Model selection by penalization in mixture of experts models with a non-asymptotic approach

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

[1] *Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France*
trung-tin.nguyen@inria.fr, florence.forbes@inria.fr
[2] *Normandie Univ, UNICAEN, CNRS, LMNO, 14000 Caen, France.*
faicel.chamroukhi@unicaen.fr
[3] *School of Mathematics and Physics, University of Queensland, St. Lucia, Australia.*
h.nguyen7@uq.edu.au

Résumé. Cette étude est consacrée au problème de la sélection de modèles parmi une collection de modèles de mélanges d'experts avec experts gaussiens et fonctions d'activations gaussiennes normalisées, caractérisés par le nombre de composantes du mélange et la complexité des experts moyens, dans un cadre d'estimation par maximum de vraisemblance pénalisée. En particulier, nous établissons des limites de risque non asymptotiques qui prennent la forme d'inégalités oracles faibles, sous une condition de limite inférieure pour la pénalité. Leur bon comportement empirique est ensuite démontré en simulation et sur des données réelles.

Mots-clés. Mélange d'experts, sélection de modèle, maximum de vraisemblance pénalisée.

Abstract. This study is devoted to the problem of model selection among a collection of Gaussian-gated localized mixtures of experts models characterized by the number of mixture components, and the complexity of Gaussian mean experts, in a penalized maximum likelihood estimation framework. In particular, we establish non-asymptotic risk bounds that take the form of weak oracle inequalities, provided that lower bounds of the penalties hold. Their good empirical behavior is then demonstrated on synthetic and real datasets.

Keywords. Mixture of experts, model selection, penalized maximum likelihood.

1 Introduction

Mixture of experts (MoE) models, originally introduced as neural network architectures in Jacobs et al. (1991), are flexible models that generalize the classical finite mixture and finite mixtures of regression models. The popularity of these conditional mixture density models arise largely due to their universal approximation properties, which have been studied in Nguyen et al. (2020b, 2021b), and which improve upon approximation

capabilities of unconditional finite mixture models, as studied in Nguyen et al. (2016), Ho et al. (2019), Nguyen et al. (2019, 2021a). Detailed reviews on the practical and theoretical aspects of MoE models can be found in Nguyen and Chamroukhi (2018), Nguyen (2021).

In this work, we study the class of MoE models with Gaussian experts and normalized Gaussian gating functions for clustering and regression, first introduced by Xu et al. (1995) and recently studied numerically in Chamroukhi et al. (2019). From hereon in, we refer to such models as *Gaussian-gated localized MoE* (GLoME) models. These models are useful to learn potentially nonlinear relationships between a multivariate output and a high-dimensional input issued from a heterogeneous population. This involves performing regression, clustering and model selection, simultaneously. While estimation can be performed using standard Expectation-Maximisation algorithms, it crucially depends and requires hyperparameter choices, including the number of mixture components (or clusters), and the degree of complexity of each Gaussian expert's mean function.

Traditional model selection criteria such as AIC, BIC, or ICL-BIC are based on asymptotic theory or Bayesian approaches. In contrast, the present contribution is to provide a finite-sample oracle inequality indicative of the quality of a data-driven selected GLoME model with respect to the true model. More specifically, we establish a non-asymptotic risk bound that takes the form of a weak oracle inequality, provided that a lower bound on the penalty holds true. Our non-asymptotic risk bound allows the number n of observations to be fixed while the dimensionality and cardinality of the models, characterized by the number of covariates and the dimension of the response, are allowed to grow with respect to n, and can be much larger than n. To the best of our knowledge, this is the most recent and advanced effort in literature to develop a finite-sample oracle inequality for the framework of MoE regression models.

From now on, we are interested in estimating the law of the random variable $\mathbf{Y}$, conditionally on $\mathbf{X}$, respectively of dimension L and D. Subsequently, $\left(\mathbf{X}_{[n]}, \mathbf{Y}_{[n]}\right) := (\mathbf{X}_i, \mathbf{Y}_i)_{i\in[n]}$, $[n] = \{1, \ldots, n\}, n \in \mathbb{N}^\star$, denotes a random sample, and $\mathbf{x}$ and $\mathbf{y}$ stands for the observed values of the random variables $\mathbf{X}$ and $\mathbf{Y}$, respectively. For a matrix $\mathbf{A}$, let $m(\mathbf{A})$ and $M(\mathbf{A})$ be, respectively, the modulus of the smallest and largest eigenvalues of $\mathbf{A}$.

2 Collection of GLoME models

We consider models with inverse conditional PDFs of the form (2.1). Such models have been considered in Xu et al. (1995), Deleforge et al. (2015) and are very useful in a high dimensional regression context, where typically $D \gg L$.

$$s_{\boldsymbol{\psi}_{K,d}}(\mathbf{x} \mid \mathbf{y}) = \sum_{k=1}^{K} g_k\left(\mathbf{y}; \boldsymbol{\omega}\right) \phi_D\left(\mathbf{x}; \boldsymbol{\upsilon}_{k,d}(\mathbf{y}), \boldsymbol{\Sigma}_k\right); \; g_k\left(\mathbf{y}; \boldsymbol{\omega}\right) = \frac{\pi_k \phi_L\left(\mathbf{y}; \mathbf{c}_k, \boldsymbol{\Gamma}_k\right)}{\sum_{j=1}^{K} \pi_j \phi_L\left(\mathbf{y}; \mathbf{c}_j, \boldsymbol{\Gamma}_j\right)}. \quad (2.1)$$

Here, $g_k(\cdot;\boldsymbol{\omega})$ and $\phi_D(\cdot;\boldsymbol{\upsilon}_{k,d}(\cdot),\boldsymbol{\Sigma}_k)$, $k \in [K]$, $K \in \mathbb{N}^\star$, $d \in \mathbb{N}^\star$, are called normalized Gaussian gating functions and Gaussian experts, respectively. Furthermore, we decompose the parameters of the model as follows: $\boldsymbol{\psi}_{K,d} = (\boldsymbol{\omega}, \boldsymbol{\upsilon}_d, \boldsymbol{\Sigma}) \in \boldsymbol{\Omega}_K \times \boldsymbol{\Upsilon}_{K,d} \times \mathbf{V}_K =: \boldsymbol{\Psi}_{K,d}$, $\boldsymbol{\omega} = (\boldsymbol{\pi}, \boldsymbol{c}, \boldsymbol{\Gamma}) \in (\boldsymbol{\Pi}_{K-1} \times \mathbf{C}_K \times \boldsymbol{V}'_K) =: \boldsymbol{\Omega}_K$, $\boldsymbol{\pi} = (\pi_k)_{k\in[K]}$, $\boldsymbol{c} = (\mathbf{c}_k)_{k\in[K]}$, $\boldsymbol{\Gamma} = (\boldsymbol{\Gamma}_k)_{k\in[K]}$, $\boldsymbol{\upsilon}_d = (\boldsymbol{\upsilon}_{k,d})_{k\in[K]} \in \boldsymbol{\Upsilon}_{K,d}$, and $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_k)_{k\in[K]} \in \mathbf{V}_K$. Note that $\boldsymbol{\Pi}_{K-1} = \left\{(\pi_k)_{k\in[K]} \in (\mathbb{R}^+)^K, \sum_{k=1}^K \pi_k = 1\right\}$ is a $K-1$ dimensional probability simplex, $\mathbf{C}_K$ is a set of K-tuples of mean vectors of size $L \times 1$, $\boldsymbol{V}'_K$ is a set of K-tuples of elements in $\mathcal{S}_L^{++}$, where $\mathcal{S}_L^{++}$ denotes the collection of symmetric positive definite matrices on $\mathbb{R}^L$, $\boldsymbol{\Upsilon}_{K,d}$ is a set of K-tuples of mean functions from $\mathbb{R}^L$ to $\mathbb{R}^D$ depending on a degree d (*e.g.,* polynomial degree) and $\mathbf{V}_K$ is a set containing K-tuples from $\mathcal{S}_D^{++}$.

In order to establish our finite-sample oracle inequality, Theorem 3.1, we need to explicitly impose some classical boundedness conditions on the parameter space. Specifically, we assume that there exist deterministic positive constants $a_{\boldsymbol{\pi}}, A_{\boldsymbol{c}}, a_{\boldsymbol{\Gamma}}, A_{\boldsymbol{\Gamma}}$, and set

$$\widetilde{\boldsymbol{\Omega}}_K = \left\{\boldsymbol{\omega} \in \boldsymbol{\Omega}_K : \forall k \in [K], \|\mathbf{c}_k\|_\infty \leq A_{\boldsymbol{c}}, a_{\boldsymbol{\Gamma}} \leq m(\boldsymbol{\Gamma}_k) \leq M(\boldsymbol{\Gamma}_k) \leq A_{\boldsymbol{\Gamma}}, a_{\boldsymbol{\pi}} \leq \pi_k\right\}. \quad (2.2)$$

The set $\boldsymbol{\Upsilon}_{K,d}$ will be chosen as a tensor product of compact sets of moderate dimension (*e.g.,* a set of polynomials of degree smaller than d, whose coefficients are smaller in absolute values than $T_{\boldsymbol{\Upsilon}}$). In particular, we focus on the bounded $\mathcal{Y} = [0,1]^L$. In this case, $\varphi_{\boldsymbol{\Upsilon},i}$ can be chosen as monomials with maximum (non-negative) degree d: $\mathbf{y}^{\mathbf{r}} = \prod_{l=1}^L \mathbf{y}_l^{\mathbf{r}_l}$. Then, $\boldsymbol{\Upsilon}_{K,d} = \boldsymbol{\Upsilon}_{p,d}^K$, where

$$\boldsymbol{\Upsilon}_{p,d} = \left\{\mathbf{y} \mapsto \left(\sum_{|\mathbf{r}|=0}^{d} \boldsymbol{\alpha}_{\mathbf{r}}^{(j)} \mathbf{y}^{\mathbf{r}}\right)_{j\in[D]} =: (\boldsymbol{\upsilon}_{d,j}(\mathbf{y}))_{j\in[D]} : \|\boldsymbol{\alpha}\|_\infty \leq T_{\boldsymbol{\Upsilon}}\right\}. \quad (2.3)$$

For GLoME models, note that any covariance matrix $\boldsymbol{\Sigma}_k$ can be decomposed into the form $B_k \mathbf{P}_k \mathbf{A}_k \mathbf{P}_k^\top$ such that: $B_k = |\boldsymbol{\Sigma}_k|^{1/D}$ is a positive scalar corresponding to the volume, $\mathbf{P}_k$ is the matrix of eigenvectors of $\boldsymbol{\Sigma}_k$ and $\mathbf{A}_k$ the diagonal matrix of normalized eigenvalues of $\boldsymbol{\Sigma}_k$; $B_- \in \mathbb{R}^+, B_+ \in \mathbb{R}^+$, $\mathcal{A}(\lambda_-, \lambda_+)$ is a set of diagonal matrices $\mathbf{A}_k$, such that $|\mathbf{A}_k| = 1$ and $\forall i \in [D], \lambda_- \leq (\mathbf{A}_k)_{i,i} \leq \lambda_+$, where $\lambda_-, \lambda_+ \in \mathbb{R}$; and $SO(D)$ is the special orthogonal group of dimension D. In this way, we obtain the classical covariance matrix parameterization, described by Celeux and Govaert (1995) for Gaussian parsimonious clustering models, defined by

$$\mathbf{V}_K = \left\{\left(B_k \mathbf{P}_k \mathbf{A}_k \mathbf{P}_k^\top\right)_{k\in[K]} : B_- \leq B_k \leq B_+, \mathbf{P}_k \in SO(D), \mathbf{A}_k \in \mathcal{A}(\lambda_-, \lambda_+)\right\}. \quad (2.4)$$

For GLoME, we need to choose the degree of polynomials d and the number of components K among finite sets $\mathcal{D}_{\boldsymbol{\Upsilon}} = [d_{\max}]$ and $\mathcal{K} = [K_{\max}]$, respectively, where $d_{\max} \in \mathbb{N}^\star$ and $K_{\max} \in \mathbb{N}^\star$ may depend on the sample size n. We wish to estimate the unknown

true conditional density s_0 by conditional densities belonging to the following collection of models $(S_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$, $\mathcal{M} = \{(K,d) : K \in \mathcal{K}, d \in \mathcal{D}_{\boldsymbol{\Upsilon}}\}$,

$$S_{\mathbf{m}} = \left\{(\mathbf{x},\mathbf{y}) \mapsto s_{\boldsymbol{\psi}_{K,d}}(\mathbf{x} \mid \mathbf{y}) =: s_{\mathbf{m}}(\mathbf{x} \mid \mathbf{y}) : \boldsymbol{\psi}_{K,d} \in \widetilde{\boldsymbol{\Omega}}_K \times \boldsymbol{\Upsilon}_{K,d} \times \mathbf{V}_K\right\}, \tag{2.5}$$

where $\widetilde{\boldsymbol{\Omega}}_K$, $\boldsymbol{\Upsilon}_{K,d}$ and $\mathbf{V}_K$ are define previously in (2.2), (2.3) and (2.4), respectively.

3 Oracle inequality for collection of GLoME models

In the maximum likelihood approach, the Kullback–Leibler divergence is the most natural loss function. However, to take into account the structure of conditional densities and the random covariates $\left(\mathbf{Y}_{[n]}\right)$, we consider a *tensorized Kullback–Leibler divergence* $\mathrm{KL}^{\otimes \mathrm{n}}$, defined as:

$$\mathrm{KL}^{\otimes \mathrm{n}}(s,t) = \mathbb{E}_{\mathbf{Y}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^{n}\mathrm{KL}\left(s\left(\cdot \mid \mathbf{Y}_i\right), t\left(\cdot \mid \mathbf{Y}_i\right)\right)\right], \tag{3.1}$$

if $s\ dy$ is absolutely continuous w.r.t. $t\ dy$, and $+\infty$ otherwise. We refer to our result as a *weak oracle inequality*, because its statement is based on a smaller divergence, when compared to $\mathrm{KL}^{\otimes \mathrm{n}}$, namely the *tensorized Jensen–Kullback–Leibler divergence* (Cohen and Le Pennec, 2011): given $\rho \in (0,1)$,

$$\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}(s,t) = \mathbb{E}_{\mathbf{Y}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\rho}\mathrm{KL}\left(s\left(\cdot \mid \mathbf{Y}_i\right), (1-\rho)\, s\left(\cdot \mid \mathbf{Y}_i\right) + \rho t\left(\cdot \mid \mathbf{Y}_i\right)\right)\right].$$

In a penalized maximum likelihood estimation context (PMLE), by adding a suitable penalty $\mathrm{pen}(\mathbf{m})$, one hopes to create a trade-off between a good data fit and model complexity. For a given choice of $\mathrm{pen}(\mathbf{m})$, the *selected model* $S_{\widehat{\mathbf{m}}}$ is chosen as the one whose index is an η'-almost minimizer of the sum of the negative log-likelihood and this penalty. That is $S_{\widehat{\mathbf{m}}} = \widehat{s}_{\widehat{\mathbf{m}}}$, satisfying

$$\sum_{i=1}^{n} -\ln\left[\widehat{s}_{\widehat{\mathbf{m}}}\left(\mathbf{x}_i \mid \mathbf{y}_i\right)\right] + \mathrm{pen}\left(\widehat{\mathbf{m}}\right) \leq \inf_{\mathbf{m}\in\mathcal{M}}\left\{\sum_{i=1}^{n} -\ln\left[\widehat{s}_{\mathbf{m}}\left(\mathbf{x}_i \mid \mathbf{y}_i\right)\right] + \mathrm{pen}(\mathbf{m})\right\} + \eta', \tag{3.2}$$

$$\text{where} \quad \sum_{i=1}^{n} -\ln\left[\widehat{s}_{\mathbf{m}}\left(\mathbf{x}_i \mid \mathbf{y}_i\right)\right] \leq \inf_{s_{\mathbf{m}}\in S_{\mathbf{m}}}\sum_{i=1}^{n} -\ln\left[s_{\mathbf{m}}\left(\mathbf{x}_i \mid \mathbf{y}_i\right)\right] + \eta. \tag{3.3}$$

Note that $\widehat{s}_{\widehat{\mathbf{m}}}$ is then called the η'-penalized likelihood estimate and depends on both the error terms η and η'. From hereon in, the term *selected model (estimate) or best data-driven model (estimate)* is used to indicate that it satisfies (3.2).

Theorem 3.1, proved in Nguyen et al. (2021d), provides a lower bound on the penalty function, $\mathrm{pen}(\mathbf{m})$, which guarantees that the PMLE selects a model in the collection that performs almost as well as the best model.

Theorem 3.1 (Finite-sample weak oracle inequality for GLoME models)**.** *Assume that we observe* $\left(\mathbf{x}_{[n]}, \mathbf{y}_{[n]}\right)$*, arising from an unknown conditional density* s_0*. Given a collection of GLoME models,* $(S_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$*, there is a constant* C *such that for any* $\rho \in (0,1)$*, for any* $\mathbf{m} \in \mathcal{M}$*,* $z_{\mathbf{m}} \in \mathbb{R}^+$*,* $\Xi = \sum_{\mathbf{m}\in\mathcal{M}} e^{-z_{\mathbf{m}}} < \infty$ *and any* $C_1 > 1$*, there is a constant* κ *depending only on* ρ *and* C_1*, such that if for every index* $\mathbf{m} \in \mathcal{M}$*,*

$$pen(\mathbf{m}) > \kappa \left[(C + \ln n) \dim\left(S_{\mathbf{m}}\right) + z_{\mathbf{m}} \right],$$

then the η'*-penalized likelihood estimate* $\widehat{s}_{\widehat{\mathbf{m}}}$*, defined in* (3.3) *and* (3.2)*, satisfies*

$$\mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}}\left[\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}\left(s_0, \widehat{s}_{\widehat{\mathbf{m}}}\right)\right] \leq C_1 \inf_{\mathbf{m}\in\mathcal{M}} \left(\inf_{s_{\mathbf{m}}\in S_{\mathbf{m}}} \mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, s_{\mathbf{m}}\right) + \frac{pen(\mathbf{m})}{n} \right) + \frac{\kappa C_1 \Xi + \eta + \eta'}{n}. \tag{3.4}$$

It is worth noting that Theorem 3.1 extends a corollary of (Montuelle et al., 2014, Theorem 1), which can be verified via Lemma 1 from Nguyen et al. (2021a), which makes explicit the relationship between softmax and Gaussian gating classes. In particular, for modeling a sample of high-dimensional regression data issued from a heterogeneous population with hidden graph-structured interaction between covariates, we refer readers to the works of Nguyen et al. (2021c) while Nguyen et al. (2020a) provides a stronger oracle inequality but slower convergence rate of the error upper bound. More precisely, in (3.4), we obtain a weak oracle inequality due to the different divergences on the left, $\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}$, and on the right $\mathrm{KL}^{\otimes \mathrm{n}}$, but with a faster convergence rate $\mathcal{O}(n^{-1})$ of the error upper bound compared to $\mathcal{O}(n^{-1/2})$ in Nguyen et al. (2020a, Theorem 3.2). A more detailed comparison can be found in Nguyen (2021, Section 1.2.12.3).

Numerical experiments are available in Nguyen et al. (2021d) and https://github.com/Trung-TinNGUYEN/NamsGLoME-Simulation, will be presented in the communication in order to investigate how well the empirical tensorized Kullback–Leibler divergence between the true model and the best data-driven model follows the finite-sample oracle inequality of Theorem 3.1, as well as the convergence rate of the error upper bound $\left(\kappa C_1 \Xi + \eta + \eta'\right)/n$.

References

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern recognition*, 28(5):781–793. 3

Chamroukhi, F., Lecocq, F., and Nguyen, H. D. (2019). Regularized Estimation and Feature Selection in Mixtures of Gaussian-Gated Experts Models. In Nguyen, H., editor, *Statistics and Data Science*, pages 42–56, Singapore. Springer Singapore. 2

Cohen, S. and Le Pennec, E. (2011). Conditional density estimation by penalized likelihood model selection and applications. *Technical report, INRIA*. 4

Deleforge, A., Forbes, F., and Horaud, R. (2015). High-dimensional regression with gaussian mixtures and partially-latent response variables. *Statistics and Computing*, 25(5):893–911. 2

Ho, N., Yang, C.-Y., and Jordan, M. I. (2019). Convergence rates for gaussian mixtures of experts. *arXiv preprint arXiv:1907.04377*. 2

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87. 1

Montuelle, L., Le Pennec, E., et al. (2014). Mixture of gaussian regressions model with logistic weights, a penalized maximum likelihood approach. *Electronic Journal of Statistics*, 8(1):1661–1695. 5

Nguyen, H. D. and Chamroukhi, F. (2018). Practical and theoretical aspects of mixture-of-experts modeling: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1246. 2

Nguyen, H. D., Chamroukhi, F., and Forbes, F. (2019). Approximation results regarding the multiple-output Gaussian gated mixture of linear experts model. *Neurocomputing*, 366:208–214. 2

Nguyen, H. D., Lloyd-Jones, L. R., and McLachlan, G. J. (2016). A universal approximation theorem for mixture-of-experts models. *Neural computation*, 28(12):2585–2593. 2

Nguyen, H. D., Nguyen, T., Chamroukhi, F., and McLachlan, G. J. (2021a). Approximations of conditional probability density functions in Lebesgue spaces via mixture of experts models. *Journal of Statistical Distributions and Applications*, 8(1):13. 2, 5

Nguyen, T. (2021). *Model Selection and Approximation in High-dimensional Mixtures of Experts Models: from Theory to Practice*. PhD Thesis, Normandie Université. 2, 5

Nguyen, T., Chamroukhi, F., Nguyen, H., and McLachlan, G. (2021b). Approximation of probability density functions via location-scale finite mixtures in Lebesgue spaces. *Communications in Statistics - Theory and Methods*. 1

Nguyen, T., Chamroukhi, F., Nguyen, H. D., and Forbes, F. (2021c). Non-asymptotic model selection in block-diagonal mixture of polynomial experts models. *arXiv preprint arXiv:2104.08959*. 5

Nguyen, T., Nguyen, H. D., Chamroukhi, F., and Forbes, F. (2021d). A non-asymptotic penalization criterion for model selection in mixture of experts models. *arXiv preprint arXiv:2104.02640*. 4, 5

Nguyen, T., Nguyen, H. D., Chamroukhi, F., and McLachlan, G. J. (2020a). An l_1-oracle inequality for the Lasso in mixture-of-experts regression models. *arXiv preprint arXiv:2009.10622*. 5

Nguyen, T., Nguyen, H. D., Chamroukhi, F., and McLachlan, G. J. (2020b). Approximation by finite mixtures of continuous density functions that vanish at infinity. *Cogent Mathematics & Statistics*, 7(1):1750861. 1

Xu, L., Jordan, M., and Hinton, G. E. (1995). An Alternative Model for Mixtures of Experts. In Tesauro, G., Touretzky, D., and Leen, T., editors, *Advances in Neural Information Processing Systems*, volume 7. MIT Press. 2