Robust non-normal mixture of experts

Faicel Chamroukhi

Laboratoire des Sciences de l’Information et des Systèmes
LSIS UMR CNRS 7296
Laboratoire de mathématiques Paul Painlevé
LPP UMR CNRS 8524

The 8th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2015)

Session: Mixture models for non-vectorial data

December 13, 2015
Regression data with atypical features

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

- Heterogeneous regression data
- Data with possible atypical observations
- Data with possibly asymmetric and heavy-tailed distributions

Objectives
- Derive robust models to fit at best the data
- Deal with other possible features like skewness, heavy tails
Scientific context

- Analysis of clustered regression data
  - exploratory analysis
  - decisional analysis: make decision and prediction for future data

Topics

- density estimation
- regression
- clustering/segmentation

Mixture modeling framework

- Mixture density: \( f(x) = \sum_{k=1}^{K} \mathbb{P}(z = k) f(x|z = k) = \sum_{k=1}^{K} \pi_k f_k(x) \)
- Generative model: \( z \sim \mathcal{M}(1; \pi_1, \ldots, \pi_k) \) then \( x|z \sim f(x|z) \)
- Derive a robust model for fitting from such data
Outline

1. Introduction and context
2. Related work
3. Non-normal mixtures of experts
4. Experiments
5. Conclusion and perspectives
Related work

Observed pairs of data \((x, y)\) where \(y \in \mathbb{R}\) is the response for some covariate \(x \in \mathbb{R}^p\) governed by a hidden categorical random variable \(Z\)

### Mixture of regressions

\[
f(y|x; \Psi) = \sum_{k=1}^{K} \pi_k f_k(y|x; \Psi_k)
\]

- Bai et al. (2012); Wei (2012): robust regression mixture based on the \(t\) distribution
- Ingrassia et al. (2012): Cluster-weighted modeling based on the \(t\) distribution
- Song et al. (2014): robust regression mixture based on the Laplace distribution

\(\rightarrow\) A mixture of experts (MoE) framework (Jacobs et al., 1991; Jordan and Jacobs, 1994)
Mixture of Experts (MoE) modeling framework

- Observed pairs of data \((x, y)\) where \(y \in \mathbb{R}\) is the response for some covariate \(x \in \mathbb{R}^p\) governed by a hidden categorical random variable \(Z\)

- Mixture of experts (MoE) (Jacobs et al., 1991; Jordan and Jacobs, 1994):

\[
f(y|x; \Psi) = \sum_{k=1}^{K} \pi_k(r; \alpha) f_k(y|x; \Psi_k)
\]

Gating function of some predictors \(r \in \mathbb{R}^q\):

\[
\pi_k(r; \alpha) = \frac{\exp(\alpha_k^T r)}{\sum_{\ell=1}^{K} \exp(\alpha_{\ell}^T r)}
\]

MoE for regression usually use normal experts \(f_k(y|x; \Psi_k)\)

Objectives

- Overcome (well-known) limitations of modeling with the normal distribution.
  
  \(\leftarrow \) Not adapted for a set of data containing a group or groups of observations with asymmetric behavior, heavy tails or atypical observations
Non-normal mixtures of experts

- Li et al. (2010): Bayesian mixture of asymmetric \( t \) experts
- Nguyen and McLachlan (2016): Mixture of Laplace experts

Non-normal mixtures of experts (NNMoE)

1. the \( t \) MoE (TMoE) (Robustness, heavy tails)
2. the skew-\( t \) MoE (STMoE) (skewness, robustness, heavy tails)

Correspond to extensions of the mixture of \( t \) distributions (Mclachlan and Peel, 1998) for regression (Bai et al., 2012; Wei, 2012) and the mixture of skew \( t \) distributions (Lin et al., 2007a) to the MoE modeling framework

\[
\pi_k = [0.4, 0.6], \mu_k = [-1, 2]; \sigma_k = [1, 1]; \nu_k = [3, 7]; \lambda_k = [14, -12];
\]
The skew $t$ mixture of experts (STMoE) model

- A $K$-component mixture of skew $t$ experts (STMoE) is defined by:
  \[
  f(y | r, x; \Psi) = \sum_{k=1}^{K} \pi_k(r; \alpha) \text{ST}(y; \mu(x; \beta_k), \sigma_k^2, \lambda_k, \nu_k)
  \]

- $k$th expert: has skew $t$ distribution (Azzalini and Capitanio, 2003):
  \[
  f(y | x; \mu(x; \beta_k), \sigma^2, \lambda, \nu) = \frac{2}{\sigma} t_\nu(d_y(x)) \frac{1}{T_{\nu+1}} \left( \lambda d_y(x) \sqrt{\frac{\nu + 1}{\nu + d_y^2(x)}} \right)
  \]
  where $d_y(x) = \frac{y - \mu(x; \beta_k)}{\sigma}$.

Model characteristics

$\leftarrow$ For $\{\nu_k\} \to \infty$, the STMoE reduces to the SNMoE

$\leftarrow$ For $\{\lambda_k\} \to 0$, the STMoE reduces to the TMoE.

$\leftarrow$ For $\{\nu_k\} \to \infty$ and $\{\lambda_k\} \to 0$, it approaches the NMoE.

$\leftarrow$ The STMoE is flexible as it generalizes the (skew)-normal and $t$ MoE 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 $E_i \sim \text{SN}(\lambda_{zi})$ and $W_i \sim \text{Gamma}(\frac{\nu_{zi}}{2}, \frac{\nu_{zi}}{2})$, and $x_i$ and $r_i$ are given covariates. A variable $Y_i$ having the following representation:

$$Y_i = \mu(x_i; \beta_{zi}) + \sigma_{zi} \frac{E_i}{\sqrt{W_i}}$$

is said to follow the STMoE distribution.

- **Hierarchical representation**

$$Y_i | u_i, w_i, Z_{ik} = 1, x_i \sim \mathcal{N}\left(\mu(x_i; \beta_k) + \delta_k |u_i|, \frac{1 - \delta_k^2}{w_i} \sigma_k^2\right),$$

$$U_i | w_i, Z_{ik} = 1 \sim \mathcal{N}\left(0, \frac{\sigma_k^2}{w_i}\right),$$

$$W_i | Z_{ik} = 1 \sim \text{Gamma}\left(\frac{\nu_k}{2}, \frac{\nu_k}{2}\right)$$

$$Z_i | r_i \sim \text{Mult}(1; \pi_1(r_i; \alpha), \ldots, \pi_K(r_i; \alpha)).$$

The variables $U_i$ and $W_i$ are hidden in this hierarchical representation.
Identifiability of the STMoE model

\( f(.; \Psi) \) is identifiable when \( f(.; \Psi) = f(.; \Psi^*) \) if and only if \( \Psi = \Psi^* \).

Ordered, initialized, and irreducible STMoEs are identifiable:

- Ordered implies that there exist a certain ordering relationship such that
  \( (\beta_1^T, \sigma_1^2, \lambda_1, \nu_1)^T < \ldots < (\beta_K^T, \sigma_K^2, \lambda_K, \nu_K)^T \);

- initialized implies that \( \alpha_K \) is the null vector, as assumed in the model

- irreducible implies that if \( k \neq k' \), then one of the following conditions holds:
  \( \beta_k \neq \beta_{k'}, \sigma_k \neq \sigma_{k'}, \lambda_k \neq \lambda_{k'} \) or \( \nu_k \neq \nu_{k'} \).

\( \Rightarrow \) 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 \( \{ \text{ST}(y; \mu(x; \beta_1), \sigma_1^2, \lambda_1, \nu_1), \ldots, \text{ST}(y; \mu(x; \beta_{4K}), \sigma_{4K}^2, \lambda_{4K}, \nu_{4K}) \} \) contains \( 4K \) linearly independent functions of \( y \), for any \( 4K \) distinct quadruplet \( (\mu(x; \beta_k), \sigma_k^2, \lambda_k, \nu_k) \) for \( k = 1, \ldots, 4K \).

\( \Rightarrow \) Thus, via Lemma 2 of Jiang and Tanner (1999) we have any ordered and initialized irreducible STMoE is identifiable.
Parameter estimation via the ECM algorithm

- Parameter vector: \( \Psi = (\alpha_1^T, \ldots, \alpha_{K-1}^T, \theta_1^T, \ldots, \theta_K^T, \nu_1, \ldots, \nu_K)^T \) where \( \theta_k = (\beta_k^T, \sigma_k^2, \lambda_k)^T \)

- Maximize the observed-data log-likelihood:

\[
\log L(\Psi) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k(r_i; \alpha)ST(y; \mu(x_i; \beta_k), \sigma_k^2, \lambda_k, \nu_k) \cdot
\]

\( \leftrightarrow \) iteratively by the ECM algorithm (Meng and Rubin, 1993)

- The complete-data log-likelihood:

\[
\log L_c(\Psi) = \log L_{1c}(\alpha) + \sum_{k=1}^{K} \left[ \log L_{2c}(\theta_k) + \log L_{3c}(\nu_k) \right]
\]

where

\[
\log L_{1c}(\alpha) = \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{ik} \log \pi_k(r_i; \alpha),
\]

\[
\log L_{2c}(\theta_k) = \sum_{i=1}^{n} Z_{ik} \left[ -\log(2\pi\sigma_k^2) - \frac{1}{2} \log(1 - \delta_k^2) - \frac{W_i d_{ik}^2}{2(1 - \delta_k^2)} + \frac{W_i U_i \delta_k}{d_{ik}} - \frac{W_i U_i^2}{2(1 - \delta_k^2)\sigma_k^2} \right]
\]

\[
\log L_{3c}(\nu_k) = \sum_{i=1}^{n} Z_{ik} \left[ -\log \Gamma \left( \frac{\nu_k}{2} \right) + \left( \frac{\nu_k}{2} \right) \log \left( \frac{\nu_k}{2} \right) + \left( \frac{\nu_k}{2} \right) \log(W_i) - \left( \frac{\nu_k}{2} \right) W_i \right].
\]
MLE via the ECM algorithm: E-Step

- **E-Step** Calculates the conditional expectation of the complete-data log-likelihood, given the observed data \( \{y_i, x_i, r_i\}_{i=1}^n \) and a current parameter estimation \( \Psi^{(m)} \):

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

where

\[
Q_1(\alpha; \Psi^{(m)}) = \sum_{i=1}^n \sum_{k=1}^K \tau_{ik}^{(m)} \log \pi_k(r_i; \alpha),
\]

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

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

1. **E-Step:** requires the following conditional expectations:

\[
\begin{align*}
\tau_{ik}^{(m)} &= \mathbb{E}_{\Psi^{(m)}} [Z_{ik} | y_i, x_i, r_i], \\
w_{ik}^{(m)} &= \mathbb{E}_{\Psi^{(m)}} [W_i | y_i, Z_{ik} = 1, x_i, r_i], \\
e_{1,ik}^{(m)} &= \mathbb{E}_{\Psi^{(m)}} [W_i U_i | y_i, Z_{ik} = 1, x_i, r_i], \\
e_{2,ik}^{(m)} &= \mathbb{E}_{\Psi^{(m)}} [W_i U_i^2 | y_i, Z_{ik} = 1, x_i, r_i], \\
e_{3,ik}^{(m)} &= \mathbb{E}_{\Psi^{(m)}} [\log(W_i) | y_i, Z_{ik} = 1, x_i, r_i].
\end{align*}
\]

\( \leftarrow \) Calculated analytically except \( e_{3,ik}^{(m)} \) \( \leftarrow \) I adopted a one-step-late (OSL) approach as in Lee and McLachlan (2014)

\( \leftarrow \) Note that Lee and McLachlan (2015) presented an exact series-based truncation approach for the multivariate skew t mixture models

2. **CM-Steps:** \( \Psi^{(m+1)} = \arg \max_{\Psi \in \Omega} Q(\Psi; \Psi^{(m)}) \)
CM-Step 1 update the mixing parameters $\alpha^{(m+1)}$ by:

$$\alpha^{(m+1)} = \arg \max_{\alpha} \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(m)} \log \pi_k(r_i; \alpha)$$

$\rightarrow$ Iteratively Reweighted Least Squares (IRLS) algorithm

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

- A convex optimization problem
- Analytic calculation of the Hessian and the gradient

CM-Step 2 Update the regression params $(\beta_k^{T(m+1)}, \sigma_k^{2(m+1)})$: For the polynomial regressors: $\mu(x; \beta_k) = \beta_k^T x$ we have analytic weighted regressions updates:

$$\beta_k^{(m+1)} = \left[ \sum_{i=1}^{n} \tau_{ik}^{(q)} w_{ik}^{(m)} x_i x_i^T \right]^{-1} \sum_{i=1}^{n} \tau_{ik}^{(q)} \left( w_{ik}^{(m)} y_i - e_{1,ik}^{(m)} \right) x_i ,$$

$$\sigma_k^{2(m+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(m)} \left( y_i - \beta_k^{T(m+1)} x_i \right)^2 - 2 \delta_k^{(m+1)} e_{1,ik}^{(m)} (y_i - \beta_k^{T(m+1)} x_i) + e_{2,ik}^{(m)} }{2 \left( 1 - \delta_k^{2(m)} \right) \sum_{i=1}^{n} \tau_{ik}^{(m)} }$$
ECM algorithm for the STMoE: M-Step

- **CM-Step 3** Update the skewness parameters $\lambda_k$ as solution of

$$
\delta_k (1 - \delta_k^2) \sum_{i=1}^n \tau^{(m)}_{ik} + (1 + \delta_k^2) \sum_{i=1}^n \tau^{(m)}_{ik} \frac{d^{(m+1)}_{ik} e^{(m)}_{1,ik}}{\sigma_{k}^{(m+1)}} - \delta_k \sum_{i=1}^n \tau^{(m)}_{ik} \left[ w^{(m)}_{ik} d^{2 (m+1)}_{ik} + \frac{e^{(m)}_{2,ik}}{\sigma_{k}^{2 (m+1)}} \right] = 0.
$$

- **CM-Step 4** Update the degree of freedom $\nu_k$ as solution of:

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

$\hookrightarrow$ Use a root finding algorithm, such as Brent’s method (Brent, 1973)
ECM algorithm for the STMoE: M-Step

- **CM-Step 3** Update the skewness parameters $\lambda_k$ as solution of

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

- **CM-Step 4** Update the degree of freedom $\nu_k$ as solution of:

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

$\rightarrow$ Use a root finding algorithm, such as Brent's method (Brent, 1973)

- **Prediction** Predicted response: $\hat{y} = \mathbb{E}_{\hat{\Psi}}(Y|r, x)$ for $\hat{\nu}_k > 1$:

$$
\mathbb{E}_{\hat{\Psi}}(Y|r, x) = \sum_{k=1}^{K} \pi_k(r; \hat{\alpha}) \left( \hat{\beta}_k^T x + \hat{\sigma}_k \hat{\delta}_k \xi(\hat{\nu}_k) \right) \text{ where } \xi(\hat{\nu}_k) = \sqrt{\frac{\hat{\nu}_k}{\pi}} \Gamma \left( \frac{\frac{\hat{\nu}_k}{2} - 1}{\hat{\nu}_k} \right)
$$

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

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

- **Model selection** The value of $(K, p)$ can be computed by using BIC, ICL

Number of free parameters: $\eta_{\Psi} = K(p + 6) - 2$ for the STMoE model.
Temperature anomalies data set

- Data have been analyzed earlier by Hansen et al. (1999, 2001) and recently by Nguyen and McLachlan (2016) by using Laplace mixture of linear experts
- \( n = 135 \) yearly measurements of the global annual temperature anomalies for the period of 1882 – 2012.

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

- Data have been analyzed earlier by Hansen et al. (1999, 2001) and recently by Nguyen and McLachlan (2016) by using Laplace mixture of linear experts.

- $n = 135$ yearly measurements of the global annual temperature anomalies for the period of 1882 – 2012.

![Data Analysis Diagrams](image)

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 (2016) who used a Laplace mixture of linear experts.

Model selection: Except the result provided by AIC for the NMoE model which overestimates the number of components, all the others results provide evidence for two components in the data.

<table>
<thead>
<tr>
<th>K</th>
<th>NMoE</th>
<th>SNMoE</th>
<th>TMoE</th>
<th>STMoE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BIC</td>
<td>AIC</td>
<td>ICL</td>
<td>BIC</td>
</tr>
<tr>
<td>1</td>
<td>46.0623</td>
<td>50.4202</td>
<td>46.0623</td>
<td>43.6096</td>
</tr>
<tr>
<td>2</td>
<td>79.9163</td>
<td>91.5374</td>
<td>79.6241</td>
<td>75.0116</td>
</tr>
<tr>
<td>3</td>
<td>71.3963</td>
<td>90.5374</td>
<td>58.4874</td>
<td>63.9254</td>
</tr>
<tr>
<td>4</td>
<td>66.7276</td>
<td>92.8751</td>
<td>54.7524</td>
<td>55.4731</td>
</tr>
<tr>
<td>5</td>
<td>59.5100</td>
<td>92.9206</td>
<td>51.2429</td>
<td>45.3469</td>
</tr>
</tbody>
</table>

Table: Choosing the number of expert components $K$ for the temperature anomalies data by using the information criteria BIC, AIC, and ICL.
Tone perception data set

- Recently studied by Bai et al. (2012) and Song et al. (2014) by using, respectively, robust $t$ regression mixture and Laplace regression mixture.

- Data consist of $n = 150$ pairs of “tuned” variables, considered here as predictors ($x$), and their corresponding “strecth ratio” variables considered as responses ($y$).

![Figure: Fitting the MoE models to the tone data set](image)
## Model selection

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

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

Experimental protocol as in Nguyen and McLachlan (2016)

Figure: Fitted MoE to $n = 500$ observations generated according to the NMoE: NMoE fit (top), TMoE fit (middle), STMoE fit (bottom).
Robustness of the NNMoE

Experimental protocol as in Nguyen and McLachlan (2016)

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

The MSE is given by:

$$\frac{1}{n} \sum_{i=1}^{n} \left\| \mathbb{E}_{\Psi}(Y_i \mid r_i, x_i) - \mathbb{E}_{\hat{\Psi}}(Y_i \mid r_i, x_i) \right\|^2$$

for different noise levels.

<table>
<thead>
<tr>
<th>Model</th>
<th>Outliers</th>
<th>0%</th>
<th>1%</th>
<th>2%</th>
<th>3%</th>
<th>4%</th>
<th>5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMoE</td>
<td>0%</td>
<td>0.0001783</td>
<td>0.001057</td>
<td>0.001241</td>
<td>0.003631</td>
<td>0.013257</td>
<td>0.028966</td>
</tr>
<tr>
<td>SNMoE</td>
<td>1%</td>
<td>0.0001798</td>
<td>0.003479</td>
<td>0.004258</td>
<td>0.015288</td>
<td>0.022056</td>
<td>0.028967</td>
</tr>
<tr>
<td>TMoE</td>
<td>2%</td>
<td>0.0001685</td>
<td>0.000566</td>
<td>0.000464</td>
<td>0.000221</td>
<td>0.000263</td>
<td>0.000045</td>
</tr>
<tr>
<td>STMoE</td>
<td>3%</td>
<td>0.0002586</td>
<td>0.000741</td>
<td>0.000794</td>
<td>0.000696</td>
<td>0.000697</td>
<td>0.000626</td>
</tr>
<tr>
<td></td>
<td>4%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table: MSE between the estimated mean function and the true one
Consider 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 \((0, 4)\).

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

→ In this noisy case the \(t\) mixture of regressions fails (is affected severely by the outliers) as showed in Song et al. (2014)
Temporal railway data

- \( n = 562 \) temporal data
- 30 added artificial outliers
Outline

1 Introduction and context
2 Related work
3 Non-normal mixtures of experts
4 Experiments
5 Conclusion and perspectives
Summary

- The STMoE model is suggested for possibly noisy and heterogeneous regression data
- It also dedicated to accommodate regression data with possibly possibly non-symmetric and heavy tailed distribution
- Outputs: density estimation, non-linear regression function approximation and clustering for regression data
- The model selection using information criteria tends to promote using BIC and ICL against AIC

Perspectives

- Here we only considered the MoE in their standard (non-hierarchical) version. One interesting future direction is therefore to extend it to the hierarchical mixture of experts framework (Jordan and Jacobs, 1994).
- Extension to the multiple regression regression setting
Thank you!


References II


