Robust non-normal mixture of experts

Faicel Chamroukhi

Laboratoire des Sciences de l'Information et des Sysèmes
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

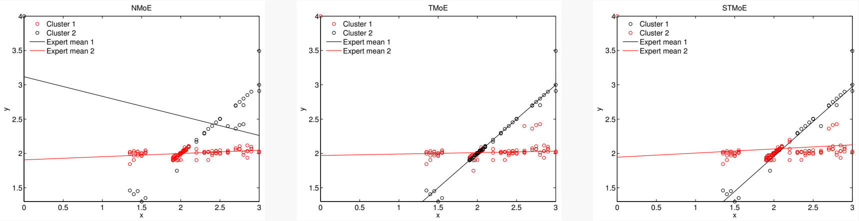


Figure: Fitting MoLE to the tone data set with ten outliers $(0, 4)$.

- 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

  $\hookrightarrow$ exploratory analysis

  $\hookrightarrow$ 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

Related work

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

Mixture of regressions

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

- 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

$\hookrightarrow$ A mixture of experts (MoE) framework (Jacobs et al., 1991; Jordan and Jacobs, 1994)

Mixture of Experts (MoE) modeling framework

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

$$f(y|\boldsymbol{x};\boldsymbol{\Psi}) = \sum_{k=1}^{K} \underbrace{\pi_k(\boldsymbol{r};\boldsymbol{\alpha})}_{\text{Gating network}} \underbrace{f_k(y|\boldsymbol{x};\boldsymbol{\Psi}_k)}_{\text{Experts}}$$

- Gating function of some predictors $\boldsymbol{r} \in \mathbb{R}^q$: $\pi_k(\boldsymbol{r};\boldsymbol{\alpha}) = \frac{\exp(\boldsymbol{\alpha}_k^T \boldsymbol{r})}{\sum_{\ell=1}^{K} \exp(\boldsymbol{\alpha}_\ell^T \boldsymbol{r})}$
- MoE for regression usually use normal experts $f_k(y|\boldsymbol{x};\boldsymbol{\Psi}_k)$

Objectives

- Overcome (well-known) limitations of modeling with the normal distribution.

  $\hookrightarrow$ Not adapted for a set of data containing a group or groups of observations with asymmetric behavior, heavy tails or atypical observations

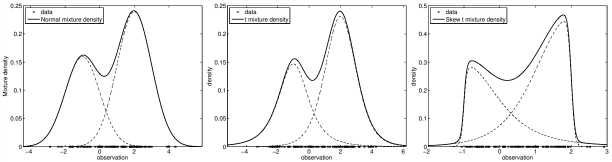
Non-normal mixtures of experts

- 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:

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

- kth expert: has skew t distribution (Azzalini and Capitanio, 2003):

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

where $d_y(\boldsymbol{x}) = \frac{y-\mu(\boldsymbol{x};\boldsymbol{\beta}_k)}{\sigma}$.

Model characteristics

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

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

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

$\hookrightarrow$ 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 $E_i \sim \mathsf{SN}(\lambda_{z_i})$ 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**

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

The variables U_i and W_i are hidden in this hierarchical representation

Identifiability of the STMoE model

$f(.;\boldsymbol{\Psi})$ is identifiable when $f(.;\boldsymbol{\Psi}) = f(.;\boldsymbol{\Psi}^\star)$ if and only if $\boldsymbol{\Psi} = \boldsymbol{\Psi}^\star$.
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 $\boldsymbol{\alpha}_K$ 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.

Parameter estimation via the ECM algorithm

- Parameter vector: $\boldsymbol{\Psi} = (\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_{K-1}^T, \boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_K^T, \nu_1, \ldots, \nu_K)^T$ where $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k^T, \sigma_k^2, \lambda_k)^T$
- Maximize the observed-data log-likelihood:

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

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

- The complete-data log-likelihood:

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

where

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

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

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

MLE via the ECM algorithm: E-Step

- **E-Step** Calculates the 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 $\boldsymbol{\Psi}^{(m)}$:

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

where

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

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

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

Parameter estimation via the ECM algorithm

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

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

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

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

2. CM-Steps: $\boldsymbol{\Psi}^{(m+1)} = \arg\max_{\boldsymbol{\Psi}\in\boldsymbol{\Omega}} Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(m)})$

STоME: ECM algorithm: M-Step

- **CM-Step 1** update the mixing parameters $\boldsymbol{\alpha}^{(m+1)}$ by:

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

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

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

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

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

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

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

ECM algorithm for the STMoE: M-Step

- **CM-Step 3** Update the skewness parameters λ_k as solution of

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

- **CM-Step 4** Update the degree of freedom ν_k as solution of:

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

$\hookrightarrow$ 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 λ_k as solution of

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

- **CM-Step 4** Update the degree of freedom ν_k as solution of:

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

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

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

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

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

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

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

  Number of free parameters: $\eta_{\boldsymbol{\Psi}} = K(p+6)-2$ 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 $1882 - 2012$.

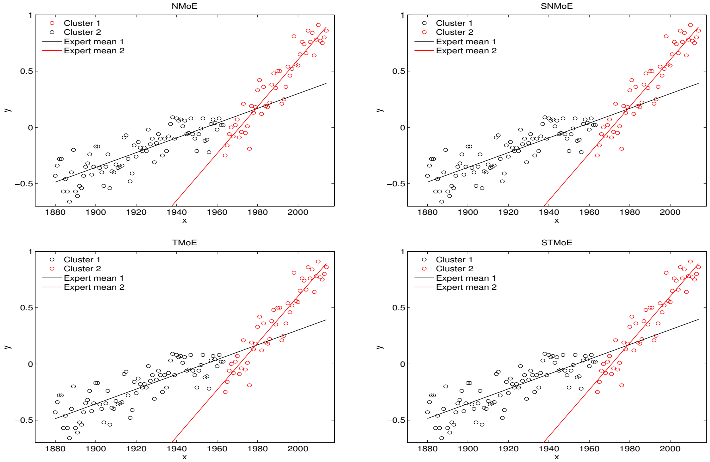


Figure: Fitting the MoLE models to the temperature anomalies data set.

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 $1882 - 2012$.

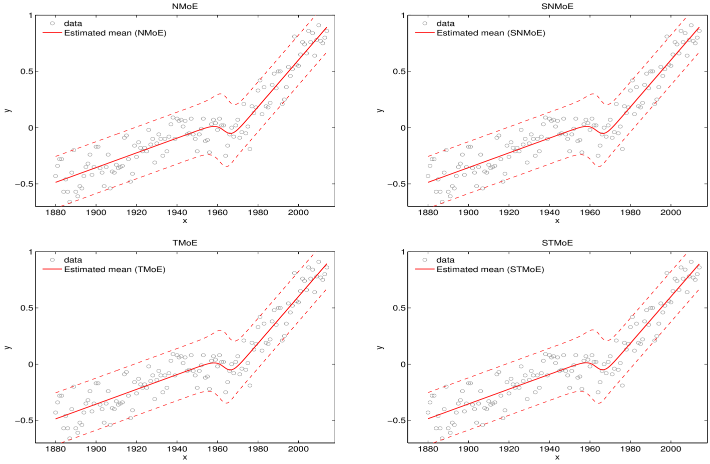


Figure: Fitting the MoLE models to the temperature anomalies data set.

- 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.

	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	79.9163	91.5374	79.6241	75.0116	89.5380	74.7395	74.7960	89.3224	74.5279	69.6382	87.0698	69.3416
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	92.9206	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.

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 "strech ratio" variables considered as responses (y).

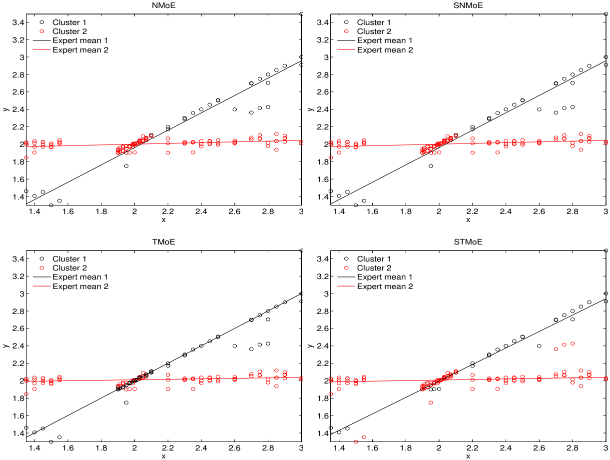


Figure: Fitting the MoE models to the tone data set

Model selection

K	NMoE			SNMoE			TMoE			STMoE		
	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	<u>122.8725</u>	132.8471	<u>102.4049</u>	<u>204.8241</u>	219.8773	186.8415	<u>92.4352</u>	110.4990	<u>82.4552</u>
3	118.1939	137.7630	76.5249	117.7939	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	<u>234.9216</u>	<u>187.7673</u>	77.7092	116.8474	56.3654
5	<u>141.6961</u>	<u>176.3184</u>	<u>123.6550</u>	107.2795	<u>149.4284</u>	96.6832	187.8652	230.0141	164.9629	79.0439	<u>128.7194</u>	67.7485

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)

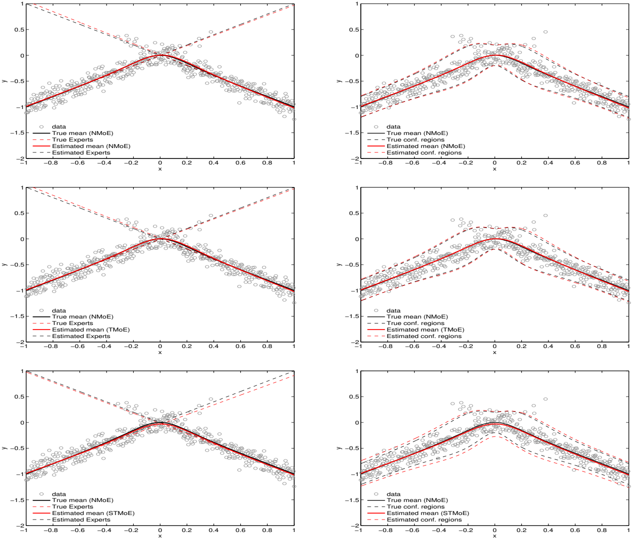


Figure: Fitted MoE to $n = 500$ observations generated according to the NMoE: NMoE fit (top), TMoE fit (middle), STMoE fit (bottom).

Robustness of the NNMoE

Experimental protocol as in Nguyen and McLachlan (2016)

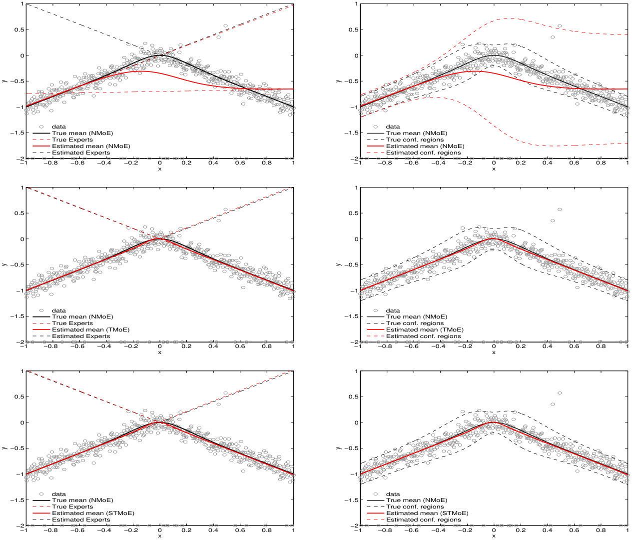


Figure: Fitted MoE to $n = 500$ observations generated according to the NMoE with 5% of outliers $(x; y = -2)$: NMoE fit (top), TMoE fit (middle), STMoE fit (bottom).

Robustness of the NNMoE

MSE $\frac{1}{n}\sum_{i=1}^{n}\|\mathbb{E}_{\boldsymbol{\Psi}}(Y_i|\boldsymbol{r}_i,\boldsymbol{x}_i) - \mathbb{E}_{\hat{\boldsymbol{\Psi}}}(Y_i|\boldsymbol{r}_i,\boldsymbol{x}_i)\|^2$ for different noise levels

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

Tone perception data set (noisy case)

- 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 $(0, 4)$

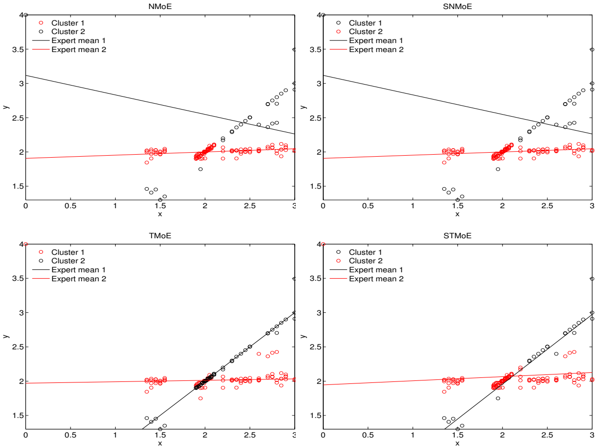
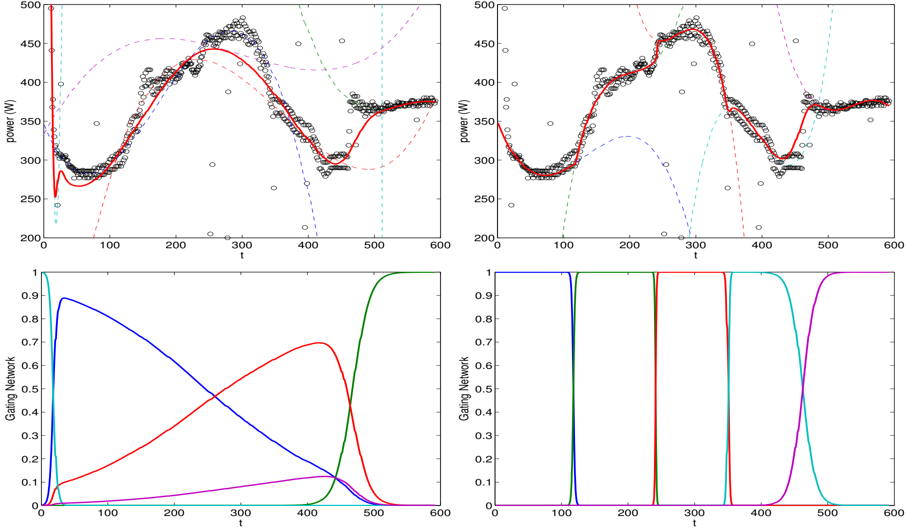


Figure: Fitting MoLE to the tone data set with ten added outliers $(0, 4)$.

$\hookrightarrow$ 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

Summary

- The STMoE model is suggested for possibly noisy and heterogeneous regression data
- it also dedicated to acoomodate 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. $\hookrightarrow$ 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 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.

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.

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.

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.

References II

Feng Li, Mattias Villani, and Robert Kohn. Flexible modeling of conditional distributions using smooth mixtures of asymmetric student t densities. *Journal of Statistical Planning and Inference*, 140(12):3638 – 3654, 2010. ISSN 0378-3758. doi: http://dx.doi.org/10.1016/j.jspi.2010.04.031.

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 Geoffrey J. McLachlan. Laplace mixture of linear experts. *Computational Statistics & Data Analysis*, 93: 177–191, 2016. doi: http://dx.doi.org/10.1016/j.csda.2014.10.016.

Ankit B. Patel, Tan Nguyen, and Richard G. Baraniuk. A probabilistic theory of deep learning. Technical Report Technical Report No 2015-1, Rice University Electrical and Computer Engineering Dept., April 2015. URL http://arxiv.org/abs/1504.00641v1.

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

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.