Regularized Estimation and Feature Selection in Mixtures of Gaussian-Gated Experts Models

Faïcel Chamroukhi[1], Florian Lecocq[2], and Hien D. Nguyen[3]

[1] Department of Mathematics, University of Queensland
Brisbane, 4072 Queensland, Australia
[2] University of Caen, Laboratory of Mathematics Nicolas Oresme
LMNO - UMR CNRS, Unicaen Campus 2, 14000 Caen, France
[3] Department of Mathematics and Statistics, La Trobe University
Melbourne Victoria 3086, Australia

Abstract. Mixtures-of-Experts models and their maximum likelihood estimation (MLE) via the EM algorithm have been thoroughly studied in the statistics and machine learning literature. They are subject of a growing investigation in the context of modeling with high-dimensional predictors with regularized MLE. We examine MoE with Gaussian gating network, for clustering and regression, and propose an ℓ_1-regularized MLE to encourage sparse models and deal with the high-dimensional setting. We develop an EM-Lasso algorithm to perform parameter estimation and utilize a BIC-like criterion to select the model parameters, including the sparsity tuning hyperparameters. Experiments conducted on simulated data show the good performance of the proposed regularized MLE compared to the standard MLE with the EM algorithm.

Keywords: Mixtures-of-Experts · Clustering · Feature selection · EM algorithm · Lasso · High-dimensional data.

1 Introduction

Mixture-of-experts (MoE), originally introduced in [12,13], form a class of conditional mixture models [16] for modeling, clustering and prediction in the presence of heterogeneous data. Their construction rely on conditional mixture models [16] in which both the gating network, formed by the mixing proportions, and the experts network formed by the mixture components, depend on the predictors or the inputs. The most popular choices for the gating network are the softmax gating functions [12] or the Gaussian gating functions; the latter is a particular case of the exponential family gating functions introduced in [24].

Different choices are now common for the expert network model, depending on the type of the observed responses. For instance, a model for normal observations for regression and clustering was introduced in [5] or non-normally distributed expert models like in [1] to deal with skewed data distributions [3], to ensure robustness to outliers [2,19], or to accommodate both skewness and robustness as in [4]. A detailed review on MoE models can be found in [17]

Fitting MoE is generally performed by maximum-likelihood estimation (MLE) via the EM algorithm or its variants [8,15]. In a high-dimensional setting, the regularization of the MLE, to perform parameter estimation under a sparsity hypothesis and hence to simultaneously perform feature selection, has been studied in [14] and more recently in [7,11]. These approaches consider ℓ_1 and ℓ_2 penalties for the log-likelihood function, and are constructed upon softmax gating functions.

In this paper, we consider MoE with Gaussian gated functions, and propose an ℓ_1-regularized MLE via and EM-Lasso algorithm. We study the performance of the proposal on an experimental setup. The remainder of this paper is organized as follows. Section 2 describes the MoE modeling framework, and the Gaussian-gated MoE and its MLE with the EM algorithm. Then, Section 3 presents the proposed regularized MLE and the EM-Lasso algorithm. Finally, Section 4 is dedicated to numerical experiments.

2 Gaussian-Gated Mixture-of-Experts

2.1 MoE modeling framework

We consider mixtures-of-experts model to relate a high-dimensional predictor $\boldsymbol{X} \in \mathbb{R}^p$ to a response $\boldsymbol{Y} \in \mathbb{R}^d$, potentially multivariate $d \geq 1$. We assume that the pair $(\boldsymbol{X}, \boldsymbol{Y})$ is generated from a heterogeneous population governed by a hidden structure represented by a latent categorical variable $Z \in [K] = \{1, \ldots, K\}$. Assume that we observe a random sample $\{(\boldsymbol{X}_i, \boldsymbol{Y}_i)\}_{i=1,\ldots,n}$ of n independently and identically distributed (i.i.d) pairs $(\boldsymbol{X}_i, \boldsymbol{Y}_i)$ from $(\boldsymbol{X}, \boldsymbol{Y})$, and let $\mathcal{D} = ((\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n))$ be an observed data sample. Assume that the pair $(\boldsymbol{X}, \boldsymbol{Y})$ follows a MoE distribution, then the MoE model can be defined as

$$f(\boldsymbol{y}_i|\boldsymbol{x}_i; \boldsymbol{\Psi}) = \sum_{k=1}^{K} g_k(\boldsymbol{x}_i; \mathbf{w}) f(\boldsymbol{y}_i|\boldsymbol{x}_i; \boldsymbol{\theta}_k) \tag{1}$$

where $g_k(\boldsymbol{x}; \mathbf{w}) = \mathbb{P}(Z = k|\boldsymbol{X} = \boldsymbol{x}; \mathbf{w})$ is the distribution of the hidden variable Z given the predictor $\boldsymbol{x}$ with parameters $\mathbf{w}$, which represents the gating network, and the conditional component densities $f(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\theta}_k) = f(\boldsymbol{y}_i|\boldsymbol{X} = \boldsymbol{x}, Z = k; \boldsymbol{\theta})$ represent the experts network whose parameters are $\boldsymbol{\theta}_k$.

2.2 Gaussian-Gated Mixture-of-Experts

Let us define by $\phi_m(\boldsymbol{v}; \boldsymbol{m}, \mathbf{C}) = (2\pi)^{-m/2}|\mathbf{C}|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{v} - \boldsymbol{m})^\top \mathbf{C}^{-1}(\boldsymbol{v} - \boldsymbol{m})\right)$ the probability density function of a Gaussian random vector $\boldsymbol{V}$ of dimension m with mean $\boldsymbol{m}$ and covariance matrix $\mathbf{C}$. We consider mixture-of-experts for clustering and regression of heterogeneous data. In this case, the mixture of Gaussian-gated experts models, we abbreviate as MoGGE, for multivariate real responses, is defined by (1) where the experts are (multivariate) Gaussian regressions, given by

$$f(\boldsymbol{y}_i|\boldsymbol{x}_i; \boldsymbol{\theta}_k) = \phi_d(\boldsymbol{y}_i; \mathbf{a}_k + \mathbf{B}_k^T \boldsymbol{x}_i, \boldsymbol{\Sigma}_k) \tag{2}$$

and the gating network $g(\boldsymbol{x}_i; \mathbf{w})$ is defined by Gaussian gating function of the form:

$$g_k(\boldsymbol{x}_i; \mathbf{w}) = \frac{\mathbb{P}(Z_i = k) f(\boldsymbol{x}_i | Z_i = k; \boldsymbol{w}_k)}{\sum_{\ell=1}^{K} \mathbb{P}(Z_i = \ell) f(\boldsymbol{x}_i | Z_i = \ell; \boldsymbol{w}_\ell)} = \frac{\alpha_k \phi_p(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \mathbf{R}_k)}{\sum_{\ell=1}^{K} \alpha_\ell \phi_p(\boldsymbol{x}_i; \boldsymbol{\mu}_\ell, \mathbf{R}_\ell)} \tag{3}$$

with $\mathbb{P}(Z_i = k) = \alpha_k$, $f(\boldsymbol{x}_i | Z_i = k; \mathbf{w}) = \phi_p(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \mathbf{R}_k)$ $(k = 1, \ldots, K)$. This Gaussian gating network was introduced in [24] to sidestep the need for a non-linear optimization routine in the inner loop of the EM algorithm in the case of a softmax function for the gating network. The MoGGE model is thus parameterized by the parameter vector $\boldsymbol{\Psi} = (\mathbf{w}^T, \boldsymbol{\theta}^T)^T$ where $\mathbf{w} = (\boldsymbol{w}_1^T, \ldots, \boldsymbol{w}_K^T)^T$ is the parameter vector of the gating network and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_K^T)^T$ is the parameter vector of experts network, with $\boldsymbol{w}_k = (\alpha_k, \boldsymbol{\mu}_k^T, \text{vech}(\mathbf{R}_k)^T)^T$ and $\boldsymbol{\theta}_k = (\mathbf{a}_k^T, \mathbf{B}_k^T, \text{vech}(\boldsymbol{\Sigma}_k)^T)^T$ for $k = 1, \ldots, K$. The approximation capabilities of this model have been studied very recently in [18].

2.3 Maximum likelihood estimation via the EM algorithm

Mixtures-of-experts of the form (1) with softmax gating functions are in general estimated by maximizing the (conditional) log-likelihood $\sum_{i=1}^{n} \log f(\boldsymbol{y}_i | \boldsymbol{x}_i; \boldsymbol{\Psi})$ by using the EM algorithm , in which the M-step requires an internal iterative numerical optimization procedure (eg. a Newton-Raphson algorithm) to update the softmax parameters. We follow the approach of estimating MoGGE in [24], which relies on maximizing the joint loglikelihood, and in the MLE, the M-Step can then be solved in a closed form. Indeed, based on equations (1), (2), and (3), we the MoGGE conditional density is given by:

$$f(\boldsymbol{y}_i | \boldsymbol{x}_i; \boldsymbol{\Psi}) = \sum_{k=1}^{K} \frac{\alpha_k \phi_p(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \mathbf{R}_k)}{\sum_{\ell=1}^{K} \alpha_\ell \phi_p(\boldsymbol{x}_i; \boldsymbol{\mu}_\ell, \mathbf{R}_\ell)} \phi_d(\boldsymbol{y}_i; \mathbf{a}_k + \mathbf{B}_k^T \boldsymbol{x}_i, \boldsymbol{\Sigma}_k) \cdot \tag{4}$$

Then we can write the joint density as:

$$\begin{aligned} f(\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\Psi}) = f(\boldsymbol{x}_i; \mathbf{w}) f(\boldsymbol{y}_i | \boldsymbol{x}_i; \boldsymbol{\theta}) &= \sum_{k=1}^{K} \mathbb{P}(Z_i = k) f(\boldsymbol{x}_i; \boldsymbol{w}_k) f(\boldsymbol{y}_i | \boldsymbol{x}_i; \boldsymbol{\theta}_k) \\ &= \sum_{k=1}^{K} \alpha_k \phi_p(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \mathbf{R}_k) \phi_d(\boldsymbol{y}_i; \mathbf{a}_k + \mathbf{B}_k^T \boldsymbol{x}_i, \boldsymbol{\Sigma}_k) \cdot \end{aligned} \tag{5}$$

The joint log-likelihood to be maximized by EM is therefore given by:

$$L(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log f(\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \alpha_k \phi_p(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \mathbf{R}_k) \phi_d(\boldsymbol{y}_i; \mathbf{a}_k + \mathbf{B}_k^T \boldsymbol{x}_i, \boldsymbol{\Sigma}_k) \cdot \tag{6}$$

2.4 The EM algorithm for the MoGGE model

The complete-data log-likelihood upon which the EM principle is constructed is then defined by

$$L_c(\boldsymbol{\Psi}) = \sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik} \log \left[\alpha_k \phi_p(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \mathbf{R}_k)\phi_d(\boldsymbol{y}_i; \mathbf{a}_k + \mathbf{B}_k^T \boldsymbol{x}_i, \boldsymbol{\Sigma}_k)\right] \quad (7)$$

where Z_{ik} being an indicator binary-valued variable such that $Z_{ik} = 1$ if $Z_i = k$ (i.e., if the ith pair $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ is generated from the kth expert and $Z_{ik} = 0$ otherwise. The EM algorithm, after starting with an initial solution $\boldsymbol{\Psi}^{(0)}$, alternates between the E- and the M- Steps until convergence (when there is no longer a significant change in the log-likelihood (6)).

E-step: Compute the expectation of the complete-data log-likelihood (7), given the observed data $\mathcal{D}$ and the current parameter vector estimate $\boldsymbol{\Psi}^{(q)}$:

$$\begin{aligned} Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(q)}) &= \mathbb{E}\left[L_c(\boldsymbol{\Psi})|\mathcal{D}; \boldsymbol{\Psi}^{(q)}\right] \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}^{(q)} \log \left[\alpha_k \phi_p(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \mathbf{R}_k)\phi_d(\boldsymbol{y}_i; \mathbf{a}_k + \mathbf{B}_k^T \boldsymbol{x}_i, \boldsymbol{\Sigma}_k)\right], \quad (8) \end{aligned}$$

where:

$$\tau_{ik}^{(q)} = \mathbb{P}(Z_i = k|\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\Psi}^{(q)}) = \frac{\alpha_k^{(q)} \phi_p(\boldsymbol{x}_i; \boldsymbol{\mu}_k^{(q)}, \mathbf{R}_k^{(q)})\phi_d(\boldsymbol{y}_i; \mathbf{a}_k^{(q)} + \mathbf{B}_k^{(q)^T} \boldsymbol{x}_i, \boldsymbol{\Sigma}_k^{(q)})}{f(\boldsymbol{x}_i, \boldsymbol{y}_i; \boldsymbol{\Psi}^{(q)})}, \quad (9)$$

is the posterior probability that the observed pair $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ is generated by the kth expert. This step therefore only requires the computation of the posterior component membership probabilities $\tau_{ik}^{(q)}$ $(i = 1, \ldots, n)$, for $k = 1, \ldots, K$.

M-step: Calculate the parameter vector update $\boldsymbol{\Psi}^{(q+1)}$ by maximizing the Q-function (8), i.e, $\boldsymbol{\Psi}^{(q+1)} = \arg\max_{\boldsymbol{\Psi}} Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(q)})$. By decomposing the $Q-$function (8) as

$$Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(q)}) = \sum_{k=1}^{K} Q(\boldsymbol{w}_k; \boldsymbol{\Psi}^{(q)}) + Q(\boldsymbol{\theta}_k; \boldsymbol{\Psi}^{(q)}) \quad (10)$$

where

$$Q(\boldsymbol{w}_k; \boldsymbol{\Psi}^{(q)}) = \sum_{i=1}^{n} \tau_{ik}^{(q)} \log \left[\alpha_k \phi_p(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \mathbf{R}_k)\right] \quad (11)$$

and

$$Q(\boldsymbol{\theta}_k; \boldsymbol{\Psi}^{(q)}) = \sum_{i=1}^{n} \tau_{ik}^{(q)} \log \phi_d(\boldsymbol{y}_i; \mathbf{a}_k + \mathbf{B}_k^T \boldsymbol{x}_i, \boldsymbol{\Sigma}_k), \quad (12)$$

the maximization can then be done by performing K separate maximizations w.r.t the gating network parameters and the experts network parameters.

Updating the the gating networks' parameters: Maximizing (11) w.r.t $\boldsymbol{w}_k$'s corresponds to the M-Step of a Gaussian Mixture Model [16]. The closed-form expressions for updating the parameters are given by:

$$\alpha_k^{(q+1)} = \sum_{i=1}^{n} \tau_{ik}^{(q)} \Big/ n, \tag{13}$$

$$\boldsymbol{\mu}_k^{(q+1)} = \sum_{i=1}^{n} \tau_{ik}^{(q)} \boldsymbol{x}_i \Big/ \sum_{i=1}^{n} \tau_{ik}^{(q)}, \tag{14}$$

$$\mathbf{R}_k^{(q+1)} = \sum_{i=1}^{n} \tau_{ik}^{(q)} (\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(q+1)})(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(q+1)})^T \Big/ \sum_{i=1}^{n} \tau_{ik}^{(q)}. \tag{15}$$

Updating the experts' network parameters Maximizing (12) w.r.t $\boldsymbol{\theta}_k$'s corresponds to the M-Step of standard MoE with multivariate Gaussian regression experts, see e.g [6]. The closed-form updating formulas are given by:

$$\mathbf{a}_k^{(q+1)} = \sum_{i=1}^{n} \tau_{ik}^{(q)} (\boldsymbol{y}_i - \mathbf{B}_k^{(q)^T} \boldsymbol{x}_i) \Big/ \sum_{i=1}^{n} \tau_{ik}^{(q)}, \tag{16}$$

$$\mathbf{B}_k^{(q+1)} = \Big[\sum_{i=1}^{n} \tau_{ik}^{(q)} \boldsymbol{x}_i \boldsymbol{x}_i^T \Big]^{-1} \sum_{i=1}^{n} \tau_{ik}^{(q)} \boldsymbol{x}_i (\boldsymbol{y}_i - \mathbf{a}_k^{(q+1)})^T, \tag{17}$$

$$\boldsymbol{\Sigma}_k^{(q+1)} = \sum_{i=1}^{n} \tau_{ik}^{(q)} (\boldsymbol{y}_i - (\mathbf{a}_k^{(q+1)} + \mathbf{B}_k^{(q+1)^T} \boldsymbol{x}_i))(\boldsymbol{y}_i - (\mathbf{a}_k^{(q+1)} + \mathbf{B}_k^{(q+1)^T} \boldsymbol{x}_i))^T \Big/ \sum_{i=1}^{n} \tau_{ik}^{(q)}. \tag{18}$$

However, in a high dimensional setting, MLE may be unstable or even unfeasible. One possible way to proceed in such a context is the regularization of the objective function. In the context of MoE models, this has been studied namely in [14,7,11] where ℓ_1 and ℓ_2 regularization for the log-likelihood function of the standard MoE model with softmax gating network. This penalized MLE allow an efficient estimation for simultaneous parameter estimation and feature selection.

3 Penalized maximum likelihood parameter estimation

Here we study the regularized estimation of the MoGGE model. We first consider the case when $d = 1$ (univariate response $\boldsymbol{y}_i$). The expert densities are thus defined by $f(y_i|\boldsymbol{x}_i; \boldsymbol{\theta}_k) = \phi(y_i; \beta_{k,0} + \boldsymbol{\beta}_k^T \boldsymbol{x}_i, \sigma_k^2)$ with $\boldsymbol{\theta}_k = (\beta_{k,0}, \boldsymbol{\beta}_k^T, \sigma_k^2)^T$.

In our proposed approach, rather than maximizing the joint log-likelihood (6), we attempt to maximize its ℓ_1-regularized version, to encourage sparse models and to perform estimation and feature selection. The resulting penalized log-likelihood can then be defined by:

$$\mathcal{L}(\boldsymbol{\Psi}) = L(\boldsymbol{\Psi}) - \mathrm{Pen}_{\lambda,\gamma}(\boldsymbol{\Psi}) \tag{19}$$

where $L(\boldsymbol{\Psi})$ is the observed-data log-likelihood of $\boldsymbol{\Psi}$ defined by (6) and $\mathrm{Pen}_{\lambda,\gamma}(\boldsymbol{\Psi})$ is a Lasso [22] regularization term encouraging sparsity for the expert network

parameters and the gating network parameters, with λ and γ positive real values representing tuning hyperparameters. For regularizing the expert parameters, the penalty is naturally applied to the regression coefficient vectors $\boldsymbol{\beta}_k$. For the gating network, since the estimates are those of a Gaussian mixture, we then follow the strategy of feature selection in model-based clustering in [20] in which we apply the penalty to the Gaussian mean vectors $\boldsymbol{\mu}_k$ and assume that the Gaussian covariance matrices of the gating network are diagonal, ie. $\mathbf{R}_k = \text{diag}(\nu_1^2, \ldots, \nu_K^2)$. The penalty function is then given by:

$$\text{Pen}_{\lambda,\gamma}(\boldsymbol{\Psi}) = \lambda \sum_{k=1}^{K} \|\boldsymbol{\beta}_k\|_1 + \gamma \sum_{k=1}^{K} \|\boldsymbol{\mu}_k\|_1. \tag{20}$$

We now derive an EM-Lasso algorithm to maximize (19).

3.1 The EM-Lasso algorithm for the MoGGE model

Lets first define the penalized joint complete-data log-likelihood, which is given by

$$\mathcal{L}_c(\boldsymbol{\Psi}) = L_c(\boldsymbol{\Psi}) - \text{Pen}_{\lambda,\gamma}(\boldsymbol{\Psi}) \tag{21}$$

where $L_c(\boldsymbol{\Psi})$ is the non-regularized joint complete-data log-likelihood defined by (7). The EM-Lasso algorithm then alternates between the two following steps until convergence (when there is no significant change in (19).

E-step. This step computes the expectation of the complete-data log-likelihood (21), given the observed data $\mathcal{D}$, using the current parameter vector $\boldsymbol{\Psi}^{(q)}$:

$$\mathcal{Q}_{\lambda,\gamma}(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(q)}) = \mathbb{E}\left[\mathcal{L}_c(\boldsymbol{\Psi}) | \mathcal{D}; \boldsymbol{\Psi}^{(q)}\right] = Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(q)}) - \text{Pen}_{\lambda,\gamma}(\boldsymbol{\Psi}) \tag{22}$$

which only requires the computation of the posterior probabilities of component membership $\tau_{ik}^{(q)}$ $(i = 1, \ldots, n)$, for each of the K experts as defined by (9).

M-step. This step updates the value of the parameter vector $\boldsymbol{\Psi}$ by maximizing the Q-function (8) with respect to $\boldsymbol{\Psi}$, that is, by computing the parameter vector update $\boldsymbol{\Psi}^{(q+1)} = \arg\max_{\boldsymbol{\Psi}} \mathcal{Q}_{\lambda,\gamma}(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(q)})$. Now we have this decomposition

$$\mathcal{Q}_{\lambda,\gamma}(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(q)}) = \sum_{k=1}^{K} \mathcal{Q}_{\gamma}(\boldsymbol{w}_k; \boldsymbol{\Psi}^{(q)}) + \mathcal{Q}_{\lambda}(\boldsymbol{\Psi}_k; \boldsymbol{\Psi}^{(q)}) \tag{23}$$

and the maximization is performed by K separate maximizations of the penalized Q-functions $\mathcal{Q}_{\gamma}(\boldsymbol{w}_k; \boldsymbol{\Psi}^{(q)})$ and $\mathcal{Q}_{\lambda}(\boldsymbol{\Psi}_k; \boldsymbol{\Psi}^{(q)})$.

Coordinate Ascent for updating the gating network Updating the gating network parameters consists of maximizing w.r.t $\boldsymbol{w}_k$ the following penalized Q-function

$$\begin{aligned}\mathcal{Q}_\gamma(\boldsymbol{w}_k;\boldsymbol{\Psi}) &= \sum_{i=1}^{n}\tau_{ik}^{(q)}\log\left[\alpha_k\phi_p(\boldsymbol{x}_i;\boldsymbol{\mu}_k,\mathbf{R}_k)\right]-\gamma\sum_{j=1}^{p}|\mu_{k,j}| \\ &= \sum_{i=1}^{n}\tau_{ik}^{(q)}\log\alpha_k+\sum_{i=1}^{n}\tau_{ik}^{(q)}\log\phi_p(\boldsymbol{x}_i;\boldsymbol{\mu}_k,\mathbf{R}_k)-\gamma\sum_{j=1}^{p}|\mu_{k,j}|.\end{aligned}$$

It can be seen that the updates of the α_k's are unchanged compared to the standard algorithm and are given by (13). For the mean vectors, updating the coefficients $\mu_{k,j}$ corresponds to weighted version or and ℓ_1-regularized maximum likelihood estimation a Gaussian mean; The coefficients $\mu_{k,j}$ can then be updated in a cyclic way by using a Coordinate ascent algorithm until (24) is maximized. Coordinate ascent (CA) [10, sec. 5.4] [9,23] is indeed an efficient way to solve Lasso-regularization problems. For each coefficient index $j = 1,\ldots,p$, it can be easily shown that, after starting with the previous EM-Lasso estimate as initial value, i.e, $\mu_{kj}^{(0,q)} = \mu_{kj}^{(q)}$, each iteration t of the CA algorithm updates are given by the following updating formulas (see eg. [20]), written in a scalar and a vector form:

$$\begin{aligned}\mu_{kj}^{(t+1,q)} &= \text{sign}(\tilde{\mu}_{kj}^{(q+1)})\left(|\tilde{\mu}_{kj}^{(q+1)}|-\frac{\gamma}{\sum_{i=1}^{n}\tau_{ik}^{(q)}}\nu_{kj}^{2(q)}\right)_+ \\ &= \mathcal{S}\left(\sum_{i=1}^{n}\tau_{ik}^{(q)}x_{ij};\gamma\nu_{kj}^{2(q)}\right)\Big/\sum_{i=1}^{n}\tau_{ik}^{(q)} \\ &= \mathcal{S}\left(\mathbf{X}_j^T\boldsymbol{\tau}_k^{(q)};\gamma\nu_{kj}^{2(q)}\right)/\mathbf{1}_n^T\boldsymbol{\tau}_k^{(q)} \qquad (24)\end{aligned}$$

with, $\tilde{\mu}_{kj}^{(q+1)} = \sum_{i=1}^{n}\tau_{ik}^{(q)}x_{ij}\Big/\sum_{i=1}^{n}\tau_{ik}^{(q)}$ is the usual non-regularized MLE update for μ_k (Eq. (14)), $\mathbf{X}_j$ the jth column of $\mathbf{X}$, $\mathbf{1}_n$ is a vector of ones of size n, $\boldsymbol{\tau}_k^{(q)} = (\tau_{1k}^{(q)},\ldots,\tau_{nk}^{(q)})^T$, and $\mathcal{S}(u;\eta) := \text{sign}(u)(|u|-\eta)_+$ is the soft-thresholding operator with $(.)_+ = \max\{.,0\}$. The CA procedure is iterated until no significant change in (24) is observed. We then take the update at convergence of the CA algorithm, i.e $\mu_{kj}^{(q+1)} = \mu_{kj}^{(t+1,q)}$. Finally, the updates of the diagonal elements of the co-variance matrices are given by:

$$\nu_{kj}^{2(q+1)} = \sum_{i=1}^{n}\tau_{ik}^{(q)}(x_{ij}-\mu_{kj}^{(q+1)})^2\Big/\sum_{i=1}^{n}\tau_{ik}^{(q)}. \qquad (25)$$

Coordinate Ascent for updating the experts network The maximization step for updating the expert parameters $\boldsymbol{\theta}_k$ consists of maximizing the function

$\mathcal{Q}_\lambda(\boldsymbol{\theta}_k; \boldsymbol{\Psi}^{(q)})$ given by:

$$\mathcal{Q}_\lambda(\boldsymbol{\theta}_k; \boldsymbol{\Psi}^{(q)}) = Q(\boldsymbol{\Psi}_k; \boldsymbol{\Psi}^{(q)}) - \lambda \sum_{j=1}^{p} |\beta_{k,j}|$$
$$= -\frac{1}{2\sigma_k^2} \sum_{i=1}^{n} \tau_{ik}^{(q)} \left(y_i - (\beta_{k,0} + \boldsymbol{\beta}_k^T \boldsymbol{x}_i)\right)^2 - \frac{n_k^{(q)}}{2} \log(2\pi\sigma_k^2) - \lambda \sum_{j=1}^{p} |\beta_{k,j}|.$$

Updating $\boldsymbol{\beta}_k$, for each component k, consists of solving an independent weighted Lasso problem where the weights are the posterior component membership probabilities $\tau_{ik}^{(q)}$. Each of these weighted Lasso problems is then separately solved by Coordinate Ascent. The CA algorithm, after starting from the previous EM-Lasso estimate as initial values, i.e $\beta_{kj}^{(0,q)} = \beta_{kj}^{(q)}$, calculates, at each iteration t, the following coordinate updates, until no significant change in (26):

$$\beta_{kj}^{(t+1,q)} = \mathcal{S}\left(\sum_{i=1}^{n} \tau_{ik}^{(q)} r_{ikj}^{(t,q)} x_{ij}; \lambda \sigma_k^{(q)2}\right) \Big/ \sum_{i=1}^{n} \tau_{ik}^{(q)} x_{ij}^2 \tag{26}$$
$$= \mathcal{S}\left(\mathbf{X}_j^T \mathbf{W}_k^{(q)} \boldsymbol{r}_{kj}^{(q)}; \lambda {\sigma_k^{(q)}}^2\right) / (\mathbf{X}_j^T \mathbf{W}_k^{(q)} \mathbf{X}_j), \tag{27}$$

with $r_{ikj}^{(t,q)} = y_i - \beta_{k0}^{(q)} - \boldsymbol{x}_i^T \boldsymbol{\beta}_k^{(t,q)} + \beta_{kj}^{(t,q)} x_{ij}$, $\boldsymbol{r}_{kj}^{(t,q)} = \boldsymbol{y} - \beta_{k0}^{(q)} \mathbf{1}_n - \mathbf{X}\boldsymbol{\beta}_k^{(t,q)} + \beta_{kj}^{(t,q)} \mathbf{X}_j$ is the residual without considering the contribution of the j-th coefficient, and $\mathbf{W}_k^{(q)} = \text{diag}(\boldsymbol{\tau}_k^{(q)})$. The parameter vector update is then taken at convergence of the CA algorithm, i.e $\boldsymbol{\beta}_k^{(q+1)} = \boldsymbol{\beta}_k^{(t+1,q)}$. Then, the intercept and the variance, have the following standard updates:

$$\beta_{k,0}^{(q+1)} = \sum_{i=1}^{n} \tau_{ik}^{(q)} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_k^{(q+1)}) \Big/ \sum_{i=1}^{n} \tau_{ik}^{(q)} = {\boldsymbol{\tau}_k^{(q)}}^T (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}_k^{(q+1)}) / \mathbf{1}_n^T \boldsymbol{\tau}_k^{(q)} \tag{28}$$

$$\sigma_k^{2(q+1)} = \sum_{i=1}^{n} \tau_{ik}^{(q)} \left(y_i - (\beta_{k,0}^{(q+1)} + \boldsymbol{x}_i^T \boldsymbol{\beta}_k^{(q+1)})\right)^2 \Big/ \sum_{i=1}^{n} \tau_{ik}^{(q)} \tag{29}$$
$$= \left\| \sqrt{\mathbf{W}_k^{(q)}} \left(\boldsymbol{y} - \boldsymbol{\beta}_{k,0}^{(q+1)} \mathbf{1}_n - \mathbf{X}\boldsymbol{\beta}_k^{(q+1)}\right) \right\|_2^2 / \mathbf{1}_n^T \boldsymbol{\tau}_k^{(q)}. \tag{30}$$

3.2 Algorithm tuning and model selection

In practice, appropriate values of the tuning parameters (λ, γ) as well as the number of experts K should be chosen. In order to select them, we use a modified BIC based on a grid of candidate values for K, λ and γ. This modified BIC is an extension of the criterion used in [21] for regularized mixture of regressions and was used in [7,11] and is defined as:

$$\text{BIC}(K, \lambda, \gamma) = L(\widehat{\boldsymbol{\Psi}}_{K,\lambda,\gamma}) - \text{df}(K, \lambda, \gamma) \frac{\log n}{2}, \tag{31}$$

where $\widehat{\boldsymbol{\Psi}}_{K,\lambda,\gamma}$ is the penalized log-likelihood estimator obtained by the EM-Lasso algorithm, and $\text{df}(K, \lambda, \gamma)$ is the estimated number of non-zero coefficients

in the model, interpreted as the degrees of freedom. Let's assume that $K_0 \in \{K_1, \ldots, K_M\}$, whith K_0 the true number of expert components. For each value of K, we define grids of tuning parameters $\{\lambda_1, \ldots, \lambda_{M_1}\}$ and $\{\gamma_1, \ldots, \gamma_{M_2}\}$. For each triplet (K, λ, γ), we calculated the penalized log-likelihood estimators $\widehat{\boldsymbol{\Psi}}_{K,\lambda,\gamma}$ and compute BIC(K, λ, γ). Finally, the model with parameters (K, λ, γ) having the highest BIC value, is then selected.

4 Experimental study

In this section, we study the performance of our approach on simulated data. The codes are written in Matlab and in R and will be made publicly available on https://github.com/fchamroukhi. Different evaluation criteria are used to assess the model's performance, including sparsity, estimation of parameters and clustering accuracy.

Sparsity performance In order to evaluate the sparsity of the model, we calculate the specificity/sensitivity defined by:

- Sensitivity: proportion of correctly estimated zero coefficients;
- Specificity: proportion of correctly estimated nonzero coefficients.

Clustering performance For measuring the clustering performance, we calculate the correct classification rate and the Adjusted Rate index (ARI) between the true simulated partition and the partition estimated by the EM algorithms. The estimated cluster labels are obtained by plugin the Baye's allocation rule for the estimated model, which consists of maximizing the posterior probabilities defined in 9 and calculated with the estimated parameters. That is, the estimated class label $\hat{z}_i$ for the i-th pair $(\boldsymbol{X}_i, \boldsymbol{Y}_i)$ is given by

$$\hat{z}_i = \arg\max_{k=1}^{K} \tau_{ik}(\hat{\boldsymbol{\Psi}}) \quad (i = 1, \ldots, n). \tag{32}$$

For calculating the classification rate, we evaluate all the possible permutations of the obtained partition, and the one giving the best rate is then retained.

4.1 Simulation study

The data are generated according to the following generative hierarchical process:

$$\begin{aligned} Z_i &\sim \text{Mult}(1; \alpha_1, \ldots, \alpha_K) \\ \boldsymbol{X}_i | Z_i = z_i &\sim \mathcal{N}_p(.; \boldsymbol{\mu}_{z_i}, \mathbf{R}_{z_i}) \\ \boldsymbol{Y}_i | \boldsymbol{X}_i = \boldsymbol{x}_i, Z_i = z_i &\sim \mathcal{N}_d(.; \beta_{z_i,0} + \boldsymbol{\beta}_{z_i}^T \boldsymbol{x}_i, \sigma_{z_i}^2). \end{aligned}$$

We consider a MoGGE model of $K = 2$ expert components. The parameters of the Gaussian gating function, whose prior probabilities are $\alpha_1 = \alpha_2 = 0.5$, are $\boldsymbol{\mu}_1 = (0, 1, -1, -1.5, 0, 0.5, 0, 0)^T$, $\boldsymbol{\mu}_2 = (2, 0, 1, -1.5, 0, -0.5, 0, 0)^T$ and

$\mathbf{R}_1 = \mathbf{R}_2 = \text{diag}(\nu_1^2, \ldots, \nu_K^2)$ with $\nu_1^2 = \ldots = \nu_K^2 = 1$. The parameters of the Gaussian expert regressors are $\boldsymbol{\beta}_1 = (0, 1.5, 0, 0, 0, 1, 0, -0.5)^T$, $\boldsymbol{\beta}_2 = (1, -1.5, 0, 0, 2, 0, 0, 0.5)$, and $\sigma_1 = \sigma_2 = 1$. For each data set, we sample $n = 300$ data pairs, and for each experiment, 100 datasets were generated to average the results and provide error bars. In order to get the best model for each sample in the sense of the BIC criterion, we estimated the penalized model with the following grids of values for the parameters: $\lambda = (0, 1, 2, \ldots, 25)$, $\gamma = (0, 1, 2, \ldots, 25)$; The minimum and maximum values selected for λ and γ are respectively 4, 20 and 3, 18. Then we selected the penalized model which maximizes the modified BIC value (31). The results will be provided in the parts below.

Obtained results

Parameter estimation accuracy Figure 1 shows the estimated parameters for the gating network, with the error bars, for the proposed approach and for the standard MoGGE model. Similarly, Figure 2 shows the estimated parameters

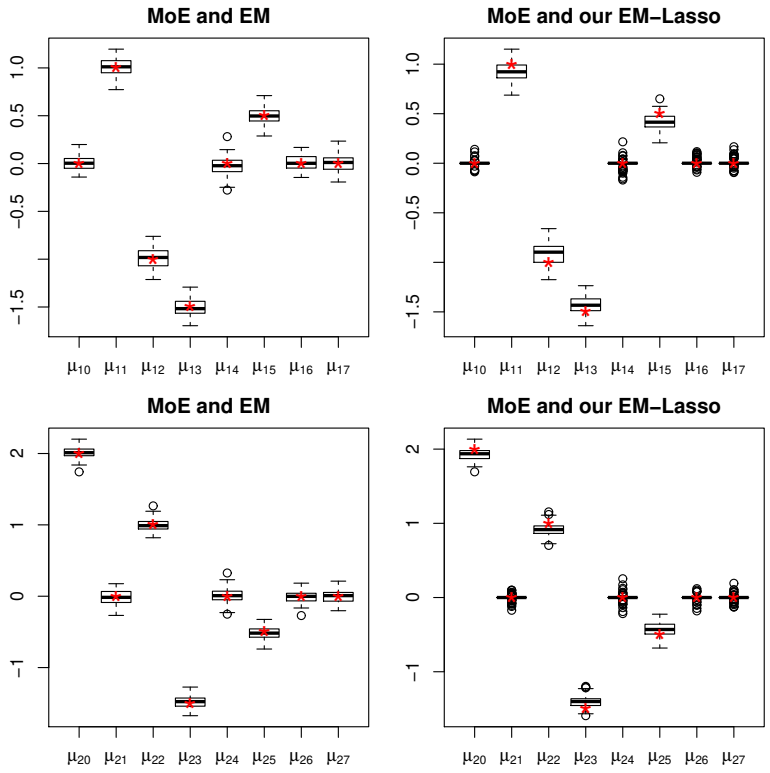

Fig. 1: Boxplots of the estimated gating network parameters $\mu_{k,j}$: component $k = 1$, top, and component $k = 2$, bottom. The red stars are the true values.

of the gating network. It can be seen on the two figures that, as expected, the

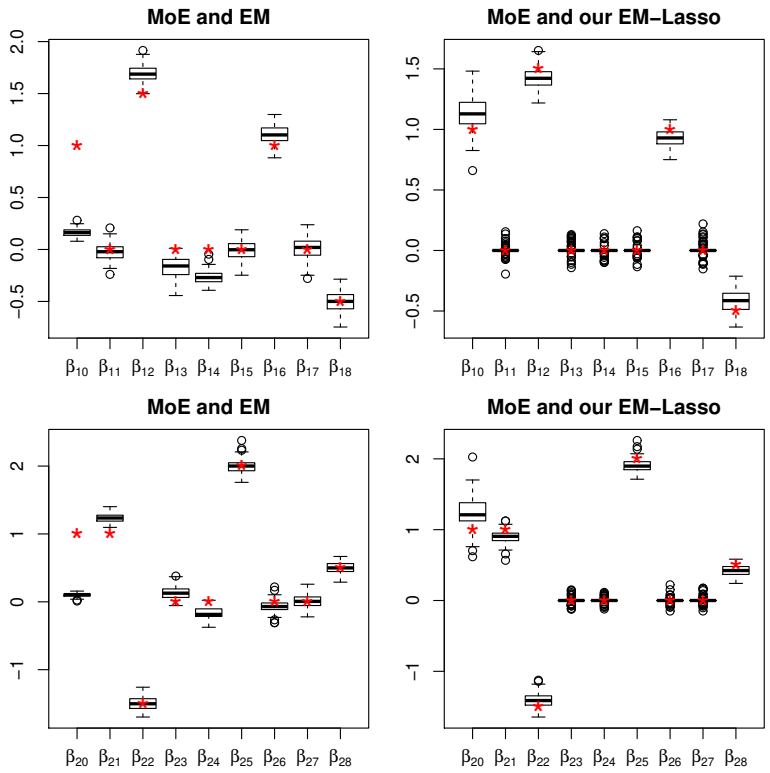

Fig. 2: Boxplots of the estimated expert network parameters $\beta_{k,j}$: component $k = 1$, top, and component $k = 2$, bottom. The red stars are the true values.

proposed lasso-regularization approach with the proposed EM-Lasso algorithm, clearly provides models that are sparser, compared to the standard approach with EM, where the zero-coefficients are not precisely recovered. This is observed for both the gating function parameters, and the expert function parameters. While the penalized version we can see that it may be subject of a bias in estimating the non-zero coefficients, the parameter estimated and the bias are still reasonable. Hence, if one would to encourage sparsity, and to still have a good performance in density estimation, then the penalized MoGGE is a better choice, compared to the standard MLE of the MoGGE model.

Sensitivity/specificity results Table 1 gives the sensitivity (S_1) and specificity (S_2) results for the two compared approaches. Note that here since we have two components, then only the estimation of one Gaussian gating function is considered, as the parameters of the other one are zeros. It can be seen that,

Table 1: Sensitivity (S_1) and specificity (S_2) results.

Method	Expert 1		Expert 2		Gate	
	S_1	S_2	S_1	S_2	S_1	S_2
MoGGE-EM	0.000	1.000	0.000	1.000	0.000	1.000
MoGGE-EMLasso-BIC	0.790	1.000	0.785	1.000	0.779	1.000

none of the parameters in the non penalized model has a null value. The penalized model provides naturally sparser models compared to the standard non-penalized one.

Clustering results We calculate the accuracy of clustering for each data set. The results in terms of correct classification rate and ARI values are provided in Table 2. We can see that the classification rate as well as as the Adjusted Rand Index are very close for the two methods, with a slight advantage to the proposed approach.

Table 2: Clustering results: correct classification rate and Adjusted Rand Index.

Model	C.rate	ARI
MoGGE - EM	$97.25\%_{(0.8770\%)}$	$89.28\%_{(3.325\%)}$
MoGGE-EMLasso-BIC	$97.43\%_{(0.8521\%)}$	$89.99\%_{(3.231\%)}$

Selecting the sparsity tuning parameters We compute the Lasso path for a sample with same parameters as presented at the beginning of the section. On Figure 3, we observe that even with very small values (null value as well, i.e. non penalized MoE) of γ, the true zero parameters have values very close to zero. We also note that for values of ratio close to 0.8 for both λ and γ, almost every true zero parameters have null values and the slight bias introduced in the true nonzero parameters is reasonable.

5 Conclusion and future work

In this paper, the mixture of Gaussian-gated experts is studied towards modeling and clustering of heterogeneous regression data with high-dimensional predictors. A regularized MLE approach is proposed to simultaneously perform parameter estimation and feature selection. The developed EM-Lasso algorithm to fit the model relies on coordinate ascent updates of the regularized parameters, and its application in numerical experiments clearly shows it provides sparse models. Its performance is also compared to the state-of-the art fitting with the EM algorithm, shows its good performance, in particular in terms of sparsity. The diagonal hypothesis of the covariance matrix to derive the regularization (19) is now being relaxed, so that the regularization is on the elements of the

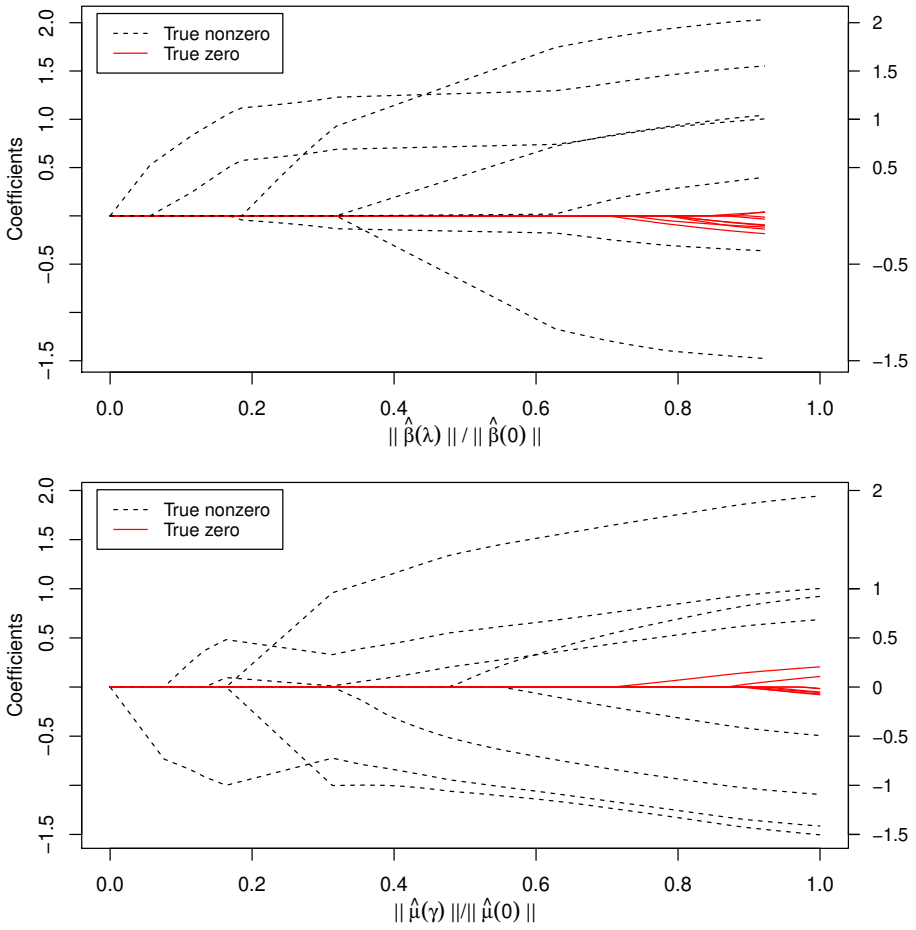

Fig. 3: Lasso paths of the estimated gating network parameters (top) and expert network parameters (bottom). The solid line represents the values of the true non-zero values, and the dashed line represents the true zero values.

precision matrix, i.e a graphical Lasso regularization. A future extension will also consider multivariate response with dedicated sparsity on the matrices of regression coefficients.

Acknowledgments

This research is supported by Ethel Raybould Fellowship (Univ. of Queensland), ANR SMILES ANR-18-CE40-0014, and Région Normandie RIN AStERiCs.

References

1. Chamroukhi, F.: Non-normal mixtures of experts (July 2015), arXiv:1506.06707

2. Chamroukhi, F.: Robust mixture of experts modeling using the t-distribution. Neural Networks - Elsevier **79**, 20–36 (2016)
3. Chamroukhi, F.: Skew-normal mixture of experts. In: The International Joint Conference on Neural Networks (IJCNN). Vancouver, Canada (July 2016)
4. Chamroukhi, F.: Skew t mixture of experts. Neurocomputing **266**, 390–408 (2017)
5. Chamroukhi, F., Samé, A., Govaert, G., Aknin, P.: A regression model with a hidden logistic process for feature extraction from time series. In: International Joint Conference on Neural Networks (IJCNN). pp. 489–496 (2009)
6. Chamroukhi, F., Trabelsi, D., Mohammed, S., Oukhellou, L., Amirat, Y.: Joint segmentation of multivariate time series with hidden process regression for human activity recognition. Neurocomputing **120**, 633–644 (2013)
7. Chamroukhi, F., Huynh, B.T.: Regularized Maximum Likelihood Estimation and Feature Selection in Mixtures-of-Experts Models. Journal de la Société Française de Statistique **160**(1), 57–85 (March 2019)
8. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. JRSS, B **39(1)**, 1–38 (1977)
9. Friedman, J., Hastie, T., Hfling, H., Tibshirani, R.: Pathwise coordinate optimization. Tech. rep., Annals of Applied Statistics (2007)
10. Hastie, T., Tibshirani, R., Wainwright, M.: Statistical Learning with Sparsity: The Lasso and Generalizations. Chapman & Hall/CRC (2015)
11. Huynh, T., Chamroukhi, F.: Estimation and feature selection in mixtures of generalized linear experts models. arXiv:1907.06994 (July 2019)
12. Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive mixtures of local experts. Neural Computation **3**(1), 79–87 (1991)
13. Jordan, M.I., Jacobs, R.A.: Hierarchical mixtures of experts and the EM algorithm. Neural Computation **6**, 181–214 (1994)
14. Khalili, A.: New estimation and feature selection methods in mixture-of-experts models. Canadian Journal of Statistics **38**(4), 519–539 (2010)
15. McLachlan, G.J., Krishnan, T.: The EM algorithm and extensions. New York: Wiley, second edn. (2008)
16. McLachlan, G.J., Peel., D.: Finite mixture models. New York: Wiley (2000)
17. Nguyen, H.D., Chamroukhi, F.: Practical and theoretical aspects of mixture-of-experts modeling: An overview. WIREs: Data Mining and Knowledge Discovery pp. e1246–n/a (Feb 2018). https://doi.org/10.1002/widm.1246
18. Nguyen, H.D., Chamroukhi, F., Forbes, F.: Approximation results regarding the multiple-output mixture of linear experts model. Neurocomputing **doi:10.1016/j.neucom.2019.08.014** (2019)
19. Nguyen, H.D., McLachlan, G.J.: Laplace mixture of linear experts. Computational Statistics & Data Analysis **93**, 177–191 (2016)
20. Pan, W., Shen, X.: Penalized model-based clustering with application to variable selection. J. Mach. Learn. Res. **8**, 1145–1164 (May 2007)
21. Städler, N., Bühlmann, P., van de Geer, S.: Rejoinder: l1-penalization for mixture regression models. TEST **19**(2), 280–285 (2010)
22. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B **58**(1), 267–288 (1996)
23. Wu, T.T., Lange, K.: Coordinate descent algorithms for lasso penalized regression. Ann. Appl. Stat. **2**(1), 224–244 (03 2008). https://doi.org/10.1214/07-AOAS147
24. Xu, L., Jordan, M.I., Hinton, G.E.: An alternative model for mixtures of experts. In: Tesauro, G., Touretzky, D.S., Leen, T.K. (eds.) Advances in Neural Information Processing Systems 7, pp. 633–640. MIT Press (1995)