

Approximation and non-asymptotic model selection in mixture of experts models

TrungTin Nguyen¹, Faicel Chamroukhi¹, Hien Duy Nguyen², Florence Forbes³

¹UNICAEN, LMNO UMR CNRS, France, ²University of Queensland, Australia, ³Inria Grenoble Rhone-Alpes, France.



Learning nonlinear regression models from complex data using GLoME models

Random sample: $(\mathbf{X}_i, \mathbf{Y}_i)_{i \in [n]} \in (\mathcal{X} \times \mathcal{Y})^n \subset (\mathbb{R}^D \times \mathbb{R}^L)^n$ 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 $(\mathbf{x}_i, \mathbf{y}_i)_{i \in [n]}$, $[n] := \{1, \dots, n\}$ (potentially $D \gg L$), 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]:

$$s_{\psi_K}(\mathbf{x}|\mathbf{y}) = \sum_{k=1}^K \underbrace{\mathbf{g}_k(\mathbf{y}; \boldsymbol{\omega})}_{\text{Gaussian-gated network}} \times \underbrace{\Phi_D(\mathbf{x}; \mathbf{v}_k(\mathbf{y}), \boldsymbol{\Sigma}_k)}_{\text{Gaussian expert}}, \quad \mathbf{g}_k(\mathbf{y}; \boldsymbol{\omega}) = \frac{\pi_k \Phi_L(\mathbf{y}; \mathbf{c}_k, \boldsymbol{\Gamma}_k)}{\sum_{j=1}^K \pi_j \Phi_L(\mathbf{y}; \mathbf{c}_j, \boldsymbol{\Gamma}_j)}, \forall k \in [K], K \in \mathbb{N}^*, \text{ where:}$$

$\psi_K = (\boldsymbol{\omega}, \mathbf{v}, \boldsymbol{\Sigma}) \in \Omega_K \times \Upsilon_K \times \mathbf{V}_K =: \Psi_K$, $\boldsymbol{\omega} = (\boldsymbol{\pi}, \mathbf{c}, \boldsymbol{\Gamma}) \in (\Pi_{K-1} \times \mathbf{C}_K \times \mathbf{V}'_K) =: \Omega_K$, $\Pi_{K-1} = \left\{ (\pi_k)_{k \in [K]} \in (\mathbb{R}^+)^K, \sum_{k=1}^K \pi_k = 1 \right\}$, $\mathbf{C}_K / \Upsilon_K$: 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:** choosing number of mixture components and mean functions' degree 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

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

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

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

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

$$\Upsilon_K = \Upsilon_b^K, T_\Upsilon \in \mathbb{R}^+,$$

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

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

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

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

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

$m \in \mathcal{M} = \{(K, d_\Upsilon) : K \in [K_{\max}], K_{\max}, d_\Upsilon \in \mathbb{N}^*\}$,

$S_m = \left\{ \mathcal{X} \times \mathcal{Y} \ni (\mathbf{x}, \mathbf{y}) \mapsto s_{\psi_K}(\mathbf{x}|\mathbf{y}) =: s_m(\mathbf{x}|\mathbf{y}) : \right.$

$$\left. \psi_K \in \tilde{\Omega}_K \times \Upsilon_K \times \mathbf{V}_K =: \tilde{\Psi}_K \right\}.$$

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. $\forall m \in \mathcal{M}$, $\text{pen}(m) \geq \kappa(\rho, C_1) [(C + \ln n) \dim(S_m) + z_m]$. Then, the η' -PMLE $\hat{s}_{\hat{m}}$, defined by $\hat{m} = \arg \min_{m \in \mathcal{M}} (\sum_{i=1}^n -\ln(\hat{s}_m(\mathbf{x}_i|\mathbf{y}_i)) + \text{pen}(m)) + \eta'$, $\hat{s}_m = \arg \min_{s_m \in S_m} \sum_{i=1}^n -\ln(s_m(\mathbf{x}_i|\mathbf{y}_i))$, with the loss $\text{JKL}_\rho^{\otimes n}(s, t) = \mathbb{E}_{\mathbf{Y}} \left[\frac{1}{n} \sum_{i=1}^n \frac{1}{\rho} \text{KL}(s(\cdot|\mathbf{Y}_i), (1-\rho)s(\cdot|\mathbf{Y}_i) + \rho t(\cdot|\mathbf{Y}_i)) \right]$, satisfies

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

Numerical experiments

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

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

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

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

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

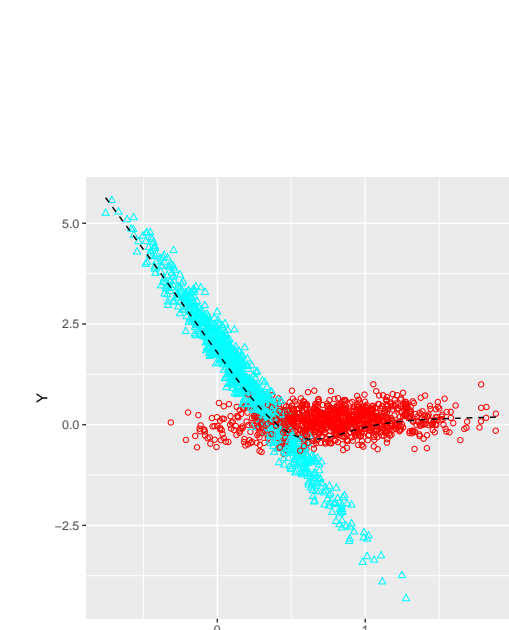
Numerical results:

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.

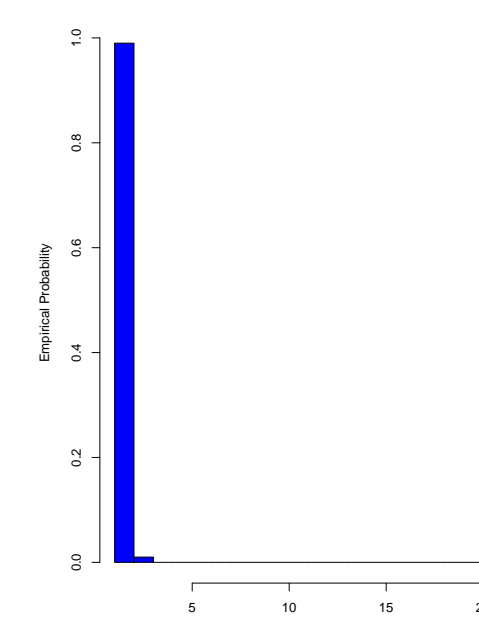
Fig.2: Histogram of selected K using slope heuristic over 100 trials.

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

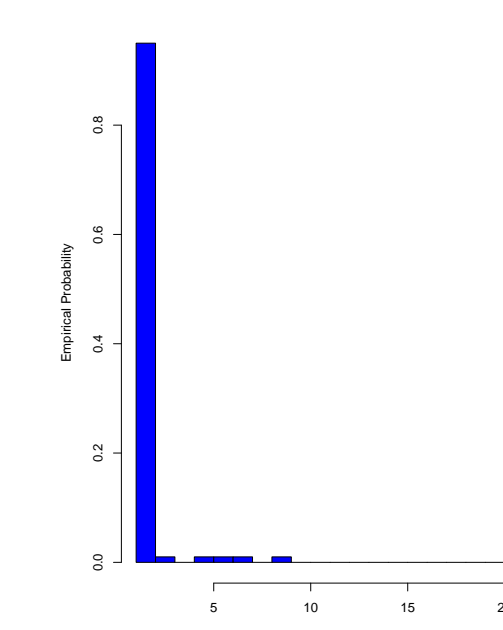
Fig.4: Rate of error decay in a log-log scale, using 30 trials.



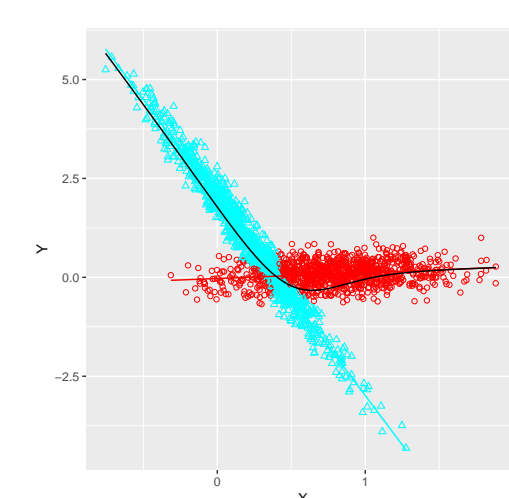
1.1 WS realization



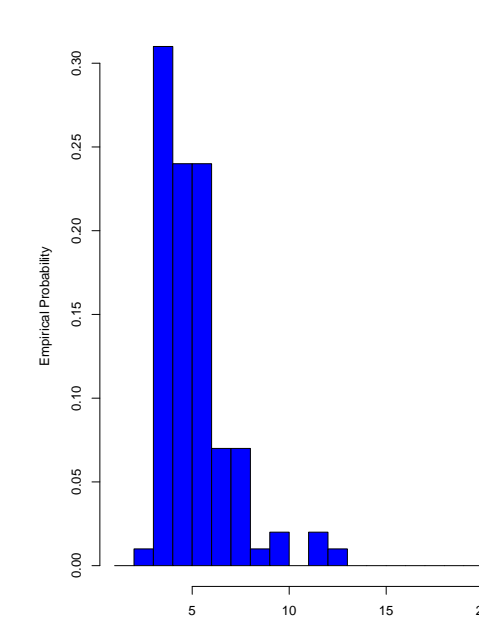
2.1 WS with $n = 2000$



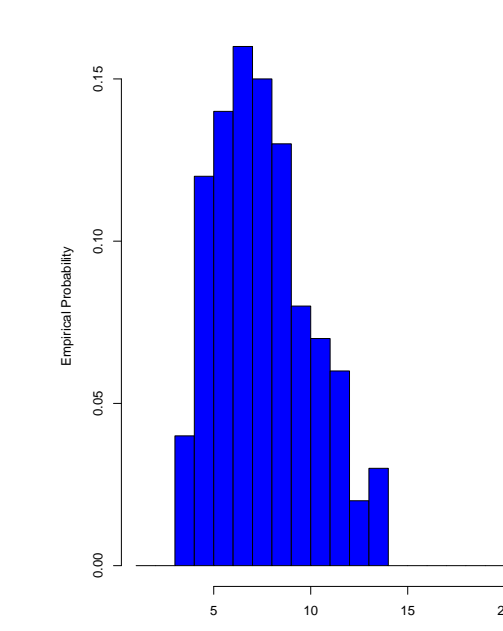
2.2 WS with $n = 10000$



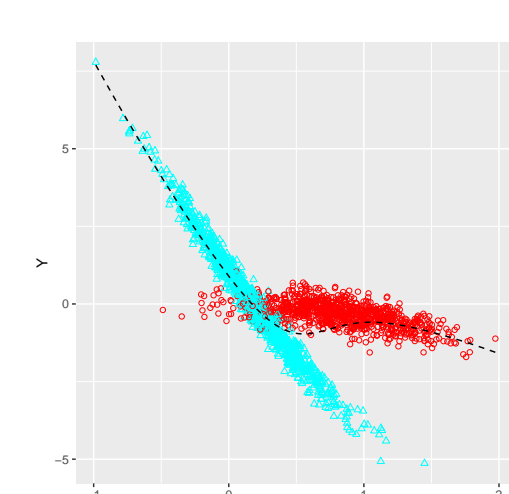
1.2 GLoME clustering



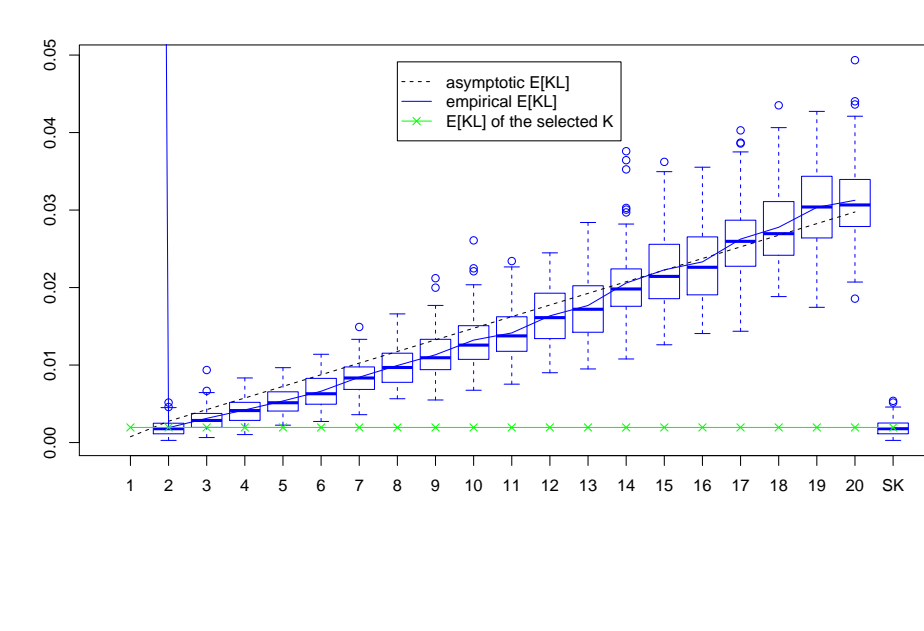
2.3 MS with $n = 2000$



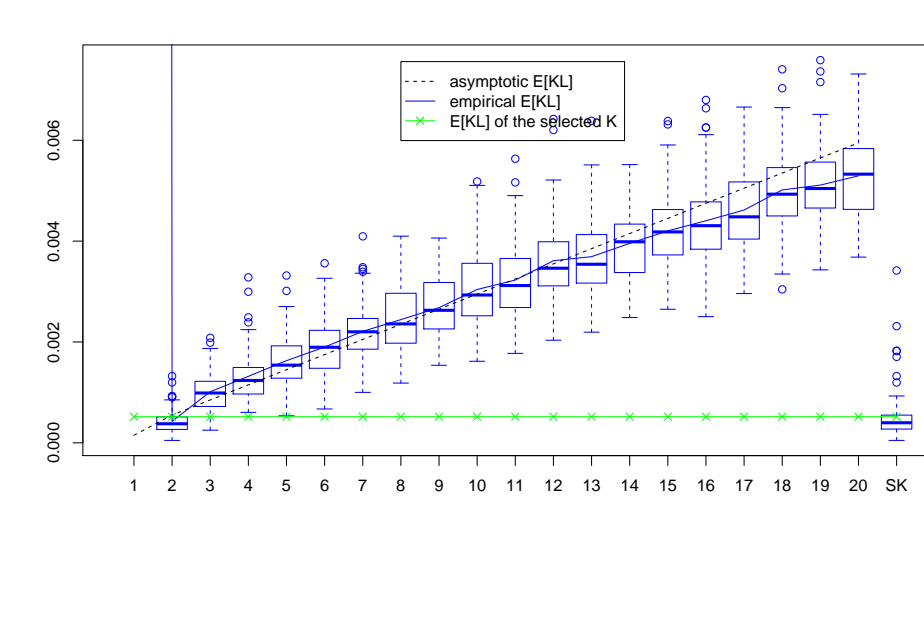
2.4 MS with $n = 10000$



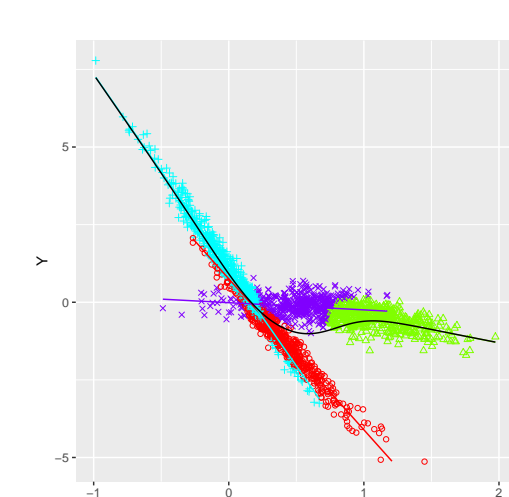
1.3 MS realization



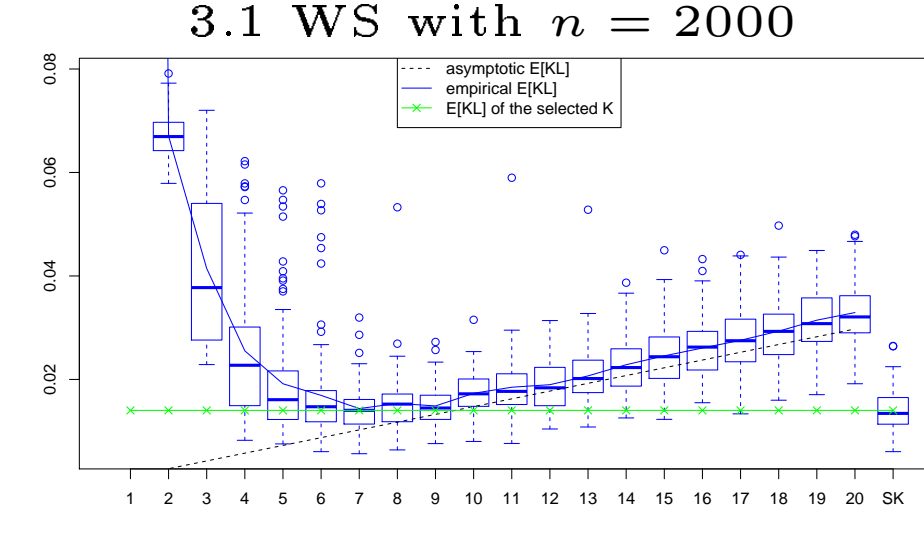
3.1 WS with $n = 2000$



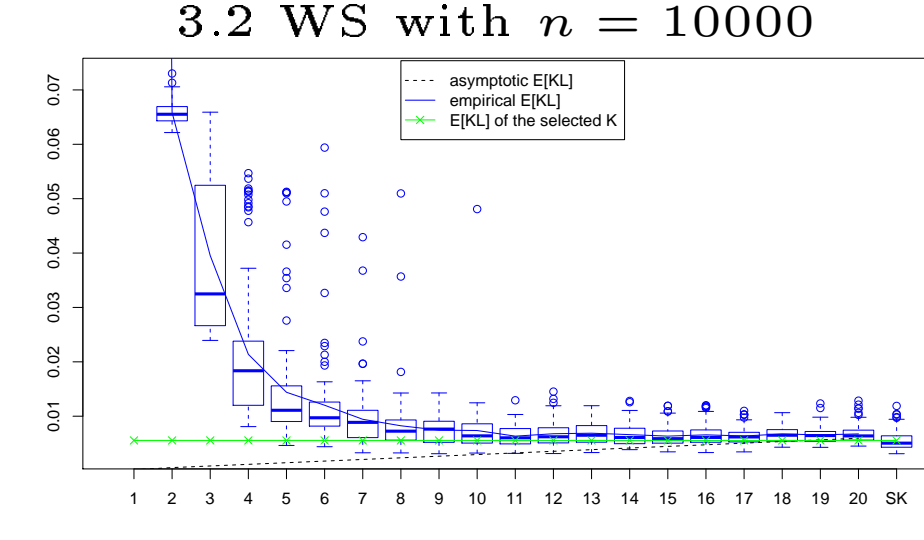
3.2 WS with $n = 10000$



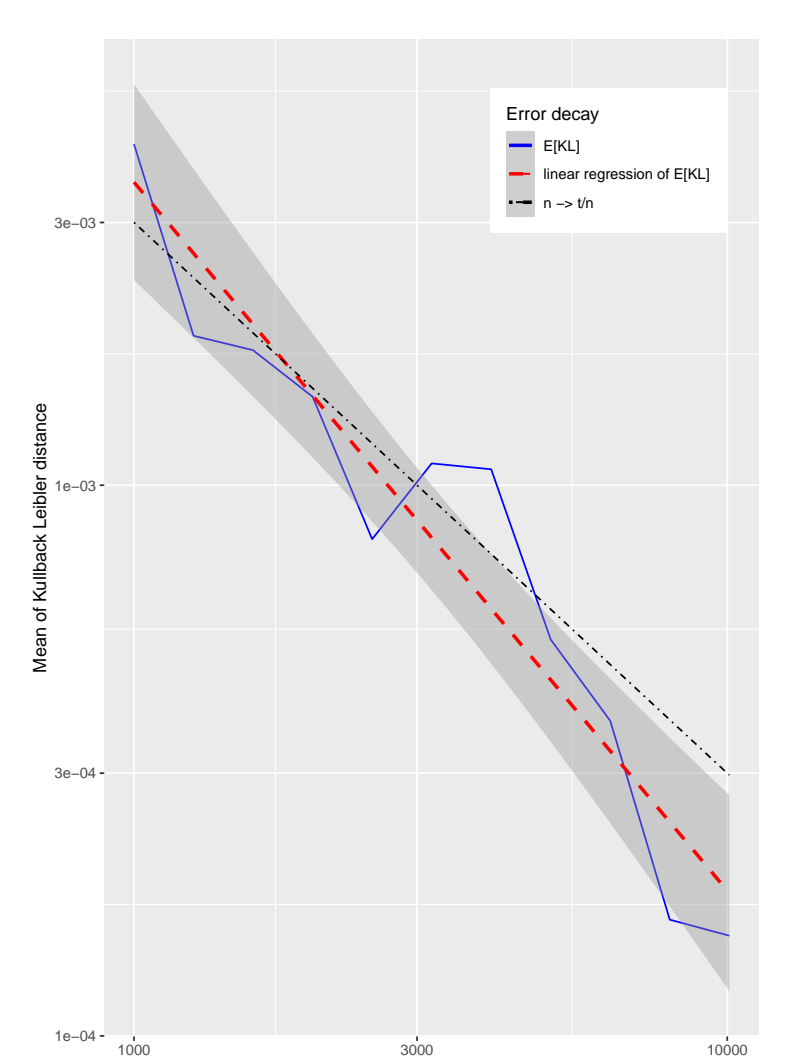
1.4 GLoME clustering



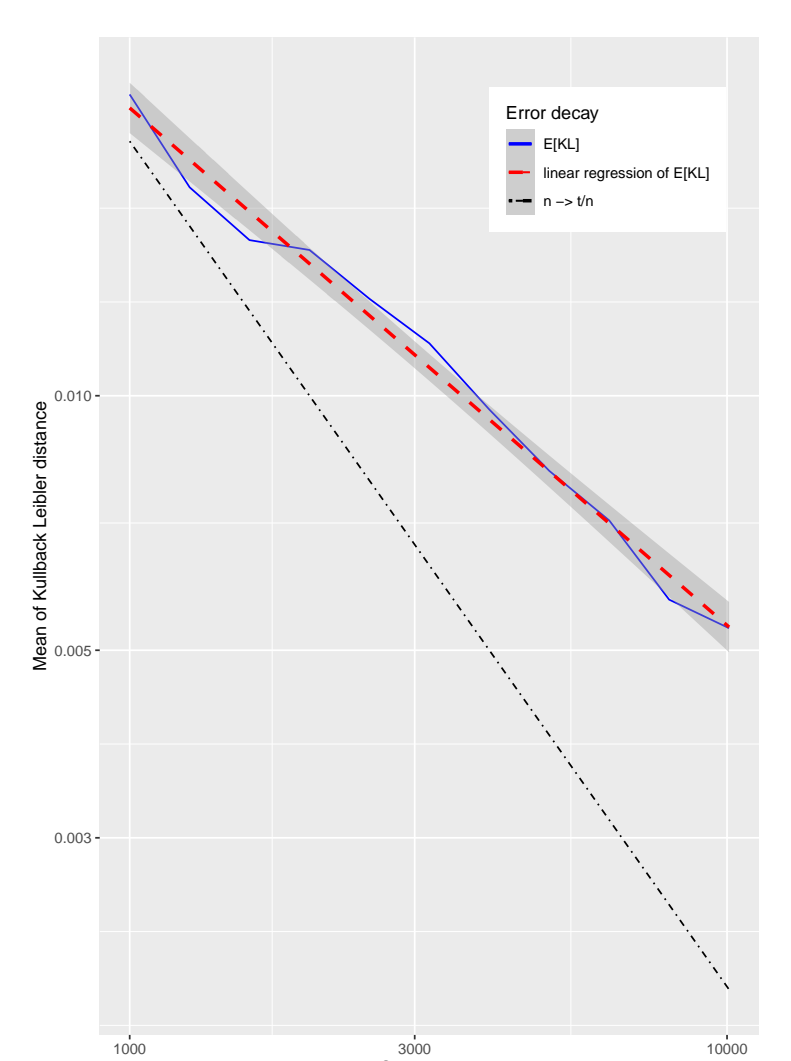
3.3 MS with $n = 2000$



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 , $t = 20$.

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.