# On some new Mixtures-of-Experts Models

Faicel Chamroukhi

Research Summer School on Statistics for Data Science S4D 2018

June 22, 2018

- Data with possible atypical observations, skewed

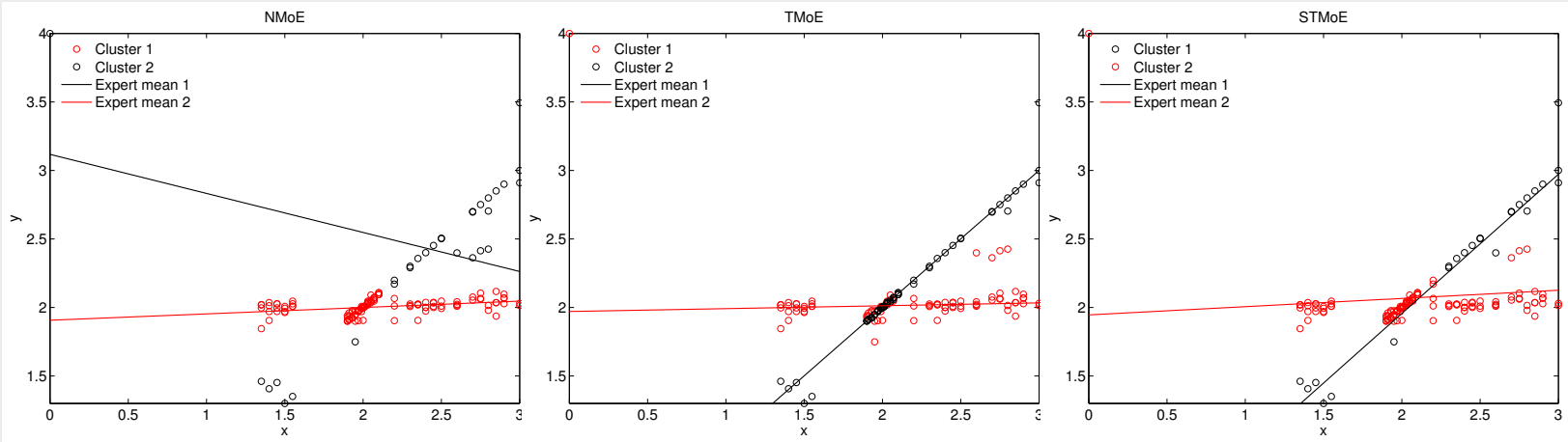



Figure: Fitting MoLE to the tone data set with ten outliers $(0, 4)$.

## Objectives

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

# Outline

1. Introduction
2. Non-normal mixtures of experts

1 Introduction

2 Non-normal mixtures of experts
- The skew-normal mixture of experts model
- The $t$ mixture of experts model
- The skew $t$ mixture of experts model
- Prediction, clustering and model selection with the non-normal MoE
- Experiments
- An illustrative example

# Non-normal mixtures of experts

## Problem

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

## Objectives

- Overcome these (well-known) limitations of MoE modeling with the normal distribution.
- We proposed three non-normal derivations including two robust mixture of experts (MoE) models. $\hookrightarrow$ suitable to accommodate data which exhibit additional features such as skewness, heavy-tails and which may be affected by atypical data **??**Chamroukhi (2015)

# Mixture of experts for continuous data

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

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

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

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

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

# The normal mixture of experts model and its MLE

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

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

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

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

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

# Contribution

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

# The skew-normal mixture of experts model

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

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

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

# The skew-normal mixture of experts model

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

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

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

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

# The skew-normal mixture of experts model

The SNMoE model is characterized as follows.

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

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

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

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

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

# The skew-normal mixture of experts model

The SNMoE model is characterized as follows.

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

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

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

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

# MLE via the ECM algorithm

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

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

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

# Maximum likelihood estimation via the ECM algorithm

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

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

with

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

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

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

# ECM for the SNMoE: E-Step

**E-Step** calculates the $Q$-function

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

with

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

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

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

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

**The Iteratively Reweighted Least Squares (IRLS) algorithm:**

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

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

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

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

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

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

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

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

# The $t$ mixture of experts model

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

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

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

# The $t$ mixture of experts model

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

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

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

# The $t$ mixture of experts model

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

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

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

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

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

# MLE of the TMoE model

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

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

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

# MLE of the TMoE model

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

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

where

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

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

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

# MLE of the TMoE model: E-Step

**E-Step** Calculate the $Q$-function:

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

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

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

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

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

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

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

# MLE of the TMoE model: M-Step

**M-Step 1** Calculate $\boldsymbol{\alpha}^{(m+1)}$ by maximizing $Q_1(\boldsymbol{\alpha}; \boldsymbol{\Psi}^{(m)})$ w.r.t $\boldsymbol{\alpha}$. $\Rightarrow$ Iteratively via IRLS (16) as for the mixture of SNMoE.

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

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

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

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

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

# The skew $t$ mixture of experts model

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

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

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

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

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

# The skew $t$ mixture of experts (STMoE) model

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

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

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

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

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

# Representation of the STMoE model

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

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

  is said to follow the STMoE distribution

- **Hierarchical representation**

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

  The variables $U_i$ and $W_i$ are hidden in this hierarchical representation

# Identifiability of the STMoE model

Ordered, initialized, and irreducible STMoEs are identifiable:

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

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

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

# MLE via the ECM algorithm

- Maximize the observed-data log-likelihood:

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

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

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

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

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

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

# MLE via the ECM algorithm: E-Step

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

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

where

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

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

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

# MLE via the ECM algorithm: E-Step

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

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

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

# MLE via the ECM algorithm: M-Step

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

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

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

- **CM-Step 3** Update the skewness parameters $\lambda_k$ by solving the following equation:

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

- **CM-Step 4** Update the degree of freedom $\nu_k$ by solving of the following equation:

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

# Prediction, clustering and model selection

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

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

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

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

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

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

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

# Illustation on Bishop's data set

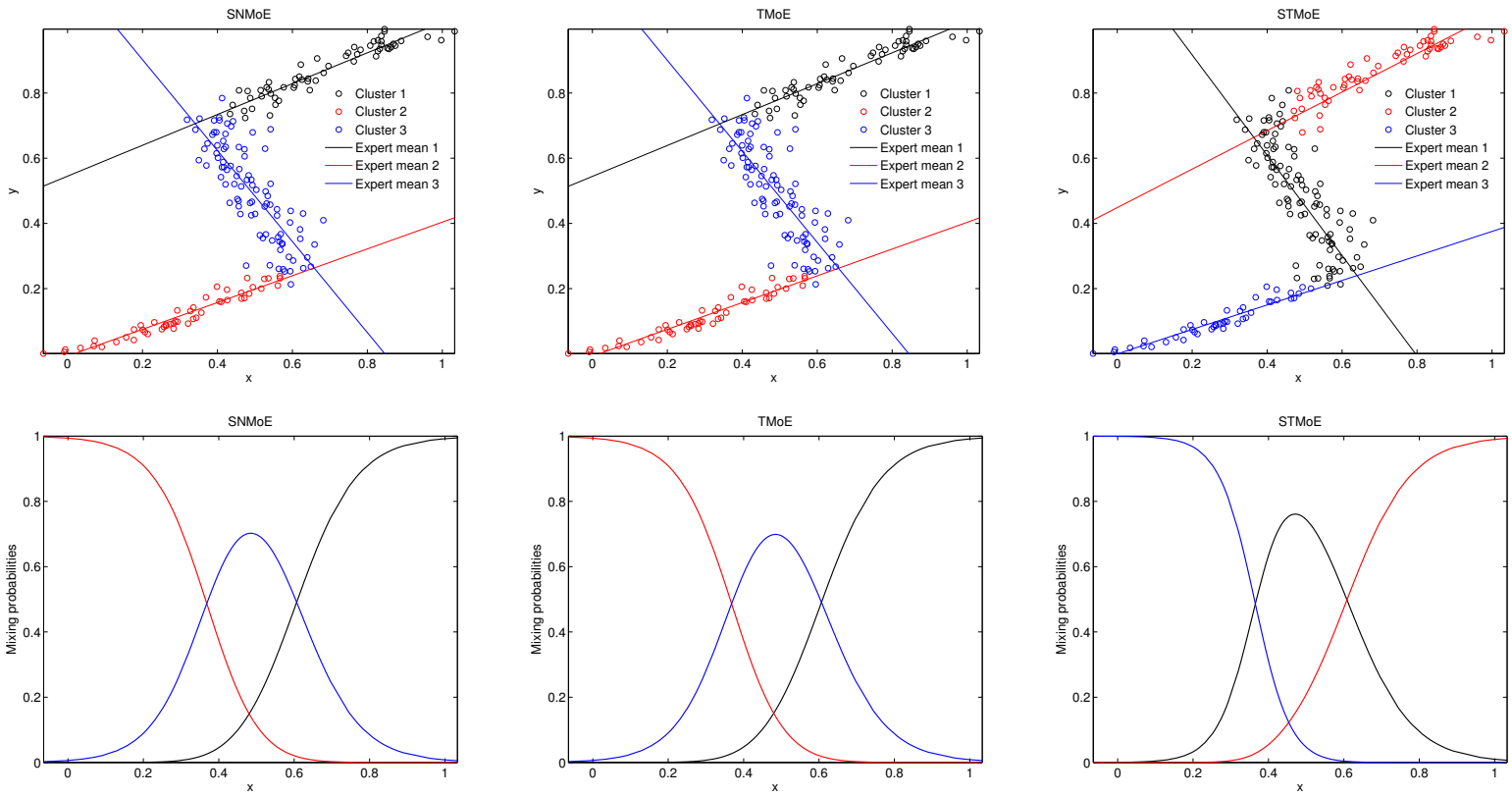


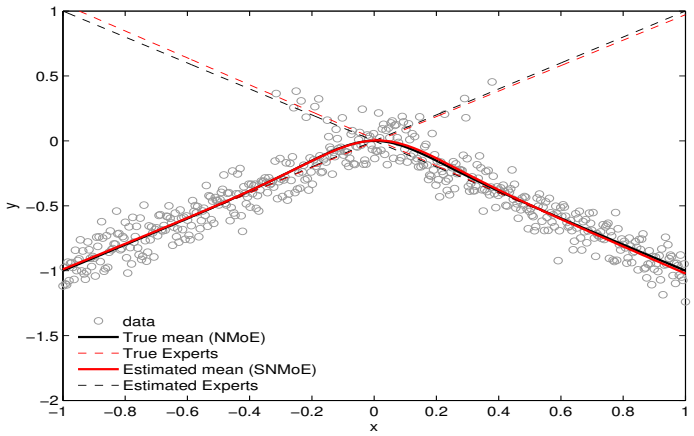

Figure: Fitting the the non-normal mixture of experts models (SNMoE, TNMoE, STMoE) to the toy data set analyzed in Bishop and Svensén (2003): $n = 250$ values of input variables $x_i$ generated uniformly in $(0, 1)$ and output variables $y_i$ generated as $y_i = x_i + 0.3 \sin(2\pi x_i) + \epsilon_i$,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

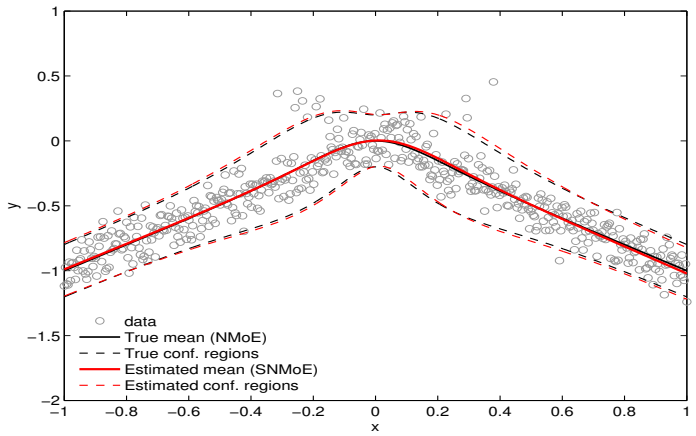
# Experiments: Robustness of the NNMoE

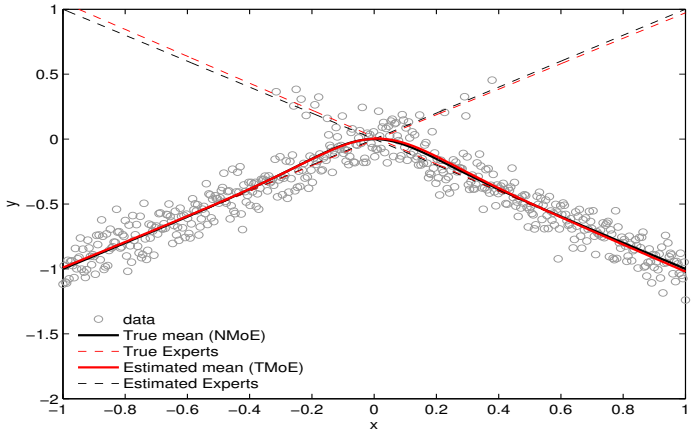
Experimental protocol as in Nguyen and McLachlan (2014)

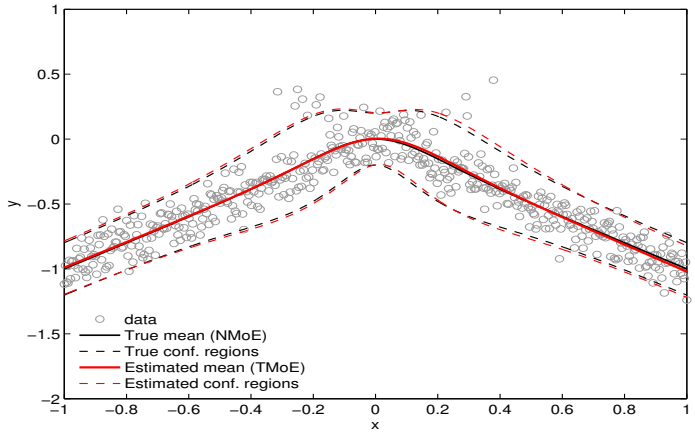
| Model | Outliers | 0% | 1% | 2% | 3% | 4% | 5% |
|---|---|---|---|---|---|---|---|
| NMoE | NMoE | 0.0001783 | 0.001057 | 0.001241 | 0.003631 | 0.013257 | 0.028966 |
| | SNMoE | 0.0001798 | 0.003479 | 0.004258 | 0.015288 | 0.022056 | 0.028967 |
| | TMoE | 0.0001685 | 0.000566 | 0.000464 | 0.000221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

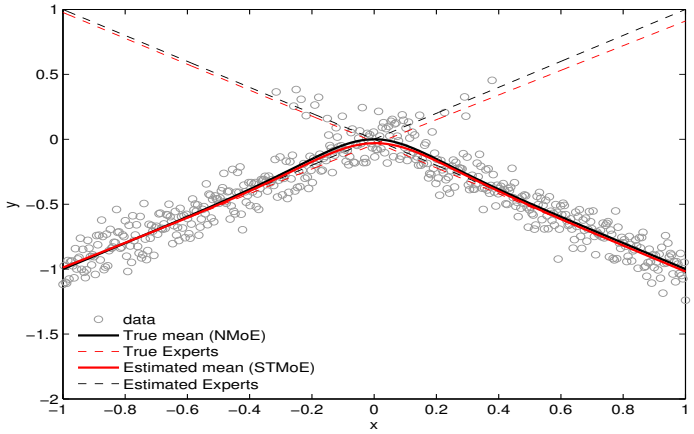
Table: MSE between the estimated mean function and the true one

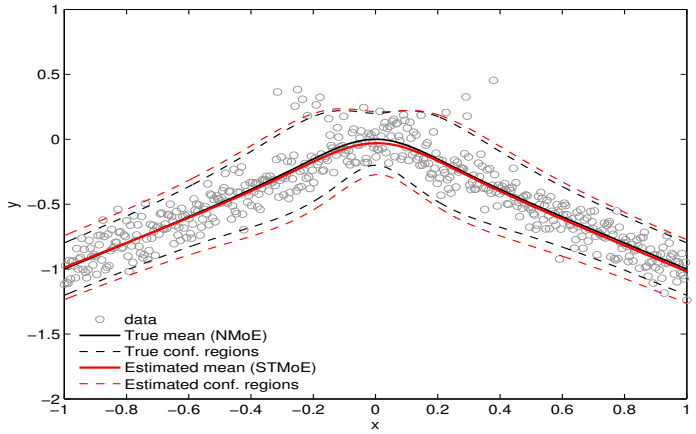

data
True mean (NMoE)
True Experts
Estimated mean (SNMoE)
Estimated Experts
x
y


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


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


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


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


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

# Robustness of the NNMoE

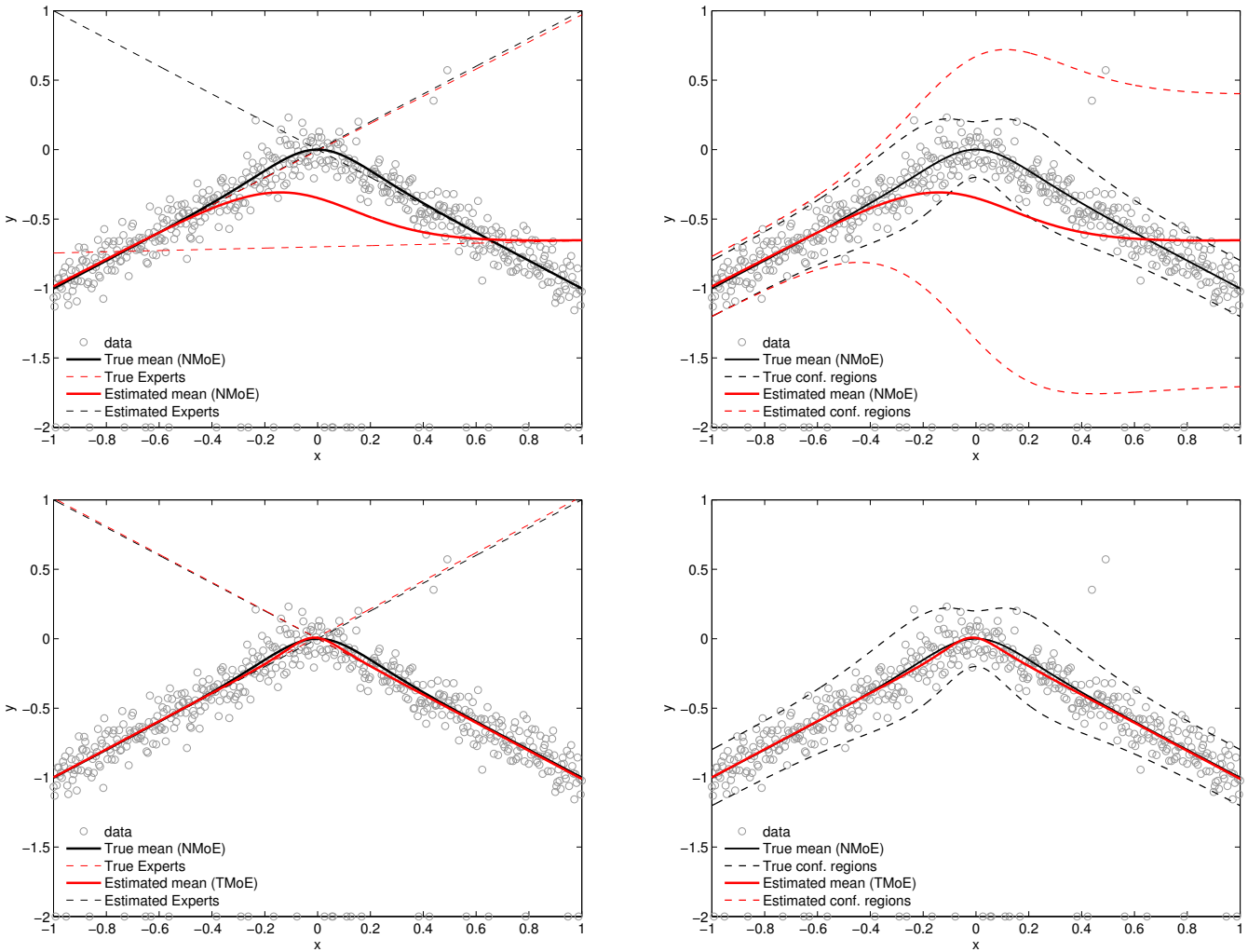



Figure: Fitted MoE to $n = 500$ observations generated according to the NMoE with $5\%$ of outliers $(x; y = -2)$: NMoE fit (top), TMoE fit (bottom).

# Robustness of the NNMoE

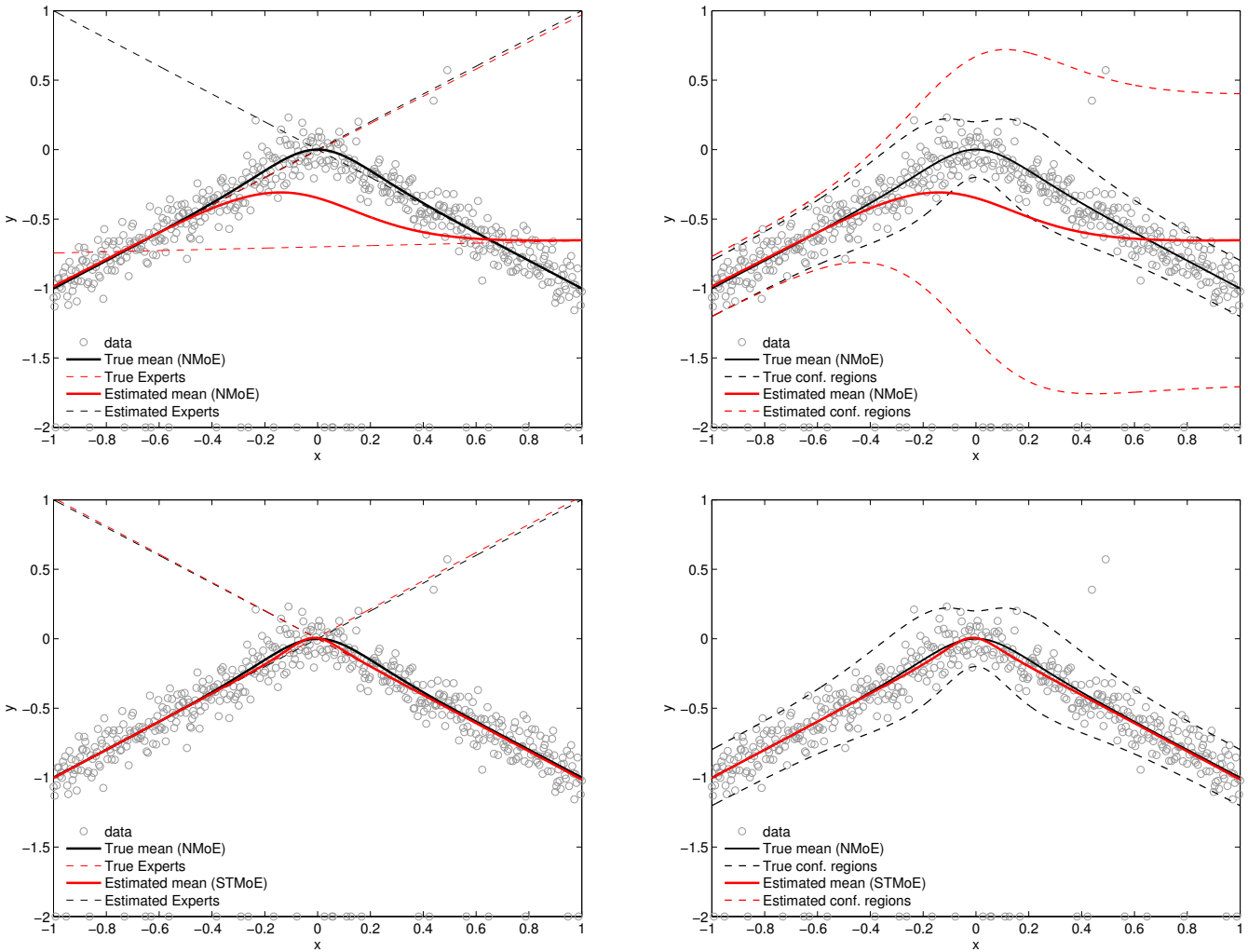



Figure: Fitted MoE to $n = 500$ observations generated according to the NMoE with $5\%$ of outliers $(x; y = -2)$: NMoE fit (top), STMoE fit (bottom).

# Experiments

## Application to two real-world data sets

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

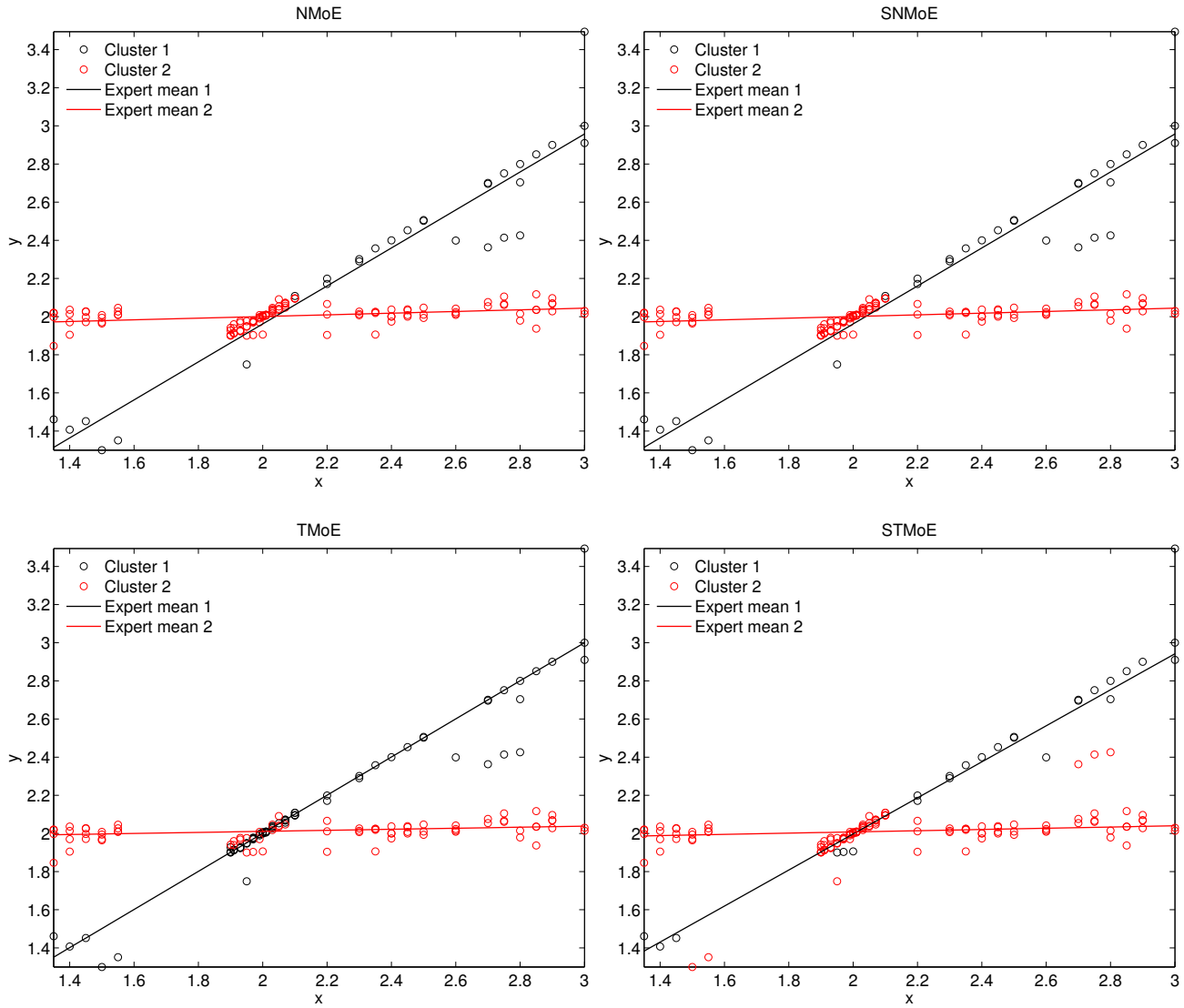



Figure: Fitting the MoLE to the tone data set studied by Bai et al. (2012) and Song et al. (2014) by using robust regression mixture models based on, respectively, the $t$ distribution and the Laplace distribution: $n = 150$ pairs of "tuned" predictors ($x$), and their corresponding "strech ratio" responses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

## Model selection

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

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

## Model's Robustness

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

# Robustness to outliers

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

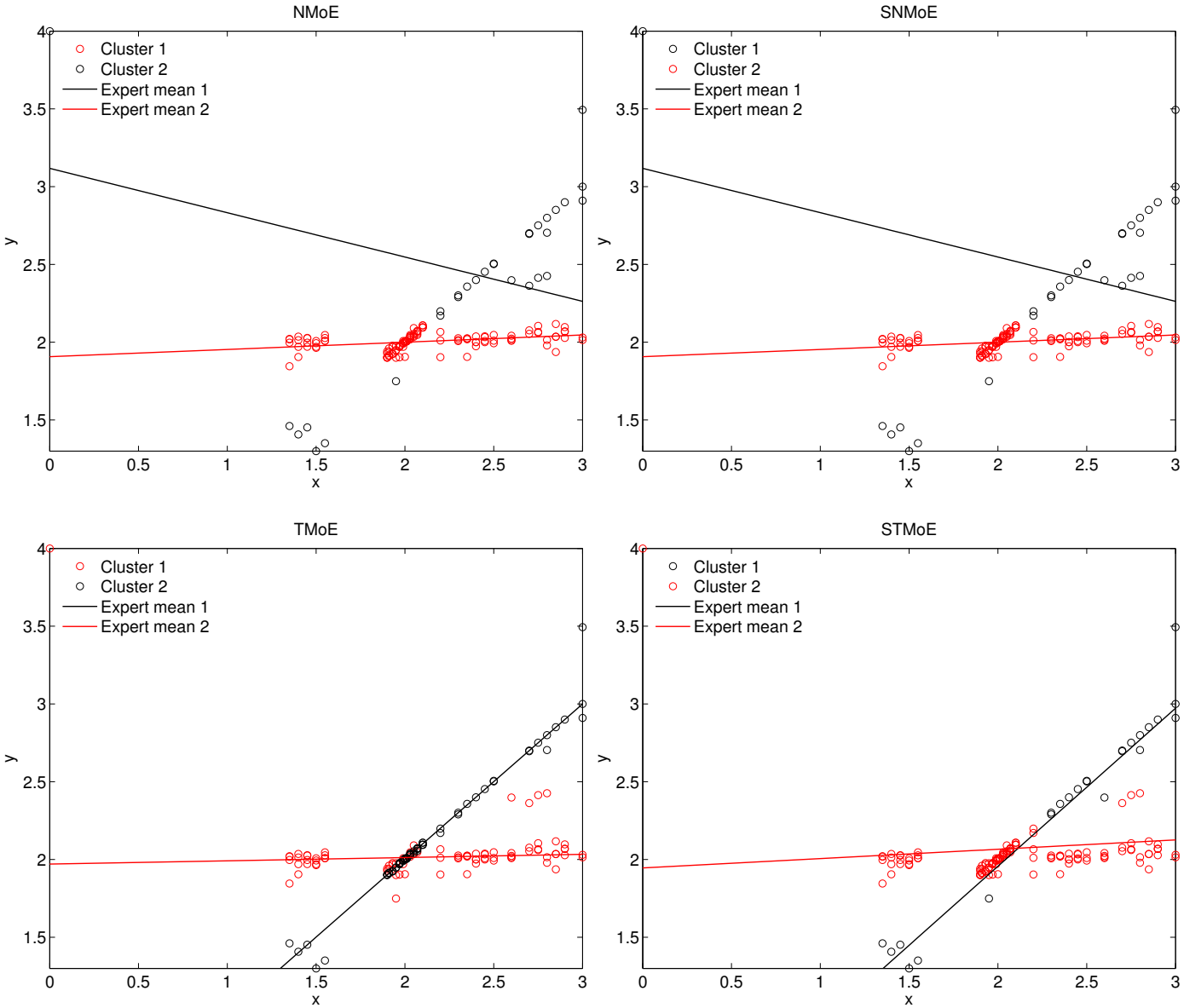



Figure: Fitting MoLE to the tone data set with ten added outliers $(0, 4)$.

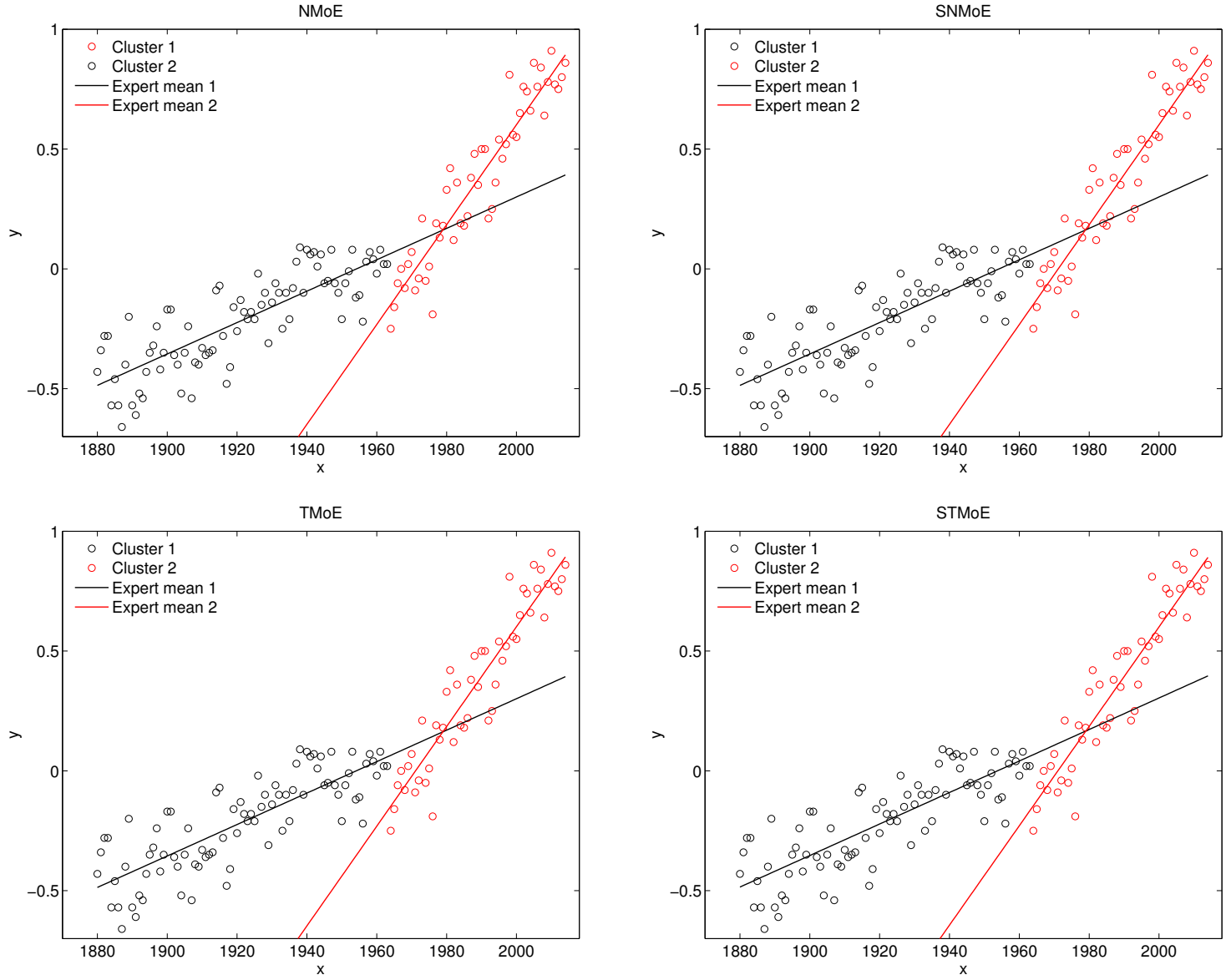



Figure: Fitting the MoLE models to the temperature anomalies data set.

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

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

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

# Some references


- Hien D. Nguyen and Faicel Chamroukhi. Practical and theoretical aspects of mixture-of-experts modeling: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, pages e1246–n/a, Feb 2018. ISSN 1942-4795. doi: 10.1002/widm.1246. URL http://dx.doi.org/10.1002/widm.1246
- F. Chamroukhi. Skew $t$ mixture of experts. *Neurocomputing - Elsevier*, 266:390–408, 2017. URL https://chamroukhi.users.lmno.cnrs.fr/papers/STMoE.pdf
- F. Chamroukhi. Robust mixture of experts modeling using the $t$-distribution. *Neural Networks - Elsevier*, 79:20–36, 2016c. URL https://chamroukhi.users.lmno.cnrs.fr/papers/TMoE.pdf
- Faicel Chamroukhi. Piecewise regression mixture for simultaneous functional data clustering and optimal segmentation. *Journal of Classification*, 33(3):374–411, 2016d. URL https://chamroukhi.users.lmno.cnrs.fr/papers/Chamroukhi-PWRM-JournalClassif-2016.pdf
- F. Chamroukhi. Skew-normal mixture of experts. In *The International Joint Conference on Neural Networks (IJCNN)*, Vancouver, Canada, July 2016b
- F. Chamroukhi. Non-normal mixtures of experts. *arXiv:1506.06707*, July 2015. URL http://arxiv.org/pdf/1506.06707.pdf. Report (61 pages)
- F. Chamroukhi. Unsupervised learning of regression mixture models with unknown number of components. *Journal of Statistical Computation and Simulation*, 86:2308 – 2334, 2016a. URL https://chamroukhi.users.lmno.cnrs.fr/papers/Chamroukhi-JSCS-2015.pdf
- F. Chamroukhi, D. Trabelsi, S. Mohammed, L. Oukhellou, and Y. Amirat. Joint segmentation of multivariate time series with hidden process regression for human activity recognition. *Neurocomputing*, 120:633–644, November 2013b. URL https://chamroukhi.users.lmno.cnrs.fr/papers/chamroukhi_et_al_neucomp2013b.pdf
- F. Chamroukhi, H. Glotin, and A. Samé. Model-based functional mixture discriminant analysis with hidden process regression for curve classification. *Neurocomputing*, 112:153–163, 2013a. URL https://chamroukhi.users.lmno.cnrs.fr/papers/chamroukhi_et_al_neucomp2013a.pdf
- A. Samé, F. Chamroukhi, Gérard Govaert, and P. Aknin. Model-based clustering and segmentation of time series with changes in regime. *Advances in Data Analysis and Classification*, 5:301–321, 2011. URL https://chamroukhi.users.lmno.cnrs.fr/papers/adac-2011.pdf
- F. Chamroukhi, A. Samé, G. Govaert, and P. Aknin. A hidden process regression model for functional data description. application to curve discrimination. *Neurocomputing*, 73(7-9):1210–1221, 2010. URL https://chamroukhi.users.lmno.cnrs.fr/papers/chamroukhi_neucomp_2010.pdf
- F. Chamroukhi, A. Samé, G. Govaert, and P. Aknin. Time series modeling by a regression approach based on a latent process. *Neural Networks*, 22(5-6):593–602, 2009. URL https://chamroukhi.users.lmno.cnrs.fr/papers/Chamroukhi_Neural_Networks_2009.pdf

# References I


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

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

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

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

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

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

F. Chamroukhi. Non-normal mixtures of experts. *arXiv:1506.06707*, July 2015. URL `http://arxiv.org/pdf/1506.06707.pdf`. Report (61 pages).

F. Chamroukhi. Unsupervised learning of regression mixture models with unknown number of components. *Journal of Statistical Computation and Simulation*, 86:2308 – 2334, 2016a. URL `https://chamroukhi.users.lmno.cnrs.fr/papers/Chamroukhi-JSCS-2015.pdf`.

F. Chamroukhi. Skew-normal mixture of experts. In *The International Joint Conference on Neural Networks (IJCNN)*, Vancouver, Canada, July 2016b.

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

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

# References II


F. Chamroukhi, A. Samé, G. Govaert, and P. Aknin. Time series modeling by a regression approach based on a latent process. *Neural Networks*, 22(5-6):593–602, 2009. URL https://chamroukhi.users.lmno.cnrs.fr/papers/Chamroukhi_Neural_Networks_2009.pdf.

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

F. Chamroukhi, H. Glotin, and A. Samé. Model-based functional mixture discriminant analysis with hidden process regression for curve classification. *Neurocomputing*, 112:153–163, 2013a. URL https://chamroukhi.users.lmno.cnrs.fr/papers/chamroukhi_et_al_neucomp2013a.pdf.

F. Chamroukhi, D. Trabelsi, S. Mohammed, L. Oukhellou, and Y. Amirat. Joint segmentation of multivariate time series with hidden process regression for human activity recognition. *Neurocomputing*, 120:633–644, November 2013b. URL https://chamroukhi.users.lmno.cnrs.fr/papers/chamroukhi_et_al_neucomp2013b.pdf.

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

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

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

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

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

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

# References III


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

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

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

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

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

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

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

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

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

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

A. Samé, F. Chamroukhi, Gérard Govaert, and P. Aknin. Model-based clustering and segmentation of time series with changes in regime. *Advances in Data Analysis and Classification*, 5:301–321, 2011. URL `https://chamroukhi.users.lmno.cnrs.fr/papers/adac-2011.pdf`.

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

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

Thank you!