Regularized Maximum-Likelihood Estimation of Mixture-of-Experts

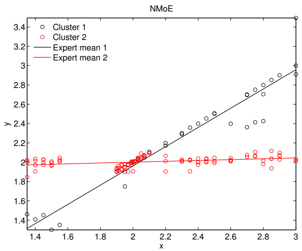
Faicel Chamroukhi & Bao-Tuyen Huynh

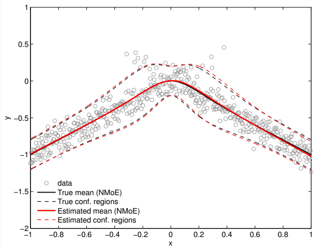
https://chamroukhi.users.lmno.cnrs.fr

IJCNN 2018, Rio

July 11, 2018

Scientific context





- Heterogeneous regression data $\hookrightarrow$ underlying unknown partition
- Data issued from non-linear regression function

Modeling framework

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

  generative formulation :

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

Outline

1. Mixture-of-Experts (MoE) Modeling and MLE
2. Regularized MLE of the MoE
3. Proposed EM algorithm with block corrdinate ascent
4. Experimental study

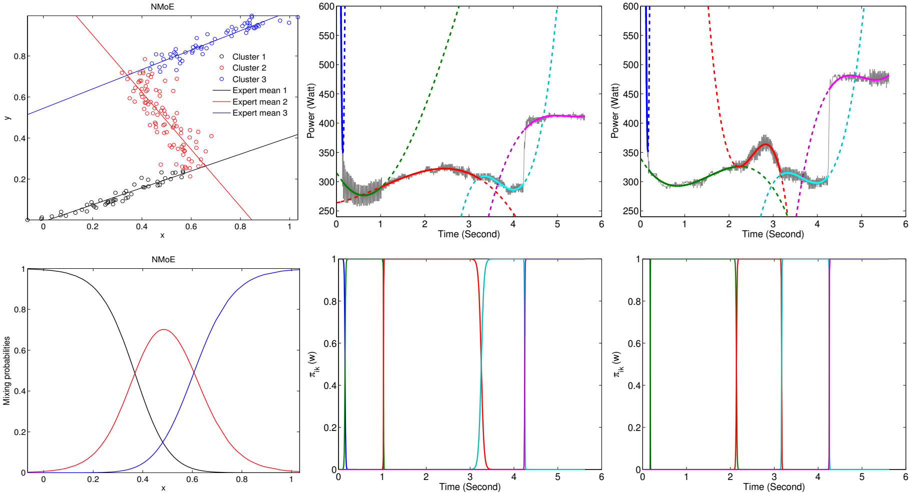
Mixture-of-Experts (MoE) modeling framework

- Observed pairs of data $(\boldsymbol{x}, y)$ where the response $y \in \mathbb{R}$ for the predictors $\boldsymbol{x} \in \mathbb{R}^p$ governed by a hidden categorical random variable Z
- Mixture of experts (MoE) (Jacobs et al., 1991; Jordan and Jacobs, 1994) :

$$f(y|\boldsymbol{x};\boldsymbol{\theta}) = \sum_{k=1}^{K} \underbrace{\pi_k(\boldsymbol{x};\mathbf{w})}_{\text{Gating network}} \underbrace{f_k(y|\boldsymbol{x};\boldsymbol{\theta}_k)}_{\text{Expert Network}}$$

- Gating network (e.g softmax): $\pi_k(\boldsymbol{x};\mathbf{w}) = \frac{\exp(w_{k0}+\boldsymbol{w}_k^T\boldsymbol{r})}{1+\sum_{\ell=1}^{K-1}\exp(w_{\ell 0}+\boldsymbol{w}_\ell^T\boldsymbol{r})}$
- Experts network (e.g Gaussian regressors): $f_k(y|\boldsymbol{x};\boldsymbol{\theta}_k) = \phi\big(y;\mu(\boldsymbol{x};\boldsymbol{\beta}_k),\sigma_k^2\big)$ with parametric (non-)linear regression functions $\mu(\boldsymbol{x};\boldsymbol{\beta}_k)$
- is parameterized by $\boldsymbol{\theta} = (\mathbf{w}^T, \boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_K^T)^T$
- Non-normal MoE, for data with atypical observations, and with possible heavy tailed and asymmetric distributions: Chamroukhi (2016, 2017); Nguyen and Chamroukhi (2018)

Illustrations



Standard MLE of the MoE model

- MLE: $\boldsymbol{\theta}$ is commonly estimated by maximizing the observed-data log-likelihood:

$$\widehat{\boldsymbol{\theta}}_n \in \arg\max_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta})$$

with

$L(\boldsymbol{\theta}) = \ln f((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_1); \boldsymbol{\theta}) = \sum_{i=1}^{n} \ln \sum_{k=1}^{K} \pi_k(\boldsymbol{x}_i; \boldsymbol{w}) f(\boldsymbol{y}_i|\boldsymbol{x}_i; \boldsymbol{\theta}_k).$

$\hookrightarrow$ the EM algorithm (Dempster et al. (1977))

Standard MLE of the MoE model

- MLE: $\boldsymbol{\theta}$ is commonly estimated by maximizing the observed-data log-likelihood:

$$\widehat{\boldsymbol{\theta}}_n \in \arg\max_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta})$$

with
$L(\boldsymbol{\theta}) = \ln f((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_1); \boldsymbol{\theta}) = \sum_{i=1}^{n} \ln \sum_{k=1}^{K} \pi_k(\boldsymbol{x}_i; \boldsymbol{w}) f(\boldsymbol{y}_i|\boldsymbol{x}_i; \boldsymbol{\theta}_k).$
$\hookrightarrow$ the EM algorithm (Dempster et al. (1977))

$\hookrightarrow$ The standard MLE of MoE when p is large (high-dimensional setting)
$\hookrightarrow$ the features are possibly correlated and sparse
$\hookrightarrow$ Looking for a sparse models

Regularized MLE of the MoE

RMLE: $\boldsymbol{\theta}$ is estimated by maximizing a penalized observed-data log-likelihood:

$$\widehat{\boldsymbol{\theta}}_n \in \arg\max_{\boldsymbol{\theta}\in\Theta} PL(\boldsymbol{\theta})$$

with $PL(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) - \mathrm{Pen}(\boldsymbol{\theta})$

- $\hookrightarrow$ Pen$(\boldsymbol{\theta})$ should encourage sparsity
- parameter estimation and selection problem

Proposed Regularized Mixture of Experts model

$$\mathrm{Pen}(\boldsymbol{\theta}) = \underbrace{\sum_{k=1}^{K} \lambda_k \|\boldsymbol{\beta}_k\|_1}_{\text{Lasso-like pen.}} + \underbrace{\sum_{k=1}^{K-1} \gamma_k \|\boldsymbol{w}_k\|_1 + \frac{\rho}{2}\|\boldsymbol{w}_k\|_2^2}_{\text{Elastic-Net like pen.}}$$

- Lasso penalty for the experts $\hookrightarrow$ encourage a sparse solution
- The elastic net penalty (Zou and Hastie (2005)) for the gating network:
  $\hookrightarrow$ reduce the norm of the estimated values of the gating network parameters by using the L_2 penalties;
  $\hookrightarrow$ the Lasso penalty to recover a sparse solution
- The convexity of L_1 and L_2 penalties have also advantageous numerical properties.
- If the correlation between the features is high, one can add L_2 penalties for the expert network.

Regularized MLE via an EM algorithm

- The penalized log-likelihood function:

$$PL(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) - \sum_{k=1}^{K} \lambda_k \|\boldsymbol{\beta}_k\|_1 - \sum_{k=1}^{K-1} (\gamma_k \|\boldsymbol{w}_k\|_1 + \frac{\rho}{2} \|\boldsymbol{w}_k\|_2^2) \tag{1}$$

- The penalized complete-data log-likelihood function:

$$PL_c(\boldsymbol{\theta}) = L_c(\boldsymbol{\theta}) - \sum_{k=1}^{K} \lambda_k \|\boldsymbol{\beta}_k\|_1 - \sum_{k=1}^{K-1} (\gamma_k \|\boldsymbol{w}_k\|_1 + \frac{\rho}{2} \|\boldsymbol{w}_k\|_2^2) \tag{2}$$

with

$$L_c(\boldsymbol{\theta}) = \ln f((\boldsymbol{x}_1, y_1, z_1)), \ldots, (\boldsymbol{x}_n, y_n, z_n); \boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log [\pi_k(\boldsymbol{x}_i; \boldsymbol{w}) f(\boldsymbol{y}_i | \boldsymbol{x}_i; \boldsymbol{\theta}_k$$

such that $z_{ik} = 1$ iff $z_i = k$ (the data pair $(\boldsymbol{x}_i, y_i)$ originates from expert k

Statistical inference for RMoE

Theorem (Khalili (2010))

Let $(\boldsymbol{V}_i)_{i=1,\ldots,n} = (\boldsymbol{X}_i, Y_i)_{i=1,\ldots,n}$ be a random sample from a density function $f(\boldsymbol{v};\boldsymbol{\theta})$ ($\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_h)$) which satisfies some regularity conditions:
The joint density of $\boldsymbol{V}_i$ is given by

$$f(\boldsymbol{v}_i;\boldsymbol{\theta}) = f(\boldsymbol{x}_i)\sum_{k=1}^{K}\pi_k(\boldsymbol{x}_i;\boldsymbol{w})p(\boldsymbol{y}_i|\boldsymbol{x}_i;\boldsymbol{\theta}_k).$$

Assume that $\rho/\sqrt{n} \to 0$ as $n \to \infty$. Then, there exists a local maximizer $\widehat{\boldsymbol{\theta}}_n$ of the regularized log-likelihood function $PL(\boldsymbol{\theta})$ (1) for which

$$\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| = O(\frac{1}{\sqrt{n}}(1 + q^*_{1n} + q_{1n})),$$

where

$$q^*_{1n} = \max_{k,j}\{\lambda_k/\sqrt{n} : \beta^0_{kj} \neq 0\};\; q_{1n} = \max_{k,j}\{\gamma_k/\sqrt{n} : w^0_{kj} \neq 0\}.$$

- By choosing $\max_k \gamma_k = O(\sqrt{n}), \max_k \lambda_k = O(\sqrt{n})$ we have the root-n consistent property for $\widehat{\boldsymbol{\theta}}_n$.

Parameter estimation for RMoE

Khalili's method:

- Approximates the L_1 penalty function in a some neighborhood by an ε -local quadratic function

$$\eta|t| \approx \eta|t_0| + \frac{\eta}{2(|t_0| + \varepsilon)}(t^2 - t_0^2).$$

  $\hookrightarrow$ Almost surely none of the components will be exactly zero.
- Needs using a threshold to recover the zero coefficients
  $\hookrightarrow$ The size of threshold affects the degree of sparsity of the solution.
- The Newton-Raphson algorithm is used to update the M-step of the EM algorithm.
  $\hookrightarrow$ This approach still require computing the inverse matrix.

In our proposal:

- A block EM algorithm with coordinate ascent algorithm to estimate the parameters:
  $\hookrightarrow$ Exact L_1 penalty regularization;
  $\hookrightarrow$ Avoids computing matrix inversion;
  $\hookrightarrow$ Avoids using a threshold to recover the zero coefficients.

Block EM algorithm with coordinate ascent

E-step

- Compute the conditional expectation of the penalized complete-data log-likelihood

$$
\begin{aligned}
Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(q)}) &= \mathbb{E}\left[PL_c(\boldsymbol{\theta})|\mathcal{D};\boldsymbol{\theta}^{(q)}\right]\\
&= \sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}^{(q)}\log\left[\pi_k(\boldsymbol{x}_i;\boldsymbol{w})f_k(\boldsymbol{y}_i|\boldsymbol{x}_i;\boldsymbol{\theta}_k)\right]\\
&\quad -\sum_{k=1}^{K}\lambda_k\|\boldsymbol{\beta}_k\|_1-\sum_{k=1}^{K-1}(\gamma_k\|\boldsymbol{w}_k\|_1-\frac{\rho}{2}\|\boldsymbol{w}_k\|_2^2).
\end{aligned}
$$

Block EM algorithm with coordinate ascent

E-step

- Compute the conditional expectation of the penalized complete-data log-likelihood

$$
\begin{aligned}
Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(q)}) &= \mathbb{E}\left[PL_c(\boldsymbol{\theta})|\mathcal{D};\boldsymbol{\theta}^{(q)}\right] \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}^{(q)}\log\left[\pi_k(\boldsymbol{x}_i;\boldsymbol{w})f_k(\boldsymbol{y}_i|\boldsymbol{x}_i;\boldsymbol{\theta}_k)\right] \\
&\quad -\sum_{k=1}^{K}\lambda_k\|\boldsymbol{\beta}_k\|_1 - \sum_{k=1}^{K-1}(\gamma_k\|\boldsymbol{w}_k\|_1 - \frac{\rho}{2}\|\boldsymbol{w}_k\|_2^2).
\end{aligned}
$$

$\hookrightarrow$ Calculate the posterior component probabilities:

$$
\tau_{ik}^{(q)} = \mathbb{P}(Z_i = k|\boldsymbol{y}_i,\boldsymbol{x}_i;\boldsymbol{\theta}^{(q)}) = \frac{\pi_k(\boldsymbol{x}_i;\boldsymbol{w}^{(q)})\mathcal{N}(y_i;\beta_{k0}^{(q)} + \boldsymbol{x}_i^T\boldsymbol{\beta}_k^{(q)},\sigma_k^{(q)2})}{\sum_{l=1}^{K}\pi_l(\boldsymbol{x}_i;\boldsymbol{w}^{(q)})\mathcal{N}(y_i;\beta_{l0}^{(q)} + \boldsymbol{x}_i^T\boldsymbol{\beta}_l^{(q)},\sigma_l^{(q)2})}.
$$

$\hookrightarrow$ As in standard MoE

Block EM algorithm with coordinate ascent (cont.)

M-step

- Maximizing the Q function: $\boldsymbol{\theta}^{(q+1)} \in \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(q)})$ with

$$Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(q)}) = Q(\boldsymbol{w};\boldsymbol{\theta}^{(q)}) + Q(\boldsymbol{\beta},\sigma;\boldsymbol{\theta}^{(q)}),$$

where

$$Q(\boldsymbol{w};\boldsymbol{\theta}^{(q)}) = \sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}^{(q)}\log\pi_k(\boldsymbol{x}_i;\boldsymbol{w}) - \sum_{k=1}^{K-1}\left(\gamma_k\|\boldsymbol{w}_k\|_1 - \frac{\rho}{2}\|\boldsymbol{w}_k\|_2^2\right), \tag{3}$$

$\hookrightarrow$ a weighted regularized multiclass logistic regression problem

and

$$Q(\boldsymbol{\beta},\sigma;\boldsymbol{\theta}^{(q)}) = \sum_{k=1}^{K}\sum_{i=1}^{n}\tau_{ik}^{(q)}\log\mathcal{N}(y_i;\beta_{k0}+\boldsymbol{x}_i^T\boldsymbol{\beta}_k,\sigma_k^2) - \sum_{k=1}^{K}\lambda_k\|\boldsymbol{\beta}_k\|_1 \tag{4}$$

$\hookrightarrow$ K independent weighted LASSO problems

Updating the gating network parameters

- Coordinate ascent algorithm to update $\boldsymbol{w}$ Tseng (1988, 2001)
- w_{kj} is updated by maximizing the component (k, j) of (3) given by

$$Q(w_{kj};\boldsymbol{\theta}^{(q)}) = \begin{cases} F(w_{kj};\boldsymbol{\theta}^{(q)}) - \gamma_k w_{kj} & \text{, if } w_{kj} > 0 \quad (F_1) \\ F(0;\boldsymbol{\theta}^{(q)}) & \text{, if } w_{kj} = 0 \\ F(w_{kj};\boldsymbol{\theta}^{(q)}) + \gamma_k w_{kj} & \text{, if } w_{kj} < 0 \quad (F_2) \end{cases},$$

$$F(w_{kj};\boldsymbol{\theta}^{(q)}) = \sum_{i=1}^{n} \tau_{ik}^{(q)}(w_{k0} + \boldsymbol{w}_k^T \boldsymbol{x}_i) - \sum_{i=1}^{n} \log\Big(1 + \sum_{l=1}^{K-1} e^{w_{l0} + \boldsymbol{w}_l^T \boldsymbol{x}_i}\Big) - \frac{\rho}{2} w_{kj}^2. \quad (5)$$

Univariate Newton-Raphson algorithm

- F_1 and F_2 are smooth univariate concave functions in w_{kj}. $\hookrightarrow$ Univariate Newton-Raphson algorithm can be used to update w_{kj}

$$w_{kj}^{(s+1)} = w_{kj}^{(s)} - \left(\frac{\partial^2 F(w_{kj};\boldsymbol{\theta}^{(q)})}{\partial^2 w_{kj}}\right)^{-1}\Bigg|_{w_{kj}^{(s)}} \left(\frac{\partial F(w_{kj};\boldsymbol{\theta}^{(q)})}{\partial w_{kj}} - \gamma_k \text{sign}(w_{kj})\right)\Bigg|_{w_{kj}^{(s)}},$$

where $\frac{\partial^2 F(w_{kj};\boldsymbol{\theta}^{(q)})}{\partial^2 w_{kj}}$ and $\frac{\partial F(w_{kj};\boldsymbol{\theta}^{(q)})}{\partial w_{kj}}$ have closed-form.

Updating the expert parameters

M-step (cont.)

- Update β_{kj} using coordinate ascent algorithm with soft-thresholding operator

$$\beta_{kj}^{[s+1]} = \mathcal{S}_{\lambda_k \sigma_k^{(q)2}} \Big(\sum_{i=1}^{n} \tau_{ik}^{(q)} r_{ikj}^{[s]} x_{ij}\Big) \Big/ \sum_{i=1}^{n} \tau_{ik}^{(q)} x_{ij}^{2},$$

where $r_{ikj}^{[s]} = y_i - \beta_{k0}^{[s]} - \boldsymbol{\beta}_k^{[s]T} \boldsymbol{x}_i + \beta_{kj}^{[s]} x_{ij}$, $[\mathcal{S}_{\gamma}(u)]_j = \mathsf{sign}(u_j)(|u_j| - \gamma)_+$ and $(x)_+ = \max\{x, 0\}$ in the sth loop of the coordinate ascent algorithm.

$$\beta_{k0}^{[s+1]} = \sum_{i=1}^{n} \tau_{ik}^{(q)} (y_i - \boldsymbol{x}_i^{\top} \boldsymbol{\beta}_k^{[s+1]}) \Big/ \sum_{i=1}^{n} \tau_{ik}^{(q)}.$$

Updating the expert parameters

M-step (cont.)

- Update β_{kj} using coordinate ascent algorithm with soft-thresholding operator

$$\beta_{kj}^{[s+1]} = \mathcal{S}_{\lambda_k \sigma_k^{(q)2}}\Big(\sum_{i=1}^{n} \tau_{ik}^{(q)} r_{ikj}^{[s]} x_{ij}\Big) \Big/ \sum_{i=1}^{n} \tau_{ik}^{(q)} x_{ij}^2,$$

where $r_{ikj}^{[s]} = y_i - \beta_{k0}^{[s]} - \boldsymbol{\beta}_k^{[s]T} \boldsymbol{x}_i + \beta_{kj}^{[s]} x_{ij}$, $[\mathcal{S}_\gamma(u)]_j = \mathsf{sign}(u_j)(|u_j| - \gamma)_+$ and $(x)_+ = \max\{x, 0\}$ in the sth loop of the coordinate ascent algorithm.

$$\beta_{k0}^{[s+1]} = \sum_{i=1}^{n} \tau_{ik}^{(q)} (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_k^{[s+1]}) \Big/ \sum_{i=1}^{n} \tau_{ik}^{(q)}.$$

- Rerun the E-step, keep

$$(w_{k0}^{(q+2)}, \boldsymbol{w}_k^{(q+2)}) = (w_{k0}^{(q+1)}, \boldsymbol{w}_k^{(q+1)});\ (\beta_{k0}^{(q+2)}, \boldsymbol{\beta}_k^{(q+2)}) = (\beta_{k0}^{(q+1)}, \boldsymbol{\beta}_k^{(q+1)}),$$

and update $\sigma_k^{2(q+2)}$ as follows

$$\sigma_k^{2(q+2)} = \sum_{i=1}^{n} \tau_{ik}^{(q+1)} (y_i - \beta_{k0}^{(q+2)} - \boldsymbol{x}_i^\top \boldsymbol{\beta}_k^{(q+2)})^2 \Big/ \sum_{i=1}^{n} \tau_{ik}^{(q+1)}.$$

Simulation study

Simulation protocol

- $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}; \boldsymbol{\Sigma})$ with $\mathrm{corr}(x_{ij}, x_{ij'}) = 0.5^{|j-j'|}$; $K = 2$
- Sample size: $n = 300$, 100 different data sets;
- The regression coefficients:

$$
\begin{aligned}
(\beta_{10}, \boldsymbol{\beta}_1)^T &= (0, 0, 1.5, 0, 0, 0, 1)^T; \sigma_1 = 1 \\
(\beta_{20}, \boldsymbol{\beta}_2)^T &= (0, 1, -1.5, 0, 0, 2, 0)^T; \sigma_2 = 1 \\
(w_{10}, \boldsymbol{w}_1)^T &= (1, 2, 0, 0, -1, 0, 0)^T; \sigma_3 = 1
\end{aligned}
$$

Considered approaches for comparison

- The standard MoE;
- MoE+L_2 (MoE with L_2 penalties in the gating network);
- MoE-BIC (MoE with model selection using BIC criterion - 100 submodels);
- MIXLASSO (MLR with Lasso penalties) (see Khalili and Chen (2007));

Evaluation criteria

- The sensitivity/specificity (sparsity);
- The parameter estimation (density estimation);
- The misclassification error: Adjust rand index - ARI (clustering).

Sensitivity/specificity result

- *Sensitivity (S_1):* proportion of correctly estimated zero coefficients;
- *Specificity:* proportion of correctly estimated nonzero coefficients.

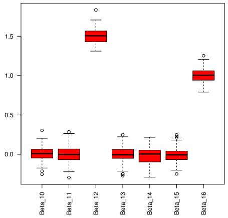
Method	Expert 1		Expert 2		Gate	
	S_1	S_2	S_1	S_2	S_1	S_2
MoE	0.000	1.000	0.000	1.000	0.000	1.000
MoE+L_2	0.000	1.000	0.000	1.000	0.000	1.000
MoE-BIC	0.920	1.000	0.930	1.000	0.850	1.000
MIXLASSO	0.775	1.000	0.693	1.000	N/A	N/A
Our MoE-Lasso+L_2	0.700	1.000	0.803	1.000	0.853	0.945

Table: Sensitivity (S_1) and specificity (S_2) results.

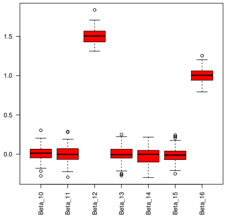
- MoE and MoE+L_2 could not be considered as model selection methods since their sensitivity equal zero.
- MIXLASSO can detect the zero coefficients in the experts. However, this model has a poor result when clustering the data.
- The MoE-Lasso+L_2 model can detect the zero coefficients in the experts and the gating network.

Parameter estimation for expert 1

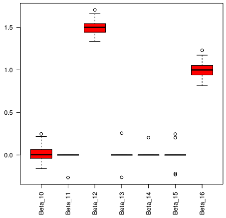
- $(\beta_{10}, \boldsymbol{\beta}_1)^T = (0, 0, 1.5, 0, 0, 0, 1)^T$.



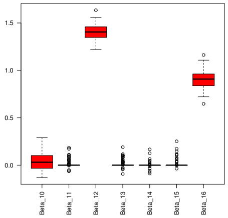
MoE



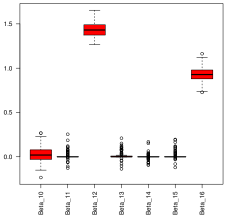
MoE-L_2



MoE-BIC



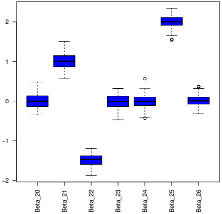
MIXLASSO



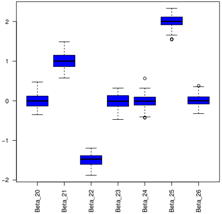
MoE-Lasso + L_2

Parameter estimation for expert 2

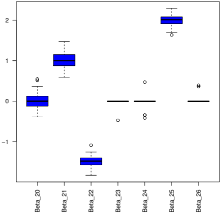
- $(\beta_{20}, \boldsymbol{\beta}_2)^T = (0, 1, -1.5, 0, 0, 2, 0)^T$.



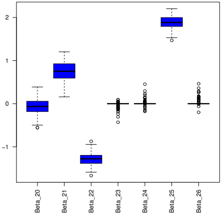
MoE



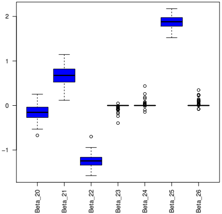
MoE-L_2



MoE-BIC



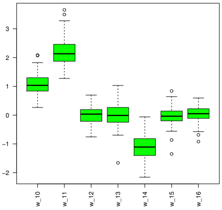
MIXLASSO



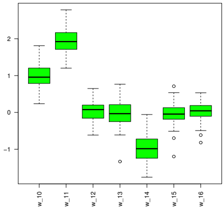
MoE-Lasso + L_2

Parameter estimation for gating network

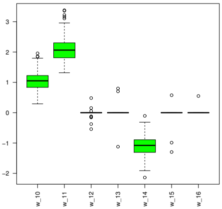
- $(w_{10}, \boldsymbol{w}_1)^T = (1, 2, 0, 0, -1, 0, 0)^T$.



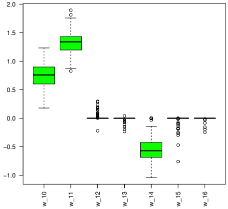
MoE



MoE-L_2



MoE-BIC



MoE-Lasso + L_2

Result for data clustering

Model	MoE	MoE+L_2	MoE-BIC	**MoE-Lasso** + L_2	MIXLASSO
C. rate	$89.57\%_{(1.65\%)}$	$89.62\%_{(1.63\%)}$	$90.05\%_{(1.65\%)}$	$89.46\%_{(1.76\%)}$	$82.89\%_{(1.92\%)}$
ARI	$0.6226_{(.053)}$	$0.6241_{(.052)}$	$0.6380_{(.053)}$	$0.6190_{(.056)}$	$0.4218_{(.050)}$

Table: clustering accuracy results (correct classification rate and Adjusted Rand Index).

Remarks

- MoE-BIC provides the best results. However, it is hard to apply BIC in reality especially for high dimensional data, since this involves a huge collection of model candidates.
- MIXLASSO can detect zero coefficients in the experts, but it provides a poor result when clustering data.
- MoE-Lasso+L_2 can detect zero coefficients in the model and provide a competitive result with MoE, MoE-L_2 in term of clustering, although it also causes bias to the non-zero coefficients.

Applications to real data sets

- For real data sets, we calculate the mean squared error and the correlation between the response variable Y with its predictor $\hat{Y}$, where

$$\hat{Y} = \sum_{k=1}^{K} \pi_k(\boldsymbol{x}; \hat{\boldsymbol{w}})(\hat{\beta}_{k0} + \boldsymbol{x}^T \hat{\boldsymbol{\beta}}_k).$$

- Housing data: 13 features, 506 observations, $K = 2$.

	MoE	MoE-Lasso+L_2 (Khalili)	**MoE-Lasso + L_2**
R^2	0.8457	0.8094	0.8221
MSE	$0.1544_{(.577)}$	$0.2044_{(.709)}$	$0.1989_{(.619)}$

Table: Results for Housing data set.

- Baseball salary data: 32 features, 337 observations, $K = 2$.

	MoE	**MoE-Lasso + L_2**	MIXLASSO
R^2	0.8099	0.8020	0.4252
MSE	$0.2625_{(.758)}$	$0.2821_{(.633)}$	$1.1858_{(2.792)}$

Table: Results for Baseball salaries data set.

The proximal Newton method

- We recently improve the proposed algorithm by using the proximal Newton method (Lee et al. (2006), Lee et al. (2014) and Friedman et al. (2010)) for updating the gating network parameters.
- The idea of the proximal Newton method:
  - Approximate the smooth part of $Q(\boldsymbol{w}; \boldsymbol{\theta}^{(q)})$ with its local quadratic form;
  - Use coordinate ascent with soft-thresholding operator to solve the resulting approximated convex optimization problem;
  - Combine with backtracking line search to update $\boldsymbol{w}$.

Extension result for proximal Newton method

- Coordinate ascent algorithm (CA) VS proximal Newton (PN) method:

Criteria	**MoE-Lasso** + L_2 (CA)	**MoE-Lasso** + L_2 (PN)
C.Rate	$89.46\%_{(1.76\%)}$	$89.53\%_{(1.65\%)}$
ARI	$0.6190_{(.056)}$	$0.6210_{(.052)}$
$PL(\boldsymbol{\theta})$ value	$-558.140_{(12.99)}$	$-558.410_{(13.03)}$

Table: Simulation results.

- Application of the proximal Newton algorithm to the residential building data set: 107 features, 372 observations, $K = 3$.

Method	Before clustering		After clustering	
	R^2	MSE	R^2	MSE
Proximal Newton	0.9887	$0.0120_{(.879)}$	0.9993	$0.000654_{(.002)}$

Table: Results for residential building data set.

Conclusion and perspectives

Conclusion

- We propose a regularized MoE which does not require using approximations as in standard MoE regularization
- A blockwise EM algorithm with coordinate ascent algorithm is proposed to monotonically maximize the RMoE objective function
- The updating of the gating network for some situations is time consuming since we don't have a closed-form
- The algorithm has been improved by using proximal Newton method to update the gating network, which has a closed-form update for each parameter and improve the running time
- Future work: Estimation and feature selection for hierarchical MoE and MoE with discrete data, ...
- Consider the case $p \gg n$

Thank you!

References I

F. Chamroukhi. Robust mixture of experts modeling using the t-distribution. *Neural Networks - Elsevier*, 79:20–36, 2016. 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`.

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

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

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.

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

A. Khalili. New estimation and feature selection methods in mixture-of-experts models. *Canadian Journal of Statistics*, 38(4): 519–539, 2010.

A. Khalili and J. Chen. Variable selection in finite mixture of regression models. *Journal of the American Statistical association*, 102(479):1025–1038, 2007.

Jason D Lee, Yuekai Sun, and Michael A Saunders. Proximal newton-type methods for minimizing composite functions. *SIAM Journal on Optimization*, 24(3):1420–1443, 2014.

Su-In Lee, Honglak Lee, Pieter Abbeel, and Andrew Y Ng. Efficient l_1 regularized logistic regression. In *AAAI*, volume 6, pages 401–408, 2006.

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`. https://arxiv.org/abs/1707.03538v1.

P. Tseng. Coordinate ascent for maximizing nondifferentiable concave functions. 1988.

P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494, 2001.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.