Regularized Mixture of Experts for High-Dimensional Data

Faicel Chamroukhi & Bao-Tuyen Huynh

Normandie Univ, UNICAEN, LMNO UMR CNRS 6139, 14000 Caen, France
Email: {chamroukhi,huynh}@unicaen.fr

Résumé. Cet article propose un modèle de mélange d'experts pour la classification automatique de données de régression hétérogènes comportant un nombre de prédicteurs potentiellement grand. L'estimation parcimonieuse des paramètres s'appuie sur une régularisation de l'estimateur du maximum de vraisemblance pour les experts et les fonctions d'activations, mise en œuvre par un algorithme EM dédié. La méthode de régularisation proposée, contrairement aux méthodes de régularisation de l'état de l'art sur les mélanges d'experts, ne se base pas sur une solution approchée et ne nécessite pas de seuillage pour retrouver la solution parcimonieuse. L'étape M de l'algorithme, effectuée par montée de coordonnées, rendant ainsi prometteur le passage de l'algorithme l'échelle. Une étude expérimentale met en évidence de bonnes performances de l'approche proposée.

Mots-clés. Mélange d'experts, Sélection de variables, Régularisation, algorithme EM.

Abstract. We consider the Mixture of Experts (MoE) modeling for clustering heterogeneous regression data with possibly high-dimensional features and propose a regularized maximum-likelihood estimation based on a dedicated EM algorithm which integrates coordinate ascent updates of the parameters. Unlike state-of-the art regularized MLE for MoE, the proposed modeling does not require an approximate of the regularization. The proposed algorithm allows to automatically obtaining sparse solutions without thresholding, and includes coordinate ascent updates avoiding matrix inversion, and can thus be scalable. An experimental study shows the good performance of the algorithm in terms of recovering sparse solutions, density estimation, and clustering.

Keywords. Mixture of experts, Feature selection, Regularization, EM algorithm.

1 Introduction

Mixture of Experts (MoE) introduced by [8] are successful models for modeling heterogeneous data in many statistical learning problems including regression, clustering and classification. A MoE is a fully conditional mixture model where both the mixing proportions (gating network) and the components densities (experts network), depend on some input covariates. This makes MoE more adapted for input-dependent data than standard unconditional mixture distributions. A general review of the MoE models and their applications can be found in [14]. While the MoE modeling with maximum likelihood inference is widely used, its application in high-dimensional problems is still challenging

due to the known problem of the ML estimation (MLE) in such a setting, and hence there is a need to select a subset of the potentially large number of features, that really explain the problem. Indeed, in high-dimensional setting, the features can be correlated and thus the actual features that explain the problem reside in a low-dimensional space. This can be achieved by regularizing the objective function so that to encourage sparse solutions.

In related models, including mixture of linear regressions (MLR), where the mixing proportions are constant, [10] proposed regularized ML inference, including MIXLASSO, MIXHARD and MIXSCAD and provided some asymptotic properties corresponding to these penalty functions. Another L_1 penalization for MLR models for high-dimensional data was proposed by [15] and an adaptive Lasso penalized estimator with oracle inequality which includes the setting $p \gg n$ was presented. [12] provided an L_1-oracle inequality by a Lasso estimator for finite mixture of Gaussian regression models. This result can be seen as a complementary result to [15], by studying the Lasso for its L_1-regularization properties rather than considering it as a variable selection procedure. This work was extended later in [4] by considering a mixture of multivariate Gaussian regression models. When the set of features can be seen as to be splitted into groups, [6] introduced the two types of penalty functions called MIXGL1 and MIXGL2 for MLR models, based on group Lasso. An MM algorithm version for MLR with Lasso penalty can be found in [11]. Their method allows to avoid matrix operations. In [9], the author extended his MLR regularization to the MoE setting and provided a root-n consistent and oracle properties for Lasso and SCAD penalties and developed an EM algorithm [3] for fitting the models. However, as we will discuss it in section 3.1, this is based on approximated penalty function, and uses a Newton-Raphson in the updates, which requires matrix inversion.

Here we consider MoE models with regularization as in [9] and propose a regularized maximum-likelihood inference which doesn't require an approximate of the regularization. We develop a hybrid EM and coordinate ascent algorithm for model fitting. The proposed algorithm allows to automatically select sparse solutions without thresholding, and includes coordinate ascent updates avoiding matrix inversion.

2 Modeling with Mixture of Experts (MoE)

Let $((\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n))$ be a random sample of n i.i.d pairs $(\boldsymbol{X}_i, Y_i)$ where $Y_i \in \mathcal{X} \subset \mathbb{R}$ is the ith response given some vector of predictors $\boldsymbol{X}_i \in \mathcal{X} \subset \mathbb{R}^p$. We consider the MoE modeling for the analysis of a heteregeneous set of such data. Let $\mathcal{D} = ((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n))$ be an observed data sample. The MoE model assumes that the pairs $(\boldsymbol{x}, y)$ are generated from K tailored probability density components (experts) governed by a hidden categorical random variable $Z \in \{1, \ldots, K\}$ that indicates the component from which a particular pair is drawn. The latter represents the gating network. Formally, the MoE an be defined by the following probability density (or mass) function:

$$f(y_i|\boldsymbol{x}_i; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k(\boldsymbol{x}_i; \boldsymbol{w}) f(y_i|\boldsymbol{x}_i; \boldsymbol{\theta}_k) \tag{1}$$

where the gating network defined by the distribution of the hidden variable Z given the predictor $\boldsymbol{x}$, $\pi_k(\boldsymbol{x}_i; \boldsymbol{w}) = \mathbb{P}(Z_i = k|\boldsymbol{X}_i = \boldsymbol{x}_i; \boldsymbol{w}) = \frac{\exp(w_{k0} + \boldsymbol{x}_i^T \boldsymbol{w}_k)}{1 + \sum_{l=1}^{K-1} \exp(w_{l0} + \boldsymbol{x}_i^T \boldsymbol{w}_l)}$ is in general a softmax

network and the parameter vector is $\boldsymbol{\theta} = (\boldsymbol{w}_1^T, \ldots, \boldsymbol{w}_{K-1}^T, \boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_K^T)^T \in \mathbb{R}^{\nu_{\boldsymbol{\theta}}}$ where $\boldsymbol{\theta}_k$ $(k = 1, \ldots, K)$ is the parameter vector of the kth expert. The experts are chosen to sufficiently represent the data for each group k. For example, MoE for non-symmetric data and robust MoE [2, 1, 13] have been introduced. For a review on MoE, types of gating networks and experts networks, the reader is refereed to [14]. For the case of univariate continuous outputs Y_i, a common choice to model the relationship between the input $\boldsymbol{x}$ and the output Y is by considering regression functions, typically Gaussian.

3 Regularized MoE modeling (RMoE)

The parameter vector $\boldsymbol{\theta}$ of the MoE (1) is commonly estimated by maximizing the log-likelihood $\log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\Big[\sum_{k=1}^{K} \pi_k(\boldsymbol{x}_i; \boldsymbol{w})\mathcal{N}(y_i; \beta_{k0} + \boldsymbol{\beta}_k^T \boldsymbol{x}_i, \sigma_k^2)\Big]$ by using the EM algorithm [3, 8] which allows to iteratively find an appropriate local maximizer of the log-likelihood. However, it is well-known that the MLE may be unstable of even infeasible in high-dimension namely due to possibly redundant and correlated features. In such a context, a regularization of the MLE is needed. Regularized maximum likelihood estimation allows the selection of a relevant subset of features for prediction and thus encourages sparse solutions. In mixture of experts modeling, one may consider both sparsity in the feature space of the gates, and of the experts. We propose to infer the MoE model by maximizing a regularized log-likelihood criterion, which encourages sparsity for both the gating network parameters and the expert parameters and does not require any approximation, along with performing the maximization by coordinate ascent, so that to avoid matrix inversion.

3.1 Regularized maximum-likelihood estimation of the MoE

The proposed regularization combines a Lasso penalty for the experts parameters, and an Elastic-Net like penalty for the gating network, defined by:

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

A similar strategy were proposed in [9] where the author proposed a regularized ML function like (2) but which is then approximated in the model inference algorithm. The devoloped EM algorithm for fitting the model follows indeed the suggestion of [5] to approximate the penalty function in a some neighbourhood by a local quadratic function. Therefore, the Newton-Raphson method could be used to update parameters in the M-step. The weakness of this design is that once a feature is set to zero, it may never reenter the model at a later stage of the algorithm. To avoid this numerical instability of the algorithm due to the small values of some of the features in the denominator of this approximation, [9] replaced that approximation by an ϵ-local quadratic function. Unfortunately, these strategies have some drawbacks. First, by approximating the penalty functions with (ϵ-)quadratic functions, almost surely none of the components will be exactly zero. Hence, a threshold should be considered to declare a coefficient is zero

and this threshold affects the degree of sparsity. Secondly, it cannot guarantee the non-decreasing property of the EM algorithm of the penalized objective function. Thus, the convergence of the EM algorithm cannot be ensured. One has also to choose ϵ as an additional tuning parameter in practice. Our proposal overcomes these limitations.

3.2 Parameter estimation with a block-wise EM algorithm

We propose a block-wise EM algorithm, which integrates a coordinate ascent algorithm for updating the model parameters, to monotonically find local maximizers of (2). After starting with an initial solution $\boldsymbol{\theta}^{(0)}$, the algorithm alternates between the two following steps until convergence (i.e., when there is no longer a significant change in the relative variation of the regularized log-likelihood (2)).

E-step: Compute the conditional expectation $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)})$ of the penalized complete-data log-likelihood, given the observed data $\mathcal{D}$ and a current parameter vector $\boldsymbol{\theta}^{(q)}$. This only requires the computation of the posterior component memberships $\tau_{ik}^{(q)}$ for each i and k:

$$\tau_{ik}^{(q)} = \mathbb{P}(Z_i = k | y_i, \boldsymbol{x}_i; \boldsymbol{\theta}^{(q)}) = \frac{\pi_k(\boldsymbol{x}_i; \boldsymbol{w}^{(q)}) \mathcal{N}(y_i; \beta_{k0}^{(q)} + \boldsymbol{\beta}_k^T \boldsymbol{x}_i^{(q)}, \sigma_k^{(q)2})}{\sum_{l=1}^{K} \pi_l(\boldsymbol{x}_i; \boldsymbol{w}^{(q)}) \mathcal{N}(y_i; \beta_{l0}^{(q)} + \boldsymbol{\beta}_l^T \boldsymbol{x}_i^{(q)}, \sigma_l^{(q)2})}. \tag{3}$$

M-step: Update the parameters by maximizing the Q-function:

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(q)}) = \underbrace{\sum_{i,k} \tau_{ik}^{(q)} \log \pi_k(\boldsymbol{x}_i; \boldsymbol{w}) - \sum_{k=1}^{K-1} (\gamma_k \|\boldsymbol{w}_k\|_1 - \frac{\rho}{2} \|\boldsymbol{w}_k\|_2^2)}_{Q(\boldsymbol{w}; \boldsymbol{\theta}^{(q)})} + \underbrace{\sum_{i,k} \tau_{ik}^{(q)} \log f(y_i | \boldsymbol{x}_i; \boldsymbol{\beta}, \sigma_k^2) - \sum_{k=1}^{K} \lambda_k \|\boldsymbol{\beta}_k\|_1}_{Q(\boldsymbol{\beta}, \sigma; \boldsymbol{\theta}^{(q)})}.$$

We propose a coordinate ascent algorithm to update $\boldsymbol{w}$ and the $\boldsymbol{\beta}$' parameters. Indeed, based on [16] with regularity conditions, then the coordinate ascent algorithm is successful in updating $\boldsymbol{w}$. The function $Q(\boldsymbol{w}, \boldsymbol{\theta}^{(q)})$ is decomposed into separate problems of weighted and smoothly regularized multinational logistic regression problems. Thus, one can use one-dimensional generalized Newton-Raphson algorithm with initial value $w_{kj}^{(0)} = w_{kj}^{(q)}$ to find the maximizers of these functions. The m^{th} iteration of this coordinate ascent algorithm is given by: $w_{kj}^{(m+1)} = w_{kj}^{(m)} - \frac{\partial Q(w_{kj}; \boldsymbol{\theta}^{(q)})}{\partial w_{kj}}\Big|_{w_{kj}^{(m)}} \left(\frac{\partial^2 Q(w_{kj}; \boldsymbol{\theta}^{(q)})}{\partial^2 w_{kj}} \right)^{-1} \Big|_{w_{kj}^{(m)}}$, where the gradient and the hessian are analytic. Next, we alternate between the update of β_{kj} and σ_k, in $Q(\boldsymbol{\beta}, \sigma; \boldsymbol{\theta}^{(q)})$. This consists in solving K weighted lasso problems and we obtain the following closed-form coordinate ascent updates:

$$\beta_{kj}^{(m+1)} = \mathcal{S}_{\lambda_k \sigma_k^{(q)2}} \sum_{i=1}^{n} \tau_{ik}^{(q)} r_{ikj}^{m} x_{ij} \Big/ \sum_{i=1}^{n} \tau_{ik}^{(q)} x_{ij}^2, \text{ and } \beta_{k0}^{(m+1)} = \sum_{i=1}^{n} \tau_{ik}^{(q)} (y_i - \boldsymbol{\beta}_k^T \boldsymbol{x}_i^{(m+1)}) \Big/ \sum_{i=1}^{n} \tau_{ik}^{(q)} \tag{4}$$

with $r_{ikj}^{(m)} = y_i - \beta_{k0}^{(m)} - \boldsymbol{\beta}_k^T \boldsymbol{x}_i^{(m)} + \beta_{kj}^{(m)} x_{ij}$ and $\mathcal{S}_{\lambda_k \sigma_k^{(q)2}}(.)$ is a soft-thresholding operator defined by $[\mathcal{S}_\gamma(u)]_j = \text{sign}(u_j)(|u_j| - \gamma)_+$ and $(x)_+$ a shorthand for $\max\{x, 0\}$, with initial values $(\beta_{k0}^{(0)}, \boldsymbol{\beta}_k^{(0)}) = (\beta_{k0}^{(q)}, \boldsymbol{\beta}_k^{(q)})$. Then, rerun the E-step, and update σ_k^2 by

$$\sigma_k^{2(q+2)} = \sum_{i=1}^{n} \tau_{ik}^{(q+1)} (y_i - \beta_{k0}^{(q+2)} - \boldsymbol{\beta}_k^{(q+2)^T} \boldsymbol{x}_i)^2 \Big/ \sum_{i=1}^{n} \tau_{ik}^{(q+1)}. \tag{5}$$

This algorithm, at each iteration, improves the optimised penalised log-likelihood function (2). Also we can directly get zero coefficients without thresholding unlike in [9, 7].

4 Experimental results

We consider predictors $\{\boldsymbol{x}_i\}$ generated from a zero-mean multivariate Gaussian distribution with correlation structure $\text{corr}(x_{ij}, x_{ij'}) = 0.5^{|j-j'|}$. The response Y is generated from a normal MoE regressors model with $K = 2$, $n = 300$ and parameters: $(\beta_{10}, \boldsymbol{\beta}_1)^T = (0, 0, 1.5, 0, 0, 0, 1)^T$; $(\beta_{20}, \boldsymbol{\beta}_2)^T = (0, 1, -1.5, 0, 0, 2, 0)^T$; $(w_{10}, \boldsymbol{w}_1)^T = (1, 2, 0, 0, -1, 0, 0)^T$. The results are averaged over 100 data sets. We consider nonpenalized MoE, MoE with ridge penalty for the gates (MoE-L_2), MoE with BIC penalty for feature selection (MoE-BIC) and MIXLASSO (see [10]). For BIC selection, we consider a pool of $5 \times 4 \times 5 = 100$ candidates to choose the best submodel. Table 1 presents the *sensitivity/specificity* results (i.e, proportion of correctly estimated zero coefficients and nonzero coefficents) and the *clustering* (including correct classification rate and the Adjusted rand index ARI). We can

Method	Sensitivity/Specificity						Correct classification rate	ARI
	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	$89.57\%_{(1.65\%)}$	$0.623_{(.053)}$
MoE-L_2	0.000	1.000	0.000	1.000	0.000	1.000	$89.62\%_{(1.63\%)}$	$0.624_{(.052)}$
MoE-BIC	0.920	1.000	0.930	1.000	0.850	1.000	$90.05\%_{(1.65\%)}$	$0.638_{(.053)}$
MIXLASSO	0.775	1.000	0.6933	1.000	N/A	N/A	$82.89\%_{(1.92\%)}$	$0.4218_{(.050)}$
MoE-Lasso+L_2	0.700	1.000	0.803	1.000	0.853	0.945	$89.46\%_{(1.76\%)}$	$0.619_{(.056)}$

Table 1: Sensitivity (S_1)/Specificity (S_2) and clustering error summaries.

see that the proposed algorithm performs very well to retrieve the actual sparse support. Actually, the L_2 and MoE models cannot be considered as model selection methods since their sensitivity criterion almost surely equal zero. The Lasso+L_2 performs quite well in terms of experts 1 and 2 while the feature selection becomes more difficult for the gate parameters $\boldsymbol{w}$ since there are correlations between features. The BIC provides very good results in general. However, in practice to obtain the best submodel we must consider a lot of cases and this restricts the application capability of BIC. The MIXLASSO, in some sense can select the actual non-zero features for the experts but this model doesn't perform well in clustering. We apply and evaluate the algorithm on the real data set of baseball salaries (in the same setting as in [10] (32 features)). The results in Table 2 provide the

Method	R^2	MSE	Exp.1	Exp.2	Gate
MoE	0.8099	$0.2625_{(.758)}$	0	0	0
MIXLASSO	0.4252	$1.1858_{(2.792)}$	19	21	N/A
MoE-Lasso+L_2	0.8020	$0.2821_{(.633)}$	17	20	29

Table 2: MSE between observations and prediction and R^2 for Baseball salaries data.

number of zero coefficients in the experts and the gates of the estimated parameters. We can see that the proposed algorithm shrinks some parameters to zero and has acceptable results compared to MoE. It also has better results compared to MIXLASSO.

5 Conclusion and future work

We proposed a regularized MLE for the MoE model which encourages sparsity, and developed a blockwise EM algorithm to monotonically maximize this regularized objective towards at least a local maximum. The proposed regularization does not require using using approximations and the proposed algorithm is based on univariate updates of the model parameters via coordinate ascent, which allows to tackle problems in high-dimensional computation and to promote its scalability. The experimental results confirm that the algorithm performs well in feature selection and clustering of heteteregeneous regression data. A future work would consist in hierarchical MoE and MoE for discrete data.

References

[1] F. Chamroukhi. Robust mixture of experts modeling using the t distribution. *Neural Networks*, 79:20–36, 2016.

[2] F. Chamroukhi. Skew t mixture of experts. *Neurocomputing*, 266:390 – 408, 2017.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. of the royal statistical society. Series B*, pages 1–38, 1977.

[4] E. Devijver. An ℓ_1-oracle inequality for the lasso in multivariate finite mixture of multivariate gaussian regression models. *ESAIM: Probability and Statistics*, 19:649–670, 2015.

[5] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

[6] F. K. Hui, D. I. Warton, S. D. Foster, et al. Multi-species distribution modeling using penalized mixture of regressions. *The Annals of Applied Statistics*, 9(2):866–882, 2015.

[7] D. R. Hunter and R. Li. Variable selection using mm algorithms. *Annals of statistics*, 33(4):1617, 2005.

[8] 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.

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

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

[11] L. R. Lloyd-Jones, H. D. Nguyen, and G. J. McLachlan. A globally convergent algorithm for lasso-penalized mixture of linear regression models. *arXiv:1603.08326*, 2016.

[12] C. Meynet. An ℓ_1-oracle inequality for the lasso in finite mixture gaussian regression models. *ESAIM: Probability and Statistics*, 17:650–671, 2013.

[13] H. D. Nguyen and G. J. McLachlan. Laplace mixture of linear experts. *Computational Statistics & Data Analysis*, 93:177–191, 2016.

[14] Hien D. Nguyen and Faicel Chamroukhi. An introduction to the practical and theoretical aspects of mixture-of-experts modeling. *ArXiv preprint arXiv:1707.03538v1*, Jul 2017.

[15] N. Städler, P. Bühlmann, and S. Van De Geer. $l1$-penalization for mixture regression models. *Test*, 19(2):209–256, 2010.

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