Neural Networks 79 (2016) 20–36

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Robust mixture of experts modeling using the t distribution

F. Chamroukhi *

Université de Toulon, CNRS, LSIS, UMR 7296, 83957 La Garde, France
Aix Marseille Université, CNRS, ENSAM, LSIS, UMR 7296, 13397 Marseille, France
Laboratoire Paul Painlevé (LPP), UMR CNRS 8524, Université Lille 1, 59650 Villeneuve d'Ascq, France

ARTICLE INFO

Article history:
Received 17 September 2015
Received in revised form 23 January 2016
Accepted 11 March 2016
Available online 31 March 2016

Keywords:
Mixture of experts
t distribution
EM algorithm
Robust modeling
Non-linear regression
Model-based clustering

ABSTRACT

Mixture of Experts (MoE) is a popular framework for modeling heterogeneity in data for regression, classification, and clustering. For regression and cluster analyses of continuous data, MoE usually uses normal experts following the Gaussian distribution. However, for a set of data containing a group or groups of observations with heavy tails or atypical observations, the use of normal experts is unsuitable and can unduly affect the fit of the MoE model. We introduce a robust MoE modeling using the t distribution. The proposed t MoE (TMoE) deals with these issues regarding heavy-tailed and noisy data. We develop a dedicated expectation–maximization (EM) algorithm to estimate the parameters of the proposed model by monotonically maximizing the observed data log-likelihood. We describe how the presented model can be used in prediction and in model-based clustering of regression data. The proposed model is validated on numerical experiments carried out on simulated data, which show the effectiveness and the robustness of the proposed model in terms of modeling non-linear regression functions as well as in model-based clustering. Then, it is applied to the real-world data of tone perception for musical data analysis, and the one of temperature anomalies for the analysis of climate change data. The obtained results show the usefulness of the TMoE model for practical applications.

1. Introduction

Mixture of experts (MoE) introduced by Jacobs, Jordan, Nowlan, and Hinton (1991) is widely studied in statistics and machine learning. They consist in a fully conditional mixture model where both the mixing proportions, known as the gating functions, and the component densities, known as the experts, are conditional on some input covariates. MoE has been investigated, in their simple form, as well as in their hierarchical form (Jordan & Jacobs, 1994) (e.g. Section 5.12 of McLachlan & Peel, 2000) for regression and model-based cluster and discriminant analyses and in different application domains. A complete review of the MoE models can be found in Yuksel, Wilson, and Gader (2012). For continuous data, which we consider here in the context of non-linear regression and model-based cluster analysis, MoE usually uses normal experts, that is, expert components following the Gaussian distribution. Along this paper, we will call it the normal mixture of experts, abbreviated NMoE. It is well-known that the normal distribution is sensitive to outliers, which makes NMoE unsuitable to noisy data. Moreover, for a set of data containing a group or groups of observations with heavy tails, the use of normal experts may be unsuitable and can unduly affect the fit of the MoE model. In this paper, we attempt to overcome these limitations in MoE by proposing a more adapted and robust MoE model which can deal with the issues of heavy-tailed and atypical data.

The problem of sensitivity of NMoE to outliers has been considered very recently by Nguyen and McLachlan (2016) where the authors proposed a Laplace mixture of linear experts (LMoLE) for a robust modeling of non-linear regression data. The model parameters are estimated by maximizing the observed-data likelihood via a minorization–maximization (MM) algorithm. Here, we propose an alternative MoE model, by relying on the t distribution. We call this proposed model the t mixture of experts, abbreviated TMoE. The t distribution provides indeed a natural robust extension of the normal distribution to model data with possible outliers and tails more heavy compared to the normal distribution. It has been considered to develop the t mixture model proposed by Mclachlan and Peel (1998) for robust cluster analysis of multivariate data. We also mention that Lin, Lee, and Hsieh (2007) also proposed a mixture of skew t distributions to deal with heavy-tailed and asymmetric distributions. However, in the skew-t mixture model of Lin et al. (2007), the mixing

* Correspondence to: Université de Toulon, LSIS, UMR CNRS 7296, Bâtiment R, BP 20132 - 83957 La Garde Cedex, France. Tel.: +33 0 4 94 14 20 06; fax: +33 0 4 94 14 28 97.
E-mail address: faicel.chamroukhi@univ-tln.fr.

http://dx.doi.org/10.1016/j.neunet.2016.03.002
0893-6080/

proportions and the components means are constant, that is, they are not predictor-depending. In the proposed TMoE, however, we consider t expert components in which both the mixing proportions and the mixture component means are predictor-depending. More specifically, we use polynomial regressors for the components, as well as multinomial logistic regressors for the mixing proportions. In the framework of regression analysis, recently, Bai, Yao, and Boyer (2012), Ingrassia, Minotti, and Vittadini (2012) proposed a robust mixture modeling of regression on univariate data, by using a univariate t-mixture model. For the general multivariate case using t mixtures, one can refer to for example the two key papers Mclachlan and Peel (1998) and Peel and Mclachlan (2000). The inference in the previously described approaches is performed by maximum likelihood estimation via expectation–maximization (EM) or extensions (Dempster, Laird, & Rubin, 1977; McLachlan & Krishnan, 2008), in particular the expectation conditional maximization (ECM) algorithm (Meng & Rubin, 1993). For the Bayesian framework, Frühwirth-Schnatter and Pyne (2010) have considered the Bayesian inference for both the univariate and the multivariate skew-normal and skew-t mixtures. For the regression context, the robust modeling of regression data has been studied namely by Ingrassia et al. (2012), Wei (2012) who considered a t-mixture model for regression analysis of univariate data, as well as by Bai et al. (2012) who relied on the M-estimate in mixture of linear regressions. In the same context of regression, Song, Yao, and Xing (2014) proposed the mixture of Laplace regressions, which has been then extended by Nguyen and McLachlan (2016) to the case of mixture of experts, by introducing the Laplace mixture of linear experts (LMoLE). However, unlike our proposed TMoE model, the regression mixture models of Bai et al. (2012), Ingrassia et al. (2012), Song et al. (2014) and Wei (2012) do not consider conditional mixing proportions, that is, mixing proportions depending on some input variables, as in the case of mixture of experts, which we investigate here.

Here we consider the MoE framework for non-linear regression problems and model-based clustering of regression data, and we attempt to overcome the limitations of the NMoE model for dealing with heavy-tailed data and which may contain outliers. We investigate the use of the t distribution for the experts, rather than the commonly used normal distribution. The t-mixture of experts model (TMoE) handles the issues regarding namely the sensitivity of the NMoE to outliers. This model is an extension of the unconditional mixture of t distributions (Mclachlan & Peel, 1998; Wei, 2012), to the mixture of experts (MoE) framework, where the mixture means are regression functions and the mixing proportions are covariate-varying. For the models inference, we develop a dedicated expectation–maximization (EM) algorithm to estimate the parameters of the proposed model by monotonically maximizing the observed data log-likelihood. The EM algorithm is indeed a very popular and successful estimation algorithm for mixture models in general and for mixture of experts in particular. Indeed, the EM algorithm for MoE has been shown by Ng and McLachlan (2004) to be monotonically maximizing the MoE likelihood. The authors have shown that the EM (with IRLS in this case) algorithm has stable convergence and the log-likelihood is monotonically increasing when a learning rate smaller than one is adopted for the IRLS procedure within the M-step of the EM algorithm. They have further proposed an expectation conditional maximization (ECM) algorithm to train MoE, which also has desirable numerical properties. Beyond the frequentist framework we consider here, we also mention the MoE has also been considered in the Bayesian framework, for example one can cite the Bayesian MoE Waterhouse (1997), Waterhouse, Mackay, and Robinson (1996) and the Bayesian hierarchical MoE Bishop and Svensén (2003). Beyond the Bayesian parametric framework, the MoE models have also been investigated within the Bayesian non-parametric framework. We cite for example the Bayesian non-parametric MoE model (Rasmussen & Ghahramani, 2001) and the Bayesian non-parametric hierarchical MoE approach of Shi et al. (2005) using Gaussian Processes experts for regression. For further models on mixture of experts for regression, the reader can refer to for example the book of Shi and Choi (2011). In this paper, we investigate semi-parametric models under the maximum likelihood estimation framework.

The remainder of this paper is organized as follows. In Section 2 we briefly recall the MoE framework, particularly the NMoE model and its maximum-likelihood estimation via EM. Then, in Section 3 we present the TMoE model and derive its parameter estimation technique using the EM algorithm in Section 4. Next, in Section 5 we investigate the use of the proposed models for fitting non-linear regression functions as well for prediction. We also show in Section 6 how the models can be used in a model-based clustering prospective. In Section 7, we discuss the model selection problem for the model. In Section 8, we perform experiments to assess the proposed models. Finally, Section 9 is dedicated to conclusions and future work.

2. Mixture of experts for continuous data

Mixture of experts (Jacobs et al., 1991; Jordan & Jacobs, 1994) is used in a variety of contexts including regression, classification and clustering. Here we consider the MoE framework for fitting (non-linear) regression functions and clustering of univariate continuous data. 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. Thanks to their great flexibility, mixture models (McLachlan & Peel, 2000) took much attention for non-linear regression problems and we distinguish in particular the classical mixture of regressions model (Faria & Soromenho, 2010; Gaffney & Smyth, 1999; Hunter & Young, 2012; Jones & McLachlan, 1992; Quandt, 1972; Quandt & Ramsey, 1978; Veaux, 1989; Viele & Tong, 2002) and mixture of experts for regression analysis (Jacobs et al., 1991; Jordan & Jacobs, 1994; Young & Hunter, 2010). The univariate mixture of regressions model assumes that the observed pairs of data $(\boldsymbol{x}, y)$ where $y \in \mathbb{R}$ is the response for some covariate $\boldsymbol{x} \in \mathbb{R}^p$, are generated from K regression functions and are governed by a hidden categorical random variable Z indicating from which component each observation is generated. Thus, the mixture of regressions decomposes the nonlinear regression model density $f(y|\boldsymbol{x})$ into a convex weighted sum of K regression components $f_k(y|\boldsymbol{x})$ and can be defined as follows:

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

where the π_k's are defined by $\pi_k = \mathbb{P}(Z = k)$ and represent the non-negative mixing proportions that sum to 1, that is, $\pi_k > 0\ \forall k$ and $\sum_{k=1}^{K} \pi_k = 1$. The model parameter vector is given by $\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_{K-1}, \boldsymbol{\Psi}_1^T, \ldots, \boldsymbol{\Psi}_K^T)^T$, $\boldsymbol{\Psi}_k$ being the parameter vector of the kth component of the mixture density.

2.1. The mixture of experts (MoE) model

Although similar, the mixture of experts (Jacobs et al., 1991) differs from regression mixture models in many aspects. One of the main differences is that the MoE model consists in a fully conditional mixture while in the regression mixture, only the component densities are conditional on some covariates. Indeed, the mixing proportions are constant for the regression mixture, while in the MoE, they are modeled as a function of some

covariates, generally modeled by logistic or a softmax function. Mixture of experts (MoE) for regression analysis (Jacobs et al., 1991; Jordan & Jacobs, 1994) extends the model (1) by modeling the mixing proportions as function of some covariates $\boldsymbol{r} \in \mathbb{R}^q$. The mixing proportions, known as the gating functions in the context of MoE, are modeled by the multinomial logistic (softmax) model and are defined by:

$$\pi_k(\boldsymbol{r}; \boldsymbol{\alpha}) = \mathbb{P}(Z = k|\boldsymbol{r}; \boldsymbol{\alpha}) = \frac{\exp(\boldsymbol{\alpha}_k^T \boldsymbol{r})}{\sum_{\ell=1}^{K} \exp(\boldsymbol{\alpha}_\ell^T \boldsymbol{r})} \tag{2}$$

where $\boldsymbol{r} \in \mathbb{R}^q$ is a covariate vector, $\boldsymbol{\alpha}_k$ is the q-dimensional coefficients vector associated with $\boldsymbol{r}$ and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_{K-1}^T)^T$ is the parameter vector of the gating network, with $\boldsymbol{\alpha}_K$ being the null vector. Thus, the MoE model consists in a fully conditional mixture model where both the mixing proportions (the gating functions) and the component densities (the experts) are conditional on predictors (respectively denoted here by $\boldsymbol{r}$ and $\boldsymbol{x}$).

2.2. The normal MoE (NMoE) model and its maximum likelihood estimation

In the case of MoE for regression, it is usually assumed that the experts are normal, that is, follow a normal distribution. A K-component normal MoE (NMoE) ($K > 1$) has the following formulation:

$$f(y|\boldsymbol{r}, \boldsymbol{x}; \boldsymbol{\Psi}) = \sum_{k=1}^{K} \pi_k(\boldsymbol{r}; \boldsymbol{\alpha}) \mathrm{N}\left(y; \mu(\boldsymbol{x}; \boldsymbol{\beta}_k), \sigma_k^2\right) \tag{3}$$

which involves, in the semi-parametric case, component means defined as parametric (non-)linear regression functions $\mu(\boldsymbol{x}; \boldsymbol{\beta}_k)$.

The NMoE model parameters are estimated by maximizing the observed data log-likelihood by using the EM algorithm (Dempster et al., 1977; Jacobs et al., 1991; Jordan & Jacobs, 1994; Jordan & Xu, 1995; McLachlan & Krishnan, 2008; Ng & McLachlan, 2004). Suppose we observe an i.i.d. sample of n individuals $(y_1, \ldots, y_n)$ with their respective associated covariates $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ and $(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n)$. Then, under the NMoE model, the observed data log-likelihood for the parameter vector $\boldsymbol{\Psi}$ is given by:

$$\log L(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k(\boldsymbol{r}_i; \boldsymbol{\alpha}) \mathrm{N}\left(y_i; \mu(\boldsymbol{x}_i; \boldsymbol{\beta}_k), \sigma_k^2\right). \tag{4}$$

The E-Step at the mth iteration of the EM algorithm for the NMoE model requires the calculation of the following posterior probability that the individual $(y_i, \boldsymbol{x}_i, \boldsymbol{r}_i)$ belongs to expert k, given a parameter estimation $\boldsymbol{\Psi}^{(m)}$:

$$\begin{aligned} \tau_{ik}^{(m)} &= \mathbb{P}(Z_i = k|y_i, \boldsymbol{x}_i, \boldsymbol{r}_i; \boldsymbol{\Psi}^{(m)}) \\ &= \frac{\pi_k(\boldsymbol{r}_i; \boldsymbol{\alpha}^{(m)}) \mathrm{N}\left(y_i; \mu(\boldsymbol{x}_i; \beta_k^{(m)}), \sigma_k^{2\,(m)}\right)}{f(y_i|\boldsymbol{r}_i, \boldsymbol{x}_i; \boldsymbol{\Psi}^{(m)})}. \end{aligned} \tag{5}$$

Then, the M-step calculates the parameter update $\boldsymbol{\Psi}^{(m+1)}$ by maximizing the well-known Q-function (the expected complete-data log-likelihood), that is:

$$\boldsymbol{\Psi}^{(m+1)} = \arg\max_{\boldsymbol{\Psi} \in \boldsymbol{\Omega}} Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(m)}) \tag{6}$$

where $\boldsymbol{\Omega}$ is the parameter space. For example, in the case of normal mixture of linear experts (NMoLE) where each expert's mean has the following linear form:

$$\mu(\boldsymbol{x}_i; \boldsymbol{\beta}_k) = \boldsymbol{\beta}_k^T \boldsymbol{x}_i, \tag{7}$$

where $\boldsymbol{\beta}_k \in \mathbb{R}^p$ is the vector of regression coefficients of expert component k, the updates for each of the expert component parameters consist in analytically solving a weighted Gaussian linear regression problem and are given by:

$$\boldsymbol{\beta}_k^{(m+1)} = \left[\sum_{i=1}^{n} \tau_{ik}^{(m)} \boldsymbol{x}_i \boldsymbol{x}_i^T\right]^{-1} \sum_{i=1}^{n} \tau_{ik}^{(q)} y_i \boldsymbol{x}_i, \tag{8}$$

$$\sigma_k^{2\,(m+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(m)} \left(y_i - \boldsymbol{\beta}_k^{T\,(m+1)} \boldsymbol{x}_i\right)^2}{\sum_{i=1}^{n} \tau_{ik}^{(m)}}. \tag{9}$$

For the gating network, the parameter update $\boldsymbol{\alpha}^{(m+1)}$ cannot however be obtained in a closed form. It can be calculated by Iteratively Reweighted Least Squares (IRLS) (Chamroukhi, Samé, Govaert, & Aknin, 2009; Chen, Xu, & Chi, 1999; Green, 1984; Jacobs et al., 1991; Jordan & Jacobs, 1994).

However, the normal distribution, used to model experts in the NMoE model, is not adapted to deal with data with heavy tailed data distribution and it is also known that the normal distribution is sensitive to outliers. In the proposed model, we propose a robust fitting of the MoE model, which is adapted to data with heavy-tailed distribution and is more robust to outliers, by using the t distribution. This is the t MoE (TMoE) model which we present in the next section.

3. The t MoE (TMoE) model

The proposed t MoE (TMoE) model is based on the t distribution, which is known as a robust generalization of the normal distribution. The t distribution is recalled in the following section. We also describe its stochastic and hierarchical representations, which will be used to derive those of the proposed TMoE model.

3.1. The t distribution

The use of the t distribution in standard mixture models has been shown to be more robust than the normal distribution to handle outliers in the data and accommodate data with heavy tailed distribution. This has been shown in terms of density modeling and cluster analysis for multivariate data (Mclachlan & Peel, 1998; Peel & Mclachlan, 2000) as well as for univariate data by using a skewed-t mixture model (Lin et al., 2007). The t-distribution with location parameter $\mu \in \mathbb{R}$, scale parameter $\sigma^2 \in (0, \infty)$ and degrees of freedom $\nu \in (0, \infty)$ has the probability density function

$$f(y; \mu, \sigma^2, \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{d_y^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \tag{10}$$

where $d_y^2 = \left(\frac{y-\mu}{\sigma}\right)^2$ denotes the squared Mahalanobis distance between y and μ (σ being the scale parameter), and Γ is the Gamma function given by $\Gamma(x) = \int_0^\infty x^{t-1} e^{-x}\, dx$. The t distribution can be characterized as follows. Let E be a univariate random variable with a standard normal distribution with pdf given by $\phi(.)$. Then, let W be a random variable independent of E and following the gamma distribution, that is $W \sim \text{gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$ where the density function of the gamma distribution is given by $f(u; a, b) = \{b^a u^{a-1}/\Gamma(a)\} \exp(-bu) \mathbb{1}_{(0,\infty)}(u)$; $(a, b) > 0$ and the indicator function $\mathbb{1}_{(0,\infty)}(u) = 1$ for $u > 0$ and is zero elsewhere. Then, a random variable Y having the following representation:

$$Y = \mu + \sigma \frac{E}{\sqrt{W}} \tag{11}$$

follows the t distribution $t_\nu(\mu, \sigma^2, \nu)$ with pdf given by (10). As given in Liu and Rubin (1995) for the multivariate case, a hierarchical representation of the t distribution in this univariate case can be expressed from the stochastic representation (11) as:

$$\begin{aligned} Y_i|w_i &\sim \mathrm{N}\left(\mu, \frac{\sigma^2}{w_i}\right) \\ W_i &\sim \text{gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right). \end{aligned} \tag{12}$$

3.2. The t MoE (TMoE) model

The proposed t MoE (TMoE) model extends the t mixture model to the MoE framework. The mixture of t distributions have been first proposed by Mclachlan and Peel (1998), Peel and Mclachlan (2000) for multivariate data. For the univariate case, a K-component t mixture model takes the following form:

$$f(y; \boldsymbol{\Psi}) = \sum_{k=1}^{K} \pi_k \, t(y; \mu_k, \sigma_k^2, \nu_k) \tag{13}$$

where each of the mixture components has a t density given by (10). Lin et al. (2007) proposed a mixture of skew t distributions to deal with heavy-tailed and asymmetric distributions. However, in the skew-t mixture model of Lin et al. (2007), the mixing proportions and the components means are constant and are not predictor-depending does not consider the regression problem and is not a mixture of experts model. Wei (2012) considered the t-mixture model for the regression context on univariate data where the means μ_k in (13) are (linear) regression functions of the form $\mu(\boldsymbol{x}; \boldsymbol{\beta}_k)$. However, this model does not explicitly model the mixing proportions as function of the inputs; they are assumed to be constant.

The proposed t MoE (TMoE) is MoE model with t-distributed experts and is defined as follows. Let $t_\nu(\mu, \sigma^2, \nu)$ denote a t distribution with location parameter μ, scale parameter σ and degrees of freedom ν, whose density is given by (10). A K-component TMoE model is then defined by:

$$f(y|\boldsymbol{r}, \boldsymbol{x}; \boldsymbol{\Psi}) = \sum_{k=1}^{K} \pi_k(\boldsymbol{r}; \boldsymbol{\alpha}) \, t\left(y; \mu(\boldsymbol{x}; \boldsymbol{\beta}_k), \sigma_k^2, \nu_k\right) \tag{14}$$

whose parameter vector 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 $\boldsymbol{\Psi}_k = (\boldsymbol{\beta}_k^T, \sigma_k^2, \nu_k)^T$ is the parameter vector for the kth expert component which has a t distribution. When the robustness parameter $\nu_k \to \infty$ for each k, each t expert component approaches a normal expert and thus the TMoE model (14) approaches the NMoE model (3).

In the following section, we present the stochastic and hierarchical characterizations of the proposed TMoE model and then derive the model maximum likelihood inference scheme.

3.2.1. Stochastic representation of the TMoE

By using the stochastic representation (11) of the t distribution, the one for the t MoE (TMoE) is derived as follows. Let E be a univariate random variable following the standard normal distribution $E \sim \phi(.)$. Suppose that, conditional on the hidden variable $Z_i = z_i$, a random variable W_i is distributed as gamma $\left(\frac{\nu_{z_i}}{2}, \frac{\nu_{z_i}}{2}\right)$. Then, given the covariates $(\boldsymbol{x}_i, \boldsymbol{r}_i)$, a random variable Y_i is said to follow the TMoE model (14) if it has the following representation:

$$Y_i = \mu(\boldsymbol{x}_i; \boldsymbol{\beta}_{z_i}) + \sigma_{z_i} \frac{E_i}{\sqrt{W_{z_i}}}, \tag{15}$$

where the categorical variable Z_i conditional on the covariate $\boldsymbol{r}_i$ follows the multinomial distribution:

$$Z_i|\boldsymbol{r}_i \sim \text{Mult}(1; \pi_1(\boldsymbol{r}_i; \boldsymbol{\alpha}), \ldots, \pi_K(\boldsymbol{r}_i; \boldsymbol{\alpha})) \tag{16}$$

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 multinomial logistic function (2). In this incomplete data framework, z_i represents the hidden label of the expert component generating the ith observation.

3.2.2. Hierarchical representation of the TMoE

By introducing the binary latent component-indicators Z_{ik} such that $Z_{ik} = 1$ iff $Z_i = k$, Z_i being the hidden class label of the ith observation, a hierarchical representation for the TMoE model can be derived from its stochastic representation and is as follows. From (12), (15), and (16), following the hierarchical representation of the mixture of multivariate t-distributions (see for example Mclachlan & Peel, 1998), the hierarchical representation of the TMoE model is written as:

$$\begin{aligned} Y_i|w_i, Z_{ik} = 1, \boldsymbol{x}_i &\sim \mathrm{N}\left(\mu(\boldsymbol{x}_i; \boldsymbol{\beta}_k), \frac{\sigma_k^2}{w_i}\right), \\ W_i|Z_{ik} = 1 &\sim \text{gamma}\left(\frac{\nu_k}{2}, \frac{\nu_k}{2}\right) \\ \boldsymbol{Z}_i|\boldsymbol{r}_i &\sim \text{Mult}\left(1; \pi_1(\boldsymbol{r}_i; \boldsymbol{\alpha}), \ldots, \pi_K(\boldsymbol{r}_i; \boldsymbol{\alpha})\right). \end{aligned} \tag{17}$$

3.3. Identifiability of the TMoE model

Jiang and Tanner (1999b) have established that ordered, initialized, and irreducible MoEs are identifiable. Ordered implies that there exists a certain ordering relationship on the experts parameters $\boldsymbol{\Psi}_k$ such that $(\boldsymbol{\alpha}_1^T, \boldsymbol{\Psi}_1^T)^T \prec \cdots \prec (\boldsymbol{\alpha}_K^T, \boldsymbol{\Psi}_K^T)^T$; initialized implies that $\boldsymbol{\alpha}_K$, the parameter vector of the Kth gating function $\pi_K(\boldsymbol{r}; \boldsymbol{\alpha})$, is the null vector, and irreducible implies that $\boldsymbol{\Psi}_k \neq \boldsymbol{\Psi}_{k'}$ for any $k \neq k'$. For the proposed TMoE model, ordered implies that there exists a certain ordering relationship such that $(\boldsymbol{\beta}_1^T, \sigma_1^2, \nu_1)^T \prec \cdots \prec (\boldsymbol{\beta}_K^T, \sigma_K^2, \nu_K)^T$; initialized implies that $\boldsymbol{\alpha}_K$ is the null vector, as assumed here in the model, and finally irreducible implies that if $k \neq k'$, then one of the following conditions holds: $\boldsymbol{\beta}_k \neq \boldsymbol{\beta}_{k'}$, $\sigma_k \neq \sigma_{k'}$, or $\nu_k \neq \nu_{k'}$. Then, we can establish the identifiability of ordered and initialized irreducible TMoE models by applying Lemma 2 of Jiang and Tanner (1999b), which requires the validation of the following nondegeneracy condition. The set $\{t(y; \mu(\boldsymbol{x}; \boldsymbol{\beta}_1), \sigma_1^2, \nu_1), \ldots, t(y; \mu(\boldsymbol{x}; \boldsymbol{\beta}_{3K}), \sigma_{3K}^2, \nu_{3K})\}$ contains $3K$ linearly independent functions of y, for any $3K$ distinct triplet $(\mu(\boldsymbol{x}; \boldsymbol{\beta}_k), \sigma_k^2, \nu_k)$ for $k = 1, \ldots, 3K$. Thus, via Lemma 2 of Jiang and Tanner (1999b) we have any ordered and initialized irreducible TMoE is identifiable.

4. Maximum likelihood estimation of the TMoE model

Given an i.i.d. sample of n observations, the unknown parameter vector $\boldsymbol{\Psi}$ can be estimated by maximizing the observed-data log-likelihood, which, under the TMoE model, is given by:

$$\log L(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k(\boldsymbol{r}_i; \boldsymbol{\alpha}) \, t\left(y_i; \mu(\boldsymbol{x}_i; \boldsymbol{\beta}_k), \sigma_k^2, \nu_k\right). \tag{18}$$

To perform this maximization, we first use the EM algorithm and then describe an extension based on the ECM algorithm (Meng & Rubin, 1993) as in Liu and Rubin (1995) for a single t distribution, and as in Mclachlan and Peel (1998) and Peel and Mclachlan (2000) for mixture of t-distributions.

4.1. The EM algorithm for the TMoE model

To maximize the log-likelihood function (18) for the TMoE model, the EM algorithm starts with an initial parameter vector

$\boldsymbol{\Psi}^{(0)}$ and alternates between the E- and M-steps until convergence. The E-step computes the expected completed data log-likelihood (the Q-function) and the M-Step maximizes it. From the hierarchical representation of the TMoE (17), the complete data consist of the responses $(y_1, \ldots, y_n)$ and their corresponding covariates $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ and $(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n)$, as well as the latent variables $(w_1, \ldots, w_n)$ and the latent component labels $(z_1, \ldots, z_n)$. Thus, the complete-data log-likelihood of $\boldsymbol{\Psi}$ is given by:

$$\begin{aligned}\log L_c(\boldsymbol{\Psi}) &= \sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik}\big[\log\left(\mathbb{P}\left(Z_i = k|\boldsymbol{r}_i\right)\right) + \log\left(f\left(w_i|Z_{ik}=1\right)\right) \\ &\quad + \log\left(f\left(y_i|w_i, Z_{ik}=1, \boldsymbol{x}_i\right)\right)\big] \\ &= \log L_{1c}(\boldsymbol{\alpha}) + \sum_{k=1}^{K}\big[\log L_{2c}(\boldsymbol{\theta}_k) + \log L_{3c}(\nu_k)\big], \end{aligned} \tag{19}$$

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

$$\log L_{1c}(\boldsymbol{\alpha}) = \sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik}\log \pi_k(\boldsymbol{r}_i;\boldsymbol{\alpha}), \tag{20}$$

$$\log L_{1c}(\boldsymbol{\theta}_k) = \sum_{i=1}^{n} Z_{ik}\Big[-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma_k^2) - \frac{1}{2}w_i d_{ik}^2\Big], \tag{21}$$

$$\begin{aligned}\log L_{3c}(\nu_k) = \sum_{i=1}^{n} Z_{ik}\Big[&-\log\Gamma\left(\frac{\nu_k}{2}\right) + \left(\frac{\nu_k}{2}\right)\log\left(\frac{\nu_k}{2}\right) \\ &+ \left(\frac{\nu_k}{2}-1\right)\log(w_i) - \left(\frac{\nu_k}{2}\right)w_i\Big]. \end{aligned} \tag{22}$$

4.2. E-step

The E-Step of the EM algorithm for the TMoE calculates the Q-function, that is the conditional expectation of the complete-data log-likelihood (19), given the observed data and a current parameter estimation $\boldsymbol{\Psi}^{(m)}$, m being the current iteration. It can be seen from (20)–(22) that computing the Q-function requires the following conditional expectations:

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

It follows that the Q-function is given by:

$$Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(m)}) = Q_1(\boldsymbol{\alpha};\boldsymbol{\Psi}^{(m)}) + \sum_{k=1}^{K}\left[Q_2(\boldsymbol{\theta}_k, \boldsymbol{\Psi}^{(m)}) + Q_3(\nu_k, \boldsymbol{\Psi}^{(m)})\right], \tag{23}$$

where

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

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

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

These conditional expectations are given as follows. First, the conditional expectation $\mathbb{E}_{\boldsymbol{\Psi}^{(m)}}\left[Z_{ik}|y_i, \boldsymbol{x}_i, \boldsymbol{r}_i\right]$, which corresponds to the posterior component memberships, is given by:

$$\tau_{ik}^{(m)} = \frac{\pi_k(\boldsymbol{r}_i;\boldsymbol{\alpha}^{(m)})\, t(y_i;\mu(\boldsymbol{x}_i;\boldsymbol{\beta}_k^{(m)}), \sigma_k^{2\,(m)}, \nu_k^{(m)})}{f(y_i|\boldsymbol{r}_i, \boldsymbol{x}_i;\boldsymbol{\Psi}^{(m)})}. \tag{24}$$

Then, it can be easily shown (see for example Liu & Rubin, 1995, Mclachlan & Peel, 1998 and Peel & Mclachlan, 2000 for details) that:

$$\mathbb{E}_{\boldsymbol{\Psi}^{(m)}}\left[W_i|y_i, Z_{ik}=1, \boldsymbol{x}_i, \boldsymbol{r}_i\right] = \frac{\nu_k^{(m)}+1}{\nu_k^{(m)} + d_{ik}^{2\,(m)}} = w_{ik}^{(m)}, \tag{25}$$

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\Psi}^{(m)}}\left[\log(W_i)|y_i, Z_{ik}=1, \boldsymbol{x}_i, \boldsymbol{r}_i\right] = \log\left(w_{ik}^{(m)}\right) + \Bigg\{\psi\left(\frac{\nu_k^{(m)}+1}{2}\right) \\ - \log\left(\frac{\nu_k^{(m)}+1}{2}\right)\Bigg\} = e_{1,ik}^{(m)}, \end{aligned} \tag{26}$$

where $\psi(x) = \{\partial\Gamma(x)/\partial x\}/\Gamma(x)$ is the Digamma function.

4.3. M-step

In the M-step, as it can be seen from (23), the Q-function can be maximized by independently maximizing $Q_1(\boldsymbol{\alpha};\boldsymbol{\Psi}^{(m)})$, and, for each k, $Q_2(\boldsymbol{\Psi}_k;\boldsymbol{\Psi}^{(m)})$, $Q_3(\nu_k;\boldsymbol{\Psi}^{(m)})$, with respect to $\boldsymbol{\alpha}$, $\boldsymbol{\Psi}_k$ and ν_k, respectively. Thus, on the $(m+1)$th iteration of the EM algorithm, the model parameters are updated as follows.

M-step 1. Calculate $\boldsymbol{\alpha}^{(m+1)}$ by maximizing $Q_1(\boldsymbol{\alpha};\boldsymbol{\Psi}^{(m)})$ w.r.t. $\boldsymbol{\alpha}$:

$$\boldsymbol{\alpha}^{(m+1)} = \arg\max_{\boldsymbol{\alpha}} Q_1(\boldsymbol{\alpha};\boldsymbol{\Psi}^{(m)}). \tag{27}$$

Unlike the case of the standard t mixture model (e.g., Mclachlan & Peel, 1998, Peel & Mclachlan, 2000) and t regression mixture model (Bai et al., 2012; Ingrassia et al., 2012; Wei, 2012), for which the mixing proportions are not predictor-depending and their update is done in closed form, for the proposed TMoE does, there is no closed form solution to update the gating network parameters. This is performed by Iteratively Reweighted Least Squares (IRLS).

The Iteratively Reweighted Least Squares (IRLS) algorithm:

The IRLS algorithm is used to maximize $Q_1(\boldsymbol{\alpha}, \boldsymbol{\Psi}^{(m)})$ with respect to the parameter $\boldsymbol{\alpha}$ in the M-Step at each iteration m of the EM algorithm. The IRLS is a Newton–Raphson algorithm and consists in starting with an initial vector $\boldsymbol{\alpha}^{(0)}$, and, at the $(l+1)$th iteration of the IRLS, updating the estimation of $\boldsymbol{\alpha}$ as follows:

$$\boldsymbol{\alpha}^{(l+1)} = \boldsymbol{\alpha}^{(l)} - \left[\frac{\partial^2 Q_1(\boldsymbol{\alpha}, \boldsymbol{\Psi}^{(m)})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^T}\right]_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^{(l)}}^{-1} \frac{\partial Q_1(\boldsymbol{\alpha}, \boldsymbol{\Psi}^{(m)})}{\partial\boldsymbol{\alpha}}\Big|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^{(l)}} \tag{28}$$

where $\frac{\partial^2 Q_1(\boldsymbol{\alpha},\boldsymbol{\Psi}^{(m)})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^T}$ and $\frac{\partial Q_1(\boldsymbol{\alpha},\boldsymbol{\Psi}^{(m)})}{\partial\boldsymbol{\alpha}}$ are respectively the Hessian matrix and the gradient vector of $Q_1(\boldsymbol{\alpha}, \boldsymbol{\Psi}^{(m)})$. At each IRLS iteration the Hessian and the gradient are evaluated at $\boldsymbol{\alpha} = \boldsymbol{\alpha}^{(l)}$ and are computed analytically similarly as in Chamroukhi et al. (2009). The parameter update $\boldsymbol{\alpha}^{(m+1)}$ in (27) is taken at convergence of the IRLS algorithm (28). Then, for $k = 1 \ldots, K$:

M-Step 2. Calculate $\boldsymbol{\theta}_k^{(m+1)}$ by maximizing $Q_2(\boldsymbol{\theta}_k;\boldsymbol{\Psi}^{(m)})$ w.r.t. $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k^T, \sigma_k^2)^T$. This is achieved by first maximizing $Q_2(\boldsymbol{\theta}_k;\boldsymbol{\Psi}^{(m)})$ w.r.t. $\boldsymbol{\beta}_k$ and then w.r.t. σ_k^2. For the t mixture of linear experts (TMoLE) case where the expert means have the form (7), this maximization is performed analytically and provides the following updates:

$$\boldsymbol{\beta}_k^{(m+1)} = \left[\sum_{i=1}^{n}\tau_{ik}^{(m)}w_{ik}^{(m)}\boldsymbol{x}_i\boldsymbol{x}_i^T\right]^{-1}\sum_{i=1}^{n}\tau_{ik}^{(q)}w_{ik}^{(m)}y_i\boldsymbol{x}_i, \tag{29}$$

$$\sigma_k^{2\,(m+1)} = \frac{1}{\sum_{i=1}^{n}\tau_{ik}^{(m)}}\sum_{i=1}^{n}\tau_{ik}^{(m)}w_{ik}^{(m)}\left(y_i - \boldsymbol{\beta}_k^{T\,(m+1)}\boldsymbol{x}_i\right)^2. \tag{30}$$

Here, we note that, following Kent, Tyler, and Vardi (1994) in the case of ML estimation for single component t distribution and Mclachlan and Peel (1998), Peel and Mclachlan (2000) for mixture of multivariate t distributions, the EM algorithm can be modified slightly by replacing the divisor $\sum_{i=1}^{n} \tau_{ik}^{(m)}$ in (30) by $\sum_{i=1}^{n} \tau_{ik}^{(m)} w_{ik}^{(m)}$. This modified algorithm may converge faster than the conventional EM algorithm.

M-Step 3. Calculate $\nu_k^{(m+1)}$ by maximizing $Q_3(\nu_k; \boldsymbol{\Psi}^{(m)})$ w.r.t. ν_k. The degrees of freedom update $\nu_k^{(m+1)}$ is therefore obtained by iteratively solving the following equation for ν_k:

$$
\begin{aligned}
&-\psi\left(\frac{\nu_k}{2}\right)+\log\left(\frac{\nu_k}{2}\right)+1\\
&+\frac{1}{\sum_{i=1}^{n}\tau_{ik}^{(m)}}\sum_{i=1}^{n}\tau_{ik}^{(m)}\left(\log(w_{ik}^{(m)})-w_{ik}^{(m)}\right)\\
&+\psi\left(\frac{\nu_k^{(m)}+1}{2}\right)-\log\left(\frac{\nu_k^{(m)}+1}{2}\right)=0. \qquad (31)
\end{aligned}
$$

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

It is obvious to see that, as mentioned previously, if the number of degrees of freedom ν_k approaches infinity for all k, then the parameter updates for the TMoE model are exactly those of the NMoE model (since w_{ik} tends to 1 in that case). The TMoE model constitutes therefore a robust generalization of the NMoE model, which is able to model data with density heaving longer tails than those of the NMoE model.

After deriving the EM algorithm for the parameter estimation of the TMoE model, now we describe an ECM extension.

4.4. The ECM algorithm for the TMoE model

Following the ECM extension of the EM algorithm for a single t distribution proposed by Liu and Rubin (1995) and the one of the EM algorithm for the t-mixture model (Mclachlan & Peel, 1998; Peel & Mclachlan, 2000), the EM algorithm for the TMoE model can also be modified to give an ECM version by adding an additional E-Step between the two M-steps 2 and 3. This additional E-step consists in taking the parameter vector $\boldsymbol{\Psi}$ with $\boldsymbol{\theta}_k = \boldsymbol{\theta}_k^{(m+1)}$ instead of $\boldsymbol{\theta}_k^{(m)}$, that is

$$Q_3(\nu_k; \boldsymbol{\Psi}^{(m)}) = Q_3(\nu_k; \boldsymbol{\alpha}^{(m)}, \boldsymbol{\theta}_k^{(m+1)}, \nu_k^{(m)}).$$

Thus, the M-Step 3 in the above is replaced by a Conditional-Maximization (CM)-Step in which the degrees of freedom update (31) is calculated with the conditional expectation (25) and (26) computed with the updated parameters $\boldsymbol{\beta}_k^{(m+1)}$ and $\sigma_k^{2(m+1)}$ respectively given by (29) and (30).

The TMoE handles therefore the problem of heavy tailed data possibly affected by outliers. It therefore provides a more robust modeling framework for fitting MoE to data. In the next section, we show how to use the TMoE in fitting regression functions and clustering, and we discuss the question of model selection.

5. Prediction using the TMoE

The goal in regression is to be able to make predictions for the response variable(s) given some new value of the predictor variable(s) on the basis of a model trained on a set of training data. In regression analysis using MoE, the aim is therefore to predict the response y given new values of the predictors $(\boldsymbol{x}, \boldsymbol{r})$, on the basis of a MoE model characterized by a parameter vector $\hat{\boldsymbol{\Psi}}$ inferred from a set of training data, here, by maximum likelihood via EM. These predictions can be expressed in terms of the predictive distribution of y, which is obtained by substituting the maximum likelihood parameter $\hat{\boldsymbol{\Psi}}$ into (1)–(2) to give:

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

Using f, we might then predict y for a given set of $\boldsymbol{x}$'s and $\boldsymbol{r}$'s as the expected value under f, that is by calculating the prediction $\hat{y} = \mathbb{E}_{\hat{\boldsymbol{\Psi}}}(Y|\boldsymbol{r}, \boldsymbol{x})$. We thus need to compute the expectation of the MoE model. It is easy to show (see for example Section 1.2.4 in Frühwirth-Schnatter, 2006) that the mean and the variance of a MoE distribution of the form (5) are respectively given by:

$$\mathbb{E}_{\hat{\boldsymbol{\Psi}}}(Y|\boldsymbol{r}, \boldsymbol{x}) = \sum_{k=1}^{K} \pi_k(\boldsymbol{r}; \hat{\boldsymbol{\alpha}}_n) \mathbb{E}_{\hat{\boldsymbol{\Psi}}}(Y|Z=k, \boldsymbol{x}), \qquad (32)$$

$$
\begin{aligned}
\mathbb{V}_{\hat{\boldsymbol{\Psi}}}(Y|\boldsymbol{r}, \boldsymbol{x}) = &\sum_{k=1}^{K} \pi_k(\boldsymbol{r}; \hat{\boldsymbol{\alpha}}_n)\left[\left(\mathbb{E}_{\hat{\boldsymbol{\Psi}}}(Y|Z=k, \boldsymbol{x})\right)^2\right.\\
&\left.+\ \mathbb{V}_{\hat{\boldsymbol{\Psi}}}(Y|Z=k, \boldsymbol{x})\right] - \left[\mathbb{E}_{\hat{\boldsymbol{\Psi}}}(Y|\boldsymbol{r}, \boldsymbol{x})\right]^2, \qquad (33)
\end{aligned}
$$

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. The mean and the variance for the MoE models described here are given as follows.

NMoE. For the NMoE model, the normal expert means and variances are respectively $\mathbb{E}_{\hat{\boldsymbol{\Psi}}}(Y|Z=k, \boldsymbol{x}) = \hat{\boldsymbol{\beta}}_k^T \boldsymbol{x}$ and $\mathbb{V}_{\hat{\boldsymbol{\Psi}}}(Y|Z=k, \boldsymbol{x}) = \hat{\sigma}_k^2$.

TMoE. For the TMoE model, by using the expressions of the mean and the variance of the t distribution, it follows that for the TMoE model, for $\hat{\nu}_k > 1$, the expert means are $\mathbb{E}_{\hat{\boldsymbol{\Psi}}}(Y|Z=k, \boldsymbol{x}) = \hat{\boldsymbol{\beta}}_k^T \boldsymbol{x}$ and, for $\hat{\nu}_k > 2$, the expert variances are $\mathbb{V}_{\hat{\boldsymbol{\Psi}}}(Y|Z=k, \boldsymbol{x}) = \frac{\hat{\nu}_k}{\hat{\nu}_k - 2}\hat{\sigma}_k^2$.

6. Model-based clustering using the TMoE

It is natural to utilize the MoE models for a model-based clustering perspective to provide a partition of the regression data into K clusters. Model-based clustering using the TMoE, as in MoE in general, consists in assuming that the observed data $\{\boldsymbol{x}_i, \boldsymbol{r}_i, y_i\}_{i=1}^n$ are generated from a K component mixture of t experts with parameter vector $\boldsymbol{\Psi}$. The mixture components can be interpreted as clusters and hence each cluster can be associated with a mixture component. The problem of clustering therefore becomes the one of estimating the MoE parameters $\boldsymbol{\Psi}$, which is performed here by using dedicated EM algorithms. Once the parameters are estimated, the provided posterior component memberships $\hat{\tau}_{ik}$ defined in (24) represent a fuzzy partition of the data. A hard partition of the data can then be obtained by applying the optimal Bayes' allocation rule, that is:

$$\hat{z}_i = \arg\max_{k=1}^{K} \hat{\tau}_{ik} \qquad (34)$$

where $\hat{z}_i$ represents the estimated cluster label for the ith observation.

7. Model selection for the NNMoE

One of the issues in mixture model-based clustering is model selection. The problem of model selection for the TMoE model

presented here in its general form, is equivalent to the one of choosing the optimal number of experts K, the degree p of the polynomial regression and the degree q for the logistic regression. The optimal value of (K, p, q) can be computed by using some model selection criteria such as the Akaike Information Criterion (AIC) (Akaike, 1974), the Bayesian Information Criterion (BIC) (Schwarz, 1978) or the Integrated Classification Likelihood criterion (ICL) (Biernacki, Celeux, & Govaert, 2000), etc. The AIC and BIC are penalized observed data log-likelihood criteria which can be defined as functions to be maximized and are respectively given by:

$$\mathrm{AIC}(K, p, q) = \log L(\hat{\boldsymbol{\Psi}}) - \frac{\eta_{\boldsymbol{\Psi}} \log(n)}{2},$$

$$\mathrm{BIC}(K, p, q) = \log L(\hat{\boldsymbol{\Psi}}) - \frac{\eta_{\boldsymbol{\Psi}} \log(n)}{2}.$$

The ICL criterion consists in a penalized complete-data log-likelihood and can be expressed as:

$$\mathrm{ICL}(K, p, q) = \log L_c(\hat{\boldsymbol{\Psi}}) - \frac{\eta_{\boldsymbol{\Psi}} \log(n)}{2}.$$

In the above, $\log L(\hat{\boldsymbol{\Psi}})$ and $\log L_c(\hat{\boldsymbol{\Psi}})$ are respectively the incomplete (observed) data log-likelihood and the complete data log-likelihood, obtained at convergence of the E(C)M algorithm for the corresponding MoE model. The number of free parameters of the model $\eta_{\boldsymbol{\Psi}}$ is given by $\eta_{\boldsymbol{\Psi}} = K(p + q + 3) - q - 1$ for the NMoE model and $\eta_{\boldsymbol{\Psi}} = K(p + q + 4) - q - 1$ for the TMoE model.

However, note that in MoE it is common to use gating functions modeled as logistic transformation of linear functions of the covariates, that is the covariate vector in (2) is given by $\boldsymbol{r}_i = (1, r_i)^T$ (corresponding to $q = 2$), r_i being a univariate covariate variable. This is what we adopted in this work. Moreover, for the case of linear experts, that is when the experts are linear regressors with parameter vector $\boldsymbol{\beta}_k$ for which the corresponding covariate vector $\boldsymbol{x}_i$ in (7) is given by $\boldsymbol{x}_i = (1, x_i)^T$ (corresponding to $p = 2$), x_i being a univariate covariate variable possibly different from r_i, the model selection reduces to choosing the number of experts K. Here in the presented experiments we mainly consider this linear case for the expert components. Notice that the overall modeling problem is still non-linear and is adapted to fit non-linear regression functions.

8. Experimental study

This section is dedicated to the evaluation of the proposed approach on simulated data and real-world data. We evaluated the performance of proposed EM algorithm by comparing it with the standard normal MoE (NMoE) model (Jacobs et al., 1991; Jordan & Jacobs, 1994) and the Laplace MoE of (Nguyen & McLachlan, 2016)[1] on both simulated and real-world data sets.

8.1. Initialization and stopping rules

The parameters $\boldsymbol{\alpha}_k$ $(k = 1, \ldots, K-1)$ of the mixing proportions are initialized randomly, including an initialization at the null vector for one run (corresponding to equal mixing proportions). Then, the common parameters $(\boldsymbol{\beta}_k, \sigma_k^2)$ $(k = 1, \ldots, K)$ are initialized from a random partition of the data into K clusters. This corresponds to fitting a normal MoE where the initial values of the parameters are respectively given by (8) and (9) with the posterior memberships τ_{ik} replaced by the hard assignments Z_{ik} issued from the random partition. For the TMoE model, the robustness parameter ν_k $(k = 1, \ldots, K)$ is initialized randomly in the range $[1, 200]$. For the LMoE model

$$f(y|\boldsymbol{r}, \boldsymbol{x}; \boldsymbol{\Psi}) = \sum_{k=1}^{K} \pi_k(\boldsymbol{r}; \boldsymbol{\alpha}) \,\mathrm{Laplace}(y; \mu(\boldsymbol{x}; \boldsymbol{\beta}_k), \lambda_k), \tag{35}$$

the scale parameter λ_k is initialized in a similar way as σ_k^2. Then, the algorithms are stopped when the relative variation of the observed-data log-likelihood $\frac{\log L(\boldsymbol{\Psi}^{(m+1)}) - \log L(\boldsymbol{\Psi}^{(m)})}{|\log L(\boldsymbol{\Psi}^{(m)})|}$ reaches a prefixed threshold (for example $\epsilon = 10^{-6}$). For each model, this process is repeated 10 times and the solution corresponding the highest log-likelihood is finally selected.

8.2. Experiments on simulation data sets

In this section we perform an experimental study on simulated data sets to apply and assess the proposed model. Two sets of experiments have been performed. The first experiment aims at observing the effect of the sample size on the estimation quality and the second one aims at observing the impact of the presence of outliers in the data on the estimation quality, that is the robustness of the models.

8.2.1. Experiment 1

For this first experiment on simulated data, each simulated sample consisted of n observations with increasing values of the sample size n : 50, 100, 200, 500, 1000. The simulated data are generated from a two component mixture of linear experts, that is $K = 2, p = q = 1$. The covariate variables $(\boldsymbol{x}_i, \boldsymbol{r}_i)$ are simulated such that $\boldsymbol{x}_i = \boldsymbol{r}_i = (1, x_i)^T$ where x_i is simulated uniformly over the interval $(-1, 1)$. We consider each of the three models (NMoE, LMoE, TMoE) for data generation, that is, given the covariates, the response $y_i|\{\boldsymbol{x}_i, \boldsymbol{r}_i; \boldsymbol{\Psi}\}$ is simulated according to the generative process of the models (3), (35) and (14). For each generated sample, we fit each of the four models. Thus, the results are reported for all the models with data generated from each of the two models. We consider the mean square error (MSE) between each component of the true parameter vector and the estimated one, which is given by $\|\boldsymbol{\Psi}_j - \hat{\boldsymbol{\Psi}}_j\|^2$. The squared errors are averaged on 100 trials. The used simulation parameters $\boldsymbol{\Psi}$ for each model are given in Table 1.

8.2.2. Obtained results

Table 2 shows the obtained results in terms of the MSE for the TMoE. One can observe that the parameter estimation error is decreasing as n increases, which illustrates the convergence property of the maximum likelihood estimator of the model. For details on the convergence property of the MLE for MoE, see for example Jiang and Tanner (1999a). One can also observe that the error decreases significantly for $n \geq 500$, especially for the regression coefficients and the scale parameters.

In addition to the previously shown results, we plotted in Figs. 1–3 the estimated quantities provided by applying the proposed model and their true counterparts for $n = 500$ for the same the data set which was generated according to the normal MoE model. The upper-left plot of each of these figures shows the estimated mean function, the estimated expert component mean functions, and the corresponding true ones. The upper-right plot shows the estimated mean function and the estimated confidence region computed as plus and minus twice the estimated (pointwise) standard deviation of the model as presented in Section 5, and their true counterparts. The bottom-left plot shows the true expert component mean functions and the true partition, and the bottom-right plot shows their estimated counterparts.

[1] All the algorithms have been implemented in Matlab and the codes are available upon request from the author.

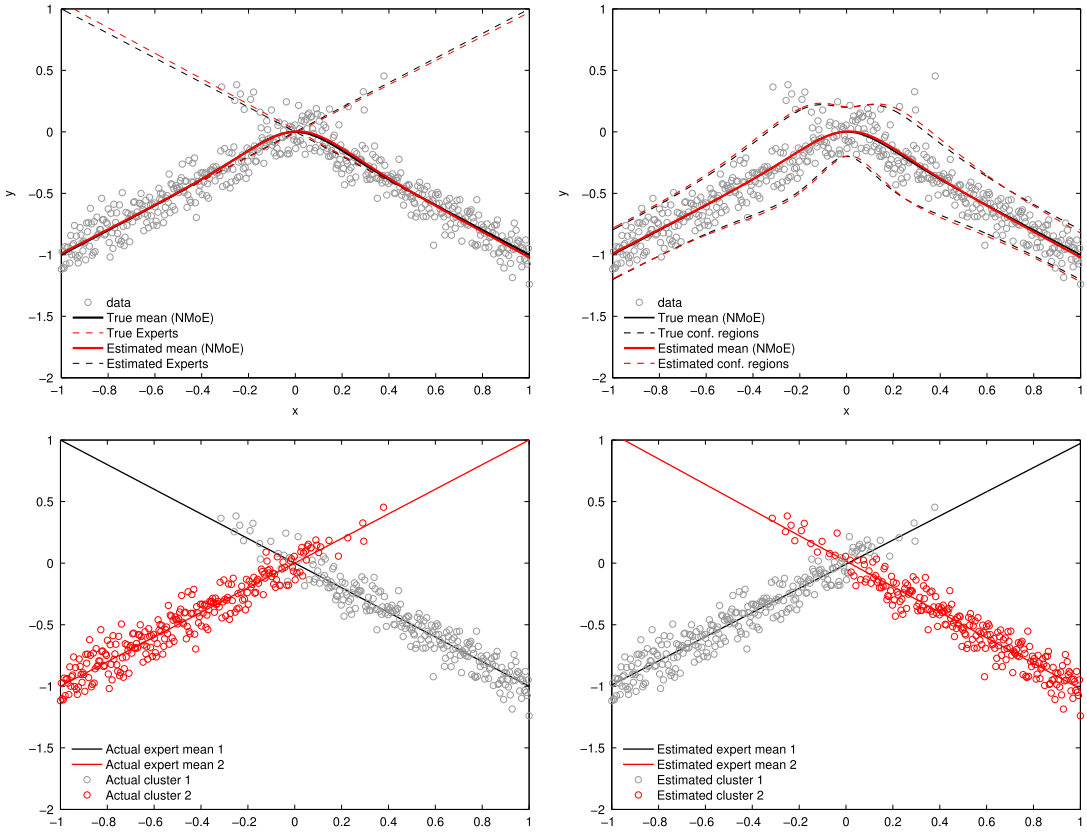

Fig. 1. Fitted NMoE model to a data set generated according to the NMoE model.

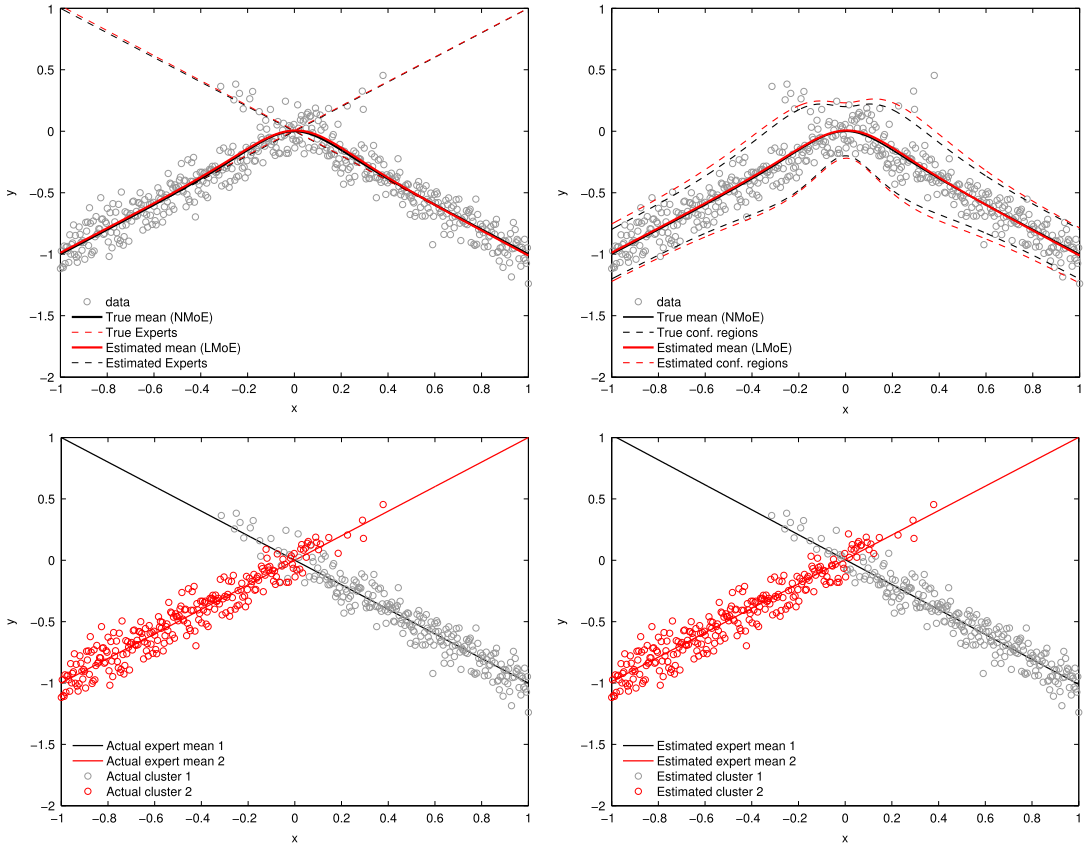

Fig. 2. Fitted LMoE model to a data set generated according to the NMoE model.

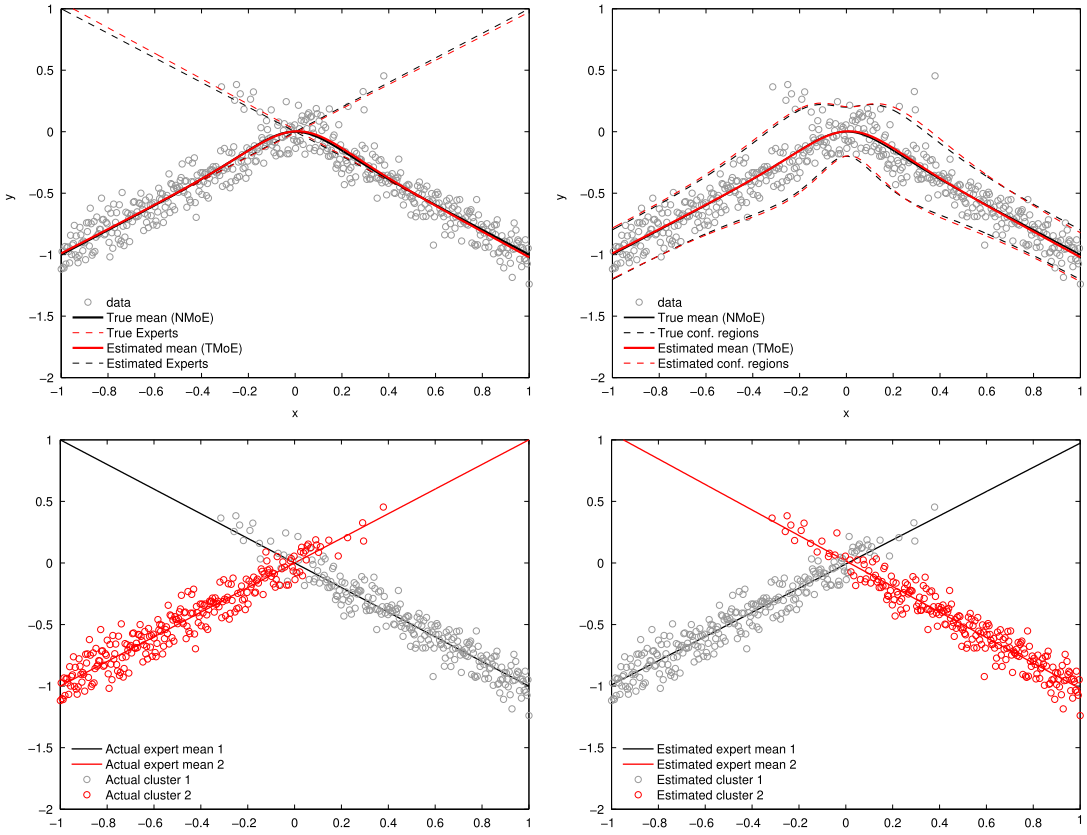

Fig. 3. Fitted TMoE model to a data set generated according to the NMoE model.

Table 1
Parameter values used in simulation.

	Parameters				
Component 1	$\boldsymbol{\alpha}_1 = (0, 10)^T$	$\boldsymbol{\beta}_1 = (0, 1)^T$	$\sigma_1 = 0.1$	$\nu_1 = 5$	$\lambda_1 = 0.1$
Component 2	$\boldsymbol{\alpha}_2 = (0, 0)^T$	$\boldsymbol{\beta}_2 = (0, -1)^T$	$\sigma_2 = 0.1$	$\nu_2 = 7$	$\lambda_2 = 0.1$

Table 2
MSE between each component of the estimated parameter vector of the TMoE model and the actual one for a varying sample size n.

n	Param.									
	α_{10}	α_{11}	β_{10}	β_{11}	β_{20}	β_{21}	σ_1	σ_2	ν_1	ν_2
50	1.3059	6.4611	0.0214130	0.0290114	0.0044140	0.0192600	0.0010655	0.0003317	37.956	11.722
100	1.2150	4.5056	0.0024706	0.0117546	0.0005275	0.0007891	0.0001450	0.0002301	6.1528	10.412
200	0.0341	3.8193	0.0001553	0.0007335	0.0002022	0.0005061	0.0000504	0.0000262	2.0975	6.3710
500	0.0356	2.2633	0.0000112	0.0000214	0.0001337	0.0002163	0.0000126	0.0000007	0.4859	5.4937
1000	0.0053	1.2510	0.0000018	0.0000258	0.0000005	0.0000427	0.0000126	0.0000004	0.0014	2.7844

One can clearly see that the estimations provided by the proposed model are quasi identical to the true ones which correspond to those of the NMoE model in this case. This provides an additional support to the fact that the proposed algorithm performs well and the proposed TMoE model is a good generalization of the normal MoE (NMoE), as it clearly approaches the NMoE as shown in these simulated examples. The proposed TMoE also provides quasi-identical results to the LMoE model.

8.2.3. Experiment 2

In this experiment we examine the robustness of the proposed model to outliers versus the standard NMoE one. For that, we considered each of the three models (NMoE, LMoE, TMoE) for data generation. For each generated sample, each of the two models is considered for the inference. The data were generated exactly in the same way as in Experiment 1, except for some observations which were generated with a probability c from a class of outliers. We considered the same class of outliers as in Nguyen and McLachlan (2016), that is, the predictor x is generated uniformly over the interval $(-1, 1)$ and the response y is set the value -2. We apply the MoE models by setting the covariate vectors as before, that is, $\boldsymbol{x} = \boldsymbol{r} = (1, x)^T$. We considered varying probability of outliers $c = 0\%, 1\%, 2\%, 3\%, 4\%, 5\%$ and the sample size of the generated data is $n = 500$. An example of simulated sample containing 5% outliers is shown in Fig. 4. As a criterion of evaluation of the impact of the outliers on the quality of the results, we considered the MSE between the true regression mean function and the estimated one. This MSE is calculated as $\frac{1}{n}\sum_{i=1}^{n} \|\mathbb{E}_{\boldsymbol{\psi}}(Y_i|\boldsymbol{r}_i, \boldsymbol{x}_i) - \mathbb{E}_{\hat{\boldsymbol{\psi}}}(Y_i|\boldsymbol{r}_i, \boldsymbol{x}_i)\|^2$ where the expectations are computed as in Section 5.

8.2.4. Obtained results

Table 3 shows, for each of the two models, the results in terms of mean squared error (MSE) between the true mean function and

Table 3
MSE between the estimated mean function and the true one for each of the four models for a varying probability c of outliers for each simulation. The first column indicates the model used for generating the data and the second one indicates the model used for inference.

Model		c					
		0%	1%	2%	3%	4%	5%
NMoE	NMoE	0.000178	0.001057	0.001241	0.003631	0.013257	0.028966
	LMoE	0.000144	0.000389	0.000686	0.000153	0.000296	0.000121
	TMoE	0.000168	0.000566	0.000464	0.000221	0.000263	0.000045
LMoE	NMoE	0.000287	0.003830	0.003740	0.010631	0.021247	0.026198
	LMoE	0.000126	0.000378	0.000125	0.000270	0.000165	0.000605
	TMoE	0.000183	0.000273	0.000236	0.000182	0.000168	0.000070
TMoE	NMoE	0.000257	0.0004660	0.002779	0.015692	0.005823	0.005419
	LMoE	0.000288	0.0004568	0.000205	0.000133	0.000146	0.000307
	TMoE	0.000252	0.0002520	0.000144	0.000157	0.000488	0.000245

the estimated one, for an increasing number of outliers in the data. First, one can see that, when there are no outliers ($c = 0\%$), the error of the TMoE is less than those of the NMoE model, for the two situations, that is including the case where the data are not generated according to the TMoE model, which is somewhat surprising. This includes the case where the data are generated according to the NMoE model, for which the TMoE error is slightly less than the one of the NMoE model. Then, it can be seen that when there are outliers, the TMoE model clearly outperforms the NMoE model for all the situations. This confirms that the TMoE model is much more robust to outliers compared to the normal one because the expert components in TMoE follow a robust distribution, that is the t distribution. Furthermore, it can be seen that, when the number of outliers is increasing, the increase in the error of the NMoE model is more pronounced compared to the one of the TMoE model. The error for the TMoE may indeed slightly increase, remains stable or even slightly decreases in some situations when the data are generated according to the TMoE model. This supports the expected robustness of the TMoE and the fact that the NMoE is severely affected by outliers. To make comparison with the LMoE, which is also clearly more robust that the NMoE, it can be seen that for some situations the LMoE provides better results compared to the TMoE, however, the overall results favor the TMoE model, namely in the situation where the noise is relatively high (5% of outliers). To highlight the robustness to noise of the TMoE model, in addition to the previously shown numerical results, Figs. 4–6 show an example of results obtained on the same data set by, respectively, the NMoE, the LMoE, and the TMoE. The data are generated by the NMoE model and contain $c = 5\%$ of outliers.

In this example, we clearly see that the NMoE model is severely affected by the outliers. It provides a rough fit especially for the second component whose estimation is affected by the outliers. However, one can see that the TMoE model provides a precise fit; the estimated mean functions and expert components are very close to the true ones. The TMoE is robust to outliers, in terms of estimating the true model as well as in terms of estimating the true partition of the data (as shown in the middle plots). The solution is also very close to the one provided by the LMoE model. Notice that for the TMoE the confidence region is not shown because for this situation the estimated degrees of freedom are less than 2 (1.5985 and 1.5253) for the TMoE; Hence the variance for the TMoE in that case is not defined (see Section 5). The TMoE model provides indeed components with small degrees of freedom corresponding to highly heavy tails, which allow to handle outliers in this noisy case.

8.3. Application to two real-world data sets

In this section, we consider an application to two real-world data sets: the tone perception data set and the temperature anomalies data set shown in Fig. 7.

8.3.1. Tone perception data set

The first analyzed data set is the real tone perception data set[2] which goes back to Cohen (1984). It was 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. In the tone perception experiment, a pure fundamental tone was played to a trained musician. Electronically generated overtones were added, determined by a stretching ratio ("stretch ratio" $= 2$) which corresponds to the harmonic pattern usually heard in traditional definite pitched instruments. The musician was asked to tune an adjustable tone to the octave above the fundamental tone and a "tuned" measurement gives the ratio of the adjusted tone to the fundamental. The obtained data consists of $n = 150$ pairs of "tuned" variables, considered here as predictors (x), and their corresponding "stretch ratio" variables considered as responses (y). To apply the MoE models, we set the response y_i ($i = 1, \ldots, 150$) as the "stretch 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. We also follow the study in Bai et al. (2012) and Song et al. (2014) by using two mixture components. The model selection results, given later in Table 5, confirm two-components are present in the data when using the TMoE model and the Bayesian Information Criterion (Schwarz, 1978).

Fig. 8 shows the scatter plots of the tone perception data and the linear expert components of the fitted NMoE model, the LMoE model, and the proposed TMoE model. One can observe that we obtain a reasonable fit with the three models. But the one of the NMoE differs slightly from the one of the LMoE and the one of the TMoE (which are quasi-identical), and which, upon a visual inspection, can be seen more adapted by better fitting the two regression lines to the data. The two regression lines may correspond to correct tuning and tuning to the first overtone, respectively, as analyzed in Bai et al. (2012) (also see Song et al., 2014 for the analysis).

Fig. 9 shows the log-likelihood profiles for each of the two models. It can namely be seen that training the t MoE for this experiment may take more iterations than the normal MoE model. The TMoE has indeed more parameters to estimate than the NMoE one, that is, the robustness parameters ν_k. However, in terms of computing time, the models converge in only few seconds on a personal laptop (with 2.9 GHz processor and 8 GB memory).

The values of estimated parameters for the tone perception data set are given in Table 4. One can see that the regression coefficients are very similar for all the models, except for the first component of the NMoE model. This can be observed on the fit in Fig. 8 where the first expert component for the NMoE model slightly differs from the corresponding one of both the LMoE model and the proposed TMoE model. In addition, it can be seen from the values of the

[2] Source: http://artax.karlin.mff.cuni.cz/r-help/library/fpc/html/tonedata.html.

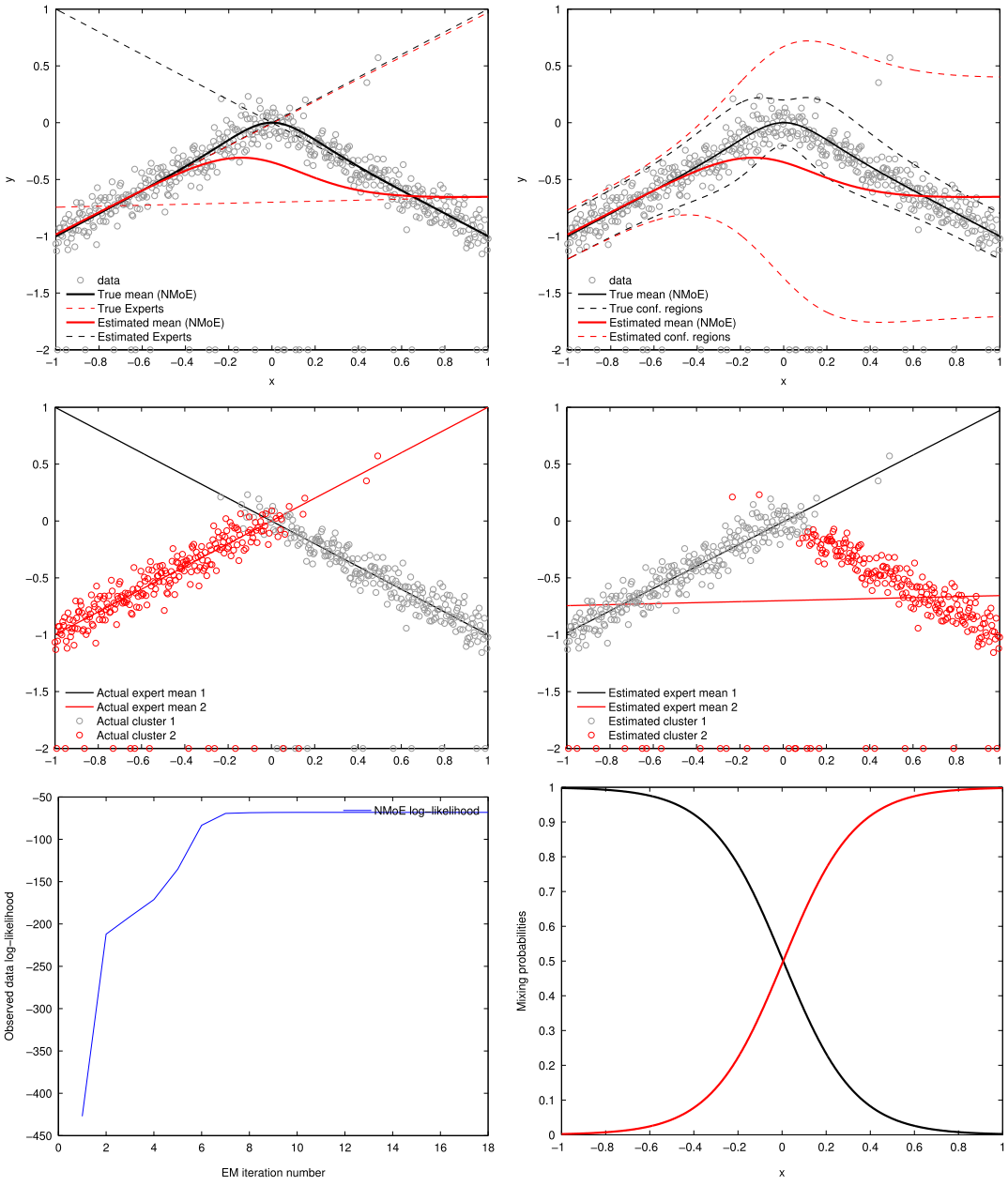

Fig. 4. Fitted NMoE model to a data set of $n = 500$ observations generated according to the NMoE model and including 5% of outliers.

Table 4
Values of the estimated MoE parameters for the original tone perception data set.

Model	Param.											
	α_{10}	α_{11}	β_{10}	β_{11}	β_{20}	β_{21}	σ_1	σ_2	λ_1	λ_2	ν_1	ν_2
NMoE	−2.690	0.796	−0.029	0.995	1.913	0.043	0.137	0.047	–	–	–	–
LMoE	−0.460	0.087	0.0036	0.998	1.961	0.023	–	–	0.049	0.030	–	–
TMoE	−0.058	−0.070	0.002	0.999	1.956	0.027	0.002	0.029	–	–	0.555	2.017

Table 5
Choosing the number of expert components K for the original tone perception data by using the information criteria BIC, AIC, and ICL. Underlined value indicates the highest value for each criterion.

K	NMoE			LMoE			TMoE		
	BIC	AIC	ICL	BIC	AIC	ICL	BIC	AIC	ICL
1	1.8662	6.3821	1.8662	36.8061	41.3220	−7.5160	71.3931	77.4143	71.3931
2	122.8050	134.8476	107.3840	149.6360	161.6786	−20.0425	204.8241	219.8773	186.8415
3	118.1939	137.7630	76.5249	209.1995	228.7687	−32.5691	199.4030	223.4880	183.0389
4	121.7031	148.7989	94.4606	204.3286	231.4244	−45.0957	201.8046	234.9216	187.7673
5	141.6961	176.3184	123.6550	141.3988	176.0211	−57.6223	187.8652	230.0141	164.9629

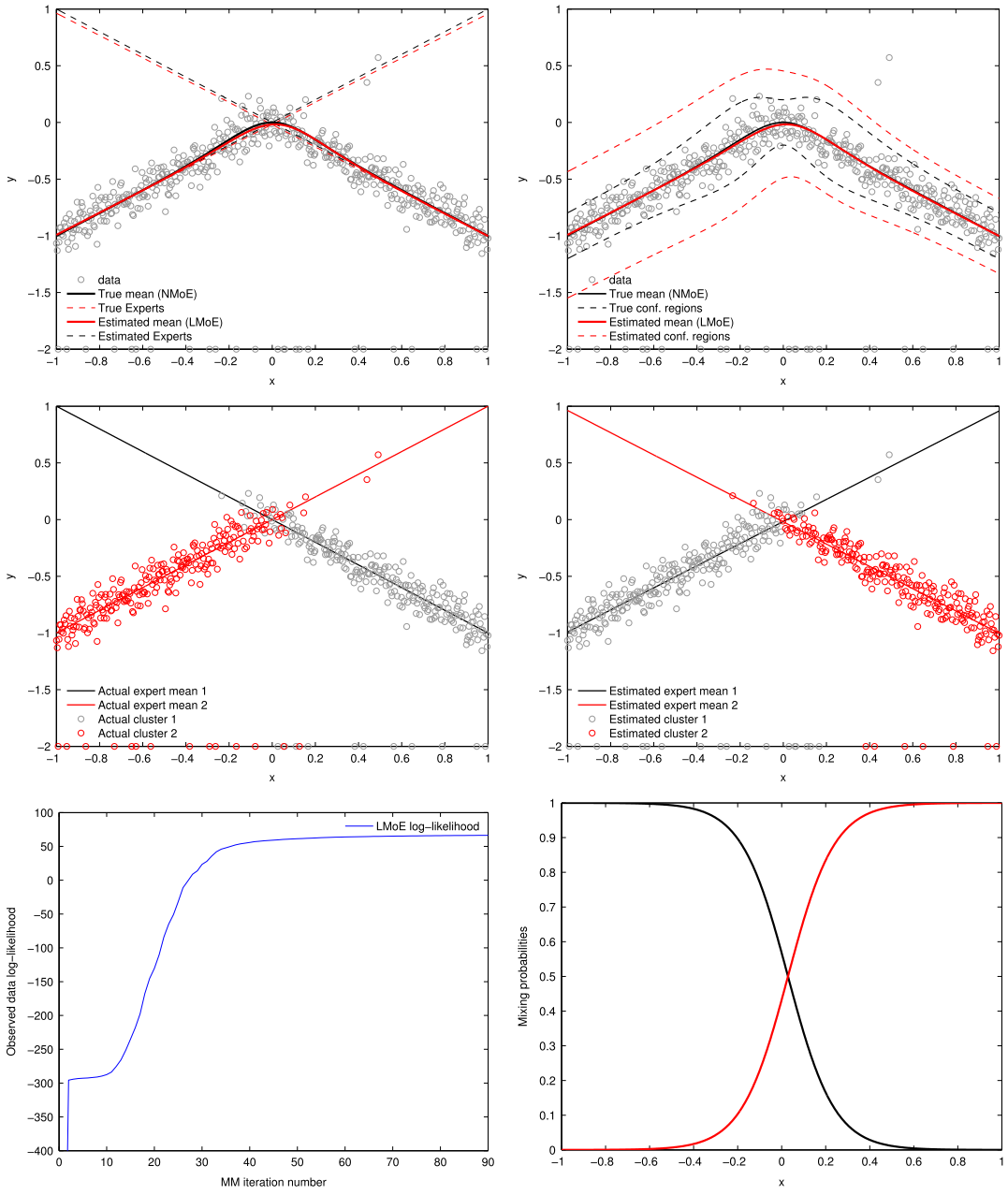

Fig. 5. Fitted LMoE model to a data set of $n = 500$ observations generated according to the NMoE model and including 5% of outliers (the same data set shown in Fig. 4).

common parameters that the LMoE and the TMoE provide very close results.

We also performed a model selection procedure on this data set to choose the best number of MoE components for a number of components between 1 and 5. We used BIC, AIC, and ICL. Table 5 gives the obtained values of the model selection criteria. One can see that the NMoE model overestimates the number of components. AIC performs poorly for all the models. BIC provides the correct number of components for the proposed TMoE model but seems to overestimate the number of components for the LMoE model (provides evidence for 3 components). ICL hesitates between 2 (the correct number) and 4 components for the TMoE model. One can conclude that the BIC is the criterion to be suggested for the analysis. Thus, from this experiment, it would be more adapted to use BIC with the proposed TMoE model.

Robustness to outliers. Now we examine the sensitivity of the MoE models to outliers based on this real data set. For this, we adopt 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. We apply the MoE models in the same way as before.

The left plot in Fig. 10 shows that the normal MoE is sensitive to outliers. However, 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 is 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). One can also see that, even the regression mean functions are affected severely by the outliers, the provided partitions are still reasonable and similar to those provided in the previous non-noisy case. Then, the middle plot of in Fig. 10 shows that the LMoE model is more robust to outliers compared to the NMoE model, however, the regression line is not very well adjusted

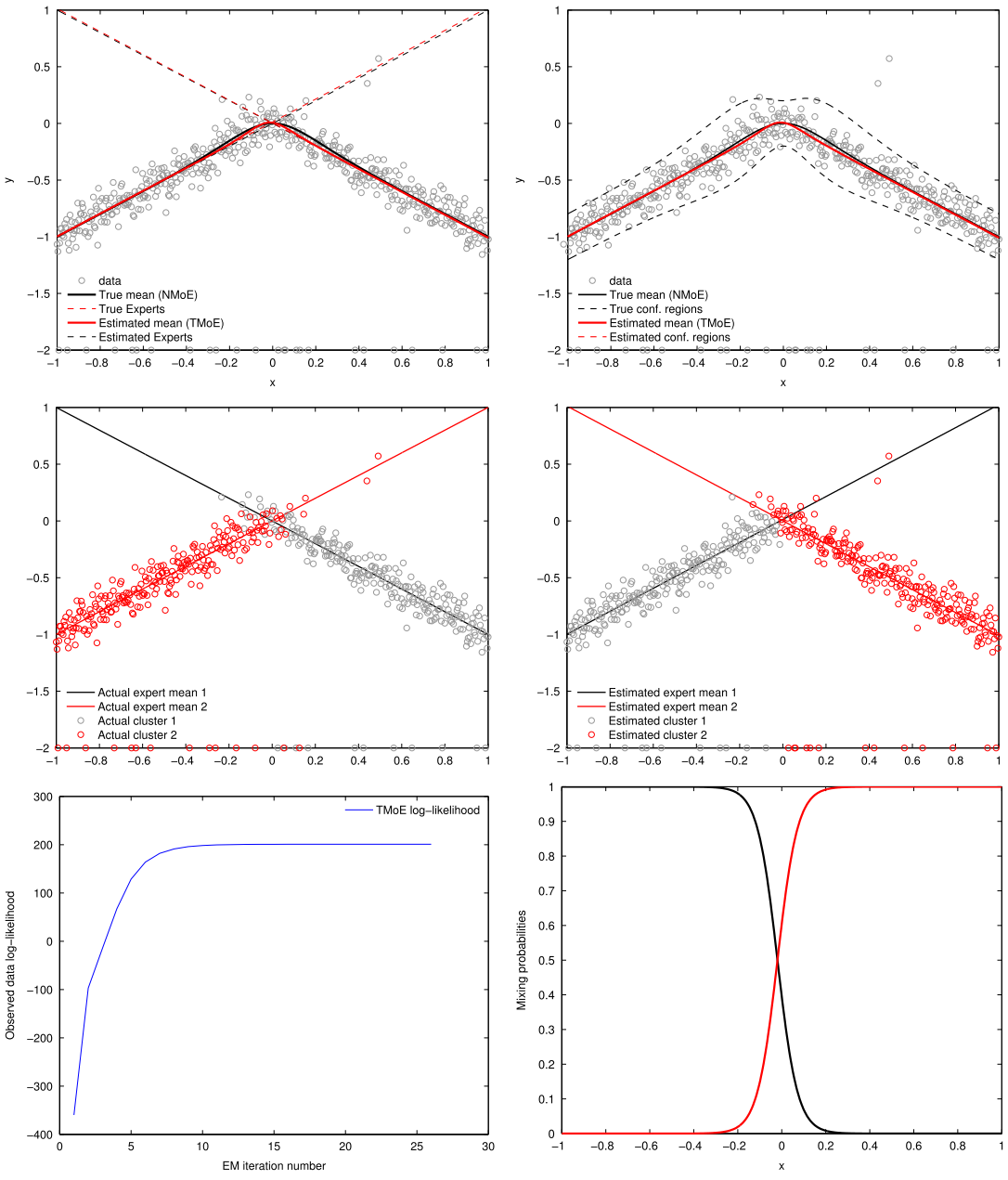

Fig. 6. Fitted TMoE model to a data set of $n = 500$ observations generated according to the NMoE model and including 5% of outliers (the same data set shown in Fig. 4).

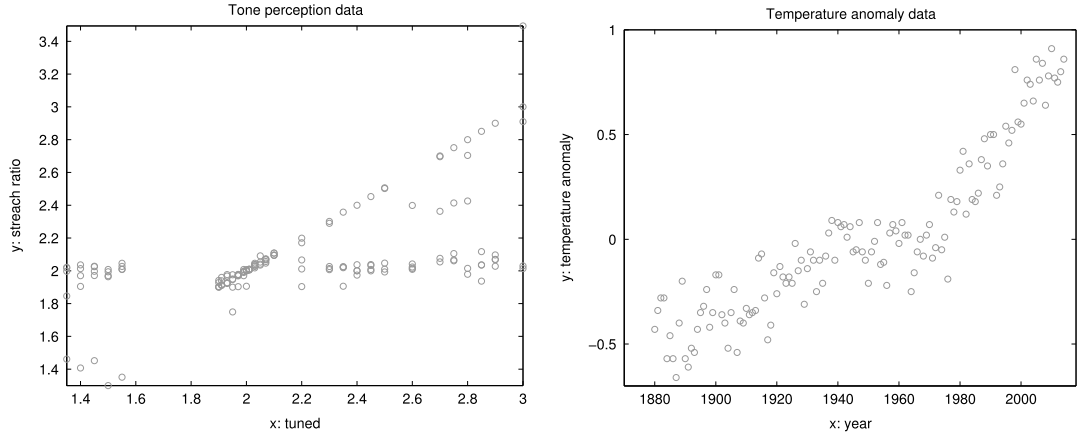

Fig. 7. Scatter plots of the tone perception data and the temperature anomalies data.

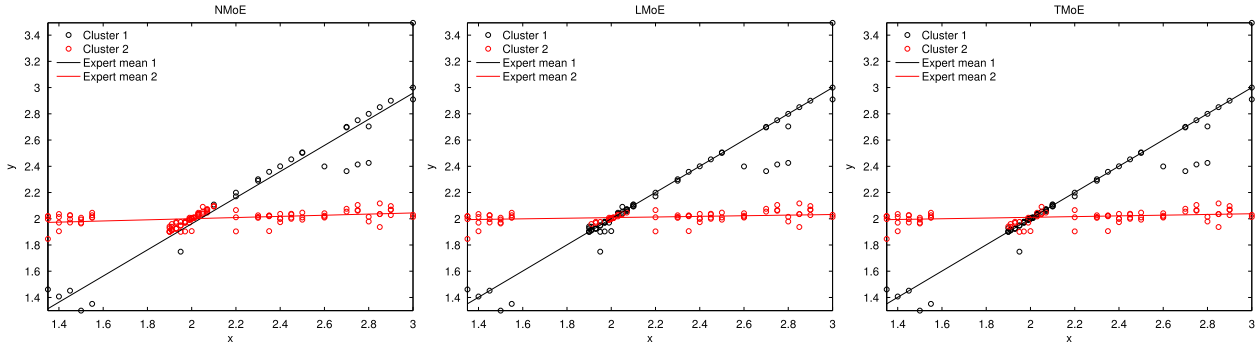

Fig. 8. The fitted MoLE to the original tone data set with left: NMoE solution, middle: LMoE solution, and right: TMoE model solution. The predictor x is the actual tone ratio and the response y is the perceived tone ratio.

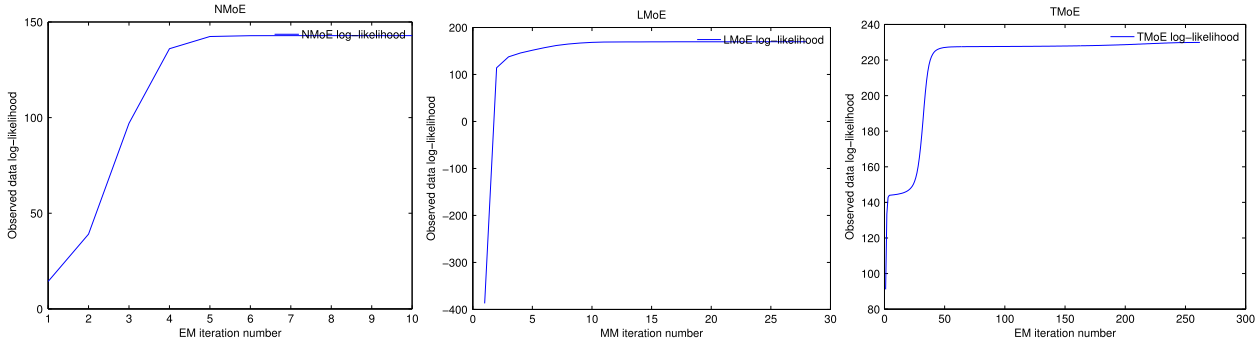

Fig. 9. The log-likelihood during the iterations when fitting the MoLE models to the original tone data set. Left: NMoE model, Middle: LMoE model, Right: TMoE model.

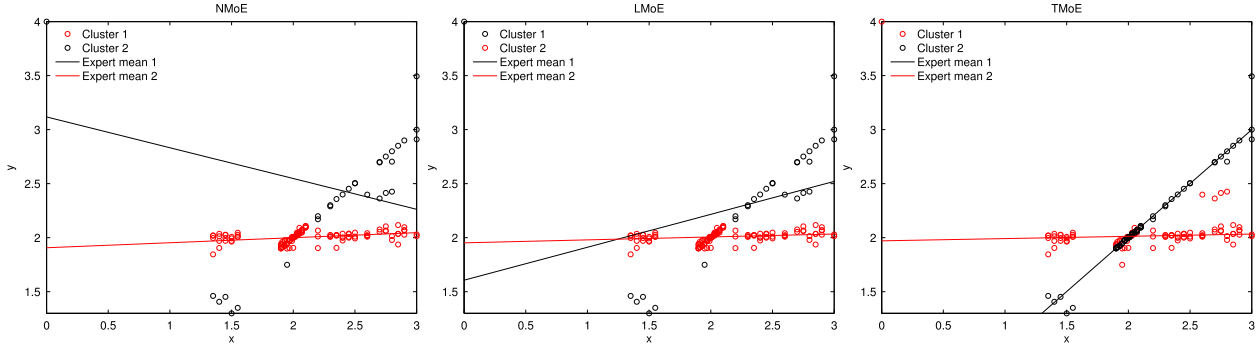

Fig. 10. Fitting MoLE to the tone data set with ten added outliers (0, 4). Left: NMoE model fit, Middle: LMoE model fit, Right: TMoE model fit. The predictor x is the actual tone ratio and the response y is the perceived tone ratio.

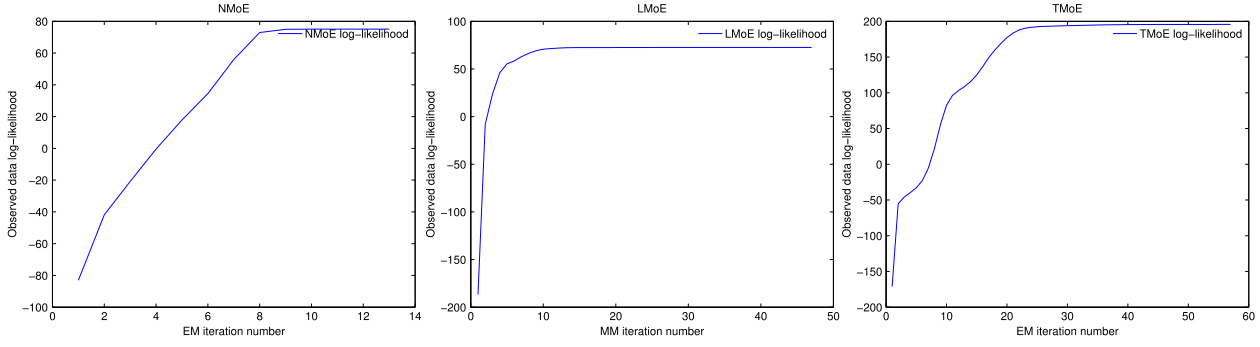

Fig. 11. The log-likelihood during the EM iterations when fitting the MoLE models to the tone data set with ten added outliers (0, 4). Left: NMoE model, Middle LMoE, and Right: TMoE model.

to the data. However, the right plot in Fig. 10 clearly shows that the TMoE provides a robust good fit, which is preferred to the LMoE solution. For the TMoE, the obtained fit is quasi-identical to the first one on the original data without outliers, shown in the right plot of Fig. 8. Moreover, we notice that, as shown in Song et al. (2014), for this situation with outliers, the t mixture of regressions fails; The fit is affected severely by the outliers. However, for the proposed TMoE model, the ten high leverage outliers have no significant impact on the fitted experts. This is because here the mixing proportions depend on the inputs, which is not the case for the regression mixture model described in Song et al. (2014).

Fig. 11 shows the log-likelihood profiles for each of the three models, which, while showing a similar behavior than the one in the case without outliers, show that the maximum likelihood value

Table 6
Values of the estimated MoE parameters for the tone perception data set with added outliers.

Model	Param.											
	α_{10}	α_{11}	β_{10}	β_{11}	β_{20}	β_{21}	σ_1	σ_2	λ_1	λ_2	ν_1	ν_2
NMoE	0.811	0.150	3.117	−0.285	1.907	0.046	0.700	0.050	–	–	–	–
LMoE	−0.557	−0.232	1.606	0.3047	1.9524	0.027	–	–	0.546	0.038	–	–
TMoE	0.888	−0.236	0.002	0.999	1.971	0.020	0.002	0.024	–	–	0.682	0.812

Table 7
Values of the estimated MoE parameters for the temperature anomalies data set.

Model	Param.											
	α_{10}	α_{11}	β_{10}	β_{11}	β_{20}	β_{21}	σ_1	σ_2	λ_1	λ_2	ν_1	ν_2
NMoE	946.483	−0.481	−12.805	0.006	−41.073	0.020	0.115	0.110	–	–	–	–
LMoE	354.076	−0.180	−13.026	0.006	−40.796	0.020	–	–	0.092	0.088	–	–
TMoE	947.225	−0.482	−12.825	0.006	−41.008	0.020	0.114	0.108	–	–	70.82	54.38

Table 8
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.

K	NMoE			LMoE			TMoE		
	BIC	AIC	ICL	BIC	AIC	ICL	BIC	AIC	ICL
1	46.0623	50.4202	46.0623	39.2617	43.6196	−7.3579	43.5521	49.3627	43.5521
2	79.9163	91.5374	79.6241	71.0153	82.6364	−19.6211	74.7960	89.3224	74.5279
3	71.3963	90.2806	58.4874	61.9639	80.8482	−31.8843	63.9709	87.2131	47.3643
4	66.7276	92.8751	54.7524	49.9480	76.0955	−44.1475	56.8410	88.7990	45.1251
5	59.5100	92.9206	51.2429	40.3062	73.7169	−56.4107	43.7767	84.4505	29.3881

for the NMoE model is significantly less than the one in the case without outliers, compared to the best solution which is provided by the TMoE model.

The values of estimated MoE parameters in this case with outliers are given in Table 6. The regression coefficients for the second expert component are very similar for the three models. For the first component, the TMoE model retrieved a more heavy tailed component. Finally, for this data set, we can conclude that the TMoE provides the best solution.

8.3.2. Temperature anomalies data set

In this experiment, we examine another real-world data set related to climate change analysis. The NASA GISS Surface Temperature (GISTEMP) analysis provides a measure of the changing global surface temperature with monthly resolution for the period since 1880, when a reasonably global distribution of meteorological stations was established. The GISS analysis is updated monthly, however the data presented here[3] are updated annually as issued from the Carbon Dioxide Information Analysis Center (CDIAC), which has served as the primary climate-change data and information analysis center of the US Department of Energy since 1982. 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. These data have been analyzed earlier by Hansen, Ruedy, Glascoe, and Sato (1999); Hansen et al. (2001) and recently by Nguyen and McLachlan (2016) by using the Laplace mixture of linear experts (LMoLE).

To apply the proposed t mixture of expert model, we consider a mixture of two experts as in Nguyen and McLachlan (2016). This number of components is also the one provided by the model selection criteria as shown later in Table 8. Indeed, as mentioned by Hansen et al. (2001), Nguyen and McLachlan (2016) found that the data could be segmented into two periods of global warming (before 1940 and after 1965), separated by a transition period where there was a slight global cooling (i.e. 1940–1965). Documentation of the basic analysis method is provided by Hansen et al. (1999) and Hansen et al. (2001). We set the response $y_i (i = 1, \ldots, 135)$ 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.

Figs. 12–14 respectively show, for each of the three compared models, the fitted linear expert components, the corresponding means and confidence regions computed as plus and minus twice the estimated (pointwise) standard deviation as presented in Section 5, and the log-likelihood profiles. One can observe that the three models are successfully applied on the data set and provide very similar results.

The values of estimated MoE parameters for the temperature anomalies data set are given in Table 7. One can see that the parameters common for the three models are quasi-identical, with a slight difference for the gating network parameters provided by the LMoE model. This slight difference results in the slight difference in the shape of the estimated mean curve. The TMoE provides high degrees of freedom, which tends to approach a normal distribution. This can also be seen on the log-likelihood profiles, which converges to almost the same value, meaning that the hypothesis of normality may be likely for this data set. On the other hand, the regression coefficients are also similar to those found by Nguyen and McLachlan (2016) who used LMoE.

We performed a model selection procedure on the temperature anomalies data set to choose the best number of MoE components from values between 1 and 5. Table 8 gives the obtained values of the used model selection criteria, that is BIC, AIC, and ICL. One can see that, except the result provided by AIC for the NMoE model which provides a high number of components, and the one provided by ICL of the LMoE model, which underestimates the number of components, all the others results provide evidence for two components in the data.

9. Conclusion and future work

In this paper, we proposed a new robust non-normal MoE model, which generalizes the standard normal MoE. It is based on

[3] From Ruedy, Sato, and Lo (2015), http://cdiac.ornl.gov/ftp/trends/temp/hansen/gl_land.txt.

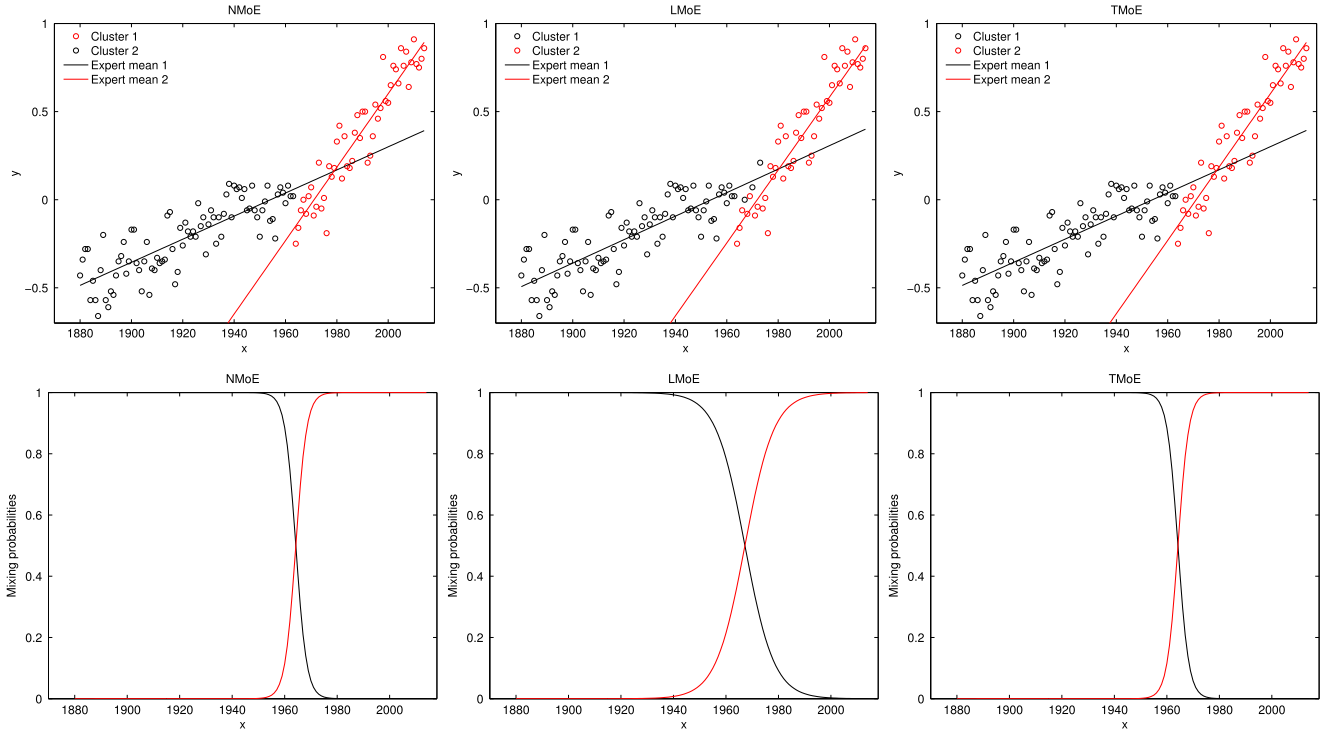

Fig. 12. Fitting the MoLE models to the temperature anomalies data set. Left: NMoE model fit; Middle: LMoE model; Right: TMoE model. The predictor x is the year and the response y is the temperature anomaly.

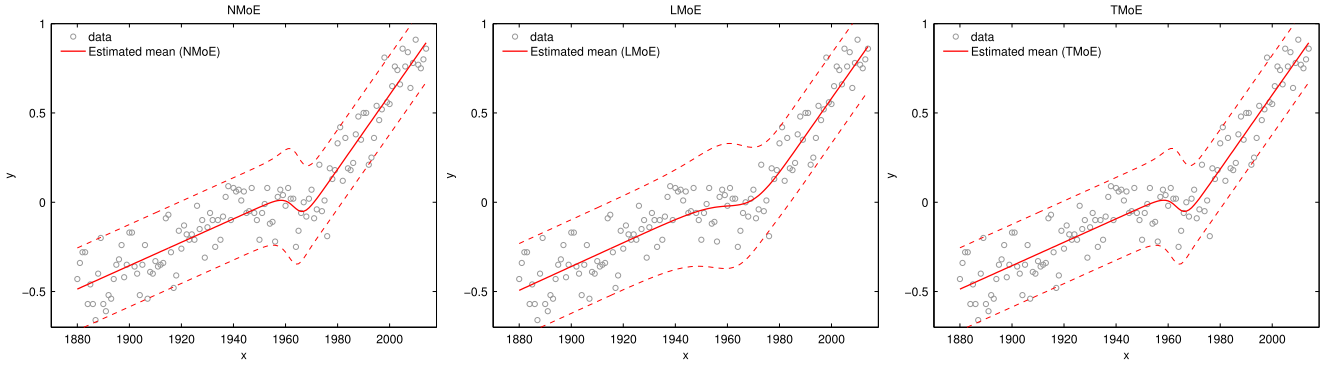

Fig. 13. The fitted MoLE models to the temperature anomalies data set. Left: NMoE model; Middle: LMoE; Right: TMoE model. The predictor x is the year and the response y is the temperature anomaly. The shaded region represents plus and minus twice the estimated (pointwise) standard deviation as presented in Section 5.

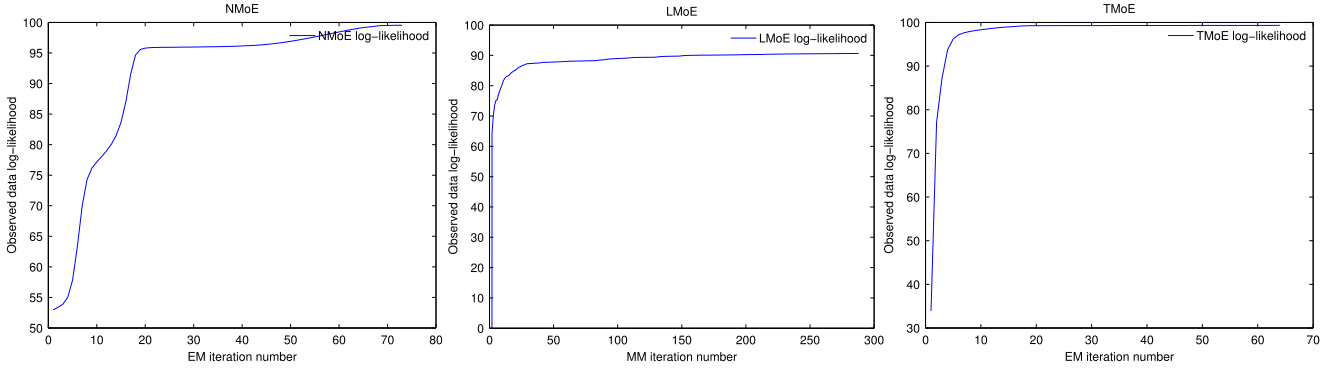

Fig. 14. The log-likelihood during the EM iterations when fitting the MoLE models to the temperature anomalies data set. Left: NMoE model; Middle: LMoE; Right: TMoE model.

the t distribution and named TMoE. The TMoE model is suggested for data with possibly outliers and heavy tail. We developed an EM algorithm and ECM extension to infer the proposed model and described its use in non-linear regression and prediction, as well as in model-based clustering. The developed model is successfully applied and validated on simulated and real data sets. The results

obtained on simulated data confirm the good performance of the model in terms of density estimation, non-linear regression function approximation and clustering. In addition, the simulation results provide evidence of the robustness of the TMoE model to outliers, compared to the normal alternative model. The proposed model is also successfully applied to two different real data sets, including a situation with outliers. The model selection using information criteria tends to promote using BIC and also ICL against AIC which performed poorly in the analyzed data. The obtained results support the benefit of the proposed approach for practical applications. Furthermore, compared to the LMoE model, the TMoE has been revealed to be more adapted in several situations.

In this paper, we only considered the MoE in their standard (non-hierarchical) version. One interesting future direction is therefore to extend the proposed models to the hierarchical MoE framework (Jordan & 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.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Bai, X., Yao, W., & Boyer, J. E. (2012). Robust fitting of mixture regression models. *Computational Statistics & Data Analysis*, *56*(7), 2347–2359.

Biernacki, C., Celeux, G., & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(7), 719–725.

Bishop, C., & Svensén, M. (2003). Bayesian hierarchical mixtures of experts. In *Uncertainty in artificial intelligence*.

Brent, R. P. (1973). *Prentice-Hall series in automatic computation*, *Algorithms for minimization without derivatives*. Englewood Cliffs, N.J: Prentice-Hall.

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

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

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

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

Faria, S., & Soromenho, G. (2010). Fitting mixtures of linear regressions. *Journal of Statistical Computation and Simulation*, *80*(2), 201–225.

Frühwirth-Schnatter, S. (2006). *Springer series in statistics*, *Finite mixture and Markov switching models*. New York: Springer Verlag.

Frühwirth-Schnatter, S., & Pyne, S. (2010). Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-*t* distributions. *Biostatistics*, *11*(2), 317–336.

Gaffney, S., & Smyth, P. (1999). Trajectory clustering with mixtures of regression models. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 63–72). ACM Press.

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

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

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

Hunter, D., & Young, D. (2012). Semiparametric mixtures of regressions. *Journal of Nonparametric Statistics*, *24*(1), 19–38.

Ingrassia, S., Minotti, S., & Vittadini, G. (2012). Local statistical modeling via a cluster-weighted approach with elliptical distributions. *Journal of Classification*, *29*(3), 363–401.

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

Jiang, W., & Tanner, M. A. (1999a). On the asymptotic normality of hierarchical mixtures-of-experts for generalized linear models. *IEEE Transactions on Information Theory*, *46*, 1005–1013.

Jiang, W., & Tanner, M. A. (1999b). On the identifiability of mixtures-of-experts. *Neural Networks*, *12*, 197–220.

Jones, P. N., & McLachlan, G. J. (1992). Fitting finite mixture models in a regression context. *Australian Journal of Statistics*, *34*(2), 233–240.

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

Jordan, M. I., & Xu, L. (1995). Convergence results for the EM approach to mixtures of experts architectures. *Neural Networks*, *8*(9), 1409–1431.

Kent, J., Tyler, D., & Vardi, Y. (1994). A curious likelihood identity for the multivariate *t*-distribution. *Communications in Statistics—Simulation and Computation*, *23*, 441–453.

Lin, T. I., Lee, J. C., & Hsieh, W. J. (2007). Robust mixture modeling using the skew *t* distribution. *Statistics and Computing*, *17*(2), 81–92.

Liu, C., & Rubin, D. B. (1995). ML estimation of the *t* distribution using EM and its extensions, ECM and ECME. *Statistica Sinica*, *5*, 19–39.

McLachlan, G. J., & Krishnan, T. (2008). *The EM algorithm and extensions* (2nd ed.). New York: Wiley.

Mclachlan, G. J., & Peel, D. (1998). *Lecture notes in computer science*, *Robust cluster analysis via mixtures of multivariate t-distributions* (pp. 658–666). Springer-Verlag.

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

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

Ng, S.-K., & McLachlan, G. J. (2004). Using the em algorithm to train neural networks: misconceptions and a new algorithm for multiclass classification. *IEEE Transactions on Neural Networks*, *15*(3), 738–749.

Nguyen, H. D., & McLachlan, G. J. (2016). Laplace mixture of linear experts. *Computational Statistics & Data Analysis*, *93*, 177–191.

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

Quandt, R. E. (1972). A new approach to estimating switching regressions. *Journal of the American Statistical Association*, *67*(338), 306–310.

Quandt, R. E., & Ramsey, J. B. (1978). Esimating mixtures of normal distributions and switching regressions. *Journal of the American Statistical Association*, *73*(364), 730–738.

Rasmussen, C. E., & Ghahramani, Z. (2001). Infinite mixtures of Gaussian process experts. In *Advances in neural information processing systems, Vol. 14* (pp. 881–888). MIT Press.

Ruedy, R., Sato, M., & Lo, K. (2015). NASA GISS surface temperature (GISTEMP) analysis. http://dx.doi.org/10.3334/CDIAC/cli.001. Center for Climate Systems Research, NASA Goddard Institute for Space Studies 2880 Broadway, New York, NY 10025 USA.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Shi, J. Q., & Choi, T. (2011). *Gaussian process regression analysis for functional data*. Chapman & Hall/CRC Press.

Shi, J. Q., Murray-Smith, R., & Titterington, D. M. (2005). Hierarchical Gaussian process mixtures for regression. *Statistics and Computing*, *15*(1), 31–41.

Song, W., Yao, W., & Xing, Y. (2014). Robust mixture regression model fitting by laplace distribution. *Computational Statistics & Data Analysis*, *71*(0), 128–137.

Veaux, R. D. D. (1989). Mixtures of linear regressions. *Computational Statistics and Data Analysis*, *8*(3), 227–245.

Viele, K., & Tong, B. (2002). Modeling with mixtures of linear regressions. *Statistics and Computing*, *12*, 315–330.

Waterhouse, S. R. (1997). *Classification and regression using mixtures of experts*. (Ph.D. thesis), Department of Engineering, Cambridge University.

Waterhouse, S., Mackay, D., & Robinson, T. (1996). *Bayesian methods for mixtures of experts*. MIT Press, pp. 351–357.

Wei, Y. (2012). *Robust mixture regression models using t-distribution. Tech. rep., Master Report*. Department of Statistics, Kansas State University.

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

Yuksel, S. E., Wilson, J. N., & Gader, P. D. (2012). Twenty years of mixture of experts. *IEEE Transactions on Neural Networks and Learning Systems*, *23*(8), 1177–1193.

Faicel Chamroukhi received his Master degree of Engineering Sciences, in the area of signals, images and robotics from Pierre & Marie Curie (Paris 6) University in 2007. Then, he received his Ph.D. degree in applied mathematics and computer science, in the area of statistical leaning and data analysis from Compiègne University of Technology in 2010. In 2011, he was qualified for the position of Associate Professor in applied mathematics (CNU 26), computer science (CNU 27), and signal processing (CNU 61). Since september 2011, he is an Associate Professor at University of Toulon and the Information Sciences and Systems Lab (LSIS) UMR CNRS 7296. In 2015, he received his Accreditation to Supervise Research (HDR) in applied mathematics and computer science, in the area of statistical learning and data analysis, from Toulon University. Since 2016 he is qualified for the position of Professor in the three area of applied mathematics, computer science, and signal processing (CNU 26, 27, 61). In 2015, he was awarded a CNRS research leave and since september he moved to the Lab of mathematics Paul Painlevé (LPP) UMR CNRS 7296, probability and statistics team, in Lille, where he is also invited at INRIA - Modal team. His multidisciplinary research is in the area of Data Science and includes statistics, machine learning and statistical signal processing, with a particular focus of the statistical methodology and inference of latent data models for complex heterogeneous high-dimensional and massive data, temporal data, functional data, and their application to real-world problems including dynamical systems, acoustic/speech processing, life sciences (medicine, biology), information retrieval, social networks.