Non-asymptotic model selection in mixture of experts models

TrungTin Nguyen[1], Faicel Chamroukhi[1], Hien Duy Nguyen[2], Florence Forbes[3]

[1] *UNICAEN, LMNO UMR CNRS, France,* [2] *La Trobe University, Australia,* [3] *Inria Grenoble Rhone-Alpes, France.*

Learning nonlinear regression models from complex data using GLoME models

Random sample: $(\mathbf{X}_i, \mathbf{Y}_i)_{i\in[n]} \in (\mathcal{X}\times\mathcal{Y})^n \subset \left(\mathbb{R}^D \times \mathbb{R}^L\right)^n$ of the multivariate response $\mathbf{Y} = (\mathbf{Y}_j)_{j\in[L]}$ and the set of covariates $\mathbf{X} = (\mathbf{X}_j)_{j\in[D]}$ with the corresponding observed values $(\mathbf{x}_i, \mathbf{y}_i)_{i\in[n]}$, $[n] := \{1,\ldots,n\}$, arising from an unknown conditional density s_0.

Our proposal: approximating s_0 by a **Gaussian-gated localized mixture of experts (GLoME)** model due to its flexibility and effectiveness [3, 4, 5]:

$$s_{\boldsymbol{\psi}_K}(\mathbf{x}|\mathbf{y}) = \sum_{k=1}^{K} \underbrace{\mathrm{g}_k(\mathbf{y};\boldsymbol{\omega})}_{\text{Gaussian-gated function}} \underbrace{\Phi_D(\mathbf{x};\boldsymbol{\upsilon}_k(\mathbf{y}),\boldsymbol{\Sigma}_k)}_{\text{Gaussian expert}}, \quad \mathrm{g}_k(\mathbf{y};\boldsymbol{\omega}) = \frac{\boldsymbol{\pi}_k \Phi_L(\mathbf{y};\mathbf{c}_k,\boldsymbol{\Gamma}_k)}{\sum_{j=1}^K \boldsymbol{\pi}_j \Phi_L(\mathbf{y};\mathbf{c}_j,\boldsymbol{\Gamma}_j)}, \forall k\in[K], K\in\mathbb{N}^\star, \text{where:}$$

$\boldsymbol{\psi}_K = (\boldsymbol{\omega},\boldsymbol{\upsilon},\boldsymbol{\Sigma}) \in \boldsymbol{\Omega}_K \times \boldsymbol{\Upsilon}_K \times \mathbf{V}_K =: \boldsymbol{\Psi}_K$, $\boldsymbol{\omega} = (\boldsymbol{\pi},\boldsymbol{c},\boldsymbol{\Gamma}) \in (\boldsymbol{\Pi}_{K-1}\times\mathbf{C}_K\times\mathbf{V}'_K) =: \boldsymbol{\Omega}_K$, $\boldsymbol{\Pi}_{K-1} = \left\{(\boldsymbol{\pi}_k)_{k\in[K]} \in (\mathbb{R}^+)^K, \sum_{k=1}^K \boldsymbol{\pi}_k = 1\right\}$, $\mathbf{C}_K/\boldsymbol{\Upsilon}_K$: K-tuples of mean vectors/functions of size $L\times 1/D\times 1$, $\mathbf{V}'_K/\mathbf{V}_K$: K-tuples of elements in $\mathcal{S}_L^{++}/\mathcal{S}_D^{++}$ (space of symmetric positive-definite matrices).

Contributions:

- **Model selection criterion**: estimating the number of mixture components via a penalized maximum likelihood estimator.
- **Non-asymptotic oracle inequality**: providing a lower bound on the penalty such that our estimator satisfies an oracle inequality.

Boundedness assumptions

$$S_m = \Big\{\mathcal{X}\times\mathcal{Y} \ni (\mathbf{x},\mathbf{y}) \mapsto s_{\boldsymbol{\psi}_K}(\mathbf{x}|\mathbf{y}) =: s_m(\mathbf{x}|\mathbf{y}) : \boldsymbol{\psi}_K = (\boldsymbol{\omega},\boldsymbol{\upsilon},\boldsymbol{\Sigma}) \in \widetilde{\boldsymbol{\Omega}}_K \times \boldsymbol{\Upsilon}_K \times \mathbf{V}_K =: \widetilde{\boldsymbol{\Psi}}_K\Big\},$$

$m \in \mathcal{M} = \{K \in [K_{\max}], K_{\max}\in\mathbb{N}^\star\}$,

$\widetilde{\boldsymbol{\Omega}}_K = \{\boldsymbol{\omega}\in\boldsymbol{\Omega}_K : \forall k\in[K], \|\mathbf{c}_k\|_\infty \le A_{\mathbf{c}}, 0 < a_{\boldsymbol{\Gamma}} \le m(\boldsymbol{\Gamma}_k) \le M(\boldsymbol{\Gamma}_k) \le A_{\boldsymbol{\Gamma}}, 0 < a_{\boldsymbol{\pi}} \le \boldsymbol{\pi}_k\}$,

$m(\mathbf{A})/M(\mathbf{A})$: smallest/largest eigenvalues of matrix $\mathbf{A}$,

$\boldsymbol{\Upsilon}_K = \boldsymbol{\Upsilon}_b^K, d_{\boldsymbol{\Upsilon}}\in\mathbb{N}^\star, T_{\boldsymbol{\Upsilon}}\in\mathbb{R}^+$,

$(\varphi_{\boldsymbol{\Upsilon},i})_{i\in[d_{\boldsymbol{\Upsilon}}]}$: collection of bounded functions on $\mathcal{Y}$,

$$\boldsymbol{\Upsilon}_b = \left\{\mathbf{y}\mapsto\left(\sum_{i=1}^{d_{\boldsymbol{\Upsilon}}} \boldsymbol{\alpha}_i^{(j)}\varphi_{\boldsymbol{\Upsilon},i}(\mathbf{y})\right)_{j\in[D]} : \|\boldsymbol{\alpha}\|_\infty \le T_{\boldsymbol{\Upsilon}}\right\},$$

$$\mathbf{V}_K = \Big\{\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_k)_{k\in[K]} = \left(B_k\mathbf{P}_k\mathbf{A}_k\mathbf{P}_k^\top\right)_{k\in[K]} : B_- \le B_k \le B_+,\ \mathbf{P}_k\in SO(D),\ \mathbf{A}_k\in\mathcal{A}(\lambda_-,\lambda_+)\Big\},$$

$B_k = |\boldsymbol{\Sigma}_k|^{1/D}$: volume, $B_-\in\mathbb{R}^+, B_+\in\mathbb{R}^+$,

$\mathbf{P}_k$: eigenvectors of $\boldsymbol{\Sigma}_k \in$ special orthogonal $SO(D)$,

$\mathbf{A}_k$: diagonal matrix of normalized eigenvalues of $\boldsymbol{\Sigma}_k$,

such that $|\mathbf{A}_k| = 1$ and $0 < \forall i\in[D], \lambda_- \le (\mathbf{A}_k)_{i,i} \le \lambda_+$.

Model selection procedure

Goal: find the best model among $(S_m^*)_{m\in\mathcal{M}}$, $\mathcal{M} = [K_{\max}]$ based one $(\mathbf{x}_i,\mathbf{y}_i)_{i\in[n]}$ arising from an forward conditional density s_0^*.

1. For each $m\in\mathcal{M}$: estimate the forward MLE $(\widehat{s}_m^*(\mathbf{y}_i|\mathbf{x}_i))_{i\in[N]}$ by inverse MLE $\widehat{s}_m$ via an inverse regression trick by GLLiM-EM algorithm.
2. Calculate η'-PMLE $\widehat{m}$ with $\mathrm{pen}(m) = \kappa\dim(S_m^*)$.
3. **Large enough but not explicit value for κ!** Asymptotic criteria: AIC: $\kappa = 1$; BIC: $\kappa = \frac{\ln n}{2}$. Non-asymptotic criterion: strong justification for slope heuristic approach in a finite sample setting.

References

[1] Jean-Patrick Baudry, Cathy Maugis, and Bertrand Michel. Slope heuristics: overview and implementation. *Statistics and Computing*, 22(2):455–470, 2012.

[2] Antoine Deleforge, Florence Forbes, and Radu Horaud. High-dimensional regression with gaussian mixtures and partially-latent response variables. *Statistics and Computing*, 25(5):893–911, 2015.

[3] Nhat Ho, Chiao-Yu Yang, and Michael I Jordan. Convergence Rates for Gaussian Mixtures of Experts. *arXiv preprint arXiv:1907.04377*, 2019.

[4] Hien Duy Nguyen, TrungTin Nguyen, Faicel Chamroukhi, and Geoffrey McLachlan. Approximations of conditional probability density functions in Lebesgue spaces via mixture of experts models. *arXiv preprint arXiv:2012.02385*, 2020.

[5] Trung Tin Nguyen, Hien Duy Nguyen, Faicel Chamroukhi, and Florence Forbes. A non-asymptotic penalization criterion for model selection in mixture of experts models. *arXiv preprint arXiv:2104.02640*, 2021.

Non-asymptotic oracle inequality [5]

Theorem. Given a collection $(S_m)_{m\in\mathcal{M}}$ of GLoME models, $\rho\in(0,1)$, $C_1 > 1$, assume that $\Xi = \sum_{m\in\mathcal{M}} e^{-z_m} < \infty, z_m\in\mathbb{R}^+, \forall m\in\mathcal{M}$, and there exist constants C and $\kappa(\rho,C_1) > 0$ s.t. $\forall m\in\mathcal{M}$, $\mathrm{pen}(m) \ge \kappa(\rho,C_1)\left[(C+\ln n)\dim(S_m)+z_m\right]$. Then, the η'-PMLE $\widehat{s}_{\widehat{m}}$, defined by $\widehat{m} = \mathrm{argmin}_{m\in\mathcal{M}}\left(\sum_{i=1}^n -\ln(\widehat{s}_m(\mathbf{x}_i|\mathbf{y}_i)) + \mathrm{pen}(m)\right) + \eta'$, $\widehat{s}_m = \mathrm{argmin}_{s_m\in S_m}\sum_{i=1}^n -\ln(s_m(\mathbf{x}_i|\mathbf{y}_i))$, satisfies, for any $\rho\in(0,1)$, $\mathrm{JKL}_\rho^{\otimes n}(s,t) = \mathbb{E}_{\mathbf{Y}}\left[\frac{1}{n}\sum_{i=1}^n \frac{1}{\rho}\mathrm{KL}\left(s(\cdot|\mathbf{Y}_i), (1-\rho)s(\cdot|\mathbf{Y}_i)+\rho t(\cdot|\mathbf{Y}_i)\right)\right]$,

$$\mathbb{E}\left[\mathrm{JKL}_\rho^{\otimes n}(s_0,\widehat{s}_{\widehat{m}})\right] \le C_1 \inf_{m\in\mathcal{M}}\left(\inf_{s_m\in S_m}\mathrm{KL}^{\otimes n}(s_0,s_m) + \frac{\mathrm{pen}(m)}{n}\right) + \frac{\kappa(\rho,C_1)C_1\Xi}{n} + \frac{\eta+\eta'}{n}.$$

Numerical experiments

➠**Well-Specified (WS)** : $s_0^*\in S_m^*$,

$$s_0^*(y|x) = \frac{\Phi(x;0.2,0.1)\Phi(y;\mathbf{-5x+2},0.09) + \Phi(x;0.8,0.15)\Phi(y;\mathbf{0.1x},0.09)}{\Phi(x;0.2,0.1)+\Phi(x;0.8,0.15)},$$

➠**Misspecified (MS)** : $s_0^*\notin S_m^*$,

$$s_0^*(y|x) = \frac{\Phi(x;0.2,0.1)\Phi(y;\mathbf{x^2-6x+1},0.09) + \Phi(x;0.8,0.15)\Phi(y;\mathbf{-0.4x^2},0.09)}{\Phi(x;0.2,0.1)+\Phi(x;0.8,0.15)}.$$

Estimation by EM (xLLiM package [2]) and model selection via the slope heuristic (capushe package [1]).

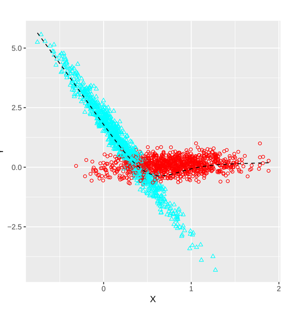
Numerical results:

Fig.1: Clustering deduced from the estimated conditional density of GLoME via the Bayes' optimal allocation rule with 2000 data points. The dash and solid black curves present the true and estimated mean functions.

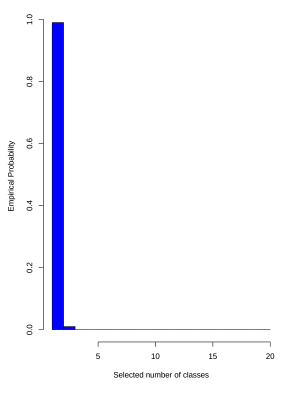
Fig.2: Histogram of selected K using slope heuristic over 100 trials.

Fig.3: Box-plot of the Kullback-Leibler divergence over 100 trials.

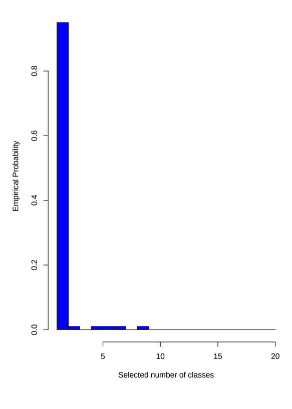
Fig.4: Rate of error decay in a log-log scale, using 30 trials.



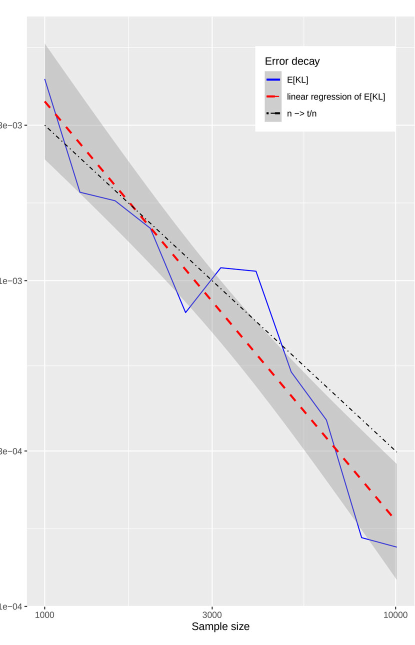
1.1 WS realization



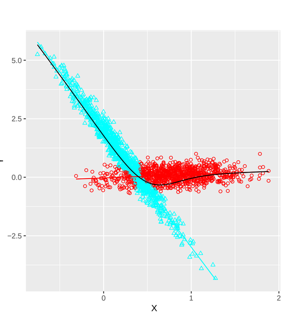
2.1 WS with $n = 2000$



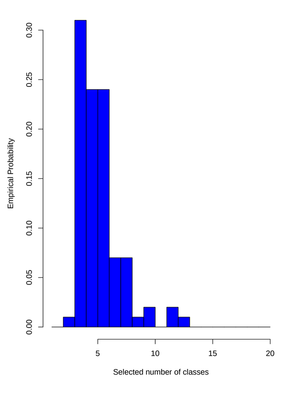
2.2 WS with $n = 10000$



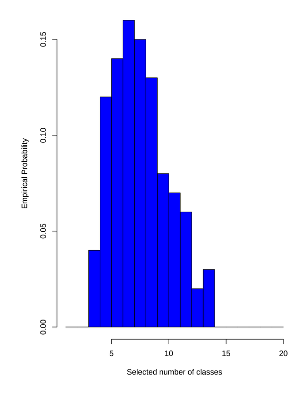
4.1 WS: free regression's slope ≈ -1.287 and $t = 3$.



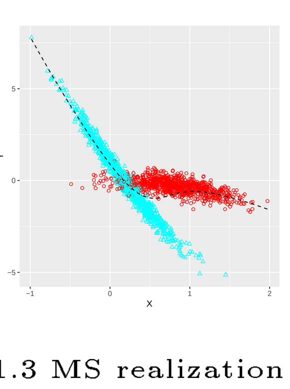
1.2 GLoME clustering



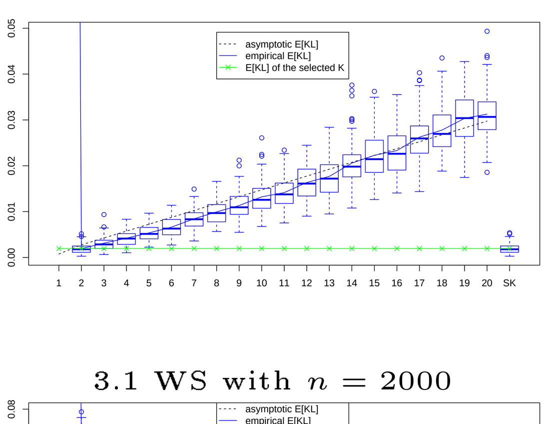
2.3 MS with $n = 2000$



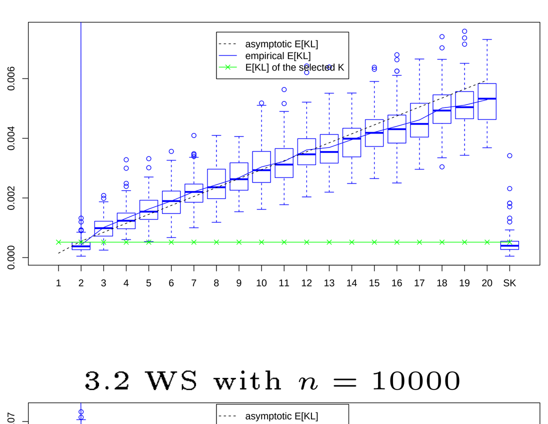
2.4 MS with $n = 10000$



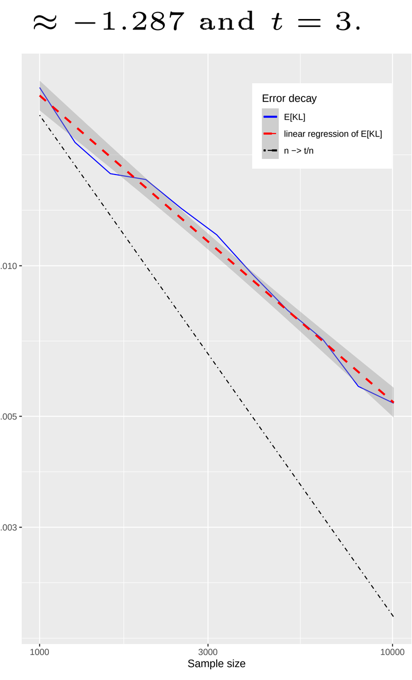
1.3 MS realization



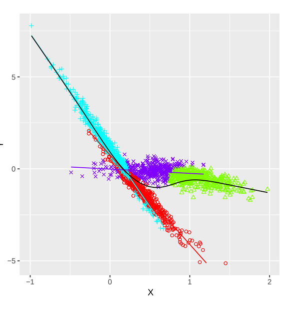
3.1 WS with $n = 2000$



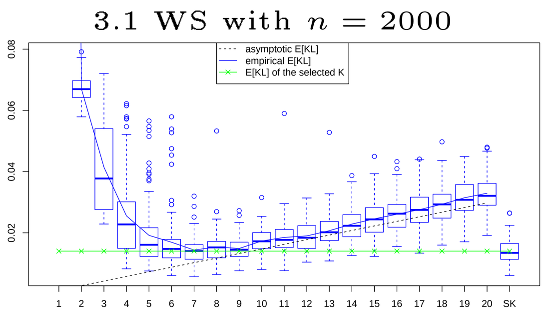
3.2 WS with $n = 10000$



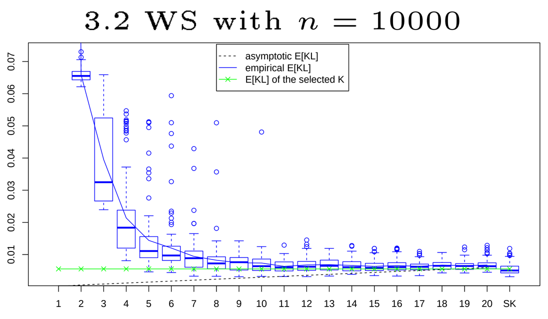
4.2 MS: free regression's slope ≈ -0.6120, $t = 20$.



1.4 GLoME clustering



3.3 MS with $n = 2000$



3.4 MS with $n = 10000$