Journal of the Royal Statistical Society Series B: Statistical Methodology, 2024, **00**, 1–22
https://doi.org/10.1093/jrsssb/qkae117

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series B
Statistical Methodology
B

Original Article

Prediction sets for high-dimensional mixture of experts models

Adel Javanmard[1], Simeng Shao[2] and Jacob Bien[1]

[1]Data Sciences and Operations Department, Marshall School of Business, University of Southern California, 3670 Trousdale Parkway, Los Angeles, CA 90089, USA
[2]Amazon, Seattle, WA, USA

Address for correspondence: Adel Javanmard, Data Sciences and Operations Department, Marshall School of Business, University of Southern California, 3670 Trousdale Parkway, Los Angeles, CA 90089, USA. Email: ajavanma@usc.edu

Abstract

Large datasets make it possible to build predictive models that can capture heterogenous relationships between the response variable and features. The mixture of high-dimensional linear experts model posits that observations come from a mixture of high-dimensional linear regression models, where the mixture weights are themselves feature-dependent. In this article, we show how to construct valid prediction sets for an ℓ_1-penalized mixture of experts model in the high-dimensional setting. We make use of a debiasing procedure to account for the bias induced by the penalization and propose a novel strategy for combining intervals to form a prediction set with coverage guarantees in the mixture setting. Synthetic examples and an application to the prediction of critical temperatures of superconducting materials show our method to have reliable practical performance.

Keywords: high-dimensional statistics, mixture of experts models, prediction set, expectation-maximization, debiasing

1 Introduction

In traditional statistics, we imagine a universal relationship between variables that holds across an entire population; observations not following this relationship are dismissed as outliers. However, we know that reality is more complex, with numerous subpopulations likely exhibiting distinct behaviours. As datasets grow in size, we become better able to detect and properly model this heterogeneity. The mixture of regressions model (Quandt & Ramsey, 1978) is an important tool for extending linear regression to this heterogeneity-aware setting. For a random response $y \in \mathbb{R}$ and a random vector of predictors $\boldsymbol{x} \in \mathbb{R}^p$, we imagine a latent subgroup membership $z \in \{1, \ldots, K\}$ that determines the conditional distribution of y given $\boldsymbol{x}$:

$$y \,|\, \boldsymbol{x}, z = k \sim N(\boldsymbol{x}^T \boldsymbol{\beta}_k, \sigma_k^2). \tag{1}$$

Making predictions with this model requires estimating for each subgroup a coefficient vector $\boldsymbol{\beta}_k$, an error variance σ_k^2, and a group membership probability $\pi_k = \mathbb{P}(z = k)$. Our focus in this work is on the mixture of experts model (MoE, Jordan & Jacobs, 1994), which is even more flexible in that it allows these group membership probabilities to depend on the predictors as well:

$$z \,|\, \boldsymbol{x} \sim \text{Multinomial}[\pi(\boldsymbol{x})] \quad \text{with} \quad \pi_k(\boldsymbol{x}) = \frac{\exp(\boldsymbol{x}^T \boldsymbol{\alpha}_k)}{\sum_{\ell=1}^K \exp(\boldsymbol{x}^T \boldsymbol{\alpha}_\ell)}. \tag{2}$$

Received: October 31, 2022. **Revised:** November 8, 2024. **Accepted:** November 15, 2024

The MoE is a widely used model throughout many fields. Yuksel et al. (2012) provide an extensive survey of the model and its diverse applications, citing over thirty applied papers in fields including climate sciences, biomedicine, robotics, astronomy, and finance (to name just a few). These authors note that the MoE model is closely connected to many of the most influential algorithms in machine learning. The popularity of the MoE model within machine learning has persisted to the present, which is evident by recent work in using MoEs in deep learning and to make large language models more efficient (Fournier et al., 2023). In such machine learning settings, it is common to think of the MoE as an ensemble method in which an adaptively weighted average of different 'expert' predictions forms the main prediction. These experts provide estimates of $\mathbb{E}[y \mid \boldsymbol{x}, z = k]$, and often neural networks are used in place of the linear regression model given in (1). Gormley and Frühwirth-Schnatter (2019) note the diversity of perspectives in how MoE models are conceived of. A more statistical viewpoint, which aligns with our focus here, uses MoE as a flexible framework for mixture modelling in which covariates control the mixture weights (Gormley et al., 2023).

Our choice to focus on experts that are high-dimensional linear regression models stems from two considerations. First, the ever increasing complexity of data sets have driven us toward the data regime where one has far more attributes (e.g. genetic and laboratory measurements, detailed user history) measured than observations (e.g. patients, users). The high-dimensional regression model is foundational to the field of statistics and has attracted a significant interest and tremendous research effort as it allows fitting high-dimensional parametric models from order of magnitude smaller sample size, by using regularization terms which promote potential natural structure in the model (e.g. sparsity or low-rankness). Second, a recent application of this model to an oceanographic application sparked our interest in the inferential questions we study here. Hyun et al. (2023) develop a high-dimensional mixture of experts approach to modelling phytoplankton subpopulations as a function of environmental covariates in the ocean. The (log) diameter of the phytoplankton cells within each specific subpopulation are taken to be Gaussian with mean depending on environmental covariates (expressed through nonzero values of $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$); however, the prevalence of the different subpopulations also depends on these covariates (expressed through nonzero values of $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K$). Their goal is to predict an entire mixture of Gaussians model based on a covariate vector $\boldsymbol{x}$.

When making predictions, it is valuable to be able to quantify one's level of uncertainty. In this article, we develop the machinery necessary to do so in the context of the high-dimensional mixture of the experts model described above. In particular, given a sample of n observations from the model in (1)–(2), a confidence level $q \in (0, 1)$, and a new predictor vector of interest $\boldsymbol{x}_{\text{new}}$, we show how to form a properly calibrated *prediction set* $\Omega_q(\boldsymbol{x}_{\text{new}})$. That is, given a new draw $(\boldsymbol{x}_{\text{new}}, y_{\text{new}})$ from (1)–(2), we have that

$$\mathbb{P}(y_{\text{new}} \in \Omega_q(\boldsymbol{x}_{\text{new}}) \mid \boldsymbol{x}_{\text{new}}) \geq 1 - q. \tag{3}$$

In words, a prediction set is a range or set of values within which future responses are expected to fall, with a specified level of confidence $(1 - q)$. Of course, shorter prediction sets are preferred. As an extreme case, $(-\infty, \infty)$ is a prediction set with 100% coverage but of course is uninformative. As an illustration, consider the toy example shown in Figure 1. There are $K = 2$ subpopulations and the covariates are given by $\boldsymbol{x} = (1, t, t^2)^T$, with $t \in \mathbb{R}$ (so $p = 3$). In one subpopulation, the response y follows a quadratic relation in t and in the other it is constant. Note that in both cases, the response can be written as a linear function of $\boldsymbol{x}$ with different coefficients (for details of the construction, see Section 4). This parametrization allows us to visualize the response as a function of t. The upper panel shows that the mixture weight on the quadratic subpopulation decreases with increasing t. This can also be seen by inspecting the scatterplot and noting that for large t, most of the points are in the constant subpopulation. The solid lines show our estimated mixture of experts model fit based on the points in the scatterplot. Now suppose we are about to observe a new point y_{new} at, say, $t_{\text{new}} = 0.5$. Can we form a set $\Omega_{0.95}(t_{\text{new}})$ that is guaranteed to capture y_{new} at least 95% of the time? Given that we do not know which subpopulation y_{new} will be drawn from, $\Omega_{0.95}(t_{\text{new}})$ will be a union of two intervals. The size and location of these intervals will depend on our estimates of the subpopulation means and variances (governed by t_{new} and our estimates of $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1$, and σ_2) as well as the estimated mixture weights (governed by t_{new} and our estimates

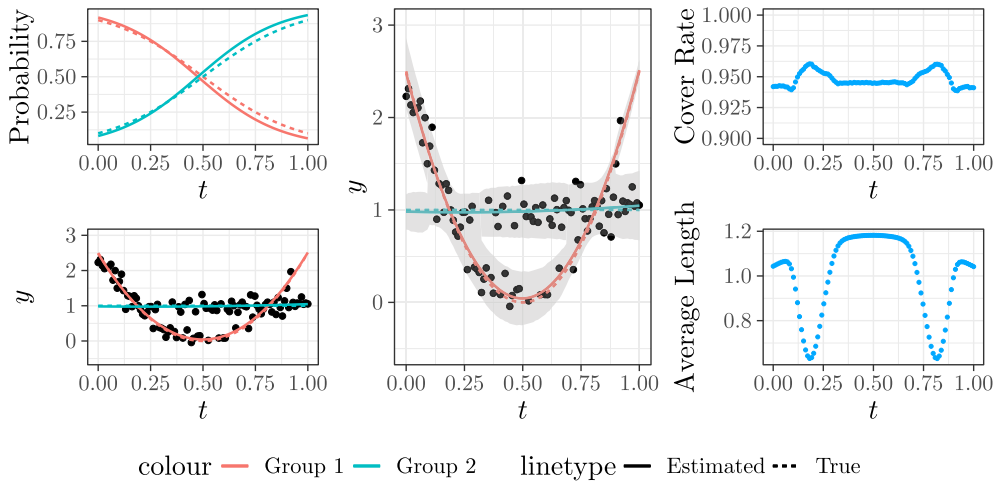

Figure 1. (Top left) True (dashed) and estimated (solid) class probabilities $\pi_k(t)$ and $\hat{\pi}_k(t)$. (Bottom left) Data set of $n = 100$ points $\{(t, y_t)\}$ (black dots) with true (dashed) and estimated (solid) mean functions $\mu_k(t)$ and $\hat{\mu}_k(t)$. (Middle) Prediction set at 95% confidence level $\Omega_{0.05}(t)$ for $t \in [0, 1]$ (shaded area). (Right) Monte Carlo estimate of coverage and length of prediction sets $\Omega_{0.05}(t)$ across 500 training sets and 1,000 y_{new} per training set.

of $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$). The grey band in the middle panel shows our constructed $\Omega_{0.95}(t_{\text{new}})$ as we vary t_{new} from 0 to 1. When the means of the populations are far apart, the prediction set is a union of two intervals, while the set becomes a single interval when the means are close to each other. The larger a subpopulation's mixture weight, the wider that interval becomes. This 'strategy' is reasonable because if it's unlikely that a point will fall in a certain subpopulation, one can get away with less good coverage of that subpopulation (and thereby reduce the overall size of the prediction set). The rightmost panel shows the result of a simulation in which we generated 500 training sets, each time constructing $\Omega_{0.95}(t)$ as a function of t; then, we generated 1,000 y_{new} at each t and computed the coverage rate (averaging over the $500 \cdot 1{,}000$ repetitions for each t). This verifies that our procedure approximately attains the nominal 95% coverage. The bottom panel shows the average size of the prediction set. Quite intuitively, the prediction set is smallest when the two subpopulations are very close to each other. One observes a bit of overcoverage when the subpopulations are close, which makes sense since in this situation the interval from one subpopulation can sometimes cover a point from the opposite subpopulation.

Constructing prediction sets in the context of a high-dimensional mixture of experts model is challenging. In fact, making any sort of precise statements about even the most simple mixture of regression models is nontrivial. For example, much effort has gone into understanding the convergence and estimation error of the expectation–maximization (EM) estimator (Dempster et al., 1977) used in fitting such models (Balakrishnan et al., 2017; Klusowski et al., 2019; Kwon & Caramanis, 2020; Kwon et al., 2019; Yi et al., 2014) as well as being able to test whether there are two groups versus one (H.-T. Zhu & Zhang, 2004). Adding high dimensionality to the study of mixture of regression models brings additional challenges. Städler et al. (2010) and Yi and Caramanis (2015) proposed different ℓ_1-regularized maximum likelihood estimators with accompanying estimation error results. Wang et al. (2015) take this a step further and develop a truncation-based high-dimensional estimator with both estimation error results and the ability to construct confidence intervals for low-dimensional components of the parameter vector. While their results hold for general latent variable models, their application to mixture of regression models is more of a proof of concept, with $K = 2$, $\sigma_1 = \sigma_2$ assumed known, $\pi_1 = \pi_2 = 0.5$, and $\boldsymbol{\beta}_1 = -\boldsymbol{\beta}_2$. Zhang et al. (2020) provide inference for individual coefficients and differences of the form $\boldsymbol{\beta}_{1j} - \boldsymbol{\beta}_{2j}$ within the context of this model using an ℓ_1 penalty. They generalize to the case of unknown mixture weight π_1, $\boldsymbol{\beta}_1 \neq -\boldsymbol{\beta}_2$, and an unknown covariance matrix for X (which they take to be multivariate normal), but they still assume $K = 2$ and that the value $\sigma_1 = \sigma_2$ is known. While we are able to adopt in part a similar debiased approach, we will highlight later

why their technique, which works in the $K = 2$ mixture of regression setting (which involves a single unknown mixture parameter π_1) does not easily generalize to our setting of a mixture of experts model, in which mixture weights depend on the unknown parameter vectors $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K \in \mathbb{R}^p$.

Furthermore, our interest in forming a prediction set (3) requires the ability to make inferential statements about $\boldsymbol{x}_{\text{new}}^T \boldsymbol{\beta}_k$, not just low-dimensional components of $\boldsymbol{\beta}_k$. In this sense, T. Cai et al. (2021) pursue a similar goal in performing inference for individualized treatment effects $\boldsymbol{x}_{\text{new}}^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)$; however, unlike the mixture of regression setting, the group memberships are known in their context.

To summarize, to the best of our knowledge, predictive inference in mixture of expert models has not been addressed in the literature. Furthermore, we address this problem for general K and in the high-dimensional setting. We make use of ideas from the debiased lasso literature. The debiasing approach for constructing confidence intervals for coefficients has been widely used in linear regression models in high-dimensional settings (Javanmard & Montanari, 2014a, 2014b, 2018; Van de Geer et al., 2014; Zhang & Zhang, 2014). In recent years, there has been work on inference for general linear functions (T. T. Cai & Guo, 2017; Guo et al., 2021; Javanmard & Lee, 2020; Y. Zhu & Bradic, 2018). In terms of debiasing in the non-mixture setting, T. T. Cai and Guo (2017), and Athey et al. (2018) proposed bias-corrected estimators for a single linear regression model while, as we have noted above, T. Cai et al. (2021) considers inference for $\boldsymbol{x}_{\text{new}}^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)$ in the case of two observed (i.e. non-latent) groups.

Interest in predictive inference has led to an active area of work on conformal prediction. These approaches are attractive for being distribution-free and providing finite-sample coverage (see, e.g. Lei et al., 2018; Papadopoulos et al., 2002; Romano et al., 2019; Vovk et al., 2005). They rely on very general ideas such as the exchangeability of draws from the distribution. However, the coverage that these conformal methods attain is not conditional on $\boldsymbol{x}_{\text{new}}$ as in (3) but rather holds marginally over $\boldsymbol{x}_{\text{new}}$:

$$\mathbb{P}(y_{\text{new}} \in \Omega_q(\boldsymbol{x}_{\text{new}})) \geq 1 - q.$$

Indeed, it has been proven that to obtain finite-length sets with conditional coverage, one needs to make stronger assumptions (Lei & Wasserman, 2014; Vovk, 2012); for asymptotic conditional coverage, one can attain finite-length intervals, but these still require certain smoothness assumptions (see, e.g. Assumptions 1 and 2 in Lei & Wasserman, 2014). In certain applications, one specifically desires coverage of the form (3) and assuming the parametric form (1)–(2) can be a small price to pay. For example, in the oceanographic example of mixture of experts (Hyun et al. 2023), the mixture of Gaussians structure is visually well-supported. Prediction sets with conditional coverage are desirable because we would like to be able to say that for a given set of environmental conditions (e.g. at a specific temperature and salinity level) our prediction set for the phytoplankton diameters will have a 95% coverage guarantee. Marginal coverage would mean that if we make predictions over many randomly sampled environments, our coverage would average out to 95%. The latter means, for example, that a procedure could be overconfident (i.e. undercovering) at high temperatures and underconfident (i.e. overcovering) at low temperatures. Setting conditional coverage (3) as the goal guards against this undesirable property.

The rest of the article is organized as follows. In Section 2, we describe our approach to constructing prediction sets. This involves estimating parameters using a penalized EM approach (Section 2.1), then using a debiasing technique on the coefficient vectors from each of the component distributions (Section 2.2), and finally combining K intervals into a prediction set in a fashion that maintains proper coverage (Section 2.3). In Section 3, we provide theoretical guarantees that establish the asymptotic validity of our constructed prediction sets and provide insight into the conditions under which we expect nominal coverage to hold. In Section 4 we investigate the empirical performance of our prediction sets in a variety of settings. Section 5 shows our sets and evaluates their performance empirically in predicting the critical temperatures of superconducting materials. Finally, in Section 6, we highlight the differences between mixture of experts models linear mixed-effects models (LMMs). We also discuss how our proposed methodology could be extended to models involving continuous random effects. Before proceeding, we introduce some notation.

Notation. Throughout the article, we use $[p]$ for the set of integers $1, \ldots, p$. We use $\boldsymbol{e}_i$ to denote the ith standard basis vector. For a vector $\boldsymbol{x} \in \mathbb{R}^p$, we denote $\|\boldsymbol{x}\|_q = (\sum_{j=1}^p |x_j|^q)^{1/q}$ for $q > 0$ and $\|\boldsymbol{x}\|_0 = |\text{supp}(\boldsymbol{x})|$, $\|\boldsymbol{x}\|_\infty = \max_{j\in[p]} |x_j|$. For a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, we use $|\boldsymbol{A}|_\infty = \max_{1\le i,j\le n} |A_{i,j}|$. For sequences a_n and b_n, we use the notation $a_n \asymp b_n$ to indicate that a_n is bounded both above and below by b_n asymptotically, i,e, for some constants C_0, C_1 and for all $n \ge n_{n_0}$ we have $C_0 \le |a_n/b_n| \le C_1$. In addition, we write $a_n = O_p(b_n)$ if for any $\varepsilon > 0$, there exists $C_\varepsilon > 0$ and large enough n_ε such that $\mathbb{P}(|a_n/b_n| > C_\varepsilon) < \varepsilon$, for all $n \ge n_\varepsilon$. we write $a_n = o_p(b_n)$ if a_n/b_n converges to zero in probability, i.e. $\lim_{n\to\infty} \mathbb{P}(|a_n/b_n| \ge \varepsilon) = 0$, $\forall\ \varepsilon > 0$. The notation $\overset{d}{\Rightarrow}$ indicates convergence in distribution.

We use the notation $\|\cdot\|_{\psi_2}$, $\|\cdot\|_{\psi_1}$ to refer to the sub-Gaussian and sub-exponential norms respectively. Specifically, for a random variable X, we let

$$\|X\|_{\psi_1} = \sup_{q\ge 1} q^{-1}(\mathbb{E}|X|^q)^{1/q}, \quad \|X\|_{\psi_2} = \sup_{q\ge 1} q^{-1/2}(\mathbb{E}|X|^q)^{1/q}.$$

For a random vector $\boldsymbol{x}$, its sub-Gaussian and sub-exponential norms are defined as

$$\|\boldsymbol{x}\|_{\psi_1} = \sup_{\|\boldsymbol{u}\|_2\le 1} \|\langle \boldsymbol{x}, \boldsymbol{u}\rangle\|_{\psi_2}, \quad \|\boldsymbol{x}\|_{\psi_1} = \sup_{\|\boldsymbol{u}\|_2\le 1} \|\langle \boldsymbol{x}, \boldsymbol{u}\rangle\|_{\psi_1}.$$

2 Methodology

Before delving into our methodology, we provide an overview and motivation for our general strategy. We begin (in Section 2.1) with a review of a penalized maximum-likelihood procedure for the MoE model (1)–(2) such as is used in Hyun et al. (2023). Our use of penalization allows for good estimation performance in the high-dimensional setting we consider. However, such a penalty introduces bias in the estimated model, which makes it difficult to construct prediction intervals. Therefore, in Section 2.2 we employ a debiasing strategy. Finally, in Section 2.3, we introduce a novel methodology for forming prediction sets in this context.

Our setting provides a flexible framework to do predictive inference in a model that is of larger dimension than the sample size. It is worth noting that non-parametric density estimation methods would fail in this high-dimensional regime as they suffer from the curse of dimensionality.

2.1 Penalized EM-based estimator

The EM algorithm (Dempster et al., 1977) is a common heuristic when faced with maximum-likelihood problems involving missing data, especially in the form of latent variables. The algorithm operates in an iterative fashion, alternating between the E (expectation) step and M (maximization) step, while managing to increase the objective function.

Assume n data points $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ are drawn independently from the MoE model (1)–(2). The EM algorithm aims at maximizing the log-likelihood

$$\ell(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \log\left[\sum_{k=1}^K \pi_k(\boldsymbol{x}_i)\cdot\phi_k(\boldsymbol{x}_i, y_i)\right], \tag{4}$$

where $\boldsymbol{\theta} = (\{\boldsymbol{\beta}_k, \boldsymbol{\alpha}_k, \sigma_k\}_{k\in[K]}) \in \mathbb{R}^{(2p+1)K}$ represents the model parameters, $\phi_k(\boldsymbol{x}_i, y_i)$ is given by

$$\phi_k(\boldsymbol{x}_i, y_i) = \frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_k)^2}{2\sigma_k^2}\right), \tag{5}$$

and $\pi_k(\boldsymbol{x}_i)$ is given in (2). The log-likelihood $\ell(\boldsymbol{\theta})$ is not concave even as a function of $\boldsymbol{\beta}_k$, treating $\boldsymbol{\alpha}_k$ and σ_k as fixed. This is due to the marginalizing over the latent cluster memberships z_i. The EM

algorithm instead employs a minorize–maximize approach (Hunter & Lange, 2004) in which a minorizer to the log-likelihood is repeatedly constructed and maximized. More specifically, given some fixed $\widehat{\boldsymbol{\theta}}$ it maximizes a lower bound function $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})$ over $\boldsymbol{\theta}$ to make $\ell(\boldsymbol{\theta}) - \ell(\widehat{\boldsymbol{\theta}})$ large. We refer to Hastie et al. (2009) for a more detailed introduction to EM algorithm and derivation of the function $Q(\boldsymbol{\theta}\,|\,\widehat{\boldsymbol{\theta}})$ and here only provide the description of the EM algorithm for the MoE model.

Let $\gamma_{i,k}(\boldsymbol{\theta})$ be the probability that $z_i = k$ conditioned on the observed variable $(\boldsymbol{x}_i, y_i)$, i.e.

$$\gamma_{i,k}(\boldsymbol{\theta}) : = \mathbb{P}(z_i = k \,|\, \boldsymbol{x}_i, y_i) = \frac{\pi_k(\boldsymbol{x}_i) \cdot \phi_k(\boldsymbol{x}_i, y_i)}{\sum_{\ell=1}^{K} \pi_\ell(\boldsymbol{x}_i) \cdot \phi_\ell(\boldsymbol{x}_i, y_i)}.$$

The function $\gamma_{i,k}(\boldsymbol{\theta})$ is sometimes referred to as *responsibilities* in the literature (see, e.g. Hyun et al. 2023) because it quantifies how 'responsible' group k is for point i.

Consider the function $Q(\boldsymbol{\theta}\,|\,\widehat{\boldsymbol{\theta}})$ defined as

$$Q(\boldsymbol{\theta}\,|\,\widehat{\boldsymbol{\theta}}) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\gamma_{i,k}(\widehat{\boldsymbol{\theta}})\big[\log \pi_k(\boldsymbol{x}_i) + \log \phi_k(\boldsymbol{x}_i, y_i)\big] + \sum_{k=1}^{K}\lambda_{\boldsymbol{\alpha}}\|\boldsymbol{\alpha}_k\|_1 + \sum_{k=1}^{K}\lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}_k\|_1, \tag{6}$$

where the regularization terms $\|\boldsymbol{\beta}_k\|_1$, $\|\boldsymbol{\alpha}_k\|_1$ are added to enforce sparsity on the estimated parameters and allow for applications in the high-dimensional sparse regime.

The EM algorithm iterates between estimating the conditional membership probabilities $\gamma_{i,k}(\widehat{\boldsymbol{\theta}})$, and reducing (6) by updating the parameters $\widehat{\boldsymbol{\theta}}$. The details of the updates are given below:

- E step: Estimating the conditional responsibility of membership based on the latest parameter estimate.

$$\gamma_{i,k}(\widehat{\boldsymbol{\theta}}) = \frac{\frac{\hat{\pi}_{i,k}}{\widehat{\sigma}_k}\exp\left(-\frac{(y_i - \boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}_k)^2}{2\widehat{\sigma}_k^2}\right)}{\sum_{\ell=1}^{K}\frac{\hat{\pi}_{i,\ell}}{\widehat{\sigma}_\ell}\exp\left(-\frac{(y_i - \boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}_\ell)^2}{2\widehat{\sigma}_\ell^2}\right)}.$$

- M-step: Updating $\widehat{\boldsymbol{\theta}}$ to lower the objective value in (6).
  - For each $k = 1, \ldots, K$, update $\widehat{\boldsymbol{\beta}}_k$:

$$\widehat{\boldsymbol{\beta}}_k = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n}\frac{\gamma_{i,k}(\widehat{\boldsymbol{\theta}})}{\widehat{\sigma}_k^2}\frac{(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2}{2n} + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_1. \tag{7}$$

  - Update $\{\widehat{\boldsymbol{\alpha}}_k\}_{k=1}^K$:

$$\{\widehat{\boldsymbol{\alpha}}_k\} = \arg\min_{\{\boldsymbol{\alpha}_k\}} -\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{K}\gamma_{i,k}(\widehat{\boldsymbol{\theta}})\boldsymbol{x}_i^T\boldsymbol{\alpha}_k - \log\left(\sum_{\ell=1}^{K}\exp(\boldsymbol{x}_i^T\boldsymbol{\alpha}_\ell)\right)\right) + \lambda_{\boldsymbol{\alpha}}\sum_{k=1}^{K}\|\boldsymbol{\alpha}_k\|_1. \tag{8}$$

  - For each $k = 1, \ldots, K$, update $\widehat{\sigma}_k$:

$$\widehat{\sigma}_k^2 = \frac{\sum_{i=1}^{n}\gamma_{i,k}(\widehat{\boldsymbol{\theta}})(y_i - \boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}_k)^2}{\sum_{i=1}^{n}\gamma_{i,k}(\widehat{\boldsymbol{\theta}})}. \tag{9}$$

As noted in Hyun et al. (2023), this M-step update represents a decreasing in the objective but not a minimization since $(\boldsymbol{\beta}_k, \sigma_k)$ would need to be jointly optimized. The algorithm terminates when the improvement in (6) is below some threshold.

In practice, we use the R package `flowmix` (Hyun, 2022) to carry out the EM algorithm.

2.2 Debiased prediction

Before describing our debiasing proposal, we discuss challenges of statistical inference with regularized models in high-dimensional settings, followed by a brief background on the debiasing techniques.

A common practice in estimating models in the high-dimensional regime ($p > n$) is to impose regularization terms which promote specific types of structure in the model (e.g. sparse structure through an ℓ_1-penalty as in the Lasso or low-rank structure through a nuclear norm regularization as is often used in matrix completion problems). The regularized M-estimators are given by the solution of an optimization problem of the following form

$$\widehat{\boldsymbol{\theta}} := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{x}_i, y_i; \boldsymbol{\theta}) + \lambda\mathcal{R}(\boldsymbol{\theta}), \tag{10}$$

for some loss function ℓ and a regularization function $\mathcal{R}$. There exists an abundant literature on the success of regularized M-estimators, showing that imposing relevant structures lead to significant reduction in the sample complexity needed for point estimation and point predictions (Candes & Tao, 2007; Raskutti et al., 2011; Zhao & Yu, 2006). However, the problems of statistical inference and uncertainty quantification go beyond point estimation and requires characterizing the distribution of the model estimates. This is a notoriously challenging task under high-dimensional settings and has been done only in some specific cases (e.g. Lasso with Gaussian designs, Bayati & Montanari, 2011.)

In classical statistics, where the number of parameters p is fixed while the sample size n grows to infinity, one can use *local asymptotic normality* (LAN), by which the Maximum Likelihood Estimator (MLE) undergoes an unbiased normal distribution (see e.g. Van der Vaart, 2000). However, LAN theory does not apply to the high-dimensional settings where p exceeds the sample size. Indeed, due to the added regularization penalty in the loss function (10), the estimator $\widehat{\boldsymbol{\theta}}$ is biased toward models $\boldsymbol{\theta}$ with small $\mathcal{R}(\boldsymbol{\theta})$. Formally, $\mathbb{E}\widehat{\boldsymbol{\theta}} \neq \boldsymbol{\theta}_0$, and characterizing the distribution of $\widehat{\boldsymbol{\theta}}$ is notoriously challenging in this regime.

Resampling methods such as the bootstrap and jackknife (Efron & Hastie, 2021) are also fraught with problems in the high-dimensional setting. An early work of Bickel and Freedman (1983) considers the residual bootstrap in the setting of $p/n \to \kappa \in (0, 1)$ and show that when $\kappa > 0$, there exists a data-dependent direction such that the projection of the bootstrap estimate along it does not have the correct asymptotic distribution. El Karoui and Purdom (2018) complement this negative result by focusing on fixed, non-data-dependent projection directions (e.g. a single coordinate of the model) and show that both of the most commonly used methods of bootstrapping (residual bootstrap and pairs bootstrap) give poor inference on the model as the ratio p/n grows. Likewise, jackknife resampling severely overestimates the variance in high dimensions. Note that these results apply when p grows at the same rate of n. The problem becomes exacerbated when $p \gg n$ and for penalized likelihood methods (similar to the setting of interest in our work). For example, Chatterjee and Lahiri (2010) consider the Lasso estimate and show that the asymptotic distribution of the bootstrapped lasso estimator is a random measure, which is inconsistent when at least one of the model parameters is zero. Chatterjee and Lahiri (2011) propose a modified bootstrap method to overcome this problem, but this line of work requires case-by-case modification and does not apply to more complicated penalized estimators such as MoE. Another issue that we would like to highlight is the bias in the penalized likelihood estimate. While there are bias-corrected bootstrap methods in the literature (Efron & Tibshirani, 1994) their validity holds under strong assumptions (e.g. that the difference of the model estimate and the true model being a pivotal quantity, meaning that its distribution does not depend on the model parameters).

One potential attempt to cope with the issue of bias is to try to estimate it and then enlarge the prediction sets accordingly to still contain the true response. However, this may result in suboptimally large prediction sets. For example, in the case of the Lasso with regularization parameter λ, Javanmard and Montanari (2014a) show that $\|\text{Bias}(\widehat{\boldsymbol{\theta}})\|_\infty \gtrsim \lambda$, while an alternative proposed debiased Lasso estimator $\widehat{\boldsymbol{\theta}}^d$ satisfies $\|\text{Bias}(\widehat{\boldsymbol{\theta}}^d)\|_\infty \lesssim s\lambda\sqrt{(\log p)/n}$ (see Corollary 2.7, 2.8 in

Javanmard & Montanari, 2014a for a formal statement). Therefore, in the regime of $s \ll \sqrt{n/(\log p)}$ the Lasso estimate has a bias substantially larger than the debiased estimator. We refer to Appendix E of the online supplementary material for further discussion on the bias of regularized M-estimators.

This prelude motivates us to follow a debiasing approach to first remove the bias of the regularized M-estimator (asymptotically) and then construct the prediction sets based on the resulting debiased estimator. We follow a similar technique proposed in a series of work (Javanmard & Montanari, 2014a; Van de Geer et al., 2014; Zhang & Zhang, 2014), where the idea is to first compute the regularized estimator $\widehat{\boldsymbol{\theta}}$ and then add a term to compensate for the bias introduced by the regularization $\mathcal{R}(\boldsymbol{\theta})$. This is done by moving the estimator in a direction based on the (sub)gradient of the regularization. Note that by the stationary condition for $\widehat{\boldsymbol{\theta}}$, we have $\lambda \nabla R(\widehat{\boldsymbol{\theta}}) = -\nabla \mathcal{L}(\widehat{\boldsymbol{\theta}})$, where we use the shorthand $\mathcal{L}(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{x}_i, y_i, \boldsymbol{\theta})$ to refer to the total loss. The debiased estimator then takes the form

$$\widehat{\boldsymbol{\theta}}^d := \widehat{\boldsymbol{\theta}} - \boldsymbol{M} \nabla \mathcal{L}(\widehat{\boldsymbol{\theta}}),$$

for some matrix $\boldsymbol{M}$. While the focus of Javanmard and Montanari (2014a), Zhang and Zhang (2014) and Van de Geer et al. (2014) has been on linear regression, the above description is a natural extension to general loss functions and regularizers. There have been different proposals for constructing $\boldsymbol{M}$. In Van de Geer et al. (2014) it is constructed using node-wise Lasso on the features matrix. In Javanmard and Montanari (2014a), it is constructed through a quadratic optimization which aims to minimize the variance of the resulting debiased estimator while controlling its bias.

For linear regression, the asymptotic properties of the above debiasing technique and its minimax optimality were studied in Javanmard and Montanari (2014a) (see also T. Cai et al., 2021 for the minimax optimality of the length of confidence intervals built upon the debiased estimator). In addition, the asymptotic optimality in terms of semiparametric efficiency was established in Van de Geer et al. (2014).

It is worth noting that the goal in Javanmard and Montanari (2014a), Zhang and Zhang (2014) and T. T. Cai and Guo (2017) is to establish inference for coefficients, which differs from our goal of establishing inference for prediction. Closer to our aim here, we follow the framework of T. Cai et al. (2021) which was proposed to construct prediction intervals under a linear regression model; however, our setting of a MoE model is more complex and requires novel methodology and analysis.

We next proceed by describing our debiasing approach for the MoE model. From (1)–(2), we have that

$$y_{\text{new}} \,|\, \boldsymbol{x}_{\text{new}} \sim N(\boldsymbol{x}_{\text{new}}^T \boldsymbol{\beta}_k, \sigma_k^2) \text{ with probability } \pi_k(\boldsymbol{x}_{\text{new}}).$$

Because of the penalization, $\hat{\Gamma}_k := \boldsymbol{x}_{\text{new}}^T \widehat{\boldsymbol{\beta}}_k$ is a biased estimate of $\Gamma_k := \boldsymbol{x}_{\text{new}}^T \boldsymbol{\beta}_k$, and thus would give biased predictions for y_{new} even if we knew that $z_{\text{new}} = k$. We therefore propose a debiasing procedure. Given an arbitrary vector $\boldsymbol{u}_k$, we construct a debiased prediction

$$\hat{\Gamma}_k^d := \boldsymbol{x}_{\text{new}}^T \widehat{\boldsymbol{\beta}}_k + \boldsymbol{u}_k^T \frac{1}{n} \sum_i \frac{\partial \ell_i(\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\beta}_k}, \tag{11}$$

where $\partial / \partial \boldsymbol{\beta}_k$ denotes the gradient with respect to $\boldsymbol{\beta}_k$, and $\ell_i(\widehat{\boldsymbol{\theta}})$ are the summands in (4) evaluated at $\widehat{\boldsymbol{\theta}} = (\{\widehat{\boldsymbol{\beta}}_k, \widehat{\boldsymbol{\alpha}}_k, \widehat{\sigma}_k\}_{k \in [K]})$.

Our proposed choice for $\boldsymbol{u}_k$ depends on the estimated sample Fisher information matrix, which is given by

$$\tilde{\boldsymbol{I}}(\widehat{\boldsymbol{\theta}}) = \frac{1}{n} \sum_{i=1}^{n} \nabla \ell_i(\widehat{\boldsymbol{\theta}}) \nabla \ell_i^T(\widehat{\boldsymbol{\theta}}), \tag{12}$$

with

$$\nabla \ell_i = \left(\left\{ \frac{\partial \ell_i}{\partial \boldsymbol{\beta}_k} \right\}_{k=1}^{K}, \left\{ \frac{\partial \ell_i}{\partial \boldsymbol{\alpha}_k} \right\}_{k=1}^{K}, \left\{ \frac{\partial \ell_i}{\partial \sigma_k} \right\}_{k=1}^{K} \right)^T.$$

We would like a choice of $\boldsymbol{u}_k$ that will lead to a narrower interval. This intuition (which is described in greater detail below) suggests choosing $\boldsymbol{u}_k$ as the solution to the optimization problem

$$\begin{aligned} \min_{\boldsymbol{u}} \quad & \boldsymbol{u}^T \tilde{\boldsymbol{I}}_k^{\beta}(\widehat{\boldsymbol{\theta}}) \boldsymbol{u} \\ \text{s.t.} \quad & \sup_{\omega \in \mathcal{C}} \left| \langle \omega, \tilde{\boldsymbol{\Sigma}}_k \boldsymbol{u} - \boldsymbol{x}_{\text{new}} \rangle \right| \le \lambda_k \|\boldsymbol{x}_{\text{new}}\|_2, \quad \|\boldsymbol{u}\|_1 \le L \|\boldsymbol{x}_{\text{new}}\|_2, \end{aligned} \tag{13}$$

where we define $\tilde{\boldsymbol{\Sigma}}_k := \frac{1}{n} \sum_{i=1}^{n} \frac{\gamma_{i,k}(\widehat{\boldsymbol{\theta}})}{\widehat{\sigma}_k^2} \boldsymbol{x}_i \boldsymbol{x}_i^T$, L is a sufficiently large constant, and λ_k is a tuning parameter (in Section 3, we discuss the proper rate of λ_k and the choice of L). In addition, $\mathcal{C} = \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p, \boldsymbol{x}_{\text{new}}/\|\boldsymbol{x}_{\text{new}}\|_2\}$, where $\boldsymbol{e}_i$ denotes the ith standard Euclidean basis vector. The matrix $\tilde{\boldsymbol{I}}_k^{\beta}(\widehat{\boldsymbol{\theta}})$ is the estimated Fisher information matrix, constrained to $\boldsymbol{\beta}_k$, which is given by

$$\tilde{\boldsymbol{I}}_k^{\beta}(\widehat{\boldsymbol{\theta}}) = \frac{1}{n} \sum_{i=1}^{n} \left(\gamma_{i,k}(\widehat{\boldsymbol{\theta}}) \frac{y_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}_k}{\widehat{\sigma}_k^2} \boldsymbol{x}_i \right) \left(\gamma_{i,k}(\widehat{\boldsymbol{\theta}}) \frac{y_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}_k}{\widehat{\sigma}_k^2} \boldsymbol{x}_i \right)^T. \tag{14}$$

The above characterization follows from the definition of $\tilde{I}(\widehat{\boldsymbol{\theta}})$ given by (12) along with identity (30) of the online supplementary material, restricted to class k.

Remark 1 Because $\tilde{\boldsymbol{\Sigma}}_k$ and $\tilde{\boldsymbol{I}}_k^{\beta}(\widehat{\boldsymbol{\theta}})$ depend on both the response vector $\mathbf{y}$ and the covariate matrix $\mathbf{X}$, the solution $\boldsymbol{u}_k$ to (13) is also dependent on them. This is in contrast to the optimization problem proposed by Javanmard and Montanari (2014a) for constructing the debiasing direction, as it only involved covariates and hence conditional on the covariate matrix $\mathbf{X}$, the debiasing direction was independent of the response vector $\mathbf{y}$.

To deal with the complications resulting from the dependence of $\boldsymbol{u}_k$ on $(\mathbf{X}, \mathbf{y})$, we do sample splitting. Specifically, we split the data into two sets, $\mathcal{D}_1$ and $\mathcal{D}_2$, where optimization (13) is solved on $\mathcal{D}_1$, and $\widehat{\boldsymbol{\beta}}_k$, $\hat{\Gamma}_k^d$ are calculated using samples in $\mathcal{D}_2$, per (11).

The above approach to construct the direction $\boldsymbol{u}_k$ is generalized from Javanmard and Montanari (2014a), Zhang and Zhang (2014) and T. T. Cai and Guo (2017) and aims to find a direction that minimizes the variance while controlling the bias. We would like to stress that our optimization problem differs from T. Cai et al. (2021) in that we use a constrained Fisher information matrix, $\tilde{\boldsymbol{I}}_k^{\beta}(\widehat{\boldsymbol{\theta}})$, in the objective of (13), and we allow the matrix $\tilde{\boldsymbol{\Sigma}}_k$ in the constraint of (13) to be different from the matrix in the objective, while T. Cai et al. (2021) keeps them the same.

The first constraint in (13) can be decomposed as

$$\left\| \tilde{\boldsymbol{\Sigma}}_k \boldsymbol{u} - \boldsymbol{x}_{\text{new}} \right\|_{\infty} \le \lambda_k \|\boldsymbol{x}_{\text{new}}\|_2 \tag{15}$$

$$\left| \langle \boldsymbol{x}_{\text{new}}, \tilde{\boldsymbol{\Sigma}}_k \boldsymbol{u} - \boldsymbol{x}_{\text{new}} \rangle \right| \le \lambda_k \|\boldsymbol{x}_{\text{new}}\|_2^2 \tag{16}$$

Similar to the general intuition behind the quadratic optimization of Javanmard and Montanari (2014a), the objective value $\boldsymbol{u}^T \tilde{\boldsymbol{I}}_k^{\beta}(\widehat{\boldsymbol{\theta}}) \boldsymbol{u}$ is related to the variance of $\hat{\Gamma}_k^d$, and the constraint (15) relates to its bias. So the optimization is indeed aiming to minimize the variance (and hence the length of prediction intervals which will be constructed based on $\hat{\Gamma}_k^d$), while controlling the bias of $\hat{\Gamma}_k^d$. That said, in the analysis we need to show that the bias is dominated by the variance term and therefore

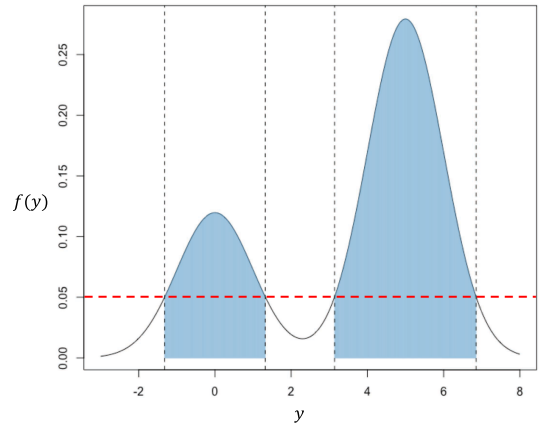

Figure 2. An illustration of the mixture density f, given by (19) for $K = 2$ groups. The cut-off level (indicated by the red line) is the highest level such that the region shaded in blue has area at least $1 - q$.

need to establish a lower bound on the variance. The constraint (16) is added for this step. It makes the feasible set of the optimization problem smaller and makes it possible to lower bound the optimal value of the objective (see Proposition A3 for technical arguments). This idea originates from T. Cai et al. (2021), which introduced the 'variance-enhancement projection direction'. While the general intuition carries over to our current setting, characterizing the statistical properties of $\hat{\Gamma}_k^d$ under the MoE requires a rather intricate and technical analysis.

We denote the estimated variance for $\hat{\Gamma}_k^d$ as

$$\widehat{V}_k = \frac{1}{n} \boldsymbol{u}_k^T \tilde{\boldsymbol{I}}_k^{\beta}(\widehat{\boldsymbol{\theta}}) \boldsymbol{u}_k, \quad k = 1, \ldots, K, \tag{17}$$

and the prediction variance estimate (conditional on $z_{\text{new}} = k$) as

$$b_k^2 := \widehat{V}_k + \widehat{\sigma}_k^2, \quad k = 1, \ldots, K. \tag{18}$$

Before proceeding, we note that solving (13) involves optimizing a quadratic program for each point $\boldsymbol{x}_{\text{new}}$ for which a prediction set is desired. This can become computationally expensive when the number of $\boldsymbol{x}_{\text{new}}$ is large. In such a case, a fruitful line of future work may look into warm-starting schemes (in which the solution from a neighbouring $\boldsymbol{x}_{\text{new}}$ is used in initialization), which could be performed either on (13) directly or on its dual.

2.3 Prediction sets

Recall that our goal is to construct a $100(1 - q)\%$ prediction set $\Omega_q(\boldsymbol{x}_{\text{new}})$ satisfying (3). By the nature of the mixture, we seek a prediction set of the form

$$\Omega_q(\boldsymbol{x}_{\text{new}}) = \cup_{k=1}^K \mathcal{L}_k,$$

where each $\mathcal{L}_k = [l_k, u_k]$ is centred at a debiased estimator $\hat{\Gamma}_k^d$. Regarding the length of $\Omega_q(\boldsymbol{x}_{\text{new}})$ as our budget it is clear at an intuitive level that we should spend more on the mixture components to which $(\boldsymbol{x}_{\text{new}}, y_{\text{new}})$ is more likely to belong, i.e. to those groups k with larger $\pi_k(\boldsymbol{x}_{\text{new}})$. To this end, we form a probability density function using a weighted mixture of Gaussian densities:

$$f(y) = \sum_{k=1}^K \frac{\hat{\pi}_k(\boldsymbol{x}_{\text{new}})}{b_k} \cdot \phi\left(\frac{y - \hat{\Gamma}_k^d}{b_k}\right), \tag{19}$$

where $\phi(\cdot)$ is the standard normal pdf. The particular form of this density is justified in the proof of Theorem 2 given in Section B of the online supplementary material. We give a schematic illustration of f in Figure 2.

We seek a set of intervals $[y_1^-, y_1^+], \ldots, [y_K^-, y_K^+]$, such that

$$\sum_{k=1}^{K} \int_{y_k^-}^{y_k^+} f(y)dy = 1 - q, \tag{20}$$

while minimizing $\sum_{k=1}^{K} |y_k^+ - y_k^-|$. For this, we start from a large cutoff (the horizontal line in the figure) and decrease that until the area under f and corresponding to y with $f(y)$ above the cutoff (the blue region in the figure) is $1 - q$. To approximate the area we take a discretization approach as outlined below.

Assume without loss of generality that $\hat{\Gamma}_1^d \leq \cdots \leq \hat{\Gamma}_K^d$. We start by considering the interval

$$\mathcal{Q} = [\hat{\Gamma}_1^d - b_1 z_{1-q/2}, \quad \hat{\Gamma}_K^d + b_K z_{1-q/2}],$$

which we know has probability at least $1 - q$. We then divide $\mathcal{Q}$ into segments of size δ, denoted as $\mathcal{Q}_1, \ldots, \mathcal{Q}_{\frac{|\mathcal{Q}|}{\delta}}$. The area under the curve $f(y)$ confined to the segment $\mathcal{Q}_i$ is approximately δh_i, where h_i is the density f evaluated at the midpoint point of $\mathcal{Q}_i$. We next sort h_i's corresponding to each segment in decreasing order, i.e.

$$h_{(1)} \geq h_{(2)} \geq \cdots \geq h_{(\frac{|\mathcal{Q}|}{\delta})},$$

and find the smallest N such that

$$\delta \sum_{i=1}^{N} h_{(i)} \geq 1 - q. \tag{21}$$

We return $\Omega_q(\boldsymbol{x}_{\text{new}}) := \cup_{i=1}^{N} \mathcal{Q}_{(i)}$ as the prediction set.

Algorithm 1: Constructing prediction set $\Omega_q(\boldsymbol{x}_{\text{new}})$ with $(1 - q)$ coverage

Input: Confidence level $1 - q$, discretization scale δ, debiased estimate of each centre $\hat{\Gamma}_k^d$, prediction standard error estimate b_k, membership probability estimate $\hat{\pi}_k(\boldsymbol{x}_{\text{new}})$, $k = 1, \ldots, K$.

Output: Prediction set with $(1 - q)$ coverage

1: Form a weighted mixture of Gaussian densities

$$f(y) = \sum_{k=1}^{K} \frac{\hat{\pi}_k(\boldsymbol{x}_{\text{new}})}{b_k} \cdot \phi\left(\frac{y - \hat{\Gamma}_k^d}{b_k}\right).$$

2: Choose a large enough interval $\mathcal{Q}$ so that the integral $\int_{\mathcal{Q}} f(y) \geq 1 - q$, i.e.

$$\mathcal{Q} = [\hat{\Gamma}_1^d - b_1 z_{1-q/2}, \quad \hat{\Gamma}_K^d + b_K z_{1-q/2}].$$

3: Divide $\mathcal{Q}$ into segments of size δ. Let y_i be the midpoint of $\mathcal{Q}_i$ and $h_i = f(y_i)$.

4: Sort h_i's in decreasing order,

$$h_{(1)} \geq h_{(2)} \geq \cdots \geq h_{(\frac{|\mathcal{Q}|}{\delta})}.$$

5: Find the smallest N such that

$$\delta \sum_{i=1}^{N} h_{(i)} \geq 1 - q.$$

6: Return the union of the corresponding segments

$$\Omega_q(\boldsymbol{x}_{\text{new}}) = \cup_{i=1}^{N} \quad \mathcal{Q}_{(i)}. \tag{22}$$

3 Theoretical guarantees

We consider a sequence of problems where the sample size $n \to \infty$ and covariate dimension $p = p(n) \to \infty$, while the number of groups K is bounded, and we establish asymptotic validity of our prediction sets for the MoE model (1)–(2). For the sake of concreteness, we have thus far focused on the penalized EM estimator; however, in this section we describe our theoretical results in terms of a general estimator $\widehat{\boldsymbol{\theta}}$. Doing so elucidates the specific condition needed by an estimator for our theoretical results to hold. This condition, along with other assumptions about the random covariate vectors $\{\boldsymbol{x}_i\}_{i\in[n]}$ and the model parameters $\boldsymbol{\theta}$, are summarized below.

- (A1) *Parameter estimation* η_{est}. Suppose that

$$\max_{k\in[K]} \left\| \widehat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k \right\|_1 = \max_{k\in[K]} \left(\left\| \widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k \right\|_1 + \left\| \widehat{\boldsymbol{\alpha}}_k - \boldsymbol{\alpha}_k \right\|_1 + |\widehat{\sigma}_k - \sigma_k| \right) = O_p(\eta_{est}),$$

  where η_{est} scales with n, p and potentially other structure associated with the parameters (e.g. sparsity levels). We assume that

$$\sqrt{n} \log(np) \eta_{\text{est}}^2 = o(1). \tag{23}$$

- (A2) *Distribution of features*. We have a positive-semidefinite matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ (or more precisely a sequence of matrices of growing dimension) with bounded operator norm, $\|\boldsymbol{\Sigma}\|_{\text{op}} \le C_\Sigma$, for a constant C_Σ as $p \to \infty$. Suppose that $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}_i$ are independent sub-Gaussian vectors, with mean zero and sub-Gaussian norm $\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}_i\|_{\psi_2} = O(1)$.
- (A3) *Bounded noise and signal*. The noise variances σ_k^2 are strictly positive and bounded constants for $k \in [K]$. We also assume that $\max_{\ell,k\in[K]} \|\boldsymbol{\beta}_k - \boldsymbol{\beta}_\ell\|_2 = O(1)$.

Condition (A1) assumes an ℓ_1-consistency rate for the estimate $\widehat{\boldsymbol{\theta}}$. Consistency presupposes identifiability, which in the case of the $\{\boldsymbol{\alpha}_k\}$ may require additional assumptions (since $\{\boldsymbol{\alpha}_k + \mathbf{c} : k \in [K]\}$ corresponds to the same $\pi_k(\boldsymbol{x})$ as $\{\boldsymbol{\alpha}_k : k \in [K]\}$). Our theoretical results apply to any estimator which satisfy condition (23). The proposed EM estimator in Section 2.1 is just one specific choice. Instead of ℓ_1-regularization, one can follow other variants based on iterative truncation. For example, Wang et al. (2015) analyses a mixture of regression model with two groups and proposes a truncated EM algorithm (with a gradient ascent implementation) which achieves $\eta_{est} = s\sqrt{\log(p)\log(n)/n}$, with s the sparsity level of model parameters. The work of Zhang et al. (2020) derives a similar ℓ_1-consistency rate for the high-dimensional mixed linear regression with two groups, for an iterative EM procedure which performs ℓ_1 regularization at each step. While the mixture of experts model is more complicated, we conjecture that a similar rate for η_{est} carries over to this setting. Under such conjecture, condition (23) simplifies to

$$\frac{s^2 \log(np) \log(p) \log(n)}{\sqrt{n}} = o(1).$$

Condition (A2) is on the random covariate vectors $\boldsymbol{x}_i$ and is a common assumption in high-dimensional statistical estimation; see e.g. Bühlmann and Van De Geer (2011). Condition (A3) on the pairwise distances $\|\boldsymbol{\beta}_k - \boldsymbol{\beta}_\ell\|_2$ and noise variance σ_k^2 is to control the heterogeneity of data coming from different groups.

Our first theorem is on asymptotic normality of the bias-corrected estimators $\hat{\Gamma}_k^d$ defined in (11) and involves the matrix

$$\boldsymbol{\Sigma}_k = \mathbb{E}\left[\frac{\gamma_{1k}(\boldsymbol{\theta})}{\sigma_k^2} \boldsymbol{x}_1 \boldsymbol{x}_1^T \right].$$

Theorem 1 Suppose that $\frac{\|\Sigma_k^{-1}x_{\text{new}}\|_1}{\|x_{\text{new}}\|_2} = O(1)$ and $\log(p) = o(n^{1/4}/\sqrt{\log(n)})$. Let u_k be the solution to the optimization problem (13) with constant $L \geq \frac{\|\Sigma_k^{-1}x_{\text{new}}\|_1}{\|x_{\text{new}}\|_2}$ and $\lambda_k \asymp \eta_{\text{est}}\log(np) + \sqrt{\log(p)/n}$. Under Assumptions (A1), (A2), (A3), and Remark 1, we have

$$\frac{\sqrt{n}(\hat{\Gamma}_k^d - \Gamma_k)}{\sqrt{u_k^T \tilde{I}_k^\beta(\widehat{\theta})u_k}} \overset{d}{\Rightarrow} N(0, 1), \tag{24}$$

where $\tilde{I}_k^\beta(\widehat{\theta})$, given by (14), is the sample Fisher information matrix constrained to entries corresponding to β_k.

Remark 2 In Section 2.1, we proposed to use an EM estimator with ℓ_1 penalization to obtain the initial estimator $\widehat{\theta}$. It is worth noting that in constructing the debiased predictions (Section 2.2) we do not use the explicit form of the regularization (it is only based on the loss function ℓ). Also Theorem 1 establishes unbiased normality of the predictions $\hat{\Gamma}_k^d$ under Assumption (A1) which only involves the estimation error of the initial estimator $\widehat{\theta}$, but not the specific form of the regularization used to construct it. Therefore, we do not explicitly enforce sparsity (or any other specific form of regularization), which speaks to the generality of our result.

Now that we have established the asymptotic normality of our debiased estimators, we are ready to prove that our prediction sets provide proper asymptotic coverage.

Theorem 2 Under the assumptions of Theorem 1, the prediction set $\Omega_q(x_{\text{new}})$ has asymptotically valid coverage. Specifically, fix $\gamma > 0$ arbitrarily small and in Algorithm 2.3 set the discretization scale δ so that

$$\delta \leq 11\sqrt{\gamma \min_{k\in[K]}(b_k)/\text{Len}(\mathcal{Q})},$$

with $\text{Len}(\mathcal{Q})$ representing the length of interval $\mathcal{Q}$. Then, for any x_{new} and its response y_{new} generated according to MoE, we have

$$\lim_{n\to\infty} \mathbb{P}\left(y_{\text{new}} \in \Omega_q(x_{\text{new}}) \mid x_{\text{new}}\right) \geq 1 - q - \gamma.$$

Note that choosing γ arbitrarily small we get a coverage arbitrarily close to $1 - q$. We refer to Section A of the online supplementary material for the proof of Theorems 1 and 2.

4 Numerical study

In Section 4.1, we return to the low-dimensional example given in Section 1 and consider several variations to build greater understanding of the behaviour of our intervals. In Section 4.2, we assess the performance of our procedures in a high-dimensional example.

4.1 A low-dimensional example

Figure 1 shows a two-group example where the mean functions of the two groups are

$$\mu_1(t) = 10(t - 0.5)^2 \quad \text{and} \quad \mu_2(t) = 1,$$

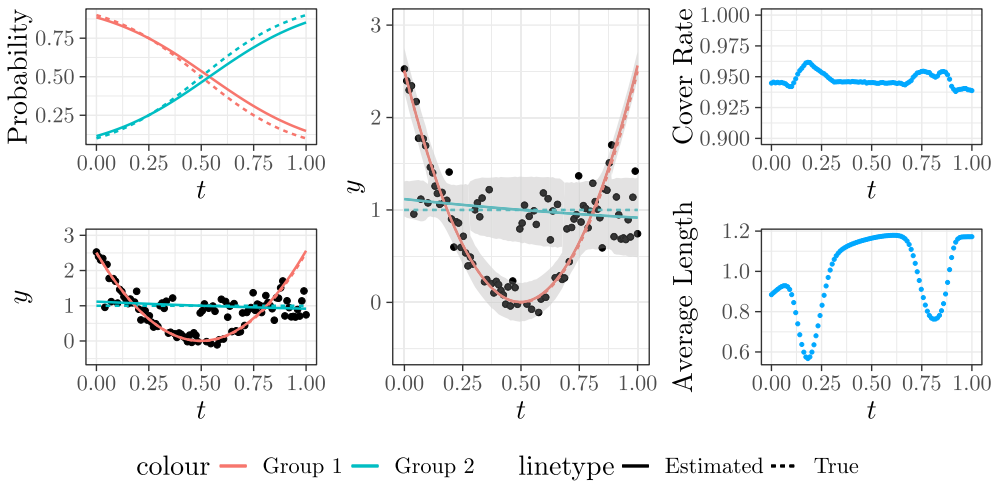

Figure 3. Unequal error variance case ($\sigma_1 < \sigma_2$). See caption of Figure 1 for description of panels.

the error variances are $\sigma_k^2 = 0.15^2$, and the log odds of being in the first group is given by

$$\log\left[\frac{\pi_1(t)}{1-\pi_1(t)}\right] = \log(9) - 2\log(9)\cdot t,$$

for $n = 100$ equally spaced values of t ranging from 0 to 1. This means that $\pi_1(t)$ decreases from 0.9 to 0.1. Writing this MoE model in the notation of (1)–(2), we have $\boldsymbol{x} = (1, t, t^2)^T$, $\boldsymbol{\beta}_1 = (2.5, -10, 10)^T$, $\boldsymbol{\beta}_2 = (1, 0, 0)^T$, $\boldsymbol{\alpha}_1 = (\log 9, -2\log 9, 0)^T$, and $\boldsymbol{\alpha}_2 = (0, 0, 0)^T$. This initial setting is perfectly symmetric in the two groups other than the difference in mean functions. For example, the first group at $t = 0$ is as common as the second group is at $t = 1$. This symmetry is manifest in the width of the intervals: In the middle panel, the width of the interval around the first group at $t = 0$ matches that of the second group at $t = 1$. The symmetry is also apparent in the third panel plots.

To build our intuition, we explore the effect of breaking this symmetry between groups. For example, suppose that the error variances in the two groups are not equal. In Figure 3, we take $\sigma_1 = 0.1$ and $\sigma_2 = 0.2$. While the empirical coverage is maintained around the nominal 0.95 level, the average length of the prediction set $\Omega_{0.05}(t)$ is no longer symmetric around $t = 0.5$. This makes intuitive sense since the class imbalance means that there is more uncertainty in estimating $\boldsymbol{\beta}_2$ than $\boldsymbol{\beta}_1$; thus, at $t = 1$, where more of the prediction set is devoted to the second group (since $\pi_2(1) > \pi_1(1)$), the average length of $\Omega_{0.5}(1)$ will be larger compared to that of $\Omega_{0.5}(0)$.

Another way to break the symmetry would be by considering class imbalance. We again assume $\sigma_1 = \sigma_2 = 0.15$, but now suppose that overall about 60% of observations belong to the first group. In particular, we take instead $\boldsymbol{\alpha}_1 = \log 10 \cdot (1, -\log 4, 0)^T$, so that $\pi_1(t)$ ranges from about 0.91 at $t = 0$ to about 0.29 at $t = 1$. Figure 4 shows the effect of this class imbalance. We see the same increasing length as in the previous example despite the error variances being equal. In this example, it is the class imbalance that leads to greater uncertainty in estimating $\boldsymbol{\beta}_2$.

4.2 High-dimensional case

We consider a high-dimensional case where $K = 2$ and $p = 501$ (including the intercept). The oceanographic application in Hyun et al. (2023) has repeated observations at multiple $\boldsymbol{x}_t$, and we mimic that setup with $T = 150$ feature vectors $\boldsymbol{x}_t$, and $n_t = 5$ observations per $\boldsymbol{x}_t$ for a total of $n = 750$ measurements. This is high-dimensional since $\boldsymbol{\theta}$ has dimension $K(2p+1) = 2002$. We take $x_{t1} = 1$ and x_{tj} for $j > 1$ to be independent standard Gaussians. We generate n_t responses y_{it} for each $\boldsymbol{x}_t$ according to the MoE model (1)–(2) with $\sigma_1 = \sigma_2 = 1$ and

$$\boldsymbol{\beta}_1 = (-2, 4, -2, -4, 6, 2, 0, 0, 0, 0, 0, \mathbf{0}_{490}^T)^T,$$
$$\boldsymbol{\beta}_2 = (2, 0, 0, 0, 0, 0, 4, -2, -4, 6, 2, \mathbf{0}_{490}^T)^T,$$

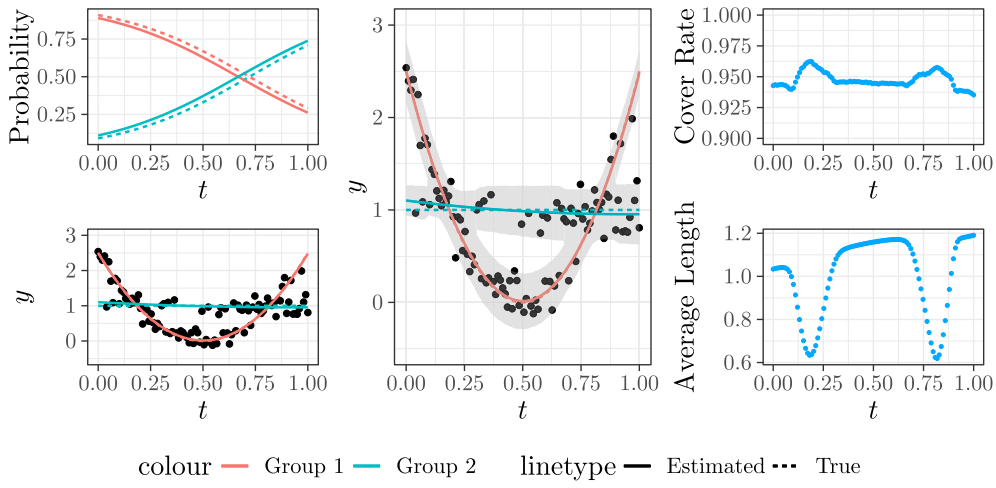

Figure 4. Class imbalance case ($\mathbb{E}[\pi_1(\boldsymbol{x})] > \mathbb{E}[\pi_2(\boldsymbol{x})]$). See caption of Figure 1 for description of panels.

and

$$\boldsymbol{\alpha}_1 = (\mathbf{0}_{491}, 0, 0, 0, 0, 0, 0.7, 0.7, 0.7, 0.7, 0.7)^T,$$
$$\boldsymbol{\alpha}_2 = (\mathbf{0}_{491}, -0.7, -0.7, -0.7, -0.7, -0.7, 0, 0, 0, 0, 0)^T.$$

We estimate the model using the penalized EM algorithm described in Section 2.1 as implemented in the `flowmix` R package (Hyun, 2022). We perform fivefold cross-validation to choose the parameters $(\lambda_{\boldsymbol{\alpha}}, \lambda_{\boldsymbol{\beta}})$ from a 10×10 logarithmically spaced grid. While in Section 2.2, we remark that sample splitting avoids the complications resulting from $\boldsymbol{u}_k$'s dependence on $(\mathbf{X}, \mathbf{y})$, empirically we find that coverage is attained even if we ignore this dependence. Therefore, in this and all numerical results we do not use sample splitting for the debiasing step.

To evaluate our method, we generate 100 independent $\boldsymbol{x}_{\text{new}} \in \mathbb{R}^p$. For each $\boldsymbol{x}_{\text{new}}$, we compute $\Omega_{0.05}(\boldsymbol{x}_{\text{new}})$ and record its length. We then generate 100 independent $y_{\text{new},i}$'s for each $\boldsymbol{x}_{\text{new}}$ and record the proportion of $y_{\text{new},i}$'s falling into the prediction set for that $\boldsymbol{x}_{\text{new}}$.

We repeat the above procedure 500 times, keeping the 100 $\boldsymbol{x}_{\text{new}}$'s the same. For each $\boldsymbol{x}_{\text{new}}$, we compute

1. the average length of prediction sets (across the 500 runs), and
2. the coverage probability (proportion across the 500 runs and 100 $y_{\text{new},i}$).

In the left and middle panels of Figure 5, we plot these quantities as a function of $\pi_1(\boldsymbol{x}_{\text{new}})$, the true probability of being drawn from cluster 1 for each $\boldsymbol{x}_{\text{new}}$. As desired, the coverage rate of our prediction sets meet the 95% nominal level, regardless of $\pi_1(\boldsymbol{x}_{\text{new}})$. We observe that the prediction sets tend to be twice as long when $\pi_1(\boldsymbol{x}_{\text{new}}) \approx 0.5$ compared to at the extremes. This is likely because at an extreme the less common group's interval can be very narrow without hurting coverage whereas when the two groups are balanced, both intervals are needed.

In the right panel of Figure 5, we plot the average length of prediction intervals against the distances between the two group means, i.e.

$$\left|\boldsymbol{x}_{\text{new}}^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)\right|.$$

Recall that in the low-dimensional examples (Figures 1, 3, and 4) we observed a marked decrease in the prediction set length when the means crossed each other. A similar phenomenon is apparent

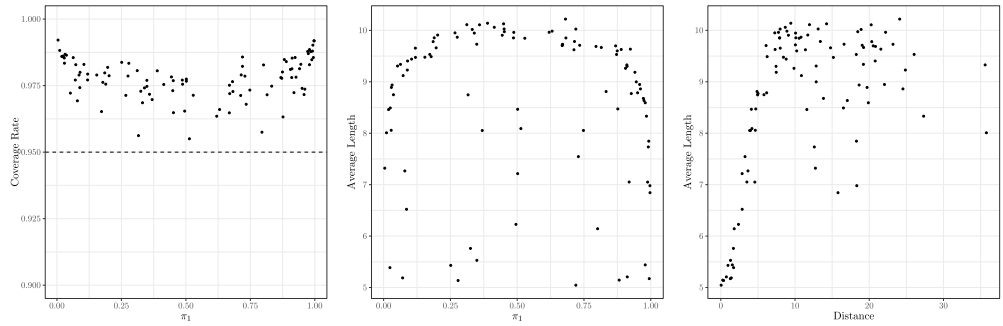

Figure 5. High-dimensional simulation. Every point corresponds to a different $x_{\text{new}} \in \mathbb{R}^p$. The empirical coverage and average length of the prediction set $\Omega_{0.05}(x_{\text{new}})$ is computed over 500 training sets and 100 y_{new} values for each x_{new}.

here. There is a linear increasing trend when $|x_{\text{new}}^T(\beta_1 - \beta_2)|$ goes from 0 to 10. As suggested in the middle panel, the individual intervals have average length around 5. Thus, this linear increase represents the two overlapping intervals gradually being pulled apart. At a distance of 10, they no longer overlap, which explains the levelling of this trend. The variability in length seen for distances greater than 10 can be explained, for example, by differing values of $\pi_1(x_{\text{new}})$.

5 Superconductivity data application

We apply our method to the superconductivity data provided in Hamidieh (2018a). This dataset contains the critical temperature (in Kelvin) and a set of $p = 81$ attributes for about 21,000 materials. The attributes used as predictors are elemental property statistics and electronic structures of attributes. We centre and scale each predictor column, and we take the response to be $\log(1 + \text{temperature})$. The log transform makes the data less skewed right and adding 1 Kelvin to each temperature can be thought of as replacing the log of extremely low temperatures (some are less than $1mK$) with 0.

We randomly split the observations into a training set of $n = 200$ (used for estimating model parameters, cross-validation of λ_α and λ_β), a validation set of size 1000 (used to choose K), and a test set of about 20,000 observation (for measuring the coverage of our prediction sets). Table 1 shows the mean squared prediction error, computed on the validation set, for K ranging from 1 to 5. To make predictions at a given $x_{\text{validation}}$, we use

$$x_{\text{validation}}^T \widehat{\beta}_{\hat{k}(x_{\text{validation}})} \quad \text{where} \quad \hat{k}(x_{\text{validation}}) = \arg\max_k \hat{\pi}_k(x_{\text{validation}})$$

is the class with highest estimated probability. The prediction errors on the validation set suggest that $K = 2$ may be a suitable choice (where the lowest error is highlighted in bold).

For each observation $(x_{\text{new}}, y_{\text{new}})$ in the test set, we form $\Omega_{0.05}(x_{\text{new}})$ and note whether $y_{\text{new}} \in \Omega_{0.05}(x_{\text{new}})$. Figure 6 displays the prediction sets for a random subset of 100 of the 20,000 intervals formed on the test set. We see that $\Omega_{0.05}(x_{\text{new}})$ is often a single interval, although it also occasionally the union of two intervals. The overall coverage on the test set is 97.1%. The average length of $\Omega_{0.05}(x_{\text{new}})$ (after being transformed back from log-values) is around 42 Kelvin.

The 97.1% coverage is averaged over all x_{new} and yet our prediction sets are designed for conditional coverage in the sense of (3). This stronger form of coverage implies that we can get coverage on subsets of observations defined by x_{new}, i.e.

$$\mathbb{P}(y_{\text{new}} \in \Omega_q(x_{\text{new}}) \mid x_{\text{new}} \in \mathcal{S}) \geq 1 - q.$$

The predictor that is most correlated with critical temperature is the weighted standard deviation of thermal conductivity. We divide the range of this variable into five equally spaced sub-intervals and divide test data points into 5 subgroups accordingly. Table 2 confirms that our prediction sets meet the nominal level within each subgroup.

Table 1. Prediction errors computed on a validation set for fitted models where $K = 1, 2, 3, 4, 5$

K	Prediction error
1	8.953
2	**0.867**
3	0.978
4	1.318
5	0.869

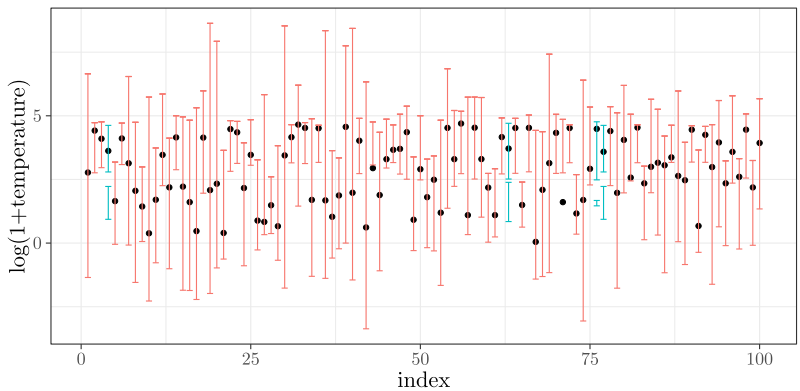

Figure 6. Plot of prediction sets for 100 randomly sampled data points in the test set. The black points are the log (1 + temperature) and the bars correspond to the prediction sets for each observation. Note that for some of the observations the prediction set consists of two intervals.

Table 2. Coverage rates of 95% prediction sets conditional on subgroups defined by the predictor 'weighted standard deviation of thermal conductivity'

Subgroup	Number of data	Coverage rate (%)
1	3,098	95.9
2	4,039	97.2
3	3,960	97.1
4	3,999	95.6
5	4,000	94.8

6 Discussion and extensions

MoE models are useful for modelling heterogeneity in regression relationships. We consider here the extent to which they are connected to mixed-effects models, another popular approach to modelling heterogeneity in data. In this section, we first highlight the major differences between the MoE model, studied in this article, and the mixed-effects model. We then discuss an extension of our methodology to cases where the heterogeneity in data is formulated by a continuous random effect, rather than a discrete group structure as in the MoE model.

6.1 Linear mixed-effects models

In various settings, data have an inherent grouping structure where observations within groups are dependent and those between groups are independent. LMMs (Henderson Jr, 1982; Pinheiro & Bates, 2006) provide a flexible tool for analysing such clustered data, including settings with

repeated measurements on the same unit, panel (longitudinal) data, multi-level data (data with nested or hierarchical structure), or data collected across different environments.

In the LMM setup, it is assumed that observations belong to N groups, indexed by $i = 1, , N$, each of size n_i. For each group, we observe a $n_i \times 1$ vector of responses $\boldsymbol{y}_i$ generated according to the following model

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\gamma}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, N, \tag{25}$$

where $\boldsymbol{X}_i \in \mathbb{R}^{n_i \times p}$ is the observed fixed-effects design matrix, $\boldsymbol{Z}_i \in \mathbb{R}^{n_i \times m}$ is the observed random effects design matrix, $\boldsymbol{\gamma}_i$ is an unknown group-specific vector of random effects, and $\boldsymbol{\beta}$ is an unknown vector of fixed effects. It is often assumed that $\boldsymbol{\epsilon}_i \sim N_{n_i}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_{n_i})$ and uncorrelated for $i \in [N]$. In addition, $\boldsymbol{\gamma}_i \sim N_m(\boldsymbol{0}, \boldsymbol{\Psi})$ and uncorrelated for $i \in [N]$, and also independent of noises $\boldsymbol{\epsilon}_i$.

We explain the main differences between the MoE model and the LMM by first trying to write the MoE (1) as an LMM. A natural attempt would be to define $\bar{\boldsymbol{\beta}} := (\sum_{k=1}^{K} \boldsymbol{\beta}_k)/K$ as the fixed effect and the deviations $\boldsymbol{\Delta}_k := \boldsymbol{\beta}_k - \bar{\boldsymbol{\beta}}$ as the group-specific effects, which brings us to

$$\boldsymbol{y}_k = \boldsymbol{X}_k\bar{\boldsymbol{\beta}} + \boldsymbol{X}_k\boldsymbol{\Delta}_k + \boldsymbol{\epsilon}_k. \tag{26}$$

Now comparing (25) and (26) already exhibits two main differences:

1. In LMMs, $\boldsymbol{\gamma}_i$ are random (often drawn from a normal distribution), which are different among groups. However, in (26), $\boldsymbol{\Delta}_k$ are unknown but deterministic vectors. The randomness comes from the group memberships of the samples, and as described by (2) these membership probabilities depend on the predictors.
2. In LMMs, the group structure is typically known; for example, it is known which observations are repeated measures on the same unit, or which observations are collected in the same environment. In contrast, the main flexibility of an MoE model is to allow for latent group structure in data.

In summary, while both the MoE model and the LMM are flexible tools to model group structure in the data, they do it in very different ways. The LMM is often used when the groups are known and allows for dependence among observations from the same group by incorporating a random effect in the model. The MoE on the other hand deals with unknown groups and assumes that the group memberships are random and independent across samples, but with a probability that depends on the predictors of the sample.

6.2 Extension to continuous random effects

In the mixture-of-experts model, the heterogeneity in data is captured by a *discrete* cluster structure, where different clusters follow different linear models. Here, we discuss an extension where *continuous* random effects drive the regression coefficients of each sample. Concretely, we consider the model where the responses are generated as

$$y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\boldsymbol{u}_i + \varepsilon_j \quad \text{for} \quad i = 1, \ldots, n,$$

with $\boldsymbol{x}_i \in \mathbb{R}^p$ and $\boldsymbol{z}_i \in \mathbb{R}^m$ are the observed features, $\boldsymbol{\beta} \in \mathbb{R}^p$ is the unknown fixed vector and $\boldsymbol{u}_i \in \mathbb{R}^m$ is the unknown (continuous) random effect. We assume $(\boldsymbol{u}_i, \boldsymbol{x}_i, y_j)$ are i.i.d, but allow $\boldsymbol{u}_i$ to depend on $\boldsymbol{x}_i$, i.e. it is generated from a density function $f(\boldsymbol{u} \mid \boldsymbol{x}_i)$.

As discussed earlier, the goal of prediction inference is that for a given confidence level $q \in (0, 1)$ and new predictor vectors $\boldsymbol{x}_{\text{new}}, \boldsymbol{z}_{\text{new}}$, we want to build a prediction set $\Omega_q(\boldsymbol{x}_{\text{new}}, \boldsymbol{z}_{\text{new}})$ such that $\mathbb{P}(y_{\text{new}} \in \Omega_q(\boldsymbol{x}_{\text{new}}, \boldsymbol{z}_{\text{new}}) \mid \boldsymbol{x}_{\text{new}}, \boldsymbol{z}_{\text{new}}) \geq 1 - q$. Here, we discuss how our developed methodology can be used for this setting. We will sketch the generalization and focus the presentation on the main ideas.

Constructing predictions sets. Define $\Gamma := \boldsymbol{x}_{\text{new}}^T\boldsymbol{\beta}$. Our procedure has two main steps:

- *Debiased estimation of* Γ: Following our procedure we first estimate $\boldsymbol{\beta}$ given the training sets and then use the debiasing technique to obtain an asymptotically unbiased estimate $\hat{\Gamma}^d$ of Γ. Note that the log-likelihood function of the data $\{(\boldsymbol{x}_i, \boldsymbol{z}_i, y_i)\}$ reads

$$\ell(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \log\left[\int f_{\boldsymbol{v}}(\boldsymbol{u} \mid \boldsymbol{x}_i)\frac{1}{\sigma}\phi\left(\frac{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta} - \boldsymbol{z}_i^T\boldsymbol{u}}{\sigma}\right)\mathrm{d}\boldsymbol{u}\right], \tag{27}$$

  with $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{v})$ denoting the parameters in the model. Here, $f_{\boldsymbol{v}}(\boldsymbol{u}|\boldsymbol{x})$ denotes the conditional density function of random effect $\boldsymbol{u}$, coming from a family of distribution parametrized by $\boldsymbol{v}$. By comparing the above equation with (4) it is clear that $f_{\boldsymbol{v}}(\boldsymbol{u}|\boldsymbol{x})$ plays the role of $\pi_k(\boldsymbol{x}_i)$ in the MoE model. Similar to our developed procedure, we use an EM algorithm on the ℓ_1-penalized log-likelihood function to get estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{v}}$ of $\boldsymbol{\beta}$ and $\boldsymbol{v}$. We then construct the debiased estimate $\hat{\Gamma}^d$ of Γ_d similar to (11).
- *Mixture of Gaussian densities:* Similar to Equation (19), we form a probability density function using a weighted mixture of Gaussian densities:

$$f(y) = \int \frac{f_{\hat{\boldsymbol{v}}}(\boldsymbol{u}|\boldsymbol{x}_{\text{new}})}{\sigma}\phi\left(\frac{y - \hat{\Gamma}^d - z_{\text{new}}^T\boldsymbol{u}}{\sigma}\right)\mathrm{d}\boldsymbol{u}.$$

  This is the conditional density of y_{new} given $\boldsymbol{x}_{\text{new}}$, z_{new} and $\hat{\Gamma}^d$. We then follow the same lines (2–6) of Algorithm 2.3 to construct $\Omega_q(\boldsymbol{x}_{\text{new}}, z_{\text{new}})$.

7 Conclusion

We have shown how to construct prediction sets for the high-dimensional mixture of experts model. Mixture models are important for capturing the heterogeneity that is present in many real-world situations. While in small data samples, it was common to dismiss deviations from the norm as outliers, in large data sets it becomes possible to use models that can identify and model these subgroups. While mixture of regression models allow for such heterogeneity-aware predictive modelling, they assume that the relative sizes of the subgroups are fixed. Importantly, mixture of experts models remove this assumption and allow the prevalence of different subgroups to depend on the features. This generalization is essential in many situations from ecology, where the relative proportions of different subpopulations depends on environmental covariates (Hyun et al. 2023), to politics, where the political composition depends on demographic and geographic variables.

Our focus on conditional coverage can be crucial in certain applications. For example, Romano et al. (2020) emphasizes the importance of ensuring that all subpopulations enjoy the same coverage guarantees and cast this as a fairness issue when the subpopulations are defined based on a protected attribute.

Acknowledgments

We would like to thank the Editor, the Associate Editor, and anonymous reviewers for thoughtful comments and suggestions which have improved the content and presentation of the article. Part of the work reported in the article was conducted while Simeng Shao was a PhD student at the Data Sciences and Operations department at the University of Southern California.

Conflicts of interest: None declared.

Funding

A.J. was partially supported by the Sloan Research Fellowship in mathematics, Adobe Data Science Faculty Research Awards, Amazon Faculty Research Award, National Science Foundation CAREER Award DMS-1844481 and National Science Foundation Award DMS-2311024. J.B. was supported in part by National Science Foundation CAREER Award DMS-1653017.

Data availability

The superconductor dataset used in the numerical experiments of this article is available in the UCI Machine Learning Repository (Hamidieh, 2018b).

References

Athey S., Imbens G. W., & Wager S. (2018). Approximate residual balancing: Debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *80*(4), 597–623. https://doi.org/10.1111/rssb.12268

Balakrishnan S., Wainwright M. J., & Yu B. (2017). Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Annals of Statistics*, *45*(1), 77–120. https://doi.org/10.1214/16-AOS1435

Bayati M., & Montanari A. (2011). The lasso risk for Gaussian matrices. *IEEE Transactions on Information Theory*, *58*(4), 1997–2017. https://doi.org/10.1109/TIT.2011.2174612

Bickel P. J., & Freedman D. A. (1983). Bootstrapping regression models with many parameters. In *Festschrift for Erich L. Lehmann. Wadsworth Statist./Probab. Ser* (pp. 28–48). Wadsworth.

Bühlmann P., & Van De Geer S. (2011). *Statistics for high-dimensional data: Methods, theory and applications*. Springer Science & Business Media.

Cai T., Tony Cai T., & Guo Z. (2021). Optimal statistical inference for individualized treatment effects in high-dimensional models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *83*(4), 669–719. https://doi.org/10.1111/rssb.12426

Cai T. T., & Guo Z. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *Annals of Statistics*, *45*(2), 615–646. https://doi.org/10.1080/02331888.2016.1265969

Candes E., & Tao T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, *35*(6), 2313–2351. https://doi.org/10.1214/009053606000001523

Chatterjee A., & Lahiri S (2010). Asymptotic properties of the residual bootstrap for lasso estimators. *Proceedings of the American Mathematical Society*, *138*(12), 4497–4509. https://doi.org/10.1090/proc/2010-138-12

Chatterjee A., & Lahiri S. N. (2011). Bootstrapping lasso estimators. *Journal of the American Statistical Association*, *106*(494), 608–625. https://doi.org/10.1198/jasa.2011.tm10159

Dempster A. P., Laird N. M., & Rubin D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Efron B., & Hastie T (2021). *Computer age statistical inference, student edition: Algorithms, evidence, and data science*. (Vol. 6). Cambridge University Press.

Efron B., & Tibshirani R. J. (1994). *An introduction to the bootstrap*. Chapman and Hall/CRC.

El Karoui N., & Purdom E. (2018). Can we trust the bootstrap in high-dimensions? The case of linear models. *Journal of Machine Learning Research*, *19*(5), 1–66. https://doi.org/10.4855

Fournier Q., Caron G. M., & Aloise D (2023). A practical survey on faster and lighter transformers. *ACM Computing Surveys*, *55*(14s), 1–40. https://doi.org/10.1145/3586074

Gormley I. C., & Frühwirth-Schnatter S. (2019). Mixture of experts models. In *Handbook of mixture analysis* (pp. 271–307). CRC Press.

Gormley I. C., Murphy T. B., & Raftery A. E. (2023). Model-based clustering. *Annual Review of Statistics and Its Application*, *10*, 573–595. https://doi.org/10.1146/statistics.2023.10.issue-1

Guo Z., Renaux C., Bühlmann P., & Cai T. (2021). Group inference in high dimensions with applications to hierarchical testing. *Electronic Journal of Statistics*, *15*(2), 6633–6676. https://doi.org/10.1214/21-EJS1955

Hamidieh K. (2018a). A data-driven statistical model for predicting the critical temperature of a superconductor. *Computational Materials Science*, *154*, 346–354. https://doi.org/10.1016/j.commatsci.2018.07.052

Hamidieh K (2018b). Superconductivty Data Data Set. https://doi.org/10.24432/C53P47

Hastie T., Tibshirani R., Friedman J. H., & Friedman J. H (2009). *The elements of statistical learning: Data mining, inference, and prediction*. (Vol. 2). Springer.

Henderson Jr C. R (1982). Analysis of covariance in the mixed model: Higher-level, nonhomogeneous, and random regressions. *Biometrics*, *38*(3), 623–640. https://doi.org/10.2307/2530044

Hunter D. R., & Lange K. (2004). A tutorial on mm algorithms. *The American Statistician*, *58*(1), 30–37. https://doi.org/10.1198/0003130042836

Hyun S (2022). *flowmix: Ocean Flow Cytometry Analysis*. R package version 0.0.0.9000.

Hyun S., Cape M. R., Ribalet F., & Bien J. (2023). Modeling cell populations measured by flow cytometry with covariates using sparse mixture of regressions. *The Annals of Applied Statistics*, *17*(1), 357–377. https://doi.org/10.1214/22-AOAS1631

Javanmard A., & Lee J. D. (2020). A flexible framework for hypothesis testing in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *82*(3), 685–718. https://doi.org/10.1111/rssb.12373

Javanmard A., & Montanari A (2014a). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, *15*(1), 2869–2909.

Javanmard A., & Montanari A (2014b). Hypothesis testing in high-dimensional regression under the Gaussian random design model: Asymptotic theory. *IEEE Transactions on Information Theory*, *60*(10), 6522–6554. https://doi.org/10.1109/TIT.2014.2343629

Javanmard A., & Montanari A (2018). Debiasing the lasso: Optimal sample size for gaussian designs. *Annals of Statistics*, *46*(6A), 2593–2622. https://doi.org/10.1214/17-AOS1630

Jordan M. I., & Jacobs R. A. (1994). Hierarchical mixtures of experts and the em algorithm. *Neural Computation*, *6*(2), 181–214. https://doi.org/10.1162/neco.1994.6.2.181

Klusowski J. M., Yang D., & Brinda W. (2019). Estimating the coefficients of a mixture of two linear regressions by expectation maximization. *IEEE Transactions on Information Theory*, *65*(6), 3515–3524. https://doi.org/10.1109/TIT.18

Kwon J., & Caramanis C (2020). Em converges for a mixture of many linear regressions. In *International Conference on Artificial Intelligence and Statistics* (pp. 1727–1736). PMLR.

Kwon J., Qian W., Caramanis C., Chen Y., & Davis D (2019). Global convergence of the em algorithm for mixtures of two component linear regression. In Conference on Learning Theory (pp. 2055–2110). PMLR.

Lei J., G'Sell M., Rinaldo A., Tibshirani R. J., & Wasserman L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, *113*(523), 1094–1111. https://doi.org/10.1080/01621459.2017.1307116

Lei J., & Wasserman L. (2014). Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(1), 71–96. https://doi.org/10.1111/rssb.12021

Papadopoulos H., Proedrou K., Vovk V., & Gammerman A (2002). Inductive confidence machines for regression. In *European Conference on Machine Learning* (pp. 345–356). Springer.

Pinheiro J., & Bates D. (2006). *Mixed-effects models in S and S-PLUS*. Springer Science & Business Media.

Quandt R. E., & Ramsey J. B. (1978). Estimating mixtures of normal distributions and switching regressions. *Journal of the American Statistical Association*, *73*(364), 730–738. https://doi.org/10.1080/01621459.1978.10480085

Raskutti G., Wainwright M. J., & Yu B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, *57*(10), 6976–6994. https://doi.org/10.1109/TIT.2011.2165799

Romano Y., Barber R. F., Sabatti C., & Cands E. J. (2020). With malice towards none: Assessing uncertainty via equalized coverage. *Harvard Data Science Review*, *2*(2), https://doi.org/10.1162/99608f92.03f00592

Romano Y., Patterson E., & Candes E. (2019). Conformalized quantile regression. *Advances in Neural Information Processing Systems*, *32*, 3543–3553. https://doi.org/10.4855

Städler N., Bühlmann P., & van de Geer S. A. (2010). L1-penalization for mixture regression models. *TEST*, *19*, 209–256. https://doi.org/10.1007/s11749-010-0197-z

Van de Geer S., Bühlmann P., Ritov Y., & Dezeure R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, *42*(3), 1166–1202. https://doi.org/10.1214/14-AOS1221

Van der Vaart A. W (2000). *Asymptotic statistics*. (Vol. *3*). Cambridge University Press.

Vovk V (2012). Conditional validity of inductive conformal predictors. In Asian conference on machine learning (pp. 475–490). PMLR.

Vovk V., Gammerman A., & Shafer G. (2005). *Algorithmic learning in a random world*. Springer Science & Business Media.

Wang Z., Gu Q., Ning Y., & Liu H. (2015). High dimensional expectation-maximization algorithm: Statistical optimization and asymptotic normality. *NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems*, *2*, 2521–2529.

Yi X., & Caramanis C. (2015). Regularized EM algorithms: A unified framework and statistical guarantees. *Advances in Neural Information Processing Systems*, *28*, 1567–1575. https://doi.org/10.4855

Yi X., Caramanis C., & Sanghavi S (2014). Alternating minimization for mixed linear regression. In *International Conference on Machine Learning* (pp. 613–621). PMLR.

Yuksel S. E., Wilson J. N., & Gader P. D. (2012). Twenty years of mixture of experts. *IEEE transactions on neural networks and learning systems*, *23*(8), 1177–1193. https://doi.org/10.1109/TNNLS.2012.2200299

Zhang C.-H., & Zhang S. S (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(1), 217–242. https://doi.org/10.1111/rssb.12026

Zhang L., Ma R., Cai T. T., & Li H (2020). 'Estimation, confidence intervals, and large-scale hypotheses testing for high-dimensional mixed linear regression', arXiv, arXiv:2011.03598, preprint: not peer reviewed. https://doi.org/10.48550/arXiv.2011.03598

Zhao P., & Yu B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research*, *7*, 2541–2563.

Zhu H.-T., & Zhang H. (2004). Hypothesis testing in mixture regression models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *66*(1), 3–16. https://doi.org/10.1046/j.1369-7412.2003.05379.x

Zhu Y., & Bradic J. (2018). Linear hypothesis testing in dense high-dimensional linear models. *Journal of the American Statistical Association*, *113*(524), 1583–1600. https://doi.org/10.1080/01621459.2017.1356319