Journal of Statistical Planning and Inference 237 (2025) 106250

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Estimation and group-feature selection in sparse mixture-of-experts with diverging number of parameters

Abbas Khalili [a,*], Archer Yi Yang [a,b], Xiaonan Da [c]

[a] Department of Mathematics and Statistics, McGill University, Montreal, Canada
[b] Mila - Quebec AI Institute, Montreal, Canada
[c] Statistics Canada, Ottawa, Canada

ARTICLE INFO

Keywords:
Mixture-of-experts
Regularization
Variable selection

ABSTRACT

Mixture-of-experts provide flexible statistical models for a wide range of regression (supervised learning) problems. Often a large number of covariates (features) are available in many modern applications yet only a small subset of them is useful in explaining a response variable of interest. This calls for a feature selection device. In this paper, we present new group-feature selection and estimation methods for sparse mixture-of-experts models when the number of features can be nearly comparable to the sample size. We prove the consistency of the methods in both parameter estimation and feature selection. We implement the methods using a modified EM algorithm combined with proximal gradient method which results in a convenient closed-form parameter update in the M-step of the algorithm. We examine the finite-sample performance of the methods through simulations, and demonstrate their applications in a real data example on exploring relationships in body measurements.

1. Introduction

High-dimensional data arises in many research fields such as biology, medicine, engineering, social science and econometrics (Rish and Grabarnik, 2014; Wainwright, 2019). At the beginning of a study, data often consists of observations on a large number of features, yet only a small subset of which is important in explaining the behavior of a response variable. Sparse regularization can help select important features to form a more parsimonious model while alleviating overfitting brought by high-dimensionality, thus improves interpretability and prediction accuracy of the resulting model (Simon et al., 2013). The seminal works of Tibshirani (1996) on the least absolute shrinkage operator (Lasso), Fan and Li (2001) on the smoothly clipped absolute deviation (SCAD), Zou (2006) on the adaptive Lasso (AdaLasso), and Yuan and Lin (2006) on the group Lasso have led to astonishing amounts of research developments over the last two decades for estimation and feature selection in various high-dimensional supervised/unsupervised learning problems; see the two books (Hastie et al., 2019) and Fan et al. (2020) for a comprehensive review of the topic.

Estimation and feature selection become even more complex when the relationship between a response variable and potential features varies across multiple sub-populations – due to the existence of an unobservable heterogeneity in a population or data generation process. Mixture-of-experts (MOE) models, originally introduced by Jacobs et al. (1991), are composed of several functions which are referred to as experts and a gating network which assigns observations to an expert with a certain probability. The MOE models can be viewed as a decision tree with its branches as experts and the decision process governed by the gating network of e.g. multinomial logit probabilistic models. As a generalization of finite mixture of regression (FMR) models (McLachlan and Peel, 2000), MOEs provide a rich class of statistical models to deal with unobserved heterogeneity in the data. These models were originally

* Corresponding author.
E-mail addresses: abbas.khalili@mcgill.ca (A. Khalili), archer.yang@mcgill.ca (A.Y. Yang), xiaonan.da@statcan.gc.ca (X. Da).

https://doi.org/10.1016/j.jspi.2024.106250
Received 29 August 2022; Received in revised form 30 July 2024; Accepted 9 November 2024
Available online 19 November 2024

proposed in problem decomposition context, where a complex problem is divided into a set of simpler subproblems based on a divide-and-conquer principle, and then one or more specialized problem-solving experts are assigned to each of the subproblems (Yuksel et al., 2012). This supervised learning technique have been widely applied in many regression and classification problems due to its flexibility in capturing complex relationship between variables of interest; see Nguyen and Chamroukhi (2018) and the references therein. However, despite their popularity in applications, very limited studies are conducted on estimation and feature selection in high-dimensional MOEs. This is the focus of our paper.

To the best of our knowledge, there are currently only a few statistical papers that study estimation and feature selection problems in FMR models as a special case of MOEs. Städler et al. (2010) elegantly studied feature selection in Gaussian FMR models using Lasso when the dimension of the parameter space exceeds the number of observations. Guo et al. (2010) introduced a pairwise variable selection method for high-dimensional Gaussian mixture model, with simplification that the expectations of the mixture components are not modeled as functions of covariates. Khalili and Lin (2013) proposed a general theory for feature selection in FMR models when the number of features can grow similar to $n^{\frac{1}{4}}$. Khalili and Chen (2007), Khalili (2010), and Chamroukhi and Huynh (2019) studied feature selection problems in FMR and MOEs under the standard setting of fixed-p-large-n. Nguyen et al. (2020) studied statistical error of the high-dimensional Lasso estimators in MOEs .

In this paper, we study estimation and feature selection in MOEs with potentially a large number of covariates using a grouped regularization technique. Motivated by the regularization techniques in regression, we propose a computationally efficient estimation and feature selection method for a general class of MOEs. In an MOE model with more than $K = 2$ mixture components and when the number of features is large, the standard regularization of individual regression parameters, such as a penalty on the individual experts' parameters, may result in different subsets of selected features for different mixing probabilities in the gating network, rendering the ending model difficult to interpret. To overcome this issue, we apply a group regularization on the gating network parameters. As a result, the effects of a feature on different mixing probabilities $\{g_1, \dots, g_k\}$, $\sum_{k=1}^{K} g_k = 1$, will share the same sparsity. Thus, the grouping is on all the regression parameters corresponding to each feature in the gating network, which results in a more interpretable sparse MOE model. We study conditions under which the proposed methods are consistent in estimation and feature selection. We examine the finite-sample performance of the methods via simulations, and demonstrate their applications by analyzing a data on exploring relationships in body measurements.

The rest of the paper is organized as follows. In Section 2, we introduce MOE models and their sparsity structure. In Section 3, we outline our new estimation and feature selection method. We study theoretical properties of the proposed methods in Section 4. Numerical algorithm and implementation details of the methods are given in Sections 5 and 6, respectively. Our simulation study and a real data example are given in Sections 7 and 8, respectively. Some discussion and closing remarks are given in Section 9. Regularity conditions, tables, and some figures are given in Appendix. The proofs are given in our Supplementary Material.

2. Mixture-of-experts (MOE) and their sparsity

Let $Y \in \mathcal{Y} \subset \mathbb{R}$ be a response variable of interest and $\boldsymbol{x} = (x_1, x_2, \dots, x_p)^\top \in \mathcal{X} \subset \mathbb{R}^p$ be a p-dimensional vector of features which may be related to Y. Further, let $\mathcal{F} = \{h(y; \eta, \phi) : \eta \in \mathbb{R}, \text{ and } \phi \in \mathbb{R}^+\}$ be a known parametric family of probability (mass) density functions with respect to a σ-finite measure ν. In an MOE model with K components, the conditional density (mass) function of Y given $\boldsymbol{x}$ is

$$f(y; \boldsymbol{x}, \boldsymbol{\theta}) = \sum_{k=1}^{K} g_k(\boldsymbol{x}; \boldsymbol{\alpha})\ h(y; \eta_k(\boldsymbol{x}), \phi_k) \tag{1}$$

with $\eta_k(\boldsymbol{x}) = \beta_{0k} + \boldsymbol{\beta}_k^\top \boldsymbol{x}$, and $\boldsymbol{\beta}_k = (\beta_{k1}, \beta_{k2}, \dots, \beta_{kp})^\top$, for $k = 1, 2, \dots, K$. Here, h is called expert and the mixing probabilities g_k are referred to as the gating network (Jacobs et al., 1991). The g_k is commonly modeled using a conditional multinomial regression function

$$\log\left(\frac{g_k(\boldsymbol{x}; \boldsymbol{\alpha})}{g_K(\boldsymbol{x}; \boldsymbol{\alpha})}\right) = \alpha_{0k} + \boldsymbol{\alpha}_k^\top \boldsymbol{x} \quad \text{for } k = 1, \dots, K-1, \tag{2}$$

and $g_K(\boldsymbol{x}; \boldsymbol{\alpha}) = 1 - \sum_{k=1}^{K-1} g_k(\boldsymbol{x}; \boldsymbol{\alpha})$, where $\boldsymbol{\alpha} = (\alpha_{01}, \boldsymbol{\alpha}_1, \alpha_{02}, \boldsymbol{\alpha}_2, \dots, \alpha_{0,K-1}, \boldsymbol{\alpha}_{K-1})^\top$, with $\boldsymbol{\alpha}_k = (\alpha_{k1}, \alpha_{k2}, \dots, \alpha_{kp})^\top$. The vector of all parameters is denoted by

$$\boldsymbol{\theta} = (\beta_{01}, \boldsymbol{\beta}_1, \beta_{02}, \boldsymbol{\beta}_2, \dots, \beta_{0K}, \boldsymbol{\beta}_K, \alpha_{01}, \boldsymbol{\alpha}_1, \alpha_{02}, \boldsymbol{\alpha}_2, \dots, \alpha_{0,K-1}, \boldsymbol{\alpha}_{K-1}, \boldsymbol{\phi}),$$

where $\boldsymbol{\phi} = (\phi_1, \phi_2, \dots, \phi_K)^\top$ is the vector of dispersion parameters. Note that $\dim(\boldsymbol{\theta}) = d = (2K-1)(p+1)+K$, and K is fixed. Denote $\boldsymbol{\Theta} \subseteq \mathbb{R}^d$ as the parameter space.

One may interpret an MOE model as follows: given the input variable $\boldsymbol{x}$, with probability $g_k(\boldsymbol{x}; \boldsymbol{\alpha})$, the random variable Y is generated according to the distribution $h(y; \eta_k(\boldsymbol{x}), \phi_k), k = 1, \dots, K$.

Identifiability is essential for statistical inference in MOE models: if $f(y; \boldsymbol{x}, \boldsymbol{\theta}_1) = f(y; \boldsymbol{x}, \boldsymbol{\theta}_2)$, for all $(y, \boldsymbol{x}) \in \mathcal{Y} \times \mathcal{X}$, then we must have $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$, up to a mixture component permutation. The unique representation of an MOE depends on the density $h(y; \eta, \phi)$, the maximum possible order K, and the design matrix $(\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n)^\top$. Jiang and Tanner (1999b) studied the identifiability of MOE models under a random design matrix, where $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ are a random sample from a marginal density $m(\boldsymbol{x})$ that does not depend on $\boldsymbol{\theta}$. The density $m(\boldsymbol{x})$ must not have all of its mass concentrated in up to K of $(p-1)$-dimensional linear subspaces. We restate their main result as follows.

Proposition 1 (*Jiang and Tanner, 1999b*). *Assume that* $\{h(y;\eta_j,\phi_j); j=1,2,\dots,2K\}$ *are linearly independent functions of* y, *for any* $2K$ *distinct parameters* η_j *and* ϕ_j, *referred to as non-degeneracy condition. If for any two parameter vectors* $\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\in\boldsymbol{\Theta}$, $f(y;\boldsymbol{x},\boldsymbol{\theta}_1)=f(y;\boldsymbol{x},\boldsymbol{\theta}_2)$, *for all* $(y,\boldsymbol{x})\in\mathcal{Y}\times\mathcal{X}$, *then* $\boldsymbol{\theta}_1=\boldsymbol{\theta}_2$, *up to permutation of the entries of the two parameter vectors.*

The non-degeneracy condition is applicable to MOE models based on Gaussian, Poisson, and Binomial with number of trials $m>2K-1$. Hennig (2000) showed that for fixed designs, in addition to the non-degeneracy condition on experts $h(y;\eta,\phi)$, a sufficient condition for identifiability is that the design points $\boldsymbol{x}_i$ do not fall in the union of any K linear subspaces of $(p-1)$-dimension.

Sparse MOE models: In many applications of MOEs, there are often a large number of features present in the data. To avoid over-parameterization when fitting an MOE model to such data, it is necessary to assume certain structure for the model. A common practice is to assume sparsity, under which many of the elements of the vectors $\boldsymbol{\alpha}_k$ and $\boldsymbol{\beta}_k$ are zero, resulting in a parsimonious and more interpretable MOE model. Specifically, let $\mathcal{S}=\{1,2,\dots,p\}$ be the index set representing the full feature vector $\boldsymbol{x}$. For any index subsets $A_1,A_2,\dots,A_K\subset\mathcal{S}$, with cardinality $|A_k|$ for $k=1,2,\dots,K$, denote $\boldsymbol{\beta}_k[A_k]$ and $\boldsymbol{x}[A_k]$ as subvectors of $\boldsymbol{\beta}_k$ and $\boldsymbol{x}$, respectively, such that $\beta_{kj}\neq 0,\forall j\in A_k$, for $k=1,2,\dots,K$. In regards to the sparsity of the gating network, for each $j=1,2,\dots,p$, let $\boldsymbol{\alpha}_{\cdot j}=(\alpha_{1j},\alpha_{2j},\dots,\alpha_{K-1,j})^\top$, which represents the grouping effect of a feature x_j on the whole gating network. For any $A^*\subset\mathcal{S}$, with cardinality $|A^*|$, denote $\boldsymbol{\alpha}[A^*]$ and $\boldsymbol{x}[A^*]$ as subvectors of $\boldsymbol{\alpha}$ and $\boldsymbol{x}$, respectively, such that for any $j\in A^*$, we have $\boldsymbol{\alpha}_{\cdot j}\neq 0$. Equivalently, we assume that for any $j\notin A^*$, we have $\alpha_{kj}=0$, for all $k=1,2,\dots,K-1$. This formulation leads to the grouping effect among the regression coefficients α_{kj} in the gating network $g_1,\dots,g_k$. In Section 4, these subsets are referred to as active sets. For $A_1,A_2,\dots,A_K,A^*\subset\mathcal{S}$, we denote a *sparse* MOE or *submodel* as

$$f_{[A_1,\dots,A_K;A^*]}(y;\boldsymbol{x},\boldsymbol{\theta})=\sum_{k=1}^{K} g_k(\boldsymbol{x}[A^*];\boldsymbol{\alpha}[A^*])\ h(y;\eta_k(\boldsymbol{x}[A_k]),\phi_k). \tag{3}$$

We assume that the true model underlying data is a sparse MOE of the form in (3). The goal is to correctly recover the supports of the nonzero coefficients in the true model and accurately estimate their values, based on the given data. Note that for a given value of K and p, the total number of MOE submodels of the form (3) is $2^{[(K+1)p]}$, which could be very large even for moderate values of K and p. Hence, all-subset selection methods such as AIC, BIC and their variants (Konishi and Kitagawa, 2008) are clearly not practical in this scenario. In this paper, we investigate the use of regularization techniques for sparse learning in MOEs.

3. Simultaneous estimation and feature selection in sparse MOEs

Let $(\boldsymbol{x}_i,y_i), i=1,2,\dots,n$, be an observed random sample from a true sparse MOE defined in (3). The (conditional) log-likelihood of the parameter vector $\boldsymbol{\theta}$ based on the full model (1) is given by

$$l_n(\boldsymbol{\theta})=\sum_{i=1}^{n}\log\left\{\sum_{k=1}^{K} g_k(\boldsymbol{x}_i;\boldsymbol{\alpha})\ h(y_i;\eta_k(\boldsymbol{x}_i),\phi_k)\right\}. \tag{4}$$

The maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$, i.e. the maximizer of $l_n(\boldsymbol{\theta})$, when the dimension of $\boldsymbol{\theta}$ is small relative to the sample size n, is well-studied in the literature (Jiang and Tanner, 1999a). However, the MLE does not have the sparsity property as postulated by (3) when the dimension of $\boldsymbol{\theta}$ is large. Thus, we focus on a penalized maximum likelihood estimator of $\boldsymbol{\theta}$ as outlined below.

To select features for each expert $h(y;\eta_k(\boldsymbol{x}),\phi_k)$, we penalize individual regression coefficients β_{kj}'s by introducing a Lasso-type regularization function (to be described below). This allows potentially different subsets of features to be selected in different experts. For the gating network $\{g_1,g_2,\dots,g_K\}$, instead, we aim to select the same features across the gating network which also enhances interpretability of the resulting model. More specifically, for each $j=1,2,\dots,p$, let $\boldsymbol{\alpha}_{\cdot j}=(\alpha_{1j},\alpha_{2j},\dots,\alpha_{K-1,j})^\top$, which represents the effect of a feature x_j on the whole gating network, we hope that x_j is selected when the corresponding $\boldsymbol{\alpha}_{\cdot j}\neq\mathbf{0}$. According to the structural sparsity assumption of the true model defined in (3), we apply group penalization on the entire vector $\boldsymbol{\alpha}_{\cdot j}$ instead of using coordinate-separable penalization on α_{kj}'s. Denote

$$\|\boldsymbol{\alpha}_{\cdot j}\|_2=\left(\sum_{k=1}^{K-1}\alpha_{kj}^2\right)^{1/2},\quad j=1,\dots,p.$$

We can see that $\|\boldsymbol{\alpha}_{\cdot j}\|_2=0$ if and only if $\alpha_{kj}=0$, for all $k=1,2,\dots,K-1$, thus can preserve (remove) the same features for (from) the whole gating network.

We are now ready to tackle the feature selection problem in MOEs. We estimate $\boldsymbol{\theta}$ by maximizing the penalized log-likelihood function

$$L_n(\boldsymbol{\theta})=l_n(\boldsymbol{\theta})-\boldsymbol{R}_n(\boldsymbol{\theta}), \tag{5}$$

where

$$\boldsymbol{R}_n(\boldsymbol{\theta})=\sum_{k=1}^{K}\sum_{j=1}^{p} r_n(\beta_{kj};\lambda)+\sum_{j=1}^{p}\left\{r_n(\|\boldsymbol{\alpha}_{\cdot j}\|_2;\lambda^*)+\frac{\tau^*}{2}\|\boldsymbol{\alpha}_{\cdot j}\|_2^2\right\} \tag{6}$$

for some regularization function r_n and tuning parameters $(\lambda,\lambda^*,\tau^*)$. The first regularization function allows for separate feature selection for each expert k, while the second penalty enforces groupwise feature selection across the gating network by penalizing the entire parameter vector $\boldsymbol{\alpha}_{\cdot j}$. The main purpose of using an additional ridge-type (quadratic) penalty is to improve the estimation of the model with highly correlated covariates and thus avoiding unstable estimates of the gating network parameters. In addition,

it helps with numerical stability of the computational algorithm as pointed out by Friedman et al. (2010). Examples of the penalty r_n are the Lasso, AdaLasso, SCAD, and MCP which are given in Appendix A.

The maximum penalized likelihood estimator (MPLE) of $\boldsymbol{\theta}$ is then given by

$$\widehat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L_n(\boldsymbol{\theta}). \tag{7}$$

By appropriate tuning of the parameters (λ, λ^*) together with τ^* in (6), various elements of the vector estimator $\widehat{\boldsymbol{\theta}}_n$ turn into zero. We achieve the goal of feature selection and estimation simultaneously, which is of a great computational advantage when fitting an MOE to data. In Sections 4.1 and 4.2, we discuss the selection of appropriate tuning parameters $(\lambda, \lambda^*, \tau^*)$ to ensure that the ridge penalty does not overshadow the penalty r_n, thereby enabling the method to effectively carry out variable selection. Numerical implementation of (7) is given in Section 5.

4. Large-sample study

We first introduce some notations. For each k, we assume that the parameter vectors in the experts are partitioned as $\boldsymbol{\beta}_k = (\boldsymbol{\beta}_{k,1}, \boldsymbol{\beta}_{k,2})$, such that each $\boldsymbol{\beta}_{k,1}$ contains the non-zero coefficients and $\boldsymbol{\beta}_{k,2} = \mathbf{0}$. Similarly, we assume the partitioning $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$ such that $\boldsymbol{\alpha}_1$ contains all the intercepts $\alpha_{0k}, k = 1, 2, \ldots, K-1$ and non-zero vectors $\boldsymbol{\alpha}_{\cdot j}$, and $\boldsymbol{\alpha}_2$ contains all those $\boldsymbol{\alpha}_{\cdot j} = \mathbf{0}$. Without loss of generality, we thus rearrange the elements of the master vector $\boldsymbol{\theta}$ and write $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ such that $\boldsymbol{\theta}_2$ contains all the zero regression coefficients $\boldsymbol{\beta}_{k,2}$, for $k = 1, 2, \ldots, K$, and $\boldsymbol{\alpha}_2 = \mathbf{0}$. Further, denote $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{01}, \boldsymbol{\theta}_{02})$ as the true parameter-vector of the MOE model such that $\boldsymbol{\theta}_{02} = \mathbf{0}$. We assume $\boldsymbol{\theta}_0$ is an interior point of the parameter space $\boldsymbol{\Theta}$. Also, denote the so-called active sets

$$A_{kn} = \{1 \le j \le p_n; \beta^0_{kj} \ne 0\} \quad , \quad k = 1, 2, \ldots, K \tag{8}$$

corresponding to the true non-zero regression parameters of the experts, and the active set

$$A^*_n = \{1 \le j \le p_n; \boldsymbol{\alpha}^0_{\cdot j} \ne 0\} \tag{9}$$

corresponding to the true non-zero grouping parameters $\boldsymbol{\alpha}^0_{\cdot j}$ of the gating network. Let $s_{kn} = |A_{kn}|, k = 1, \ldots, K-1$, and $s^*_n = |A^*_n|$ be the cardinalities of the above active sets. Further, let $s_n = \max\{\max_{1\le k\le K}\{s_{nk}\}, s^*_n\}$ be the maximum number of the non-zero regression coefficients in the experts and the gating network.

The following quantities help us to state the regularity conditions on the penalty r_n. Denote

$$a_{n1} = \max_{1\le k\le K}\max_{j\in A_{kn}}\{|r'_n(\beta^0_{kj};\lambda_n)|/\sqrt{n}\},\ a_{n2} = \max_{j\in A^*_n}\{|r'_n(\|\boldsymbol{\alpha}^0_{\cdot j}\|_2;\lambda^*_n)|/\sqrt{n}\}, \tag{10}$$

$$b_{n1} = \max_{1\le k\le K}\max_{j\in A_{kn}}\{|r''_n(\beta^0_{kj};\lambda_n)|/n\},\ b_{n2} = \max_{j\in A^*_n}\{|r''_n(\|\boldsymbol{\alpha}^0_{\cdot j}\|_2;\lambda^*_n)|/n\}, \tag{11}$$

$$a_n = \max(a_{n1}, a_{n2}),\ b_n = \max(b_{n1}, b_{n2}), \tag{12}$$

where $r'_n(\cdot;\lambda_n)$ and $r''_n(\cdot;\lambda_n)$ are the first and second derivatives of $r_n(\theta;\lambda_n)$ with respect to θ. In what follows, the large-sample behaviors of λ_n and λ^*_n are the same and thus we use λ_n to represent both when needed. We consider the following conditions on r_n, and the parameters $(\lambda_n, \lambda^*_n, \tau^*_n)$.

- C_0. For all n and λ_n, $r_n(0;\lambda_n) = 0$, and $r_n(\theta;\lambda_n)$ is symmetric and non-negative. It is non-decreasing and twice differentiable for all θ in $(0,\infty)$ with at most a few exceptions. In addition, there exists constants C_1 and C_2 such that when $\theta_1 > C_1\lambda_n$ and $\theta_2 > C_1\lambda_n$, then $\frac{1}{n}|r''_n(\theta_1;\lambda_n) - r''_n(\theta_2;\lambda_n)| \le C_2|\theta_1 - \theta_2|$.
- C_1. As $n\to\infty$, $\frac{\tau^*_n}{\sqrt{n}}\max_{j\in A^*_n}\|\boldsymbol{\alpha}^0_{\cdot j}\|_2 = o(1+a_n)$, $\frac{a_n/\sqrt{n}}{\min_{j\in A^*_n}\|\boldsymbol{\alpha}^0_{\cdot j}\|_2} = o(1)$, and $\tau^*_n = o(n)$. Also, $\min_{j\in A^*_n}\|\boldsymbol{\alpha}^0_{\cdot j}\|_2/\lambda^*_n \to \infty$, $\min_{j\in A_{nk}}|\beta^0_{kj}|/\lambda_n \to \infty$, $k = 1, 2, \ldots, K$.
- C_2. As $n\to\infty$, $b_n = o(1)$.
- C_3. For $T_n = \{\theta; 0 < \theta \le \sqrt{\frac{p_n}{n}}\log n\}$, $\lim_{n\to\infty}\inf_{\theta\in T_n}\frac{r'_n(\theta;\lambda_n)}{\sqrt{np_n}} = +\infty$.
- C^*_3. For $T^*_n = \{\theta; 0 < \theta \le \sqrt{\frac{s_n p_n}{n}}\log n\}$, $\lim_{n\to\infty}\inf_{\theta\in T^*_n}\frac{r'_n(\theta;\lambda_n)}{\sqrt{ns_np_n}} = +\infty$.

Conditions C_0-C^*_3 guide us on the appropriate choice of r_n and the tuning parameters $(\lambda_n, \lambda^*_n, \tau^*_n)$ in order to achieve consistency in both estimation of the non-zero regression coefficients and feature selection. More specifically, C_0 is a standard smoothness condition on the penalty r_n that facilitates obtaining estimators by differentiating the objective function $L_n(\boldsymbol{\theta})$ when solving (7) and for studying the asymptotic properties of the estimators of the true non-zero regression coefficients. Conditions C_1 and C_2 are to control the contribution of r_n with respect to the log-likelihood function $l_n(\boldsymbol{\theta})$ in (5) to guarantee the existence of consistent estimators of $\boldsymbol{\theta}_0$. The second part of Condition C_1 is often referred to as a minimum-signal assumption which is necessary to guarantee the selection consistency; please see the last paragraph in Section 4.2 for more discussion. Under conditions C_3 and C^*_3, the penalty function r_n grows sufficiently fast in a vanishing neighborhood of $\theta = 0$ resulting in feature selection consistency (sparsity) property of the MPLE. The implications of these conditions for the Lasso, AdaLasso, SCAD, and MCP are explained after each theorem in Sections 4.1 and 4.2.

To focus on the main results, regularity conditions R_1-R_5 on the family $\mathcal{F} = \{h(y;\eta,\phi) : \eta \in \mathbb{R}, \text{and } \phi \in \mathbb{R}^+\}$ are given in Appendix B. Condition R_1 is on identifiability of the model which makes the estimation problem of interest well-defined;

see Section 2 for more on identifiability of MOE models. Additionally, the common support condition facilitates interchanging differentiation and integration operations on the density. R_2 is a smoothness condition on the density required in Taylor's expansions for asymptotic analyses while R_3 guarantees the asymptotic existence of the MPLE of the model parameters. R_4 posits positive definiteness and finiteness of the Fisher information while R_5 allows interchanging of the expectation and the limits due to the dominant convergence theorem. The most popular MOEs that satisfy the conditions are with experts h belonging to the exponential family including Gaussian, Poisson, and Binomial with number of trails $m > 2K - 1$.

In what follows, we study asymptotic properties of the MPLE $\widehat{\boldsymbol{\theta}}_n$ under two scenarios when p_n grows slowly as a function of the sample size n and when p_n could be as large as n. The proofs are given in the Supplementary Material.

4.1. Dimension p_n grows slowly with n

Theorems 1 and 2 extends the results of Fan and Peng (2004) for (generalized) linear regression models to MOEs with diverging number of parameters, where we also perform group variable selection.

Theorem 1. *Let $(\boldsymbol{x}_i, Y_i), i = 1, 2, \ldots, n$, be a random sample with the conditional density in (1) and a joint density satisfying the regularity conditions R_1-R_5 in Appendix B. Assume that the penalty r_n and $(\lambda_n, \lambda_n^*, \tau_n^*)$ satisfy Conditions C_0-C_2. If $\frac{p_n^2}{\sqrt{n}} \to 0$, as $n \to \infty$, there exists a local maximizer $\widehat{\boldsymbol{\theta}}_n$ of the penalized log-likelihood $L_n(\boldsymbol{\theta})$ in (5) such that $\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_2 = O_p\{\sqrt{\frac{p_n}{n}}(1 + a_n)\}$, where a_n is in given in (12).*

Theorem 1 guaranties the existence of a $\sqrt{n/p_n}$-consistent estimator of the parameter-vector $\boldsymbol{\theta}_0$ of the sparse MOE model, similar to the ordinary MLE, as long as r_n and the tuning parameters $(\lambda_n, \lambda_n^*, \tau_n^*)$ are chosen such that $a_n = O(1)$. This is also similar to the result of Huber (1973) for M-estimators in the context of robust regression in which the number of parameters diverges. For the Lasso, AdaLasso, SCAD, and MCP this translates into the choices of the parameters (λ_n, λ_n^*) and τ_n^* according to Conditions C_0-C_2. More specifically, for the Lasso, one could choose $\sqrt{n}\max\{\lambda_n, \lambda_n^*\} = O(1)$, $\tau_n^* \max_{j \in A_n^*} \|\boldsymbol{\alpha}_{\cdot j}^0\|_2 = o(\sqrt{n})$ and $\tau_n^* = o(n)$. For SCAD and MCP, by the minimum-signal condition in C_1, we have $a_n = 0$ and τ_n^* has to satisfy the same conditions as above. For AdaLasso, basically the weights ω and ω^* coupled with (λ_n, λ_n^*) are to be chosen so that $a_n = O(1)$. This implies that we need $\sqrt{n}\lambda_n(\max_{1 \le k \le K} \max_{j \in A_{kn}} \omega_{kj}) = O_p(1)$ and $\sqrt{n}\lambda_n^*(\max_{j \in A_n^*} \omega_j^*) = O_p(1)$, where ω_{kj} and ω_j^* are (possibly random) weights in AdaLasso; more details are provided in the discussion after Theorem 2 below.

Theorem 2 investigates even more interesting properties of the estimator $\widehat{\boldsymbol{\theta}}_n$ such as the consistency in feature selection and also asymptotic normality of the estimator $\widehat{\boldsymbol{\theta}}_n$ in estimating the true non-zero regression coefficients in both the gating network and the experts. Recall the partitioning $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{01}, \boldsymbol{\theta}_{02})$ such that $\boldsymbol{\theta}_{02} = \mathbf{0}$. Also, consider the partitioning $\widehat{\boldsymbol{\theta}}_n = (\widehat{\boldsymbol{\theta}}_{n1}, \widehat{\boldsymbol{\theta}}_{n2})$ such that $\dim(\widehat{\boldsymbol{\theta}}_{n1}) = \dim(\boldsymbol{\theta}_{01})$ and $\dim(\widehat{\boldsymbol{\theta}}_{n2}) = \dim(\boldsymbol{\theta}_{02})$. Let $\boldsymbol{B}_n$ be a constant matrix of dimension $l \times \dim(\widehat{\boldsymbol{\theta}}_{n1})$, $l < \infty$, such that $\boldsymbol{B}_n \boldsymbol{B}_n^\top \to \boldsymbol{B}$ and $\boldsymbol{B}$ is a positive definite symmetric matrix. Note that $\boldsymbol{B}_n \widehat{\boldsymbol{\theta}}_{n1}$ has the fixed dimension $l \times 1$. Let $\boldsymbol{R}_n'(\boldsymbol{\theta})$ and $\boldsymbol{R}_n''(\boldsymbol{\theta})$ be the gradient and Hessian of $\boldsymbol{R}_n$ in (6) with respect to $\boldsymbol{\theta}$.

Theorem 2. *Assume that the conditions of Theorem 1 are fulfilled, and let r_n and $(\lambda_n, \lambda_n^*, \tau_n^*)$ also satisfy Condition C_3. If $\frac{p_n^{2.5}}{\sqrt{n}} \to 0$, then for any $\sqrt{n/p_n}$-consistent estimator $\widehat{\boldsymbol{\theta}}_n = (\widehat{\boldsymbol{\theta}}_{n1}, \widehat{\boldsymbol{\theta}}_{n2})$ of $\boldsymbol{\theta}_0$, we have that, as $n \to \infty$,*

(i) Sparsity: $P(\widehat{\boldsymbol{\theta}}_{n2} = \mathbf{0}) \to 1$.

(ii) Asymptotic normality:

$$\sqrt{n}\boldsymbol{B}_n \, \mathcal{I}_{n1}^{-1/2}(\boldsymbol{\theta}_{01}) \left\{ \left[\mathcal{I}_{n1}(\boldsymbol{\theta}_{01}) + \frac{\boldsymbol{R}_n''(\boldsymbol{\theta}_{01})}{n} \right] (\widehat{\boldsymbol{\theta}}_{n1} - \boldsymbol{\theta}_{01}) + \frac{\boldsymbol{R}_n'(\boldsymbol{\theta}_{01})}{n} \right\} \xrightarrow{d} N(\mathbf{0}, \boldsymbol{B}),$$

where $\mathcal{I}_{n1}(\boldsymbol{\theta}_{01})$ is the Fisher information of the true MOE with $\boldsymbol{\theta}_{02} = \mathbf{0}$.

The estimator $\widehat{\boldsymbol{\theta}}_n$ with properties in Theorems 1 and 2 is called an *oracle estimator* as defined in Fan and Peng (2004). The estimators based on the penalty functions SCAD and MCP, and AdaLasso have the oracle property but not the one based on the Lasso. To achieve sparsity for Lasso, SCAD, and MCP, according to condition C_3 we require $\sqrt{n/p_n}\lambda_n$ and $\sqrt{n/p_n}\lambda_n^* \to \infty$, as $n \to \infty$. For AdaLasso, we require $\sqrt{n/p_n}\lambda_n(\min_{1 \le k \le K} \min_{j \notin A_{kn}} \omega_{kj})$ and $\sqrt{n/p_n}\lambda_n^*(\min_{j \notin A_n^*} \omega_j^*) \to \infty$, as $n \to \infty$. For Lasso, the required choices of (λ_n, λ_n^*) lead to an explosive bias for the non-zero estimators $\widehat{\boldsymbol{\theta}}_{n1}$ as described in Theorem 2-(ii). More specifically, the bias term $\boldsymbol{R}_n'(\boldsymbol{\theta}_{01})/n \sim (\lambda_n, \lambda_n^*)$ will go to zero slower than $n^{-\frac{1}{2}}$ in the Lasso case. On the other hand, for SCAD and MCP penalties, we have $\boldsymbol{R}_n'(\boldsymbol{\theta}_{01})/n = 0$, for any $n \ge 1$, and hence the aforementioned choices of (λ_n, λ_n^*) guarantees the oracle property of the MPLE, as long as $\lambda_n, \lambda_n^* \to 0$, as $n \to \infty$. For AdaLasso, if we choose the weights such that for all $k = 1, \ldots, K, j \in A_{kn}$, $\omega_{kj} = O(1)$, and $\omega_j^* = O(1)$, for all $j \in A_n^*$, and for all $k = 1, \ldots, K, j \notin A_{kn}$, $\omega_{kj}/\sqrt{p_n} \to \infty$, and $\omega_j^*/\sqrt{p_n} \to \infty$, for all $j \notin A_n^*$, then $\lambda_n, \lambda_n^* \sim n^{-1/2}$ suffices to achieve the oracle property. In practice, we may use the weights $\omega_{kj} = (\widetilde{\beta}_{kj})^{-1}$ and $\omega_j^* = (\widetilde{\boldsymbol{\alpha}}_{\cdot j})^{-1}$, where $(\widetilde{\beta}_{kj}, \widetilde{\boldsymbol{\alpha}}_{\cdot j})$ are the MLE of the parameters obtained by maximizing the log-likelihood $l_n(\boldsymbol{\theta})$ in (4). The weights satisfy the required conditions. Note that the ridge tuning parameter τ_n^* is chosen according to condition C_1 as explained after Theorem 1 above.

4.2. Dimension p_n is comparable to the sample size n

In this section, we extend the results of Theorems 1 and 2 to the case where the dimension p_n grows much faster than $n^{1/4}$ and comparable to the sample size n. Consequently, as shown below, the rate of consistency of the MPLE in this case will depend on the sparsity factor s_n. Recall that s_n is defined as the maximum number of the true non-zero regression coefficients in the experts and the gating network of an sparse MOE model.

Theorem 3. *Assume that the conditions of Theorem 1 hold. If* $\frac{s_n^2}{\sqrt{n}} \to 0$ *and* $s_n(p_n - s_n) = o(n)$, *as* $n \to \infty$, *then there exists a local maximizer* $\widehat{\theta}_n$ *of the penalized log-likelihood* $L_n(\theta)$ *in (5) such that* $\|\widehat{\theta}_n - \theta_0\|_2 = O_p\{\sqrt{\frac{s_n}{n}}(1 + a_n)\}$, *where* a_n *is given in (12).*

Note that the rate of consistency of the MPLE under the conditions of Theorem 3 is $\sqrt{n/s_n}$, as long as $a_n = O(1)$, while the dimension p_n grows faster than what is considered in Section 4.1. In Theorem 1, however, the rate of consistency is $\sqrt{n/p_n}$. It is worth noting that the growth rate of p_n, as a function of the sample size n, in Theorem 1 is similar to that of s_n in Theorem 3. For example, in Theorem 3, we could have $p_n = o(n^{\gamma_1})$ and $s_n = O(n^{\gamma_2})$, where $\gamma_1 > \gamma_2 > 0$, $\gamma_2 < 1/4$ and $\gamma_1 + \gamma_2 \leq 1$. The discussion provided after Theorem 1 on the choices of the tuning parameters $(\lambda_n, \lambda_n^*, \tau_n^*)$, and the weights $(\omega_{kj}, \omega_j^*)$, to assure $a_n = O(1)$ for the four penalties still holds here.

Theorem 4 that follows investigates the conditions under which the MPLE has the oracle property. Let $\boldsymbol{D}_n$ be a constant matrix of dimension $l^* \times dim(\widehat{\theta}_{n1})$, $l^* < \infty$, such that $\boldsymbol{D}_n \boldsymbol{D}_n^\top \to \boldsymbol{D}$, and $\boldsymbol{D}$ is a positive definite matrix. Theorem 4 seeks the asymptotic distribution of the finite linear transformation $\boldsymbol{D}_n \widehat{\theta}_{n1}$, which has the fixed dimension $l^* \times 1$.

Theorem 4. *Assume that the conditions of Theorem 3 hold, and let* $(r_n, \lambda_n, \lambda_n^*, \tau_n^*)$ *satisfy Condition* C_3^*. *If* $\frac{s_n^{2.5}}{\sqrt{n}} \to 0$, *then for any* $\sqrt{n/s_n}$*-consistent estimator* $\widehat{\theta}_n = (\widehat{\theta}_{n1}, \widehat{\theta}_{n2})$ *of* θ_0, *as* $n \to \infty$,

(i) Sparsity: $P(\widehat{\theta}_{n2} = \mathbf{0}) \to 1$.

(ii) Asymptotic normality:

$$\sqrt{n}\boldsymbol{D}_n \, \mathcal{I}_{n1}^{-1/2}(\theta_{01})\left\{\left[\mathcal{I}_{n1}(\theta_{01}) + \frac{\boldsymbol{R}_n''(\theta_{01})}{n}\right](\widehat{\theta}_{n1} - \theta_{01}) + \frac{\boldsymbol{R}_n'(\theta_{01})}{n}\right\} \xrightarrow{d} N(\mathbf{0}, \boldsymbol{D}),$$

where $\mathcal{I}_{n1}(\theta_{01})$ *is the Fisher information of the true sparse* MOE *with* $\theta_{02} = \mathbf{0}$.

Note that Condition C_3^* in Theorem 4 is to ensure sparsity of the MPLE. The discussion provided after Theorem 3 in Section 4.1 regarding the choices of tuning parameters for the penalties under our consideration applies here except that p_n is to be replaced by $s_n p_n$. Hence, theoretically the estimator $\widehat{\theta}_n$ based on the Lasso does not have the oracle property while the one based on the AdaLasso, SCAD, or MCP does. Nevertheless, the MPLE based on all these penalties preserves the sparsity property which is important in high dimensions. It is worth noting that, as expected, Condition C_3^* in Theorem 4 compared to Condition C_3 in Theorem 3 for sparsity requires (asymptotically) larger choices of (λ_n, λ_n^*) compared to the low-dimensional case discussed in Section 4.1.

Condition C_1 is commonly referred to as a minimum-signal condition in the variable selection literature. Basically, it implies that together with condition C_3 or C_3^*, those non-zero regression coefficients that satisfy $\beta_{kj}^0 > \lambda_n$ or $\boldsymbol{\alpha}_{\cdot j}^0 > \lambda_n^*$, where $(\lambda_n, \lambda_n^*) \to 0$ as $n \to \infty$, are detectable by the proposed regularization method and will be estimated non-zero, i.e. variable selection consistency property. On the other hand, those coefficients that are below the thresholds, the weak signals, will most likely be estimated as zero by the regularization method. Without this condition, it may be possible to establish certain estimation error bounds but not really selection consistency as those weak-signal regression parameters most likely will be estimated as zero; see also Roy et al. (2023) for a recent work on weak signal recovery in high-dimensional regression. Fang et al. (2021) proposed a two-step procedure based on both variable selection and ridge regression estimators in linear regression models that were shown to be capable of detecting weak signals and providing an estimation of both strong and weak signal. This is a future research direction worthy of investigation in the context of MOE models.

According to Theorems 1 and 4, for large n, the approximate distribution of linear transformations of the sub-vector $\widehat{\theta}_{n1}$, which estimates linear transformations of θ_{01} (the true non-zero regression coefficients), is normal. For penalties such as SCAD and MCP, the terms $\boldsymbol{R}_n'(\theta_{01})/n$ and $\boldsymbol{R}_n''(\theta_{01})/n$ can be ignored. Therefore, by estimating the information matrix $\mathcal{I}_{n1}(\theta_{01})$–typically done in MOE models using the empirical information matrix derived from the complete log-likelihood function in (14) (McLachlan and Peel, 2000)–one may attempt to perform further statistical inference, such as hypothesis testing and constructing confidence intervals for the regression coefficients of the selected model. However, such inference referred to as naive inference is reserved as the true sparse structure (oracle's perspective) of the model is not known in advance and it is estimated by the penalization method. Hence, in practice due to the variable selection stage the dimension of the sub-vector $\widehat{\theta}_{n1}$ is random and may not be equal to the dimension of the sub-vector θ_{01}, and hence asymptotically normal distribution may be distorted. The extra variability due to the variable selection needs to be taken into account for a further inference and is referred to as post-selection inference (PoSI, Berk et al. (2013)). There has been a surge of research on PoSI in recent years for (generalized) linear regression models (Zhang et al., 2022). The topic of PoSI in mixture of regression models, considered as a special case of MOE with the gating network $g_k(\boldsymbol{x}; \boldsymbol{\alpha}) = g_k$ assumed to be independent of features $\boldsymbol{x}$, was studied by Khalili and Vidyashankar (2018); PoSI in general MOEs requires a careful study and is a topic of future research.

5. Numerical algorithm

To solve the optimization problem presented in (7), we develop a modified EM algorithm (Dempster et al., 1977) that features a coordinate descent type M-step adapted to our penalized likelihood. The previous studies have shown successful application of coordinate descent methods combined with the EM algorithm in FMR models. For example, Städler et al. (2010) used the coordinate descent together with the EM algorithm in high-dimensional Gaussian FMR models with the Lasso penalty. Friedman et al. (2010) developed algorithms that make use of the coordinate descent along a regularization path for variable selection problems in generalized linear models with convex penalties. These methods are especially efficient for solving high-dimensional models. In addition, due to the complexity of MOEs, the adjusted version of the EM algorithm applies the proximal gradient descent algorithm in each coordinate descent circle of the M-step to obtain an approximation to the optimization problem. We proceed as follows.

The complete log-likelihood function (McLachlan and Peel, 2000) of an MOE model is given by

$$l_n^c(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\Big\{\log g_k(\boldsymbol{x}_i;\boldsymbol{\alpha}) + \log h(y_i;\eta_k(\boldsymbol{x}_i),\phi_k)\Big\},$$

where z_{ik} is an unobservable indicator variable showing that, given $\boldsymbol{x}_i$, the observation y_i is generated from the kth expert density $h(y_i;\eta_k(\boldsymbol{x}_i),\phi_k)$. The complete penalized log-likelihood is given by

$$L_n^c(\boldsymbol{\theta}) = l_n^c(\boldsymbol{\theta}) - \boldsymbol{R}_n(\boldsymbol{\theta}). \tag{13}$$

For fixed K and the tuning parameters $(\lambda,\lambda^*,\tau^*)$, the EM algorithm maximizes (13) iteratively in two steps as follows. Data-adaptive selections of K and the tuning parameters are discussed in Section 6.

E-step: At the $(m+1)$th iteration, given the data and current estimate $\boldsymbol{\theta}^{(m)}$, we compute the conditional expectation of $L_n^c(\boldsymbol{\theta})$ with respect to the unobservable random variables Z_{ik}'s. Thus,

$$\begin{aligned} E\Big\{L_n^c(\boldsymbol{\theta})|\text{data},\boldsymbol{\theta}^{(m)}\Big\} &= Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(m)}) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}^{(m)}\Big\{\log h(y_i;\eta_k(\boldsymbol{x}_i),\phi_k) + \log g_k(\boldsymbol{x}_i;\boldsymbol{\alpha})\Big\} - \boldsymbol{R}_n(\boldsymbol{\theta}) \\ &= Q_1(\boldsymbol{\theta};\boldsymbol{\theta}^{(m)}) + Q_2(\boldsymbol{\alpha};\boldsymbol{\theta}^{(m)}) - \boldsymbol{R}_n(\boldsymbol{\theta}), \end{aligned} \tag{14}$$

where

$$\tau_{ik}^{(m)} = \boldsymbol{E}\big(Z_{ik}|\boldsymbol{\theta}^{(m)},\boldsymbol{x}_i,y_i\big) = \frac{g_k(\boldsymbol{x}_i;\boldsymbol{\alpha}^{(m)})h(y_i;\eta_k^{(m)}(\boldsymbol{x}_i),\phi_k^{(m)})}{\sum_{k=1}^{K} g_k(\boldsymbol{x}_i;\boldsymbol{\alpha}^{(m)})h(y_i;\eta_k^{(m)}(\boldsymbol{x}_i),\phi_k^{(m)})}. \tag{15}$$

The leading functions in (14) are

$$Q_1(\boldsymbol{\theta};\boldsymbol{\theta}^{(m)}) = \sum_{k=1}^{K}\sum_{i=1}^{n}\{\tau_{ik}^{(m)}\log h(y_i;\eta_k(\boldsymbol{x}_i),\phi_k)\} = \sum_{k=1}^{K} Q_{1k}(\boldsymbol{\theta}_k;\boldsymbol{\theta}^{(m)})$$

with $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k,\phi_k)$, and $Q_{1k}(\boldsymbol{\theta}_k;\boldsymbol{\theta}^{(m)})$ are the inner sums $\sum_{i=1}^{n}\{\cdot\}$. Also, using (2), we have

$$\begin{aligned} Q_2(\boldsymbol{\alpha};\boldsymbol{\theta}^{(m)}) &= \sum_{k=1}^{K}\sum_{i=1}^{n}\tau_{ik}^{(m)}\log g_k(\boldsymbol{x}_i;\boldsymbol{\alpha}) \\ &= \sum_{k=1}^{K-1}\sum_{i=1}^{n}\tau_{ik}^{(m)}\widetilde{\boldsymbol{x}}_i^{\top}\widetilde{\boldsymbol{\alpha}}_k - \sum_{i=1}^{n}\log\Big(1+\sum_{k=1}^{K-1}\exp(\widetilde{\boldsymbol{x}}_i^{\top}\widetilde{\boldsymbol{\alpha}}_k)\Big), \end{aligned}$$

where $\widetilde{\boldsymbol{x}}_i = (1,\boldsymbol{x}_i^{\top})^{\top}$ and $\widetilde{\boldsymbol{\alpha}}_k = (\alpha_{0k},\boldsymbol{\alpha}_k^{\top})^{\top}$. In summary, the **E-step** boils down to the computation of the weights in (15).

M-step: In this step, we maximize the function $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(m)})$ in (14) with respect to $\boldsymbol{\theta}$. The maximization can be done using either the proximal gradient or Newton–Raphson-type algorithms in which the leading terms Q_1 and Q_2 in (14) are locally approximated by quadratic functions of $\boldsymbol{\theta}$ (Nesterov, 2004). To handle the folded concave penalties such as SCAD and MCP, we develop a proximal gradient method combined with the local linear approximation (LLA), inspired by Zou and Li (2008). This algorithm can avoid computation of the Hessian matrix as required in the local quadratic approximation method (LQA) (Fan and Li, 2001), which is particularly slow for large dimensional vectors $(\boldsymbol{\alpha},\boldsymbol{\beta}_k), k=1,\ldots,K$. On the other hand, for AdaLasso, we use the regular gradient descent method. In what follows, we only focus on the regression parameters, as the updates for the dispersion parameters ϕ_k can also be obtained by maximizing $Q_{1k}(\boldsymbol{\theta}_k;\boldsymbol{\theta}^{(m)})$ with respect to ϕ_k at each iteration of the EM.

Thus, in the M-step by using the LLA to r_n when necessary, the updates of $\boldsymbol{\beta}_k$ are obtained separately for each $k=1,\ldots,K$, by minimizing the following function with respect to $\boldsymbol{\beta}_k$,

$$\mathcal{L}_{1k}(\boldsymbol{\beta}_k;\boldsymbol{\theta}^{(m)}) + \sum_{j=1}^{p}\omega_{kj}^{(m)}|\beta_{kj}|, \tag{16}$$

with $\mathcal{L}_{1k}(\boldsymbol{\beta}_k;\boldsymbol{\theta}^{(m)}) = -Q_{1k}(\boldsymbol{\beta}_k;\boldsymbol{\theta}^{(m)})/n$. Also, the updates of $\boldsymbol{\alpha}$ are obtained by minimizing

$$\mathcal{L}_2(\boldsymbol{\alpha};\boldsymbol{\theta}^{(m)}) + \sum_{j=1}^{p}\left\{\omega_j^{*(m)}\|\boldsymbol{\alpha}_{\cdot,j}\|_2 + \frac{\tau^*}{2n}\|\boldsymbol{\alpha}_{\cdot,j}\|_2^2\right\} \tag{17}$$

with $\mathcal{L}_2(\boldsymbol{\alpha};\boldsymbol{\theta}^{(m)}) = -Q_2(\boldsymbol{\alpha};\boldsymbol{\theta}^{(m)})/n$. The minimization is done as follows.

Given $\rho_1 > 0$, we locally majorize the function in (16) by the regularized quadratic function

$$\begin{aligned} G_1(\boldsymbol{\beta}_k,\rho_1) :=\ & \mathcal{L}_{1k}(\boldsymbol{\beta}_k^{(m)};\boldsymbol{\theta}^{(m)}) + \left[\frac{\partial\mathcal{L}_{1k}(\boldsymbol{\beta}_k^{(m)};\boldsymbol{\theta}^{(m)})}{\partial\boldsymbol{\beta}_k}\right]^\top(\boldsymbol{\beta}_k - \boldsymbol{\beta}_k^{(m)}) \\ & + \frac{\rho_1}{2}\|\boldsymbol{\beta}_k - \boldsymbol{\beta}_k^{(m)}\|_2^2 + \sum_{j=1}^{p}\omega_{kj}^{(m)}|\beta_{kj}|. \end{aligned} \tag{18}$$

Minimizing function $G_1(\boldsymbol{\beta}_k,\rho_1)$ with respect to $\boldsymbol{\beta}_k$ results in the closed-form updates

$$\boldsymbol{\beta}_k^{(m+1)} = S(\boldsymbol{z}_k^{(m)};\rho_1^{-1}\boldsymbol{\omega}_k^{(m)}), \tag{19}$$

for all $k = 1,\dots,K$, where

$$\boldsymbol{z}_k^{(m)} = \boldsymbol{\beta}_k^{(m)} - \rho_1^{-1}\left(\frac{\partial\mathcal{L}_{1k}(\boldsymbol{\beta}_k^{(m)};\boldsymbol{\theta}^{(m)})}{\partial\boldsymbol{\beta}_k}\right)$$

and $S(\mathbf{z};\mathbf{w}) = [S(z_1;w_1),\dots,S(z_p;w_p)]^\top$ with $S(z;w) = (1-\frac{w}{|z|})_+ z$ as the soft-thresholding operator (Breheny and Huang, 2015; Donoho and Johnstone, 1994). The weights in (16) are for SCAD, MCP, the weights are $\omega_{kj}^{(m)} = r'(\beta_{kj}^{(m)};\lambda)/n$, where r_n' is the first derivative of r_n with respect to β_{kj}. For the Lasso and AdaLasso, we do not need to use the LLA procedure. Hence, we fix the weight $\omega_{kj}^{(m)} = \lambda$ for the lasso. For AdaLasso, the weights are chosen as λ multiplied by the reciprocal of the absolute value of the MLE of β_{kj}'s, as suggested by Zou (2006). When the dimension of $\boldsymbol{x}$ is large and the MLE is not feasible, one may use ridge-type estimates of β_{kj}'s to construct the weights.

A similar method is used to obtain updates of $\boldsymbol{\alpha}$ in the M-step. Given $\rho_2 > 0$, we locally majorize the function in (17) by the regularized quadratic function (up to some constants)

$$\begin{aligned} G_2(\boldsymbol{\alpha}_{\cdot,j},\rho_2) :=\ & \mathcal{L}_2(\boldsymbol{\alpha}^{(m)};\boldsymbol{\theta}^{(m)}) + \sum_{j=1}^{p}\left\{\left[\frac{\partial\mathcal{L}_2(\boldsymbol{\alpha};\boldsymbol{\theta}^{(m)})}{\partial\boldsymbol{\alpha}_{\cdot,j}}\right]^\top\left(\boldsymbol{\alpha}_{\cdot,j} - \boldsymbol{\alpha}_{\cdot,j}^{(m)}\right)\right. \\ & \left. + \frac{\rho_2}{2}\|\boldsymbol{\alpha}_{\cdot,j} - \boldsymbol{\alpha}_{\cdot,j}^{(m)}\|_2^2\right\} + \sum_{j=1}^{p}\left\{\omega_j^{*(m)}\|\boldsymbol{\alpha}_{\cdot,j}\|_2 + \frac{\tau^*}{2n}\|\boldsymbol{\alpha}_{\cdot,j}\|_2^2\right\}. \end{aligned} \tag{20}$$

Minimizing this function with respect to $\boldsymbol{\alpha}_{\cdot,j}$ results in the closed form updates

$$\boldsymbol{\alpha}_{\cdot,j}^{(m+1)} = S(\boldsymbol{z}_j^{(m)};(\rho_2+\tau^*/n)^{-1}\omega_j^{*(m)}),\ j = 1,\dots,p, \tag{21}$$

where

$$\boldsymbol{z}_j^{(m)} = (\rho_2+\tau^*/n)^{-1}\left[\rho_2\boldsymbol{\alpha}_{\cdot,j}^{(m)} - \left(\frac{\partial\mathcal{L}_2(\boldsymbol{\alpha};\boldsymbol{\theta}^{(m)})}{\partial\boldsymbol{\alpha}_{\cdot,j}}\right)\right]$$

and $S(\mathbf{z};\omega^*) = (1-\frac{\omega^*}{\|\mathbf{z}\|_2})_+\mathbf{z}$ is the multivariate soft-thresholding operator (Breheny and Huang, 2015; Donoho and Johnstone, 1994) for group Lasso. Note that the weights $\omega_j^{*(m)}$ in (20) are chosen in a similar fashion to the weights $\omega_{kj}^{(m)}$ as described above where λ is replaced by λ^*.

Line-search In each iteration of the EM, the two parameters ρ_1 and ρ_2 in the M-step are chosen using a backtracking line search (Boyd and Vandenberghe, 2004) such that the functions in (16) and (17), when evaluated at the updating values using the chosen stepsizes, are less than or equal to, respectively, their majorizing functions in (18) and (20).

Specifically, to determine step size ρ_1 in (18), we first initialize ρ_1 with some $\rho_1^{\max} > 0$ and repeatedly shrink ρ_1 with $\rho_1 \leftarrow \epsilon^{-1}\rho_1$ for some pre-chosen $0 < \epsilon < 1$ until the following condition holds

$$G_1(\boldsymbol{\beta}_k^{(m+1)},\rho_1) \le G_1(\boldsymbol{\beta}_k^{(m)},\rho_1), \tag{22}$$

where G_1 is defined in (18) and $\boldsymbol{\beta}_k^{(m+1)}$ is given in (19). For determining step size ρ_2 in (20), we initialize ρ_2 with some $\rho_2^{\max} > 0$ and repeatedly shrink ρ_2 with $\rho_2 \leftarrow \epsilon^{-1}\rho_2$ for some pre-chosen $0 < \epsilon < 1$ until the following condition holds

$$G_2(\boldsymbol{\alpha}^{(m+1)},\rho_2) \le G_2(\boldsymbol{\alpha}^{(m)},\rho_2), \tag{23}$$

where G_2 is defined in (20) and $\boldsymbol{\alpha}^{(m+1)}$ is given in (21). We summarize our algorithm in Algorithm 1.

Algorithm 1: Modified EM Algorithm.

Initialization: Choose initial values $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\beta}^{(0)}, \boldsymbol{\alpha}^{(0)}, \boldsymbol{\phi}^{(0)})$; Set tuning parameters $(\lambda, \lambda^*, \tau^*)$; Set $m = 0$ and convergence criterion $(\delta, max.iter)$;
while $|pl_n(\boldsymbol{\theta}^{(m+1)}) - pl_n(\boldsymbol{\theta}^{(m)})| \geq \delta$ *and* $m \leq max.iter$ **do**
  E-step: Compute weights $\tau_{ik}^{(m)}$ in (15), for all i, k;
  M-step:
  for $k = 1, 2, \dots K$ *and* $j = 1, 2, \dots p$ **do**

$$\boldsymbol{\beta}_k^{(m+1)} \leftarrow S(\boldsymbol{z}_k^{(m)}; \rho_1^{-1}\, \boldsymbol{\omega}_k).$$
$$\boldsymbol{\alpha}_{\cdot,j}^{(m+1)} \leftarrow S(\boldsymbol{z}_j^{(m)}; (\rho_2 + \tau^*/n)^{-1}\omega_j^*)$$
$$\phi_k^{(m+1)} \leftarrow \arg\max_{\phi_k} Q_{1k}(\boldsymbol{\theta}_k; \boldsymbol{\beta}_k^{(m+1)}, \boldsymbol{\alpha}^{(m+1)}).$$

  end
  Update the iteration counter $m = m + 1$;
end

6. Implementation details

Initialization In our study, we adopt the following procedure for initialization: Firstly, we perform univariate clustering on the response variable to partition the data into K groups, corresponding to the number of components of the mixture model. For initialization of the regression coefficient vectors $(\beta_{01}, \boldsymbol{\beta}_1, \dots, \beta_{0K}, \boldsymbol{\beta}_K)$, we fit separate (generalized) linear regression models to each cluster using the data points assigned to that cluster. This ensures that the initial means are informed by the actual distribution of the data. Regarding the initialization of the coefficient vectors $(\alpha_{01}, \boldsymbol{\alpha}_1, \dots, \alpha_{0,K-1}, \boldsymbol{\alpha}_{K-1})$ for the mixing probabilities, we use the cluster memberships of all the data points as a categorical response variable and fit a multinomial logistic regression model to the data. This provides a starting point for the EM algorithm by reflecting the preliminary groupings of the data. To initialize the variance parameters $(\phi_1, \phi_2, \dots, \phi_K)$, we employ the estimated variances of the response variable within each cluster.

Convergence of the M-step using the LLA procedure During the M-step of the EM algorithm, the iterative procedure using the LLA provides convergence guarantees. Within the M-step, we encounter two convex optimization sub-problems: the minimization of (16) and (17) with respect to $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K$ and $\boldsymbol{\alpha}$, respectively. Both are convex problems. We perform this by first majorizing the aforementioned two objective functions by the two convex functions (18) and (20), respectively. This majorization process involves applying quadratic majorization to the leading (likelihood) terms and employing LLA majorization for the penalty function. This approach is an instance of the majorization-minimization (MM) algorithm, the convergence of which has been extensively studied in the literature, as demonstrated by Heiser (1995) and Lange et al. (2000).

Selection of tuning parameters As discussed in Section 3, the main purpose of the ridge penalty on $\boldsymbol{\alpha}$ is to help improving stability of the numerical algorithm. For the ridge tuning parameter, we use $\tau^* = C \log n$, for some constant $C > 0$ which was taken $C = 0.01$ in our simulations. This value satisfies the conditions required in our theory and it further works in our simulations. We next discuss data-driven selection of (λ, λ^*) and the mixture order K.

For a fixed mixture order K, we use a BIC-type criterion (Wang et al., 2007) to choose (λ, λ^*) from a two-dimensional grid expanded over $[0, \lambda_{\max}]^2$ for some pre-specified value $\lambda_{\max}$. More specifically, let $a_1, \dots, a_L$ be a grid of the interval $[0, \lambda_{\max}]$. For each pair $(a_l, a_{l'}), l, l' = 1, \dots, L$, corresponding to (λ, λ^*), let $\widehat{\boldsymbol{\theta}}_{ll'}$ be the MPLE in (7). Due to the group selection of the gating parameters, we calculate the total number of estimated non-zero regression parameters in $\widehat{\boldsymbol{\theta}}_{ll'}$ as $\text{DF}(l, l') = \sum_{k=1}^{K}\sum_{j=1}^{p} 1\{\widehat{\beta}_{kj}(l, l') \neq 0\} + \sum_{k=1}^{K-1}\sum_{j=1}^{p} 1\{\widehat{\alpha}_{kj}(l, l') \neq 0\}$. We compute the BIC value

$$\text{BIC}_1(l, l') = -2l_n(\widehat{\boldsymbol{\theta}}_{ll'}) + (\log n)\ \text{DF}(l, l') \tag{24}$$

where l_n is the log-likelihood in (4). We choose a pair $(a_l, a_{l'})$ over the two-dimensional discrete grid as the optimal value of (λ, λ^*) that minimizes the BIC, that is, $(\widehat{\lambda}, \widehat{\lambda}^*) = \text{argmin}_{\{(a_l, a_{l'}): l, l' = 1, \dots, L\}}\ \text{BIC}_1(l, l')$. Let $\widehat{\boldsymbol{\theta}}(K)$ be the final parameter estimate corresponding to $(\widehat{\lambda}, \widehat{\lambda}^*)$, for any given mixture order K.

We estimate the mixture order as follows. The above process is repeated for each $K = 1, \dots, \mathcal{K}$, for some pre-specified upper bound $\mathcal{K}$. We then compute the total number of non-zero elements of $\widehat{\boldsymbol{\theta}}(K)$ as $\text{DF}(K) = \sum_{k=1}^{K}\sum_{j=1}^{p} 1\{\widehat{\beta}_{kj} \neq 0\} + \sum_{k=1}^{K-1}\sum_{j=1}^{p} 1\{\widehat{\alpha}_{kj} \neq 0\}$. We compute the BIC value

$$\text{BIC}_2(K) = -2l_n(\widehat{\boldsymbol{\theta}}(K)) + (\log n)\big(\text{DF}(K) + 3K - 1\big)\ ,\quad K = 1, \dots, \mathcal{K}, \tag{25}$$

where $3K - 1$ is the total number of estimated intercepts and dispersion parameters $(\widehat{\alpha}_{0k}, \widehat{\beta}_{0k}, \widehat{\phi}_k)$. The estimated order is $\widehat{K} = \text{argmin}_{1 \leq K \leq \mathcal{K}}\ \text{BIC}_2(K)$. Its performance is studied in the next section.

7. Simulation study

We carry out a simulation study to examine the finite-sample performance of the proposed methods. Each feature vector $\boldsymbol{x}$ is generated from a p-variate Gaussian distribution with mean zero and a covariance matrix with (i,j)-th element being $.5^{|i-j|}$. The corresponding design matrix will remain fixed throughout the data generation process. Given $\boldsymbol{x}$, the response y is generated from the MOE

$$f(y;\boldsymbol{x},\boldsymbol{\theta}) = \sum_{k=1}^{3} g_k(\boldsymbol{x};\boldsymbol{\alpha})\ \mathcal{N}(y;\mu_k(\boldsymbol{x}),\sigma_k^2),$$

where $\mathcal{N}(\cdot;\mu,\sigma^2)$ is the density function of Gaussian distribution with mean μ and variance σ^2. Here, we have $\mu_k(\boldsymbol{x}) = \beta_{0k} + \boldsymbol{\beta}_k^\top \boldsymbol{x}$, for $k = 1,2,3$, and the gating network

$$\log\left(\frac{g_k(\boldsymbol{x};\boldsymbol{\alpha})}{g_3(\boldsymbol{x};\boldsymbol{\alpha})}\right) = \alpha_{0k} + \boldsymbol{\alpha}_k^\top \boldsymbol{x}, \quad k = 1,2, \quad \text{and} \quad \sum_{k=1}^{3} g_k(\boldsymbol{x};\boldsymbol{\alpha}) = 1.$$

We have considered the following parameter settings with dimensions $p_n = n^\gamma$ for $\gamma = .5, .6, .7$, and the sample sizes $n = 200, 300, 400$. In all the cases, we set $\sigma_k^2 = 1$ for $k = 1,2,3$.

Setting (I): $n = 200$, $p_n = \{15, 24, 42\}$

$$\widetilde{\boldsymbol{\alpha}}_1 = (\alpha_{01}, \boldsymbol{\alpha}_1{}^\top)^\top = (-1,0,0,-1.5,0,0,-1.9,\ldots,0)^\top$$
$$\widetilde{\boldsymbol{\alpha}}_2 = (\alpha_{02}, \boldsymbol{\alpha}_2{}^\top)^\top = (-1.5,0,0,1.8,0,0,1.2,\ldots,0)^\top$$
$$\boldsymbol{\beta}_1 = (2.5,0,0,2.4,0,0,\ldots,0)^\top\ ,\ \boldsymbol{\beta}_2 = (-2,1.9,0,0,1.5,0,\ldots,0)^\top,$$
$$\boldsymbol{\beta}_3 = (0,0,-2.0,1.8,0,0,\ldots,0)^\top.$$

Setting (II): $n = 300$, $p_n = \{17, 30, 60\}$

$$\widetilde{\boldsymbol{\alpha}}_1 = (\alpha_{01}, \boldsymbol{\alpha}_1{}^\top)^\top = (-1,0,0,-1.5,0,0,-1.9,\ldots,0)^\top$$
$$\widetilde{\boldsymbol{\alpha}}_2 = (\alpha_{02}, \boldsymbol{\alpha}_2{}^\top)^\top = (-1.5,0,0,1.8,0,0,1.2,\ldots,0)^\top$$
$$\boldsymbol{\beta}_1 = (2.5,0,0,2.4,0,-1.5,\ldots,0)^\top\ ,\ \boldsymbol{\beta}_2 = (-2,1.9,0,0,1.5,0,2.0,\ldots,0)^\top,$$
$$\boldsymbol{\beta}_3 = (0,0,-2.0,1.8,0,0,-1.9,\ldots,0)^\top.$$

Setting (III) : $n = 400$, $p_n = \{20, 36, 70\}$

$$\widetilde{\boldsymbol{\alpha}}_1 = (\alpha_{01}, \boldsymbol{\alpha}_1{}^\top)^\top = (-1,0,0,-1.5,0,0,-1.9,0,1.0,\ldots,0)^\top$$
$$\widetilde{\boldsymbol{\alpha}}_2 = (\alpha_{02}, \boldsymbol{\alpha}_2{}^\top)^\top = (-1.5,0,0,1.8,0,0,1.2,0,-1.0,\ldots,0)^\top$$
$$\boldsymbol{\beta}_1 = (2.5,0,0,2.4,0,-1.5,0,1.5,\ldots,0)^\top\ ,\ \boldsymbol{\beta}_2 = (-2,1.9,0,0,1.5,0,2.0,-1.8,\ldots,0)^\top$$
$$\boldsymbol{\beta}_3 = (0,0,-2.0,1.8,0,0,-1.9,1.8,\ldots,0)^\top.$$

The total number of true non-zero regression coefficients in the above three settings are respectively 11,14 and 19. The dimension of the parameter vector $\boldsymbol{\theta}$ is $d_n = (2K-1)(p_n+1)+K$, see Section 2. The values of d_n corresponding to each of the settings are given in Tables 1 and 2 in the Appendix.

In the discussion below, let CIZ = # Correctly Identified Zero, CIN = # Correctly Identified Nonzero, IIZ = # Incorrectly Identified Zero and IIN = # Incorrectly Identified Nonzero regression coefficients. The specificity (SP) and sensitivity (SE) are respectively defined as SP = CIZ/(CIZ+IIN) and SE = CIN/(CIN+IIZ). We also report the empirical mean squared error (MSE) for each estimated regression parameter vector. Our results are based on $R = 200$ simulated samples from each of the above models, and are summarized in Tables 1 and 2 in the Appendix. We report the results based on the Lasso, AdaLasso, and SCAD; the MCP results were similar to SCAD and thus not reported here.

From Table 1, we can see that for each sample size and setting, as the dimension d_n as a function of $p_n = n^\gamma$ with $\gamma = .5, .6, .7$, increases, the MSE also increases which is expected. The MSE, corresponding to the same dimension d_n when fixing γ at each value .5, .6, .7 and increasing n, decreases. Estimation of the gating parameters $\boldsymbol{\alpha}_k$'s is more difficult than that of the experts parameters $\boldsymbol{\beta}_k$'s, which is mainly due to the multinomial nature of the gating network. Overall, the method based on the SCAD performs better than the Lasso and AdaLasso in terms of the MSE.

From Table 2, we can see that the method based on all the three penalties performs well in terms of both specificity (SP) and sensitivity (SE). For the largest dimension considered, corresponding to $p_n = n^{.7}$, the Lasso outperforms the other two penalties in terms of both (SP, SE) corresponding to the experts parameters. For the smaller dimensions corresponding to $p_n = n^{.5 \text{ or } .6}$, the three penalties perform more or less similarly. When the SE values are low, the corresponding MSE tends to be higher which is also expected. In summary, the performance of the proposed method shows that it provides a reliable new estimation and feature selection method for MOES when the number p of features is comparable to the sample size.

Finally, we assess the performance of the BIC in (25) for estimation of the mixture order K. For each simulated sample from the above model with correct order $K = 3$, we fit MOE models with $K = 1,\ldots,5$, and estimate the order using the BIC. Our results

are averaged over $R = 200$ simulated samples and are reported in Table 3. For the sample size $n = 200$, we can see that the BIC based on SCAD outperforms the other two penalties by detecting the correct $K = 3$ about 76% of times for all the three dimensions $d_n = 83, 128, 218$. As the sample size increases to $n = 300, 400$, the BIC based on the three penalties performs well corresponding to dimensions $d_n = 93,108$, by detecting the correct mixture order about 86% to 98%. For $d_n = 158, 188$, the Lasso and AdaLasso outperform SCAD, and as dimension increases to $d_n = 308, 358$ the BIC based on Lasso is the winner but clearly the order estimation becomes much harder for higher dimensions unless the sample size n increases.

8. Real data analysis

In this section we demonstrate the proposed methodology by analyzing a dataset available at http://jse.amstat.org/jse_data_archive.htm. The data contains trunk and limb body girth measurements at 12 well-defined sites, skeletal diameter measurements at nine well-defined body sites, as well as age, weight, height, and sex for 507 individuals; see Fig. 1, and the list of variables are given in Table 4. Heinz et al. (2003) used linear regression models to analyze relationship between weight (the response variable) and the aforementioned covariates. We re-analyze the data using sparse MOEs, as a generalization of linear models, allowing for potential heterogeneity of the effects of the covariates ($p = 24$) on the response variable weight. To avoid numerical issues, we standardize all the covariates to mean zero and variance one. We fit the sparse Gaussian MOEs with $K = 1, 2, 3, 4$ components and compare the fitted models using the BIC. Note that $K = 1$ corresponds to a linear model as fitted in Heinz et al. (2003) with more covariates. The BIC values for models with $K = 3, 4$ are larger than the ones corresponding to the models with $K = 1, 2$ components. On the other hand, the former models are not very different and thus we report the selected sparse MOE with $K = 3$. The selected model is based on the SCAD penalty which results in a more interpretable and spare model compared to the other penalties. The fitted Gaussian MOE model is

$$f(y; \boldsymbol{x}, \widehat{\boldsymbol{\theta}}) = \sum_{k=1}^{3} g_k(\boldsymbol{x}; \widehat{\boldsymbol{\alpha}})\ \mathcal{N}(y; \widehat{\mu}_k(\boldsymbol{x}), \widehat{\sigma}_k^2),$$

where $\widehat{\sigma}_1 = 1.68, \widehat{\sigma}_2 = 1.25, \widehat{\sigma}_3 = 1.23$, and

$$\begin{aligned}
\widehat{\mu}_1(\boldsymbol{x}) &= 68.3 + 5.12x_{12} + 2.43x_{15} + 2.81x_{17} + 1.57x_{19} + 3.18x_{23} \\
\widehat{\mu}_2(\boldsymbol{x}) &= 68.3 + 3.35x_{10} + 2.37x_{11} + 1.70x_{12} + 3.57x_{14} + 2.87x_{18} + 2.96x_{23} \\
\widehat{\mu}_3(\boldsymbol{x}) &= 68.3 - 2.77x_1 + 2.05x_5 + 5.47x_{10} + 4.68x_{12} + 4.07x_{15} - 3.98x_{16} \\
&\quad + 3.86x_{17} + 3.87x_{23}
\end{aligned} \tag{26}$$

and the gating network

$$\begin{aligned}
\log\left(\frac{g_1(\boldsymbol{x}; \widehat{\boldsymbol{\alpha}})}{g_3(\boldsymbol{x}; \widehat{\boldsymbol{\alpha}})}\right) &= 1.70 - .632x_2 - .970x_{11} \\
\log\left(\frac{g_2(\boldsymbol{x}; \widehat{\boldsymbol{\alpha}})}{g_3(\boldsymbol{x}; \widehat{\boldsymbol{\alpha}})}\right) &= .401 + .302x_2 - .302x_{11}.
\end{aligned} \tag{27}$$

We also compute the so-called posterior probabilities (15) of each observation belonging to any of the three groups (experts) indicated by the fitted model. Based on these probabilities, approximately, 66.4% of individuals were classified to group 1, 22.2% to group 2, and 11.4% to group 3. Fig. 2 shows the scatter plots of the probabilities versus the weight (response), and Fig. 3 shows the boxplots of the weights of individuals classified to any of the three groups according to the posterior probabilities. The average weights in the three groups are: 64.59, 81.46, and 90.57 kg, respectively, which shows significant differences between the three groups in terms of weights. Also, the percentages of female and male in the three groups are: (62%, 38%), (22%, 78%), and (15%, 85%), respectively. It thus makes sense why the average weight in group 1 is smaller than the other two groups as the majority in this group are female, whereas in the other two groups male are the majority. Since one of the selected covariates affecting the mean weight of the three groups is height (x_{23}), the average height of the three groups are respectively: 168.8, 179.2, and 177.6 cm. We may conclude that, with respect to the weight and height, the individuals in group 1 which are the majority in this data are living a healthy life, whereas those in group 2 may be considered as slightly overweight, and those in group 3 as obese.

We may interpret the selected covariates as follows. From (26), we can see that most of the selected covariates have positive estimated effects on the mean response variable weight, and they are mostly girth measurements. The covariates Waist girth (x_{12}) and Height (x_{23}) with positive estimated effects are selected in all the three mixture components. The two covariates x_1 and x_{16} are selected with negative effects in the third component which could be due to an artefact of their high correlation with the other selected covariates in the model. From (27), the only two covariates selected in the gating network $\{g_1, g_2, g_3\}$ are Biiliac diameter or pelvic breadth (x_2) and Chest girth (x_{11}). We can see that the larger the values of either (x_2, x_{11}), the less likely that the corresponding individual belongs to group 1, which is referred to as the group with a healthy life style. More specifically, individuals with larger values of x_2 are more likely in group 2, and those with larger values of x_{11} are more likely in group 3 (obese) which makes sense as x_{11} shows the chest size.

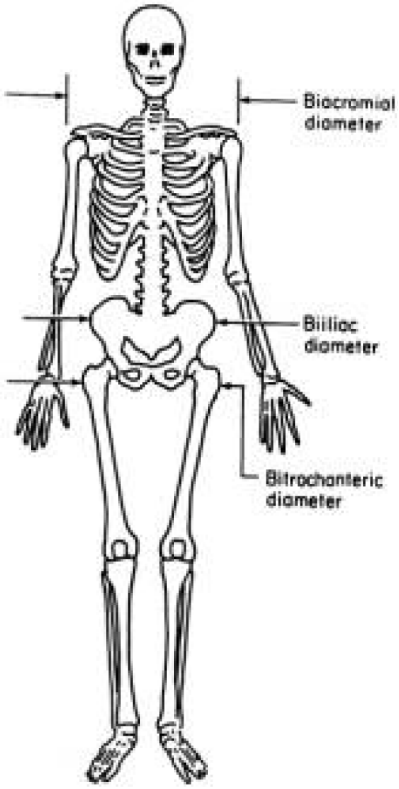

Fig. 1. Biacromial, Biiliac, and Bitrochanteric diameters (Heinz et al., 2003).

9. Discussion

We have introduced a penalized likelihood method for parameter estimation and feature selection in MOE models together with a grouped regularization technique for the gating network parameters. The proposed method is particularly useful when the number of features is large. The grouping technique has provided a new perspective on how to obtain a sparse MOE model, along with its improved interpretability. We have established consistency of the methods in both estimation and feature selection. Numerically, we have employed a modified EM algorithm combined with the proximal gradient method and LLA (Nesterov, 2004), which results in a convenient closed-form parameter update in the M-step of the algorithm.

As much as development has been made in this paper, there remains more work to be done in the study of sparse MOEs. Our current theory studies conditions under which the proposed estimators are statistically optimal, i.e. consistency in both estimation and feature selection. On the other hand, theoretical guarantees for converges of the EM algorithm is also an important research question. Several seminal works have provided theoretical insights into the convergence of the EM algorithm. For example Xu and Jordan (1996) studies convergence of the EM for Gaussian mixture models. Yi and Caramanis (2015) analyze the convergence and consistency properties of a regularized EM algorithm toward understanding regularization techniques. Balakrishnan et al. (2017) develops a theoretical framework for quantifying when and how fast EM-type iterates converge within a small neighborhood of a global optimum of the population likelihood. Zhao et al. (2020) studies the convergence behavior of the EM algorithm in Gaussian mixture models with an arbitrary number of mixture components and mixing weights. In our simulation study, we did not encounter convergence issues of the proposed EM algorithm, and the results show reasonable performance of the algorithm. Nevertheless, theoretical guarantees for converges of the proposed EM algorithm requires new theoretical tools beyond the scope of the current work and is a topic of future research.

Extension of our theoretical results to high-dimensional settings when both p_n and s_n grow to infinity faster than the rates $s_n^4 = o(n)$ and $s_n p_n = o(n)$ considered in Theorems 1–4, and growing values of the regression parameters $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K, \boldsymbol{\alpha})$, as n grows, are subjects of future research. It is also valuable to investigate non-asymptotic error bounds (Städler et al., 2010) as well as minimax rate of convergence of the proposed estimators. In addition, it is interesting to investigate estimation of the number of experts K simultaneously with feature selection. Information criterion such as the BIC is commonly used for estimation of K. Its finite sample performance in our simulation study (Section 7) is satisfactory. Although this method theoretically does not underestimate K (Leroux, 1992), its consistency in estimating K is yet to be studied. Other potential future directions are statistical inference such as hypothesis testing and confidence intervals for post-selection targets in sparse MOEs which is a topic of post-selection inference

(PoSI, Berk et al., 2013; Javanmard and Montanari, 2014; Zhang et al., 2022). These developments will contribute to the study of MOEs and their applications in real data analysis in various fields.

Acknowledgments

The authors would like to thank the editor, associate editor, and two referees for their thoughtful comments and suggestions. A. Khalili and A.Y. Yang are supported by the Natural Science and Engineering Research Council of Canada (NSERC RGPIN-2020-05011) and (NSERC RGPIN-2016-05174) and Fonds de Recherche du Québec-Nature et Technologie (FRQ-NT 327788).

Appendix A. Examples of the penalty function r_n

Common choices of $r_n(\cdot)$ includes the Lasso, AdaLasso, SCAD, and MCP:

$$\begin{aligned}
\text{Lasso} &: r_n(\theta;\lambda) = n\lambda|\theta| \\
\text{AdaLasso} &: r_n(\theta;\lambda) = n\lambda\omega|\theta| \\
\text{SCAD} &: r_n(\theta;\lambda) = \begin{cases} n\lambda|\theta| & ,|\theta| \le \lambda \\ -n(\theta^2 - 2a\lambda|\theta| + \lambda^2)/[2(a-1)] & ,\lambda < |\theta| \le a\lambda \\ n\lambda^2(a+1)/2 & ,|\theta| > a\lambda \end{cases} \\
\text{MCP} &: r_n(\theta;\lambda) = \begin{cases} n\lambda(|\theta| - \frac{|\theta|^2}{2\gamma\lambda}) & ,|\theta| < \gamma\lambda \\ n\lambda^2\gamma/2 & ,|\theta| \ge \gamma\lambda \end{cases}
\end{aligned}$$

for some constants $a > 2$, $\gamma > 0$, and $\lambda \ge 0$ is a tuning parameter that controls how light or heavy the penalty is on θ. Fan and Li (2001) suggested that the value $a = 3.7$ as a good choice in SCAD. The parameter γ in MCP controls the concavity of the penalty, such that when $\gamma \to \infty$ the penalty becomes Lasso, and if $\gamma \to 0^+$ then it becomes the L_0 penalty. In AdaLasso, ω is some pre-specified (possibly random) weights.

Appendix B. Regularity conditions

Let $f(\boldsymbol{v};\boldsymbol{\theta})$ be the joint density of $\boldsymbol{V} = (\boldsymbol{x}, Y)$, with the parameter space $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Note that the conditional density function of Y given $\boldsymbol{x}$ follows the MOE model (1). In the regularity conditions that follow we write $\boldsymbol{\theta} = (\psi_1, \psi_2, \ldots, \psi_{d_n})$, where d_n is the total number of parameters in the model. The expected value E_0 is with respect to the true distribution of $\boldsymbol{V}$ with the corresponding parameter of interest $\boldsymbol{\theta}_0$.

R_1 : The density $f(\boldsymbol{v};\boldsymbol{\theta})$ has common support in $\boldsymbol{v}$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, and $f(\boldsymbol{v};\boldsymbol{\theta})$ is identifiable with respect to $\boldsymbol{\theta}$.

R_2 : There exists an open subset $\boldsymbol{\Theta}^* \subset \boldsymbol{\Theta}$ containing the true parameter $\boldsymbol{\theta}_0$ such that for almost all $\boldsymbol{v}$, $f(\boldsymbol{v};\boldsymbol{\theta})$ admits third partial derivatives with respect to $\boldsymbol{\theta} \in \boldsymbol{\Theta}^*$.

R_3 : For all $j, l = 1, 2, \ldots, d_n$, the first and second derivatives of $f(\boldsymbol{v};\boldsymbol{\theta})$ satisfy:

$$E_0\left\{\frac{\partial}{\partial\psi_j}\log f(\boldsymbol{v};\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right\} = 0;$$

$$E_0\left\{\frac{\partial}{\partial\psi_j}\log f(\boldsymbol{v};\boldsymbol{\theta})\frac{\partial}{\partial\psi_l}\log f(\boldsymbol{v};\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right\} = E_0\left\{-\frac{\partial^2}{\partial\psi_j\partial\psi_l}\log f(\boldsymbol{v};\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right\}.$$

R_4 : The Fisher information matrix is finite and positive definite at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$:

$$\mathcal{I}_n(\boldsymbol{\theta}) = E_0\left\{\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log f(\boldsymbol{v};\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log f(\boldsymbol{v};\boldsymbol{\theta})\right)^{\top}\right\},$$

and it has finite eigenvalues $0 < m < \rho_{min}\{\mathcal{I}_n(\boldsymbol{\theta})\} < \rho_{max}\{\mathcal{I}_n(\boldsymbol{\theta})\} < M < \infty$, for some finite constant m and M. Furthermore, for $j, l = 1, 2, \ldots, d_n$, and for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}^*$ in a neighborhood of $\boldsymbol{\theta}_0$,

$$E_0\left\{\frac{\partial^2}{\partial\psi_j\partial\psi_l}\log f(\boldsymbol{v};\boldsymbol{\theta})\right\}^2 < M_2\ ,\ E_0\left\{\frac{\partial\log f(\boldsymbol{v};\boldsymbol{\theta})}{\partial\psi_j}\frac{\partial\log f(\boldsymbol{v};\boldsymbol{\theta})}{\partial\psi_l}\right\}^2 < M_3$$

for some finite constants M_2 and M_3.

R_5 : There exists integrable functions $\mathcal{B}_j(\boldsymbol{v}), \mathcal{B}_{jl}(\boldsymbol{v})$ and $\mathcal{B}_{jlm}(\boldsymbol{v})$ (possibly depending on $\boldsymbol{\theta}_0$), such that $\int_{-\infty}^{\infty}\mathcal{B}_{jlm}(\boldsymbol{v})f(\boldsymbol{v};\boldsymbol{\theta}_0)d\boldsymbol{v} < \infty$, and for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}^*$ in a neighborhood of $\boldsymbol{\theta}_0$, we have

$$\left|\frac{\partial f(\boldsymbol{v};\boldsymbol{\theta})}{\partial\psi_j}\right| \le \mathcal{B}_j(\boldsymbol{v}),\quad \left|\frac{\partial^2 f(\boldsymbol{v};\boldsymbol{\theta})}{\partial\psi_j\partial\psi_l}\right| \le \mathcal{B}_{jl}(\boldsymbol{v}),\quad \left|\frac{\partial^3\log f(\boldsymbol{v};\boldsymbol{\theta})}{\partial\psi_j\partial\psi_l\partial\psi_m}\right| \le \mathcal{B}_{jlm}(\boldsymbol{v}).$$

Appendix C. Tables and figures

See Tables 1–4 and Figs. 2 and 3.

Table 1
Average empirical mean squared errors.

	d_n	Method	$\widehat{\alpha}_1$	$\widehat{\alpha}_2$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$
		Lasso	2.41	2.16	.090	.477	.514
$n = 200$	83	AdaLasso	1.54	1.37	.117	.563	.817
		SCAD	.497	.466	.036	.099	.194
		Lasso	2.45	2.19	.108	.547	.666
Setting (I)	128	AdaLasso	2.19	1.78	.635	1.68	2.29
		SCAD	.690	.791	.038	.129	.664
		Lasso	2.51	2.25	.160	.753	1.02
	218	AdaLasso	2.20	2.41	.394	2.49	5.30
		SCAD	1.39	1.70	.040	.183	1.99
		Lasso	2.14	1.54	.156	.178	.313
$n = 300$	93	AdaLasso	1.05	.704	.061	.165	.476
		SCAD	.218	.187	.034	.068	.065
		Lasso	2.15	1.53	.223	.185	.351
Setting (II)	158	AdaLasso	1.30	.918	.165	.646	1.83
		SCAD	.382	.262	.058	.070	.545
		Lasso	2.38	1.57	.890	.220	.523
	308	AdaLasso	2.48	2.48	1.62	2.26	2.43
		SCAD	2.13	1.70	.387	.409	1.05
		Lasso	3.04	2.46	.149	.246	.324
$n = 400$	108	AdaLasso	1.78	1.37	.063	.250	.259
		SCAD	.315	.275	.036	.071	.056
		Lasso	3.06	2.46	.216	.270	.351
Setting (III)	188	AdaLasso	1.81	1.41	.228	.524	1.28
		SCAD	.471	.424	.036	.071	.106
		Lasso	3.17	2.48	.597	.331	.448
	358	AdaLasso	2.55	2.73	.544	4.75	3.28
		SCAD	2.49	2.36	.080	.798	4.70

Table 2
Average specificity and sensitivity.

	d_n	Method	$\widehat{\boldsymbol{\alpha}} = (\widehat{\boldsymbol{\alpha}}_1, \widehat{\boldsymbol{\alpha}}_2)$		$\widehat{\boldsymbol{\beta}}_1$		$\widehat{\boldsymbol{\beta}}_2$		$\widehat{\boldsymbol{\beta}}_3$	
			SP	SE	SP	SE	SP	SE	SP	SE
		Lasso	.990	1.00	.978	1.00	.912	1.00	.947	.985
$n = 200$	83	AdaLasso	.986	.988	.952	.998	.896	.985	.919	.948
		SCAD	.985	.973	1.00	1.00	.998	1.00	.993	.988
		Lasso	.993	1.00	.984	1.00	.926	.998	.949	.975
Setting (I)	128	AdaLasso	.981	.970	.938	.988	.900	.947	.929	.815
		SCAD	.985	.935	1.00	1.00	.997	.997	.987	.930
		Lasso	.995	.993	.987	.998	.932	.997	.953	.960
	218	AdaLasso	.986	.935	.914	.994	.922	.912	.950	.571
		SCAD	.990	.788	1.00	1.00	.996	.995	.980	.763
		Lasso	.999	1.00	.999	1.00	.977	1.00	.985	1.00
$n = 300$	93	AdaLasso	.999	1.00	.986	1.00	.955	.998	.938	.995
		SCAD	.999	1.00	1.00	1.00	1.00	1.00	.999	1.00
		Lasso	.999	1.00	.998	1.00	.985	1.00	.985	1.00
Setting (II)	158	AdaLasso	.999	1.00	.957	.998	.924	.988	.932	.920
		SCAD	.999	.980	.999	.997	1.00	1.00	.995	.962
		Lasso	.999	1.00	.999	.967	.985	1.00	.979	.997
	308	AdaLasso	.997	.922	.899	.942	.932	.844	.953	.370
		SCAD	.903	.997	.998	.977	.993	.984	.969	.585
		Lasso	.999	.962	.997	1.00	.988	1.00	.994	1.00
$n = 400$	108	AdaLasso	1.00	.980	.994	1.00	.973	.997	.966	1.00
		SCAD	.999	.982	1.00	1.00	1.00	1.00	1.00	1.00
		Lasso	1.00	.963	.998	1.00	.992	1.00	.995	1.00
Setting (III)	188	AdaLasso	1.00	.987	.980	1.00	.942	.997	.960	.986
		SCAD	1.00	.942	1.00	1.00	1.00	1.00	1.00	.998
		Lasso	1.00	.943	1.00	.990	.992	1.00	.993	.998
	358	AdaLasso	.999	.908	.952	.994	.929	.911	.968	.497
		SCAD	.962	.977	1.00	.999	.998	.972	.980	.699

Table 3
Order selection based on BIC: $K = 3$ is the correct mixture order.

	d_n	Method	$\widehat{K}$				
			1	2	**3**	4	5
		Lasso	.000	.445	.515	.035	.005
$n = 200$	83	AdaLasso	.005	.120	.675	.155	.045
		SCAD	.000	.115	.755	.120	.010
		Lasso	.000	.435	.520	.035	.010
Setting (I)	128	AdaLasso	.010	.020	.695	.200	.075
		SCAD	.005	.115	.750	.130	.000
		Lasso	.005	.270	.600	.125	.000
	218	AdaLasso	.115	.095	.560	.170	.060
		SCAD	.040	.055	.765	.140	.000
		Lasso	.000	.000	.950	.040	.010
$n = 300$	93	AdaLasso	.000	.010	.855	.110	.025
		SCAD	.000	.000	.940	.040	.020
		Lasso	.000	.035	.710	.100	.155
Setting (II)	158	AdaLasso	.000	.055	.700	.100	.145
		SCAD	.005	.165	.480	.195	.155
		Lasso	.000	.050	.720	.185	.045
	308	AdaLasso	.090	.260	.200	.225	.225
		SCAD	.060	.475	.245	.190	.030
		Lasso	.000	.020	.960	.015	.005
$n = 400$	108	AdaLasso	.000	.005	.900	.085	.010
		SCAD	.000	.000	.975	.010	.015
		Lasso	.000	.150	.780	.025	.045
Setting (III)	188	AdaLasso	.000	.100	.650	.180	.070
		SCAD	.000	.075	.660	.180	.085
		Lasso	.000	.220	.455	.285	.040
	358	AdaLasso	.015	.070	.330	.325	.260
		SCAD	.005	.145	.300	.390	.160

Table 4
List of the variables in the real data example (Heinz et al., 2003).

Covariates	Description
	Skeletal measurements:
x_1	Biacromial diameter (see Fig. 1)
x_2	Biiliac diameter, or "pelvic breadth" (see Fig. 1)
x_3	Bitrochanteric diameter (see Fig. 1)
x_4	Chest depth between spine and sternum at nipple level, mid-expiration
x_5	Chest diameter at nipple level, mid-expiration
x_6	Elbow diameter, sum of two elbows
x_7	Wrist diameter, sum of two wrists
x_8	Knee diameter, sum of two knees
x_9	Ankle diameter, sum of two ankles
	Girth measurements:
x_{10}	Shoulder girth over deltoid muscles
x_{11}	Chest girth, nipple line in males and just above breast tissue in females, mid-expiration
x_{12}	Waist girth, narrowest part of torso below the rib cage, average of contracted and relaxed position
x_{13}	Navel (or "Abdominal") girth at umbilicus and iliac crest, iliac crest as a landmark
x_{14}	Hip girth at level of bitrochanteric diameter
x_{15}	Thigh girth below gluteal fold, average of right and left girths
x_{16}	Bicep girth, flexed, average of right and left girths
x_{17}	Forearm girth, extended, palm up, average of right and left girths
x_{18}	Knee girth over patella, slightly flexed position, average of right and left girths
x_{19}	Calf maximum girth, average of right and left girths
x_{20}	Ankle minimum girth, average of right and left girths
x_{21}	Wrist minimum girth, average of right and left girths
	Other measurements:
x_{22}	Age (years)
x_{23}	Height (cm)
x_{24}	Sex (male = 1, female = 0)
y	Weight (kg)

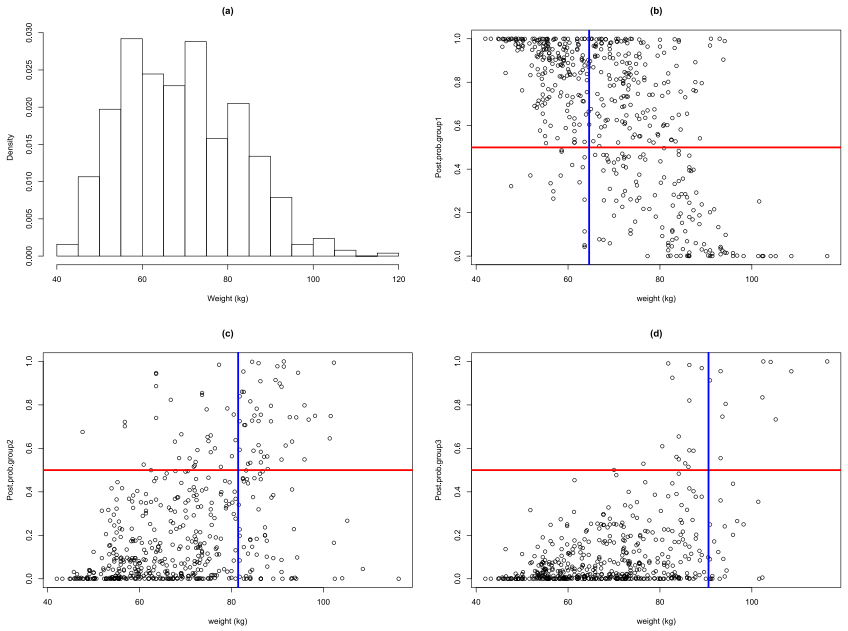

Fig. 2. (a) Histogram of the weight; (b)–(d) Posterior probabilities of observations belonging to each of the three groups represented by the fitted MOE model. The blue vertical line indicates the average weight within each group. The red line shows probability value 0.5. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

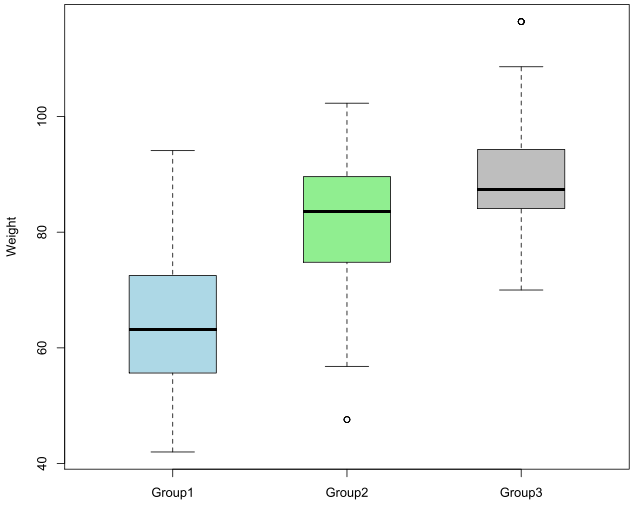

Fig. 3. Boxplots of the weights of the three identified groups.

Appendix D. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jspi.2024.106250.

References

Balakrishnan, S., Wainwright, M.J., Yu, B., 2017. Statistical guarantees for the EM algorithm: From population to sample-based analysis. Ann. Statist. 45 (1), 77–120. http://dx.doi.org/10.1214/16-AOS1435.

Berk, R., Brown, L., Buja, A., Zhang, K., Zhao, L., 2013. Valid post-selection inference. Ann. Statist. 41 (2), 802–837.

Boyd, S., Vandenberghe, L., 2004. Convex Optimization. Cambridge University Press.

Breheny, P., Huang, J., 2015. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. Statist. Comput. 25, 173–187.

Chamroukhi, F., Huynh, B.T., 2019. Regularized maximum likelihood estimation and feature selection in mixtures-of-experts models. J. Soc. Francaise Stat. 160, 57–85.
Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B Stat. Methodol. 39, 1–38.
Donoho, D.L., Johnstone, J.M., 1994. Ideal spatial adaptation by wavelet shrinkage. Biometrika 81, 425–455.
Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96, 1348–1360.
Fan, J., Li, R., Zhang, C.-H., Zou, H., 2020. Statistical Foundations of Data Science. CRC Press.
Fan, J., Peng, H., 2004. Nonconcave penalized likelihood with a diverging number of parameters. Ann. Statist. 32, 928–961.
Fang, F., Jiwei, Z., S. Ejaz, A., Qu, A., 2021. A weak-signal-assisted procedure for variable selection and statistical inference with an informative subsample. Biometrics 77, 996–1010.
Friedman, J., Hastie, T., Tibshirani, 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33, 1–22.
Guo, J., Levina, E., Michailidis, G., Zhu, J., 2010. Pairwise variable selection for high-dimensional model-based clustering. Biometrics 66, 793–804.
Hastie, T., Tibshirani, R., Wainwright, M.J., 2019. Statistical Learning with Sparsity: The Lasso and Generalizations. Chapman and Hall/CRC.
Heinz, G., Peterson, L.J., Johnson, R.W., Kerk, C.J., 2003. Exploring relationships in body dimensions. J. Stat. Educ. 11.
Heiser, W.J., 1995. Convergent computation by iterative majorization. In: Recent advances in descriptive multivariate analysis. pp. 157–189.
Hennig, C., 2000. Identifiability of models for clusterwise linear regression. J. Classification 17, 273–296.
Huber, P.J., 1973. Robust regression: Asymptotics, conjectures and Monte Carlo. Ann. Statist. 1, 799–821.
Jacobs, R., Jordan, M., Nowlan, S., Hinton, G., 1991. Adaptive mixtures of local experts. Neural Comput. 3 (79–87).
Javanmard, A., Montanari, A., 2014. Confidence intervals and hypothesis testing for high-dimensional regression. J. Mach. Learn. Res. 15 (82), 2869–2909.
Jiang, W., Tanner, M., 1999a. Hierarchical mixtures-of-experts for exponential family regression models: Approximation and maximum likelihood estimation. Ann. Statist. 27, 987–1011.
Jiang, W., Tanner, M., 1999b. On the identifiability of mixtures-of-experts. Neural Nerw. 12, 1253–1258.
Khalili, A., 2010. New estimation and feature selection methods in mixture-of-experts models. Canad. J. Statist. 38, 519–539.
Khalili, A., Chen, J., 2007. Variable selection in finite mixture of regression models. J. Amer. Statist. Assoc. 102, 1025–1038.
Khalili, A., Lin, S., 2013. Regularization in finite mixture of regression models with diverging number of parameters. Biometrics 69, 436–446.
Khalili, A., Vidyashankar, A.N., 2018. Hypothesis testing in finite mixture of regressions: Sparsity and model selection uncertainty. Canad. J. Statist. 46 (3), 429–457.
Konishi, S., Kitagawa, G., 2008. Information Criteria and Statistical Modeling. Springer New York.
Lange, K., Hunter, D.R., Yang, I., 2000. Optimization transfer using surrogate objective functions. J. Comput. Graph. Stat. 9 (1), 1–20.
Leroux, B.G., 1992. Consistent estimation of a mixing distribution. Ann. Statist. 20, 1350–1360.
McLachlan, G., Peel, D., 2000. Finite Mixture Models. John Wiley & Sons.
Nesterov, Y., 2004. Introductory Lectures on Convex Optimization. Kluwer Academic Publishers, New York.
Nguyen, H.D., Chamroukhi, F., 2018. Practical and Theoretical Aspects of Mixture-Of-Experts Modeling: An Overview. Wiley Periodicals, Inc..
Nguyen, T., Nguyen, H., Chamroukhi, F., McLachlan, G.J., 2020. An l1-oracle inequality for the lasso in mixture-of-experts regression models. https://arxiv.org/abs/2009.10622.
Rish, I., Grabarnik, G., 2014. Sparse Modelling: Theory, Algorithms, and Applications. CRC Press.
Roy, S., Tewari, A., Zhu, Z., 2023. High-dimensional variable selection with heterogeneous signals: A precise asymptotic perspective. https://arxiv.org/pdf/2201.01508.pdf.
Simon, N., Friedman, J., Hastie, T., Tibshirani, R., 2013. A sparse-group lasso. J. Comput. Graph. Statist. 22, 231–245.
Städler, N., Bühlmann, P., Van De Geer, S., 2010. l_1-Penalization for mixture regression models. Test 19 (209–256).
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 58, 267–288.
Wainwright, M.J., 2019. High-Dimensional Statistics: A Non-Asymptotic Viewpoint. Cambridge University Press.
Wang, H., Li, R., Tsai, C.-L., 2007. Tuning parameter selectors for the smoothly clipped absolute deviation method. Biometrika 94, 553–568.
Xu, L., Jordan, M.I., 1996. On convergence properties of the EM algorithm for Gaussian mixtures. Neural Comput. 8 (1), 129–151.
Yi, X., Caramanis, C., 2015. Regularized EM algorithms: A unified framework and statistical guarantees. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. Vol. 28, Curran Associates, Inc..
Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Ser. B Stat. Methodol. 68 (1), 49–67.
Yuksel, S., Wilson, J., Gader, P., 2012. Twenty years of mixture of experts. IEEE Trans. Neural Netw. Learn. Syst. 23, 1177–1193.
Zhang, D., Khalili, A., Asgharian, M., 2022. Post-model-selection inference in linear regression models: An integrated review. Stat. Surv. 16, 86–136.
Zhao, R., Li, Y., Sun, Y., 2020. Statistical convergence of the EM algorithm on Gaussian mixture models. Electron. J. Stat. 14 (1), 632–660. http://dx.doi.org/10.1214/19-EJS1660.
Zou, H., 2006. The adaptive lasso and its oracle properties. J. Amer. Statist. Assoc. 101, 1418–1429.
Zou, H., Li, R., 2008. One-step sparse estimates in nonconcave penalized likelihood models. Ann. Statist. 36, 1509—1533.