A non-asymptotic approach fo
via penalization in mixture o

TrungTin Nguyen[1], Hien Duy Nguyen[2], Faicel Cham
[1] *Inria Grenoble Rhone-Alpes, France,* [2] *University of Queensland, Australia*

Learning nonlinear regression models fr

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 multiva
the corresponding observed values $(\mathbf{x}_i, \mathbf{y}_i)_{i\in[n]}$, $[n] := \{1, \dots, n\}$ (potentially
Our proposal: approximating s_0 by a **Gaussian-gated Localized Mix**
[3, 4, 5]:

$$s_{\psi_{K,d}}(\mathbf{x}|\mathbf{y}) = \sum_{k=1}^{K} \underbrace{\mathbf{g}_k(\mathbf{y};\boldsymbol{\omega})}_{\text{Gaussian-gated network}} \times \underbrace{\mathcal{N}_D(\mathbf{x};\boldsymbol{\upsilon}_{k,d}(\mathbf{y}), \boldsymbol{\Sigma}_k}_{\text{Gaussian expert}}$$

$\psi_{K,d} = (\boldsymbol{\omega}, \boldsymbol{\upsilon}, \boldsymbol{\Sigma}) \in \boldsymbol{\Omega}_K \times \boldsymbol{\Upsilon}_{K,d} \times \mathbf{V}_K =: \boldsymbol{\Psi}_{K,d}$, $\boldsymbol{\omega} = (\boldsymbol{\pi}, \boldsymbol{c}, \boldsymbol{\Gamma}) \in (\boldsymbol{\Pi}_{K-1} \times$

K tuples of mean vectors/functions of size $L \times 1/D \times 1$, $\mathbf{V}'$ $/\mathbf{V}_K$: K tupl

or model selection
f experts models

aroukhi[3], Florence Forbes[1]
a, [3] *UNICAEN, LMNO UMR CNRS, France.*

rom complex data using GLoME models

ariate response $\mathbf{Y} = (\mathbf{Y}_j)_{j\in[L]}$ and the set of covariates $\mathbf{X} = (\mathbf{X}_j)_{j\in[D]}$ with
y $D \gg L$), arising from an unknown conditional density s_0.
xture of Experts (GLoME) model due to its flexibility and effectiveness

$$), \quad \mathbf{g}_k(\mathbf{y};\boldsymbol{\omega}) = \frac{\boldsymbol{\pi}_k \mathcal{N}_L(\mathbf{y};\mathbf{c}_k,\boldsymbol{\Gamma}_k)}{\sum_{l=1}^K \boldsymbol{\pi}_l \mathcal{N}_L(\mathbf{y};\mathbf{c}_l,\boldsymbol{\Gamma}_l)}, \forall k \in [K], K \in \mathbb{N}^\star, \text{where:}$$

$$\mathbf{C}_K \times \mathbf{V}'_K) =: \boldsymbol{\Omega}_K, \boldsymbol{\Pi}_{K-1} = \left\{(\boldsymbol{\pi}_k)_{k\in[K]} \in (\mathbb{R}^+)^K, \sum_{k=1}^K \boldsymbol{\pi}_k = 1\right\}, \mathbf{C}_K/\boldsymbol{\Upsilon}_{K,d}:$$

es of elements in $\mathcal{S}^{++}/\mathcal{S}^{++}$ (space of symmetric positive definite matrices)

K-tuples of mean vectors/functions of size $L \times 1$/$D \times 1$, $\mathbf{V}_K$/$\mathbf{V}_K$. K-tupl

Main contributions:

- **Model selection criterion**: choosing number of mixture component
- **Finite-sample oracle inequality**: establishing non-asymptotic risk

Boundedness assumptions

$$\widetilde{\mathbf{\Omega}}_K = \{\boldsymbol{\omega} \in \mathbf{\Omega}_K : \forall k \in [K], \; \|\mathbf{c}_k\|_\infty \leq A_{\mathbf{c}},$$

$$0 < a_{\mathbf{\Gamma}} \leq m\left(\mathbf{\Gamma}_k\right) \leq M\left(\mathbf{\Gamma}_k\right) \leq A_{\mathbf{\Gamma}}, 0 < a_{\boldsymbol{\pi}} \leq \boldsymbol{\pi}_k\},$$

$m(\mathbf{\Gamma}_k)/M(\mathbf{\Gamma}_k)$: the smallest/largest eigenvalues of $\mathbf{\Gamma}_k$,

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

$\mathbf{\Upsilon}_{K,d} = \otimes_{k \in [K]} \mathbf{\Upsilon}_{k,d} = \mathbf{\Upsilon}_{b,d}^K$, $T_{\mathbf{\Upsilon}} \in \mathbb{R}^+$,

$\left(\varphi_{\mathbf{\Upsilon},i}\right)_{i \in [d]}$: collection of bounded functions on $\mathcal{Y}$,

$$\mathbf{V}_K = \Big\{ \left(\mathbf{\Sigma}_k\right)_{k \in [K]} = \left(B_k \mathbf{P}_k \mathbf{A}_k \mathbf{P}_k^\top\right)_{k \in [K]} :$$

$$0 < B_- \leq B_k \leq B_+, \; \mathbf{P}_k \in SO(D), \; \mathbf{A}_k \in \mathcal{A}\left(\lambda_-, \lambda_+\right) \Big\},$$

$B_k = |\mathbf{\Sigma}_k|^{1/D}$: volume, $SO(D)$: eigenvectors of $\mathbf{\Sigma}_k$,

Theorem. Given a c

$\Xi = \sum_{\mathbf{m} \in \mathcal{M}} e^{-z_{\mathbf{m}}} < \infty$

$\forall \mathbf{m} \in \mathcal{M}$, $\mathrm{pen}(\mathbf{m}) \geq$

$\arg\min_{\mathbf{m} \in \mathcal{M}} \left(\sum_{i=1}^{n} - \ln\right.$

the loss $\mathrm{JKL}_\rho^{\otimes n}(s, t) = $

$$\mathbb{E}_{\mathbf{Y}_{[n]}} \left[\mathrm{JKL}_\rho^{\otimes n}\left(s_0\right.\right.$$

➠**Well-Speci**

$s_0^*(y|x) = \frac{\mathcal{N}(}{}$

es of elements in S_L^{++} / S_D^{++} (space of symmetric positive-definite matrices).

s and mean functions' degree via a penalized maximum likelihood estimator.
bounds provided a lower bound on the penalty holds.

Non-asymptotic oracle inequality [5]

ollection $(S_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$ of GLoME models, $\rho \in (0,1)$, $C_1 > 1$, assume that $\infty, z_{\mathbf{m}} \in \mathbb{R}^+, \forall \mathbf{m} \in \mathcal{M}$, and there exist constants C and $\kappa(\rho, C_1) > 0$ s.t. $\kappa(\rho, C_1)\left[(C + \ln n)\dim(S_{\mathbf{m}}) + z_{\mathbf{m}}\right]$. Then, a PMLE-$\widehat{s}_{\widehat{\mathbf{m}}}$, defined by $\widehat{\mathbf{m}} =$ $\ln\left(\widehat{s}_{\mathbf{m}}(\mathbf{x}_i|\mathbf{y}_i)\right) + \mathrm{pen}(\mathbf{m})\big)$, $\widehat{s}_{\mathbf{m}} = \arg\min_{s_{\mathbf{m}}\in S_{\mathbf{m}}} \sum_{i=1}^n -\ln\left(s_{\mathbf{m}}(\mathbf{x}_i|\mathbf{y}_i)\right)$, with $\mathbb{E}_{\mathbf{Y}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^n \frac{1}{\rho}\mathrm{KL}\left(s(\cdot|\mathbf{Y}_i), (1-\rho)s(\cdot|\mathbf{Y}_i) + \rho t(\cdot|\mathbf{Y}_i)\right)\right]$, satisfies

$$, \widehat{s}_{\widehat{\mathbf{m}}})\big] \leq C_1 \inf_{\mathbf{m}\in\mathcal{M}}\left(\inf_{s_{\mathbf{m}}\in S_{\mathbf{m}}} \mathrm{KL}^{\otimes n}(s_0, s_{\mathbf{m}}) + \frac{\mathrm{pen}(\mathbf{m})}{n}\right) + \frac{\kappa(\rho, C_1)\, C_1 \Xi}{n}.$$

Numerical experiments

fied (WS) : $s_0^* \in S_{\mathbf{m}}^*$,

$x; 0.2, 0.1)\mathcal{N}(y; \mathbf{-5x+2}, 0.09) + \mathcal{N}(x; 0.8, 0.15)\mathcal{N}(y; \mathbf{0.1x}, 0.09)$

$\mathcal{A}(\lambda_-, \lambda_+)$: set of diagonal matrices of normalized eigenvalues of $\boldsymbol{\Sigma}_k$ s.t. $\forall i \in [D], 0 < \lambda_- \leq (\mathbf{A}_k)_{i,i} \leq \lambda_+$,

$\mathbf{m} \in \mathcal{M} = \{(K,d) : K \in [K_{\max}], d \in [d_{\max}]\}$,

$$S_{\mathbf{m}} = \Big\{(\mathbf{x},\mathbf{y}) \mapsto s_{\boldsymbol{\psi}_{K,d}}(\mathbf{x}|\mathbf{y}) =: s_{\mathbf{m}}(\mathbf{x}|\mathbf{y}) : \boldsymbol{\psi}_{K,d} \in \widetilde{\boldsymbol{\Omega}}_K \times \boldsymbol{\Upsilon}_{K,d} \times \mathbf{V}_K =: \widetilde{\boldsymbol{\Psi}}_K\Big\}.$$

Model selection procedure

GLLiM model: finding the best data-driven model among $(S^*_{\mathbf{m}})_{m\in\mathcal{M}}$, $\mathcal{M} = [K_{\max}] \times \{1\}$, based on $(\mathbf{x}_i, \mathbf{y}_i)_{i\in[n]}$ arising from a forward conditional density s^*_0.

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

$s_0(y|x)$

➠ **Misspecifi**

$s^*_0(y|x) = \frac{\mathcal{N}(x}{}$

Estimation by EM (xLLi

Numerical results:

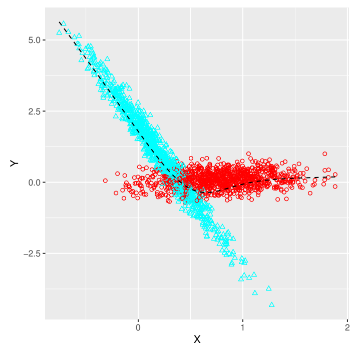
Fig.1: Clustering deduced

rule with 2000 data point

Fig.2: Histogram of selec

Fig.3: Box-plot of the Ku

Fig.4: Rate of error uppe



1.1 WS realization

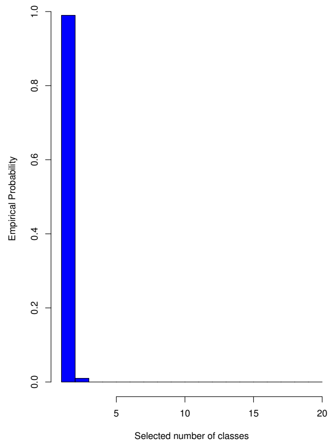
$$\mathcal{N}(x;0.2,0.1)+\mathcal{N}(x;0.8,0.15)$$,

ed (MS) : $s_0^* \notin S_{\mathbf{m}}^*$,

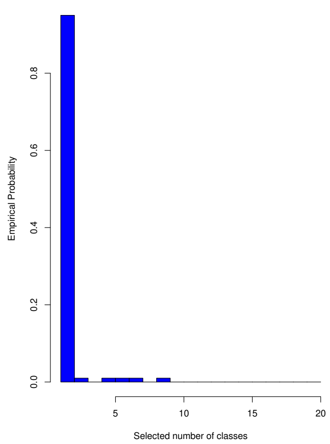
$$\frac{x;0.2,0.1)\mathcal{N}(y;\boldsymbol{x^2-6x+1},0.09)+\mathcal{N}(x;0.8,0.15)\mathcal{N}(y;\boldsymbol{-0.4x^2},0.09)}{\mathcal{N}(x;0.2,0.1)+\mathcal{N}(x;0.8,0.15)}.$$

M package [2]) and model selection via the slope heuristic (capushe package [1]).

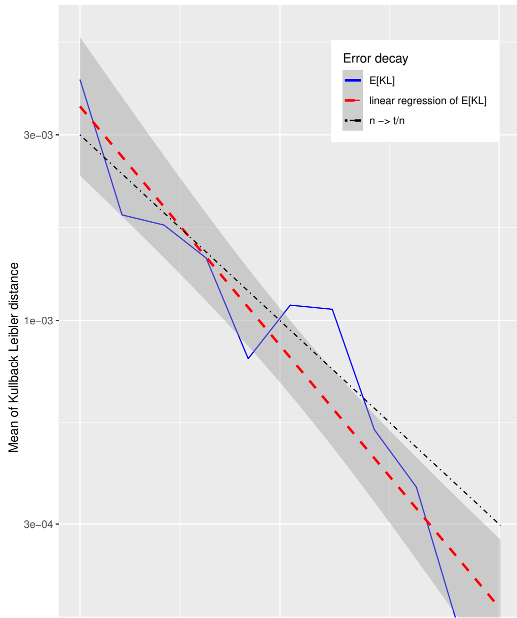
l from the estimated conditional density of GLoME via the Bayes' optimal allocation
s. The dash and solid black curves present the true and estimated mean functions.
ted K using slope heuristic over 100 trials.
ullback–Leibler divergence over 100 trials.
r bound decay in a log-log scale, using 30 trials.

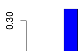


2.1 WS with $n = 2000$



2.2 WS with $n = 10000$





References

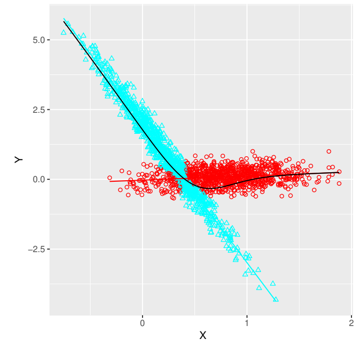
[1] Jean-Patrick Baudry, Cathy Maugis, and Bertrand Michel. Slope heuristics: overview and implementation. *Statistics and Computing*, 22(2):455–470, 2012.

[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.

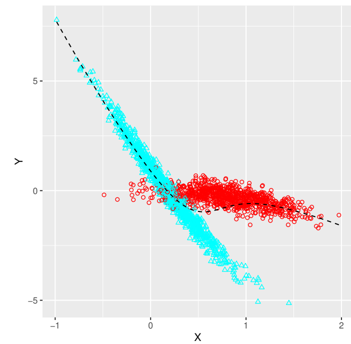
[3] Nhat Ho, Chiao-Yu Yang, and Michael I Jordan. Convergence Rates for Gaussian Mixtures of Experts. *arXiv preprint arXiv:1907.04377*, 2019.

[4] Hien Duy Nguyen, TrungTin Nguyen, Faicel Chamroukhi, and Geoffrey John McLachlan. Approximations of conditional probability density functions in Lebesgue spaces via mixture of experts models. *Journal of Statistical Distributions and Applications*, 8(1):13, 2021.

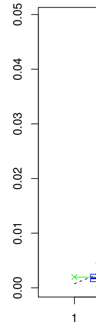
[5] Trung Tin Nguyen, Hien Duy Nguyen, Faicel Chamroukhi, and Florence Forbes. A non-asymptotic penalization criterion for model selection in mixture of experts models. *arXiv preprint arXiv:2104.02640*, 2021.

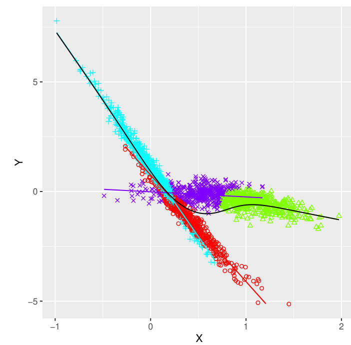


1.2 GLoME clustering

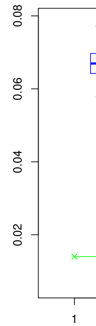


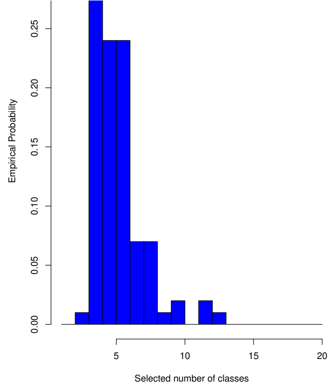
1.3 MS realization



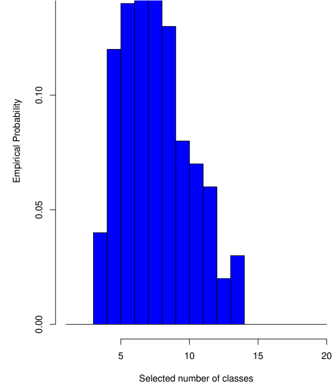


1.4 GLoME clustering

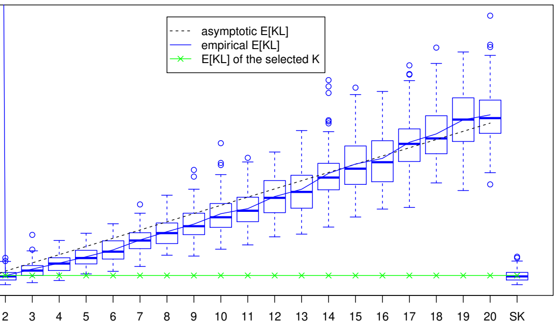




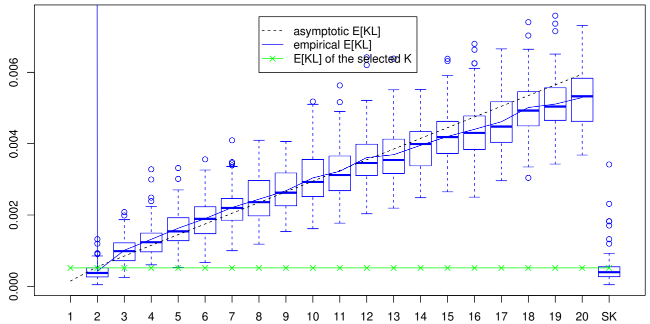
2.3 MS with $n = 2000$



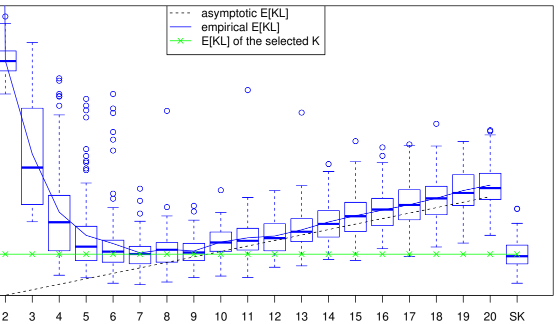
2.4 MS with $n = 10000$



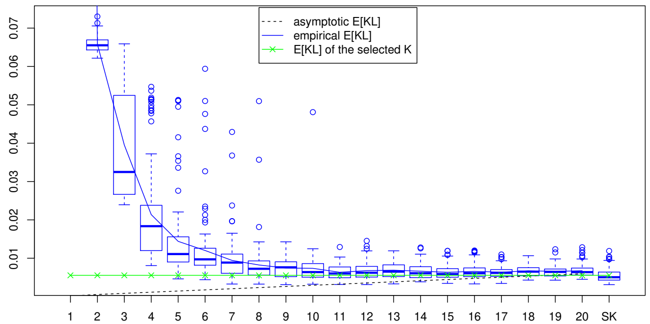
3.1 WS with $n = 2000$



3.2 WS with $n = 10000$



3.3 MS with $n = 2000$

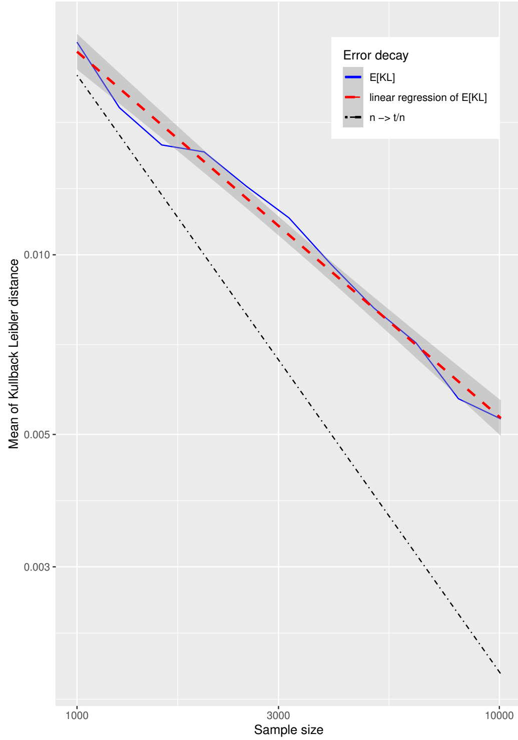


3.4 MS with $n = 10000$

4.1 WS:

free regression's slope

≈ -1.287 and $t = 3$.



4.2 MS:

free regression's slope

≈ -0.6120, $t = 20$.