theta}) + \epsilon_i$ by covering regression functions of the form $f(t_i;\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k(t_i;\mathbf{w})\boldsymbol{\beta}_k^T \boldsymbol{x}_i$,

2. Curve segmentation:

$$\hat{z}_i = \arg\max_{1\leq k\leq K} \mathbb{E}[z_i|t_i;\hat{\mathbf{w}}] = \arg\max_{1\leq k\leq K} \pi_k(t_i;\hat{\mathbf{w}})$$

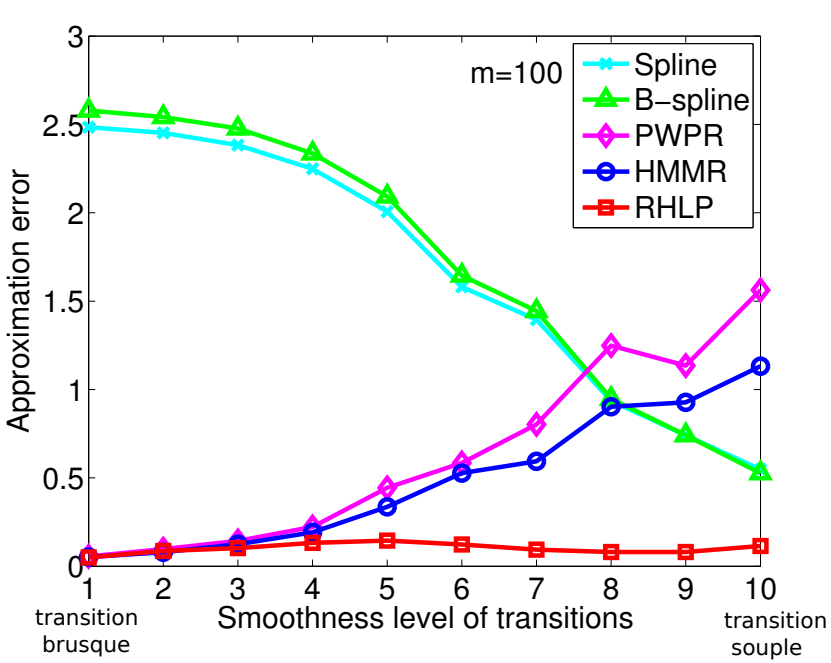
3. **Model selection** Application of BIC, ICL

   $\text{BIC}(K,p) = \log L(\hat{\boldsymbol{\theta}}) - \frac{\nu_{\boldsymbol{\theta}} \log(n)}{2}$; $\text{ICL}(K,p) = \log L_c(\hat{\boldsymbol{\theta}}) - \frac{\nu_{\boldsymbol{\theta}} \log(n)}{2}$

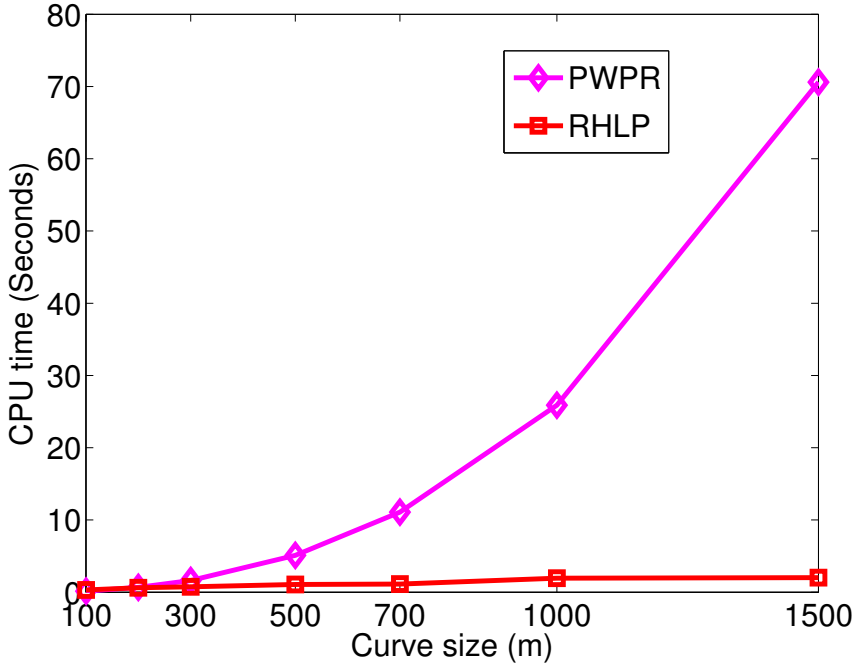
   where $\nu_{\boldsymbol{\theta}} = K(p+4) - 2$.

Evaluation in modeling and segmentation

Approximation error as a function of the speed of transitions

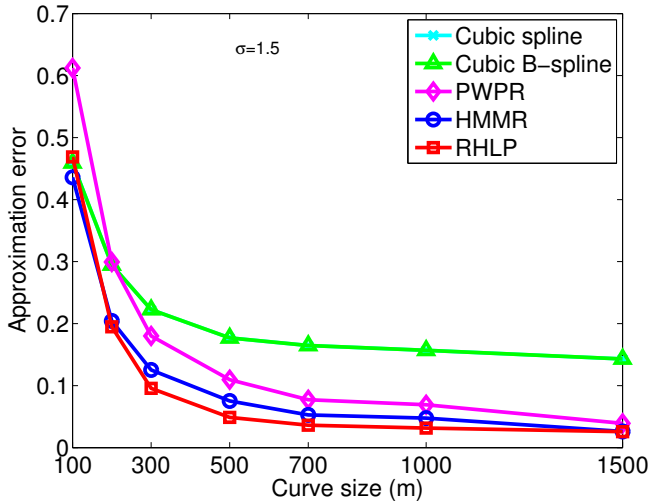


Computing time

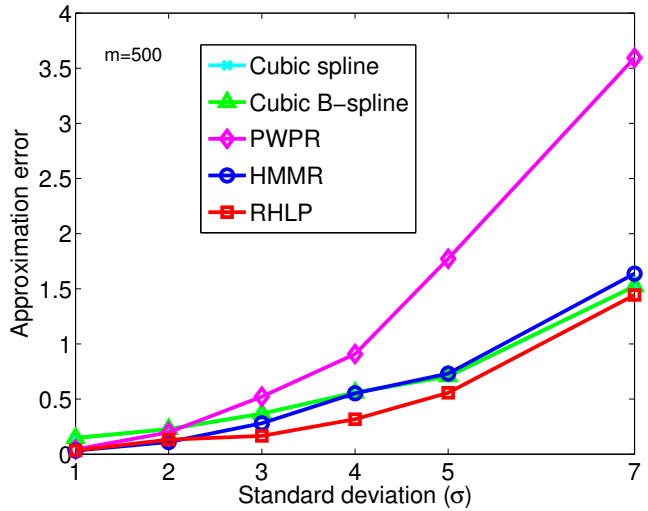


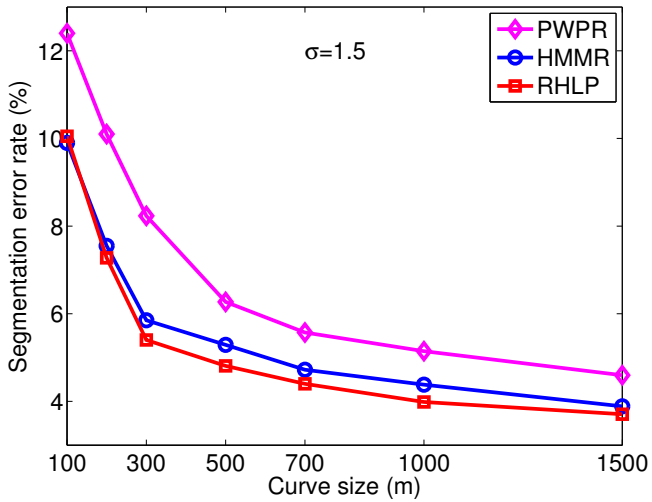
Evaluation in approximation and segmentation

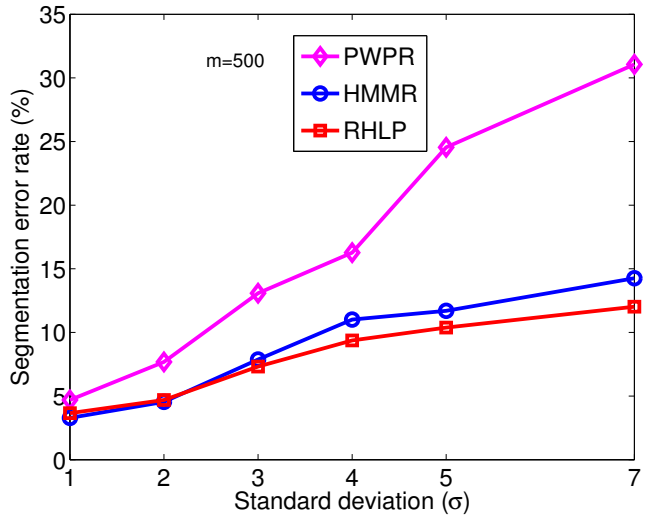
varying m



varying σ

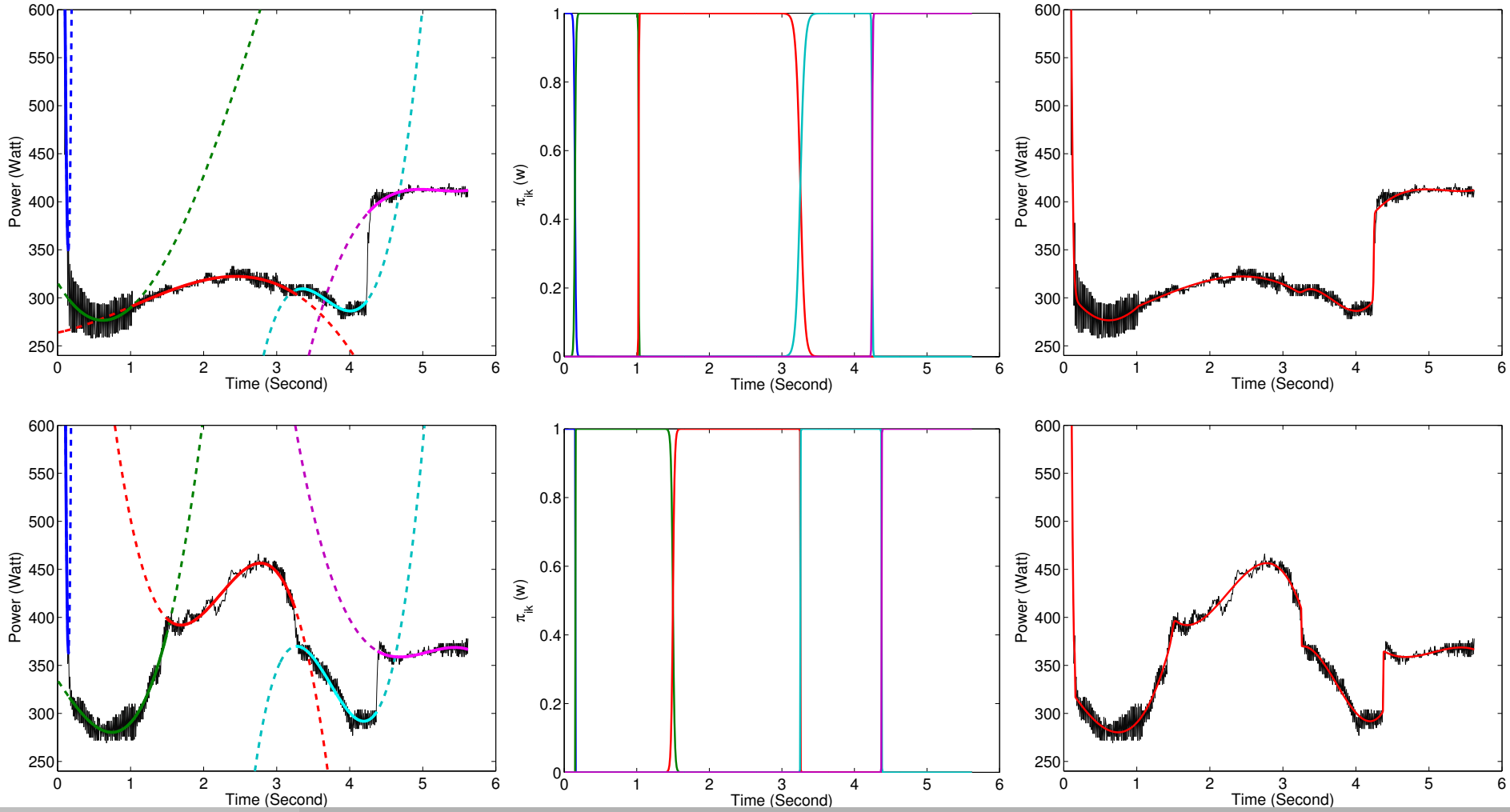






Application of the RHLP to real data

- Approximation of real time series issued from railway diagnosis application
- The data are the power signals during high-speed railway switch operations, each operation signal is composed of five successive movements



Multiple hidden process regression for joint segmentation of multivariate time series

- Extend the previous framework to the joint segmentation of multivariate time series with regime changes
- Let $\mathbf{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ be a time series of n multidimensional observations $\boldsymbol{y}_i = (y_i^{(1)}, \ldots, y_i^{(d)})^T \in \mathbb{R}^d$ observed at the time points $\mathbf{t} = (t_1, \ldots, t_n)$.
- The univariate components of the multivariate time series are simultaneously governed by a hidden process and thus the problem of segmentation becomes the one of recovering the hidden process.
- $\Rightarrow$ Multiple regression with hidden process: Multiple RHLP [J-6] and Multiple hidden Markov model regression [J-7]

Multiple hidden process regression

- The multiple regression with hidden process model:

$$\begin{aligned} y_i^{(1)} &= \boldsymbol{\beta}_{z_i}^{(1)T}\boldsymbol{x}_i + \sigma_{z_i}^{(1)}\epsilon_i \\ y_i^{(2)} &= \boldsymbol{\beta}_{z_i}^{(2)T}\boldsymbol{x}_i + \sigma_{z_i}^{(2)}\epsilon_i \\ \vdots & \quad \vdots \\ y_i^{(d)} &= \boldsymbol{\beta}_{z_i}^{(d)T}\boldsymbol{x}_i + \sigma_{z_i}^{(d)}\epsilon_i \end{aligned}$$

  which can be written in a matrix form as

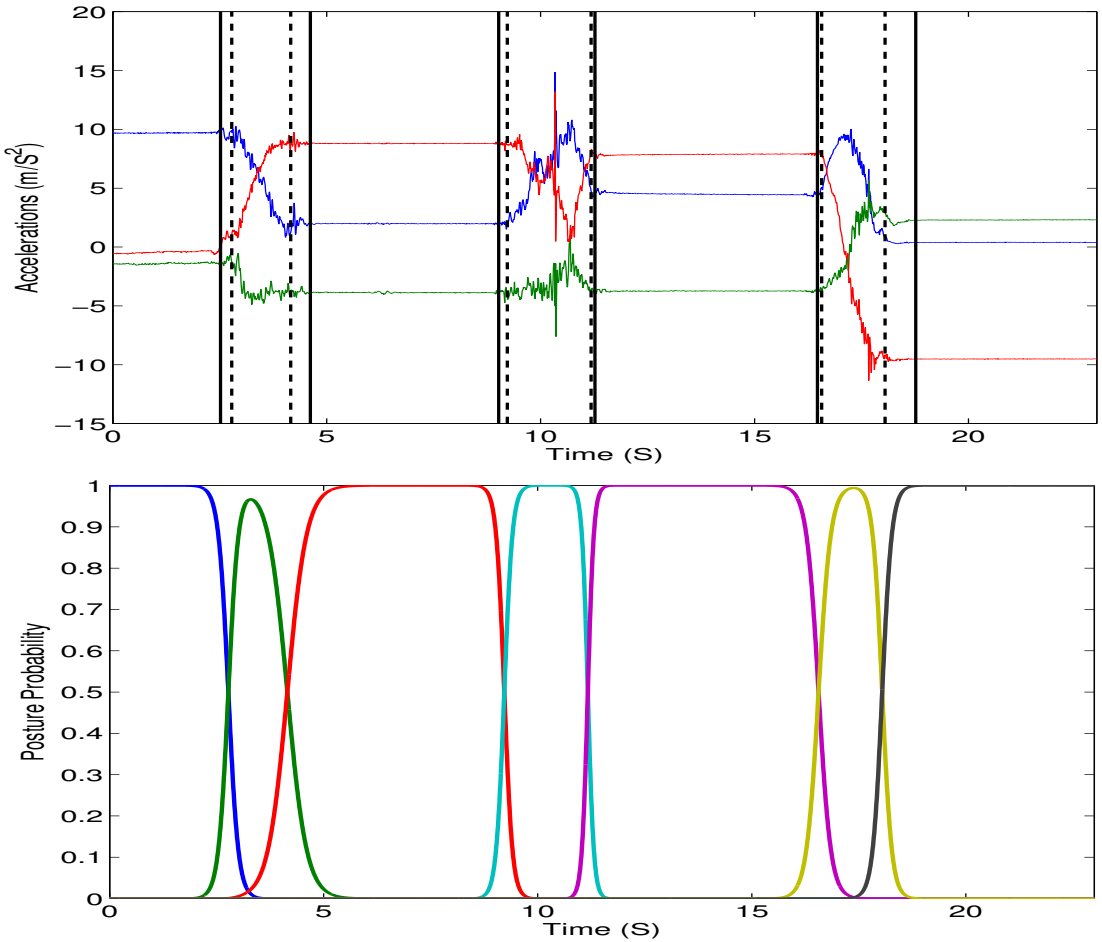
$$\boldsymbol{y}_i = \mathbf{B}_{z_i}^T\boldsymbol{x}_i + \mathbf{e}_i \quad ; \quad \mathbf{e}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{z_i}), \quad (i = 1, \ldots, n)$$

  where $\mathbf{B}_k = \left[\boldsymbol{\beta}_k^{(1)}, \ldots, \boldsymbol{\beta}_k^{(d)}\right]$ is a $(p+1) \times d$ matrix of regression parameters of regime $Z_i = k$ and $\boldsymbol{\Sigma}_{z_i}$ its corresponding $d \times d$ covariance matrix.

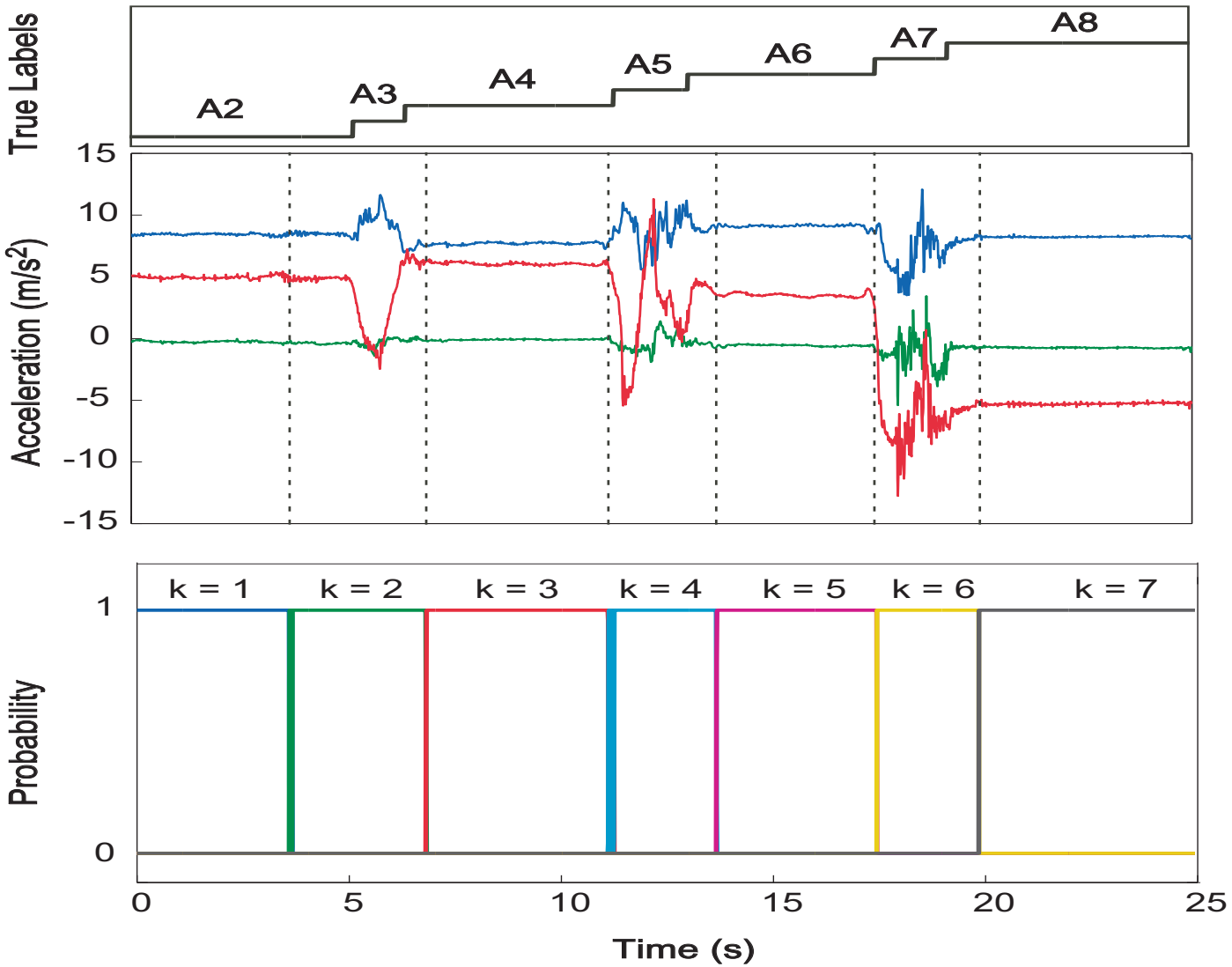
- The latent process $\mathbf{z}$ that simultaneously governs the univariate time series components controls the regime change during time
- We investigated the case where this process is logistic (MRHLP) [J-6], and where it is a Markov chain (MHMMR) [J-7]

Application on human activity time series

MRHLP Segmentation of acceleration data issued from three body-worn sensors (Data acquired at the LISSI Lab/University of Paris 12)



Multiple hidden Markov model regression



Summary

- The RHLP, thanks to its generative modeling, is naturally tailored to deal with the problem of modeling regime changing time series
- The parameter estimates are computed by maximizing the log-likelihood by using an efficient EM algorithm.
- Particularly useful for situations with smooth regime transitions
- Good performance on various real data segmentation and approximation

Outline

Model-based curve clustering

- The aim curve clustering is to cluster n iid unlabeled curves $((\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n))$ into K clusters
- We assume that each curve consists of m observations $\mathbf{y}_i = (y_{i1}, \ldots, y_{im})$ regularly observed at the inputs $\mathbf{x}_i = (x_{i1}, \ldots, x_{im})$
- $\Rightarrow$ find the unknown cluster labels $\mathbf{z} = (z_1, \ldots, z_n)$, with $z_i \in \{1, \ldots, K\}$, K being the number of clusters
- $\Rightarrow$ the curve clustering can be performed based on regression mixture models including polynomial regression mixtures (PRM) and polynomial spline regression mixtures (PSRM) (Gaffney, 2004; Chamroukhi, 2010).

Regression mixtures for model-based curve clustering

- The mixture of polynomial, spline, or B-spline regressions is defined by

$$f(\mathbf{y}_i|\mathbf{x}_i;\mathbf{\Psi}) \quad = \quad \sum_{k=1}^{K} \pi_k \, \mathcal{N}(\mathbf{y}_i;\mathbf{X}_i\boldsymbol{\beta}_k,\sigma_k^2\mathbf{I}_m).$$

- Parameter vector: $\mathbf{\Psi} = (\pi_1,\ldots,\pi_K,\mathbf{\Psi}_1^T,\ldots,\mathbf{\Psi}_K^T)^T$: where $\mathbf{\Psi}_k = (\boldsymbol{\beta}_k^T,\sigma_k^2)^T$ are respectively the regression coefficients and the noise variance

- $\mathbf{\Psi}$ is estimated by maximizing the log-likelihood:

$$\mathcal{L}(\mathbf{\Psi}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \, \mathcal{N}(\mathbf{y}_i;\mathbf{X}_i\boldsymbol{\beta}_k,\sigma_k^2\mathbf{I}_m).$$

- The maximization can be performed iteratively via the EM algorithm (eg. (Gaffney, 2004))

Limitations

1. The standard EM algorithm for regression mixture model is sensitive to initialization $\Rightarrow$ requires careful initialization
2. It requires the number of clusters to be supplied by the user $\Rightarrow$ requires to deal with the model selection

In general, theses two issues have been considered each *separately*:

- Initialization techniques: randomly, K-means, CEM, etc
- Choosing the number of clusters via an afterward model selection procedure: BIC, AIC, ICL, etc

Idea of the proposed approach [J-8]

- $\Rightarrow$ Here we attempt to overcome these limitations simultaneously in this case of model-based curve clustering
- $\Rightarrow$ We propose an EM-like algorithm which is robust with regard initialization and automatically estimates the number of clusters as the learning proceeds
- $\Rightarrow$A fully unsupervised fitting of regression mixtures

Penalized maximum likelihood estimation

- For estimating the regression mixture model $\Rightarrow$ maximize a penalized log-likelihood function rather than the standard log-likelihood (31)
- penalize the log-likelihood by a term accounting for the model complexity

Regularization

- As the model complexity is mainly governed by the number of clusters (the hidden variables z_i) $\Rightarrow$ use as penalty the entropy of the hidden variable z_i
- The (differential) entropy of one variable ($z_i \in \{1, \ldots, K\}$):

$$H(z_i) = -\mathbb{E}[\log p(z_i)] = -\sum_{k=1}^{K} \pi_k \log \pi_k.$$

- The variables $\mathbf{z} = (z_1, \ldots, z_n)$ are i.i.d, $\Rightarrow$ the whole entropy for $\mathbf{z}$ is:

$$H(\mathbf{z}) \quad = \quad -n \sum_{k=1}^{K} \pi_k \log \pi_k \, .$$

Penalized maximum likelihood estimation

- The objective function we propose to maximize is thus given bu the following penalized log-likelihood:

$$\begin{aligned}\mathcal{J}(\lambda, \boldsymbol{\Psi}) &= \mathcal{L}(\boldsymbol{\Psi}) - \lambda H(\mathbf{z}), \quad \lambda \geq 0 \\ &= \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{y}_i; \mathbf{X}_i \boldsymbol{\beta}_k, \sigma_k^2 \mathbf{I}_m) + \lambda n \sum_{k=1}^{K} \pi_k \log \pi_k\end{aligned}$$

- $\mathcal{L}(\boldsymbol{\Psi})$ is the observed-data log-likelihood maximized by the standard EM algorithm for regression mixtures
- When the entropy is large, the fitted model is rougher, and when it is small, the fitted model is smoother.
- $\lambda \geq 0$ is a smoothing parameter for establishing a trade-off between closeness of fit to the data and a smooth fit

- the model parameters $\boldsymbol{\Psi}$ are estimated by maximizing the penalized observed-data log-likelihood (1) $\mathcal{J}(\lambda, \boldsymbol{\Psi})$ given an i.i.d dataset of n curves $\mathcal{D} = ((\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n))$
- $\mathcal{J}(\lambda, \boldsymbol{\Psi})$ is iteratively maximized by using a dedicated EM-like algorithm

$\Rightarrow$ The complete-data log-likelihood of $\boldsymbol{\Psi}$ in this penalized case is given y:

$$\mathcal{J}_c(\lambda, \boldsymbol{\Psi}) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} \log\left[\pi_k \mathcal{N}(\mathbf{y}_i; \mathbf{X}_i\boldsymbol{\beta}_k, \sigma_k^2 \mathbf{I}_m)\right] + \lambda n \sum_{k=1}^{K} \pi_k \log \pi_k \,.$$

- z_{ik} is an indicator binary variable such that $z_{ik} = 1$ iff $z_i = k$ (i.e., if the ith curve $(\mathbf{x}_i, \mathbf{y}_i)$ is generated by cluster k)

Robust EM-like algorithm for regression mixtures

Start with an initial solution (parameter $\boldsymbol{\Psi}^{(0)}$ and a number of clusters K)

1. **E-step** Compute the expected penalized complete-data log-likelihood (1)

$$Q(\lambda, \boldsymbol{\Psi}; \boldsymbol{\Psi}^{(q)}) = \mathbb{E}\big[\mathcal{J}_c(\lambda, \boldsymbol{\Psi}) | \mathcal{D}; \boldsymbol{\Psi}^{(q)}\big]$$
$$= \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}^{(q)} \log\left[\pi_k \mathcal{N}(\mathbf{y}_i; \mathbf{X}_i\boldsymbol{\beta}_k, \sigma_k^2 \mathbf{I}_m)\right] + \lambda n \sum_{k=1}^{K} \pi_k \log \pi_k$$

$\Rightarrow$ simply consists in computing the posterior cluster probabilities:

$$\tau_{ik}^{(q)} = \frac{\pi_k^{(q)} \mathcal{N}\left(\mathbf{y}_i; \mathbf{X}_i\boldsymbol{\beta}_k^{(q)}, \sigma_k^{2(q)} \mathbf{I}_m\right)}{\sum_{h=1}^{K} \pi_h^{(q)} \mathcal{N}(\mathbf{y}_i; \mathbf{X}_i\boldsymbol{\beta}_h^{(q)}, \sigma_h^{2(q)} \mathbf{I}_m)}\,.$$

2. **M-step** Updating step: $\boldsymbol{\Psi}^{(q+1)} = \arg\max_{\boldsymbol{\Psi}} Q(\lambda, \boldsymbol{\Psi}; \boldsymbol{\Psi}^{(q)})$.

1 The mixing proportions updates are obtained by maximizing the function

$$Q_\pi(\lambda; \boldsymbol{\Psi}^{(q)}) = \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}^{(q)} \log \pi_k + \lambda \sum_{i=1}^{n}\sum_{k=1}^{K} \pi_k \log \pi_k$$

$\Rightarrow$ This can be solved using Lagrange multipliers :

$$\pi_k^{(q+1)} = \frac{1}{n}\sum_{i=1}^{n} \tau_{ik}^{(q)} + \lambda \pi_k^{(q)} \left(\log \pi_k^{(q)} - \sum_{h=1}^{K} \pi_h^{(q)} \log \pi_h^{(q)} \right)$$

2 The regression parameters for each class k are updated by maximizing

$$Q_{\boldsymbol{\Psi}_k}(\lambda, \boldsymbol{\beta}_k, \sigma_k^2; \boldsymbol{\Psi}^{(q)}) = \sum_{i=1}^{n} \tau_{ik}^{(q)} \log \mathcal{N}(\mathbf{y}_i; \mathbf{X}_i\boldsymbol{\beta}_k, \sigma_k^2 \mathbf{I}_m)$$

$\Rightarrow$ consists in analytic solutions of K weighted least-squares problems:

$$\boldsymbol{\beta}_k^{(q+1)} = \left[\sum_{i=1}^{n} \tau_{ik}^{(q)} \mathbf{X}_i^T \mathbf{X}_i\right]^{-1} \sum_{i=1}^{n} \tau_{ik}^{(q)} \mathbf{X}_i^T \mathbf{y}_i \quad \sigma_k^{2(q+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(q)} \|\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_k\|^2}{m \sum_{i=1}^{n} \tau_{ik}^{(q)}}$$

- for very small value of λ: the update of the mixing proportions is close to the one in the standard EM update
- however for a large value of λ : the penalization term will play its role in order to make clusters competitive $\Rightarrow$ allows for discarding invalid clusters and enhancing actual clusters
- A cluster k can be discarded if its proportion is less than $\frac{1}{n}$
- The penalization coefficient λ is set in an adaptive way to be large for enhancing competition

Initialization and stopping rule

- initialization of the number of clusters : $K^{(0)} = n$
- initialization of the mixing proportions : $\pi_k^{(0)} = \frac{1}{K^{(0)}}$, $(k = 1, \ldots, K^{(0)})$,
- to initialize the regression parameters $\boldsymbol{\beta}_k$ and the noise variances $\sigma_k^{2(0)}$, fit a polynomial regression model to each curve k :

  $\boldsymbol{\beta}_k^{(0)} = \left(\mathbf{X}^T\mathbf{X}_k\right)^{-1}\mathbf{X}_k\mathbf{y}_k$ and $\sigma_k^{2(0)} = \frac{1}{m}\|\mathbf{y}_k - \mathbf{X}_k\boldsymbol{\beta}_k^{(0)}\|^2$.

  However, to avoid singularities at the starting point, we set $\sigma_k^{2(0)}$ as a middle value in the following sorted range $\|\mathbf{y}_i - \mathbf{X}\boldsymbol{\beta}_k^{(0)}\|^2$ for $i = 1, \ldots, n$.
- $\Rightarrow \boldsymbol{\Psi}_k^{(0)} = (\boldsymbol{\beta}_k^{(0)}, \sigma_k^{2(0)})$.
- The proposed EM algorithm is stopped when the maximum variation of the estimated regression parameters between two iterations $\max_{1\leq k\leq K^{(q)}} \|\boldsymbol{\beta}_k^{(q+1)} - \boldsymbol{\beta}_k^{(q)}\|$ is less than a threshold ϵ (e.g., 10^{-6}).

Experimental study: Waveform curves of Brieman

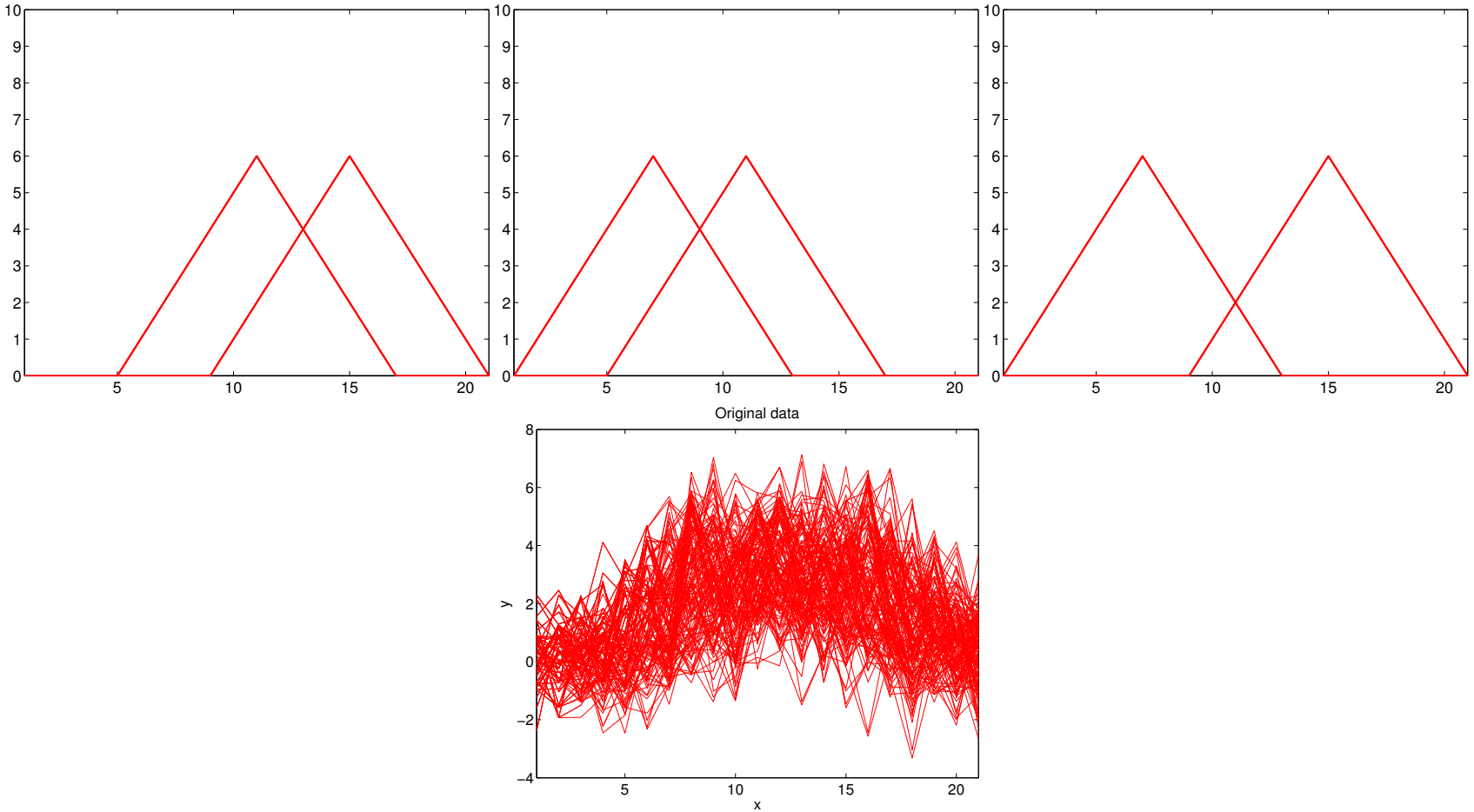


Figure: Waveform mean functions from the generative model before the Gaussian noise is added, and a sample of 150 waveforms.

EM-PRM clustering results

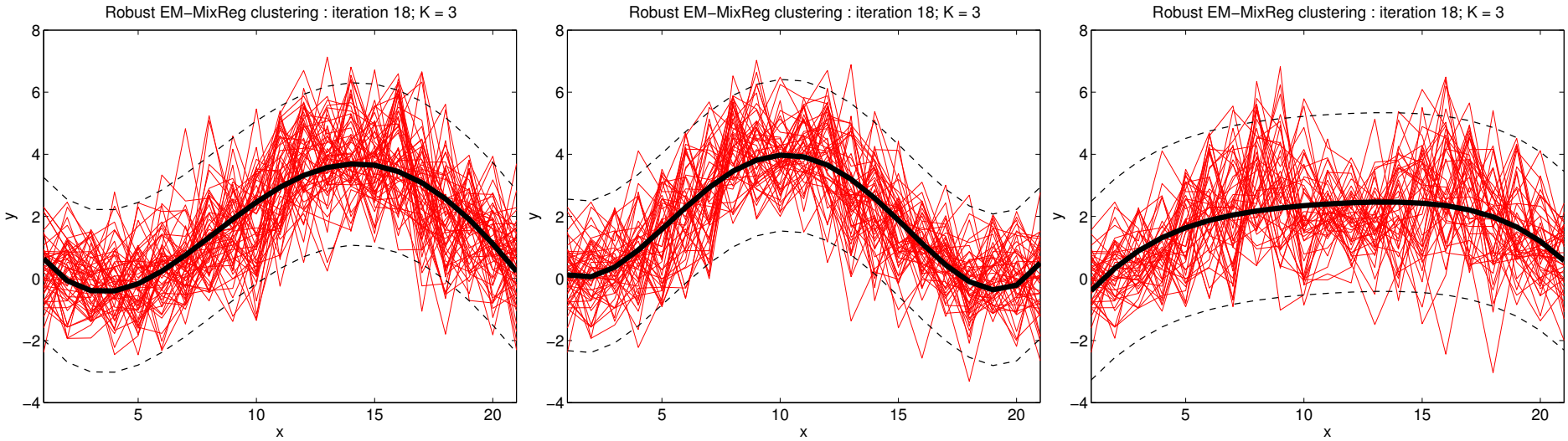


Figure: Clustering results obtained by the proposed robust EM algorithm and the PRM (polynomial degree $p = 4$) model for the waveform data.

EM-PSRM clustering results

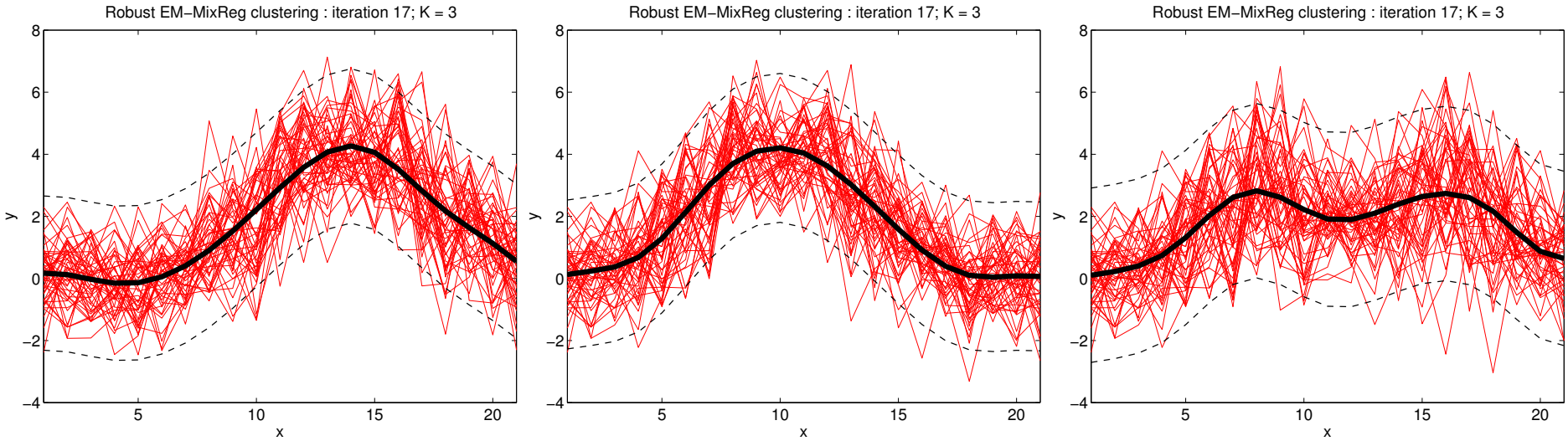


Figure: Clustering results obtained by the proposed robust EM algorithm and the SRM with a cubic-spline of three knots for the waveform data.

EM-PbSRM clustering results

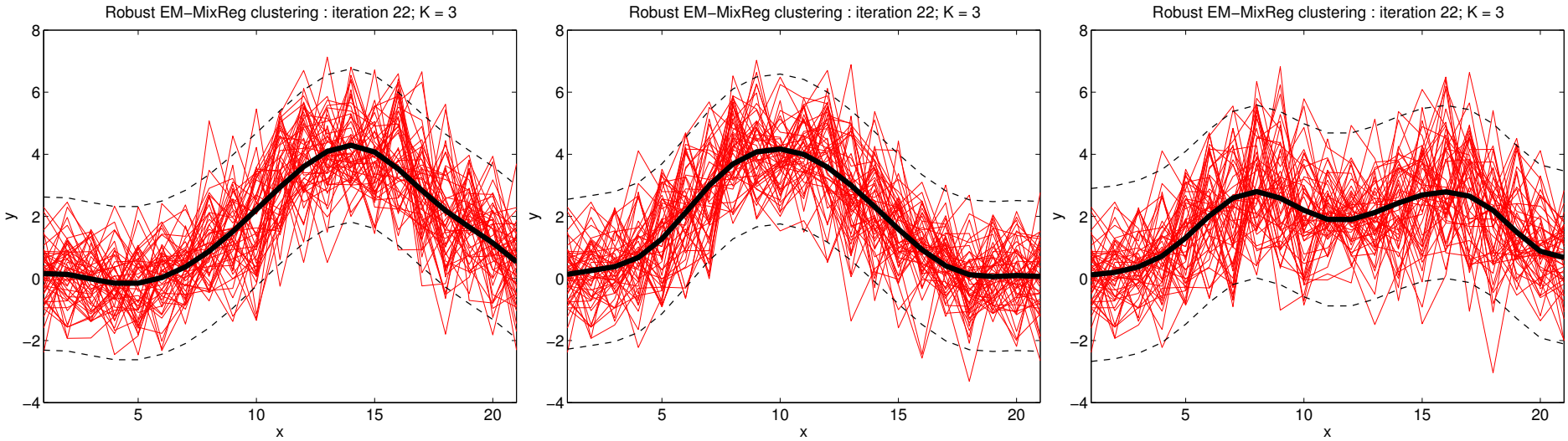


Figure: Clustering results obtained by the proposed robust EM algorithm and the bSRM with a cubic B-spline of three knots for the waveform data.

Clustering results

Estimated number of clusters, misclassification error rate and the absolute error between the true clusters proportions and variances and the estimated ones.

	actual	EM-PRM	EM-SRM	EM-bSRM
K	3	3	3	3
misc. error	-	4.31 $\pm$ (0.42)%	2.94 $\pm$ (0.88)%	2.53 $\pm$ (0.70)%
σ_1	1	0.128 $\pm$ (0.015)	0.108 $\pm$ (0.015)	0.103 $\pm$ (0.012)
σ_2	1	0.102 $\pm$ (0.015)	0.090 $\pm$ (0.011)	0.079 $\pm$ (0.010)
σ_3	1	0.223 $\pm$ (0.021)	0.180 $\pm$ (0.014)	0.141 $\pm$ (0.013)
π_1	$\frac{1}{3}$	0.0037 $\pm$ (0.0018)	0.0035 $\pm$ (0.0015)	0.0034 $\pm$ (0.0015)
π_2	$\frac{1}{3}$	0.0029 $\pm$ (0.0023)	0.0018 $\pm$ (0.0015)	0.0012 $\pm$ (0.0011)
π_3	$\frac{1}{3}$	0.0040 $\pm$ (0.0062)	0.0037 $\pm$ (0.0015)	0.0035 $\pm$ (0.0014)

Table: Clustering results over 20 different samples of 500 curves.

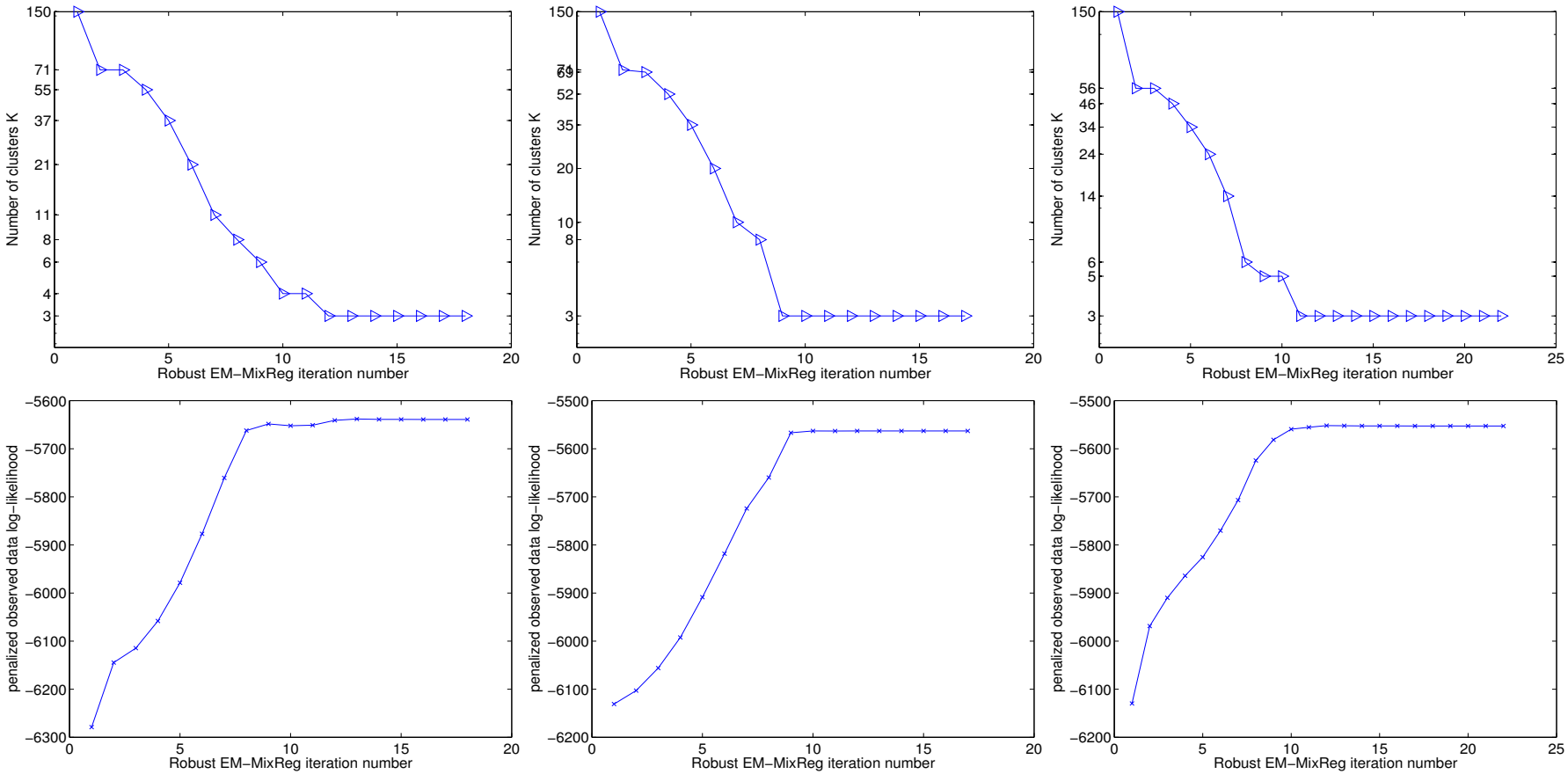


Figure: Variation of the number of clusters and the value of the objective function during the iterations of the algorithm for the PRM (left), SRM (middle) and bSRM (right) for the waveform data.

Experiments on real data

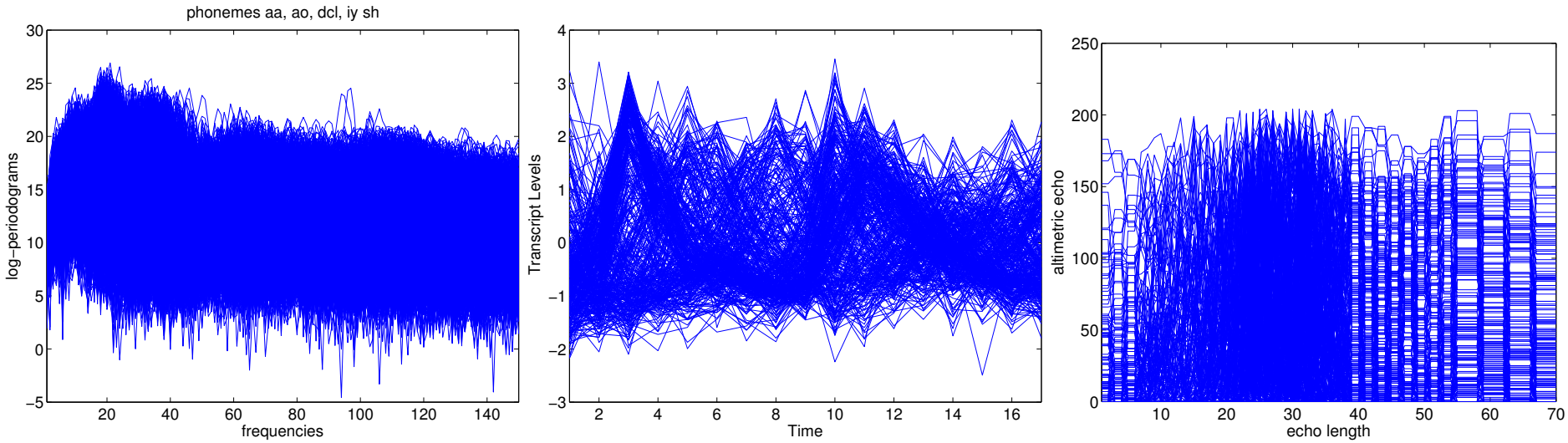


Figure: Real data: Phonemes of the classes "ao", "aa", "iy", "dcl", "sh" (left), the Yeast cell cycle data (middle) and the Topex/Poseidon satellite data (right).

Phonemes data

1000 log-periodograms (200 per cluster)

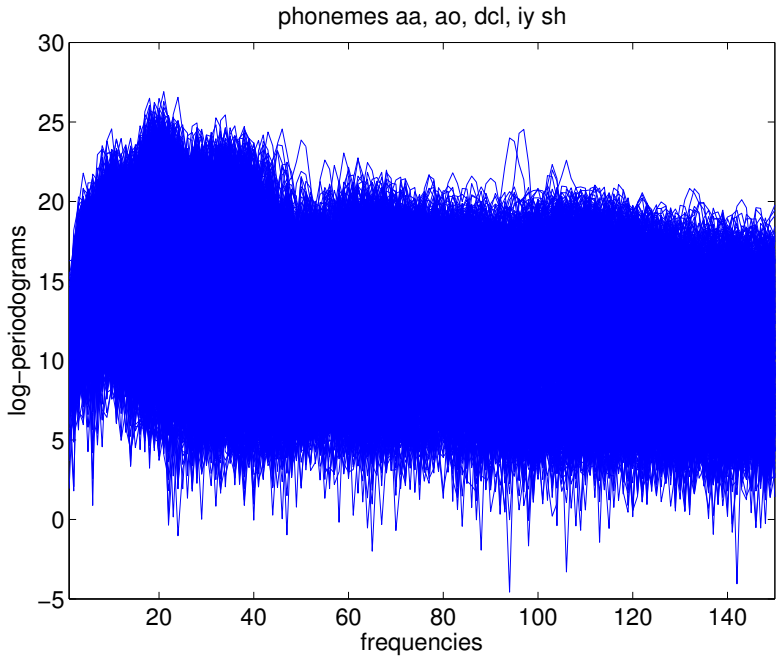


Figure: Phonemes data "ao", "aa", "yi", "dcl", "sh".

Phonemes data

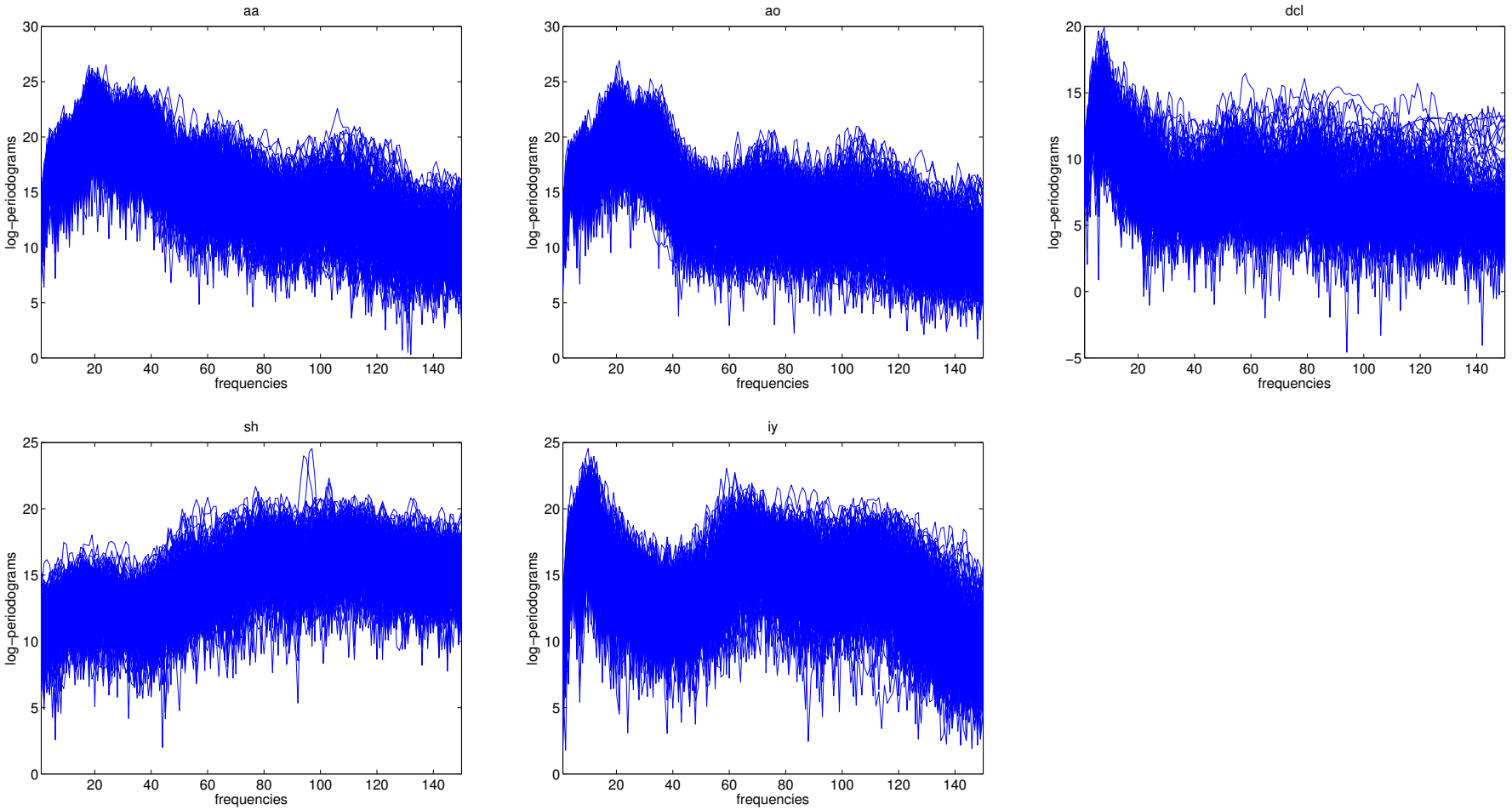


Figure: Curves of the actual five phoneme classes: "ao", "aa", "yi", "dcl", "sh".

PRM clustering results for Phonemes

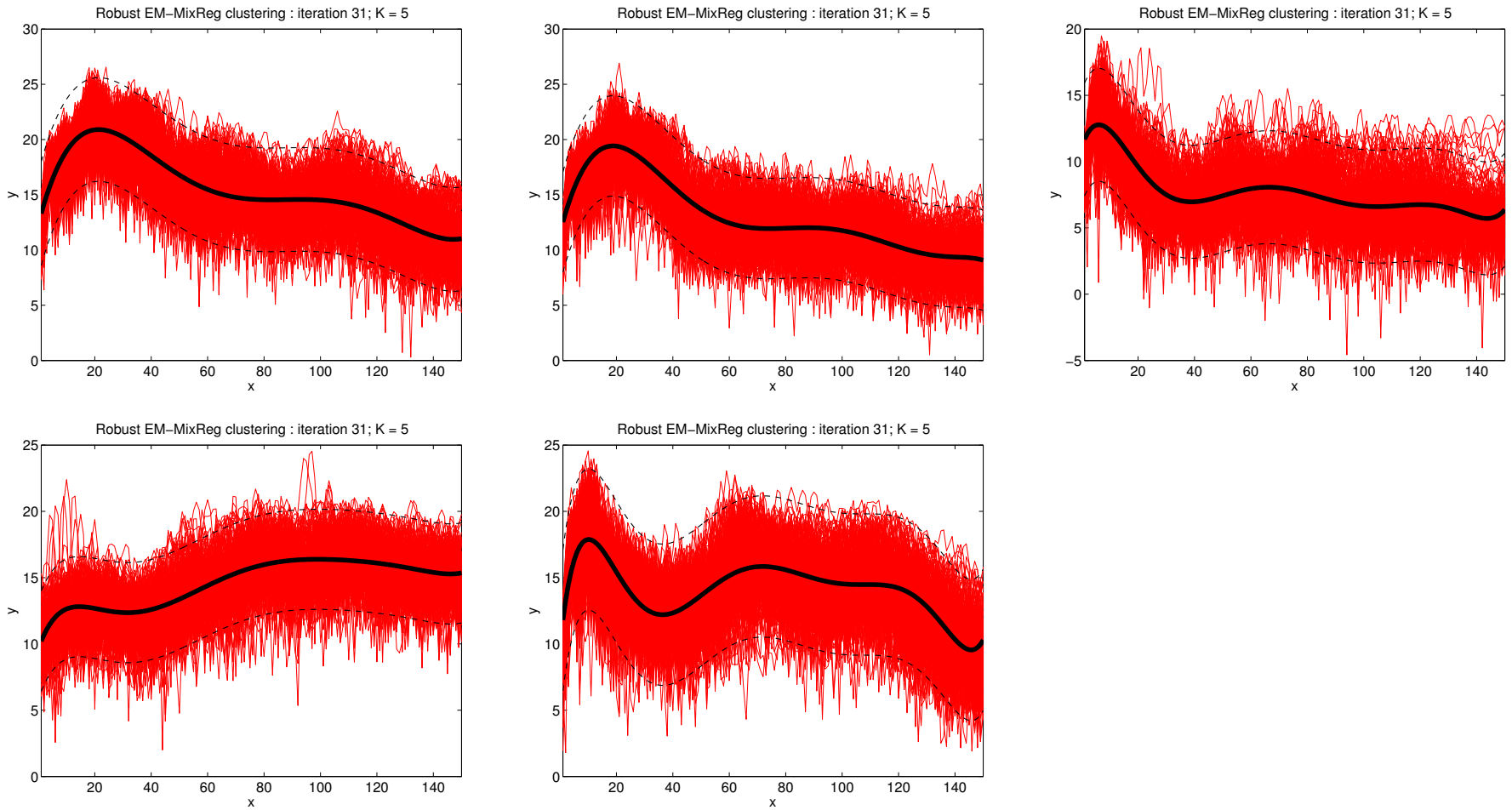


Figure: Clustering results obtained by the proposed robust EM for PRM ($p = 7$)

PbSRM clustering results for Phonemes

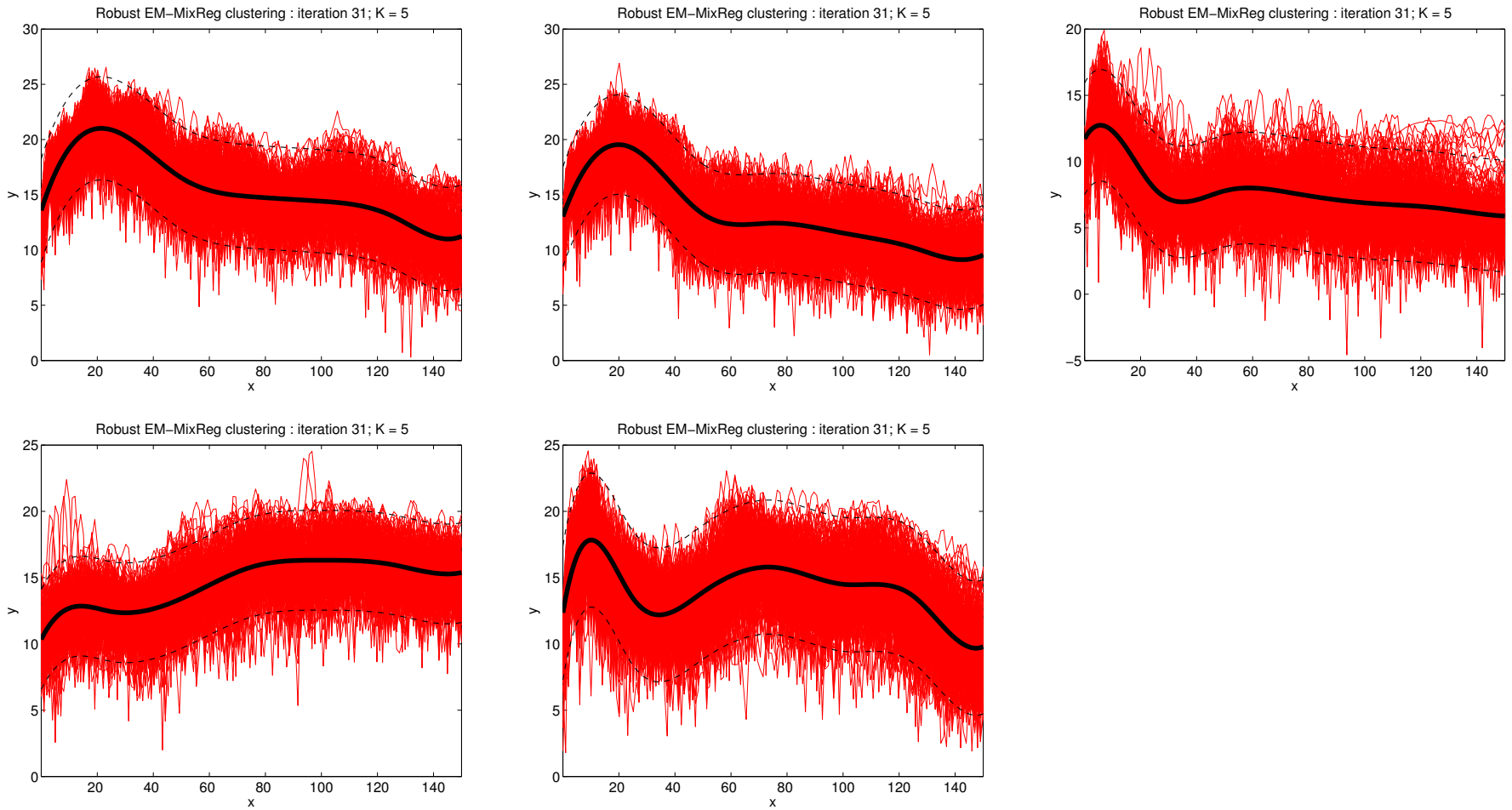


Figure: Clustering results obtained by the proposed robust EM for bSRM

Clustering results for Phonemes

- The spline regression mixture (SRM) results are closely similar to those provided by the B-spline mixture (bSRM)
- The number of phoneme classes is correctly estimated by the three models.
- The spline regression models provide better results in terms of clusters approximation than the polynomial regression mixture (here $p = 7$).
- Notice that the value of $p = 7$ correspond to the polynomial regression mixture model with the best error rate for p varying from 4 to 8.
- Values of the estimated number of clusters and the misc. error rates:

	EM-PRM	EM-SRM	EM-bSRM
Estimated K	5	5	5
Misc. error rate	14.29 %	14.09 %	14.2 %

Table: Clustering results for the phonemes data.

- The spline regressions mixture perform better than the polynomial

Clustering results for Phonemes

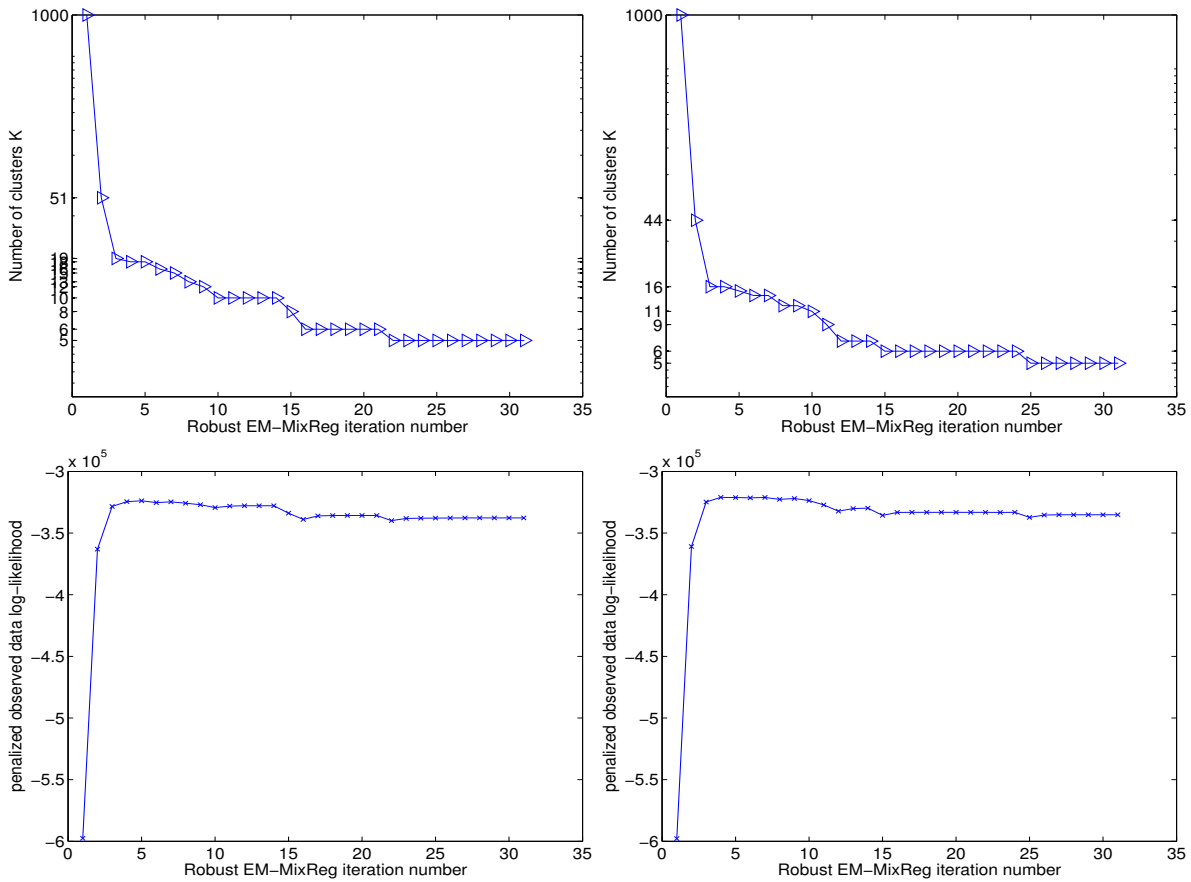


Figure: Variation of the number of clusters and the value of the objective function during the iterations of the algorithm for the PRM (left) and bSRM (right) for the phonemes data.

Yeast cell cycle data

- We consider yeast cell cycle data (time course Gene expression data) as in (Yeung et al., 2001) [2]
- This data set referred to as the subset of the 5-phase criterion in (Yeung et al., 2001) contains 384 genes expression levels over 17 time points.

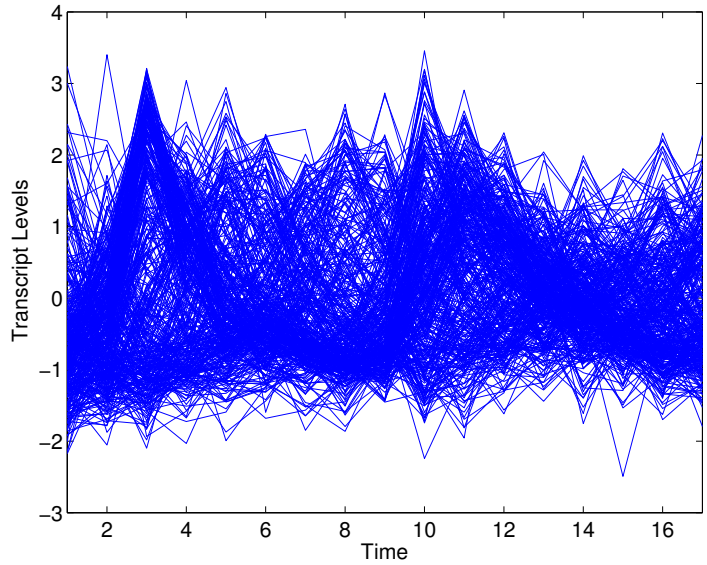


Figure: Yeast cell cycle data.

[2] http://faculty.washington.edu/kayee/model/

Yeast cell cycle data

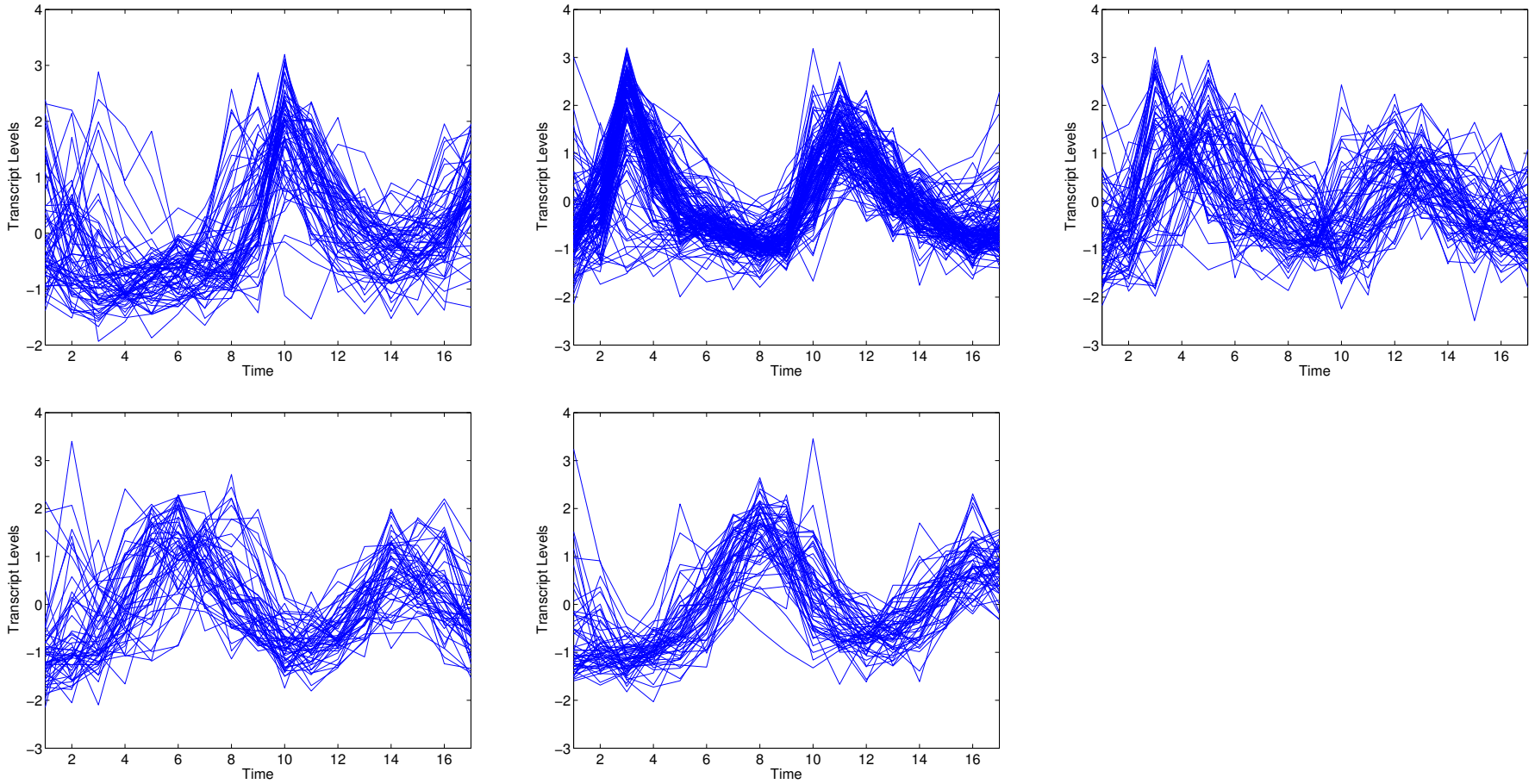


Figure: The five "actual" clusters of the used yeast cell cycle data according to Yeung et al. (2001).

SRM Clustering results for the yeast cell cycle data

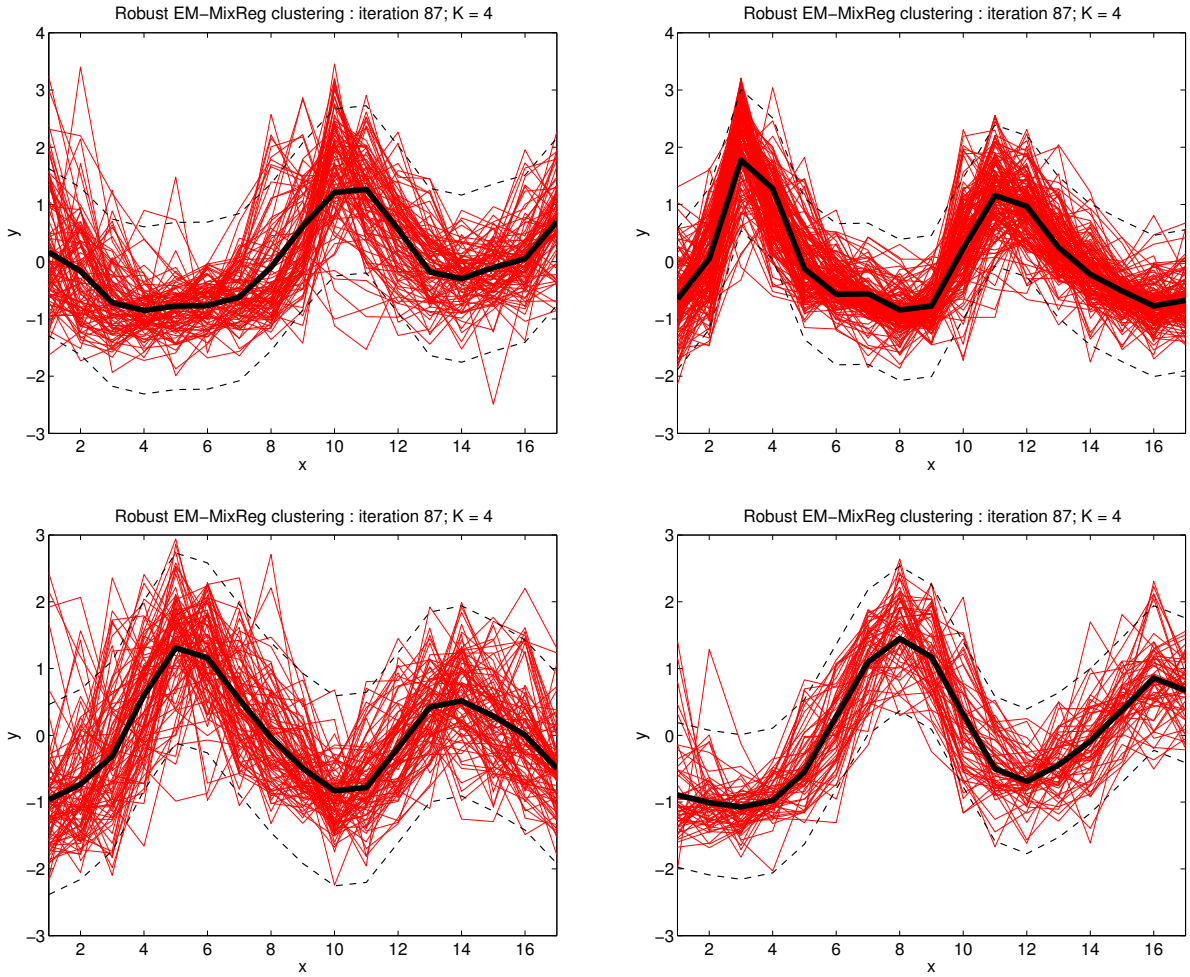


Figure: Clustering results obtained by the proposed robust EM algorithm and the SRM model with a cubic spline of 7 knots for the yeast cell cycle data.

bSRM clustering of the yeast cell cycle data

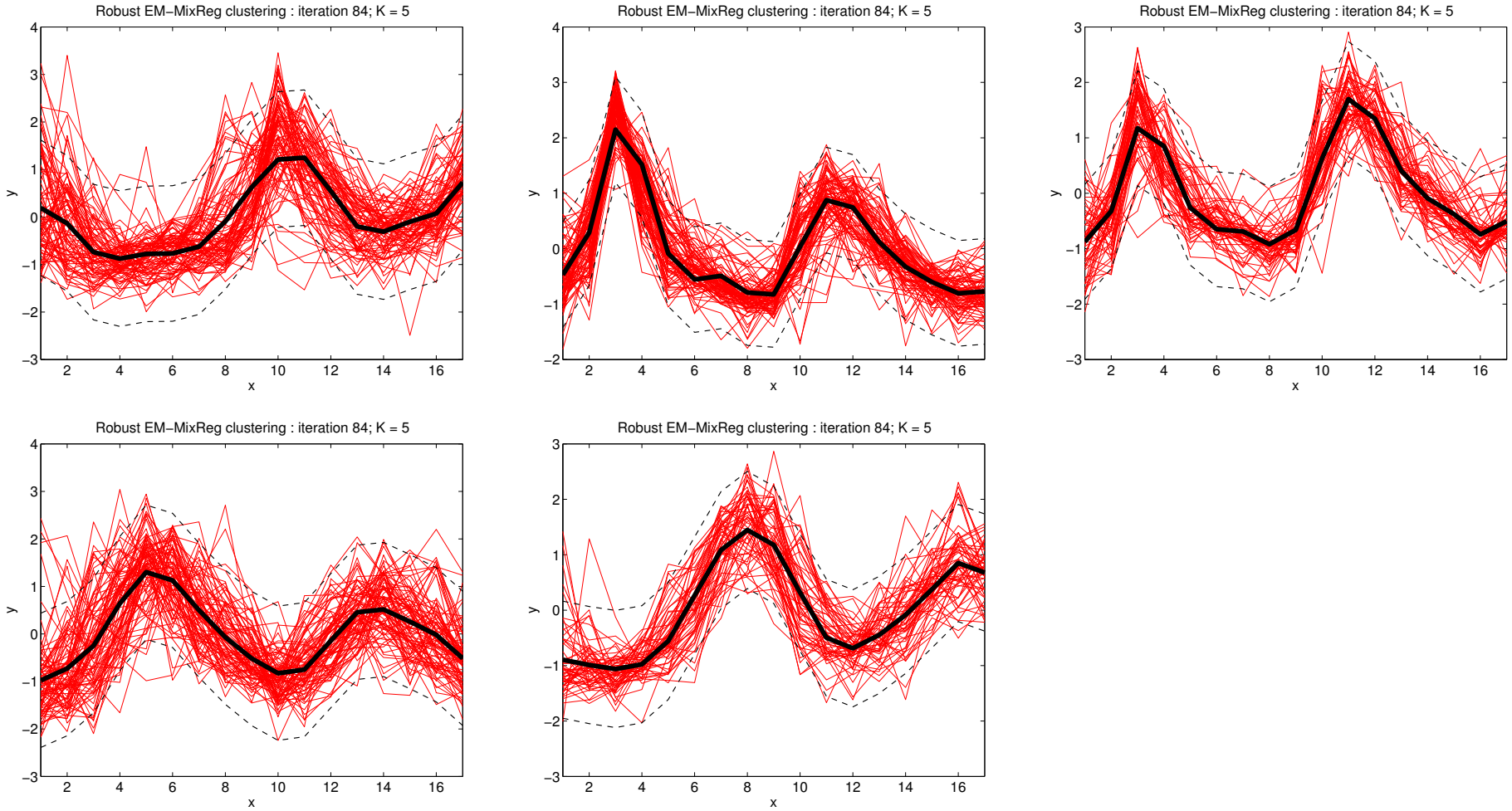


Figure: Clustering results obtained by the proposed robust EM algorithm and the bSRM model with a cubic B-spline of 7 knots for the yeast cell cycle data.

Clustering results for the yeast cell cycle data

- Both the PRM model and the SRM provide similar partitions with four clusters.
- The second and third clusters for PRM and SRM look to be merged into the second cluster for the bSRM solution and the partition of (Yeung et al., 2001) .
- Note that some model selection criteria in (Yeung et al., 2001) also provide four clusters in some situations.
- the bSRM model infers an accurate partition with the actual number of clusters. The Rand index for the obtained partition is 0.7914 which indicates that the partition is quite well defined.

Clustering results for the yeast cell cycle data

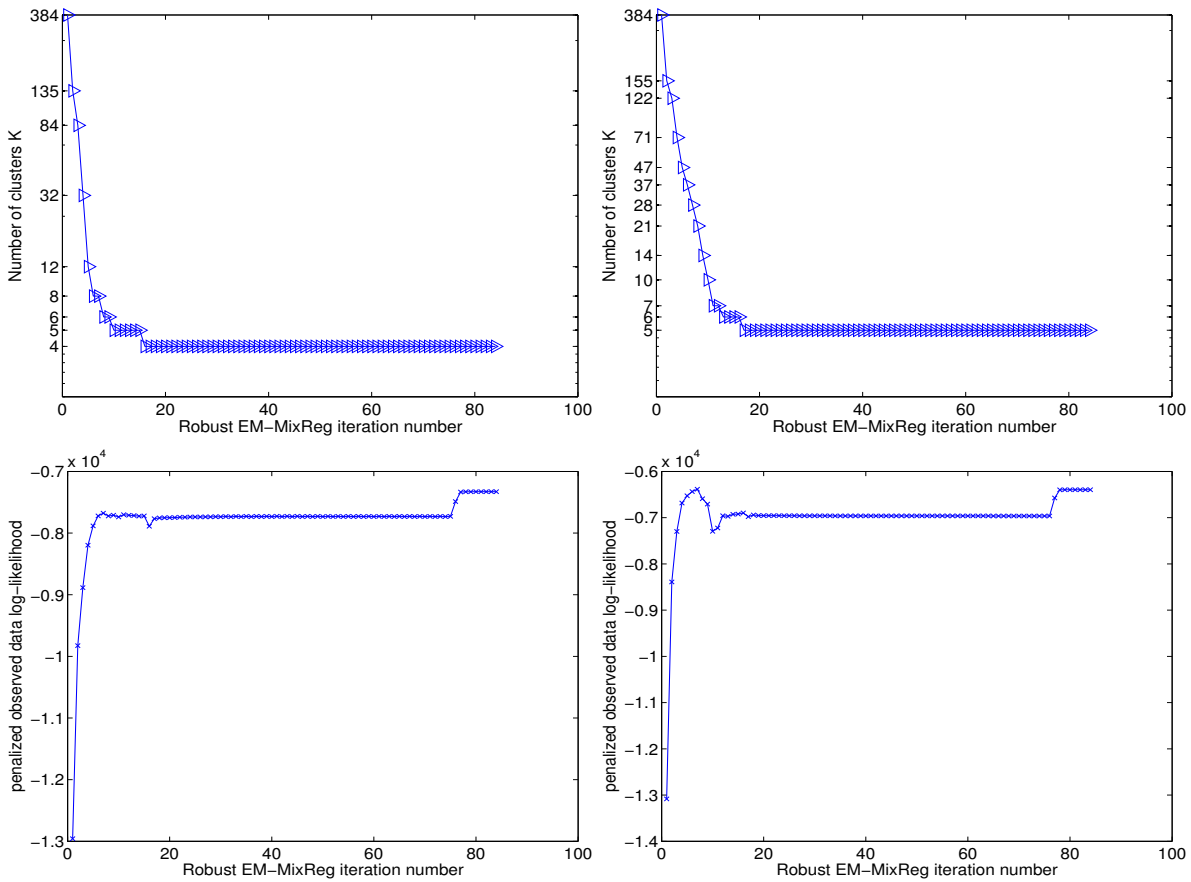


Figure: Variation of the number of clusters and the value of the objective function during the iterations of the algorithm for the PRM (left) and bSRM (right) for the yeast data.

Topex/Poseidon satellite data

- The data contain $n = 472$ waveforms of the measured echoes, sampled at $m = 70$ (number of echoes), (used in Dabo-Niang et al. (2007))
- The actual partition is unknown

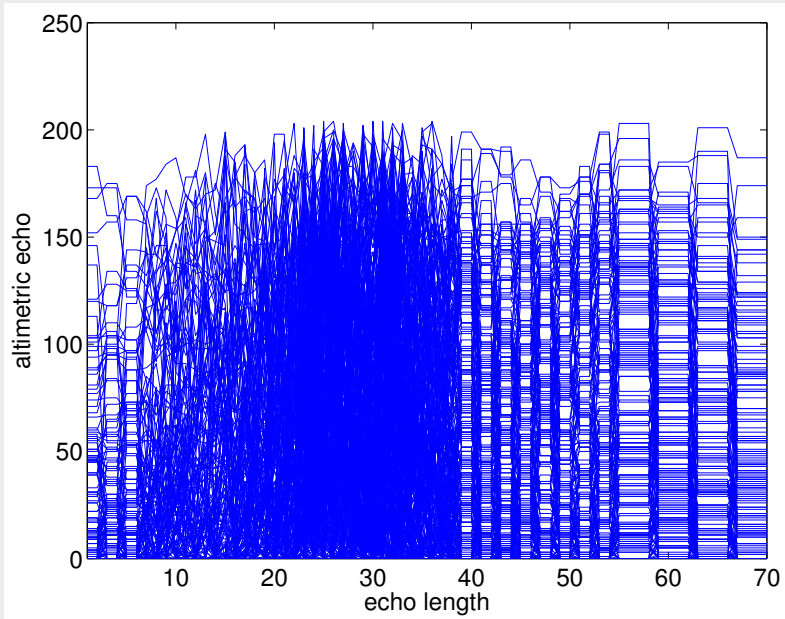


Figure: Topex/Poseidon satellite curves.

bSRM clustering results for the satellite data

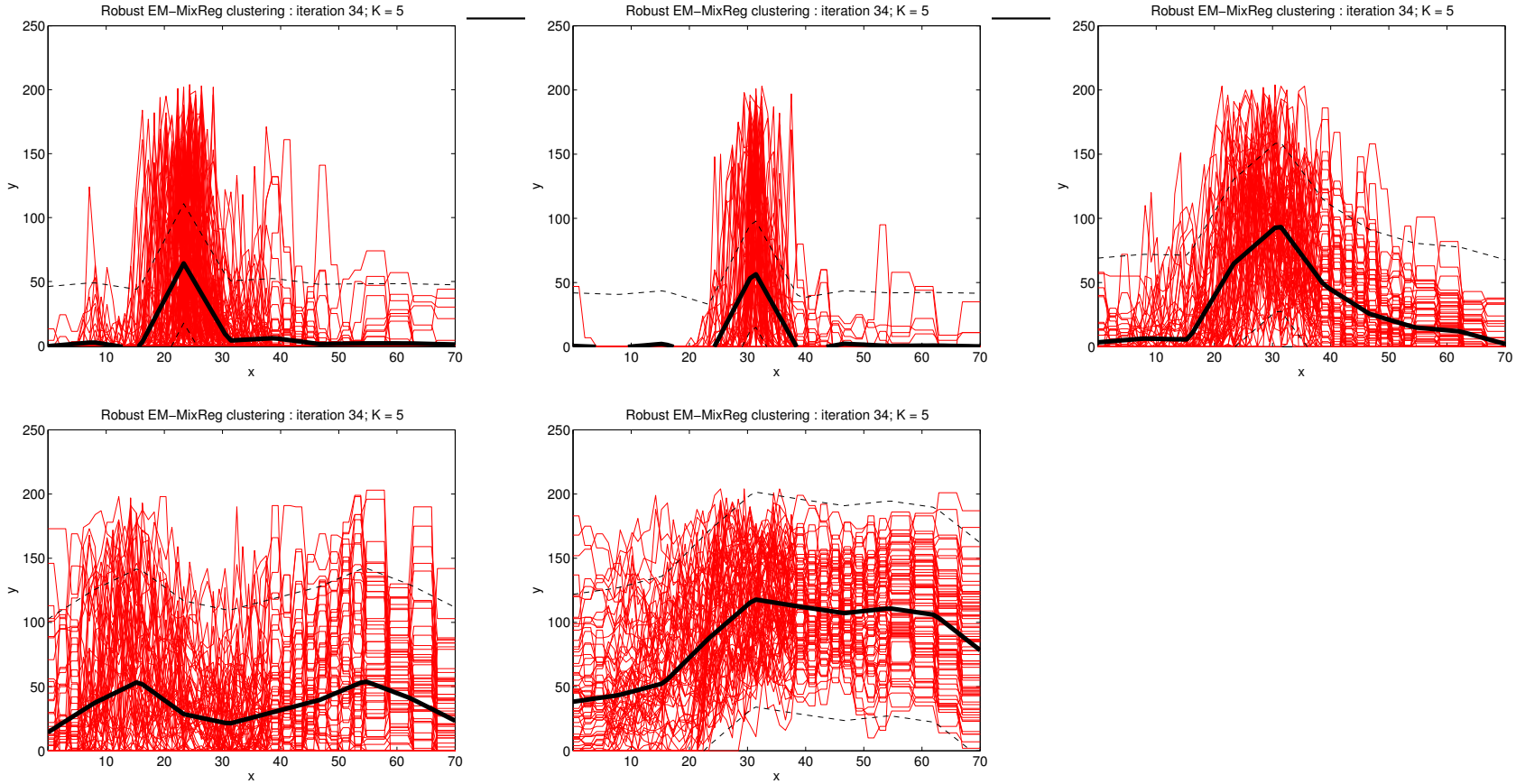


Figure: Clustering results obtained by the proposed robust EM algorithm and the bSRM model with a linear B-spline of 8 knots for the satellite data.

The estimated number of cluster (five) equals the one found by (Dabo-Niang et al., 2007) who use nonparametric kernel-based unsupervised classification technique and the partitions are quite similar

Non-normal mixtures of experts

Problem

- Mixture of experts (MoE) is a popular framework for modeling heterogeneity in data machine learning and statistics
- Investigate (MoE) for continuous data, in the case where the expert components are non-normal, (do not follow the Normal distribution)
- Indeed , for a set of data containing a group or groups of observations with asymmetric behavior, heavy tails or atypical observations, the use of normal experts may be unsuitable and can unduly affect the fit

Objectives

- Overcome these (well-known) limitations of MoE modeling with the normal distribution.
- I proposed three non-normal derivations including two robust mixture of experts (MoE) models. The proposed models are suitable to accomodate data which exhibit additional features such as skewness, heavy-tails and

Mixture of experts for continuous data

- Mixture of experts (MoE) (Jacobs et al., 1991; Jordan and Jacobs, 1994) are used in regression, classification and clustering.
- Observed pairs of data $(\boldsymbol{x}, y)$ where $y \in \mathbb{R}$ is the response for some covariate $\boldsymbol{x} \in \mathbb{R}^p$ governed by a hidden categorical random variable Z
- MoE model the component membership variable Z as a logistic function of some predictors $\boldsymbol{r} \in \mathbb{R}^q$ (the gating network)

$$\mathbb{P}(Z = k|\boldsymbol{r}; \boldsymbol{\alpha}) = \pi_k(\boldsymbol{r}; \boldsymbol{\alpha}) = \frac{\exp\left(\boldsymbol{\alpha}_k^T \boldsymbol{r}\right)}{\sum_{\ell=1}^{K} \exp\left(\boldsymbol{\alpha}_\ell^T \boldsymbol{r}\right)}$$

- MoE decompose the nonlinear regression model $f(y|\boldsymbol{x})$ as:

$$f(y|\boldsymbol{x}; \boldsymbol{\Psi}) \quad = \quad \sum_{k=1}^{K} \pi_k(\boldsymbol{r}; \boldsymbol{\alpha}) f_k(y|\boldsymbol{x}; \boldsymbol{\Psi}_k)$$

  where $f_k(y|\boldsymbol{x}; \boldsymbol{\Psi}_k)$ is the conditional density of a parametric regression function and the π_k's are covariate-varying mixing proportions.
- The model parameter vector: $\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_{K-1}, \boldsymbol{\Psi}_1^T, \ldots, \boldsymbol{\Psi}_K^T)^T$

The normal mixture of experts model and its MLE

- MoE for regression usually use normal experts $f_k(y|\boldsymbol{x};\boldsymbol{\Psi}_k)$:

$$f(y|\boldsymbol{r},\boldsymbol{x};\boldsymbol{\Psi}) = \sum_{k=1}^{K}\pi_k(\boldsymbol{r};\boldsymbol{\alpha})\mathsf{N}\big(y;\mu(\boldsymbol{x};\boldsymbol{\beta}_k),\sigma_k^2\big)$$

  where the component means are defined as parametric (non-)linear regression functions $\mu(\boldsymbol{x};\boldsymbol{\beta}_k)$.
- Given an i.i.d sample of n observations $(y_1,\ldots,y_n)$ with the covariates $(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$ and $(\boldsymbol{r}_1,\ldots,\boldsymbol{r}_n)$, the NMoE model parameters are estimated by maximizing the log-likelihood

$$\log L(\boldsymbol{\Psi}) = \sum_{i=1}^{n}\log\sum_{k=1}^{K}\pi_k(\boldsymbol{r}_i;\boldsymbol{\alpha})\mathsf{N}\big(y_i;\mu(\boldsymbol{x};\boldsymbol{\beta}_k),\sigma_k^2\big)$$

  by using the EM algorithm
- However, the normal distribution is not adapted to deal with asymmetric and heavy tailed data. It is also known that the normal distribution is sensitive to outliers

Contribution

- I introduced three new non-normal mixture of experts (NNMoE) that can better accommodate data exhibiting non-normal features, including asymmetry, heavy-tails, and the presence of outliers.
- The models rely on distributions that generalize the normal distribution:
  1. the skew-normal MoE (SNMoE) [J-12]
  2. the t MoE (TMoE) [J-13]
  3. the skew-t MoE (STMoE) [J-14]
- Dedicated E(C)M algorithms are developed to estimate the models parameters by monotonically maximizing the observed data log-likelihood.
- I describe how the presented models can be used in prediction in regression as well as in model-based clustering of regression data.

The skew-normal mixture of experts model

- The skew-normal mixture of experts (SNMoE) model uses the skew-normal distribution as density for the expert components.
- **The skew-normal distribution** (Azzalini, 1985, 1986) with location $\mu \in \mathbb{R}$, scale $\sigma^2 \in (0, \infty)$ and skewness $\lambda \in \mathbb{R}$ has density

$$f(y; \mu, \sigma^2, \lambda) = \frac{2}{\sigma}\phi(\frac{y-\mu}{\sigma})\Phi\left(\lambda(\frac{y-\mu}{\sigma})\right)$$

  where $\phi(.)$ and $\Phi(.)$ denote, respectively, the pdf and the cdf of the standard normal distribution.
- When the skewness parameter $\lambda = 0$, the skew-normal reduces to the normal distribution.
- The presented skew-normal mixture of experts (SNMoE) extends the skew-normal mixture model (Lin et al., 2007b) to the case of mixture of experts framework, by considering conditional distributions for both the mixing proportions and the means of the mixture components.

The skew-normal mixture of experts model

- The SNMoE is therefore a MoE model with skew-normal experts and is defined as follows. Let $\mathrm{SN}(\mu, \sigma^2, \lambda)$ denotes a skew-normal distribution with location parameter μ, scale parameter σ and skewness parameter λ. A K-component SNMoE is then defined by:

$$f(y|\boldsymbol{r}, \boldsymbol{x}; \boldsymbol{\Psi}) = \sum_{k=1}^{K} \pi_k(\boldsymbol{r}; \boldsymbol{\alpha}) \mathrm{SN}\big(y; \mu(\boldsymbol{x}; \boldsymbol{\beta}_k), \sigma_k^2, \lambda_k\big)$$

  where each expert component k has indeed a skew-normal distribution, whose density is defined by (1). The parameter vector of the model is $\boldsymbol{\Psi} = (\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_{K-1}^T, \boldsymbol{\Psi}_1^T, \ldots, \boldsymbol{\Psi}_K^T)^T$ with $\boldsymbol{\Psi}_k = (\boldsymbol{\beta}_k^T, \sigma_k^2, \lambda_k)^T$ the parameter vector for the kth skewed-normal expert component.

- It is obvious to see that if the skewness parameter $\lambda_k = 0$ for each k, the SNMoE model reduces to the NMoE model.

The skew-normal mixture of experts model

The SNMoE model is characterized as follows.

- **Stochastic representation of the SNMoE:** A random variable Y_i is said to follow the SNMoE model if it has the following representation:

$$Y_i = \mu(\boldsymbol{x}_i; \boldsymbol{\beta}_{z_i}) + \delta_{z_i}\sigma_{z_i}|U_i| + \sqrt{1-\delta_{z_i}^2}\,\sigma_{z_i}E_i.$$

where U and E be independent univariate random variables following the standard normal distribution $\mathsf{N}(0,1)$ with pdf $\phi(.)$, $|U|$ denotes the magnitude of U and $\delta_{z_i} = \frac{\lambda_{z_i}}{\sqrt{1+\lambda_{z_i}^2}}$ where $Z_i \in \{1,\ldots,K\}$ is a categorical variable Z_i which follows the multinomial distribution

$$Z_i|\boldsymbol{r}_i \sim \mathsf{Mult}(1; \pi_1(\boldsymbol{r}_i;\boldsymbol{\alpha}),\ldots,\pi_K(\boldsymbol{r}_i;\boldsymbol{\alpha}))$$

where each of the probabilities $\pi_{z_i}(\boldsymbol{r}_i;\boldsymbol{\alpha}) = \mathbb{P}(Z_i = z_i|\boldsymbol{r}_i)$ is given by the logistic function.

The skew-normal mixture of experts model

The SNMoE model is characterized as follows.

- The stochastic representation of the SNMoE leads to the following hierarchical representation
- **Hierarchical representation of the SNMoE**

$$\begin{aligned} Y_i|u_i, Z_{ik}=1, \boldsymbol{x}_i &\sim \mathsf{N}\Big(\mu(\boldsymbol{x}_i;\boldsymbol{\beta}_k)+\delta_k|u_i|, (1-\delta_k^2)\sigma_k^2\Big), \\ U_i|Z_{ik}=1 &\sim \mathsf{N}(0,\sigma_k^2), \\ \boldsymbol{Z}_i|\boldsymbol{r}_i &\sim \mathsf{Mult}\left(1;\pi_1(\boldsymbol{r}_i;\boldsymbol{\alpha}),\ldots,\pi_K(\boldsymbol{r}_i;\boldsymbol{\alpha})\right) \end{aligned}$$

  where Z_{ik} are the binary latent component-indicators such that $Z_{ik}=1$ iff $Z_i=k$, $\boldsymbol{Z}_i=(Z_{i1},\ldots,Z_{iK})$ and $\delta_k=\frac{\lambda_k}{\sqrt{1+\lambda_k^2}}$

- This hierarchical incomplete data representation facilitates the inference scheme by using the ECM algorithm.

MLE via the ECM algorithm

- Given an observed i.i.d sample of n observations $\{(y_i, \boldsymbol{x}_i, \boldsymbol{r}_i)\}_{i=1}^{n}$, the parameter vector $\boldsymbol{\Psi}$ of the SNMoE model can be estimated by maximizing the observed-data log-likelihood:

$$\log L(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k(\boldsymbol{r}_i; \boldsymbol{\alpha}) \mathrm{SN}\left(y; \mu(\boldsymbol{x}; \boldsymbol{\beta}_k), \sigma_k^2, \lambda_k\right).$$

- $\Rightarrow$ A dedicated Expectation Conditional Maximization (ECM) algorithm
- The ECM algorithm (Meng and Rubin, 1993) is an EM variant that mainly aims at addressing the optimization problem in the M-step of the EM algorithm. In ECM, the M-step is performed by several conditional maximization (CM) steps by dividing the parameter space into sub-spaces. The parameter vector updates are then performed sequentially, one coordinate block after another in each sub-space.

Maximum likelihood estimation via the ECM algorithm

- The complete-data log-likelihood of $\boldsymbol{\Psi}$, where the complete-data are $\{y_i, z_i, u_i, \boldsymbol{x}_i, \boldsymbol{r}_i\}_{i=1}^n$, is given by:

$$\log L_c(\boldsymbol{\Psi}) \quad = \quad \log L_c(\boldsymbol{\alpha}) + \sum_{k=1}^{K} \log L_c(\boldsymbol{\Psi}_k),$$

with

$$\log L_c(\boldsymbol{\alpha}) \;=\; \sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik} \log \pi_k(\boldsymbol{r}_i; \boldsymbol{\alpha}),$$

$$\begin{aligned}\log L_c(\boldsymbol{\Psi}_k) \;=\; \sum_{i=1}^{n} Z_{ik} \Big[& -\log(2\pi) - \log(\sigma_k^2) - \frac{1}{2}\log(1-\delta_k^2) \\ & - \frac{d_{ik}^2}{2(1-\delta_k^2)} + \frac{\delta_k \; d_{ik} \; u_i}{(1-\delta_k^2)\sigma_k} - \frac{u_i^2}{2(1-\delta_k^2)\sigma_k^2}\Big],\end{aligned}$$

where $d_{ik} = \frac{y_i - \mu(\boldsymbol{x}_i;\boldsymbol{\beta}_k)}{\sigma_k}$.

ECM for the SNMoE: E-Step

E-Step calculates the Q-function

$$Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(m)})=\mathbb{E}\big[\log L_c(\boldsymbol{\Psi})|\{y_i,\boldsymbol{x}_i,\boldsymbol{r}_i\}_{i=1}^n;\boldsymbol{\Psi}^{(m)}\big]=Q_1(\boldsymbol{\alpha};\boldsymbol{\Psi}^{(m)})+\sum_{k=1}^{K}Q_2(\boldsymbol{\Psi}_k;\boldsymbol{\Psi}^{(m)})$$

with

$$\begin{aligned}
Q_1(\boldsymbol{\alpha};\boldsymbol{\Psi}^{(m)}) &= \sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}^{(m)}\log\pi_k(\boldsymbol{r}_i;\boldsymbol{\alpha}),\\
Q_2(\boldsymbol{\Psi}_k;\boldsymbol{\Psi}^{(m)}) &= \sum_{i=1}^{n}\tau_{ik}^{(m)}\Bigg[-\log(2\pi)-\log(\sigma_k^2)-\frac{1}{2}\log(1-\delta_k^2)\\
&\qquad +\frac{\delta_k\ d_{ik}\ e_{1,ik}^{(m)}}{(1-\delta_k^2)\sigma_k}-\frac{e_{2,ik}^{(m)}}{2(1-\delta_k^2)\sigma_k^2}-\frac{d_{ik}^2}{2(1-\delta_k^2)}\Bigg]
\end{aligned}$$

where the required conditional expectations (analytic) are given by:

$$\begin{aligned}
\tau_{ik}^{(m)} &= \mathbb{E}_{\boldsymbol{\Psi}^{(m)}}\left[Z_{ik}|y_i,\boldsymbol{x}_i,\boldsymbol{r}_i\right],\\
e_{1,ik}^{(m)} &= \mathbb{E}_{\boldsymbol{\Psi}^{(m)}}\left[U_i|Z_{ik}=1,y_i,\boldsymbol{x}_i,\boldsymbol{r}_i\right],\\
e_{2,ik}^{(m)} &= \mathbb{E}_{\boldsymbol{\Psi}^{(m)}}\left[U_i^2|Z_{ik}=1,y_i,\boldsymbol{x}_i,\boldsymbol{r}_i\right].
\end{aligned}$$

CM-Step 1 Calculate $\boldsymbol{\alpha}^{(m+1)} = \arg\max_{\boldsymbol{\alpha}} Q_1(\boldsymbol{\alpha}; \boldsymbol{\Psi}^{(m)})$. does not exist in closed form (Unlike in skew-normal (regression) mixtures)

The Iteratively Reweighted Least Squares (IRLS) algorithm:

$$\boldsymbol{\alpha}^{(l+1)} = \boldsymbol{\alpha}^{(l)} - \left[\frac{\partial^2 Q_1(\boldsymbol{\alpha}, \boldsymbol{\Psi}^{(m)})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T}\right]^{-1}_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^{(l)}} \frac{\partial Q_1(\boldsymbol{\alpha}, \boldsymbol{\Psi}^{(m)})}{\partial \boldsymbol{\alpha}}\Big|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^{(l)}}$$

Then, for $k = 1 \ldots, K$,

CM-Step 2 Calculate $\boldsymbol{\beta}_k^{(m+1)}$ by maximizing $Q_2(\boldsymbol{\Psi}_k; \boldsymbol{\Psi}^{(m)})$

$$\boldsymbol{\beta}_k^{(m+1)} = \left[\sum_{i=1}^{n} \tau_{ik}^{(m)} \boldsymbol{x}_i \boldsymbol{x}_i^T\right]^{-1} \sum_{i=1}^{n} \tau_{ik}^{(q)} \left(y_i - \delta_k^{(m)} e_{1,ik}^{(m)}\right) \boldsymbol{x}_i.$$

CM-Step 3: Calculate $\sigma_k^{2(m+1)}$ by maximizing $Q_2(\boldsymbol{\Psi}_k; \boldsymbol{\Psi}^{(m)})$

$$\sigma_k^{2(m+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(m)} \left[\left(y_i - \boldsymbol{\beta}_k^{T(m+1)} \boldsymbol{x}_i\right)^2 - 2\delta_k^{(m+1)} e_{1,ik}^{(m)} (y_i - \boldsymbol{\beta}_k^{T(m+1)} \boldsymbol{x}_i) + e_{2,ik}^{(m)}\right]}{2\left(1 - \delta_k^{2(m)}\right) \sum_{i=1}^{n} \tau_{ik}^{(m)}}.$$

CM-Step 4 Calculate $\lambda_k^{(m+1)}$ by maximizing $Q_2(\boldsymbol{\Psi}_k; \boldsymbol{\Psi}^{(m)})$: Solution of:
$\sigma_k^{2(m+1)} \delta_k (1 - \delta_k^2) \sum_{i=1}^{n} \tau_{ik}^{(m)} + (1 + \delta_k^2) \sum_{i=1}^{n} \tau_{ik}^{(m)} (y_i - \boldsymbol{\beta}_k^{T(m+1)} \boldsymbol{x}_i)\, e_{1,ik}^{(m)}$
$- \delta_k \sum_{i=1}^{n} \tau_{ik}^{(m)} \left[e_{2,ik}^{(m)} + \left(y_i - \boldsymbol{\beta}_k^{T(m+1)} \boldsymbol{x}_i\right)^2\right] = 0$· root finding (Brent's method

- However, while the SNMoE model is tailored to model the skewness in the data, it may be not adapted to handle data containing groups or a group with heavy-tailed distribution.
- The NMoE and the SNMoE may thus be affected by outliers.
- $\Rightarrow$ Handle the problem of sensitivity of normal mixture of experts to outliers and heavy tails. I first propose a robust mixture of experts modeling by using the t distribution.

The t mixture of experts model

- The proposed t mixture of experts (TMoE) model is based on the t distribution, which is robust generalization of the normal distribution.
- The t distribution is more robust than the normal distribution to handle outliers in the data and to accommodate data with heavy tailed distribution.
- This has been shown in terms of density modeling and cluster analysis for multivariate data (Mclachlan and Peel, 1998; Peel and Mclachlan, 2000) as well as for univariate data (Lin et al., 2007a) and regression mixtures (Bai et al., 2012; Wei, 2012; Ingrassia et al., 2012).
- The t-distribution with location $\mu \in \mathbb{R}$, scale $\sigma^2 \in (0, \infty)$ and degrees of freedom $\nu \in (0, \infty)$ has the probability density function

$$f(y; \mu, \sigma^2, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})} \left(1 + \frac{d_y^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

where $d_y^2 = \left(\frac{y-\mu}{\sigma}\right)^2$ denotes the squared Mahalanobis distance

The t mixture of experts model

- The proposed t mixture of experts model extends the t mixture model, first proposed by Mclachlan and Peel (1998); Peel and Mclachlan (2000) for multivariate data, as well as the regression mixture model using the t-distribution as in (Bai et al., 2012; Wei, 2012; Ingrassia et al., 2012) to the MoE framework.
- A K-component TMoE model is defined by:

$$f(y|\boldsymbol{r},\boldsymbol{x};\boldsymbol{\Psi}) \;=\; \sum_{k=1}^{K}\pi_k(\boldsymbol{r};\boldsymbol{\alpha})\; t_{\nu_k}\left(y;\mu(\boldsymbol{x};\boldsymbol{\beta}_k),\sigma_k^2,\nu_k\right).$$

- The parameter vector of the TMoE model is given by $\boldsymbol{\Psi}=(\boldsymbol{\alpha}_1^T,\ldots,\boldsymbol{\alpha}_{K-1}^T,\boldsymbol{\Psi}_1^T,\ldots,\boldsymbol{\Psi}_K^T)^T$ where $\boldsymbol{\Psi}_k=(\boldsymbol{\beta}_k^T,\sigma_k^2,\nu_k)^T$
- When the robustness parameter $\nu_k\to\infty$ for each experts k, the TMoE model approaches the NMoE model

The t mixture of experts model

- **Stochastic representation for the TMoE** Let $E \sim \phi(.)$. Suppose that, conditional on the hidden variable $Z_i = z_i$, a random variable W_i is distributed as Gamma$(\frac{\nu_{z_i}}{2}, \frac{\nu_{z_i}}{2})$. Then, given the covariates $(\boldsymbol{x}_i, \boldsymbol{r}_i)$, a random variable Y_i is said to follow the TMoE model if

$$Y_i = \mu(\boldsymbol{x}_i; \boldsymbol{\beta}_{z_i}) + \sigma_{z_i} \frac{E_i}{\sqrt{W_{z_i}}},$$

  where the categorical variable $Z_i|\boldsymbol{r}_i$ is multinomial
- **Hierarchical representation of the TMoE model**

$$\begin{aligned} Y_i|w_i, Z_{ik} = 1, \boldsymbol{x}_i &\sim \mathsf{N}\left(\mu(\boldsymbol{x}_i; \boldsymbol{\beta}_k), \frac{\sigma_k^2}{w_i}\right), \\ W_i|Z_{ik} = 1 &\sim \mathsf{Gamma}\left(\frac{\nu_k}{2}, \frac{\nu_k}{2}\right) \\ \boldsymbol{Z}_i|\boldsymbol{r}_i &\sim \mathsf{Mult}\left(1; \pi_1(\boldsymbol{r}_i; \boldsymbol{\alpha}), \ldots, \pi_K(\boldsymbol{r}_i; \boldsymbol{\alpha})\right). \end{aligned}$$

- This hierarchical representation involves the hidden variables Z_i and W_i facilitates the ML inference of model parameters $\boldsymbol{\Psi}$ via E(C)M.

MLE of the TMoE model

- Given an i.i.d sample of n observations, $\boldsymbol{\Psi}$ can be estimated by maximizing the observed-data log-likelihood:

$$\log L(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k(\boldsymbol{r}_i; \boldsymbol{\alpha}) t\nu_k \left(y; \mu(\boldsymbol{x}; \boldsymbol{\beta}_k), \sigma_k^2, \nu_k\right).$$

- $\Rightarrow$ EM algorithm and then describe an ECM extension
- The complete data consist of the responses $(y_1, \ldots, y_n)$ and their corresponding predictors $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ and $(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n)$, as well as the latent variables $(w_1, \ldots, w_n)$ (in the hierarchical representation) and the latent labels $(z_1, \ldots, z_n)$.

MLE of the TMoE model

- $\Rightarrow$ The complete-data log-likelihood of $\boldsymbol{\Psi}$ is given by:

$$\log L_c(\boldsymbol{\Psi}) \quad = \quad \log L_{1c}(\boldsymbol{\alpha}) + \sum_{k=1}^{K} \big[\log L_{2c}(\boldsymbol{\Psi}_k) + \log L_{3c}(\nu_k) \big],$$

where

$$\log L_{1c}(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{ik} \log \pi_k(\boldsymbol{r}_i; \boldsymbol{\alpha}),$$

$$\log L_{2c}(\boldsymbol{\Psi}_k) = \sum_{i=1}^{n} Z_{ik} \Big[-\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma_k^2) - \frac{1}{2} w_i d_{ik}^2 \Big],$$

$$\log L_{3c}(\nu_k) = \sum_{i=1}^{n} Z_{ik} \Big[-\log \Gamma\left(\frac{\nu_k}{2}\right) + \left(\frac{\nu_k}{2}\right) \log\left(\frac{\nu_k}{2}\right) + \left(\frac{\nu_k}{2} - 1\right) \log(w_i) - \left(\frac{\nu_k}{2}\right) w_i \Big].$$

MLE of the TMoE model: E-Step

E-Step Calculate the Q-function:

$$Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(m)}) = Q_1(\boldsymbol{\alpha};\boldsymbol{\Psi}^{(m)}) + \sum_{k=1}^{K}\Big[Q_2(\boldsymbol{\theta}_k,\boldsymbol{\Psi}^{(m)}) + Q_3(\nu_k,\boldsymbol{\Psi}^{(m)})\Big],$$

where $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k^T, \sigma_k^2)^T$ and

$$Q_1(\boldsymbol{\alpha};\boldsymbol{\Psi}^{(m)}) = \sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}^{(m)}\log\pi_k(\boldsymbol{r}_i;\boldsymbol{\alpha}),$$

$$Q_2(\boldsymbol{\theta}_k;\boldsymbol{\Psi}^{(m)}) = \sum_{i=1}^{n}\tau_{ik}^{(m)}\Big[-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma_k^2) - \frac{1}{2}\,w_{ik}^{(m)}d_{ik}^2\Big].$$

$$Q_3(\nu_k;\boldsymbol{\Psi}^{(m)}) = \sum_{i=1}^{n}\tau_{ik}^{(m)}\Big[-\log\Gamma\Big(\frac{\nu_k}{2}\Big) + \Big(\frac{\nu_k}{2}\Big)\log\Big(\frac{\nu_k}{2}\Big) - \Big(\frac{\nu_k}{2}\Big)\,w_{ik}^{(m)} + \Big(\frac{\nu_k}{2}-1\Big)e_{1,ik}^{(m)}\Big]$$

$\rightarrow$ requires the following conditional expectations (analytic):

$$\begin{aligned}
\tau_{ik}^{(m)} &= \mathbb{E}_{\boldsymbol{\Psi}^{(m)}}\left[Z_{ik}|y_i,\boldsymbol{x}_i,\boldsymbol{r}_i\right],\\
w_{ik}^{(m)} &= \mathbb{E}_{\boldsymbol{\Psi}^{(m)}}\left[W_i|y_i,Z_{ik}=1,\boldsymbol{x}_i,\boldsymbol{r}_i\right],\\
e_{1,ik}^{(m)} &= \mathbb{E}_{\boldsymbol{\Psi}^{(m)}}\left[\log(W_i)|y_i,Z_{ik}=1,\boldsymbol{x}_i,\boldsymbol{r}_i\right].
\end{aligned}$$

MLE of the TMoE model: M-Step

M-Step 1 Calculate $\boldsymbol{\alpha}^{(m+1)}$ by maximizing $Q_1(\boldsymbol{\alpha}; \boldsymbol{\Psi}^{(m)})$ w.r.t $\boldsymbol{\alpha}$. $\Rightarrow$ Iteratively via IRLS (73) as for the mixture of SNMoE.

M-Step 2 Calculate $\boldsymbol{\theta}_k^{(m+1)}$ by maximizing $Q_2(\boldsymbol{\theta}_k; \boldsymbol{\Psi}^{(m)})$ w.r.t $\boldsymbol{\theta}_k$

$$\begin{aligned}
\boldsymbol{\beta}_k^{(m+1)} &= \Big[\sum_{i=1}^{n} \tau_{ik}^{(m)} w_{ik}^{(m)} \boldsymbol{x}_i \boldsymbol{x}_i^T\Big]^{-1} \sum_{i=1}^{n} \tau_{ik}^{(q)} w_{ik}^{(m)} y_i \boldsymbol{x}_i, \\
\sigma_k^{2(m+1)} &= \frac{1}{\sum_{i=1}^{n} \tau_{ik}^{(m)}} \sum_{i=1}^{n} \tau_{ik}^{(m)} w_{ik}^{(m)} \left(y_i - \boldsymbol{\beta}_k^{T(m+1)} \boldsymbol{x}_i\right)^2.
\end{aligned}$$

M-Step 3 Calculate $\nu_k^{(m+1)}$ by maximizing $Q_3(\nu_k; \boldsymbol{\Psi}^{(m)})$ w.r.t ν_k
$\Rightarrow$ iteratively solve the following equation in ν_k:

$$-\psi\left(\frac{\nu_k}{2}\right) + \log\left(\frac{\nu_k}{2}\right) + 1 + \frac{\sum_{i=1}^{n} \tau_{ik}^{(m)} \big(\log(w_{ik}^{(m)}) - w_{ik}^{(m)}\big)}{\sum_{i=1}^{n} \tau_{ik}^{(m)}} + \psi\big(\frac{\nu_k^{(m)} + 1}{2}\big) - \log\big(\frac{\nu_k^{(m)} + 1}{2}\big) = 0.$$

This scalar non-linear equation can be solved with a root finding algorithm, such as Brent's method (Brent, 1973).

The skew t mixture of experts model

- The proposed skew t mixture of experts (STMoE) model is a MoE model in which the expert components have a skew-t density
- The skew t distribution Azzalini and Capitanio (2003), can be characterized as follows. Let U be an univariate standard skew-normal variable $U \sim \text{SN}(0, 1, \lambda)$. Then, let $W \perp U \sim \text{Gamma}(\frac{\nu}{2}, \frac{\nu}{2})$. A random variable Y having the following representation:

$$Y = \mu + \sigma \frac{U}{\sqrt{W}}$$

follows the skew t distribution $\text{ST}(\mu, \sigma^2, \lambda, \nu)$ with location μ, scale σ, skewness λ and degrees of freedom ν, whose density is defined by:

$$f(y; \mu, \sigma^2, \lambda, \nu) = \frac{2}{\sigma}\, t_\nu(d_y)\, T_{\nu+1}\left(\lambda\, d_y \sqrt{\frac{\nu+1}{\nu + d_y^2}}\right)$$

where $d_y = \frac{y-\mu}{\sigma}$ and $t_\nu(.)$ and $T_\nu(.)$ respectively denote the pdf and the cdf of the standard t distribution with degrees of freedom ν.

The skew t mixture of experts (STMoE) model

- The proposed skew t mixture of experts (STMoE) model extends the univariate skew t mixture model Lin et al. (2007a), to the MoE framework.
- A K-component mixture of skew t experts (STMoE) is defined by:

$$f(y|\boldsymbol{r},\boldsymbol{x};\boldsymbol{\Psi}) \quad = \quad \sum_{k=1}^{K} \pi_k(\boldsymbol{r};\boldsymbol{\alpha})\, \mathsf{ST}(y;\mu(\boldsymbol{x};\boldsymbol{\beta}_k),\sigma_k^2,\lambda_k,\nu_k)\,.$$

- Parameter vector: $\boldsymbol{\Psi} = (\boldsymbol{\alpha}_1^T,\ldots,\boldsymbol{\alpha}_{K-1}^T,\boldsymbol{\Psi}_1^T,\ldots,\boldsymbol{\Psi}_K^T)^T$ where $\boldsymbol{\Psi}_k = (\boldsymbol{\beta}_k^T,\sigma_k^2,\lambda_k,\nu_k)^T$ is the parameter vector for the kth skew t expert component whose density is defined by

$$f\big(y|\boldsymbol{x};\mu(\boldsymbol{x};\boldsymbol{\beta}_k),\sigma^2,\lambda,\nu\big) = \frac{2}{\sigma}\; t_\nu(d_y(\boldsymbol{x}))\; T_{\nu+1}\left(\lambda\; d_y(\boldsymbol{x})\sqrt{\frac{\nu+1}{\nu+d_y^2(\boldsymbol{x})}}\right)$$

- When the robustness parameter $\{\nu_k\} \to \infty$, the STMoE reduces to the SNMoE. If the skewness parameter $\{\lambda_k\} = 0$, the STMoE reduces to the TMoE. Moreover, when $\{\nu_k\} \to \infty$ and $\{\lambda_k\} = 0$, it approaches the NMoE.
- $\Rightarrow$ The STMoE is more flexible as it generalizes the previously described models to accommodate situations with asymmetry, heavy tails, and outliers.

Representation of the STMoE model

- **Stochastic representation** Suppose that conditional on a Multinomial categorical variable Z_i, E_i and W_i are independent univariate random variables such that $E_i \sim \mathsf{SN}(\lambda_{z_i})$ and $W_i \sim \mathsf{Gamma}(\frac{\nu_{z_i}}{2}, \frac{\nu_{z_i}}{2})$, and $\boldsymbol{x}_i$ and $\boldsymbol{r}_i$ are given covariates. A variable Y_i having the following representation:

$$Y_i = \mu(\boldsymbol{x}_i; \boldsymbol{\beta}_{z_i}) + \sigma_{z_i} \frac{E_i}{\sqrt{W_i}}$$

  is said to follow the STMoE distribution

- **Hierarchical representation**

$$\begin{aligned}
Y_i|u_i, w_i, Z_{ik} = 1, \boldsymbol{x}_i &\sim \mathsf{N}\left(\mu(\boldsymbol{x}_i; \boldsymbol{\beta}_k) + \delta_k|u_i|, \frac{1-\delta_k^2}{w_i}\sigma_k^2\right), \\
U_i|w_i, Z_{ik} = 1 &\sim \mathsf{N}\left(0, \frac{\sigma_k^2}{w_i}\right), \\
W_i|Z_{ik} = 1 &\sim \mathsf{Gamma}\left(\frac{\nu_k}{2}, \frac{\nu_k}{2}\right) \\
\boldsymbol{Z}_i|\boldsymbol{r}_i &\sim \mathsf{Mult}\big(1; \pi_1(\boldsymbol{r}_i; \boldsymbol{\alpha}), \ldots, \pi_K(\boldsymbol{r}_i; \boldsymbol{\alpha})\big).
\end{aligned}$$

  The variables U_i and W_i are hidden in this hierarchical representation

Identifiability of the STMoE model

Ordered, initialized, and irreducible STMoEs are identifiable:

- Ordered implies that there exist a certain ordering relationship such that $(\boldsymbol{\beta}_1^T, \sigma_1^2, \lambda_1, \nu_1)^T \prec \ldots \prec (\boldsymbol{\beta}_K^T, \sigma_K^2, \lambda_K, \nu_K)^T$;
- initialized implies that $\mathbf{w}_K$ is the null vector, as assumed in the model
- irreducible implies that if $k \neq k\prime$, then one of the following conditions holds: $\boldsymbol{\beta}_k \neq \boldsymbol{\beta}_{k\prime}$, $\sigma_k \neq \sigma_{k\prime}$, $\lambda_k \neq \lambda_{k\prime}$ or $\nu_k \neq \nu_{k\prime}$.

$\Rightarrow$ Then, we can establish the identifiability of ordered and initialized irreducible STMoE models by applying Lemma 2 of Jiang and Tanner (1999), which requires the validation of the following nondegeneracy condition:

- The set $\{\mathsf{ST}(y; \mu(\boldsymbol{x}; \boldsymbol{\beta}_1), \sigma_1^2, \lambda_1, \nu_1), \ldots, \mathsf{ST}(y; \mu(\boldsymbol{x}; \boldsymbol{\beta}_{4K}), \sigma_{4K}^2, \lambda_{4K}, \nu_{4K})\}$ contains $4K$ linearly independent functions of y, for any $4K$ distinct quadruplet $(\mu(\boldsymbol{x}; \boldsymbol{\beta}_k), \sigma_k^2, \lambda_k, \nu_k)$ for $k = 1, \ldots, 4K$.
- Thus, via Lemma 2 of Jiang and Tanner (1999) we have any ordered and initialized irreducible STMoE is identifiable.

MLE via the ECM algorithm

- Maximize the observed-data log-likelihood:

$$\log L(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k(\boldsymbol{r}_i; \boldsymbol{\alpha}) \mathsf{ST}(y; \mu(\boldsymbol{x}_i; \boldsymbol{\beta}_k), \sigma_k^2, \lambda_k, \nu_k) \cdot$$

- $\Rightarrow$ This is performed iteratively by a dedicated ECM algorithm.
- The complete-data log-likelihood:

$$\log L_c(\boldsymbol{\Psi}) = \log L_{1c}(\boldsymbol{\alpha}) + \sum_{k=1}^{K} \big[\log L_{2c}(\boldsymbol{\theta}_k) + \log L_{3c}(\nu_k) \big]; \ \boldsymbol{\theta}_k = (\boldsymbol{\beta}_k^T, \sigma_k^2, \lambda_k)^T$$

$$\log L_{1c}(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{ik} \log \pi_k(\boldsymbol{r}_i; \boldsymbol{\alpha}),$$

$$\log L_{2c}(\boldsymbol{\theta}_k) = \sum_{i=1}^{n} Z_{ik} \Big[-\log(2\pi) - \log(\sigma_k^2) - \frac{1}{2}\log(1-\delta_k^2) - \frac{w_i \ d_{ik}^2}{2(1-\delta_k^2)} + \frac{w_i \ u_i \ \delta_k \ d_{ik}}{(1-\delta_k^2)\sigma_k} - \frac{w_i \ u_i^2}{2(1-\delta_k^2)\sigma_k^2} \Big]$$

$$\log L_{3c}(\nu_k) = \sum_{i=1}^{n} Z_{ik} \Big[-\log\Gamma\left(\frac{\nu_k}{2}\right) + \left(\frac{\nu_k}{2}\right)\log\left(\frac{\nu_k}{2}\right) + \left(\frac{\nu_k}{2}\right)\log(w_i) - \left(\frac{\nu_k}{2}\right) w_i \Big] \cdot$$

MLE via the ECM algorithm: E-Step

- **E-Step** Calculates the Q-function, that is the conditional expectation of the complete-data log-likelihood , given the observed data $\{y_i, \boldsymbol{x}_i, \boldsymbol{r}_i\}_{i=1}^n$ and a current parameter estimation $\boldsymbol{\Psi}^{(m)}$ given by:

$$Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(m)}) = Q_1(\boldsymbol{\alpha}; \boldsymbol{\Psi}^{(m)}) + \sum_{k=1}^{K} \left[Q_2(\boldsymbol{\theta}_k, \boldsymbol{\Psi}^{(m)}) + Q_3(\nu_k, \boldsymbol{\Psi}^{(m)}) \right],$$

where

$$Q_1(\boldsymbol{\alpha}; \boldsymbol{\Psi}^{(m)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(m)} \log \pi_k(\boldsymbol{r}_i; \boldsymbol{\alpha}),$$

$$Q_2(\boldsymbol{\theta}_k; \boldsymbol{\Psi}^{(m)}) = \sum_{i=1}^{n} \tau_{ik}^{(m)} \left[-\log(2\pi\sigma_k^2) - \frac{1}{2}\log(1-\delta_k^2) - \frac{w_{ik}^{(m)}\ d_{ik}^2}{2(1-\delta_k^2)} + \frac{\delta_k\ d_{ik}\ e_{1,ik}^{(m)}}{(1-\delta_k^2)\sigma_k} - \frac{e_{2,ik}^{(m)}}{2(1-\delta_k^2)\sigma_k^2} \right],$$

$$Q_3(\nu_k; \boldsymbol{\Psi}^{(m)}) = \sum_{i=1}^{n} \tau_{ik}^{(m)} \left[-\log\Gamma\left(\frac{\nu_k}{2}\right) + \left(\frac{\nu_k}{2}\right)\log\left(\frac{\nu_k}{2}\right) - \left(\frac{\nu_k}{2}\right)\ w_{ik}^{(m)} + \left(\frac{\nu_k}{2}\right) e_{3,ik}^{(m)} \right].$$

MLE via the ECM algorithm: E-Step

- $\Rightarrow$ The E-Step requires the following conditional expectations:

$$\begin{aligned}
\tau_{ik}^{(m)} &= \mathbb{E}_{\boldsymbol{\Psi}^{(m)}}\left[Z_{ik}|y_i, \boldsymbol{x}_i, \boldsymbol{r}_i\right], \\
w_{ik}^{(m)} &= \mathbb{E}_{\boldsymbol{\Psi}^{(m)}}\left[W_i|y_i, Z_{ik}=1, \boldsymbol{x}_i, \boldsymbol{r}_i\right], \\
e_{1,ik}^{(m)} &= \mathbb{E}_{\boldsymbol{\Psi}^{(m)}}\left[W_i U_i|y_i, Z_{ik}=1, \boldsymbol{x}_i, \boldsymbol{r}_i\right], \\
e_{2,ik}^{(m)} &= \mathbb{E}_{\boldsymbol{\Psi}^{(m)}}\left[W_i U_i^2|y_i, Z_{ik}=1, \boldsymbol{x}_i, \boldsymbol{r}_i\right], \\
e_{3,ik}^{(m)} &= \mathbb{E}_{\boldsymbol{\Psi}^{(m)}}\left[\log(W_i)|y_i, Z_{ik}=1, \boldsymbol{x}_i, \boldsymbol{r}_i\right].
\end{aligned}$$

- These conditional expectations are calculated analytically except $e_{3,ik}^{(m)}$ for which I adopted a one-step-late (OSL) approach as in Lee and McLachlan (2014), rather than using a Monte Carlo approximation as in Lin et al. (2007a).
- I also mention that, for the multivariate skew t mixture models, recently Lee and McLachlan (2015) presented a series-based truncation approach, which exploits an exact representation of this conditional expectation and which can also be used here.

MLE via the ECM algorithm: M-Step

- **CM-Step 1** update the mixing parameters $\boldsymbol{\alpha}^{(m+1)}$ by maximizing the function $Q_1(\boldsymbol{\alpha}; \boldsymbol{\Psi}^{(m)})$ by using IRLS. Then, for $k = 1 \ldots, K$,
- **CM-Step 2** Update the regression params $(\boldsymbol{\beta}_k^{T(m+1)}, \sigma_k^{2(m+1)})$:

$$\boldsymbol{\beta}_k^{(m+1)} = \left[\sum_{i=1}^{n} \tau_{ik}^{(q)} w_{ik}^{(m)} \boldsymbol{x}_i \boldsymbol{x}_i^T\right]^{-1} \sum_{i=1}^{n} \tau_{ik}^{(q)} \left(w_{ik}^{(m)} y_i - \boldsymbol{e}_{1,ik}^{(m)} \delta_k^{(m+1)}\right) \boldsymbol{x}_i,$$

$$\sigma_k^{2(m+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(m)} \left[w_{ik}^{(m)} \left(\boldsymbol{y}_i - \boldsymbol{\beta}_k^{T(m+1)} \boldsymbol{x}_i\right)^2 - 2\delta_k^{(m+1)} \boldsymbol{e}_{1,ik}^{(m)} (y_i - \boldsymbol{\beta}_k^{T(m+1)} \boldsymbol{x}_i) + \boldsymbol{e}_{2,ik}^{(m)}\right]}{2\left(1 - \delta_k^{2(m)}\right) \sum_{i=1}^{n} \tau_{ik}^{(m)}}$$

- **CM-Step 3** Update the skewness parameters λ_k by solving the following equation:

$$\delta_k (1 - \delta_k^2) \sum_{i=1}^{n} \tau_{ik}^{(m)} + (1 + \delta_k^2) \sum_{i=1}^{n} \tau_{ik}^{(m)} \frac{d_{ik}^{(m+1)} e_{1,ik}^{(m)}}{\sigma_k^{(m+1)}} - \delta_k \sum_{i=1}^{n} \tau_{ik}^{(m)} \left[w_{ik}^{(m)} d_{ik}^{2\,(m+1)} + \frac{e_{2,ik}^{(m)}}{\sigma_k^{2\,(m+1)}}\right] = 0 \cdot$$

- **CM-Step 4** Update the degree of freedom ν_k by solving of the following equation:

$$-\psi\left(\frac{\nu_k}{2}\right) + \log\left(\frac{\nu_k}{2}\right) + 1 + \frac{\sum_{i=1}^{n} \tau_{ik}^{(m)} \left(\boldsymbol{e}_{3,ik}^{(m)} - w_{ik}^{(m)}\right)}{\sum_{i=1}^{n} \tau_{ik}^{(m)}} = 0.$$

Prediction, clustering and model selection

- **Prediction** Predicted response: $\hat{y} = \mathbb{E}_{\hat{\boldsymbol{\Psi}}}(Y|\boldsymbol{r}, \boldsymbol{x})$ with

$$\mathbb{E}_{\hat{\boldsymbol{\Psi}}}(Y|\boldsymbol{r}, \boldsymbol{x}) = \sum_{k=1}^{K} \pi_k(\boldsymbol{r}; \hat{\boldsymbol{\alpha}}_n) \mathbb{E}_{\hat{\boldsymbol{\Psi}}}(Y|Z = k, \boldsymbol{x}),$$

$$\mathbb{V}_{\hat{\boldsymbol{\Psi}}}(Y|\boldsymbol{r}, \boldsymbol{x}) = \sum_{k=1}^{K} \pi_k(\boldsymbol{r}; \hat{\boldsymbol{\alpha}}_n) \left[\left(\mathbb{E}_{\hat{\boldsymbol{\Psi}}}(Y|Z = k, \boldsymbol{x})\right)^2 + \mathbb{V}_{\hat{\boldsymbol{\Psi}}}(Y|Z = k, \boldsymbol{x}) \right] - \left[\mathbb{E}_{\hat{\boldsymbol{\Psi}}}(Y|\boldsymbol{r}, \boldsymbol{x})\right]$$

  where $\mathbb{E}_{\hat{\boldsymbol{\Psi}}}(Y|Z = k, \boldsymbol{x})$ and $\mathbb{V}_{\hat{\boldsymbol{\Psi}}}(Y|Z = k, \boldsymbol{x})$ are respectively the component-specific (expert) means and variances.

- **Clustering of regression data** Calculate the cluster label as

$$\hat{z}_i = \arg\max_{k=1}^{K} \mathbb{E}[Z_i|\boldsymbol{r}_i, \boldsymbol{x}_i; \hat{\boldsymbol{\Psi}}] = \arg\max_{k=1}^{K} \frac{\pi_k(\boldsymbol{r}; \hat{\boldsymbol{\Psi}}) f_k\left(y_i|\boldsymbol{r}_i, \boldsymbol{x}_i; \hat{\boldsymbol{\Psi}}\right)}{\sum_{k\prime=1}^{K} \pi_{k\prime}(\boldsymbol{r}; \hat{\boldsymbol{\alpha}}) f_{k\prime}\left(y_i|\boldsymbol{r}_i, \boldsymbol{x}_i; \hat{\boldsymbol{\Psi}}_{k\prime}\right)}$$

- **Model selection** The value of (K, p) can be computed by using BIC, ICL
  Number of free parameters:
  $\eta_{\boldsymbol{\Psi}} = K(p + 4) - 2$ for the NMoE model,
  $\eta_{\boldsymbol{\Psi}} = K(p + 5) - 2$ for both the SNMoE and the TMoE models,
  $\eta_{\boldsymbol{\Psi}} = K(p + 6) - 2$ for the STMoE model.

Illustation on Bishop's data set

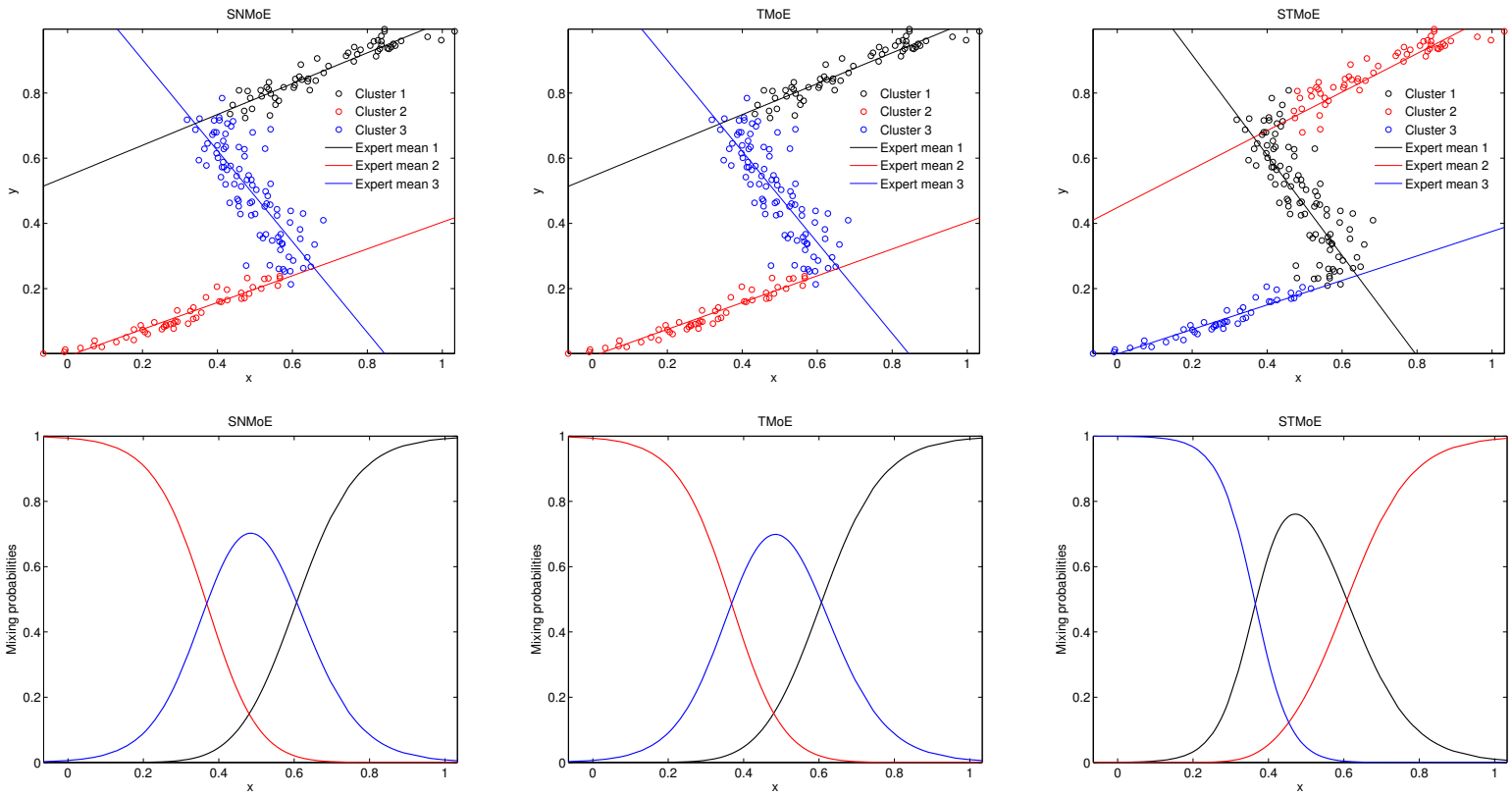


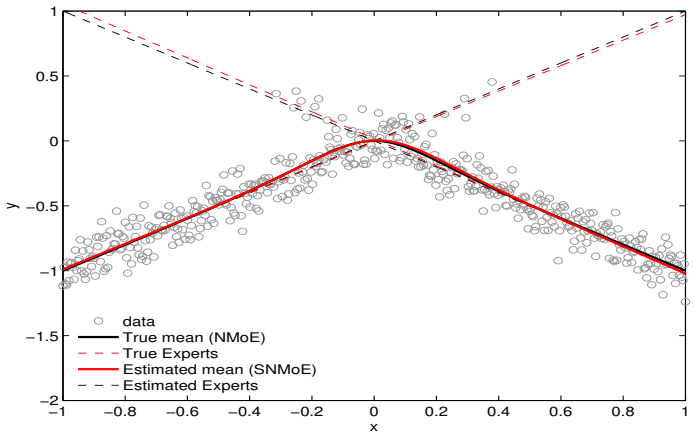
Figure: Fitting the the non-normal mixture of experts models (SNMoE, TNMoE, STMoE) to the toy data set analyzed in Bishop and Svensén (2003): $n = 250$ values of input variables x_i generated uniformly in $(0, 1)$ and output variables y_i generated as $y_i = x_i + 0.3 \sin(2\pi x_i) + \epsilon_i$, with ϵ_i drawn from a zero mean Normal distribution with standard deviation 0.05.

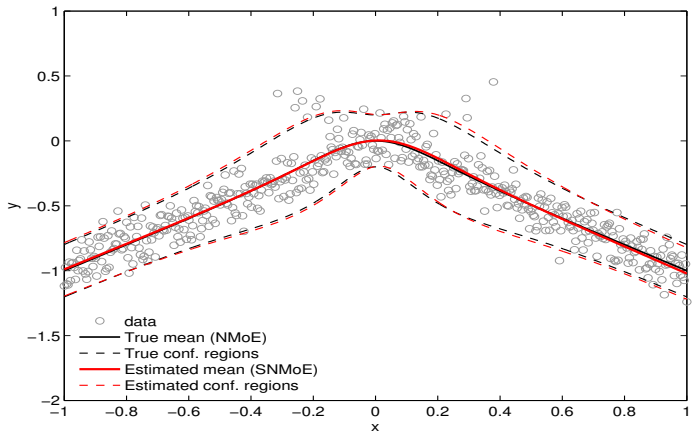
Experiments: Robustness of the NNMoE

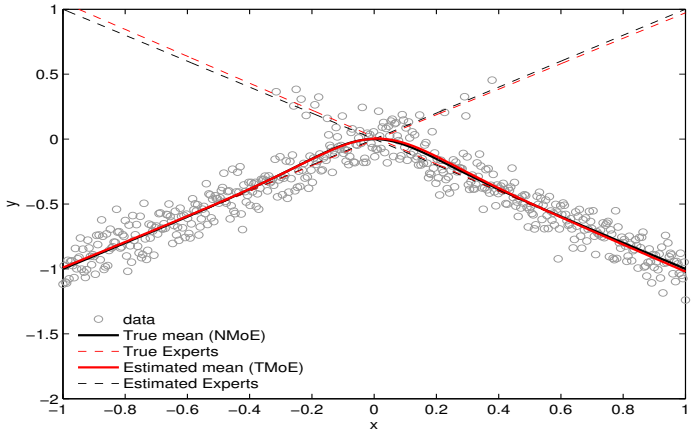
Experimental protocol as in Nguyen and McLachlan (2014)

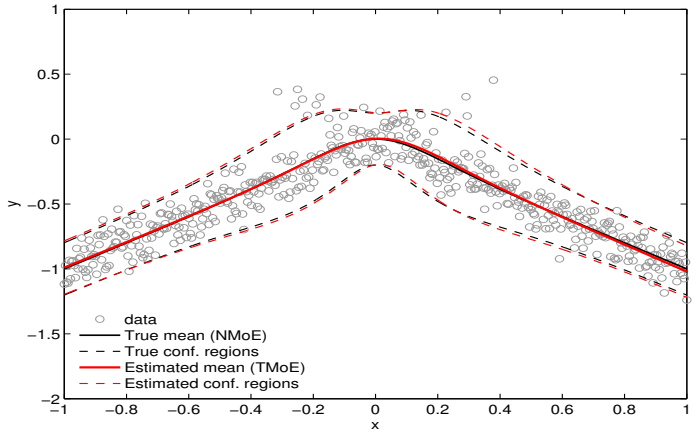
Model	Outliers	0%	1%	2%	3%	4%	5%
NMoE	NMoE	0.0001783	0.001057	0.001241	0.003631	0.013257	0.028966
	SNMoE	0.0001798	0.003479	0.004258	0.015288	0.022056	0.028967
	TMoE	0.0001685	0.000566	0.000464	0.000221	0.000263	0.000045
	STMoE	0.0002586	0.000741	0.000794	0.000696	0.000697	0.000626
SNMoE	NMoE	0.0000229	0.000403	0.004012	0.002793	0.018247	0.031673
	SNMoE	0.0000228	0.000371	0.004010	0.002599	0.018247	0.031674
	TMoE	0.0000325	0.000089	0.000130	0.000513	0.000108	0.000355
	STMoE	0.0000562	0.000144	0.000022	0.000268	0.000152	0.001041
TMoE	NMoE	0.0002579	0.0004660	0.002779	0.015692	0.005823	0.005419
	SNMoE	0.0002587	0.0004659	0.006743	0.015686	0.005835	0.004813
	TMoE	0.0002529	0.0002520	0.000144	0.000157	0.000488	0.000045
	STMoE	0.0002473	0.0002451	0.000173	0.000176	0.000214	0.000291
STMoE	NMoE	0.000710	0.0007238	0.001048	0.006066	0.012457	0.031644
	SNMoE	0.000713	0.0009550	0.001045	0.006064	0.012456	0.031644
	TMoE	0.000279	0.0003808	0.000371	0.000609	0.000651	0.000609
	STMoE	0.000280	0.0001865	0.000447	0.000600	0.000509	0.000602

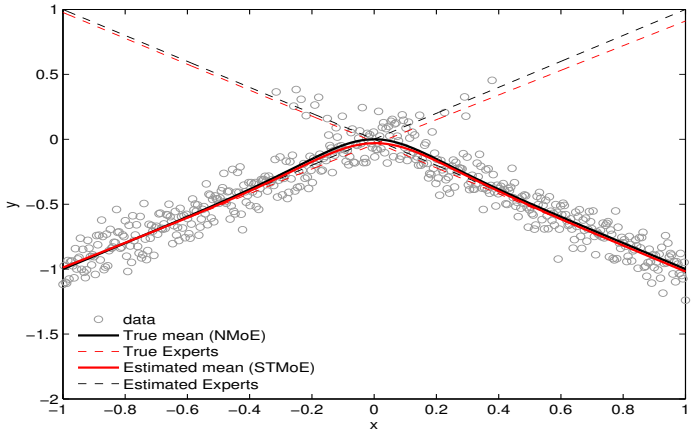
Table: MSE between the estimated mean function and the true one

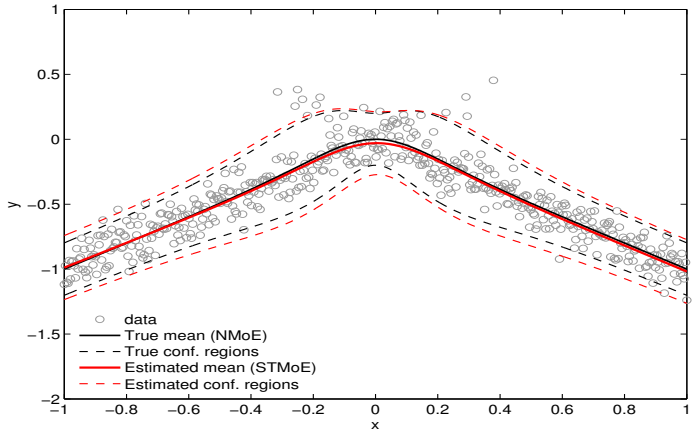

data
True mean (NMoE)
True Experts
Estimated mean (SNMoE)
Estimated Experts
x
y


data
True mean (NMoE)
True conf. regions
Estimated mean (SNMoE)
Estimated conf. regions
x
y


data
True mean (NMoE)
True Experts
Estimated mean (TMoE)
Estimated Experts
x
y


data
True mean (NMoE)
True conf. regions
Estimated mean (TMoE)
Estimated conf. regions
x
y


data
True mean (NMoE)
True Experts
Estimated mean (STMoE)
Estimated Experts
x
y


data
True mean (NMoE)
True conf. regions
Estimated mean (STMoE)
Estimated conf. regions
x
y

Robustness of the NNMoE

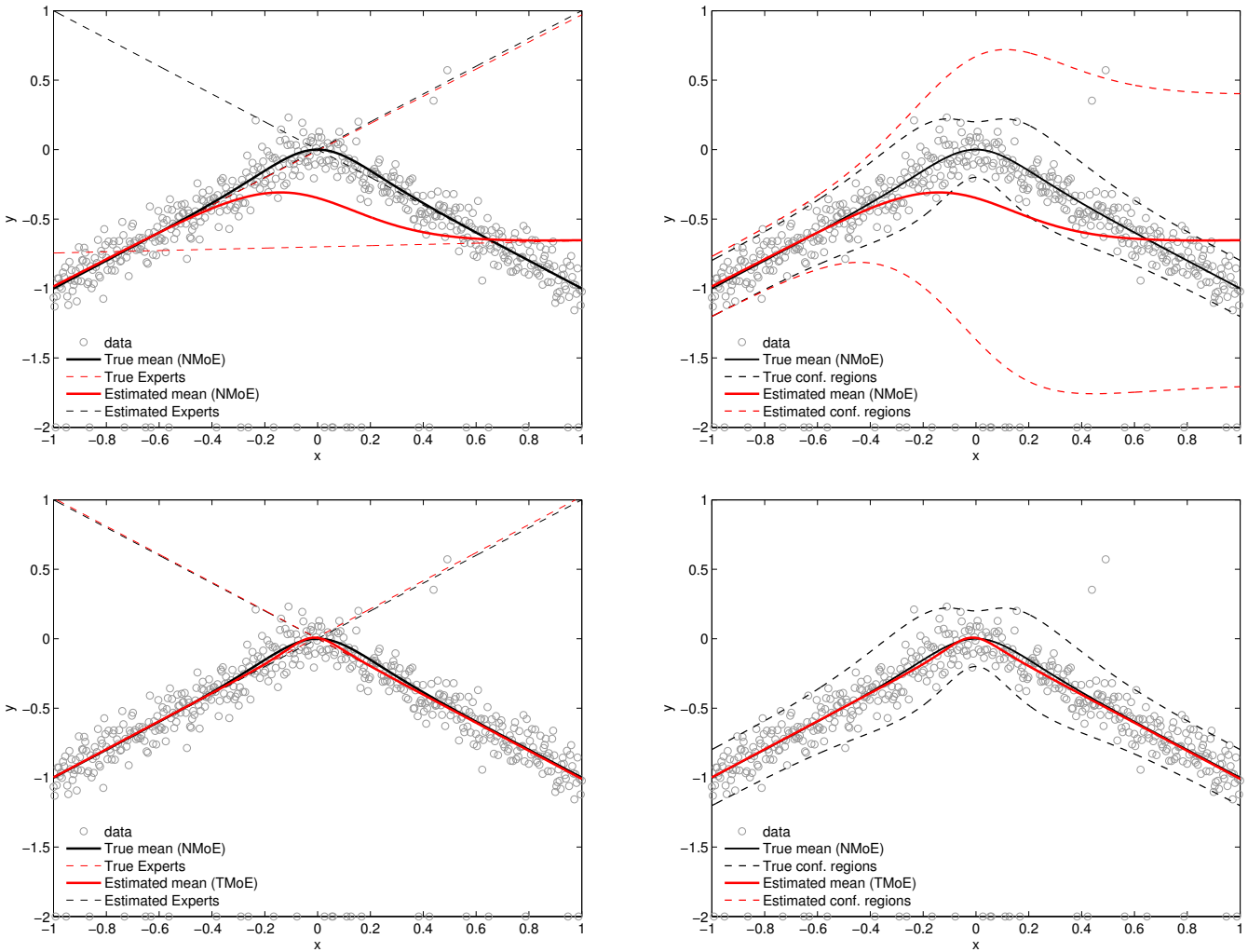


Figure: Fitted MoE to $n = 500$ observations generated according to the NMoE with 5% of outliers $(x; y = -2)$: NMoE fit (top), TMoE fit (bottom).

Robustness of the NNMoE

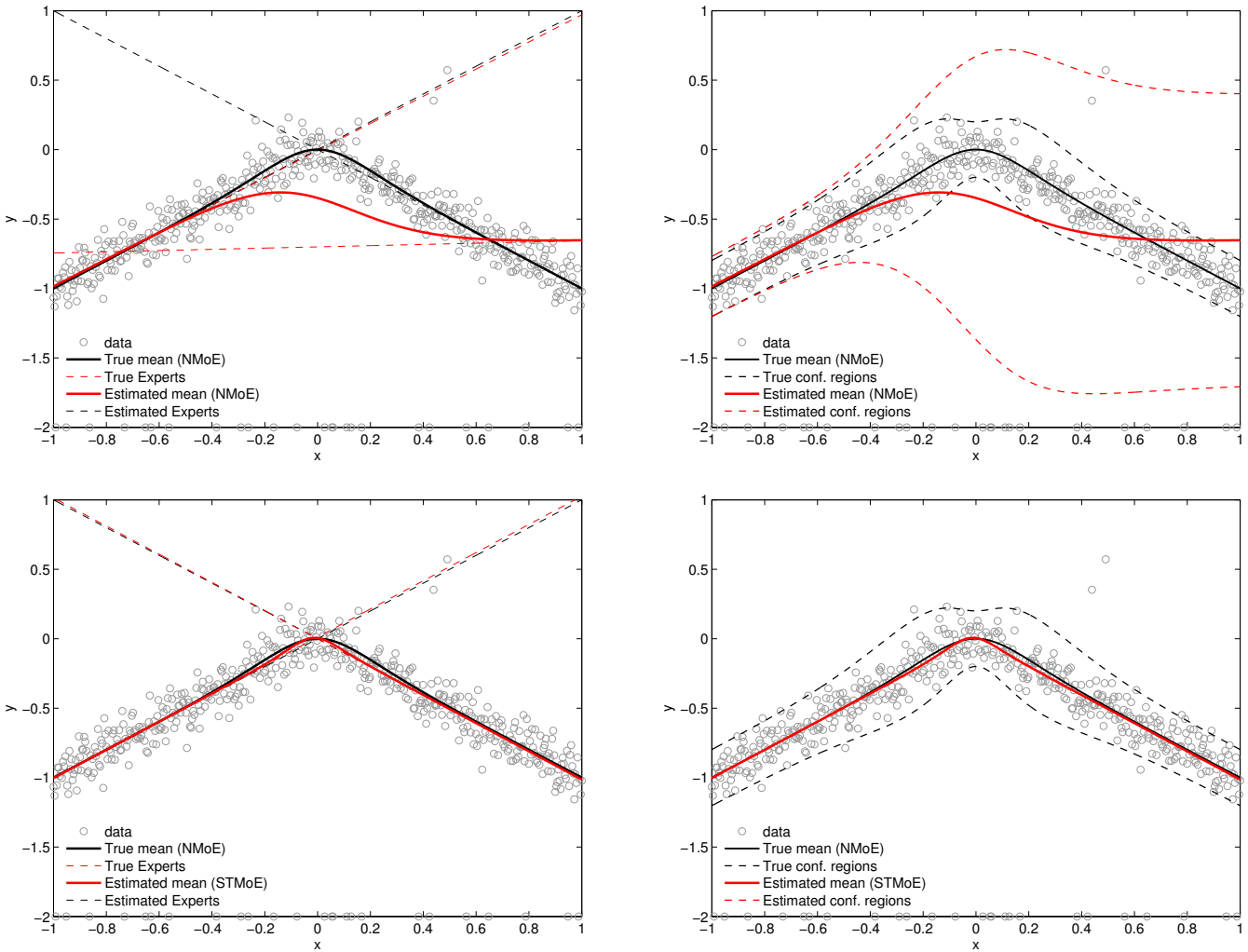


Figure: Fitted MoE to $n = 500$ observations generated according to the NMoE with 5% of outliers $(x; y = -2)$: NMoE fit (top), STMoE fit (bottom).

Experiments

Application to two real-world data sets

- **Tone perception data set** Recently studied by Bai et al. (2012) and Song et al. (2014) by using robust regression mixture models based on, respectively, the t distribution and the Laplace distribution.
- To apply the proposed MoE models, we set the response $y_i (i = 1, \ldots, 150)$ as the "strech ratio" variables and the covariates $\boldsymbol{x}_i = \boldsymbol{r}_i = (1, x_i)^T$ where x_i is the "tuned" variable of the ith observation.
- **Temperature Anomaly Data**
- The data consist of $n = 135$ yearly measurements of the global annual temperature anomalies (in degrees C) computed using data from land meteorological stations for the period of $1882 - 2012$.
- The response $y_i (i = 1, \ldots, 135)$ is set as the temperature anomalies and the covariates $\boldsymbol{x}_i = \boldsymbol{r}_i = (1, x_i)^T$ where x_i is the year of the ith observation.
- These data have been analyzed earlier by Hansen et al. (1999, 2001) and recently by Nguyen and McLachlan (2014) by using the Laplace mixture of linear experts (LMoLE)

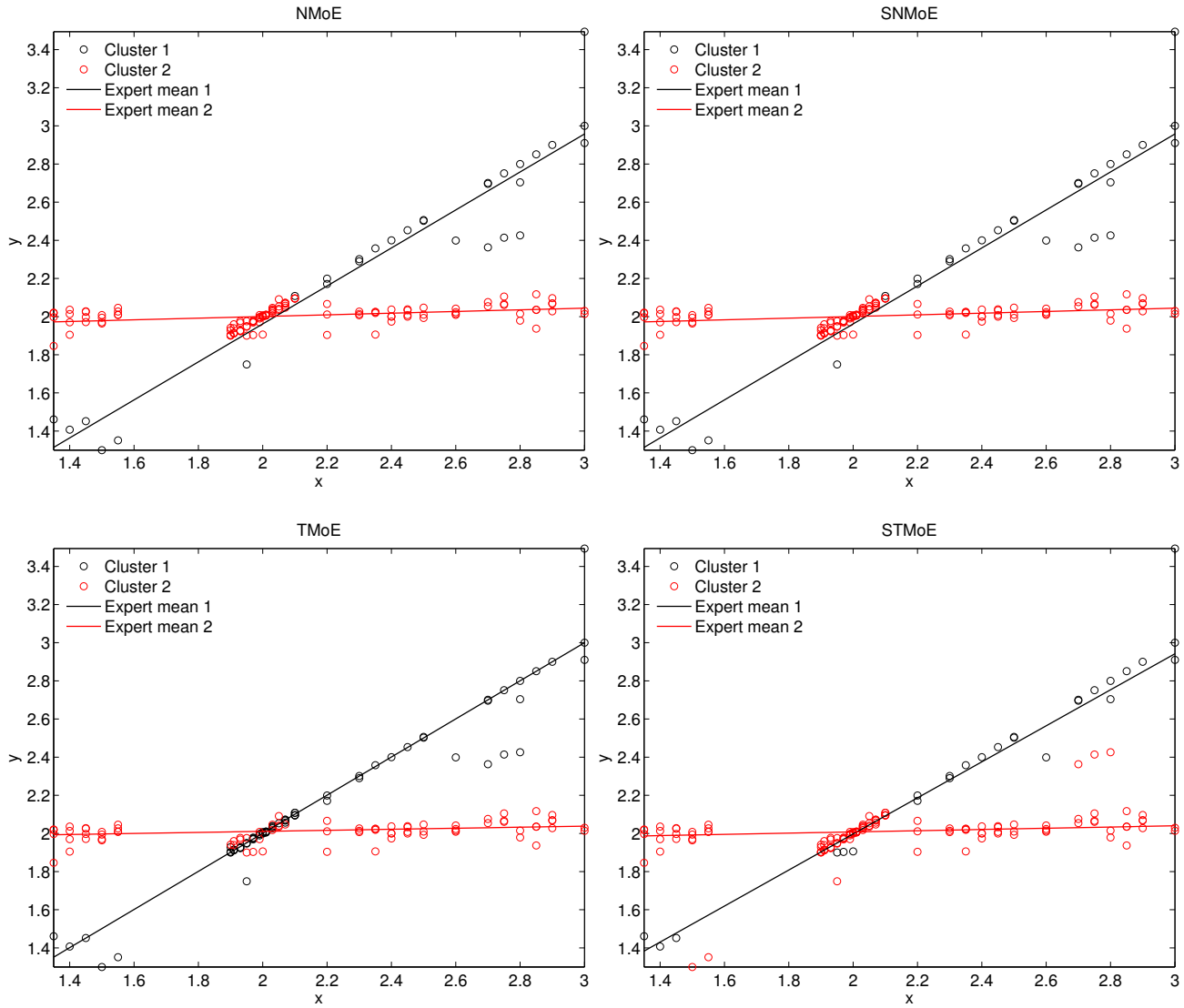


Figure: Fitting the MoLE to the tone data set studied by Bai et al. (2012) and Song et al. (2014) by using robust regression mixture models based on, respectively, the t distribution and the Laplace distribution: $n = 150$ pairs of "tuned" predictors (x), and their corresponding "strech ratio" responses (y).

Model selection

	NMoE			SNMoE			TMoE			STMoE		
K	BIC	AIC	ICL	BIC	AIC	ICL	BIC	AIC	ICL	BIC	AIC	ICL
1	1.8662	6.3821	1.8662	-0.6391	5.3821	-0.6391	71.3931	77.4143	71.3931	69.5326	77.0592	69.5326
2	122.8050	134.8476	107.3840	117.7939	132.8471	102.4049	204.8241	219.8773	186.8415	92.4352	110.4990	82.4552
3	118.1939	137.7630	76.5249	122.8725	146.9576	98.0442	199.4030	223.4880	183.0389	77.9753	106.5764	52.5642
4	121.7031	148.7989	94.4606	109.5917	142.7087	97.6108	201.8046	234.9216	187.7673	77.7092	116.8474	56.3654
5	141.6961	176.3184	123.6550	107.2795	149.4284	96.6832	187.8652	230.0141	164.9629	79.0439	128.7194	67.7485

Table: Choosing the number of experts K for the original tone perception data.

Model's Robustness

- I also examined the sensitivity of the MoE models to outliers based on this real data set.
- the same scenario used in Bai et al. (2012) and Song et al. (2014) (the last and more difficult scenario) by adding 10 identical pairs $(0, 4)$ to the original data set as outliers in the y-direction, considered as high leverage outliers.

Robustness to outliers

- $\Rightarrow$ the normal and the skew-normal mixture of experts provide almost identical fits and are sensitive to outliers.
- However, in both cases, compared to the normal regression mixture result in Bai et al. (2012), and the Laplace regression mixture and the t regression mixture results in Song et al. (2014), the fitted NMoE and SNMoE model are affected less severely by the outliers.
- This may be attributed to the fact that the mixing proportions here are depending on the predictors, which is not the case in these regression mixture models, namely the ones of Bai et al. (2012), and Song et al. (2014).
- The TMoE and the STMoE provide robust fits, which are quasi-identical to the fit obtained on the original data without outliers.
- Moreover, I notice that, as showed in Song et al. (2014), for this situation with outliers, the t mixture of regressions fails; The fit is affected severely by the outliers.

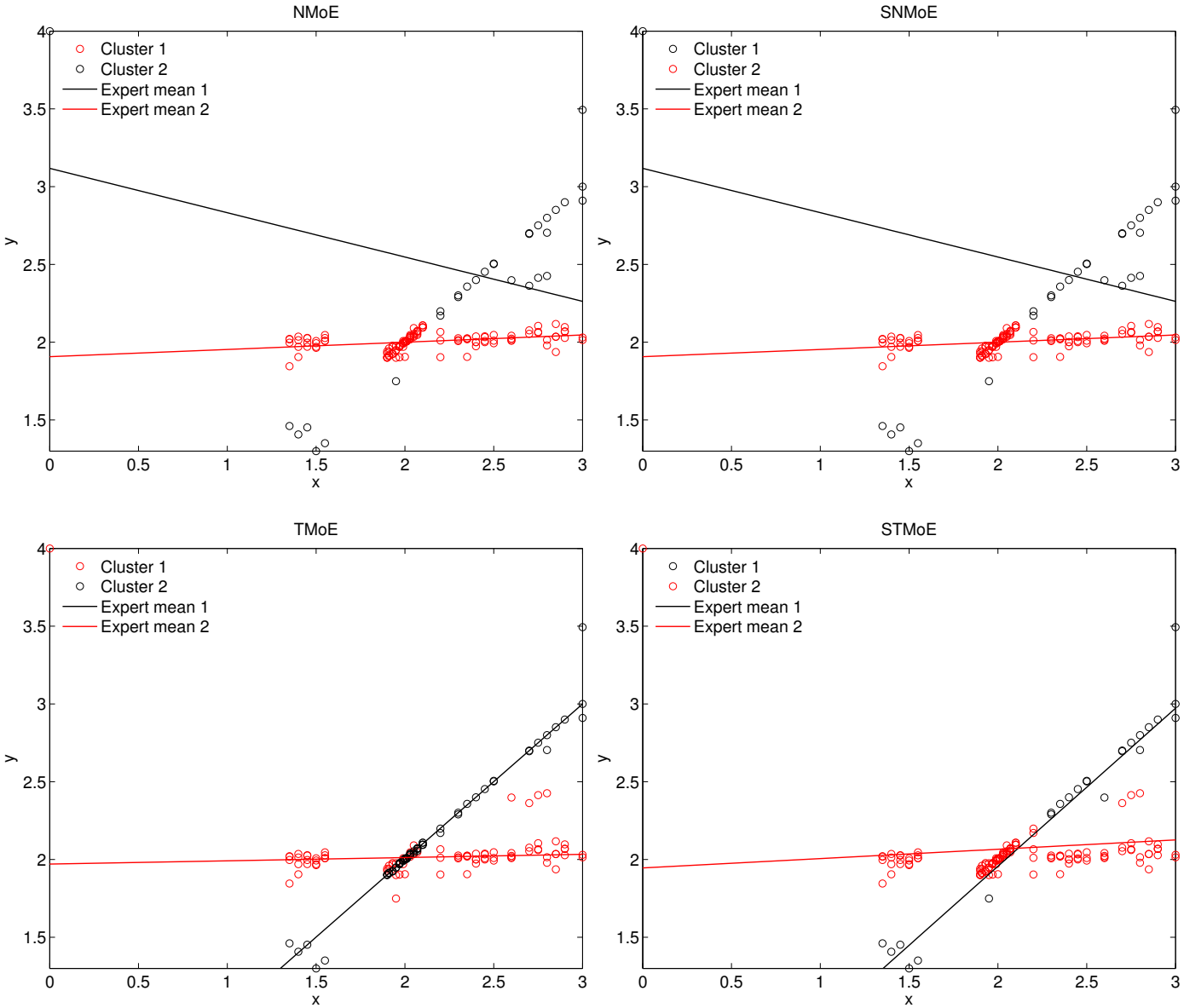


Figure: Fitting MoLE to the tone data set with ten added outliers $(0, 4)$.

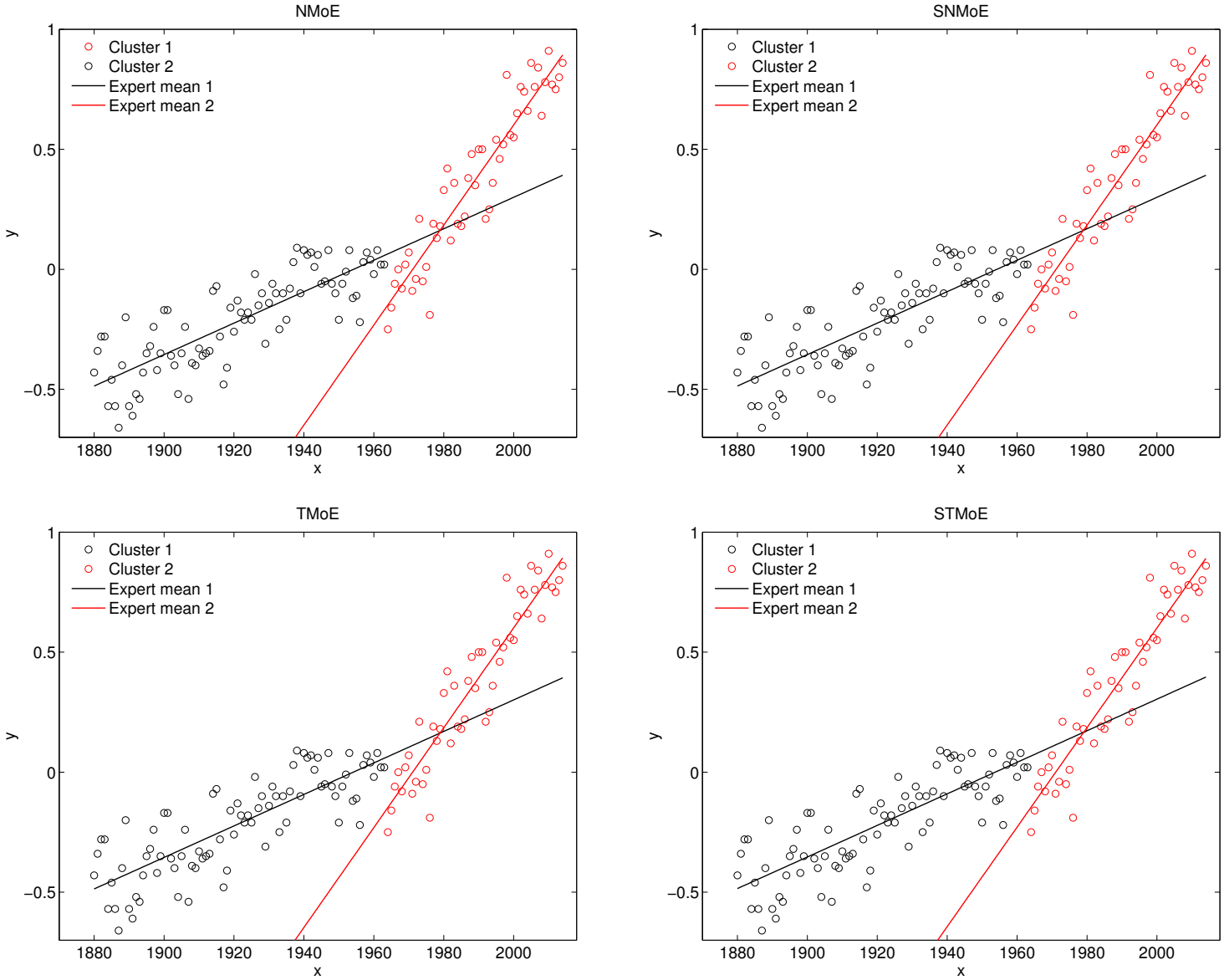


Figure: Fitting the MoLE models to the temperature anomalies data set.

- Both the TMoE and STMoE fits provide a degrees of freedom more than 17, which tends to approach a normal distribution.
- On the other hand, the regression coefficients are also similar to those found by Nguyen and McLachlan (2014) who used a Laplace mixture of linear experts.
- Model selection : Except the result provided by AIC for the NMoE model which provides overestimates the number of components, all the others results provide evidence for two components in the data.

	NMoE			SNMoE			TMoE			STMoE		
K	BIC	AIC	ICL	BIC	AIC	ICL	BIC	AIC	ICL	BIC	AIC	ICL
1	46.0623	50.4202	46.0623	43.6096	49.4202	43.6096	43.5521	49.3627	43.5521	40.9715	48.2347	40.9715
2	<u>79.9163</u>	91.5374	<u>79.6241</u>	<u>75.0116</u>	<u>89.5380</u>	<u>74.7395</u>	<u>74.7960</u>	<u>89.3224</u>	<u>74.5279</u>	<u>69.6382</u>	<u>87.0698</u>	<u>69.3416</u>
3	71.3963	90.2806	58.4874	63.9254	87.1676	50.8704	63.9709	87.2131	47.3643	54.1267	81.7268	30.6556
4	66.7276	92.8751	54.7524	55.4731	87.4312	41.1699	56.8410	88.7990	45.1251	42.3087	80.0773	20.4948
5	59.5100	<u>92.9206</u>	51.2429	45.3469	86.0207	41.0906	43.7767	84.4505	29.3881	28.0371	75.9742	-8.8817

Table: Choosing the number of expert components K for the temperature anomalies data by using the information criteria BIC, AIC, and ICL. Underlined value indicates the highest value for each criterion.

Summary

- Proposed non-normal MoE models, which generalize the normal MoE. They are based on the skew-normal, t and skew t distribution and are respectively the SNMoE, TMoE, and STMoE.
- The SNMoE model is suggested for non-symmetric data, the TMoE for data with possibly outliers and heavy tail, and the STMoE is suggested for both possibly non-symmetric, heavy tailed and noisy data.
- Here I only considered the MoE in their standard (non-hierarchical) version: One interesting future direction is therefore to extend the proposed models to the hierarchical mixture of experts framework (Jordan and Jacobs, 1994).
- Furthermore, a natural future extension of this work is to consider the case of MoE for multiple regression on multivariate data rather than simple regression on univariate data.

Outline

Conclusion

- The previous chapters presented my research during the last five years as well as my ongoing research on the problems of statistical learning of flexible models for complex data analysis.
- This involved research in statistics in the related fields of classification, high dimensional and functional data analysis, statistical signal processing, machine learning and pattern recognition, and in the field of statistical inference.
- The focus in the latter field has been on the methodology and applications of latent data models, particularly mixture models, and on maximum likelihood estimation via EM algorithms as well as maximum a posteriori estimation via Bayesian sampling, including in the Bayesian non-parametric paradigm.
- A particular attention was given to the statistical methodology and its computational aspects, which constitute a common theme of my research.

Other work and Ongoing Research

Other finished work

- Latent data models for functional data analysis [J-2][J-4][J-5]
- Bayesian non-parametric parsimonious clustering of multivariate data: PhD thesis of Marius Bartcus (2012-2015) [a]
- Bayesian regularization of spatial splines regressions [J-11] (2014 -)

[a] M. Bartcus. *Bayesian non-parametric parsimonious mixtures for model-based clustering*. Ph.D. thesis, Université de Toulon, Laboratoire des Sciences de l'Information et des Systèmes (LSIS), October 2015

Ongoing Research

- Advanced mixtures for complex data (My ongoing CNRS research leave project)
- LEarning from biG cOmplex FunctIonal daTa - LegoFit (2015 - an ANR proposal, PI with LIPN, IFSTTAR, LIPADE and AIRBUS)

  Model-based (co)-clustering for high-dimensional (functional) data

Perspectives

Perspectives

- Non-normal mixture modeling
- Feature selection in model-based clustering
- Bayesian latent variable models for sparse representations
- Unsupervised learning of feature hierarchies: Deep learning
  Patel et al. (2015), introduced a probabilistic theory of deep learning that seems to answer some key questions for deep learning from a probabilistic point of view.

Personal bibliography

Publications

[J-1] 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

[J-2] 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

[J-3] F. Chamroukhi, A. Samé, G. Govaert, and P. Aknin. Modèle à processus latent et algorithme EM pour la régression non linéaire. *Revue des Nouvelles Technologies de l'Information (RNTI)*, S1:15–32, Jan 2011

[J-4] A. Samé, F. Chamroukhi, Gérard Govaert, and P. Aknin. Model-based clustering and segmentation of time series with changes in regime. *Advances in Data Analysis and Classification*, 5:301–321, 2011

[J-5] F. Chamroukhi, H. Glotin, and A. Samé. Model-based functional mixture discriminant analysis with hidden process regression for curve classification. *Neurocomputing*, 112:153–163, 2013a

[J-6] F. Chamroukhi, D. Trabelsi, S. Mohammed, L. Oukhellou, and Y. Amirat. Joint segmentation of multivariate time series with hidden process regression for human activity recognition. *Neurocomputing*, 120:633–644, November 2013b

[J-7] D. Trabelsi, S. Mohammed, F. Chamroukhi, L. Oukhellou, and Y. Amirat. An unsupervised approach for automatic activity recognition based on hidden markov model regression. *IEEE Transactions on Automation Science and Engineering*, 3(10):829–335, 2013

[J-8] F. Chamroukhi. Unsupervised learning of regression mixture models with unknown number of components. *Journal of Statistical Computation and Simulation*, 2015e. doi: 10.1080/00949655.2015.1109096. In Press

[J-9] F. Chamroukhi. Piecewise regression mixture for simultaneous curve clustering and optimal segmentation. *Journal of Classification - Springer*, 2015d. Accepted

Submitted papers

[J-10] F. Chamroukhi, M. Bartcus, and H. Glotin. Dirichlet process parsimonious gaussian mixture for clustering. *arXiv:1501.03347*, January 2015. Submitted

[J-11] F. Chamroukhi. Bayesian mixtures of spatial spline regressions. *arXiv:1508.00635*, Aug 2015a

[J-12] F. Chamroukhi. Non-normal mixtures of experts. *arXiv:1506.06707*, July 2015c. 61 pages

[J-13] F. Chamroukhi. Robust mixture of experts modeling using the t-distribution. 2015g. submitted

[J-14] F. Chamroukhi. Robust mixture of experts modeling using the skew-t distribution. 2015f. submitted

[J-15] F. Attal, M. Dedabrishvili, S. Mohammed, F. Chamroukhi, L. Oukhellou, and Y. Amirat. Physical human activity recognition using wearable sensors. *Sensors*, 2015. submitted

Personal bibliography

Papers in preparation (2)

[J-16] F. Chamroukhi. Mixture of hidden markov model regressions for functional data clustering and segmentation. *Neural Networks*, 2015b. In preparation

[J-17] F. Chamroukhi et al. Bayesian non-parametric models for unsupervised decomposition of whale songs. *Journal of Acoustical Society of America*, 2015. In preparation

Monograph and editorials

[M-1] F. Chamroukhi. *Probabilistic Learning From Longitudinal Data: Background, Novel theoretical models, Classifiers and Algorithms*. Lap Lambert Academic Publishing, 2011. ISBN 978-3844311372

[M-2] H. Glotin, F. Chamroukhi, and T. Maillot. *Representations and Decisions in Cognitive Vision*. Proceedings of the International Summer School ERMITES 2012, 2012. ISBN 979-10-90821-00-2

[M-3] F. Chamroukhi and H. Glotin. *Unsupervised learning from big bioacoustic data (uLearnBio)*. Proceedings of the uLearnBio workshop of the International Conference on Machine Learning (ICML), 2014. ISBN 979-10-90821-06-4

Invited talks in international conferences

[C-1] F. Chamroukhi. Robust non-normal mixtures of experts. ERCIM 2015 : The 8th International Conference of the European Research Consortium for Informatics and Mathematics on Computational and Methodological Statistics, December 2015h. London, UK

[C-2] F. Chamroukhi. Model-based cluster and discriminant analysis for functional data. ERCIM 2014 : The 7th International Conference of the European Research Consortium for Informatics and Mathematics on Computational and Methodological Statistics, December 2014a. Pisa, Italy

[C-3] F. Chamroukhi. Mixture models for cluster analysis: from model-based inference to bayesian non-parametrics. uLearnBio workshop of the International Conference on Machine Learning (ICML), June 2014b

[C-4] F. Chamroukhi. Learning probabilistic latent process models from temporal data. VIIth International Summer School ERMITES 2012 on Representations and Decisions in Cognitive Vision, september 2012

Thank you!

References I

F. Attal, M. Dedabrishvili, S. Mohammed, F. Chamroukhi, L. Oukhellou, and Y. Amirat. Physical human activity recognition using wearable sensors. *Sensors*, 2015. submitted.

A. Azzalini. A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, pages 171–178, 1985.

A. Azzalini. Further results on a class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, pages 199–208, 1986.

A. Azzalini and A. Capitanio. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution. *Journal of the Royal Statistical Society, Series B*, 65:367–389, 2003.

Xiuqin Bai, Weixin Yao, and John E. Boyer. Robust fitting of mixture regression models. *Computational Statistics & Data Analysis*, 56(7):2347 – 2359, 2012.

M. Bartcus. *Bayesian non-parametric parsimonious mixtures for model-based clustering*. Ph.D. thesis, Université de Toulon, Laboratoire des Sciences de l'Information et des Systèmes (LSIS), October 2015.

C. Bishop and M. Svensén. Bayesian hierarchical mixtures of experts. In *In Uncertainty in Artificial Intelligence*, 2003.

Richard P. Brent. *Algorithms for minimization without derivatives*. Prentice-Hall series in automatic computation. Englewood Cliffs, N.J. Prentice-Hall, 1973. ISBN 0-13-022335-2.

F. Chamroukhi. *Hidden process regression for curve modeling, classification and tracking*. Ph.D. thesis, Université de Technologie de Compiègne, Compiègne, France, 2010.

F. Chamroukhi. *Probabilistic Learning From Longitudinal Data: Background, Novel theoretical models, Classifiers and Algorithms*. Lap Lambert Academic Publishing, 2011. ISBN 978-3844311372.

F. Chamroukhi. Learning probabilistic latent process models from temporal data. VIIth International Summer School ERMITES 2012 on Representations and Decisions in Cognitive Vision, september 2012.

F. Chamroukhi. Model-based cluster and discriminant analysis for functional data. ERCIM 2014 : The 7th International Conference of the European Research Consortium for Informatics and Mathematics on Computational and Methodological Statistics, December 2014a. Pisa, Italy.

References II

F. Chamroukhi. Mixture models for cluster analysis: from model-based inference to bayesian non-parametrics. uLearnBio workshop of the International Conference on Machine Learning (ICML), June 2014b.

F. Chamroukhi. Bayesian mixtures of spatial spline regressions. *arXiv:1508.00635*, Aug 2015a.

F. Chamroukhi. Mixture of hidden markov model regressions for functional data clustering and segmentation. *Neural Networks*, 2015b. In preparation.

F. Chamroukhi. Non-normal mixtures of experts. *arXiv:1506.06707*, July 2015c. 61 pages.

F. Chamroukhi. Piecewise regression mixture for simultaneous curve clustering and optimal segmentation. *Journal of Classification - Springer*, 2015d. Accepted.

F. Chamroukhi. Unsupervised learning of regression mixture models with unknown number of components. *Journal of Statistical Computation and Simulation*, 2015e. doi: $10.1080/00949655.2015.1109096$. In Press.

F. Chamroukhi. Robust mixture of experts modeling using the skew-t distribution. 2015f. submitted.

F. Chamroukhi. Robust mixture of experts modeling using the t-distribution. 2015g. submitted.

F. Chamroukhi. Robust non-normal mixtures of experts. ERCIM 2015 : The 8th International Conference of the European Research Consortium for Informatics and Mathematics on Computational and Methodological Statistics, December 2015h. London, UK.

F. Chamroukhi and H. Glotin. *Unsupervised learning from big bioacoustic data (uLearnBio)*. Proceedings of the uLearnBio workshop of the International Conference on Machine Learning (ICML), 2014. ISBN 979-10-90821-06-4.

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.

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.

F. Chamroukhi, A. Samé, G. Govaert, and P. Aknin. Modèle à processus latent et algorithme EM pour la régression non linéaire. *Revue des Nouvelles Technologies de l'Information (RNTI)*, S1:15–32, Jan 2011.

References III

F. Chamroukhi, H. Glotin, and A. Samé. Model-based functional mixture discriminant analysis with hidden process regression for curve classification. *Neurocomputing*, 112:153–163, 2013a.

F. Chamroukhi, D. Trabelsi, S. Mohammed, L. Oukhellou, and Y. Amirat. Joint segmentation of multivariate time series with hidden process regression for human activity recognition. *Neurocomputing*, 120:633–644, November 2013b.

F. Chamroukhi, M. Bartcus, and H. Glotin. Dirichlet process parsimonious gaussian mixture for clustering. *arXiv:1501.03347*, January 2015. Submitted.

K. Chen, L. Xu, and H. Chi. Improved learning algorithms for mixture of experts in multiclass classification. *Neural Networks*, 12 (9):1229–1252, 1999.

Elizabeth A. Cohen. Some effects of inharmonic partials on interval perception. *Music Perception*, 1, 1984.

Sophie Dabo-Niang, Frédéric Ferraty, and Philippe Vieu. On the using of modal curves for radar waveforms classification. *Computational Statistics & Data Analysis*, 51(10):4878 – 4890, 2007.

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. Chamroukhi et al. Bayesian non-parametric models for unsupervised decomposition of whale songs. *Journal of Acoustical Society of America*, 2015. In preparation.

S. Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models (Springer Series in Statistics)*. Springer Verlag, New York, 2006.

S. J. Gaffney. *Probabilistic Curve-Aligned Clustering and Prediction with Regression Mixture Models*. PhD thesis, Department of Computer Science, University of California, Irvine, 2004.

H. Glotin, F. Chamroukhi, and T. Maillot. *Representations and Decisions in Cognitive Vision*. Proceedings of the International Summer School ERMITES 2012, 2012. ISBN 979-10-90821-00-2.

P. Green. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of The Royal Statistical Society, B*, 46(2):149–192, 1984.

References IV

J. Hansen, R. Ruedy, J. Glascoe, and M. Sato. Giss analysis of surface temperature change. *Journal of Geophysical Research*, 104:30997–31022, 1999.

J. Hansen, R. Ruedy, Sato M., M. Imhoff, W. Lawrence, D. Easterling, T. Peterson, and T. Karl. A closer look at united states and global surface temperature change. *Journal of Geophysical Research*, 106:23947–23963, 2001.

Salvatore Ingrassia, Simona Minotti, and Giorgio Vittadini. Local statistical modeling via a cluster-weighted approach with elliptical distributions. *Journal of Classification*, 29(3):363–401, 2012.

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.

Wenxin Jiang and Martin A. Tanner. On the identifiability of mixtures-of-experts. *Neural Networks*, 12:197–220, 1999.

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

Sharon X. Lee and Geoffrey J. McLachlan. Finite mixtures of multivariate skew t-distributions: some recent and new results. *Statistics and Computing*, 24(2):181–202, 2014.

Sharon X. Lee and Geoffrey J. McLachlan. Finite mixtures of canonical fundamental skew *t*-distributions. *Statistics and Computing (To appear)*, 2015. doi: 10.1007/s11222-015-9545-x.

Tsung I. Lin, Jack C. Lee, and Wan J. Hsieh. Robust mixture modeling using the skew *t* distribution. *Statistics and Computing*, 17(2):81–92, 2007a.

Tsung I. Lin, Jack C. Lee, and Shu Y Yen. Finite mixture modelling using the skew normal distribution. *Statistica Sinica*, 17: 909–927, 2007b.

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

G. J. McLachlan and D. Peel. *Finite mixture models*. New York: Wiley, 2000.

Geoffrey J. Mclachlan and David Peel. Robust cluster analysis via mixtures of multivariate t-distributions. In *Lecture Notes in Computer Science*, pages 658–666. Springer-Verlag, 1998.

References V

X. L. Meng and D. B. Rubin. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2): 267–278, 1993.

Hien D. Nguyen and Geoffrey J. McLachlan. Laplace mixture of linear experts. *Computational Statistics & Data Analysis*, pages –, 2014. ISSN 0167-9473. doi: http://dx.doi.org/10.1016/j.csda.2014.10.016.

Ankit B. Patel, Tan Nguyen, and Richard G. Baraniuk. A probabilistic theory of deep learning. Technical Report Technical Report No 2015-1, Rice University Electrical and Computer Engineering Dept., April 2015. URL `http://arxiv.org/abs/1504.00641v1`.

D. Peel and G. J. Mclachlan. Robust mixture modelling using the t distribution. *Statistics and Computing*, 10:339–348, 2000.

L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77 (2):257–286, 1989.

L. R. Rabiner and B. H. Juang. An introduction to hidden markov models. *IEEE ASSP Magazine*, 1986.

R. Ruedy, M. Sato, and K. Lo. NASA GISS surface temperature (GISTEMP) analysis. DOI: 10.3334/CDIAC/cli.001. Center for Climate Systems Research, NASA Goddard Institute for Space Studies 2880 Broadway, New York, NY 10025 USA.

A. Samé, F. Chamroukhi, Gérard Govaert, and P. Aknin. Model-based clustering and segmentation of time series with changes in regime. *Advances in Data Analysis and Classification*, 5:301–321, 2011.

Weixing Song, Weixin Yao, and Yanru Xing. Robust mixture regression model fitting by laplace distribution. *Computational Statistics & Data Analysis*, 71(0):128 – 137, 2014.

D. Titterington, A. Smith, and U. Makov. *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons, 1985.

D. Trabelsi. *Contribution à la reconnaissance non-intrusive d'activités humaines.* Ph.D. thesis, Université Paris-Est Créteil, Laboratoire Images, Signaux et Systèmes Intelligents (LiSSi), June 2013.

D. Trabelsi, S. Mohammed, F. Chamroukhi, L. Oukhellou, and Y. Amirat. An unsupervised approach for automatic activity recognition based on hidden markov model regression. *IEEE Transactions on Automation Science and Engineering*, 3(10): 829–335, 2013.

Y. Wei. Robust mixture regression models using t-distribution. Technical report, Master Report, Department of Statistics, Kansas State University, 2012.

Ka Yee Yeung, Chris Fraley, A. Murua, Adrian E. Raftery, and Walter L. Ruzzo. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17(10):977–987, 2001.