Skew-Normal Mixture of Experts

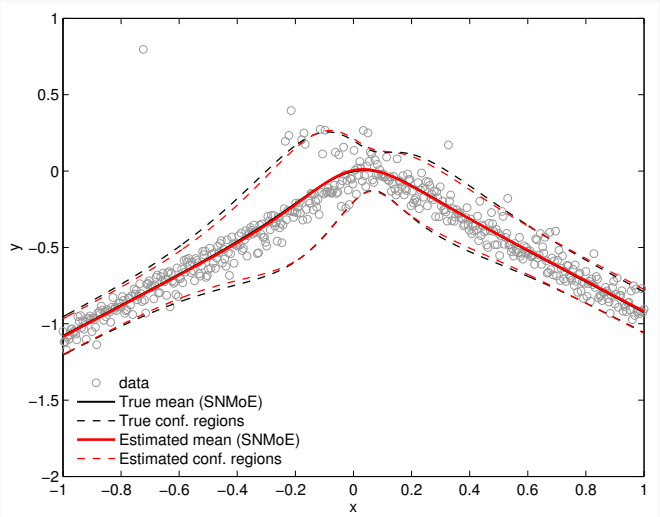
Faicel Chamroukhi

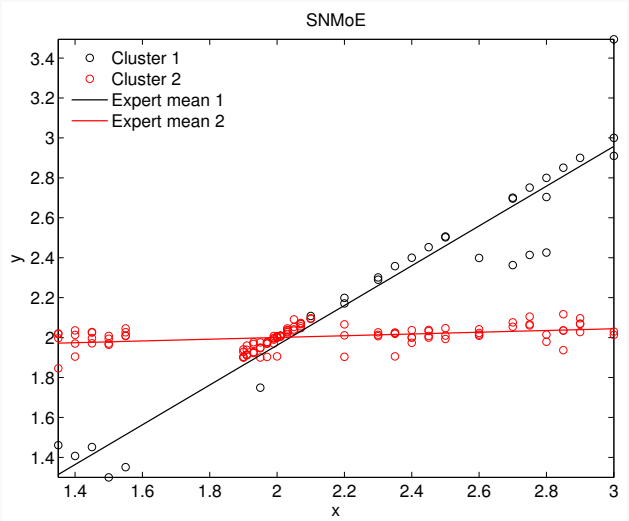
IJCNN 2016, Vancouver

July 27, 2016

Scientific context

Heterogeneous regression data





- Regression data issued from different underlying unknown processes
- Data with possibly asymmetric distributions

Objectives

- Derive a statistical model to fit at best the data
- make prediction on future observations; cluster the data
- Deal with skewness in the data distribution

Scientific context

- Analysis of clustered regression data

  $\hookrightarrow$ exploratory analysis

  $\hookrightarrow$ predictive analysis: make decision for future data

Modeling framework

- Latent variable models : $f(x|\boldsymbol{\theta}) = \int_z f(x, z|\boldsymbol{\theta})\mathrm{d}z$

  generative formulation :

$$\begin{aligned} z &\sim q(z|\boldsymbol{\theta}) \\ x|z &\sim f(x|z, \boldsymbol{\theta}) \end{aligned}$$

$\hookrightarrow$ Mixture models : $f(x|\boldsymbol{\theta}) = \sum_k \pi_k f_k(x|\boldsymbol{\theta})$

$\hookrightarrow$ density estimation for regression and clustering

$\hookrightarrow$ Infer $\boldsymbol{\theta}$ from the 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), Ingrassia et al. (2012) regression mixture based on the t distribution
- Song et al. (2014): robust regression mixture based on the Laplace distribution
- Zeller et al. (2015) : regression mixture based on scale mixtures of skew-normal distributions

$\hookrightarrow$ 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 $f_k(y|\boldsymbol{x};\boldsymbol{\Psi}_k)$

Objective

- Overcome the limitation of modeling with the normal distribution.

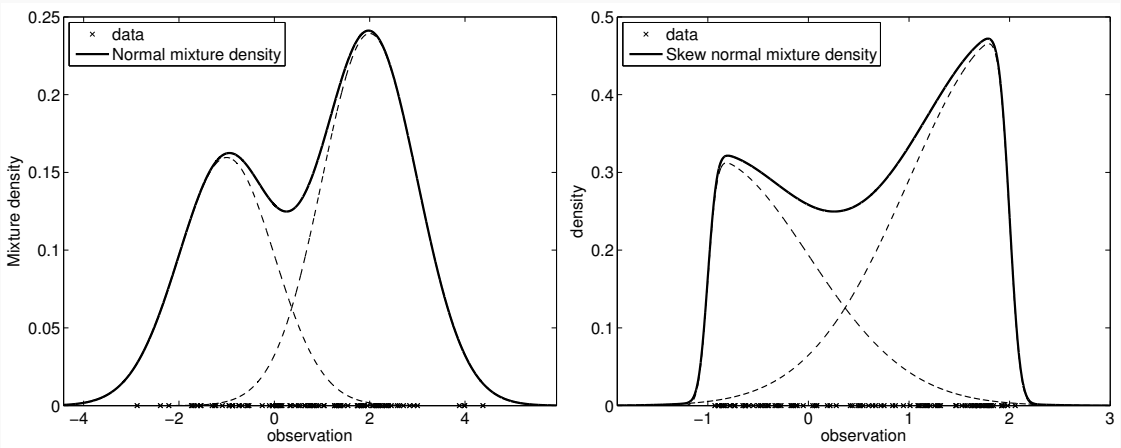
  $\hookrightarrow$ Not adapted for a set of data containing a group or groups of observations with asymmetric behavior

Non-normal mixtures of experts

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

Skew-Normal Mixtures of Experts

- the Skew-Normal MoE (SNMoE) accommodates skewness and is adapted to clustered regression data
- Corresponds to the extension of the mixture of skew-normal distributions (Lin et al., 2007) to the MoE modeling framework



$\pi_k = [0.4, 0.6]$, $\mu_k = [-1, 2]$; $\sigma_k = [1, 1]$; $\lambda_k = [14, -12]$;

The SNMoE model

- A K-component mixture of skew-normal experts (SNMoE) is defined by:

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

- kth expert: has skew-normal distribution (Azzalini, 1985, 1986):

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

where $\phi(.)$ and $\Phi(.)$ denote, respectively, the pdf and the cdf of the standard normal distribution.

$\hookrightarrow$ For $\{\lambda_k\} \to 0$, the SNMoE reduces to the NMoE.

$\hookrightarrow$ The SNMoE generalizes th normal MoE models to accommodate data with asymmetric behavior

Representation of the SNMoE model

Let Z_{ik} be the *latent* binary component-indicators such that $Z_{ik} = 1$ iff $Z_i = k$, Z_i being the hidden class label of the ith observation, we have the following generative model :

- **Hierarchical representation**

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

where $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{iK})$ is the binary indicator vector and $\delta_k = \frac{\lambda_k}{\sqrt{1+\lambda_k^2}}$ is the skewness.
The variables U_i and Z_i are hidden variables.

Parameter estimation via the ECM algorithm

- Parameter vector: $\boldsymbol{\Psi} = (\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_{K-1}^T, \boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_K^T)^T$ where $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k^T, \sigma_k^2, \lambda_k)^T$
- Maximize the observed-data log-likelihood given an observed i.i.d sample of n observations $\{y_i, \boldsymbol{x}_i, \boldsymbol{r}_i\}_{i=1}^n$:

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

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

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

$$\log L_c(\boldsymbol{\Psi}) = \log L_{1c}(\boldsymbol{\alpha}) + \sum_{k=1}^{K} \log L_{2c}(\boldsymbol{\theta}_k)$$

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

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

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

MLE via the ECM algorithm: E-Step

- **E-Step** Calculates the conditional expectation of the complete-data log-likelihood, given the observed data and a current estimation $\boldsymbol{\Psi}^{(m)}$:

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

where

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

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

↪ requires the following conditional expectations:

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

↪ Analytic solutions

SNMoE: M-Step of the ECM algorithm

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

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

ECM algorithm for the SNMoE: M-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:

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

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

3. Update the skewness parameters δ_k as solution of

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

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

Prediction, clustering, model selection

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

  Component mean : $\mathbb{E}_{\hat{\boldsymbol{\Psi}}}(Y|Z = k, \boldsymbol{x}) = \hat{\boldsymbol{\beta}}_k^T \boldsymbol{x} + \sqrt{\frac{2}{\pi}}\ \hat{\delta}_k\ \hat{\sigma}_k$

  The mean of the SNMoE model : $\mathbb{E}_{\hat{\boldsymbol{\Psi}}}(Y|\boldsymbol{x}, \boldsymbol{r}) = \sum_{k=1}^{K} \pi_k(\boldsymbol{r}; \hat{\boldsymbol{\alpha}}) \left(\hat{\boldsymbol{\beta}}_k^T \boldsymbol{x} + \sqrt{\frac{2}{\pi}} \hat{\delta}_k \hat{\sigma}_k \right)$.

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

$$\hat{z}_i = \arg\max_{k=1}^{K} \mathbb{E}[Z_i|\boldsymbol{r}_i, \boldsymbol{x}_i; \hat{\boldsymbol{\Psi}}] = \arg\max_{k=1}^{K} \mathbb{P}(Z_i = k|\boldsymbol{r}_i, \boldsymbol{x}_i; \hat{\boldsymbol{\Psi}})$$

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

  Number of free parameters: $\eta_{\boldsymbol{\Psi}} = K(p + q + 3) - q$.

A toy example

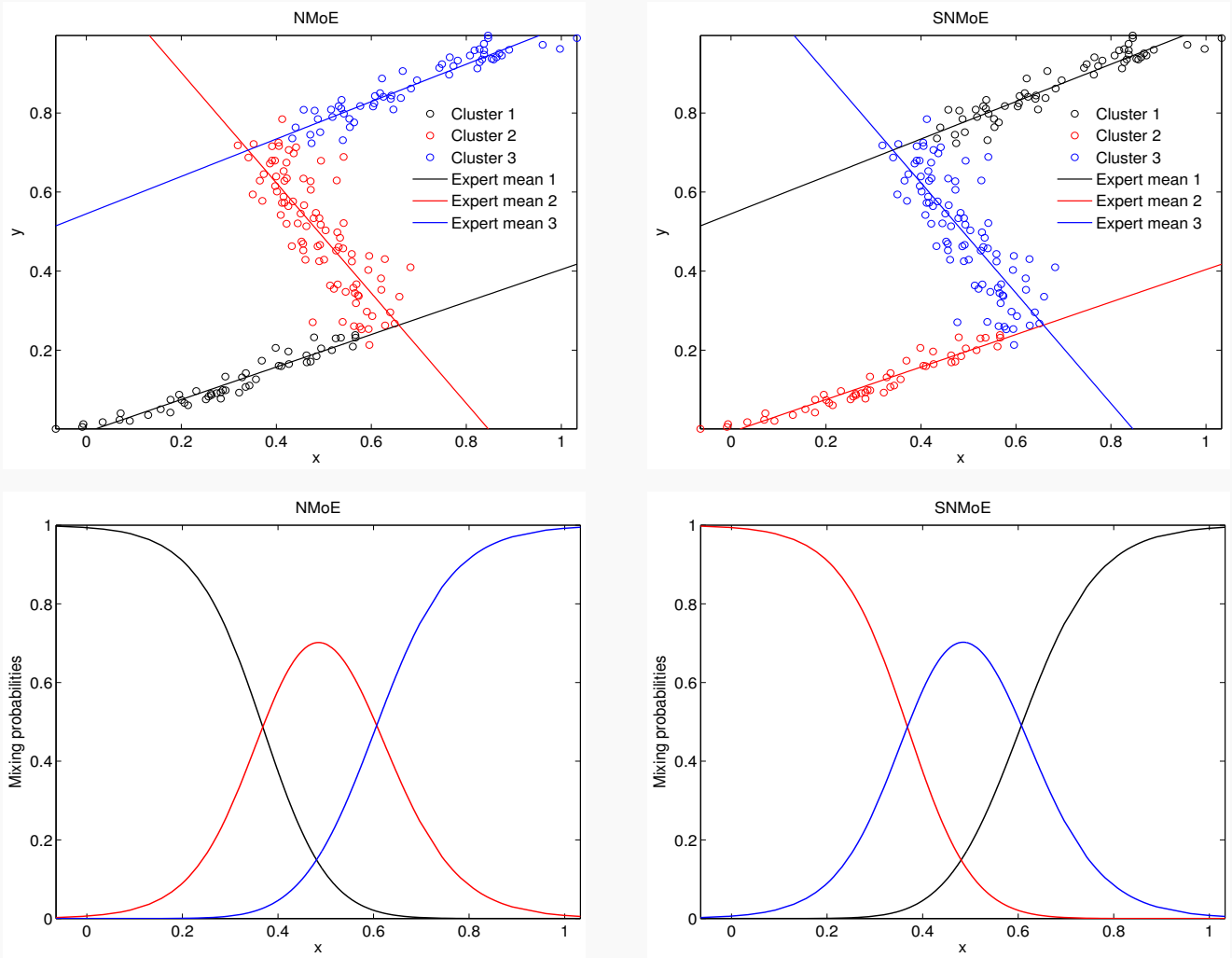
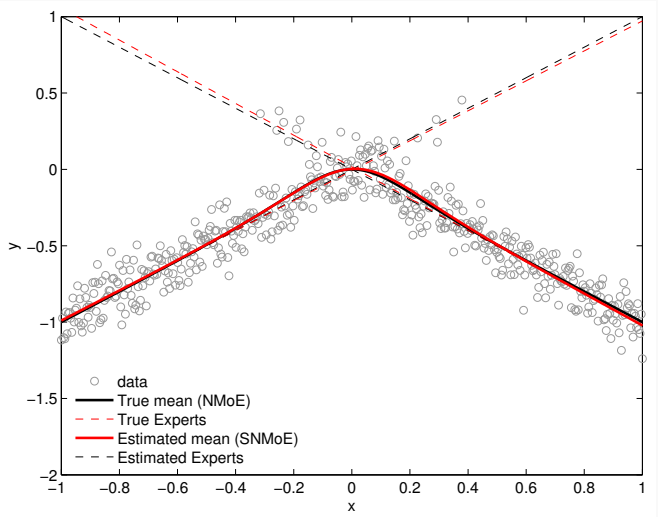
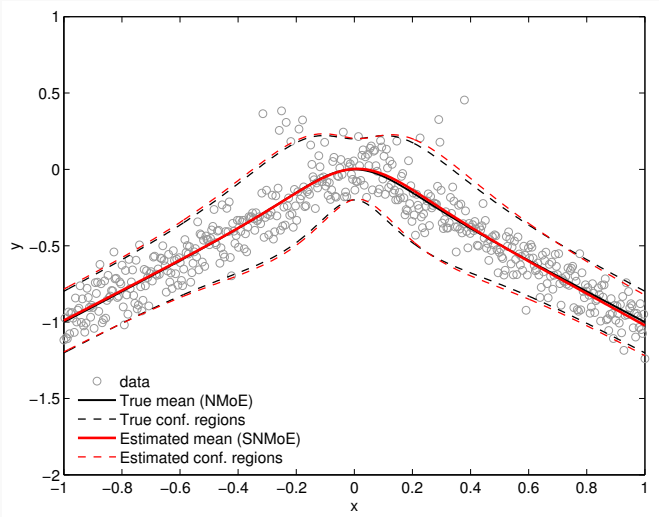
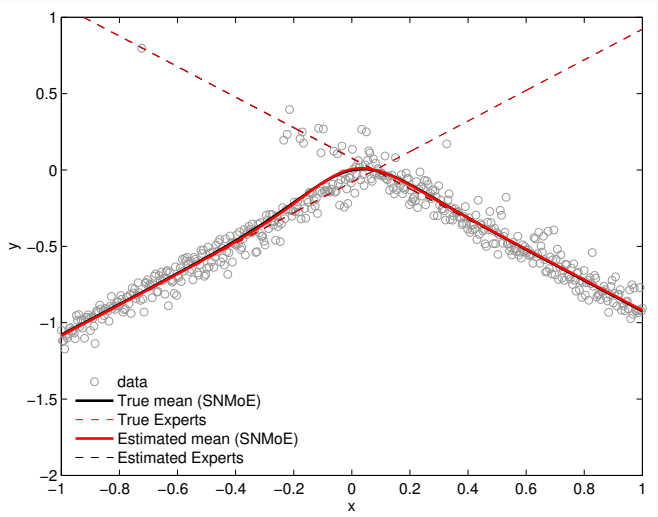


Figure: Fitting the NMoE model and the proposed SNMoE to the toy data set analyzed in Bishop and Svensén (2003): $y_i = x_i + 0.3\sin(2\pi x_i) + \epsilon_i$, with ϵ_i drawn from a zero mean Normal distribution with standard deviation 0.05 and x_i generated uniformly in $(0, 1)$

Simulated data from a mixture of two linear experts







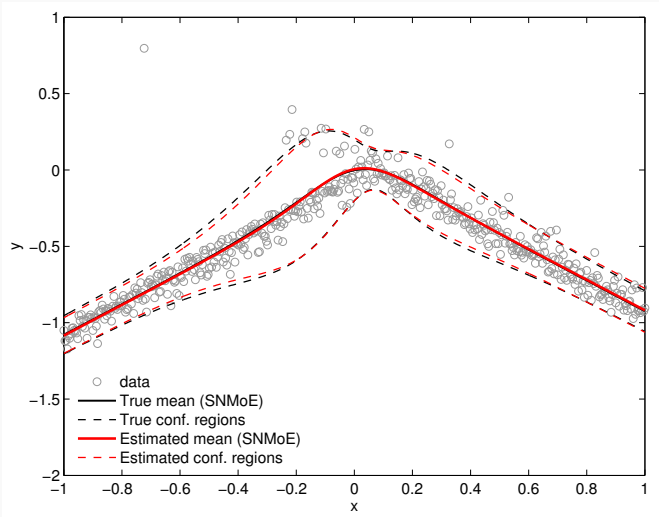
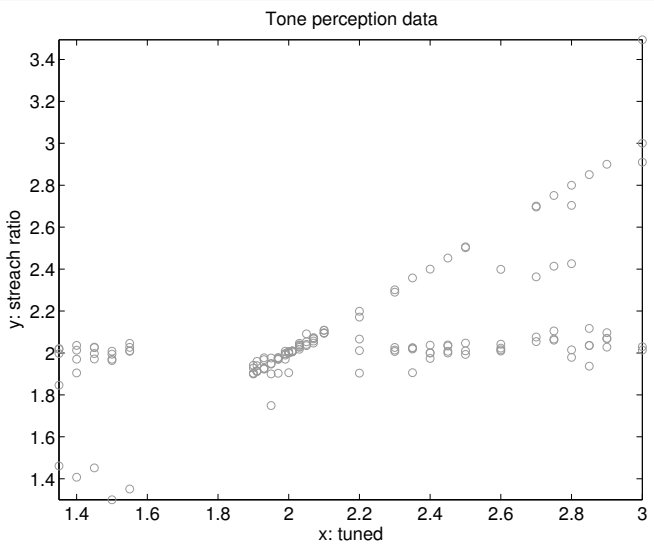


Figure: Fitted SNMoE to data generated according to the NMoE (top) and the SNMoE (bottom).

Two real datasets

- Tone perception data
- Temperature anomalies data



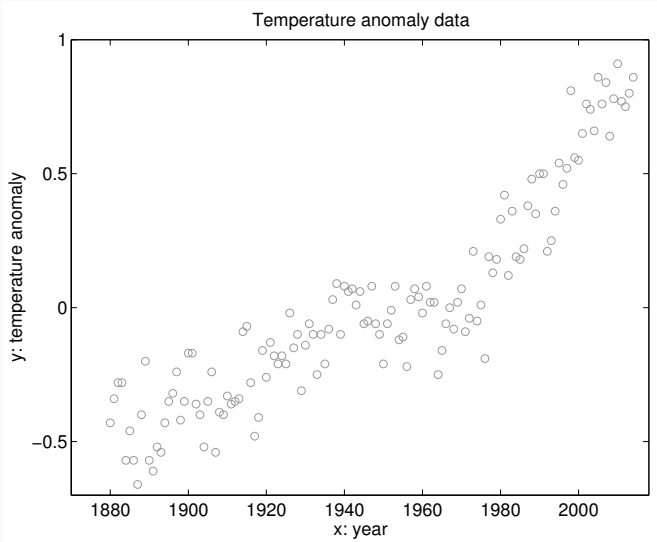


Figure: Scatter plot of the tone perception data (left) and the temperature anomalies data (right).

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 $1882 - 2012$.

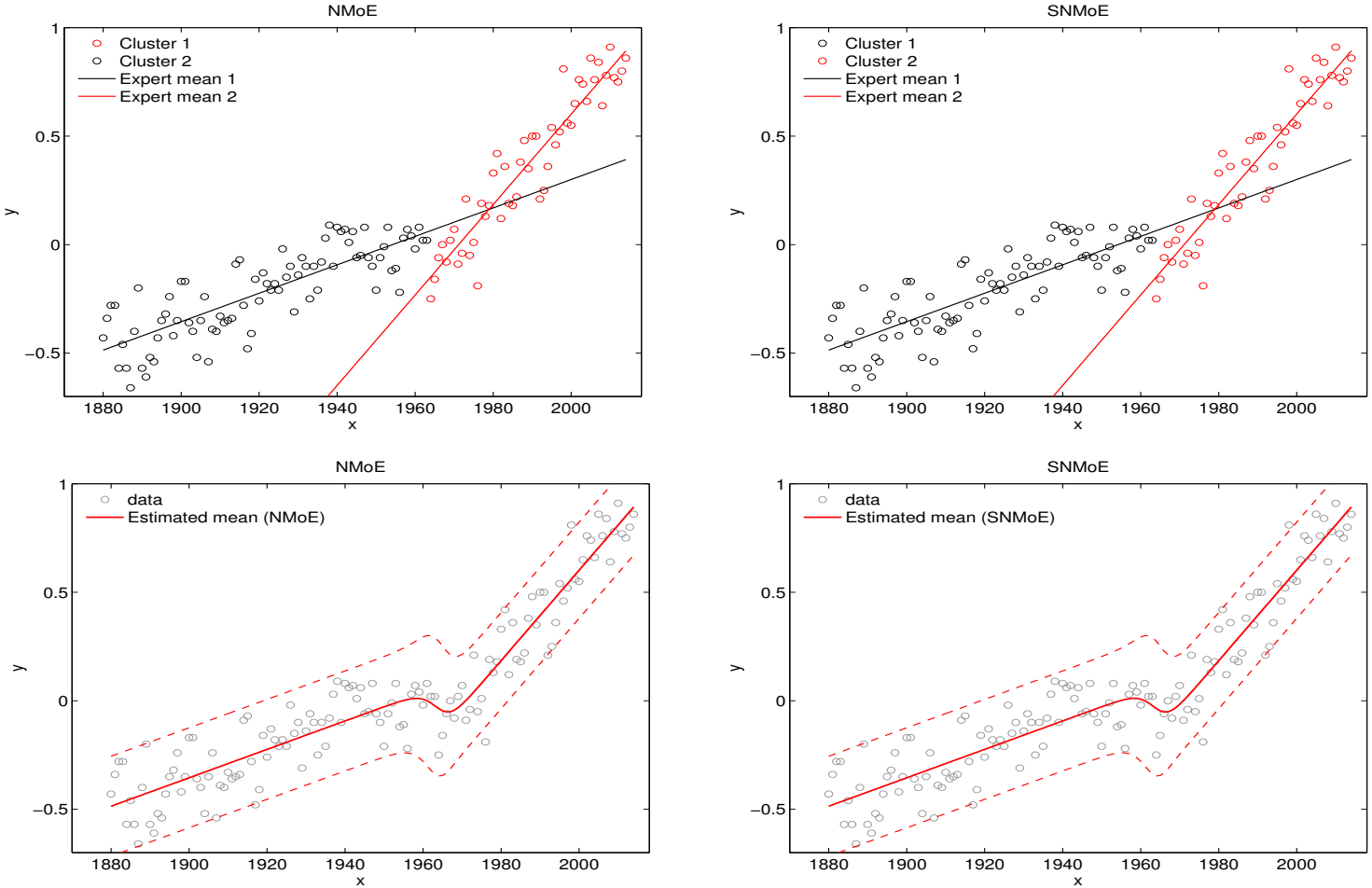


Figure: Fitting the MoLE models to the temperature anomalies data set.

- The SNMoE fit provides a skewness close to zero, which tends to approach a normal distribution.
- The regression coefficients are 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		
K	BIC	AIC	ICL	BIC	AIC	ICL
1	46.0623	50.4202	46.0623	43.6096	49.4202	43.6096
2	**79.9163**	91.5374	**79.6241**	**75.0116**	**89.5380**	**74.7395**
3	71.3963	90.2806	58.4874	63.9254	87.1676	50.8704
4	66.7276	92.8751	54.7524	55.4731	87.4312	41.1699
5	59.5100	**92.9206**	51.2429	45.3469	86.0207	41.0906

Table: Choosing the number of expert components K for the temperature anomalies data.

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).

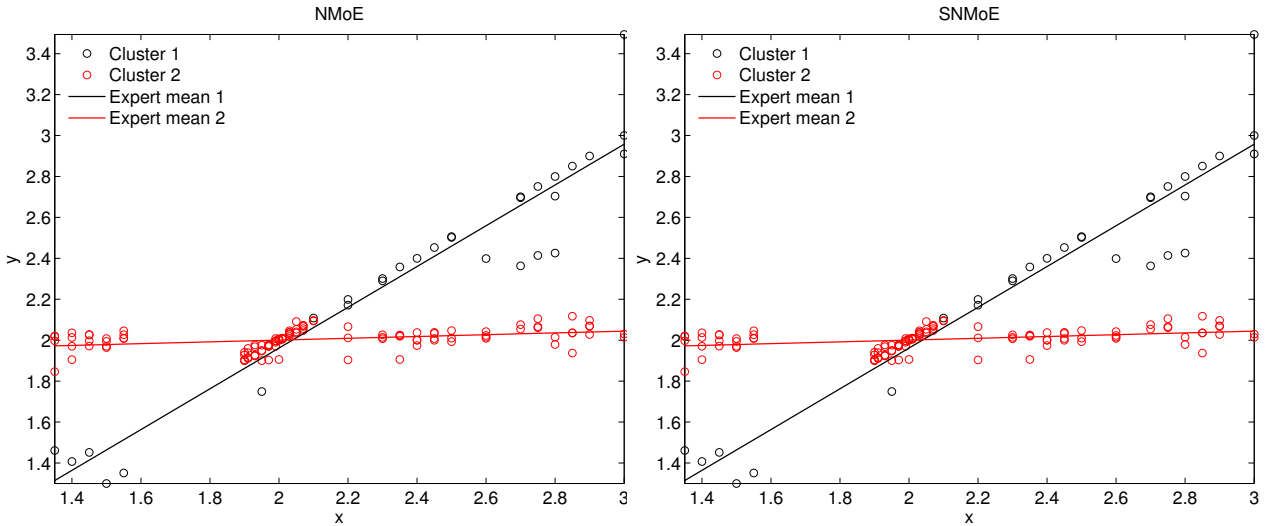


Figure: Fitting the MoE models to the tone data set

Model selection

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

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

- the number of components is overestimated with the NMoE model
- AIC performs poorly for the two models
- BIC and ICL are the suggested criteria for the analysis of this data, with the SNMoE model, which is more adapted.

Outline

Summary

- The SNMoE model is suggested for heterogeneous regression data
- it is also dedicated to accommodate regression data with possibly non-symmetric 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

- A work under review is on *robust* skew mixture of experts
- Here we only considered the MoE in their standard (non-hierarchical) version. $\hookrightarrow$ One interesting future direction is to extend it to the hierarchical MoE framework (Jordan and Jacobs, 1994).
- extension to multivariate regression

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.

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. Robust mixture of experts modeling using the t-distribution. *Neural Networks - Elsevier*, 79:20–36, 2016. URL `http://chamroukhi.univ-tln.fr/papers/TMoE.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.

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

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.

References II

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

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.

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.

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

Identifiability of the SNMoE 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 SNMoEs are identifiable:

- Ordered implies that there exist a certain ordering relationship such that $(\boldsymbol{\beta}_1^T, \sigma_1^2, \lambda_1)^T \prec \ldots \prec (\boldsymbol{\beta}_K^T, \sigma_K^2, \lambda_K)^T$;
- 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: $\boldsymbol{\beta}_k \neq \boldsymbol{\beta}_{k\prime}$, $\sigma_k \neq \sigma_{k\prime}$ or $\lambda_k \neq \lambda_{k\prime}$.

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

- The set $\{\text{SN}(y; \mu(\boldsymbol{x}; \boldsymbol{\beta}_1), \sigma_1^2, \lambda_1), \ldots, \text{SN}(y; \mu(\boldsymbol{x}; \boldsymbol{\beta}_{3K}), \sigma_{3K}^2, \lambda_{3K})\}$ contains $3K$ linearly independent functions of y, for any $3K$ distinct triplet $(\mu(\boldsymbol{x}; \boldsymbol{\beta}_k), \sigma_k^2, \lambda_k)$ for $k = 1, \ldots, 3K$.

$\hookrightarrow$ Thus, via Lemma 2 of Jiang and Tanner (1999) we have any ordered and initialized irreducible SNMoE is identifiable.