Skew-Normal Mixture of Experts

Faicel Chamroukhi

Abstract— **Mixture of Experts (MoE) is a popular framework for modeling heterogeneity in data for regression, classification and clustering. For continuous data, MoE usually uses normal experts, that is, expert components following the Gaussian distribution. However, for a set of data containing a group or groups of observations with asymmetric distribution, the use of normal experts may be unsuitable. In this paper, we introduce the skew-normal MoE (SNMoE) which can deal with the issue regarding possibly skewed data distribution. We develop a dedicated expectation conditional maximization (ECM) 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. Numerical experiments carried out on simulated data show the effectiveness of the proposed model in terms modeling non-linear regression functions as well as in model-based clustering. The proposed model is applied to two real-world data sets: the tone perception data and the temperature anomalies data.**

I. Introduction

Mixtures of experts (MoE) introduced by [14] is widely studied in statistics and machine learning. It consists 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 covariates (inputs). MoE has been investigated, in its simple form, as well as in its hierarchical form [15] (e.g Section 5.12 of [19]) 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 [26]. 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 MoE, abbreviated NMoE. However, for a set of data containing a group or groups of observations with asymmetric behavior, 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 this limitation in NMoE by proposing more adapted MoE model which can deal with skewed data distribution by relying on the skew-normal distribution. Indeed, in these last years, the use of the skew normal distribution, firstly proposed by [2], [3], has been shown beneficial in dealing with asymmetric data in various theoretic and applied problems. This problem has been also studied in the finite mixture literature by namely [17] for modeling asymmetric univariate data with the univariate skew-normal mixture. Recently, [27] introduced the scale mixtures of skew-normal distributions for mixture of regressions. The inference in the previously described mixtures and standard MoE is performed by maximum likelihood estimation via the expectation-maximization (EM) algorithm or extensions [10], [18], in particular the expectation conditional maximization (ECM) algorithm [20]. In this paper, we attempt to overcome the limitation in NMoE by proposing more adapted MoE model which can deal with skewed data distribution. We investigate the use of the skew-normal distribution for the experts and consider the MoE framework for non-linear regression problems and model-based clustering of regression data. We therefore propose the skew-normal MoE (SNMoE) to accommodate data with possible asymmetric behavior. The model corresponds to an extension of the unconditional mixture of skew-normal distributions [17] to the MoE framework, where the mixture means are regression functions and the mixing proportions are covariate-varying. We call the proposed MoE model the skew-normal MoE (SNMoE). Unlike our proposed SNMoE model, the regression mixture model of [27] does not consider conditional mixing proportions, that is, mixing proportions depending on some input variables, as in the case of MoE, which we investigate here.

For the model inference, we develop a dedicated expectation conditional maximization (ECM) algorithm to estimate the parameters of the proposed models by monotonically maximizing the observed data log-likelihood. The EM algorithms are indeed very popular and successful estimation algorithms for mixture models in general and for MoE in particular. Moreover, the EM algorithm for MoE has been shown by [21] 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. The mixture models and the MoE models have also been considered in the Bayesian framework, but in this paper, we focus on the maximum likelihood estimation framework.

This paper is organized as follows. In Section II we briefly recall the MoE framework and the NMoE model. In Section III, we present the SNMoE model and in Section IV we present its inference technique using the ECM algorithm. Section V is dedicated to the use of the MoE in regression, clustering and to the model selection problem. In Section VI, we perform experiments on simulated and real data to assess the proposed model.

Faicel Chamroukhi is with the Lab of mathematics Paul Painlevé - UMR CNRS 8524 and the Information Sciences and Systems Lab - UMR CNRS 7296. He is invited at INRIA - Modal. He would like to thank the "FUI14 SYCIE project" and "Inria-Modal" for their financial support to this work. Contact: faicel.chamroukhi@univ-tln.fr

II. Mixture of experts (MoE) for continuous data

Mixture of experts [14], [15] are used in regression, classification and clustering. Here we consider the MoE framework for fitting (non-linear) regression functions and for clustering of univariate continuous data.

A. The MoE model

The univariate MoE model for regression assumes that the observed pairs of data $(\boldsymbol{x}, y)$ where $y \in \mathbb{R}$ is the response for some covariate vector $\boldsymbol{x} \in \mathbb{R}^p$, are generated from K regressors (experts) and are governed by a hidden categorical random variable Z indicating from which component each observation is generated. MoE for regression analysis [14], [15] thus decompose the nonlinear regression model density as follows:

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

where the mixing proportions, known as the gating network, are function of some covariates $\boldsymbol{r} \in \mathbb{R}^q$ and are modeled by the multinomial logistic model as follows:

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

where $\boldsymbol{\alpha}_k$ 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. The parameter vector of the MoE model is given by $\boldsymbol{\Psi} = (\boldsymbol{\alpha}^T, \boldsymbol{\Psi}_1^T, \ldots, \boldsymbol{\Psi}_K^T)^T$, $\boldsymbol{\Psi}_k$ being the parameter vector of the kth component density (expert).

B. The normal MoE (NMoE) model

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

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

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 $\boldsymbol{\Psi}$ are estimated by maximizing the observed data log-likelihood by using the EM algorithm [10], [14], [15], [16], [21], [18].

However, the normal distribution is not adapted to deal with asymmetric data. In the proposal, we address the issue regarding the skewness by proposing the skew-normal MoE (SNMoE) model.

III. The skew-normal MoE (SNMoE) model

The skew-normal MoE (SNMoE) model uses the skew-normal distribution as density for the expert components. We first recall the skew-normal distribution and describe its stochastic and hierarchical presentations, to then derive them for the proposed SNMoE model.

A. The skew-normal distribution

As introduced by [2], [3], a random variable Y follows a univariate skew-normal distribution with location parameter $\mu \in \mathbb{R}$, scale parameter $\sigma^2 \in (0, \infty)$ and skewness parameter $\lambda \in \mathbb{R}$ if it has the density

$$f(y; \mu, \sigma^2, \lambda) = \frac{2}{\sigma} \phi\Big(\frac{y-\mu}{\sigma}\Big) \Phi\left(\lambda\Big(\frac{y-\mu}{\sigma}\Big)\right) \tag{4}$$

where $\phi(.)$ and $\Phi(.)$ denote, respectively, the probability density function (pdf) and the cumulative distribution function (cdf) of the standard normal distribution. It can be seen from (4) that when $\lambda = 0$, the skew-normal reduces to the normal distribution. As presented by [3], [13], if

$$Y = \mu + \delta|U| + \sqrt{1-\delta^2} E \tag{5}$$

where $\delta = \frac{\lambda}{\sqrt{1+\lambda^2}}$, U and E are independent random variables following the normal distribution $\mathrm{N}(0, \sigma^2)$, then Y follows the skew-normal distribution with pdf $\mathrm{SN}(\mu, \sigma^2, \lambda)$ given by (4). In the above, $|U|$ denotes the magnitude of U. This stochastic representation of the skew-normal distribution leads to the following hierarchical representation in an incomplete data framework, as presented in [17]:

$$\begin{array}{lcl} Y|u & \sim & \mathrm{N}\left(\mu + \delta|u|, (1-\delta^2)\sigma^2\right), \\ U & \sim & \mathrm{N}(0, \sigma^2). \end{array} \tag{6}$$

This hierarchical representation greatly facilitates the inference for the model, namely in the skew-normal mixture model. Introduced by [17], a K-component skew-normal mixture model is given by:

$$f(y; \boldsymbol{\Psi}) = \sum_{k=1}^{K} \pi_k \ \mathrm{SN}(y; \mu_k, \sigma_k^2, \lambda_k) \tag{7}$$

where the mixture components have a skew-normal density given by (4). For the skew-normal mixture, the mixing proportions and the means of the mixture components are assumed to be constant.

In the following section, we present the skew-normal MoE (SNMoE) which extends the skew-normal mixture model to the case of MoE framework, by considering conditional distributions for both the mixing proportions and the means of the mixture components.

B. The skew-normal MoE (SNMoE)

The proposed skew-normal MoE (SNMoE) is a K-component MoE model with skew-normal experts. It is defined by:

$$f(y|\boldsymbol{x}, \boldsymbol{r}; \boldsymbol{\Psi}) = \sum_{k=1}^{K} \pi_k(\boldsymbol{r}; \boldsymbol{\alpha}) \mathrm{SN}\big(y; \mu(\boldsymbol{x}; \boldsymbol{\beta}_k), \sigma_k^2, \lambda_k\big). \tag{8}$$

In the SNMoE model, each expert component k has a skew-normal distribution, whose density is defined by (4). The parameter vector of the model is $\boldsymbol{\Psi} = (\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_{K-1}^T, \boldsymbol{\Psi}_1^T, \ldots, \boldsymbol{\Psi}_K^T)^T$ with $\boldsymbol{\Psi}_k = (\boldsymbol{\beta}_k^T, \sigma_k^2, \lambda_k)^T$ the parameter vector for the kth skewed-normal expert component. It is obvious to see that if the skewness parameter $\lambda_k = 0$ for each k, the SNMoE model (8) reduces to the NMoE model (3). Before going on the model inference, we first present its stochastic and hierarchical representations, which will serve to derive the ECM algorithm for maximum

likelihood parameter estimation. The SNMoE model is characterized as follows.

1) Stochastic representation of the SNMoE: By using the stochastic representation (5) of the skew-normal distribution, the stochastic representation for the SNMoE is as follows. Let U and E be independent univariate random variables following the standard normal distribution $\mathrm{N}(0,1)$ with pdf $\phi(.)$. Given some covariates $\boldsymbol{x}_i$ and $\boldsymbol{r}_i$, a random variable Y_i is said to follow the SNMoE model (8) if it has the following representation:

$$Y_i = \mu(\boldsymbol{x}_i;\boldsymbol{\beta}_{z_i}) + \delta_{z_i}\sigma_{z_i}|U_i| + \sqrt{1-\delta_{z_i}^2}\,\sigma_{z_i}E_i. \tag{9}$$

In (9), we have $\delta_{z_i} = \frac{\lambda_{z_i}}{\sqrt{1+\lambda_{z_i}^2}}$ where $z_i \in \{1,\ldots,K\}$ is a realization of the categorical variable Z_i which follows the multinomial distribution, that is:

$$Z_i|\boldsymbol{r}_i \sim \text{Mult}(1;\pi_1(\boldsymbol{r}_i;\boldsymbol{\alpha}),\ldots,\pi_K(\boldsymbol{r}_i;\boldsymbol{\alpha})) \tag{10}$$

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.

The stochastic representation (9) of the SNMoE leads to the following hierarchical representation, which, as it will be presented in Section IV, facilitates the model inference.

2) Hierarchical representation of the SNMoE: By introducing the binary latent component-indicators Z_{ik} such that $Z_{ik} = 1$ iff $Z_i = k$, a hierarchical representation of the SNMoE model can be derived from its stochastic representation (9) and is as follows:

$$\begin{aligned} Y_i|u_i, Z_{ik}=1, \boldsymbol{x}_i &\sim \mathrm{N}\Big(\mu(\boldsymbol{x}_i;\boldsymbol{\beta}_k)+\delta_k|u_i|, (1-\delta_k^2)\sigma_k^2\Big), \\ U_i|Z_{ik}=1 \quad &\sim \mathrm{N}(0,\sigma_k^2), \\ \boldsymbol{Z}_i|\boldsymbol{r}_i \quad &\sim \text{Mult}\left(1;\pi_1(\boldsymbol{r}_i;\boldsymbol{\alpha}),\ldots,\pi_K(\boldsymbol{r}_i;\boldsymbol{\alpha})\right) \end{aligned} \tag{11}$$

where $\boldsymbol{Z}_i = (Z_{i1},\ldots,Z_{iK})$ and $\delta_k = \frac{\lambda_k}{\sqrt{1+\lambda_k^2}}$.

IV. Maximum likelihood estimation of the SNMoE model

The unknown parameter vector $\boldsymbol{\Psi}$ of the SNMoE model can be estimated by maximizing the observed-data log-likelihood. Given an observed i.i.d sample of n observations $(y_1,\ldots,y_n)$ with their respective associated covariates $(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$ and $(\boldsymbol{r}_1,\ldots,\boldsymbol{x}_r)$, under the SNMoE model (8), the observed data log-likelihood of $\boldsymbol{\Psi}$ is given by:

$$\log L(\boldsymbol{\Psi}) = \sum_{i=1}^{n}\log\sum_{k=1}^{K}\pi_k(\boldsymbol{r}_i;\boldsymbol{\alpha})\mathrm{SN}\left(y_i;\mu(\boldsymbol{x}_i;\boldsymbol{\beta}_k),\sigma_k^2,\lambda_k\right). \tag{12}$$

The maximization of this log-likelihood can not be performed in a closed form. However, in this latent data framework, the maximization can be performed via expectation-maximization (EM)-type algorithms [18]. More specifically, we propose a dedicated Expectation Conditional Maximization (ECM) algorithm to monotonically maximize (12). The ECM algorithm [20] is an EM variant that mainly aims at addressing the optimization problem in the M-step of the EM algorithm. In ECM, the M-step is performed by several conditional maximization (CM) steps by dividing the parameter space into sub-spaces. The parameter vector updates are then performed sequentially, one coordinate block after another in each sub-space.

A. ECM-algorithm for the SNMoE model

Deriving the ECM algorithm requires the definition of the complete-data log-likelihood. From the hierarchical representation (11) of the SNMoE, the complete-data log-likelihood $\boldsymbol{\Psi}$, where the complete-data are $\{y_i, z_i, u_i, \boldsymbol{x}_i, \boldsymbol{r}_i\}_{i=1}^n$, is given by:

$$\log L_c(\boldsymbol{\Psi}) = \log L_c(\boldsymbol{\alpha}) + \sum_{k=1}^{K}\log L_c(\boldsymbol{\Psi}_k), \tag{13}$$

with $\log L_c(\boldsymbol{\alpha}) = \sum_{i=1}^n\sum_{k=1}^K Z_{ik}\log\pi_k(\boldsymbol{r}_i;\boldsymbol{\alpha})$, $\log L_c(\boldsymbol{\Psi}_k) = \sum_{i=1}^n Z_{ik}\Big[-\log(2\pi\sigma_k^2) - \frac{1}{2}\log(1-\delta_k^2) - \frac{d_{ik}^2}{2(1-\delta_k^2)} + \frac{\delta_k\ d_{ik}\ u_i}{(1-\delta_k^2)\sigma_k} - \frac{u_i^2}{2(1-\delta_k^2)\sigma_k^2}\Big]$, where $d_{ik} = \frac{y_i-\mu(\boldsymbol{x}_i;\boldsymbol{\beta}_k)}{\sigma_k}$. Then, the proposed ECM algorithm for the SNMoE model performs as follows. It starts with an initial parameter vector $\boldsymbol{\Psi}^{(0)}$ and alternates between the E- and CM- steps until a convergence criterion is satisfied.

1) E-Step: The E-Step of the ECM algorithm for the SNMoE calculates the Q-function, that is the conditional expectation of the complete-data log-likelihood (13), given the observed data $\{(y_i,\boldsymbol{x}_i,\boldsymbol{r}_i)\}_{i=1}^n$ and a current parameter estimation $\boldsymbol{\Psi}^{(m)}$, m being the current iteration:

$$Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(m)}) = \mathbb{E}\big[\log L_c(\boldsymbol{\Psi})|\{y_i,\boldsymbol{x}_i,\boldsymbol{r}_i\}_{i=1}^n;\boldsymbol{\Psi}^{(m)}\big]. \tag{14}$$

From (13), it follows that the Q-function is given by:

$$Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(m)}) = Q_1(\boldsymbol{\alpha};\boldsymbol{\Psi}^{(m)}) + \sum_{k=1}^{K}Q_2(\boldsymbol{\Psi}_k;\boldsymbol{\Psi}^{(m)}), \tag{15}$$

with

$$Q_1(\boldsymbol{\alpha};\boldsymbol{\Psi}^{(m)}) = \sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}^{(m)}\log\pi_k(\boldsymbol{r}_i;\boldsymbol{\alpha}), \tag{16}$$

$$\begin{aligned} Q_2(\boldsymbol{\Psi}_k;\boldsymbol{\Psi}^{(m)}) = \sum_{i=1}^{n}\tau_{ik}^{(m)}\Bigg[&-\log(2\pi)-\log(\sigma_k^2)-\frac{1}{2}\log(1-\delta_k^2) \\ &+\frac{\delta_k\ d_{ik}\ e_{1,ik}^{(m)}}{(1-\delta_k^2)\sigma_k} - \frac{e_{2,ik}^{(m)}}{2(1-\delta_k^2)\sigma_k^2} - \frac{d_{ik}^2}{2(1-\delta_k^2)}\Bigg] \end{aligned} \tag{17}$$

for $k = 1,\ldots,K$, where the required conditional expectations are given by:

$$\begin{aligned} \tau_{ik}^{(m)} &= \mathbb{E}_{\boldsymbol{\Psi}^{(m)}}\left[Z_{ik}|y_i,\boldsymbol{x}_i,\boldsymbol{r}_i\right], \\ e_{1,ik}^{(m)} &= \mathbb{E}_{\boldsymbol{\Psi}^{(m)}}\left[U_i|Z_{ik}=1,y_i,\boldsymbol{x}_i,\boldsymbol{r}_i\right], \\ e_{2,ik}^{(m)} &= \mathbb{E}_{\boldsymbol{\Psi}^{(m)}}\left[U_i^2|Z_{ik}=1,y_i,\boldsymbol{x}_i,\boldsymbol{r}_i\right]. \end{aligned} \tag{18}$$

The $\tau_{ik}^{(m)}$'s represent the posterior distribution of the hidden class labels Z_i and correspond to the posterior memberships of the observed data. They are given by:

$$\tau_{ik}^{(m)} = \frac{\pi_k(\boldsymbol{r}_i;\boldsymbol{\alpha}^{(m)})\mathrm{SN}\left(y_i;\mu(\boldsymbol{x}_i;\boldsymbol{\beta}_k^{(m)}),\sigma_k^{2\,(m)},\lambda_k^{(m)}\right)}{f(y_i|\boldsymbol{x}_i,\boldsymbol{r}_i;\boldsymbol{\Psi}^{(m)})}. \tag{19}$$

The conditional expectations $e_{1,ik}^{(m)}$ and $e_{2,ik}^{(m)}$ correspond to the posterior distribution of the hidden variables U_i and U_i^2, respectively. From the hierarchical representation (11), as shown by [17] in the case of the skew-normal mixture model, by Bayes' theorem, the posterior distribution of U_i is the following half normal:

$$U_i|Z_{ik}=1,y_i,\boldsymbol{x}_i,\boldsymbol{r}_i \sim HN_{[0,\infty)}\left(\mu_{u_{ik}},\sigma_{u_k}^2\right)$$

where the posterior mean and variance in this case of SNMoE are respectively given by:

$$\mu_{u_{ik}} = \delta_k(y_i - \mu(\boldsymbol{x}_i;\boldsymbol{\beta}_k)) \quad \text{and } \sigma^2_{u_k} = (1-\delta_k^2)\sigma_k^2.$$

Then the two conditional expectations of U_i and U_i^2 are respectively given by:

$$e_{1,ik}^{(m)} = {\mu_{u_{ik}}}^{(m)} + {\sigma_{u_k}}^{(m)} \frac{\phi\left(\lambda_k^{(m)} {d_{ik}}^{(m)}\right)}{\Phi\left(\lambda_k^{(m)} {d_{ik}}^{(m)}\right)}, \tag{20}$$

$$e_{2,ik}^{(m)} = {\mu^2_{u_{ik}}}^{(m)} + {\sigma^2_{u_k}}^{(m)} + {\mu_{u_{ik}}}^{(m)} {\sigma_{u_k}}^{(m)} \frac{\phi\left(\lambda_k^{(m)} {d_{ik}}^{(m)}\right)}{\Phi\left(\lambda_k^{(m)} {d_{ik}}^{(m)}\right)} \tag{21}$$

From (15), (16), and (17), it can be seen that the Q-function is calculated by analytically calculating the conditional expectations (19), (20) and (21).

2) M-Step: Then, the M-step calculates the parameter vector $\boldsymbol{\Psi}^{(m+1)}$ by maximizing the Q-function (15) with respect to $\boldsymbol{\Psi}$. This can be performed by separately maximizing $Q_1(\boldsymbol{\alpha};\boldsymbol{\Psi}^{(m)})$ with respect to $\boldsymbol{\alpha}$ and, for each component k $(k = 1,\ldots,K)$, the function $Q(\boldsymbol{\Psi}_k;\boldsymbol{\Psi}^{(m)})$ with respect to $\boldsymbol{\Psi}_k$ where $\boldsymbol{\Psi}_k = (\boldsymbol{\beta}_k^T, \sigma_k^2, \lambda_k)^T$. We adopt the ECM extension of the EM algorithm. The M-step in this case consists of four conditional maximization (CM)-steps, corresponding to the decomposition of the vector $\boldsymbol{\Psi}$ into four sub-vectors $\boldsymbol{\Psi} = (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T, \boldsymbol{\sigma}^T, \boldsymbol{\lambda}^T)^T$. This leads to the following CM steps.

a) CM-Step 1: Calculate $\boldsymbol{\alpha}^{(m+1)}$ as:

$$\boldsymbol{\alpha}^{(m+1)} = \arg\max_{\boldsymbol{\alpha}} Q_1(\boldsymbol{\alpha};\boldsymbol{\Psi}^{(m)}). \tag{22}$$

Contrarily to the case of the standard skew-normal mixture model and skew-normal regression mixture model, this maximization in the case of the proposed SNMoE does not exist in closed form. It is performed iteratively by Iteratively Reweighted Least Squares (IRLS).

b) The Iteratively Reweighted Least Squares (IRLS) algorithm: The IRLS algorithm is used to maximize $Q_1(\boldsymbol{\alpha},\boldsymbol{\Psi}^{(m)})$ given by (16) with respect to the parameter $\boldsymbol{\alpha}$ in the M step at each iteration m of the ECM algorithm. The IRLS is a Newton-Raphson algorithm, which consists in starting with a vector $\boldsymbol{\alpha}^{(0)}$, and, at the $l+1$ iteration, updating the estimation of $\boldsymbol{\alpha}$ as follows:

$$\boldsymbol{\alpha}^{(l+1)} = \boldsymbol{\alpha}^{(l)} - \left[\frac{\partial^2 Q_1(\boldsymbol{\alpha},\boldsymbol{\Psi}^{(m)})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^T}\right]^{-1}_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^{(l)}} \frac{\partial Q_1(\boldsymbol{\alpha},\boldsymbol{\Psi}^{(m)})}{\partial\boldsymbol{\alpha}}\Big|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^{(l)}} \tag{23}$$

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 similarly as in [8][7]. The parameter update $\boldsymbol{\alpha}^{(m+1)}$ is taken at convergence of the IRLS algorithm (23). Then, for $k = 1\ldots,K$,

c) CM-Step 2: Calculate $\boldsymbol{\beta}_k^{(m+1)}$ by maximizing $Q_2(\boldsymbol{\Psi}_k;\boldsymbol{\Psi}^{(m)})$ given by (17) w.r.t $\boldsymbol{\beta}_k$. Here we focus on the common linear case for the experts where each expert-component mean function is the one of a linear regression model and has the form $\mu(\boldsymbol{x}_i;\boldsymbol{\beta}_k) = \boldsymbol{\beta}_k^T\boldsymbol{x}_i$. It can be easily shown that the maximization problem for the resulting skew-normal mixture of linear of experts (SNMoLE) can be solved analytically and has the following solution:

$$\boldsymbol{\beta}_k^{(m+1)} = \Big[\sum_{i=1}^n \tau_{ik}^{(m)}\boldsymbol{x}_i\boldsymbol{x}_i^T\Big]^{-1}\sum_{i=1}^n \tau_{ik}^{(q)}\left(y_i - \delta_k^{(m)}e_{1,ik}^{(m)}\right)\boldsymbol{x}_i. \tag{24}$$

d) CM-Step 3: Calculate $\sigma_k^{2\,(m+1)}$ by maximizing $Q_2(\boldsymbol{\Psi}_k;\boldsymbol{\Psi}^{(m)})$ given by (17) w.r.t σ_k^2. Similarly to the update of $\boldsymbol{\beta}_k$, the analytic solution of this problem is given by:

$$\sigma_k^{2\,(m+1)} = \frac{\sum_{i=1}^n \tau_{ik}^{(m)}\Big[\left(y_i - \boldsymbol{\beta}_k^{T\,(m+1)}\boldsymbol{x}_i\right)^2 - 2\delta_k^{(m+1)}e_{1,ik}^{(m)}(y_i - \boldsymbol{\beta}_k^{T\,(m+1)}\boldsymbol{x}_i) + e_{2,ik}^{(m)}\Big]}{2\left(1-\delta_k^{2\,(m)}\right)\sum_{i=1}^n \tau_{ik}^{(m)}}. \tag{25}$$

e) CM-Step 4: Calculate $\lambda_k^{(m+1)}$ by maximizing $Q_2(\boldsymbol{\Psi}_k;\boldsymbol{\Psi}^{(m)})$ given by (17) w.r.t λ_k, with $\boldsymbol{\beta}_k$ and σ_k^2 fixed at $\boldsymbol{\beta}_k^{(m+1)}$ and $\sigma_k^{2\,(m+1)}$, respectively. This consists in solving the following equation for δ_k to obtain $\delta_k^{(m+1)}$ $(k = 1,\ldots,K)$ as the solution of:

$$\begin{aligned}&\sigma_k^{2\,(m+1)}\delta_k(1-\delta_k^2)\sum_{i=1}^n \tau_{ik}^{(m)} + (1+\delta_k^2)\sum_{i=1}^n \tau_{ik}^{(m)}(y_i - \boldsymbol{\beta}_k^{T\,(m+1)}\boldsymbol{x}_i)\; e_{1,ik}^{(m)}\\ &-\delta_k\sum_{i=1}^n \tau_{ik}^{(m)}\Big[e_{2,ik}^{(m)} + \left(y_i - \boldsymbol{\beta}_k^{T\,(m+1)}\boldsymbol{x}_i\right)^2\Big] = 0.\end{aligned} \tag{26}$$

Then, given the update $\delta_k^{(m+1)}$, the update of the skewness parameter λ_k is calculated as $\lambda_k^{(m+1)} = \frac{\delta_k^{(m+1)}}{\sqrt{1-\delta_k^{2\,(m+1)}}}$.

It is obvious to see that when the skewness parameter $\lambda_k = \delta_k = 0$ for all k, the parameter updates for the SNMoE corresponds to those of the NMoE. Hence, compared to the standard NMoE, the SNMoE model has an additional flexibility feature, that is the one to handle possibly skewed data.

V. Prediction, clustering and model selection

In regression analysis using MoE, one can 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 ECM. These predictions can be expressed in terms of the predictive distribution of y obtained by substituting the estimated parameter $\hat{\boldsymbol{\Psi}}$ into (1) to give:

$$f(y|\boldsymbol{x},\boldsymbol{r};\hat{\boldsymbol{\Psi}}) = \sum_{k=1}^K \pi_k(\boldsymbol{r};\hat{\boldsymbol{\alpha}})f_k(y|\boldsymbol{x};\hat{\boldsymbol{\Psi}}_k). \tag{27}$$

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 $\hat{y} = \mathbb{E}_{\hat{\boldsymbol{\Psi}}}(Y|\boldsymbol{x},\boldsymbol{r})$. It is easy to show that for the NMoE model, the normal expert means and variances are respectively given by $\mathbb{E}_{\hat{\boldsymbol{\Psi}}}(Y|Z=k,\boldsymbol{x}) = \hat{\boldsymbol{\beta}}_k^T\boldsymbol{x}$ and $\mathbb{V}_{\hat{\boldsymbol{\Psi}}}(Y|Z=k,\boldsymbol{x}) = \hat{\sigma}_k^2$. Then, it follows that the mean of the NMoE is given by $\mathbb{E}_{\hat{\boldsymbol{\Psi}}}(Y|\boldsymbol{x},\boldsymbol{r}) = \sum_{k=1}^K \pi_k(\boldsymbol{r};\hat{\boldsymbol{\alpha}}_n)\hat{\boldsymbol{\beta}}_k^T\boldsymbol{x}$. Then, similarly, the expected value for the proposed SNMoE model is $\mathbb{E}_{\hat{\boldsymbol{\Psi}}}(Y|Z=k,\boldsymbol{x}) = \hat{\boldsymbol{\beta}}_k^T\boldsymbol{x} + \sqrt{\frac{2}{\pi}}\;\hat{\delta}_k\;\hat{\sigma}_k$ and the expert variance is $\mathbb{V}_{\hat{\boldsymbol{\Psi}}}(Y|Z=k,\boldsymbol{x}) = \left(1-\frac{2}{\pi}\hat{\delta}_k^2\right)\hat{\sigma}_k^2$ where $\hat{\delta}_k = \frac{\hat{\lambda}_k}{\sqrt{1+\hat{\lambda}_k^2}}$. It follows that the mean of the SNMoE

model is given by: $\mathbb{E}_{\hat{\boldsymbol{\Psi}}}(Y|\boldsymbol{x},\boldsymbol{r}) = \sum_{k=1}^{K} \pi_k(\boldsymbol{r};\hat{\boldsymbol{\alpha}})\left(\hat{\boldsymbol{\beta}}_k^T\boldsymbol{x} + \sqrt{\frac{2}{\pi}}\hat{\delta}_k\hat{\sigma}_k\right)$. Finally, the variance for each MoE model is obtained easily from these specific expert mean and variance calculated in the above.

Model-based clustering using the SNMoE consists in assuming that the observed data $\{\boldsymbol{x}_i, \boldsymbol{r}_i, y_i\}_{i=1}^n$ are generated from a K component SNMoE with parameter vector $\boldsymbol{\Psi}$. Once the parameters are estimated (here by E(C)M), the provided posterior membership probabilities τ_{ik} given by (19) represent a fuzzy partition of the data. A hard partition of the data can then be obtained from the posterior memberships by applying the Bayes' allocation rule:

$$\hat{z}_i = \arg\max_{k=1}^{K} \hat{\tau}_{ik} \tag{28}$$

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

The problem of model selection for MoE is equivalent to the one of choosing the optimal number of experts K, the value of p related to the polynomial regression and the value of 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 $\text{AIC}(K,p,q) = \log L(\hat{\boldsymbol{\Psi}}) - \eta_{\boldsymbol{\Psi}}$ [1], the Bayesian Information Criterion $\text{BIC}(K,p,q) = \log L(\hat{\boldsymbol{\Psi}}) - \frac{\eta_{\boldsymbol{\Psi}}\log(n)}{2}$ [24], or the Integrated Classification Likelihood criterion $\text{ICL}(K,p,q) = \log L_c(\hat{\boldsymbol{\Psi}}) - \frac{\eta_{\boldsymbol{\Psi}}\log(n)}{2}$ [5]. In the above, $\log L(\hat{\boldsymbol{\Psi}})$ and $\log L_c(\hat{\boldsymbol{\Psi}})$ are respectively the 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 $\eta_{\boldsymbol{\Psi}}$ is given by $\eta_{\boldsymbol{\Psi}} = K(p+q+2) - q$ for the NMoE model and $\eta_{\boldsymbol{\Psi}} = K(p+q+3) - q$ for the proposed SNMoE model.

However, note that in MoE it is common to use mixing proportions 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 also adopted in this work. Moreover, for the case of linear experts (linear regressors corresponding to $p = 2$), the model selection reduces to choosing the number of experts K.

VI. Experimental study

A. An illustrative example

We first start by an illustrative example by considering a data set generated from an arbitrary non-linear function, which was analyzed by [6] and elsewhere. This data set consists of $n = 250$ values of input variables x_i generated uniformly in $(0, 1)$ and output variables y_i generated as $y_i = x_i + 0.3\sin(2\pi x_i) + \epsilon_i$, with ϵ_i drawn from a zero mean Normal distribution with standard deviation 0.05. To apply the MoE models, we set the covariate vectors $(\boldsymbol{x}_i, \boldsymbol{r}_i)$ to $\boldsymbol{x}_i = \boldsymbol{r}_i = (1, x_i)^T$. We considered mixture of three linear experts as in [6].

Figure 1 shows the fitted expert mean functions, the corresponding partitions obtained by using the Bayes' rule, and the gating network outputs. One can observe that the SNMoE model is successfully applied and provides results very close to those obtained by the NMoE.

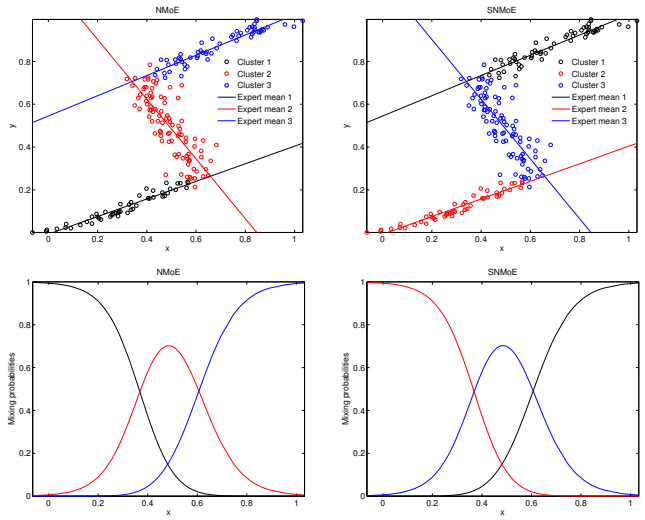

Fig. 1. Fitting the NMoE model and the proposed SNMoE to the toy data set analyzed in [6].

B. Experiments on simulation data sets

In this section we perform an experimental study on simulated data sets to apply and assess the proposed model. The aims is to observe the effect of the sample size on the estimation quality. 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 $K = 2$. The covariate variables $(\boldsymbol{x}_i, \boldsymbol{r}_i)$ are simulated such that $\boldsymbol{x}_i = \boldsymbol{r}_i = (1, x_i)^T$ (i.e $p = q = 2$) where x_i is simulated uniformly in $(-1, 1)$. We consider each of the two models for data generation (NNMoE, SNMoE), 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) and (8). 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 I.

	parameters			
comp. 1	$\boldsymbol{\alpha}_1 = (0, 10)^T$	$\boldsymbol{\beta}_1 = (0, 1)^T$	$\sigma_1 = 0.1$	$\lambda_1 = 3$
comp. 2	$\boldsymbol{\alpha}_2 = (0, 0)^T$	$\boldsymbol{\beta}_2 = (0, -1)^T$	$\sigma_2 = 0.1$	$\lambda_2 = -10$

TABLE I
PARAMETER VALUES USED IN SIMULATION.

Table II shows the obtained results in terms of the MSE for the SNMoE model. One can observe that the parameter estimation error is decreasing as n increases, which confirms the convergence property of the maximum likelihood estimator. 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 show in Figures 2 and 3 the estimated quantities provided by applying the proposed SNMoE model and their true counterparts for a data set ($n = 500$) generated according

param. n	α_{10}	α_{11}	β_{10}	β_{11}	β_{20}	β_{21}	σ_1	σ_2	λ_1	λ_2
50	1.10105	4.1882	0.00916	0.004890	0.007370	0.00348000	0.001647	0.002234	3.000	4.999
100	0.28074	1.0663	0.008301	0.0006118	0.006360	0.00007904	0.001314	0.001650	2.999	5.000
200	0.03893	0.9343	0.004709	0.0000398	0.005962	0.00005873	0.001142	0.001552	2.999	5.000
500	0.02340	0.0908	0.004475	0.0000195	0.005803	0.00000796	0.001026	0.001521	3.000	4.999
1000	0.00025	0.0613	0.003912	0.0000012	0.005499	0.00000344	0.000667	0.001517	2.999	3.999

TABLE II

MSE BETWEEN THE ESTIMATED SNMOE PARAMETERS AND THE ACTUAL ONES FOR A VARYING SAMPLE SIZE n.

the NMoE model and the SNMoE model, respectively.

One can clearly see that the estimated experts and mean functions provided by the proposed model are very close to the true ones, including when the data are generated according to the NMoE model. This provides an additional support to the fact that the proposed algorithm performs well and the proposed SNMoE model is a good generalization of the NMoE model, since it clearly approaches the NMoE as shown in this simulated example. One can also clearly see that the partitions estimated by the SNMoE model are close the actual partitions. The proposed SNMoE model can therefore be used as alternative to the NMoE model for both regression and model-based clustering.

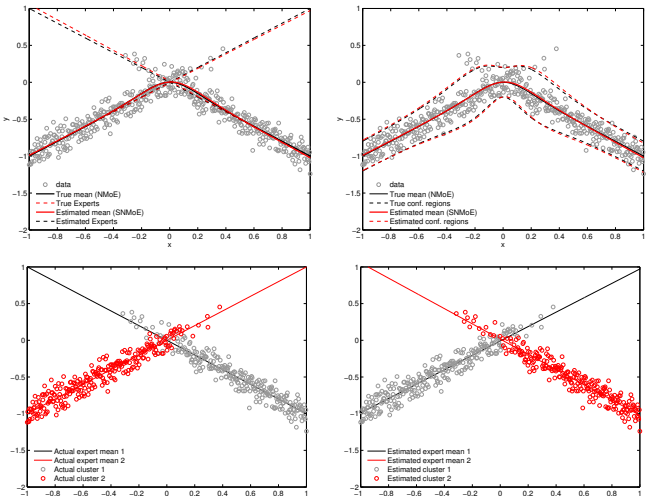

Fig. 2. Fitted SNMoE to data generated according to the NMoE.

C. 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 Figure 4.

1) Tone perception data set: The first analyzed data set is the real tone perception data set[1] which goes back to [9]. It was recently studied by [4] and [25]. 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

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

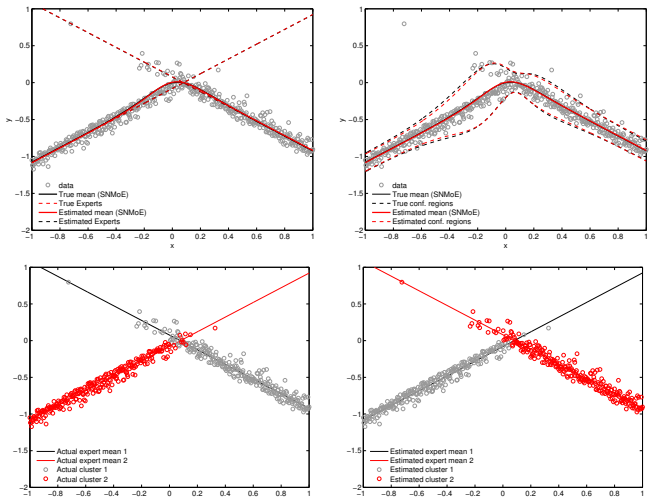

Fig. 3. Fitted SNMoE to data generated according to the SNMoE.

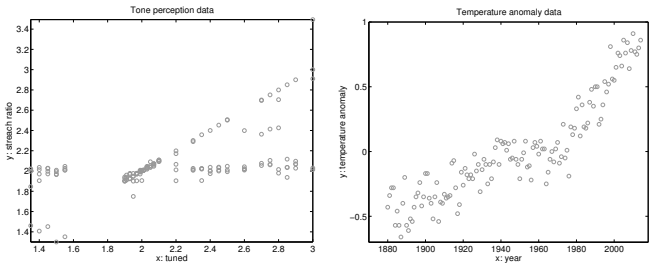

Fig. 4. Scatter plot of the tone perception data (left) and the temperature anomalies data (right).

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 first follow the study in [4] and [25] by using two expert components and then perform model selection (see Table IV).

Figure 5 shows the scatter plots of the tone perception data and the linear expert components of the fitted NMoE model and the proposed SNMoE model. One can observe that we obtain a good fit with the two models. The NMoE and SNMoE solutions are quasi-identical. The two regression lines may correspond to correct tuning and tuning to the first overtone, respectively, as analyzed in [4]. The values of estimated parameters for the tone perception data set are given in Table III. One can see that the SNMoE model parameters are identical to those of the NMoE, with a skewness close to zero, which tends to promote a non skewed distribution.

We also performed a model selection procedure on this

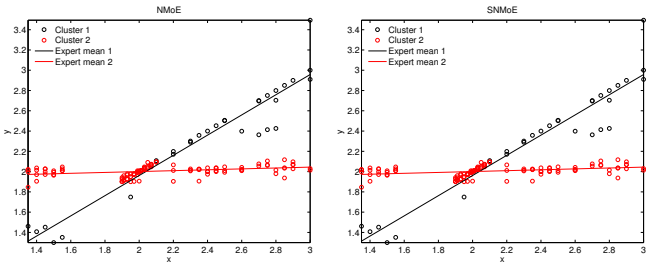

Fig. 5. Fitting the NMoE model (left) and the SNMoE model (right) to the tone perception data. The predictor x is the actual tone ratio and the response y is the perceived tone ratio.

param. model	α_{10}	α_{11}	β_{10}	β_{11}	β_{20}	β_{21}	σ_1	σ_2	λ_1	λ_2
NMoE	-2.690	0.796	-0.029	0.995	1.913	0.043	0.137	0.047	-	-
SNMoE	-2.694	0.797	-0.029	0.995	1.913	0.043	0.137	0.047	5.2e-13	-1.65e-13

TABLE III

ESTIMATED MOE PARAMETERS FOR THE TONE PERCEPTION DATA SET.

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 IV gives the obtained values of the model selection criteria. One can see that for the NMoE model overestimate the number of components. AIC performs poorly for the two models. BIC provides the correct number of components for the proposed model. ICL also estimated the correct number of components for the SNMoE model. One can conclude that the BIC and the ICL are the criteria to be suggested for the analysis of this data, with the SNMoE model, which is more adapted.

K	NMoE			SNMoE		
	BIC	AIC	ICL	BIC	AIC	ICL
1	1.8662	6.3821	1.8662	-0.6391	5.3821	-0.6391
2	122.8050	134.8476	107.3840	117.7939	132.8471	102.4049
3	118.1939	137.7630	76.5249	122.8725	146.9576	98.0442
4	121.7031	148.7989	94.4606	109.5917	142.7087	97.6108
5	141.6961	176.3184	123.6550	107.2795	149.4284	96.6832

TABLE IV

MODEL SELECTION FOR THE TONE PERCEPTION DATA.

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[2] 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 U.S. 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 [11], [12] and recently by [22] by using the Laplace mixture of linear experts (LMoLE).

To apply the two MoE models, we consider mixtures of two experts as in [22]. This number of components is also the one provided by the model selection criteria as shown later in Table VI. 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.

Figure 6 shows, for each of the two MoE models, the two fitted linear expert components, the corresponding means and confidence regions computed as plus and minus twice the estimated (pointwise) standard deviation. One can observe that the model is successfully applied on the data set and provide very similar results to the NMoE model. The values

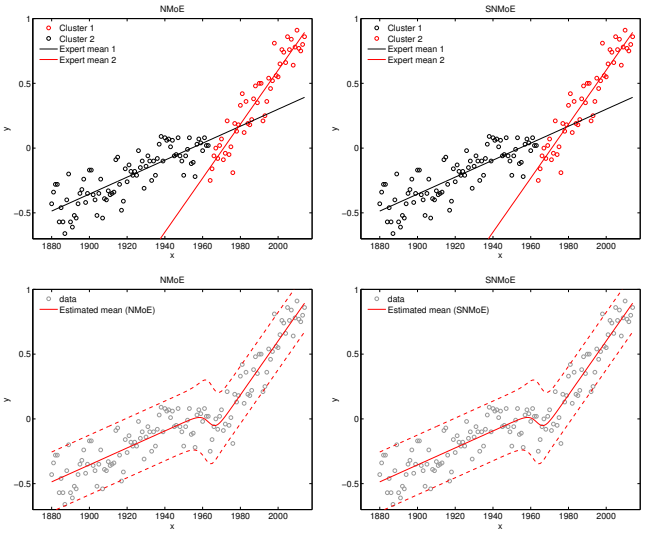

Fig. 6. Fitting the NMoE model (left) and the SNMoE model (right) to the temperature anomalies data set. The shaded region represents plus and minus twice the estimated (pointwise) standard deviation. The predictor x is the year and the response y is the temperature anomaly.

of estimated MoE parameters for this data set are given in Table V. One can see that the parameters common for the two models are quasi-identical. It can also be seen that the SNMoE model provides a fit with a skewness very close to zero. This may support the hypothesis of non-asymmetry for this data set. As mentioned by [22], [12] 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 to 1965). Documentation of the basic analysis method is provided by [11], [12].

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 VI 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

[2]source: from [23], http://cdiac.ornl.gov/ftp/trends/temp/hansen/gl_land.txt

param. model	α_{10}	α_{11}	β_{10}	β_{11}	β_{20}	β_{21}	σ_1	σ_2	λ_1	λ_2
NMoE	946.48	-0.481	-12.805	0.006	-41.073	0.020	0.115	0.110	-	-
SNMoE	950.95	-0.484	-12.805	0.006	-41.074	0.020	0.115	0.110	-8.7e-13	-9.2e-13

TABLE V

ESTIMATED MOE PARAMETERS FOR THE TEMPERATURE ANOMALIES.

high number of components, all the others results provide evidence for two components in the data.

	NMoE			SNMoE		
K	BIC	AIC	ICL	BIC	AIC	ICL
1	46.0623	50.4202	46.0623	43.6096	49.4202	43.6096
2	79.9163	91.5374	79.6241	75.0116	89.5380	74.7395
3	71.3963	90.2806	58.4874	63.9254	87.1676	50.8704
4	66.7276	92.8751	54.7524	55.4731	87.4312	41.1699
5	59.5100	92.9206	51.2429	45.3469	86.0207	41.0906

TABLE VI

MODEL SELECTION FOR THE TEMPERATURE ANOMALIES DATA.

VII. CONCLUSION AND FUTURE WORK

In this paper, we proposed a new non-normal MoE model, which generalizes the normal MoE. It is based on the skew-normal distribution. The SNMoE model is suggested for non-symmetric data. We developed an ECM algorithm to infer the model parameters and described the use of the model in non-linear regression and prediction as well as in model-based clustering. The results obtained on simulated data confirm the good performance of the model in terms of non-linear regression function approximation and clustering. The proposed model was also successfully applied to two different real data sets. Note that however both the NMoE and the proposed SNMoE can be affected by atypical observations. The use of MoE based on for example the t or the Laplace distribution is more robust. The model selection for the studied data tends to promote using BIC with the proposed SNMoE against in particular AIC which may perform poorly in the analyzed data, as well as against using BIC with the NMoE.

Here we only considered the MoE in their non-hierarchical version. One interesting future direction is therefore to extend the proposed model to the hierarchical MoE framework [15]. Furthermore, a natural future extension of this work is to consider the case of MoE with multivariate outputs.

REFERENCES

[1] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[2] A. Azzalini, "A class of distributions which includes the normal ones," *Scandinavian Journal of Statistics*, pp. 171–178, 1985.

[3] ——, "Further results on a class of distributions which includes the normal ones," *Scandinavian Journal of Statistics*, pp. 199–208, 1986.

[4] X. Bai, W. Yao, and J. E. Boyer, "Robust fitting of mixture regression models," *Computational Statistics & Data Analysis*, vol. 56, no. 7, pp. 2347 – 2359, 2012.

[5] C. Biernacki, G. Celeux, and G. Govaert, "Assessing a mixture model for clustering with the integrated completed likelihood," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 7, pp. 719–725, 2000.

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

[7] F. Chamroukhi, A. Samé, G. Govaert, and P. Aknin, "Time series modeling by a regression approach based on a latent process," *Neural Networks*, vol. 22, no. 5-6, pp. 593–602, 2009.

[8] ——, "A hidden process regression model for functional data description. application to curve discrimination," *Neurocomputing*, vol. 73, no. 7-9, pp. 1210–1221, March 2010.

[9] E. A. Cohen, "Some effects of inharmonic partials on interval perception," *Music Perception*, vol. 1, 1984.

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

[11] J. Hansen, R. Ruedy, J. Glascoe, and M. Sato, "Giss analysis of surface temperature change," *Journal of Geophysical Research*, vol. 104, pp. 30 997–31 022, 1999.

[12] J. Hansen, R. Ruedy, S. M., M. Imhoff, W. Lawrence, D. Easterling, T. Peterson, and T. Karl, "A closer look at united states and global surface temperature change," *Journal of Geophysical Research*, vol. 106, pp. 23 947–23 963, 2001.

[13] N. Henze, "A probabilistic representation of the skew-normal distribution," *Scandinavian Journal of Statistics*, pp. 271–275, 1986.

[14] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Computation*, vol. 3, no. 1, pp. 79–87, 1991.

[15] M. I. Jordan and R. A. Jacobs, "Hierarchical mixtures of experts and the EM algorithm," *Neural Computation*, vol. 6, pp. 181–214, 1994.

[16] M. I. Jordan and L. Xu, "Convergence results for the EM approach to mixtures of experts architectures," *Neural Networks*, vol. 8, no. 9, pp. 1409–1431, 1995.

[17] T. I. Lin, J. C. Lee, and S. Y. Yen, "Finite mixture modelling using the skew normal distribution," *Statistica Sinica*, vol. 17, pp. 909–927, 2007.

[18] G. J. McLachlan and T. Krishnan, *The EM algorithm and extensions*, 2nd ed. New York: Wiley, 2008.

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

[20] X. L. Meng and D. B. Rubin, "Maximum likelihood estimation via the ECM algorithm: A general framework," *Biometrika*, vol. 80, no. 2, pp. 267–278, 1993.

[21] S.-K. Ng and G. J. McLachlan, "Using the em algorithm to train neural networks: misconceptions and a new algorithm for multiclass classification." *IEEE Transactions on Neural Networks*, vol. 15, no. 3, pp. 738–749, 2004.

[22] H. D. Nguyen and G. J. McLachlan, "Laplace mixture of linear experts," *Computational Statistics & Data Analysis*, vol. 93, pp. 177–191, 2016.

[23] R. Ruedy, M. Sato, and K. Lo, "NASA GISS surface temperature (GISTEMP) analysis," DOI: 10.3334/CDIAC/cli.001, center for Climate Systems Research, NASA Goddard Institute for Space Studies.

[24] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, pp. 461–464, 1978.

[25] W. Song, W. Yao, and Y. Xing, "Robust mixture regression model fitting by laplace distribution," *Computational Statistics & Data Analysis*, vol. 71, no. 0, pp. 128 – 137, 2014.

[26] S. E. Yuksel, J. N. Wilson, and P. D. Gader, "Twenty years of mixture of experts." *IEEE Trans. Neural Netw. Learning Syst.*, vol. 23, no. 8, pp. 1177–1193, 2012.

[27] C. B. Zeller, V. H. Lachos, and C. Cabral, "Robust mixture regression modelling based on scale mixtures of skew-normal distributions," *Test (revision invited)*, 2015.