Neurocomputing 626 (2025) 129357

Contents lists available at ScienceDirect

Neurocomputing

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

Mixture of experts models for multilevel data: Modeling framework and approximation theory

Tsz Chai Fung [a],*, Spark C. Tseung [b]

[a] Maurice R. Greenberg School of Risk Science, Georgia State University, 35 Broad Street NW, Atlanta, GA 30303, United States
[b] Department of Statistical Sciences, University of Toronto, Ontario Power Building, 700 University Avenue, 9th Floor, Toronto, ON M5G 1Z5, Canada

ARTICLE INFO

Communicated by Y. Lai

Keywords:
Artificial neural network
Crossed and nested random effects
Denseness
Mixed effects models
Universal approximation theorem

ABSTRACT

Multilevel data are prevalent in many real-world applications. However, it remains an open research problem to identify and justify a class of models that flexibly capture a wide range of multilevel data. Motivated by the versatility of the mixture of experts (MoE) models in fitting regression data, in this article we extend upon the MoE and study a class of mixed MoE (MMoE) models for multilevel data. Under some regularity conditions, we prove that the MMoE is dense in the space of any continuous mixed effects models in the sense of weak convergence. As a result, the MMoE has a potential to accurately resemble almost all characteristics inherited in multilevel data, including the marginal distributions, dependence structures, regression links, random intercepts and random slopes. In a particular case where the multilevel data is hierarchical, we further show that a nested version of the MMoE universally approximates a broad range of dependence structures of the random effects among different factor levels.

1. Introduction

1.1. Background and literature review

Mixture of experts (MoE) model, which is first introduced by Jacobs et al. [1] (see also, e.g., Jordan and Jacobs [2],McLachlan and Peel [3] for details), is a probabilistic version of neural network architecture useful for flexible regression, classification and distribution modeling, with applications to various areas including healthcare, business, social and environmental science. Readers may refer to Yuksel et al. [4], Masoudnia and Ebrahimpour [5],Nguyen and Chamroukhi [6] for the literature reviews on both the theories and applications of the MoE.

The model structure of the MoE is as follows. Suppose that we have N observations $(\boldsymbol{y},\boldsymbol{x}) = \{(\boldsymbol{y}_i,\boldsymbol{x}_i)\}_{i=1,\dots,N}$, where $\boldsymbol{y}_i = (y_{i1},\dots,y_{iK})$ is a K-dimensional response variable with output space $\mathcal{Y} \subseteq \mathbb{R}^K$ and $\boldsymbol{x}_i = (x_{i1}\dots,x_{iP})$ are the P covariates or features with input space $\mathcal{X} \subseteq \mathbb{R}^P$. Under the MoE framework, the conditional distribution function of $\boldsymbol{y}_i$ given $\boldsymbol{x}_i$ is

$$F(\boldsymbol{y}_i;\boldsymbol{\alpha},\boldsymbol{\psi},g|\boldsymbol{x}_i) = \sum_{j=1}^{g} \pi_j(\boldsymbol{x}_i;\boldsymbol{\alpha}) F_0(\boldsymbol{y}_i;\boldsymbol{\psi}_j|\boldsymbol{x}_i), \tag{1}$$

where g is the number of latent classes. Here, $\pi_j(\boldsymbol{x}_i;\boldsymbol{\alpha}) > 0$ is called the gating function with $\sum_{j=1}^{g}\pi_j(\boldsymbol{x}_i;\boldsymbol{\alpha}) = 1$ and parameters $\boldsymbol{\alpha}$. The left panel of Fig. 2 graphically illustrates the MoE model architecture. While the most commonly used gating function is the logit-linear or softmax gating [1], expressed as $\pi_j(\boldsymbol{x}_i;\boldsymbol{\alpha}) = \exp\{\alpha_{j,0} + \boldsymbol{\alpha}_j^T\boldsymbol{x}_i\}/\sum_{j'=1}^{g}\exp\{\alpha_{j',0} + \boldsymbol{\alpha}_{j'}^T\boldsymbol{x}_i\}$ with $\boldsymbol{\alpha} = \{\alpha_{j0},\boldsymbol{\alpha}_j : j = 1,\dots,g\}$, various alternative gating functions have also been explored in the literature. These include Gaussian gating [7], student-t gating [8], probit gating [9], as well as more advanced "sparse" gating functions designed to reduce the complexity of MoE models, such as the Top-k gate [10] and the DSelect-k gate [11]. Also, $F_0(\boldsymbol{y}_i;\boldsymbol{\psi}_j|\boldsymbol{x}_i)$ is a probability distribution called the expert function with parameters $\boldsymbol{\psi} := \{\boldsymbol{\psi}_j : j = 1,\dots,g\}$. While a common choice for the expert function is a Gaussian distribution [12], there has been substantial developments on alternative choices of expert functions to cater for various distributional characteristics such as heavy-tailedness (Laplace by Nguyen and McLachlan [13], t-distribution by Chamroukhi [14], skewed t by Chamroukhi [15] and transformed gamma by Fung et al. [16]) and discrete distributions (Poisson by Grun and Leisch [17] and Erlang Count by Fung et al. [18]).

Model flexibility is a crucial desirable property for the class of MoE, and there are extensive research work on the approximation theory for the MoE. Zeevi et al. [19] shows that the mean function of a univariate ($K = 1$) logit-gated MoE can approximate any Sobolev class functions. This result is extended by Jiang and Tanner [20], who considers the transformed Sobolev class. Without considering the convergence rate, Nguyen et al. [21] proves that the MoE mean function is dense in the class of any continuous functions without the restriction of the

* Corresponding author.
E-mail addresses: tfung@gsu.edu (T.C. Fung), spark.tseung@mail.utoronto.ca (S.C. Tseung).

https://doi.org/10.1016/j.neucom.2025.129357
Received 3 October 2022; Received in revised form 8 September 2024; Accepted 8 January 2025
Available online 7 February 2025

Sobolev class, and Nguyen et al. [22] shows similar denseness results using a multivariate ($K > 1$) Gaussian-gated MoE.

Apart from studying the mean functions, some research works focus on conditional density approximation with respect to the Hellinger distance, Kullback–Leibler (KL) divergence, or Lebesgue space. Jiang and Tanner [23],Mendes and Jiang [24] generalize the results of Jiang and Tanner [20] by demonstrating the approximation capability of the MoE to any exponential family non-linear regression models. Norets et al. [25] shows that the logit-gated MoE with Gaussian expert function can approximate any conditional densities. Similar results are proved by Nguyen et al. [22],Norets and Pelenis [26], who consider the Gaussian gating functions. Recently, Nguyen et al. [27] proves that the class of MoE is dense in the Lebesgue space.

Another stream of distribution approximation theorems studies the denseness in the sense of Prohorov metric of weak convergence. Extending upon Tijms [28],Breuer and Baum [29] who explore denseness of finite mixture and phase-type distributions in the space of any probability distributions, Fung et al. [30] formulates the concept of "denseness" in regression settings and shows that the class of MoE is dense in the space of any regression distributions, subject to some regularity conditions such as Lipschitz continuity and distribution tightness. In contrast to other existing works on distribution approximations, the results of Fung et al. [30] are very general as: (i) they hold under a wide range of choices of expert functions (not restricted to Gaussian or other symmetric expert functions); (ii) the target distribution is not restricted to a special class (e.g., exponential family regression models).

Despite its model flexibility, the aforementioned framework implicitly assumes that input–output pairs are independent across observations. However, this assumption does not hold for multilevel data [31]. In addition to the inputs $\boldsymbol{x}_i$, the dataset includes L levels of factors, each corresponding to a categorical variable used to group or cluster data into different units or categories. Observations that share the same unit within a factor level are likely to have unobserved common characteristics, creating clear interdependencies among them. Overlooking these dependencies can result in spurious, misleading, or biased clustering and prediction outcomes [31].

Multilevel data are prevalent across many applications. The most classic one is the school problem [32–34], where "school" and "classroom within a school" act as two levels of factors affecting the performance of a student (as an observation). Multilevel data structure can also be caused by repeated measurements collected in longitudinal studies. This is common across various areas including health (e.g., Molenberghs et al. [35]) and business (e.g., Boucher and Denuit [36]). For instance, the medical outcome of a patient or the amount of insurance claims by a policyholder are measured or collected repeatedly over time. A remarkable special case of multilevel data is the hierarchical (or nested) data, where the L factor levels can be ranked from high to low. The school problem is a clear example of hierarchical data.

A popular model to account for the interdependencies among observations is the generalized linear mixed effects model (GLMM) [33,37], which assumes that the output $\boldsymbol{y}_i$ depends on the sum of fixed effects (i.e., the impact of the inputs $\boldsymbol{x}_i$) and random effects (i.e., the impact of the factors $\boldsymbol{\theta}_i$). To improve model flexibility or achieve specific clustering purposes, the GLMM framework is extended to a non-linear setting [38,39], formulated in a neural network structure [40] or integrated to a finite-mixture modeling framework [41,42]. Despite of the desirable properties of the MoE models leading to extensive applications, the research works of mixed effects models in the context of MoE framework are relatively scarce. Yau et al. [43] first proposes a two-component logit-gated Gaussian-expert MoE with random effects incorporated in both gating and expert functions. Ng and McLachlan [44] then formulates a general g-component mixed effect MoE with the use of logistic expert functions for binary classifications. Ng and McLachlan [45] considers a similar framework with random effects only incorporated to the expert functions. Nonetheless, all the aforementioned mixed effect MoE models only deal with a single level of random effect (i.e., $L = 1$).

1.2. Contributions of this paper

Driven by the growing popularity of MoE models, the widespread use of multilevel data, and the need to formally establish model flexibility through approximation theories, this paper introduces several innovative contributions to the theoretical modeling framework and approximation theory in the field of mixture of experts (MoE) models for multilevel data analysis.

First, we introduce the mixed MoE (MMoE) model for multilevel regression data, addressing a gap in the literature on MoE models for multilevel data. Our model incorporates multiple levels of random effects, providing a more comprehensive approach to multilevel data compared to the limited scope of existing works by Yau et al. [43],Ng and McLachlan [44], and Ng and McLachlan [45], which only consider a single level of random effects. Such a restriction is inadequate for handling complex multilevel data structures, such as the hierarchical structure seen in the school problem discussed in Section 1.1. Our model also features a reduced structure aimed at preserving parsimony and interpretability without sacrificing flexibility or approximation capability. Specifically, our model is simplified in two key ways: (i) we assume that random effects in the gating functions are shared across mixture latent classes and are independent, unlike the approach in Ng and McLachlan [44], which allows for varying and dependent random effects across latent classes; and (ii) we exclude the effects of observed inputs and random effects from the expert functions, as specified by Ng and McLachlan [44]. These reductions significantly decrease the number of model parameters and, as demonstrated in our subsequent work [46], are essential for feasible and effective model estimation, particularly with large datasets.

Second, we define the concept of denseness for mixed effects models in terms of weak convergence and demonstrate that the proposed MMoE class is dense in the space of continuous mixed effects models under certain mild regularity conditions. While universal approximation theories for standard input–output models, including MoE and neural network models, have been extensively explored in the statistics and machine learning literature (see Section 1.1), this paper represents the first attempt to formulate and prove approximation theories for multilevel models. Existing studies on mixed effects models within the MoE framework (e.g., Ng and McLachlan [44]) do not address the theoretical approximation capability of their models. This paper extends the setting in Fung et al. [30], which formulated denseness for regression distributions, to accommodate multilevel data. Our novel denseness theory not only illustrates the flexibility of the proposed model in capturing various characteristics of multilevel data, such as joint distributions, regression patterns, random intercepts, and random slopes, but also suggests that our model is parsimonious with a reduced structure. Compared to Fung et al. [30], this paper also relaxes several assumptions required for the denseness theorem, such as Lipschitz continuity and distribution tightness. Consequently, the proof techniques employed in this paper differ significantly from those in Fung et al. [30].

Third, in the context of hierarchical data, we prove that a nested version of the MMoE can effectively approximate a wide range of dependence structures between upper and lower-level factors, even when the MMoE is simplified to include only independent random effects across levels. This result is particularly relevant for applications where dependencies exist between factors, such as the impact of a classroom on a student's performance being influenced by the school the student attends.

The focus of this paper is to formulate the MMoE model for multilevel data and theoretically justify its versatility. In a subsequent paper [46], we will address the estimation and application problems under the proposed MMoE. A stochastic variational ECM algorithm is proposed to efficiently estimate the model parameters. Also, the MMoE is applied to an automobile insurance dataset, demonstrating its ability to reasonably predict policyholders' future claims based on their past claim histories.

Table 1
Hypothetical school problem dataset with $N = 12$ students, including scores, gender, and income.

Student i	School ID $c_1(i)$	Classroom ID $c_2(i)$	Score (year 1) y_{i1}	Score (year 2) y_{i2}	Gender (Male) x_{i1}	Income x_{i2}
1	1	1	79.48	80.18	0	10.82
2	1	1	66.16	60.96	1	11.29
3	1	2	77.28	67.48	0	12.21
4	2	3	77.30	53.53	1	10.31
5	2	3	85.91	80.29	0	11.16
6	2	4	75.65	73.99	1	10.13
7	2	5	85.02	74.35	0	11.41
8	2	5	72.84	82.08	1	11.92
9	3	6	78.21	83.41	0	11.00
10	3	7	57.64	65.50	0	11.61
11	4	8	64.47	66.21	1	10.60
12	4	8	73.38	52.63	1	11.70

1.3. Structure of the paper

This paper is structured as follows. Section 2 mathematically defines a multilevel data with an example. Section 3 defines a generalized class of mixed effect models for multilevel data, which includes nearly all mixed effect models in the literature. In Section 4, we introduce the MMoE as a candidate class of mixed effect models to flexibly capture multilevel data. Interpretation and visualization of the proposed model are also provided. Section 5 defines "denseness" in the context of mixed effect models and proves that the MMoE is a universal approximator of most mixed effect models subject to some mild conditions. In Section 6, we discuss the model formulation and denseness property in a special case where the dataset is hierarchical with nested random effects. The intuitive explanations on the proof idea of the theoretical results are demonstrated in Section 7, accompanied by a numerical illustration in Section 8. The findings are summarized in Section 9, accompanying some limitations of the denseness theory in justifying the approximation capability of the proposed MMoE.

2. Multilevel data structure

This section mathematically defines a multilevel data and provides a hypothetical example to illustrate how multilevel dataset is structured.

The observed multilevel data structure can be mathematically represented as $\mathcal{D} := \{(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{c}(i))\}_{i=1,\dots,N}$, where $\boldsymbol{y}_i$, $\boldsymbol{x}_i$, and N are defined in Section 1.1, and $\boldsymbol{c}(\cdot)$ is a function that maps each observation i to the corresponding units across all L levels of factors. Specifically, we denote $\boldsymbol{c}(i) := (c_1(i), \dots, c_L(i))$, where $c_l(\cdot) : \{1, \dots, N\} \mapsto \{1, \dots, S_l\}$ is a known function that maps observation i to its corresponding level-l unit, with S_l being the number of possible level-l units for $l = 1, \dots, L$. In other words, $c_l(i)$ identifies the level-l unit or category to which observation i belongs. The box in the top left of Fig. 1 provides a visual representation of the multilevel data structure.

Example 1. We construct a small hypothetical multilevel dataset in the context of the "school problem", consisting of $N = 12$ students, as shown in Table 1. The output of this dataset, $\boldsymbol{y}_i = (y_{i1}, y_{i2})$, is bivariate ($K = 2$), with y_{i1} and y_{i2} representing the average exam scores of student i in year 1 and year 2, respectively. The dataset also includes $P = 2$ covariates $\boldsymbol{x}_i = (x_{i1}, x_{i2})$ as inputs, where x_{i1} is a binary indicator for gender (1 for male) and x_{i2} is the log of household income. There are $L = 2$ levels of factors: "school" and "classroom". The second and third columns of the table, $c_1(i)$ and $c_2(i)$, represent the identifiers for the school and classroom to which each student belongs. For instance, observation $i = 7$ corresponds to a student in classroom 5 (a level-2 unit of the factor "classroom"), which is in school 2 (a level-1 unit of the factor "school"). This dataset includes a total of $S_1 = 4$ schools and $S_2 = 8$ classrooms.

Remark 1. One might wonder why not treat these factors as categorical variables, include them in the input $\boldsymbol{x}_i$ using one-hot encoding, and then fit a standard input–output model (e.g., regression or neural network), rather than organizing the dataset in a multilevel format and using mixed effects models. Although this approach is theoretically possible, it faces practical challenges. Typically, the number of possible level-l units, S_l, grows with the sample size at a rate of $O(N)$. When N is large, as in the case of the automobile insurance dataset discussed by Tseung et al. [46], the dimensionality of $\boldsymbol{x}_i$ would become computationally prohibitive, and the model would be highly susceptible to overfitting if the factors were treated as one-hot encoded categorical variables. Moreover, this approach ignores the interdependencies between outputs across different observations, leading to the problems discussed in Section 1.1. Additionally, standard input–output models have limitations in prediction when new units within a factor emerge (e.g., a new school or classroom not represented in the training data). Mixed effects models, which we will discuss in the next section, avoid this issue and can effectively predict future observations with previously unseen factor units.

3. Mixed effect models for multilevel data

Datasets with a multilevel structure are typically modeled using a mixed effects framework. In this approach, the known inputs $\boldsymbol{x}_i$ are considered the "fixed effects" or "hard parameter sharing". Additionally, it is assumed that there are L unobserved variables $\boldsymbol{\theta}_i = (\theta_{i1}, \dots, \theta_{iL})$ that jointly influence the output $\boldsymbol{y}_i$, where θ_{il} represents the effect of the level-l unit on observation i. We define $\theta_l^{(s)} := \theta_{il} = \theta_{i'l}$ if $c_l(i) = c_l(i') = s$ for $s = 1, \dots, S_l$ and $l = 1, \dots, L$. This means that if two observations i and i' belong to the same level-l unit, they share the same level-l unobserved variable $\theta_{il} = \theta_{i'l}$, making the observations statistically dependent. The unobserved variables $\boldsymbol{\theta}_i$ are treated as random and are specified by a probability distribution, thus $\boldsymbol{\theta}_i$ is regarded as a "random effect" or "soft parameter sharing".

Example 2. For the hypothetical "school problem" data in Table 1, the variable $\theta_1^{(s)}$ represents the unobserved characteristics shared by students from the same school, so $\theta_{i1} = \theta_{i'1}$ if students i and i' attend school s. Similarly, $\theta_2^{(s)}$ represents the shared characteristics of students in the same classroom, meaning $\theta_{i2} = \theta_{i'2}$ if they belong to classroom s. For instance, $\theta_1^{(2)} = \theta_{4,1} = \theta_{6,1}$ because students 4 and 6 attend the same school (#2), but $\theta_{4,2} \neq \theta_{6,2}$ since they are in different classrooms. The variables $\theta_1^{(s)}$ and $\theta_2^{(s)}$ can be interpreted as unobserved factors, such as the quality of teachers and the learning environment, that influence the exam performance of all students within the same school or classroom.

In this section, we will discuss some technical details regarding a generalized framework of the mixed effects models. Let $(\boldsymbol{\Omega}, \mathcal{F}, \mathbb{P})$ be the probability space and suppose that $\boldsymbol{\theta}_l^{(s)}$ is a $\mathcal{F}$-measurable map from $(\boldsymbol{\Omega}, \mathcal{F})$ to $(\boldsymbol{\Theta}_l, \mathcal{Q}_l)$ for every $l = 1, \dots, L$ and $s = 1, \dots, S_l$, where $\boldsymbol{\Theta}_l$ is the space of $\boldsymbol{\theta}_l^{(s)}$ or $\boldsymbol{\theta}_{il}$. $\boldsymbol{\theta}_l^{(s)}$ is a random variable if $(\boldsymbol{\Theta}_l, \mathcal{Q}_l) = (\mathbb{R}, \mathcal{R})$ where

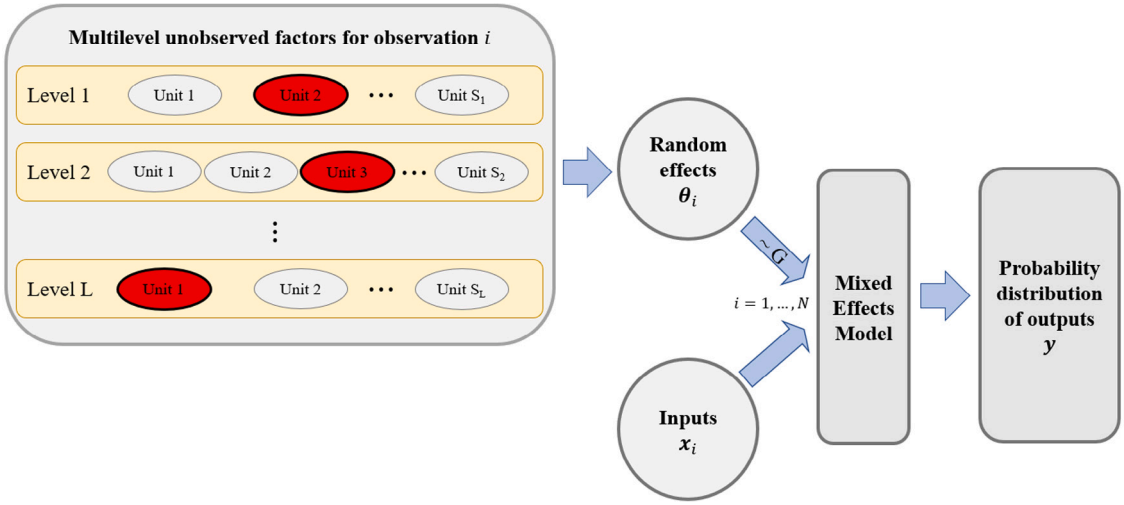


Fig. 1. Multilevel data structure and mixed effects modeling framework. The box in the top left illustrates the level-l unit to which observation i belongs across all factor levels $l = 1, \ldots, L$, corresponding to the mapping function $c(i)$. In this example, observation i is associated with the second level-1 unit, third level-2 unit, and first level-L unit. These factor unit classifications govern the random effects θ_i for observation i. For all observations $i = 1, \ldots, N$, the random effects θ_i, drawn from a distribution G, along with the inputs x_i, are used in the mixed effects model H as defined in Eq. (2). This model generates the joint probability distribution of outputs y, denoted by $\tilde{H}(y|x)$ in Eq. (4).

$\mathcal{R}$ is a Borel set, but we do not want to impose such a restriction. It is because the random effects $\theta_l^{(s)}$ are unobserved and we are unsure if they can be quantified as a real number. In the context of the school problem, summarizing the aforementioned unobserved characteristics of schools or classrooms into a single number is not feasible. The space $\boldsymbol{\Theta}_l$ also varies among different mixed effects models in literature. For example, $\boldsymbol{\Theta}_l = \mathbb{R}$ for most GLMMs [37], $\boldsymbol{\Theta}_l = \mathbb{R}^{2g-1}$ for the GLMM MoE model by Ng and McLachlan [44], and $\boldsymbol{\Theta}_l = \mathbb{R}^{gK}$ for the mixture of random effects models by Ng and McLachlan [45], where g and K are those defined in Section 1.

Under the generalized mixed effects model, we assume that y_i depends only on x_i and θ_i. Conditioned on x_i and θ_i, we further assume that $\{y_i\}_{i=1,\ldots,N}$ are mutually independent. Specifically, we denote

$$y_i|x_i, \theta_i \overset{\text{ind}}{\sim} H(\cdot|x_i, \theta_i), \qquad i = 1, \ldots, N, \tag{2}$$

where $H := H(\cdot|x_i, \theta_i)$ can be any probability distributions on y_i given x_i and θ_i. With these assumptions, the joint distribution of y given x is given by

$$\tilde{H}(y|x) = \int_{\Omega} \left[\prod_{i=1}^{N} H(y_i|x_i, \theta_i) \right] d\mathbb{P}. \tag{3}$$

Suppose that each measurable space $(\boldsymbol{\Theta}_l, \mathcal{Q}_l)$ is also equipped by a probability measure G_l, corresponding to the "distribution" of $\theta_l^{(s)}$. Denote further $(\tilde{\boldsymbol{\Theta}}, \tilde{\mathcal{Q}}, G)$ as the product of probability spaces $\{(\boldsymbol{\Theta}_l, \mathcal{Q}_l, G_l)\}_{l=1,\ldots,L;s=1,\ldots,S_l}$. Then, the joint distribution in Eq. (3) can be re-written as

$$\tilde{H}(y|x) = \int_{\tilde{\boldsymbol{\Theta}}} \left[\prod_{i=1}^{N} H(y_i|x_i, \theta_i) \right] dG(\tilde{\theta}), \tag{4}$$

where $\tilde{\theta} = \{\theta_l^{(s)}\}_{l=1,\ldots,L;s=1,\ldots,S_l}$. Fig. 1 graphically summarizes the modeling framework. The model framework above encompasses a broad range of models designed for multilevel or hierarchical data, including generalized linear mixed models (GLMM) and nonlinear mixed effects models (see, e.g., Goldstein [33],Davidian and Gallant [38]). Since there are no constraints on the functional forms of H and G, this structure is highly flexible. It naturally incorporates any potential joint distributions of $y_i|x_i, \theta_i$, regression links between x_i and y_i (nonlinear effects and interactions among covariates), the influence of random effects θ_i on y_i (random intercepts), and the interactions between θ_i and x_i (random slopes).

The following assumption has been implicitly made on $\theta_l^{(s)}$:

Assumption 1. $\{\theta_l^{(s)}\}_{l=1,\ldots,L;s=1,\ldots,S_l}$ are mutually independent.

The above assumption implies that $(\tilde{\boldsymbol{\Theta}}, \tilde{\mathcal{Q}}, G)$ is a product probability space, so $G(\tilde{\theta})$ can be written as

$$G(\tilde{\theta}) = \prod_{l=1}^{L} \prod_{s=1}^{S_l} G_l(\theta_l^{(s)}) \qquad \text{or} \qquad dG(\tilde{\theta}) = \prod_{l=1}^{L} \prod_{s=1}^{S_l} G_l(d\theta_l^{(s)}). \tag{5}$$

Note that the prior independence assumption across factors within a level (i.e., $s = 1, \ldots, S_l$) is very natural for most data structures especially for those involving repeated measurements (see, e.g., Goldstein [33],Boucher and Denuit [36],Ng et al. [41],Yau et al. [43], and Ng and McLachlan [44]). The prior independence across levels (i.e., $l = 1, \ldots, L$) is also often assumed for datasets with multilevel structure (see, e.g., Goldstein [33],McGilchrist [37]).

4. Mixture of experts model with random effect

Despite of the generality of the above mixed effect model (Eq. (3)), it is essential to appropriately specify the functional forms of H and G to model a multilevel dataset. However, this is challenging, especially when the space $\tilde{\boldsymbol{\Theta}}$ of the latent random effects $\tilde{\theta}$ defined in Section 3 is not observed from the dataset. Recall that a multilevel dataset only provides information on how each observation is classified into one of the factors for each level l, but not on what the factors are or how to quantify these factors.

In this section, we introduce the mixture of experts (MoE) model with random effects, called the mixed MoE (MMoE), as a candidate regression model to cater for multilevel data structure. The justification of the proposed model, which analyzes the ability of the MMoE to accurately approximate the generalized form of the mixed effect model (Eq. (3)), will be presented in the next section.

4.1. Model set-up

Under the MMoE, we assume that each observation i is equipped by L levels of random effects, denoted as an L-vector $w_i = (w_{i1}, \ldots, w_{iL})$. Similar to the mapping of the unobserved factors introduced in Section 3, we also have $w_{il} = w_{i'l} := w_l^{(s)}$ if $c_l(i) = c_l(i') = s$. Hence, $w_l^{(s)}$ represents the level-l random effect associated with all observations that correspond to the sth level-l factor. The only difference between θ_i and w_i is that we restrict $w_{il} \in \mathbb{R}$ into a Euclidean space instead of a general space $\tilde{\boldsymbol{\Theta}}$, which is unknown and hard to specify. Similar to

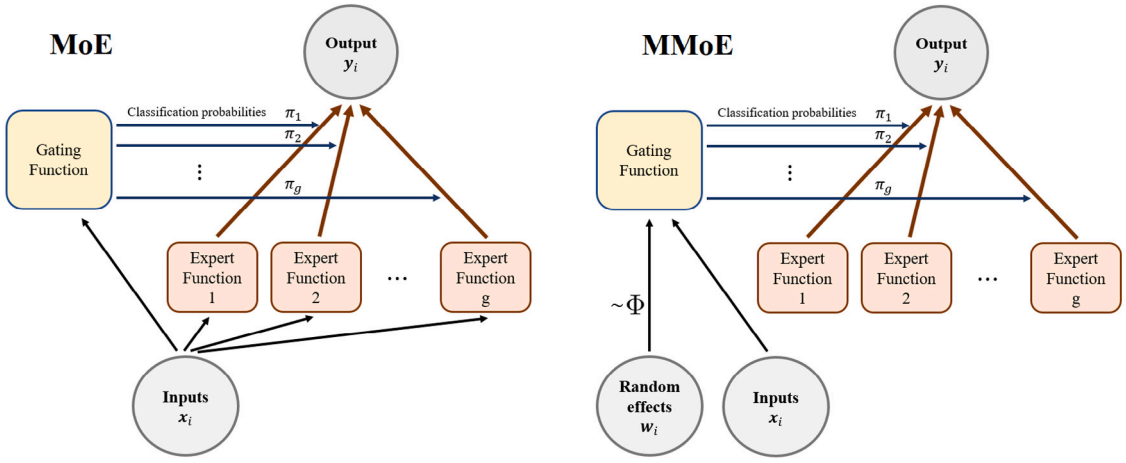


Fig. 2. Model architectures for the MoE (left panel) and MMoE (right panel) models. In the MoE model, inputs $\boldsymbol{x}_i$ are processed by both the gating and expert functions. The gating function determines the classification probabilities $\pi_j := \pi_j(\boldsymbol{x}_i;\boldsymbol{\alpha})$ $(j = 1,\dots,g)$ for each of the g expert functions. In the proposed MMoE model, the random effects $\boldsymbol{w}_i$, drawn from a distribution $\boldsymbol{\Phi}$, together with the inputs $\boldsymbol{x}_i$, are fed exclusively into the gating functions.

$\tilde{\boldsymbol{\theta}}$, we also define $\boldsymbol{w} = \{w_l^{(s)}\}_{l=1,\dots,L;s=1,\dots,S_l}$ as the random effects across all levels and factors.

The distribution function of $\boldsymbol{y}_i$ conditional on $\boldsymbol{x}_i$ and $\boldsymbol{w}_i$ is given by

$$F(\boldsymbol{y}_i;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\psi},g|\boldsymbol{x}_i,\boldsymbol{w}_i) = \sum_{j=1}^{g} \pi_j(\boldsymbol{x}_i,\boldsymbol{w}_i;\boldsymbol{\alpha},\boldsymbol{\beta}) F_0(\boldsymbol{y}_i;\boldsymbol{\psi}_j), \tag{6}$$

where g is the number of latent classes, $\pi_j(\boldsymbol{x}_i,\boldsymbol{w}_i;\boldsymbol{\alpha},\boldsymbol{\beta})$ is the mixing weight for the jth class (called the gating function) parameterized by $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, and $F_0(\boldsymbol{y}_i;\boldsymbol{\psi}_j)$ is the multivariate distribution function of $\boldsymbol{y}_i$ for the jth class (called the expert function) parameterized by $\boldsymbol{\psi}_j$. Here, we denote $\boldsymbol{\alpha} = \{\alpha_{j0},\boldsymbol{\alpha}_j : j = 1,\dots,g\} \in \mathcal{A}$ as the regression parameters of the gating function, $\boldsymbol{\beta} = \{\boldsymbol{\beta}_j : j = 1,\dots,g\} \in \mathcal{B}$ as the coefficients of the random effects, and $\boldsymbol{\psi} = \{\boldsymbol{\psi}_j : j = 1,\dots,g\} \in \boldsymbol{\Psi}$ as the parameters of the expert functions, where $\mathcal{A}$, $\mathcal{B}$ and $\boldsymbol{\Psi}$ are defined respectively as the parameter spaces for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$. Moreover, we specify $\pi_j(\boldsymbol{x}_i,\boldsymbol{w}_i;\boldsymbol{\alpha},\boldsymbol{\beta})$ as a logit linear gating function, given by

$$\pi_j(\boldsymbol{x}_i,\boldsymbol{w}_i;\boldsymbol{\alpha},\boldsymbol{\beta}) = \frac{\exp\{\alpha_{j0} + \boldsymbol{\alpha}_j^T\boldsymbol{x}_i + \boldsymbol{\beta}_j^T\boldsymbol{w}_i\}}{\sum_{j'=1}^{g}\exp\{\alpha_{j'0} + \boldsymbol{\alpha}_{j'}^T\boldsymbol{x}_i + \boldsymbol{\beta}_{j'}^T\boldsymbol{w}_i\}}, \quad j = 1,2,\dots,g. \tag{7}$$

Apart from that, the random effects $\{w_l^{(s)}\}_{l=1,\dots,L;s=1,\dots,S_l}$ are assumed to be independent across l and s, and $w_l^{(s)}$ follows a fixed pre-specified distribution $\boldsymbol{\Phi}_l$ with no extra parameters in it. Based on the above model specification, the joint distribution of $\boldsymbol{y}$ given $\boldsymbol{x}$ is

$$\tilde{F}(\boldsymbol{y};\boldsymbol{x}) := \tilde{F}(\boldsymbol{y};\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\Psi},g|\boldsymbol{x}) = \int \prod_{i=1}^{N} F(\boldsymbol{y}_i;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\Psi}|\boldsymbol{x}_i,\boldsymbol{w}_i) d\boldsymbol{\Phi}(\boldsymbol{w}), \tag{8}$$

where $\boldsymbol{\Phi}$ is the joint distribution of $\boldsymbol{w}$, given by

$$\boldsymbol{\Phi}(\boldsymbol{w}) = \prod_{l=1}^{L}\prod_{s=1}^{S_l}\boldsymbol{\Phi}_l(w_l^{(s)}) \quad \text{or} \quad d\boldsymbol{\Phi}(\boldsymbol{w}) = \prod_{l=1}^{L}\prod_{s=1}^{S_l}\boldsymbol{\Phi}_l(dw_l^{(s)}). \tag{9}$$

The model can be interpreted as follows with a visual illustration displayed in the right panel of Fig. 2. Each observation is assigned into one of the g homogeneous subgroups with classification probabilities $\pi_j(\boldsymbol{x}_i,\boldsymbol{w}_i;\boldsymbol{\alpha},\boldsymbol{\beta})$. The classification probabilities vary among observations as they depend on both inputs $\boldsymbol{x}_i$ and unobserved factors $\boldsymbol{w}_i$. Conditioned on the subgroup observation i belongs to, the outputs $\boldsymbol{y}_i$ are governed by a homogeneous probability distribution $F_0(\boldsymbol{y}_i;\boldsymbol{\psi}_j)$ independent of $\boldsymbol{x}_i$ and $\boldsymbol{w}_i$.

Example 3. The hypothetical data in Table 1 was generated using the proposed MMoE model with two subgroups ($g = 2$). The parameters are set as follows: $\alpha_{1,0} = -5.5$, $\boldsymbol{\alpha}_1 = (0,0.5)$, $\boldsymbol{\beta}_1 = (0.5,0.5)$, $\alpha_{2,0} = 0$, $\boldsymbol{\alpha}_2 = (0,0)$, and $\boldsymbol{\beta}_2 = (0,0)$. Both $\boldsymbol{\Phi}_1$ and $\boldsymbol{\Phi}_2$ are assumed to follow a standard normal distribution. The distributions of the outputs for the subgroups are modeled as $F_0(\boldsymbol{y}_i;\boldsymbol{\psi}_1) \sim \text{MVN}(\boldsymbol{\mu} = (80,80), \boldsymbol{\Sigma} = \text{diag}(5,5))$ and $F_0(\boldsymbol{y}_i;\boldsymbol{\psi}_2) \sim \text{MVN}(\boldsymbol{\mu} = (65,65), \boldsymbol{\Sigma} = \text{diag}(10,10))$, where "MVN" denotes a multivariate normal distribution, $\boldsymbol{\mu}$ is the mean vector, $\boldsymbol{\Sigma}$ is the covariance matrix, and "diag" indicates a diagonal matrix. The model can be interpreted as follows:

1. The two subgroups likely represent "good students" (subgroup 1) and "bad students" (subgroup 2), as indicated by the higher average score in subgroup 1 (80) compared to subgroup 2 (65). If it is known that student i is "good" (resp. "bad"), then $F_0(\boldsymbol{y}_i;\boldsymbol{\psi}_1)$ (resp. $F_0(\boldsymbol{y}_i;\boldsymbol{\psi}_2)$) represents the probability distribution of student i's exam scores.
2. The parameters $\boldsymbol{\alpha}$ indicate how gender and income influence the likelihood of a student being classified as "good" or "bad". The values $\boldsymbol{\alpha}_1 = (0,0.5)$ suggest that gender does not influence the probability of a student being classified as "good" or "bad". However, students from wealthier families are more likely to be categorized as "good students" (subgroup 1).
3. Students attending the same school (level-1 random effect w_{i1}) or classroom (level-2 random effect w_{i2}) share the same random effects, making them more likely to be classified into the same category. For instance, both students 11 and 12, who belong to the same school and classroom, are classified as "bad students" and have lower exam scores. The magnitudes of $\boldsymbol{\beta}$ govern the likelihood that students from the same school or classroom will be categorized within the same group.

4.2. Comparisons to the literature

Our proposed MMoE model extends the standard MoE model (e.g., Jacobs et al. [1]) described in Eq. (1) to account for multilevel data by incorporating random effects $\boldsymbol{w}_i$ into the gating function. If the coefficients of the random effects in Eq. (6) are set to $\boldsymbol{\beta} = \boldsymbol{0}$, the proposed MMoE model simplifies back to the standard MoE model.

Our proposed MMoE model builds on the limited literature concerning mixed effects MoE models, such as Yau et al. [43], Ng and McLachlan [44], and Ng and McLachlan [45], which only account for a single level of random effects (i.e., $L = 1$). By incorporating multiple levels of random effects, our model addresses the complex multilevel data structures seen in contexts like the school problem. Additionally, our approach diverges from existing mixed effects MoE models in two key ways that simplify the model structure. First, we eliminate the regression relationship in the expert functions, meaning they do not depend on the inputs $\boldsymbol{x}_i$ (see also Fig. 2), following the reduced MoE

(RMoE) model concept introduced by Fung et al. [30]. Second, our model assumes that the level-l random effect w_{il} is consistent across all g gating functions (i.e., w_{il} does not vary with j), contrasting with Ng and McLachlan [44], which allows for varying and independent level-l random effects across gating functions.

These model simplifications offer several advantages. First, they allow for a broader selection of probability distributions as the expert function F_0, including more complex non-exponential distributions (e.g., phase-type distributions) where regression modeling may be impractical or computationally intensive. Second, the simplified structure enhances interpretability, as our model can cluster observations into homogeneous subgroups and explain the variability of each level-l factor through a single source (i.e., $w_{il} \in \mathbb{R}$) rather than the multiple sources considered by Ng and McLachlan [44]. Third, the reduction in model parameters significantly decreases the computational burden during parameter estimation. Specifically, our follow-up work [46], which employs a variational ECM algorithm for parameter estimation, identifies that the primary computational challenge lies in simulating and computing the posterior samples of the random effects $\boldsymbol{w}_i$ given the observed data. By reducing the dimension of random effects from $\mathbb{R}^{2g-1}$ [44] or $\mathbb{R}^{gK}$ [45] to $\mathbb{R}$ for each level, our estimation algorithm becomes significantly more computationally efficient and scalable to large datasets.

The remaining issue is: how does such a reduced structure affect its model flexibility? To justify the proposed model, we will demonstrate the denseness property in the following section, meaning that the MMoE model structure of Eq. (8) can approximate any generalized form of mixed effect models expressed by Eq. (3). This will provide evidences suggesting that our proposed model is parsimonious. In other words, the MMoE has the simplest structure without harming its representation capability.

5. Denseness theory

This section studies the approximation capability of the class of MMoE models. Our goal is to show that the proposed MMoE is versatile enough to approximate any mixed effects models under mild regularity conditions, even if the MMoE is constructed in a reduced form: (i) the gating function is restricted to be a logit linear gating; (ii) regression link is removed in the expert functions; (iii) the random effects are restricted to follow some fixed pre-determined distributions. Before that, we need to technically formulate a class of mixed effects models and define "denseness" for mixed effects models. These definitions are the extensions of Fung et al. [30], who defines "regression distributions" and "denseness" in the regression settings without considering random effects.

Following the notations defined in Section 3, for $l = 1, \dots, L$, let $\mathcal{T}_l$ denote a collection of some spaces of level-l random effects $\boldsymbol{\Theta}_l$, and $\mathcal{G}_l$ denote a collection of probability measures G_l on $\boldsymbol{\theta}_l^{(s)}$. We also denote $\mathcal{T} := \mathcal{T}_1 \times \cdots \times \mathcal{T}_L$ as a collection of product spaces $\boldsymbol{\Theta} := \boldsymbol{\Theta}_1 \times \cdots \times \boldsymbol{\Theta}_L$, $\mathcal{H}$ as a collection of some distribution functions H on $(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta}_i)$, and $\mathcal{G} := \mathcal{G}_1 \times \cdots \times \mathcal{G}_L$ as a collection of product probability measures G on $\tilde{\boldsymbol{\theta}}$. Also, let $\mathcal{C}$ be a set containing all possible mappings $\boldsymbol{c}(\cdot)$, and define a vector $\boldsymbol{S} = (S_1, \dots, S_L)$ with $\mathcal{S} = \mathbb{N}^L$. "A class of mixed effects models" and "mixed effects distributions" are first defined as follows:

Definition 1. A class of mixed effects models $\mathcal{M}_L(\mathcal{X}; \mathcal{T}, \mathcal{H}, \mathcal{G}) := \{\tilde{H}(\cdot; \mathcal{X}; \boldsymbol{\Theta}, H, G) : \boldsymbol{\Theta}_l \in \mathcal{T}_l, H \in \mathcal{H}, G_l \in \mathcal{G}_l, l = 1, \dots, L\}$ is a collection of mixed effects distributions $\tilde{H}(\cdot; \mathcal{X}; \boldsymbol{\Theta}, H, G)$, where each mixed effects distribution $\tilde{H}(\cdot; \mathcal{X}; \boldsymbol{\Theta}, H, G) := \{\tilde{H}(\boldsymbol{y}|\boldsymbol{x}) := \tilde{H}(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\Theta}, H, G) = \int_{\tilde{\boldsymbol{\Theta}}} \left[\prod_{i=1}^{N} H(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta}_i)\right] dG(\tilde{\boldsymbol{\theta}}) : \boldsymbol{x}_i \in \mathcal{X}, i \in \{1, \dots, N\}, N \in \mathbb{N}, \boldsymbol{S} \in \mathcal{S}, \boldsymbol{c} \in C\}$ is itself a collection of joint probability distributions.

In simpler terms, a "mixed effects distribution" refers to the collection of all possible joint probability distributions, as described by Eq. (4), that can be generated by a mixed effects model outlined in Section 3, based on a specified H and a specified G. It is important to note that a single mixed effects model can produce various possible joint probability distributions since Eq. (4) depends on the covariates $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_N)$, sample size N, and the mapping $\boldsymbol{c}(\cdot)$. "A class of mixed effects models" refers to the set of all possible "mixed effects distributions" that can be generated by varying $\boldsymbol{\Theta}$, H, and G.

In the spirit of Fung et al. [30], denseness is defined in the sense of weak convergence of probability distributions. Therefore, before defining denseness, we need to define weak convergence of mixed effects distributions. Let $\{\boldsymbol{\Theta}^{(n)}\}_{n=1,2,\dots}$ denote a sequence of (product) spaces of random effects, $\{H^{(n)}\}_{n=1,2,\dots}$ denote a sequence of distribution functions on $\boldsymbol{y}_i$ given $\boldsymbol{x}_i$ and $\boldsymbol{\theta}_i$, and $\{G^{(n)}\}_{n=1,2,\dots}$ denote a sequence of product probability measures on $\tilde{\boldsymbol{\theta}}$. The definition is as follows:

Definition 2. Consider a sequence of mixed effects distributions $\tilde{\boldsymbol{H}}^{(n)} := \tilde{H}(\cdot; \mathcal{X}; \boldsymbol{\Theta}^{(n)}, H^{(n)}, G^{(n)})$ and a target mixed effects distribution $\tilde{\boldsymbol{H}} := \tilde{H}(\cdot; \mathcal{X}; \boldsymbol{\Theta}, H, G)$. We say that $\{\tilde{\boldsymbol{H}}^{(n)}\}_{n=1,2,\dots}$ weakly converges to $\tilde{\boldsymbol{H}}$ if and only if, for every $\boldsymbol{x}_i \in \mathcal{X}$ (for all $i \in \{1, \dots, N\}$), $N \in \mathbb{N}$, $\boldsymbol{S} \in \mathcal{S}$, and $\boldsymbol{c} \in C$, we have $\tilde{H}(\cdot|\boldsymbol{x}; \boldsymbol{\Theta}^{(n)}, H^{(n)}, G^{(n)}) \xrightarrow{D} \tilde{H}(\cdot|\boldsymbol{x}; \boldsymbol{\Theta}, H, G)$ as $n \to \infty$, where $\xrightarrow{D}$ denotes weak convergence or convergence in distribution. If the convergence in distribution is uniform across any compact input space $\tilde{\mathcal{X}} \subseteq \mathcal{X}$, meaning that $\tilde{H}(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\Theta}^{(n)}, H^{(n)}, G^{(n)}) \to \tilde{H}(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\Theta}, H, G)$ uniformly on $(\boldsymbol{y}, \boldsymbol{x})$ with $\boldsymbol{x}_i \in \tilde{\mathcal{X}}$ (for all $i \in \{1, \dots, N\}$), then we say that $\{\tilde{\boldsymbol{H}}^{(n)}\}_{n=1,2,\dots}$ weakly converges to $\tilde{\boldsymbol{H}}$ compactly.

In essence, the definition above requires that all distributions produced by the mixed effects models converge in order to assert that the "mixed effects distributions" themselves converge. We are now able to extend the formalism of Fung et al. [30] and define denseness in the settings of mixed effects models. Similar as above, denote $\mathcal{T}'$, $\mathcal{H}'$ and $\mathcal{G}'$, respectively, as the collections of $\boldsymbol{\Theta}$, H and G. The definition is as follows:

Definition 3. Consider two classes of mixed effects models $\mathcal{M}_L := \mathcal{M}_L(\mathcal{X}; \mathcal{T}, \mathcal{H}, \mathcal{G})$ and $\mathcal{M}'_L := \mathcal{M}_L(\mathcal{X}; \mathcal{T}', \mathcal{H}', \mathcal{G}')$. $\mathcal{M}_L$ is dense in $\mathcal{M}'_L$ if and only if for all $(\boldsymbol{\Theta}, H, G) \in \mathcal{T}' \times \mathcal{H}' \times \mathcal{G}'$, there exists a sequence of $\{(\boldsymbol{\Theta}^{(n)}, H^{(n)}, G^{(n)})\}_{n=1,2,\dots}$ with $(\boldsymbol{\Theta}^{(n)}, H^{(n)}, G^{(n)}) \in \mathcal{T} \times \mathcal{H} \times \mathcal{G}$ such that the mixed effects distributions $\{\tilde{\boldsymbol{H}}^{(n)} := \tilde{H}(\cdot; \mathcal{X}; \boldsymbol{\Theta}^{(n)}, H^{(n)}, G^{(n)})\}_{n=1,2,\dots}$ weakly converge to $\tilde{\boldsymbol{H}} := \tilde{H}(\cdot; \mathcal{X}; \boldsymbol{\Theta}, H, G)$. If $\{\tilde{\boldsymbol{H}}^{(n)}\}_{n=1,2,\dots}$ weakly converge to $\tilde{\boldsymbol{H}}$ compactly, then $\mathcal{M}_L$ is said to be compactly dense in $\mathcal{M}'_L$.

It is evident that if $\mathcal{M}_L$ includes $\mathcal{M}'_L$, then $\mathcal{M}_L$ is dense in $\mathcal{M}'_L$. However, if $\mathcal{M}_L$ is a smaller class than $\mathcal{M}'_L$, the definition of denseness implies that any mixed effects model in $\mathcal{M}'_L$ can be represented or closely approximated by models within the more extensive class $\mathcal{M}_L$. Put more simply, $\mathcal{M}_L$ can be seen as a model class that is, in practice, at least as rich or flexible as $\mathcal{M}'_L$ even if it is a theoretically smaller set.

With all the necessary definitions established, we now introduce a class of generalized mixed effects models denoted by $\mathcal{M}_L^{\text{gen}}(\mathcal{X}) := \mathcal{M}_L(\mathcal{X}; \mathcal{T}^{\text{gen}}, \mathcal{H}^{\text{gen}}, \mathcal{G}^{\text{gen}})$. Here, $\mathcal{T}^{\text{gen}} := \mathcal{T}_1^{\text{gen}} \times \cdots \times \mathcal{T}_L^{\text{gen}}$, $\mathcal{T}_l^{\text{gen}}$ $(l = 1, \dots, L)$, $\mathcal{H}^{\text{gen}}$, and $\mathcal{G}^{\text{gen}}$ represents the collections of all possible $\boldsymbol{\Theta}$, $\boldsymbol{\Theta}_l$, H, and G that satisfy the following two mild technical assumptions:

Assumption 2. Each space $\boldsymbol{\Theta}_l \in \mathcal{T}_l^{\text{gen}}$ is equipped by a complete separable metric $d_{\boldsymbol{\Theta}_l}$.

Assumption 3. For every probability distribution functions $H \in \mathcal{H}^{\text{gen}}$, $H(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta}_i)$ is continuous with respect to $(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\theta}_i)$.

Assumption 2 is a mathematical framework designed to ensure the rigor of the theoretical results and has no practical implications or constraints. Although Assumption 3 requires H to be a continuous distribution, it can be readily adapted to accommodate discrete distributions. For further details, see Remark 2.

Next, we consider the class of MMoE models with a predetermined expert function F_0 and joint distribution of random effects $\boldsymbol{\Phi}$

defined in Section 4. We denote this class as $\mathcal{M}_L^{\text{MMoE}}(\mathcal{X}; F_0, \boldsymbol{\Phi}) := \mathcal{M}_L(\mathcal{X}; \mathcal{T}^{\text{MMoE}}, \mathcal{H}^{\text{MMoE}}, \mathcal{G}^{\text{MMoE}})$. Here, $\mathcal{T}^{\text{MMoE}} := \mathcal{T}_1^{\text{MMoE}} \times \cdots \times \mathcal{T}_L^{\text{MMoE}}$ represents the set of all possible spaces for MMoE random effects $\boldsymbol{w}$, where $\mathcal{T}_l^{\text{MMoE}}$ contains the possible spaces for $w_l^{(s)}$ for level $l \in \{1, \dots, L\}$. The set $\mathcal{H}^{\text{MMoE}}$ contains all conditional distributions given by Eq. (6), and $\mathcal{G}^{\text{MMoE}}$ encompasses all possible distributions of the random effects $\boldsymbol{w}$. Given that the proposed MMoE model assumes a scalar random effect for each level, with $\mathbb{R}$ as its domain, it follows that $\mathcal{T}_l^{\text{MMoE}} = \{\mathbb{R}\}$ and $\mathcal{T}^{\text{MMoE}} = \{\mathbb{R}^L\}$ each contain only a single element. Additionally, since the distribution of $\boldsymbol{w}$ is fixed as $\boldsymbol{\Phi}$ according to Eq. (9), we have $\mathcal{G}^{\text{MMoE}} = \{\boldsymbol{\Phi}\}$, also containing just one element. Moreover, we can express $\mathcal{H}^{\text{MMoE}} = \{F(\boldsymbol{y}_i; \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\psi}, g | \boldsymbol{x}_i, \boldsymbol{w}_i) : \boldsymbol{\alpha} \in \mathcal{A}, \boldsymbol{\beta} \in \mathcal{B}, \boldsymbol{\psi} \in \boldsymbol{\Psi}, g \in \mathbb{N}\}$, where $F(\boldsymbol{y}_i; \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\psi}, g | \boldsymbol{x}_i, \boldsymbol{w}_i)$ takes the form given in Eq. (6). The following two conditions on the choices of F_0 and $\boldsymbol{\Phi}$ are crucial for establishing the denseness property of the MMoE model class:

Assumption 4. F_0 satisfies the denseness condition outlined by Proposition 3.1 of Fung et al. [30], meaning that for every $\boldsymbol{q} \in \mathbb{R}^K$, there exists a sequence of parameters $\{\boldsymbol{\psi}^{(n)}(\boldsymbol{q})\}_{n=1,2,\dots}$ such that $F_0(\cdot; \boldsymbol{\psi}^{(n)}(\boldsymbol{q})) \xrightarrow{D} \boldsymbol{q}$ as $n \to \infty$.

Assumption 5. $\boldsymbol{\Phi}_l$ is a continuous distribution function for every $l = 1, \dots, L$.

The above two assumptions are not necessarily mild. As discussed in Fung et al. [30], some common distributions, such as Pareto and exponential distributions, do not satisfy the denseness condition under Assumption 4. Moreover, Assumption 5 does not hold whenever we choose any discrete distributions for $\boldsymbol{\Phi}_l$. Nonetheless, note that the expert function F_0 and random effect distribution $\boldsymbol{\Phi}_l$ are both predetermined, so we have the control to choose suitable functions that fulfill Assumptions 4 and 5 before modeling a multilevel dataset via the MMoE. For example, one may choose Gamma, Weibull, log-normal, or inverse-Burr distributions as the expert function F_0 [30], and select a normal distribution as the random effect distribution $\boldsymbol{\Phi}_l$.

We now introduce the main result that validates the representational power of the proposed MMoE model class. The detailed technical proof is provided in Appendix A, while Section 7 offers intuitive explanations of the key steps in the proof to aid in understanding the core concepts.

Theorem 1. *Suppose that Assumptions 1 to 5 are satisfied. Then, $\mathcal{M}_L^{MMoE}(\mathcal{X}; F_0, \boldsymbol{\Phi})$ is compactly dense in $\mathcal{M}_L^{gen}(\mathcal{X})$.*

Remark 2. It is also important to investigate into the above theorem for discrete distributions (see, e.g., Jiang and Tanner [23],Fung et al. [30]) instead of continuous distributions. As discussed by Norets et al. [25], any discrete distribution can be represented by a continuous latent distribution. Then, it is obvious that Theorem 1 still holds for discrete distributions if the denseness condition in Assumption 4 is changed from $\boldsymbol{q} \in \mathbb{R}^K$ to $\boldsymbol{q} \in \mathbb{N}^K$.

Theorem 1 demonstrates that the proposed MMoE model, described in Eq. (8), can closely approximate any mixed effects models in the form of Eq. (4), which has not been addressed by the existing literature on multilevel MoE models, including Ng and McLachlan [44, 45]. This denseness property offers a theoretical foundation for the MMoE model's flexibility in capturing a wide range of model features, including joint distributions (such as multimodal distributions and dependencies among outputs), regression patterns (like non-linear links and interactions among covariates), random intercepts (reflecting unique effects of unobserved factors on the outputs), and random slopes (accounting for interactions between covariates and random effects). As discussed in Section 4, our proposed model features a reduced structure. Therefore, the result in Theorem 1, which provides approximations within a narrower class of models, is remarkably stronger compared to proving the result based on a broader class of unreduced MMoE models, such as those proposed by Ng and McLachlan [44,45]. Additionally, the denseness property suggests the parsimony of the proposed model.

6. Nested mixed effects models for hierarchical data

A nested mixed effects model is a specific case of the model discussed in Section 3, involving multiple random effects. In nested models, the L levels of factors in a multilevel dataset must be arranged hierarchically, with the first level representing the highest level and the Lth level representing the lowest. In this structure, observations that share the same unit at a lower level must also share the same unit at the higher level. If the factors across levels are not nested, the resulting model is referred to as a "crossed" mixed effects model.

Before defining a nested model, it is helpful to introduce an alternative notation for identifying observations. Denote $\boldsymbol{i} = (i_1, i_2, \dots, i_{L+1})$ as an identifier of an observation, and $\boldsymbol{i}_l = (i_1, i_2, \dots, i_l)$ as an identifier up to level l. Here, i_1 indicates the label of the level-1 unit to which the observation belongs. For $l = 2, 3, \dots, L$, i_l serves as the label for the level-l unit, given that the labels for the first $(l-1)$ levels are $\boldsymbol{i}_{l-1}$. Finally, i_{L+1} represents the label of the observation itself, given that the L units corresponding to the L levels of factors are labeled by $\boldsymbol{i}_L$.

A classic example of nested factors is the school problem (see, e.g., Aitkin and Longford [32]), where two levels of factors influence a student's performance: school and classroom. Here, since students in the same classroom must belong to the same school, we have $L = 2$, with "school" as the first level and "classroom" as the second. For example, in the school problem with $L = 2$, $\boldsymbol{i} = (2, 3, 5)$ refers to the fifth student in the third classroom of the second school.

Furthermore, denote N_0 as the number of possible level-1 factors, so that the support of i_1 is given by $\mathcal{I}_1 := \{1, \dots, N_0\}$. For $l = 2, 3, \dots, L$, define $N_{\boldsymbol{i}_{l-1}}$ as the number of possible level-l factors where the first $(l-1)$ factor levels are labeled as $\boldsymbol{i}_{l-1}$, such that the support of $\boldsymbol{i}_l$ is $\mathcal{I}_l := \{\boldsymbol{i}_l : \boldsymbol{i}_{l-1} \in \mathcal{I}_{l-1}, i_l = 1, \dots, N_{\boldsymbol{i}_{l-1}}\}$. Similarly, $N_{\boldsymbol{i}_L}$ is the number of observations having $\boldsymbol{i}_L$ as the label of the L factors. Then, the support of $\boldsymbol{i}$ is $\mathcal{I} := \{\boldsymbol{i} : \boldsymbol{i}_L \in \mathcal{I}_L, i_{L+1} = 1, \dots, N_{\boldsymbol{i}_L}\}$. Also, the total number of observations is given by $N = \sum_{i_1=1}^{N_0} \sum_{i_2=1}^{N_{i_1}} \cdots \sum_{i_L=1}^{N_{i_{L-1}}} N_{\boldsymbol{i}_L}$. Hence, the nested multilevel dataset is given by $(\boldsymbol{y}, \boldsymbol{x}) := \{(\boldsymbol{y}_{\boldsymbol{i}}, \boldsymbol{x}_{\boldsymbol{i}})\}_{\boldsymbol{i} \in \mathcal{I}}$, where $\boldsymbol{x}_{\boldsymbol{i}} \in \mathcal{X}$ and $\boldsymbol{y}_{\boldsymbol{i}} \in \mathcal{Y}$ are respectively the input and output of an observation labeled as $\boldsymbol{i} \in \mathcal{I}$.

Example 4. The school problem serves as an example of nested factors. As seen in Table 1, students in the same classroom must also belong to the same school. Therefore, we have $L = 2$, where "school" and "classroom" are the first and second levels of factors, respectively. In this dataset, there are $N_0 = 4$ schools, and the number of classrooms in each school is $(N_1, N_2, N_3, N_4) = (3, 5, 2, 2)$. Observations are relabeled from i to $\boldsymbol{i}$ to reflect this structure. For instance, student $i = 7$ is relabeled as $\boldsymbol{i} = (2, 3, 1)$, indicating the first student in the third classroom (classroom #5) of the second school (school #2). The support set for this data is $\mathcal{I} = \{(1,1,1), (1,1,2), (1,2,1), (2,1,1), (2,1,2), (2,2,1), (2,3,1), (2,3,2), (3,1,1), (3,2,1), (4,1,1), (4,1,2)\}$. To summarize the above descriptions, Fig. 3 provides a visual representation of the hierarchical data structure within the context of the school problem.

In this section, we will define the generalized class of nested mixed effects models, formulate the proposed MMoE model with nested random effects, and construct the denseness theory for the class of nested MMoE.

6.1. Generalized nested mixed effect model

Similar to the MMoE defined in Section 3, under the nested MMoE, the response $\boldsymbol{y}_{\boldsymbol{i}}$ depends on its covariates $\boldsymbol{x}_{\boldsymbol{i}}$ and L levels of latent random effects denoted as $\boldsymbol{\theta}_{i_1}, \dots, \boldsymbol{\theta}_{\boldsymbol{i}_L}$. The dependence assumption among the latent factors is stated as follows:

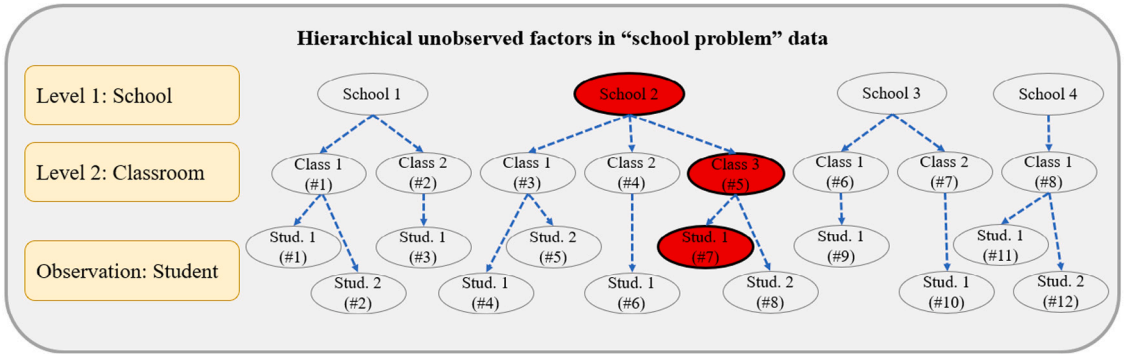


Fig. 3. Hierarchical data structure for a hypothetical "school problem" dataset, illustrating two nested levels of factors: school and classroom. Note that "Stud". means "Student". The units shaded in darker color (red) indicate the specific school and classroom to which a student belongs. In this example, the highlighted units correspond to student #7, who is the first student in the third classroom (classroom #5) within the second school (school #2). The dashed arrows depict the hierarchical organization of the data, indicating that students from the same classroom must also belong to the same school.

Assumption 6. Conditioned on $\theta_{i_1}, \dots, \theta_{i_L}$, the N observations are independent. The set of parents of θ_{i_l} is given by $\text{pa}(\theta_{i_l}) = (\theta_{i_1}, \dots, \theta_{i_{(l-1)}})$ for $l = 1, \dots, L$, meaning that θ_{i_l} may depend only on $(\theta_{i_1}, \dots, \theta_{i_{(l-1)}})$.

In other words, the lower level factor only depends directly on its corresponding upper level factors. Under a nested hierarchical data structure, we are able to relax the independence assumption in Assumption 1 by allowing the dependence of lower level random effects on their parents (upper level effects). For $l = 1, \dots, L$, we also denote G_l as the distribution of the level-l factor conditioned on $\text{pa}(\theta_{i_l})$.

The joint distribution of $\boldsymbol{y}$ given $\boldsymbol{x}$ is given by

$$\tilde{H}(\boldsymbol{y}|\boldsymbol{x}) = \int_{\tilde{\Theta}} \left[\prod_{i \in \mathcal{I}} H(\boldsymbol{y}_i; \boldsymbol{x}_i | \theta_i) \right] dG(\tilde{\theta}) \tag{10}$$

with

$$G(\tilde{\theta}) = \prod_{i_1=1}^{N_0} G_1(\theta_{i_1}) \prod_{i_2=1}^{N_{i_1}} G_2(\theta_{i_2}|\theta_{i_1}) \cdots \prod_{i_L=1}^{N_{i_{L-1}}} G_L(\theta_{i_L}|\theta_{i_1}, \dots, \theta_{i_{L-1}}), \tag{11}$$

where $\tilde{\theta} = \{\theta_{i_l} : i_l \in \mathcal{I}_l, l = 1, \dots, L\}$.

6.2. Nested mixed mixture of experts model

Analogous to the MMoE introduced in Section 4, we construct a nested MMoE for hierarchical data. Denote $\boldsymbol{w}_i = (w_{i_1}, \dots, w_{i_L}) \in \mathbb{R}^L$ as the L levels of random effects of observation $\boldsymbol{i}$. Also, define $\boldsymbol{w} = \{w_{i_l}\}_{i_l \in \mathcal{I}_l, l=1,\dots,L}$ as all the random effects aggregated across all observations. With a slight change of notations from Eqs. (6) and (7), the distribution function of $\boldsymbol{y}_i$ conditional on $\boldsymbol{x}_i$ and $\boldsymbol{w}_i$ is then

$$F(\boldsymbol{y}_i; \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\psi}, g | \boldsymbol{x}_i, \boldsymbol{w}_i) = \sum_{j=1}^{g} \pi_j(\boldsymbol{x}_i, \boldsymbol{w}_i; \boldsymbol{\alpha}, \boldsymbol{\beta}) F_0(\boldsymbol{y}_i; \boldsymbol{\psi}_j), \tag{12}$$

where the gating function given by

$$\pi_j(\boldsymbol{x}_i, \boldsymbol{w}_i; \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{\exp\{\alpha_{j0} + \boldsymbol{\alpha}_j^T \boldsymbol{x}_i + \boldsymbol{\beta}_j^T \boldsymbol{w}_i\}}{\sum_{j'=1}^{g} \exp\{\alpha_{j'0} + \boldsymbol{\alpha}_{j'}^T \boldsymbol{x}_i + \boldsymbol{\beta}_{j'}^T \boldsymbol{w}_i\}}, \quad j = 1, 2, \dots, g. \tag{13}$$

Similar to Section 4, the random effects $\{w_{i_l}\}_{l=1,\dots,L; i_l=1,\dots,N_{i_{l-1}}}$ are constructed to be independent across l and i_l with $w_{i_l} \sim \boldsymbol{\Phi}_l$. This construction is simplified from Assumption 6 of the generalized nested mixed models where the random effects may depend on their parents. Adapting from Eqs. (8) and (9), the joint distribution of $\boldsymbol{y}$ given $\boldsymbol{x}$ is

$$\tilde{F}(\boldsymbol{y}; \boldsymbol{x}) := \tilde{F}(\boldsymbol{y}; \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Psi}, g | \boldsymbol{x}) = \int \prod_{i \in \mathcal{I}} F(\boldsymbol{y}_i; \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Psi} | \boldsymbol{x}_i, \boldsymbol{w}_i) d\boldsymbol{\Phi}(\boldsymbol{w}), \tag{14}$$

where $\boldsymbol{\Phi}$ is the joint distribution of $\boldsymbol{w}$ given by

$$\boldsymbol{\Phi}(\boldsymbol{w}) = \prod_{l=1}^{L} \prod_{i_l \in \mathcal{I}_l} \boldsymbol{\Phi}_l(w_{i_l}) \quad \text{or} \quad d\boldsymbol{\Phi}(\boldsymbol{w}) = \prod_{l=1}^{L} \prod_{i_l \in \mathcal{I}_l} \boldsymbol{\Phi}_l(dw_{i_l}). \tag{15}$$

From Eq. (15) above, the random effects are still assumed to be independent under the nested MMoE. However, we will show in the following subsection that such a specification suffices to approximate the dependence of lower level random effects on their parents under a hierarchical data structure.

6.3. Denseness theory for nested MMoE

Analogous to Section 5, it is desirable to develop an approximation theory for the nested MMoE in the space of the generalized nested mixed effect models. Denote $\boldsymbol{N} = (N_0, \{N_{i_1}\}_{i_1 \in \mathcal{I}_1}, \{N_{i_2}\}_{i_2 \in \mathcal{I}_2}, \dots, \{N_{i_L}\}_{i_L \in \mathcal{I}_L})$ as the number of factors belonging to each parent factors for each level with $N_0 \in \mathbb{N}$ and $N_{i_l} \in \mathbb{N}$ for $l = 1, \dots, L$, and $\mathcal{N}$ contains all combinations of possible $\boldsymbol{N}$. Other notations, unless specified otherwise, are consistent to those defined by Section 5. The equivalent definitions analogous to Section 5 for hierarchical data structure are listed as follows:

Definition 4. A class of nested mixed effects models $\mathcal{M}_L(\mathcal{X}; \mathcal{T}, \mathcal{H}, \mathcal{G}) := \{\tilde{H}(\cdot; \mathcal{X}; \boldsymbol{\Theta}, H, G) : \boldsymbol{\Theta}_l \in \mathcal{T}_l, H \in \mathcal{H}, G_l \in \mathcal{G}_l, l = 1, \dots, L\}$ is a collection of nested mixed effects distributions $\tilde{H}(\cdot; \mathcal{X}; \boldsymbol{\Theta}, H, G)$, where each nested mixed effects distribution $\tilde{H}(\cdot; \mathcal{X}; \boldsymbol{\Theta}, H, G) := \{\tilde{H}(\boldsymbol{y}|\boldsymbol{x}) := \tilde{H}(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\Theta}, H, G) = \int_{\tilde{\Theta}} \left[\prod_{i \in \mathcal{I}} H(\boldsymbol{y}_i|\boldsymbol{x}_i, \theta_i)\right] dG(\tilde{\theta}) : \boldsymbol{x}_i \in \mathcal{X}, \boldsymbol{i} \in \mathcal{I}, \boldsymbol{N} \in \mathcal{N}\}$ is itself a collection of joint probability distributions.

Definition 5. Consider a sequence of nested mixed effects distributions $\tilde{\boldsymbol{H}}^{(n)} := \tilde{H}(\cdot; \mathcal{X}; \boldsymbol{\Theta}^{(n)}, H^{(n)}, G^{(n)})$ and a target nested mixed effects distribution $\tilde{\boldsymbol{H}} := \tilde{H}(\cdot; \mathcal{X}; \boldsymbol{\Theta}, H, G)$. We say that $\{\tilde{\boldsymbol{H}}^{(n)}\}_{n=1,2,\dots}$ weakly converge to $\tilde{\boldsymbol{H}}$ if and only if for every given $\boldsymbol{x}_i \in \mathcal{X}$ (for all $\boldsymbol{i} \in \mathcal{I}$), $\boldsymbol{N} \in \mathcal{N}$, we have $\tilde{H}(\cdot|\boldsymbol{x}; \boldsymbol{\Theta}^{(n)}, H^{(n)}, G^{(n)}) \xrightarrow{\mathcal{D}} \tilde{H}(\cdot|\boldsymbol{x}; \boldsymbol{\Theta}, H, G)$ as $n \to \infty$, where $\xrightarrow{\mathcal{D}}$ represents a weak convergence or convergence in distribution. If the distributional convergence is uniform across any compact input space $\tilde{\mathcal{X}} \subseteq \mathcal{X}$, i.e. $\tilde{H}(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\Theta}^{(n)}, H^{(n)}, G^{(n)}) \to \tilde{H}(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\Theta}, H, G)$ uniformly on $(\boldsymbol{y}, \boldsymbol{x})$ with $\boldsymbol{x}_i \in \tilde{\mathcal{X}}$ (for all $\boldsymbol{i} \in \mathcal{I}$), then we say that $\{\tilde{\boldsymbol{H}}^{(n)}\}_{n=1,2,\dots}$ weakly converge to $\tilde{\boldsymbol{H}}$ compactly.

The denseness definition for hierarchical data structure (nested mixed effects models) is exactly the same as Definition 3. Define $\mathcal{M}_L^{*\text{gen}}(\mathcal{X})$ as the class of generalized nested mixed effects models expressed in Eq. (10), subject to Assumptions 2 and 3. Also denote $\mathcal{M}_L^{*\text{MMoE}}(\mathcal{X}; F_0, \boldsymbol{\Phi})$ as the class of nested MMoE given by Eq. (14). We have the following denseness theorem:

Theorem 2. *Suppose that Assumptions 2 to 6 hold. Then, $\mathcal{M}_L^{*MMoE}(\mathcal{X}; F_0, \boldsymbol{\Phi})$ is compactly dense in $\mathcal{M}_L^{*gen}(\mathcal{X})$.*

The proof is leveraged to Appendix B. Theorem 2 suggests that the nested MMoE has a potential to approximate any generalized nested mixed effect models arbitrarily accurately, even if the random effects are restricted to be independent under the nested MMoE while the

random effects under the generalized nested mixed effect models can be dependent (Assumption 6). In contrast, under Theorem 1, the MMoE can only approximate mixed effect models with independent random effects (Assumption 1). Therefore, given that the data structure is hierarchical, Theorem 2 is a stronger theoretical result than Theorem 1.

7. Proof idea of Theorem 1

In this section, we provide a simplified illustration and intuitive explanation of the key steps involved in proving Theorem 1. The detailed technical proofs can be found in Appendix A. To demonstrate that the proposed MMoE class is dense in the space of generalized mixed effects models, we need to identify an MoE distribution $F(\boldsymbol{y}_i;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\Psi}|\boldsymbol{x}_i,\boldsymbol{w}_i)$ in Eq. (6) such that the MMoE joint distribution $\tilde{F}(\boldsymbol{y};\boldsymbol{x})$ described in Eq. (8) closely approximates the joint distribution $\tilde{H}(\boldsymbol{y}|\boldsymbol{x})$ given by Eq. (4), i.e.,

$$\int\left[\prod_{i=1}^{N}F(\boldsymbol{y}_i;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\Psi}|\boldsymbol{x}_i,\boldsymbol{w}_i)\right]d\Phi(\boldsymbol{w})\approx\int_{\tilde{\boldsymbol{\Theta}}}\left[\prod_{i=1}^{N}H(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}_i)\right]dG(\tilde{\boldsymbol{\theta}}) \tag{16}$$

The main challenge in finding a suitable $F(\boldsymbol{y}_i;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\Psi}|\boldsymbol{x}_i,\boldsymbol{w}_i)$ that satisfies Eq. (16) is that the two sides of the formula integrate two different spaces. Specifically, the left side of Eq. (16) involves a Lebesgue integration on Euclidean space, while the right side integrates with respect to an arbitrary abstract measure. One potential approach to address this issue is to apply an integral transformation on the right side. To ensure that Eq. (16) holds, we construct a mapping $M(\cdot):\tilde{\boldsymbol{\Theta}}\to\tilde{\mathcal{W}}$ and identify an appropriate MoE distribution such that $F(\boldsymbol{y}_i;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\Psi}|\boldsymbol{x}_i,\boldsymbol{w}_i)\approx H(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}_i)$ for all $i=1,\dots,N$ whenever $M(\tilde{\boldsymbol{\theta}})=\boldsymbol{w}$ with $\tilde{\boldsymbol{\theta}}\in\tilde{\boldsymbol{\Theta}}$ and $\boldsymbol{w}\in\tilde{\mathcal{W}}$, where $\tilde{\mathcal{W}}$ represents the space of $\boldsymbol{w}$. The technical details correspond to Appendix A.2 (Step 2). For simplicity, we consider only a single level of factors ($L=1$), where the random effect in the generalized mixed effects models is $\boldsymbol{\theta}_i=\boldsymbol{\theta}_{i1}$, and the MMoE random effect is $\boldsymbol{w}_i=w_{i1}$. Here, $\boldsymbol{\Theta}_1$ and $\mathcal{W}_1$ represent the spaces for $\boldsymbol{\theta}_{i1}$ and w_{i1} respectively. This result can be extended to cases with multiple levels of factors, with rigorous derivations provided in Appendices A.1 and A.2 (Steps 1 and 2).

To construct a mapping $M(\cdot)$ that effectively links the two spaces $\boldsymbol{\Theta}_1$ and $\mathcal{W}_1$, we propose discretizing both $\boldsymbol{\Theta}_1$ and $\mathcal{W}_1$ into D_1 subspaces. Let $\{\boldsymbol{\Theta}_{1,d_1}\}_{d_1=1,\dots,D_1}$ denote the disjoint partitions on $\boldsymbol{\Theta}_1$, and $\{\mathcal{W}_{1,d_1}\}_{d_1=1,\dots,D_1}$ denote the disjoint partitions on $\mathcal{W}_1$. These partitions are selected so that the probabilities match for each partition, i.e., $G_1(\boldsymbol{\Theta}_{1,d_1})=\Phi_1(\mathcal{W}_{1,d_1})$ for $d_1=1,2,\dots,D_1$. For each subspace $\boldsymbol{\Theta}_{1,d_1}$, we select a representative point $\boldsymbol{\theta}^*_{d_1}\in\boldsymbol{\Theta}_{1,d_1}$ such that $H(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}^*_{d_1})$ serves as a reasonable approximation of $H(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}_i)$ for any $\boldsymbol{\theta}_i\in\boldsymbol{\Theta}_{1,d_1}$. Rigorous technical details regarding the space partitioning procedures are presented in Appendices A.1 and A.2 (Steps 1 and 2). We then approximate $H(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}_i)$ using an MoE model with a representation analogous to $F(\boldsymbol{y}_i;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\Psi}|\boldsymbol{x}_i,\boldsymbol{w}_i)$ in Eq. (6), denoted as

$$F^{**(u)}(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{w}_i)=\sum_{d_1=1}^{D_1}\pi_{d_1}(\boldsymbol{x}_i,\boldsymbol{w}_i;\boldsymbol{\alpha},\boldsymbol{\beta})H(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}^*_{d_1}), \tag{17}$$

where $\pi_{d_1}(\boldsymbol{x}_i,\boldsymbol{w}_i;\boldsymbol{\alpha},\boldsymbol{\beta})$ is given by Eq. (7). Note that Eq. (17) corresponds exactly to $F^{**(u)}(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{w}_i)$ in Eq. (22) of the appendix, using simplified notations. The only distinction between $F^{**(u)}(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{w}_i)$ and $F(\boldsymbol{y}_i;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\Psi}|\boldsymbol{x}_i,\boldsymbol{w}_i)$ lies in the expert functions they employ: $F^{**(u)}(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{w}_i)$ uses $H(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}^*_{d_1})$, which depends on $\boldsymbol{x}_i$, while $F(\boldsymbol{y}_i;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\Psi}|\boldsymbol{x}_i,\boldsymbol{w}_i)$ uses $F_0(\boldsymbol{y}_i;\boldsymbol{\psi}_j)$, which is independent of $\boldsymbol{x}_i$. Nevertheless, following the approach outlined in Fung et al. [30], it is shown in Appendices A.3 and A.4 (Steps 3 and 4) that $H(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}^*_{d_1})$ can itself be approximated by an MoE distribution, with $F_0(\boldsymbol{y}_i;\boldsymbol{\psi}_j)$ serving as the expert function. This approximation process is not illustrated here to avoid unnecessary complexity, as it is unrelated to random effects and has already been demonstrated by Fung et al. [30]. Therefore, if $H(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}_i)$ can be well approximated by $F^{**(u)}(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{w}_i)$, it follows that it can also be effectively approximated by $F(\boldsymbol{y}_i;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\Psi}|\boldsymbol{x}_i,\boldsymbol{w}_i)$.

To perform approximations, we first observe that the softmax gating function $\pi_{d_1}(\boldsymbol{x}_i,\boldsymbol{w}_i;\boldsymbol{\alpha},\boldsymbol{\beta})$ in Eq. (17) is shown by Lemma 3.2 of Fung et al. [30] to be "fully flexible", allowing an observation i to be assigned to any of the D_1 mixture components based on the random effects $\boldsymbol{w}_i$. Specifically, by expressing $\boldsymbol{\beta}=u\tilde{\boldsymbol{\beta}}$, where $u>0$ is a tuning parameter that controls how "hard" the softmax gating function assigns observations to the mixture components, one can carefully select parameters $\tilde{\boldsymbol{\beta}}$ such that for any $\boldsymbol{w}_i\in\mathcal{W}_{1,d_1}$ and sufficiently large u, we have $\pi_{d_1}(\boldsymbol{x}_i,\boldsymbol{w}_i;\boldsymbol{\alpha},\boldsymbol{\beta})\approx 1$ and $\pi_{d_1'}(\boldsymbol{x}_i,\boldsymbol{w}_i;\boldsymbol{\alpha},\boldsymbol{\beta})\approx 0$ for any $d_1'\neq d_1$, effectively assigning the observation to mixture component d_1 when $\boldsymbol{w}_i\in\mathcal{W}_{1,d_1}$. This process interprets Lemma 2 in the appendix. Hence, we have $F^{**(u)}(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{w}_i)\approx H(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}^*_{d_1})$ from Eq. (17). If the partition of spaces $\boldsymbol{\Theta}_1$ and $\mathcal{W}_1$ are sufficiently fine by selecting a large enough D_1, then each subspace $\boldsymbol{\Theta}_{1,d_1}$ will become sufficiently small, ensuring that $H(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}^*_{d_1})\approx H(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}_i)$. Summarizing, we have

$$F(\boldsymbol{y}_i;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\Psi}|\boldsymbol{x}_i,\boldsymbol{w}_i)\approx F^{**(u)}(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{w}_i)\approx H(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}^*_{d_1})\approx H(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}_i), \tag{18}$$

and thus Eq. (16) holds, leading to the desired result.

8. Numerical illustration

We provide a numerical example to illustrate the denseness property of the proposed MMoE model, using the approximation procedures discussed in Section 7. For simplicity, the output considered in this example is one-dimensional, meaning $\boldsymbol{y}_i:=y_i\in\mathbb{R}$ with $K=1$. Additionally, we exclude the covariates $\boldsymbol{x}_i$ from this illustration, i.e., $P=1$, so that $\tilde{\boldsymbol{\beta}}$, as defined in Section 7, can be expressed as $\tilde{\boldsymbol{\beta}}=\{(\tilde{\beta}_{d_1,0},\tilde{\beta}_{d_1,1})\}_{d_1=1,2,\dots,D_1}$. Here, $\beta_{d_1,0}:=u\tilde{\beta}_{d_1,0}$ and $\beta_{d_1,1}:=u\tilde{\beta}_{d_1,1}$ correspond, respectively, to the intercept and random effect coefficients of the d_1-th latent class within $\pi_{d_1}(\boldsymbol{x}_i,\boldsymbol{w}_i;\boldsymbol{\alpha},\boldsymbol{\beta})$ in Eq. (17), which can be notationally simplified as $\pi_{d_1}(\boldsymbol{x}_i,\boldsymbol{w}_i;\boldsymbol{\alpha},\boldsymbol{\beta}):=\exp\{\beta_{d_1,0}+\beta_{d_1,1}w_1\}/\sum_{d_1'=1}^{D_1}\exp\{\beta_{d_1',0}+\beta_{d_1',1}w_1\}$ in this example. Considering only $L=1$, we choose $\boldsymbol{\Theta}_1=[-2,2]\subseteq\mathbb{R}$ as the space for generalized random effect $\boldsymbol{\theta}_i:=\theta_{i1}$. We construct θ_{i1} to follow a uniform distribution over the range -2 to 2, so that the distribution function of θ_{i1} is $(2+\theta_{i1})/4$. The conditional distribution of y_i given θ_{i1} is assumed to be normal with a mean of θ_{i1} and a standard deviation of 1, such that the density of y_i given θ_{i1} is $\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(y_i-\theta_{i1})^2}$.

In this context, many of the notations introduced in Section 7 can be greatly simplified. Specifically, we express $F^{**(u)}(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{w}_i)$ as $F^{**(u)}(y_i|w_{i1})$, $H(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}_{i1})$ as $H(y_i|\theta_{i1})$, and $H(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}^*_{d_1})$ as $H(y_i|\theta^*_{1,d_1})$, where θ^*_{1,d_1} is a scalar representative point in $\boldsymbol{\Theta}_{1,d_1}$ as defined in the previous section. To improve clarity, we restate the approximation relations in Eq. (18) using these simplified notations as follows:

$$F^{**(u)}(y_i|w_{i1})\overset{u\to\infty}{\approx}H(y_i|\theta^*_{1,d_1})\overset{D_1\to\infty}{\approx}H(y_i|\theta_{i1}). \tag{19}$$

As a concrete example, Table 2 illustrates the procedure described in Section 7 with $D_1=5$ partitions for $\boldsymbol{\Theta}_1$ and $\mathcal{W}_1$. We select the midpoint θ^*_{1,d_1} as the representative value for the subspace $\boldsymbol{\Theta}_{1,d_1}$. The values of $\{(\tilde{\beta}_{d_1,0},\tilde{\beta}_{d_1,1})\}_{d_1=1,2,\dots,D_1}$ are carefully chosen following Lemma 2 in the appendix (or see Section 7 for brief description). This example is visually represented in the second row of Fig. 4. Specifically, in Fig. 4, each vertical slice of the blue 2D images plots $F^{**(u)}(y_i|w_{i1})$ against y_i (vertical axis) and w_{i1} (horizontal axis) for $u=1,10,100$ and 1000 from left to right and for $D_1=2,5,10$ and 100 from top to bottom, while the green 2D images in the rightmost panel plot $H(y_i|\theta^*_{1,d_1})$ against y_i (vertical axis) and θ_{i1} for $D_1=2,5,10$ and 100 from top to bottom, where θ^*_{1,d_1} is the representative point of the subspace $\boldsymbol{\Theta}_{1,d_1}$ to which θ_{i1} belongs. For example, under Table 2, if $\theta_{i1}=-0.6$, then $\theta_{i1}\in\boldsymbol{\Theta}_{1,2}$ and thus $\theta^*_{1,d_1}=\theta^*_{1,2}=-0.8$. Two key observations can be made:

- As u increases, i.e., moving from the left panels to the right panels, $F^{**(u)}(y_i|w_{i1})$ visually converges towards $H(y_i|\theta^*_{1,d_1})$ as

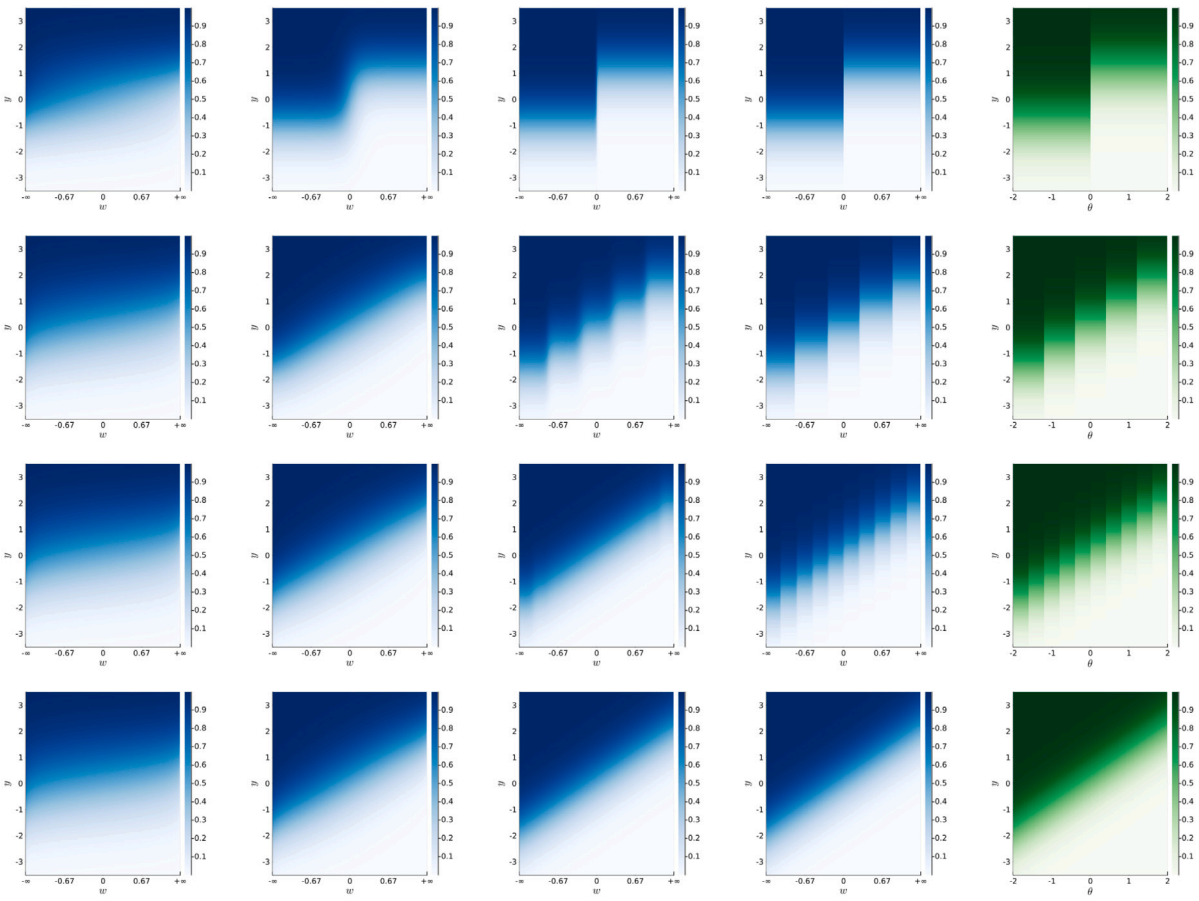


Fig. 4. Illustration of using $F^{**(u)}(y_i|w_{i1})$ (blue) to approximate $H(y_i|\theta^*_{1,d_1})$ (green) for different numbers of partitions of $\boldsymbol{\Theta}_1$ and $\mathcal{W}_1$. Each vertical slice corresponds to the conditional density of $F^{**(u)}(y_i|w_{i1})$ or $H(y_i|\theta^*_{1,d_1})$. From left to right, the control parameter u ranges from 1, 10, 100 to 1000 in the blue plots, and eventually $F^{**(u)}(y_i|w_{i1}) = H(y_i|\theta^*_{1,d_1})$ for all $w_1 \in \mathcal{W}_{1,d_1}$ as $u \to \infty$. From top to bottom, the number of partition D_1 ranges from 2, 5, 10 to 100. When the partition becomes infinitely fine and $u \to \infty$, each $H(y_i|\theta_{1,d_1})$ is approximated arbitrarily well by some $F^{**(\infty)}(y_i|w_{i1})$, e.g., $H(y_i|\theta_{1,d_1} = 1) = F^{**(\infty)}(y_i|w_1 = 0.67)$.

Table 2
Example of approximation with five partitions of $\boldsymbol{\Theta}_1$ and $\mathcal{W}_1$.

d_1	1	2	3	4	5
$\boldsymbol{\Theta}_{1,d_1}$	$[-2.00, -1.20]$	$(-1.20, -0.40]$	$(-0.40, 0.40]$	$(0.40, 1.20]$	$(1.20, 2.00]$
$\mathcal{W}_{1,d_1}$	$(-\infty, -0.84]$	$(-0.84, -0.25]$	$(-0.25, 0.25]$	$(0.25, 0.84]$	$(0.84, +\infty)$
θ^*_{1,d_1}	-1.60	-0.80	0.00	0.80	1.60
$\tilde{\beta}_{d_1,0}$	0.00	0.23	0.29	0.23	0.00
$\tilde{\beta}_{d_1,1}$	0.00	0.25	0.50	0.75	1.00

shown in the rightmost panel. This provides empirical support for the convergence $F^{**(u)}(y_i|w_{i1}) \to H(y_i|\theta^*_{1,d_1})$ as $u \to \infty$.

- As D_1 increases, making the partitions finer (moving from top to bottom in the rightmost panel), $H(y_i|\theta^*_{1,d_1})$ transitions from resembling a step function to becoming a smooth function of θ_{i1}. The bottom-right panel closely resembles $H(y_i|\theta_{i1})$, which is a continuous function of (y_i, θ_{i1}). This supports the convergence $H(y_i|\theta^*_{1,d_1}) \to H(y_i|\theta_{i1})$ as $D_1 \to \infty$.

In summary, this numerical example supports the approximation relationship given by Eq. (19), thereby demonstrating the denseness property of the proposed MMoE model as outlined in Section 7.

9. Discussions

In this paper, we introduce a class of the mixed mixture of experts models (MMoE) for multilevel regression data. We prove that the MMoE is dense in the space of generalized mixed effects models, which is a rich class containing almost all models in the literature having independent random effects, in the sense of weak convergence. We further study a special case where the data is hierarchical in structure. In this case, the proposed nested MMoE is shown to be dense in the space of generalized nested mixed effects models where the random effects can possibly be dependent. The two denseness theorems justify the versatility of the MMoE in catering for a broad range of multilevel data characteristics, including the marginal distribution, dependence, regression link, random intercept and random slope.

This paper aims to prove the most general results imposing minimal assumptions. The only practical restriction is that the expert function $F_0(\boldsymbol{y}_i; \boldsymbol{\psi}_j)$ in Eq. (6) needs to approximate any degenerate distributions (Assumption 4). This assumption is much weaker than those applied to the existing approximation theorems (see, e.g., Nguyen et al. [22], Norets and Pelenis [26]), which require that the expert density function is a scalable symmetric function (Equation (3.1) of Norets and Pelenis [26]), and that the target density function does not change abruptly w.r.t. $\boldsymbol{y}_i$ and $\boldsymbol{x}_i$ (Equation (3.2) of Norets and Pelenis [26]). On the other hand, there are several limitations of the denseness theorems formulated in this paper. Firstly, weak convergence does not guarantee the approximation capability in terms of moments (e.g., the mean function studied by Nguyen et al. [21]) or some distance metrics (e.g., the KL divergence studied by Jiang and Tanner [20],Nguyen et al. [22],Norets et al. [25]). To establish such denseness theorems, one needs to further assume that the moments of the MMoE expert functions and the target distributions are finite, and that the conditions indicated by Equations (3.1) and (3.2) of Norets and Pelenis [26] are fulfilled. Secondly, as described in Section 3.4 of Fung et al. [30], the denseness theorems

do not provide a convergence rate, so there is no control on the mixture components g to approximate any generalized mixed effects distributions at a desired level of accuracy. To establish the rate results, one needs to impose further assumptions on the target distribution $\tilde{H}(\boldsymbol{y}|\boldsymbol{x})$ in Eq. (4) and the MMoE distribution $\tilde{F}(\boldsymbol{y};\boldsymbol{x})$ in Eq. (8). We leave these technical establishments as a future research direction.

CRediT authorship contribution statement

Tsz Chai Fung: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Spark C. Tseung:** Formal analysis, Visualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Proof of Theorem 1

We begin by introducing the following functions, with detailed notation to be clarified later in the proof.

$$\tilde{H}(\boldsymbol{y}|\boldsymbol{x}) = \int_{\tilde{\boldsymbol{\Theta}}} \left[\prod_{i=1}^{N} H(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}_i)\right] dG(\tilde{\boldsymbol{\theta}}), \tag{20}$$

$$\tilde{H}^*(\boldsymbol{y}|\boldsymbol{x}) = \sum_{\tilde{\boldsymbol{d}}\in\tilde{D}} \prod_{i=1}^{N} H(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}^*_{\boldsymbol{d}^{(c(i))}}) G(\tilde{\boldsymbol{\Theta}}_{\tilde{\boldsymbol{d}}}), \tag{21}$$

$$\tilde{F}^{**(u)}(\boldsymbol{y}|\boldsymbol{x}) = \int \prod_{i=1}^{N} F^{**(u)}(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{w}_i) d\boldsymbol{\Phi}(\boldsymbol{w}) \tag{22}$$

with $F^{**(u)}(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{w}_i) = \sum_{\boldsymbol{d}\in D^+} \xi^{(u)}_{\boldsymbol{d}}(\boldsymbol{w}_i) H(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}^*_{\boldsymbol{d}})$,

$$\tilde{F}^{*(M,Q,t,u)}(\boldsymbol{y}|\boldsymbol{x}) = \int \prod_{i=1}^{N} F^{*(M,Q,t,u)}(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{w}_i) d\boldsymbol{\Phi}(\boldsymbol{w}) \tag{23}$$

with $F^{*(M,Q,t,u)}(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{w}_i) = \sum_{\boldsymbol{m}\in\mathcal{M}} \sum_{\boldsymbol{q}\in Q} \sum_{\boldsymbol{d}\in D^+} \pi_j^{(t,u)}(\boldsymbol{x}_i,\boldsymbol{w}_i;\tilde{\boldsymbol{\alpha}}^Q_{\boldsymbol{q}},\tilde{\boldsymbol{\beta}}_{\boldsymbol{d}}) 1\{\boldsymbol{y}_i \geq \lfloor \boldsymbol{y} \rfloor^M_{\boldsymbol{m}}\}$,

$$\tilde{F}^{(M,Q,t,u,v)}(\boldsymbol{y}|\boldsymbol{x}) = \int \prod_{i=1}^{N} F^{(M,Q,t,u,v)}(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{w}_i) d\boldsymbol{\Phi}(\boldsymbol{w}) \tag{24}$$

with $F^{(M,Q,t,u,v)}(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{w}_i) = \sum_{\boldsymbol{m}\in\mathcal{M}} \sum_{\boldsymbol{q}\in Q} \sum_{\boldsymbol{d}\in D^+} \pi_j^{(t,u)}(\boldsymbol{x}_i,\boldsymbol{w}_i;\tilde{\boldsymbol{\alpha}}^Q_{\boldsymbol{q}},\tilde{\boldsymbol{\beta}}_{\boldsymbol{d}}) F_0(\boldsymbol{y}_i;\boldsymbol{\psi}^{M(v)}_{\boldsymbol{m}})$. Note that Eq. (20) is in the form of the generalized mixed effects models under Eq. (4), while Eq. (24) is in the MMoE framework under Eq. (8). The main proof idea is to bound the approximation errors between any two consecutive functions from Eqs. (20) to (24):

A.1. Step 1: Approximating Eq. (20) by Eq. (21)

As the metric space $(\boldsymbol{\Theta}_l, d_{\boldsymbol{\Theta}_l})$ is complete separable (Assumption 2), the probability measure G_l on $\boldsymbol{\theta}^{(s)}_l$ is tight, i.e. $\forall \epsilon_1 > 0$, $\exists \bar{\boldsymbol{\Theta}}_l \subseteq \boldsymbol{\Theta}_l$ compact such that $G_l(\bar{\boldsymbol{\Theta}}_l) := \mathbb{P}(\boldsymbol{\theta}^{(s)}_l \in \bar{\boldsymbol{\Theta}}_l) \geq 1-\epsilon_1$. Since $\bar{\boldsymbol{\Theta}}_l$ is compact, for any $\delta_1 > 0$ we can find subspaces $\{\boldsymbol{\Theta}_{l,d_l}\}_{d_l=1,\dots,D_l}$ (d_l is called the subspace index) and points $\{\boldsymbol{\theta}^*_{l,d_l}\}_{d_l=1,\dots,D_l}$ such that $\cup_{d_l=1,\dots,D_l} \boldsymbol{\Theta}_{l,d_l} = \bar{\boldsymbol{\Theta}}_l$, $\boldsymbol{\theta}^*_{l,d_l} \in \boldsymbol{\Theta}_{l,d_l}$ for every $d_l = 1,\dots,D_l$ and $\boldsymbol{\Theta}_{l,d_l}$ is covered by the ball with radius δ_1/L centered at $\boldsymbol{\theta}^*_{l,d_l}$. Due to the uniform continuity of H w.r.t. $\boldsymbol{\theta}_i$ on $\bar{\boldsymbol{\Theta}}_l$ (Assumption 3 implies uniform convergence in compact space), for any $\epsilon_2 > 0$ we can choose sufficiently small δ_1 with the aforementioned subspace partitioning such that $|H(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}_i) - H(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}^*_{\boldsymbol{d}^{(c(i))}})| \leq \epsilon_2$ if $\boldsymbol{\theta}_{il} \in \boldsymbol{\Theta}_{l,d_l^{(c(i))}}$ for every $l = 1,\dots,L$, where $\boldsymbol{\theta}^*_{\boldsymbol{d}^{c(i)}} = \{\boldsymbol{\theta}^*_{1,d_l^{(c(i))}}\}_{l=1,\dots,L}$ and $\boldsymbol{d}^{(c(i))} = \{d_l^{(c(i))}\}_{l=1,\dots,L}$ with $d_l^{(c(i))} \in \{1,\dots,D_l\}$. Note here the superscript $(c(i))$ of $d_l^{(c(i))}$ represents the subspace index d_l corresponding to the ith observation.

Define $\tilde{H}^*(\boldsymbol{y}|\boldsymbol{x})$ in the form of Eq. (21), where $\tilde{D} = \prod_{l=1}^{L}\prod_{s=1}^{S_l} D_l^{(s)}$ with $D_l^{(s)} = \{1,\dots,D_l\}$, $\tilde{\boldsymbol{d}} = \{d_l^{(s)}\}_{l=1,\dots,L;s=1,\dots,S_l}$ and $\tilde{\boldsymbol{\Theta}}_{\tilde{\boldsymbol{d}}} = \prod_{l=1}^{L}\prod_{s=1}^{S_l} \boldsymbol{\Theta}_{l,d_l^{(s)}}$. We first introduce the following technical lemma which can be easily proved by induction:

Lemma 1. *Given $0 \leq a_i, b_i \leq 1$ and $|a_i - b_i| \leq \epsilon$ for all $i = 1,\dots,N$, then $|\prod_{i=1}^{N} a_i - \prod_{i=1}^{N} b_i| \leq N\epsilon$.*

Since $\mathbb{P}(\tilde{\boldsymbol{\theta}} \notin \cup_{\tilde{\boldsymbol{d}}\in\tilde{D}} \tilde{\boldsymbol{\Theta}}_{\tilde{\boldsymbol{d}}}) \leq \sum_{l=1,\dots,L;s=1,\dots,S_l} \mathbb{P}(\boldsymbol{\theta}^{(s)}_l \notin \bar{\boldsymbol{\Theta}}_l) \leq NL\epsilon_1$, we can now rewrite $\tilde{H}(\boldsymbol{y}|\boldsymbol{x})$ as

$$\tilde{H}(\boldsymbol{y}|\boldsymbol{x}) = \sum_{\tilde{\boldsymbol{d}}\in\tilde{D}} \int_{\tilde{\boldsymbol{\Theta}}_{\tilde{\boldsymbol{d}}}} \prod_{i=1}^{N} H(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}_i) dG(\tilde{\boldsymbol{\theta}}) + NL\mathcal{O}_1(\epsilon_1), \tag{25}$$

and $\tilde{H}^*(\boldsymbol{y}|\boldsymbol{x})$ as

$$\tilde{H}^*(\boldsymbol{y}|\boldsymbol{x}) = \sum_{\tilde{\boldsymbol{d}}\in\tilde{D}} \int_{\tilde{\boldsymbol{\Theta}}_{\tilde{\boldsymbol{d}}}} \prod_{i=1}^{N} H(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}^*_{\boldsymbol{d}^{(c(i))}}) dG(\tilde{\boldsymbol{\theta}}), \tag{26}$$

where $0 \leq \mathcal{O}_1(\epsilon_1) \leq \epsilon_1$, and we have the following approximation error bound between $\tilde{H}(\boldsymbol{y}|\boldsymbol{x})$ and $\tilde{H}^*(\boldsymbol{y}|\boldsymbol{x})$:

$$\begin{aligned} |\tilde{H}^*(\boldsymbol{y}|\boldsymbol{x}) - \tilde{H}(\boldsymbol{y}|\boldsymbol{x})| &\leq \sum_{\tilde{\boldsymbol{d}}\in\tilde{D}} \int_{\tilde{\boldsymbol{\Theta}}_{\tilde{\boldsymbol{d}}}} \left| \prod_{i=1}^{N} H(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}^*_{\boldsymbol{d}^{(c(i))}}) - \prod_{i=1}^{N} H(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}_i) \right| dG(\tilde{\boldsymbol{\theta}}) + NL\epsilon_1 \\ &\leq N\epsilon_2 + NL\epsilon_1, \end{aligned} \tag{27}$$

where the second inequality is resulted from Lemma 1.

A.2. Step 2: Approximating Eq. (21) by Eq. (22)

Partition the space of w_{il} (i.e. $\mathbb{R}$) into $D_l + 2$ adjacent half open half closed intervals $\mathcal{W}_{l,d_l} = (w^*_{l,d_l-1}, w^*_{l,d_l}]$ (for $d_l = 1,\dots,D_l$), $\mathcal{W}_{l,0} = (-\infty, w^*_{l,0}]$ and $\mathcal{W}_{l,D_l+1} = (w^*_{l,D_l}, \infty)$ such that $\boldsymbol{\Phi}_l(\mathcal{W}_{l,d_l}) = G_l(\boldsymbol{\Theta}_{l,d_l})$ for every $d_l = 1,\dots,D_l$. Note that such a partitioning always exists as $\boldsymbol{\Phi}_l$ is a continuous distribution. Also denote the domain of (the L-vector) $\boldsymbol{d} = (d_1,\dots,d_L)$ as $D = \prod_{l=1}^{L}\{1,\dots,D_l\}$ and the corresponding extended domain $D^+ = \prod_{l=1}^{L}\{0,1,\dots,D_l+1\}$.

Denote $\tilde{F}^{**(u)}(\boldsymbol{y}|\boldsymbol{x})$ as the form of Eq. (22) with $\xi^{(u)}_{\boldsymbol{d}}(\boldsymbol{w}_i)$ given by

$$\xi^{(u)}_{\boldsymbol{d}}(\boldsymbol{w}_i) = \exp\{u(\tilde{\beta}_{\boldsymbol{d},0} + \tilde{\boldsymbol{\beta}}^T_{\boldsymbol{d}}\boldsymbol{w}_i)\} / \sum_{\boldsymbol{d}'\in D^+} \exp\{u(\tilde{\beta}_{\boldsymbol{d}',0} + \tilde{\boldsymbol{\beta}}^T_{\boldsymbol{d}'}\boldsymbol{w}_i)\}. \tag{28}$$

We have the following technical lemma to help us choose suitable parameters $\{(\tilde{\beta}_{\boldsymbol{d},0}, \tilde{\boldsymbol{\beta}}_{\boldsymbol{d}})\}_{\boldsymbol{d}\in D^+}$ of $\xi^{(u)}_{\boldsymbol{d}}(\boldsymbol{w}_i)$:

Lemma 2. *There exists parameters $\{(\tilde{\beta}_{\boldsymbol{d},0}, \tilde{\boldsymbol{\beta}}_{\boldsymbol{d}})\}_{\boldsymbol{d}\in D^+}$ of $\xi^{(u)}_{\boldsymbol{d}}(\boldsymbol{w}_i)$ such that $\xi^{(u)}_{\boldsymbol{d}}(\boldsymbol{w}_i) \xrightarrow{u\to\infty} \prod_{l=1}^{L} 1^*_{w_{il}}(\mathcal{W}_{l,d_l})$ for every $\boldsymbol{d} \in D^+$, where $1^*_w(\mathcal{W})$ is an indicator which equals to 1 if w falls inside the interior of $\mathcal{W}$, equals to a constant $c \in [0,1]$ if W is on the boundary of $\mathcal{W}$, and equals to 0 otherwise.*

Proof. With a slight (notational) variant of Lemma 3.1 of Fung et al. [30], one can easily show the existence of parameters $\{(\tilde{\beta}_{\boldsymbol{d},0}, \tilde{\boldsymbol{\beta}}_{\boldsymbol{d}})\}_{\boldsymbol{d}\in D^+}$ such that $\tilde{\beta}_{\boldsymbol{d},0} + \tilde{\boldsymbol{\beta}}^T_{\boldsymbol{d}}\boldsymbol{w}_i > \max_{\boldsymbol{d}'\neq\boldsymbol{d};\boldsymbol{d}'\in D^+} \tilde{\beta}_{\boldsymbol{d}',0} + \tilde{\boldsymbol{\beta}}^T_{\boldsymbol{d}'}\boldsymbol{w}_i$ if and only if $w_{il} \in \mathcal{W}^*_{l,d_l}$ for every $\boldsymbol{d} \in D^+$, where $\mathcal{W}^*_{l,d_l}$ is the interior of $\mathcal{W}_{l,d_l}$. Under a slight variation of Lemma 3.2 of Fung et al. [30], we have $\xi^{(u)}_{\boldsymbol{d}}(\boldsymbol{w}_i) \xrightarrow{u\to\infty} 1\{\tilde{\beta}_{\boldsymbol{d},0} + \tilde{\boldsymbol{\beta}}^T_{\boldsymbol{d}}\boldsymbol{w}_i > \max_{\boldsymbol{d}'\neq\boldsymbol{d};\boldsymbol{d}'\in D^+} \tilde{\beta}_{\boldsymbol{d}',0} + \tilde{\boldsymbol{\beta}}^T_{\boldsymbol{d}'}\boldsymbol{w}_i\} + \mathcal{O}_1(1) \times 1\{\tilde{\beta}_{\boldsymbol{d},0} + \tilde{\boldsymbol{\beta}}^T_{\boldsymbol{d}}\boldsymbol{w}_i = \max_{\boldsymbol{d}'\neq\boldsymbol{d};\boldsymbol{d}'\in D^+} \tilde{\beta}_{\boldsymbol{d}',0} + \tilde{\boldsymbol{\beta}}^T_{\boldsymbol{d}'}\boldsymbol{w}_i\} = \prod_{l=1}^{L} 1^*_{w_{il}}(\mathcal{W}_{l,d_l})$, where $0 \leq \mathcal{O}_1(1) \leq 1$. □

Choosing the parameters indicated by the above lemma, we have the following approximation result between $\tilde{H}^*(\boldsymbol{y}|\boldsymbol{x})$ and $\tilde{F}^{**(u)}(\boldsymbol{y}|\boldsymbol{x})$:

$$\tilde{F}^{**(u)}(\boldsymbol{y}|\boldsymbol{x}) \xrightarrow{u\to\infty} \int \prod_{i=1}^{N} \sum_{\boldsymbol{d}\in D^+} \left(\prod_{l=1}^{L} 1^*_{w_{il}}(\mathcal{W}_{l,d_l})\right) H(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}^*_{\boldsymbol{d}}) d\boldsymbol{\Phi}(\boldsymbol{w})$$

$$\pi_j^{(t,u)}(\boldsymbol{x}_i, \boldsymbol{w}_i; \tilde{\boldsymbol{\alpha}}_{\boldsymbol{q}}^{Q}, \tilde{\boldsymbol{\beta}}_{\boldsymbol{d}})$$

$$= \frac{\exp\{\log H(\mathcal{Y}_{\boldsymbol{m}}^{M}|\boldsymbol{x}_{\boldsymbol{q}}^{*Q}, \boldsymbol{\theta}_{\boldsymbol{d}}^{*}) + t(\tilde{\alpha}_{\boldsymbol{q},0}^{Q} + \tilde{\boldsymbol{\alpha}}_{\boldsymbol{q}}^{QT}\boldsymbol{x}_i) + u(\tilde{\beta}_{\boldsymbol{d},0} + \tilde{\boldsymbol{\beta}}_{\boldsymbol{d}}^{T}\boldsymbol{w}_i)\}}{\sum_{\boldsymbol{m}'\in\mathcal{M}}\sum_{\boldsymbol{q}'\in\mathcal{Q}}\sum_{\boldsymbol{d}'\in\mathcal{D}^+}\exp\{\log H(\mathcal{Y}_{\boldsymbol{m}'}^{M}|\boldsymbol{x}_{\boldsymbol{q}'}^{*Q}, \boldsymbol{\theta}_{\boldsymbol{d}'}^{*}) + t(\tilde{\alpha}_{\boldsymbol{q}',0}^{Q} + \tilde{\boldsymbol{\alpha}}_{\boldsymbol{q}'}^{QT}\boldsymbol{x}_i) + u(\tilde{\beta}_{\boldsymbol{d}',0} + \tilde{\boldsymbol{\beta}}_{\boldsymbol{d}'}^{T}\boldsymbol{w}_i)\}} \tag{30}$$

$$= \frac{\exp\{t(\tilde{\alpha}_{\boldsymbol{q},0}^{Q} + \tilde{\boldsymbol{\alpha}}_{\boldsymbol{q}}^{QT}\boldsymbol{x}_i)\}}{\sum_{\boldsymbol{q}'\in\mathcal{Q}}\exp\{t(\tilde{\alpha}_{\boldsymbol{q}',0}^{Q} + \tilde{\boldsymbol{\alpha}}_{\boldsymbol{q}'}^{QT}\boldsymbol{x}_i)\}} \frac{\exp\{u(\tilde{\beta}_{\boldsymbol{d},0} + \tilde{\boldsymbol{\beta}}_{\boldsymbol{d}}^{T}\boldsymbol{w}_i)\}}{\sum_{\boldsymbol{d}'\in\mathcal{D}^+}\exp\{u(\tilde{\beta}_{\boldsymbol{d}',0} + \tilde{\boldsymbol{\beta}}_{\boldsymbol{d}'}^{T}\boldsymbol{w}_i)\}} H(\mathcal{Y}_{\boldsymbol{m}}^{M}|\boldsymbol{x}_{\boldsymbol{q}}^{*Q}, \boldsymbol{\theta}_{\boldsymbol{d}}^{*})$$

$$:= \gamma_{\boldsymbol{q}}^{(t)}(\boldsymbol{x}_i) \times \xi_{\boldsymbol{d}}^{(u)}(\boldsymbol{w}_i) \times H(\mathcal{Y}_{\boldsymbol{m}}^{M}|\boldsymbol{x}_{\boldsymbol{q}}^{*Q}, \boldsymbol{\theta}_{\boldsymbol{d}}^{*}), \tag{31}$$

Box I.

$$\begin{aligned}
&= \int_{\bar{\mathcal{W}}} \prod_{i=1}^{N} \sum_{\boldsymbol{d}\in\mathcal{D}^+} \left(\prod_{l=1}^{L} 1_{w_{il}}^{*}(\mathcal{W}_{l,d_l})\right) H(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta}_{\boldsymbol{d}}^{*}) d\boldsymbol{\Phi}(\boldsymbol{w}) + \mathcal{O}_1(\epsilon_1) \\
&= \sum_{\tilde{\boldsymbol{d}}\in\tilde{\mathcal{D}}} \int_{\mathcal{W}_{\tilde{\boldsymbol{d}}}} \prod_{i=1}^{N} \sum_{\boldsymbol{d}\in\mathcal{D}} \left(\prod_{l=1}^{L} 1_{w_{il}}^{*}(\mathcal{W}_{l,d_l})\right) H(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta}_{\boldsymbol{d}}^{*}) d\boldsymbol{\Phi}(\boldsymbol{w}) + \mathcal{O}_1(\epsilon_1) \\
&= \sum_{\tilde{\boldsymbol{d}}\in\tilde{\mathcal{D}}} \int_{\mathcal{W}_{\tilde{\boldsymbol{d}}}} \prod_{i=1}^{N} H(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta}_{\boldsymbol{d}^{(c(i))}}^{*}) d\boldsymbol{\Phi}(\boldsymbol{w}) + \mathcal{O}_1(\epsilon_1) \\
&= \sum_{\tilde{\boldsymbol{d}}\in\tilde{\mathcal{D}}} \prod_{i=1}^{N} H(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta}_{\boldsymbol{d}^{(c(i))}}^{*}) G(\tilde{\boldsymbol{\Theta}}_{\tilde{\boldsymbol{d}}}) + \mathcal{O}_1(\epsilon_1) = \tilde{H}^{*}(\boldsymbol{y}|\boldsymbol{x}) + \mathcal{O}_1(\epsilon_1),
\end{aligned} \tag{29}$$

where $\bar{\mathcal{W}}_l^{(s)} = \cup_{d_l=1}^{D_l} \mathcal{W}_{l,d_l}$, $\bar{\mathcal{W}} = \prod_{l=1}^{L}\prod_{s=1}^{S_l} \bar{\mathcal{W}}_l^{(s)}$ and $\mathcal{W}_{\tilde{\boldsymbol{d}}} = \prod_{l=1}^{L}\prod_{s=1}^{S_l} \bar{\mathcal{W}}_{l,d_l^{(s)}}$. The convergence is resulted from the Dominated Convergence Theorem (DCT), which is obviously uniform on $(\boldsymbol{y}, \boldsymbol{x})$ as $\xi_{\boldsymbol{d}}^{(u)}(\boldsymbol{w}_i)$ (the only term in $\tilde{F}^{**(u)}(\boldsymbol{y}|\boldsymbol{x})$ which depends on u) does not depend on $(\boldsymbol{y}, \boldsymbol{x})$ and $H(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta}_{\boldsymbol{d}}^{*})$ is bounded above at 1. The third last equality holds as the events of the indicator functions only (jointly) hold if $\boldsymbol{d} = \boldsymbol{d}^{(c(i))}$ when $\boldsymbol{w} \in \mathcal{W}_{\tilde{\boldsymbol{d}}}$. The second last equality holds as the integrand is constant on $\boldsymbol{w} \in \mathcal{W}_{\tilde{\boldsymbol{d}}}$ and $\boldsymbol{\Phi}(\mathcal{W}_{\tilde{\boldsymbol{d}}}) = \prod_{l=1}^{L}\prod_{s=1}^{S_l} \boldsymbol{\Phi}_l(\mathcal{W}_{l,d_l^{(s)}}) = \prod_{l=1}^{L}\prod_{s=1}^{S_l} G_l(\boldsymbol{\Theta}_{l,d_l^{(s)}}) = G(\tilde{\boldsymbol{\Theta}}_{\tilde{\boldsymbol{d}}})$.

A.3. Step 3: Approximating Eq. (22) by Eq. (23)

Partition the space of $y_{ik} \in \mathbb{R}$ into adjacent half open half closed intervals $\{\mathcal{Y}_{k,m}^{M} = ((m-1/2)h_M^y, (m+1/2)h_M^y]\}_{m=-M,\dots,M-1,M}$. Define $\mathcal{Y}^{M} := \prod_{k=1}^{K} \cup_{m=-M}^{M} \mathcal{Y}_{k,m}^{M} := \prod_{k=1}^{K} \mathcal{Y}_k^{M} = (-(M+1/2)h_M^y, (M+1/2)h_M^y]^K$, $\mathcal{Y}_{k,-(M+1)}^{M} = (-\infty, -(M+1/2)h_M^y]$ and $\mathcal{Y}_{k,M+1}^{M} = ((M+1/2)h_M^y, \infty)$. Choose h_M^y such that $h_M^y \downarrow 0$ and $Mh_M^y \uparrow \infty$ as $M \to \infty$. Also denote $\mathcal{Y}_{\boldsymbol{m}}^{M} = \mathcal{Y}_{1,m_1}^{M} \times \cdots \times \mathcal{Y}_{K,m_K}^{M}$ as a hypercube with $\boldsymbol{m} := (m_1, \dots, m_K) \in \mathcal{M} := \{-(M+1), \dots, (M+1)\}^K$.

Similarly, partition the space of $x_{ip} \in \mathbb{R}$ into adjacent half open half closed intervals $\{\mathcal{X}_{p,q}^{Q} = ((q-1/2)h_Q^x, (q+1/2)h_Q^x]\}_{q=-Q,\dots,Q-1,Q}$. Define $\mathcal{X}^{Q} := \prod_{p=1}^{P} \cup_{q=-Q}^{Q} \mathcal{X}_{p,q}^{Q} := \prod_{p=1}^{P} \mathcal{X}_p^{Q} = (-(Q+1/2)h_Q^x, (Q+1/2)h_Q^x]^P$, $\mathcal{X}_{p,-(Q+1)}^{Q} = (-\infty, -(Q+1/2)h_Q^x]$ and $\mathcal{X}_{p,Q+1}^{Q} = ((Q+1/2)h_Q^x, \infty)$. Choose h_Q^x such that $h_Q^x \downarrow 0$ and $Qh_Q^x \uparrow \infty$ as $Q \to \infty$. Also denote $\mathcal{X}_{\boldsymbol{q}}^{Q} = \mathcal{Y}_{1,q_1}^{Q} \times \cdots \times \mathcal{X}_{P,q_P}^{Q}$ as a hypercube with $\boldsymbol{q} := (q_1, \dots, q_P) \in \mathcal{Q} := \{-(Q+1), \dots, (Q+1)\}^P$.

Denote $\tilde{F}^{*(M,Q,t,u)}(\boldsymbol{y}|\boldsymbol{x})$ as the form of Eq. (23), where $\lfloor \boldsymbol{y} \rfloor_{\boldsymbol{m}}^{M} := (\lfloor y \rfloor_{m_1}^{M}, \dots, \lfloor y \rfloor_{m_K}^{M})$ (inside the expression of $F^{*(M,Q,t,u)}(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{w}_i)$) represents the leftmost vertex of the hypercube $\mathcal{Y}_{\boldsymbol{m}}^{M}$. Also, the function $\pi_j^{(t,u)}(\boldsymbol{x}_i, \boldsymbol{w}_i; \tilde{\boldsymbol{\alpha}}_{\boldsymbol{q}}^{Q}, \tilde{\boldsymbol{\beta}}_{\boldsymbol{d}})$ is a logit linear gating function given by Eq. (31) (see Box I), where $\boldsymbol{x}_{\boldsymbol{q}}^{*Q} = (x_{1,q_1}^{*Q}, \dots, x_{P,q_P}^{*Q})$ is the mid-point of the hypercube $\mathcal{X}_{\boldsymbol{q}}^{Q}$. Here, "mid-points" for $\mathcal{X}_{p,-(Q+1)}$ and $\mathcal{X}_{p,(Q+1)}$ are respectively defined as $-(Q+1)h_Q^x$ and $(Q+1)h_Q^x$.

Further denote $R^{Q}(\boldsymbol{x}_i) = \{\boldsymbol{q} : \forall p \text{ we have } |pj^{Q}(x_{ip}) - x_{p,q_p}^{*Q}| \le h_Q^x\}$ and $R'^{Q}(\boldsymbol{x}_i) = \{\boldsymbol{q} : \exists p \text{ such that } |pj^{Q}(x_{ip}) - x_{p,q_p}^{*Q}| > h_Q^x\}$, where $pj^{Q}(x_{ip})$ is the projection of x_{ip} on the interval $\mathcal{X}_p^{Q}$ (i.e. $pj^{Q}(x_{ip})$ is the point in $\mathcal{X}_p^{Q}$ where x_{ip} is closest to). To derive the approximation results, we first introduce the following two technical lemmas.

Lemma 3. *Given any fixed Q and h_Q^x, there exists parameters $\{\tilde{\alpha}_{\boldsymbol{q},0}^{Q}, \tilde{\boldsymbol{\alpha}}_{\boldsymbol{q}}^{QT}\}_{\boldsymbol{q}\in\mathcal{Q}}$, such that for any $\epsilon_3 > 0$, we have $\sum_{\boldsymbol{q}\in R'^{Q}(\boldsymbol{x}_i)} \gamma_{\boldsymbol{q}}^{(t)}(\boldsymbol{x}_i) \le \epsilon_3$ for all $\boldsymbol{x}_i \in \mathcal{X}$ with sufficiently large t.*

Proof. This follows directly from the proof of Theorem 3.2 of Fung et al. [30]. □

Lemma 4. *The probability distributions $\{H(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta}_i)\}_{\boldsymbol{x}_i\in\bar{\mathcal{X}}; \boldsymbol{\theta}_i\in\bar{\boldsymbol{\Theta}}}$ are tight for any compact spaces $\bar{\mathcal{X}}$ and $\bar{\boldsymbol{\Theta}}$.*

Proof. Divide the compact space $\bar{\mathcal{Z}} := \bar{\boldsymbol{X}} \times \bar{\boldsymbol{\Theta}}$ into D subspaces $\{\mathcal{Z}_d\}_{d=1,\dots,D}$, where each subspace $\mathcal{Z}_d$ is small enough to be covered by a ball with radius δ. Define $\boldsymbol{z}_d^{*} := (\boldsymbol{x}_d^{*}, \boldsymbol{\theta}_d^{*})$ as an arbitrary interior point of $\mathcal{Z}_d$ for $d = 1, \dots, D$. For each $d = 1, \dots, D$, we choose a response space $\tilde{\mathcal{Y}}_d \in \mathcal{Y}$ such that $H(\tilde{\mathcal{Y}}_d|\boldsymbol{x}_d^{*}, \boldsymbol{\theta}_d^{*}) \ge 1 - \epsilon/2$. Select a compact space $\tilde{\mathcal{Y}}^{*}$ which covers all $\{\tilde{\mathcal{Y}}_d\}_{d=1,\dots,D}$, and we have $H(\tilde{\mathcal{Y}}^{*}|\boldsymbol{x}_d^{*}, \boldsymbol{\theta}_d^{*}) \ge 1 - \epsilon/2$ true for all $d = 1, \dots, D$. Uniform continuity of H on any compact space implies that $|H(\tilde{\mathcal{Y}}^{*}|\boldsymbol{x}_d^{*}, \boldsymbol{\theta}_d^{*}) - H(\tilde{\mathcal{Y}}^{*}|\boldsymbol{x}_i, \boldsymbol{\theta}_i)| \le \epsilon/2$ for sufficient small δ if $(\boldsymbol{x}_i, \boldsymbol{\theta}_i) \in \mathcal{Z}_d$. Overall, we have $H(\tilde{\mathcal{Y}}^{*}|\boldsymbol{x}_i, \boldsymbol{\theta}_i) \ge 1 - \epsilon$, and hence the result follows. □

As a result, for any compact covariates space of $\bar{\mathcal{X}}$ and any $\epsilon_4 > 0$, we can use Lemma 4 to select a rectangular output space $\bar{\mathcal{Y}}$ such that $H(\bar{\mathcal{Y}}|\boldsymbol{x}_i, \boldsymbol{\theta}_i) \ge 1 - \epsilon_4$ for any $\boldsymbol{x}_i \in \bar{\mathcal{X}}$ and $\boldsymbol{\theta}_i \in \bar{\boldsymbol{\Theta}}$, where $\bar{\boldsymbol{\Theta}} = \prod_{l=1}^{L} \bar{\boldsymbol{\Theta}}_l$ and note that $\bar{\boldsymbol{\Theta}}_l$ is defined in Appendix A.1. Then we have for all $\boldsymbol{x}_i, \boldsymbol{x}_{\boldsymbol{q}}^{*Q} \in \bar{\mathcal{X}}$ and $\boldsymbol{\theta}_{\boldsymbol{d}}^{*} \in \bar{\boldsymbol{\Theta}}$:

$$|H(\lceil \boldsymbol{y} \rceil_{\boldsymbol{m}}^{M}|\boldsymbol{x}_{\boldsymbol{q}}^{*Q}, \boldsymbol{\theta}_{\boldsymbol{d}}^{*}) - H(pj^{M}(\lceil \boldsymbol{y} \rceil_{\boldsymbol{m}}^{M})|\boldsymbol{x}_{\boldsymbol{q}}^{*Q}, \boldsymbol{\theta}_{\boldsymbol{d}}^{*})| \le \epsilon_4, \tag{32}$$

and

$$|H(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta}_{\boldsymbol{d}}^{*}) - H(pj^{M}(\boldsymbol{y}_i)|\boldsymbol{x}_i, \boldsymbol{\theta}_{\boldsymbol{d}}^{*})| \le \epsilon_4, \tag{33}$$

where $pj^{M}(\boldsymbol{y}_i)$ is the projection of $\boldsymbol{y}_i$ on $\bar{\mathcal{Y}}$, and $\lceil \boldsymbol{y} \rceil_{\boldsymbol{m}}^{M} := (\lceil y \rceil_{m_1}^{M}, \dots, \lceil y \rceil_{m_K}^{M})$ is the rightmost vertex of hypercube $\mathcal{Y}_{\boldsymbol{m}}^{M}$. Further, because of the uniform continuity of $H(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta}_{\boldsymbol{d}}^{*})$ on $(\boldsymbol{x}_i, \boldsymbol{\theta}_{\boldsymbol{d}}^{*})$ within a compact support, for any $\epsilon_5 > 0$, we can choose sufficient large Q (to make h_Q^x small while $\mathcal{X}^{Q}$ covers $\bar{\mathcal{X}}$) and M (to make h_M^y small while $\mathcal{Y}^{M}$ covers $\bar{\mathcal{Y}}$) such that for any $\boldsymbol{y}_i \in \mathcal{Y}_{\boldsymbol{m}}^{M}$ and $\boldsymbol{q} \in R^{Q}(\boldsymbol{x}_i)$, we have:

$$|H(pj^{M}(\boldsymbol{y}_i)|\boldsymbol{x}_i, \boldsymbol{\theta}_{\boldsymbol{d}}^{*}) - H(pj^{M}(\lceil \boldsymbol{y} \rceil_{\boldsymbol{m}}^{M})|\boldsymbol{x}_{\boldsymbol{q}}^{*Q}, \boldsymbol{\theta}_{\boldsymbol{d}}^{*})| \le \epsilon_5. \tag{34}$$

Summarizing the above three equations, we have:

$$|H(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta}_{\boldsymbol{d}}^{*}) - H(\lceil \boldsymbol{y} \rceil_{\boldsymbol{m}}^{M}|\boldsymbol{x}_{\boldsymbol{q}}^{*Q}, \boldsymbol{\theta}_{\boldsymbol{d}}^{*})| \le 2\epsilon_4 + \epsilon_5. \tag{35}$$

Note that $F^{*(M,Q,t,u)}(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{w}_i)$ can be re-written as

$$\begin{aligned}
F^{*(M,Q,t,u)}(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{w}_i) &= \sum_{\boldsymbol{m}\in\mathcal{M}} \sum_{\boldsymbol{q}\in\mathcal{Q}} \sum_{\boldsymbol{d}\in\mathcal{D}^+} \gamma_{\boldsymbol{q}}^{(t)}(\boldsymbol{x}_i) \xi_{\boldsymbol{d}}^{(u)}(\boldsymbol{w}_i) H(\lceil \boldsymbol{y} \rceil_{\boldsymbol{m}}^{M}|\boldsymbol{x}_{\boldsymbol{q}}^{*Q}, \boldsymbol{\theta}_{\boldsymbol{d}}^{*}) 1\{\boldsymbol{y}_i \in \mathcal{Y}_{\boldsymbol{m}}^{M}\} \\
&= \sum_{\boldsymbol{m}\in\mathcal{M}} \sum_{\boldsymbol{q}\in R^{Q}(\boldsymbol{x}_i)} \sum_{\boldsymbol{d}\in\mathcal{D}^+} \gamma_{\boldsymbol{q}}^{(t)}(\boldsymbol{x}_i) \xi_{\boldsymbol{d}}^{(u)}(\boldsymbol{w}_i) H(\lceil \boldsymbol{y} \rceil_{\boldsymbol{m}}^{M}|\boldsymbol{x}_{\boldsymbol{q}}^{*Q}, \boldsymbol{\theta}_{\boldsymbol{d}}^{*}) 1\{\boldsymbol{y}_i \in \mathcal{Y}_{\boldsymbol{m}}^{M}\} + \mathcal{O}_1(\epsilon_3),
\end{aligned} \tag{36}$$

where $0 \le \mathcal{O}_1(\epsilon_3) \le \epsilon_3$. The last equality is resulted from Lemma 3. The approximation result is:

$$|\tilde{F}^{*(M,Q,t,u)}(\boldsymbol{y}|\boldsymbol{x}) - \tilde{F}^{**(u)}(\boldsymbol{y}|\boldsymbol{x})|$$

$$\leq \int \left| \prod_{i=1}^{N} F^{*(M,Q,t,u)}(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{w}_i) - \prod_{i=1}^{N} F^{**(u)}(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{w}_i) \right| d\Phi(\boldsymbol{w})$$
$$\leq N \int \left| F^{*(M,Q,t,u)}(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{w}_i) - F^{**(u)}(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{w}_i) \right| d\Phi(\boldsymbol{w})$$
$$\leq N \int \sum_{\boldsymbol{m}\in\mathcal{M}} \sum_{\boldsymbol{q}\in R^{Q}(\boldsymbol{x}_i)} \sum_{\boldsymbol{d}\in D^{+}} \gamma_{\boldsymbol{q}}^{(t)}(\boldsymbol{x}_i)\xi_{\boldsymbol{d}}^{(u)}(\boldsymbol{w}_i)$$
$$\times \left| H(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}_{\boldsymbol{d}}^{*}) - H(\lceil \boldsymbol{y} \rceil_{\boldsymbol{m}}^{M}|\boldsymbol{x}_{\boldsymbol{q}}^{*Q},\boldsymbol{\theta}_{\boldsymbol{d}}^{*}) \right| 1\{\boldsymbol{y}_i \in \mathcal{Y}_{\boldsymbol{m}}^{M}\} d\Phi(\boldsymbol{w}) + 2\epsilon_3$$
$$\leq N(2\epsilon_4 + \epsilon_5 + 2\epsilon_3), \tag{37}$$

where the second inequality is resulted from Lemma 1, the third and last inequalities are respectively resulted from Eqs. (35) and (36).

A.4. Step 4: Approximating Eq. (23) by Eq. (24)

Write $\tilde{F}^{(M,Q,t,u,v)}(\boldsymbol{y}|\boldsymbol{x})$ as in Eq. (24), where $F_0(\boldsymbol{y}_i;\boldsymbol{\psi}_{\boldsymbol{m}}^{M(v)})$ in $F^{(M,Q,t,u,v)}(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{w}_i)$ is chosen in the way that $F_0(\boldsymbol{y}_i;\boldsymbol{\psi}_{\boldsymbol{m}}^{M(v)}) \xrightarrow{D} 1\{\boldsymbol{y}_i \geq \lfloor \boldsymbol{y} \rfloor_{\boldsymbol{m}}^{M}\}$ as $v \to \infty$. Due to the distributional convergence as well as $\mathcal{M}$ is a finite set, for any $\epsilon_6 > 0$ we can find a sufficient large v such that for every $\boldsymbol{m} \in \mathcal{M}$, we have:

$$|F_0(\boldsymbol{y}_i;\boldsymbol{\psi}_{\boldsymbol{m}}^{M(v)}) - 1\{\boldsymbol{y}_i \geq \lfloor \boldsymbol{y} \rfloor_{\boldsymbol{m}}^{M}\}| \leq \epsilon_6 + \sum_{k=1}^{K} 1\{y_{ik} \in \mathcal{L}_k^{\delta^*}(\lfloor y \rfloor_{m_k}^{M})\}, \tag{38}$$

where δ^* is chosen to be $0 < \delta^* < h_M^y/2$, and $\mathcal{L}_k^{\delta^*}(\lfloor y \rfloor_{m_k}^{M})$ represents non-overlapping intervals for $m_k = -(M+1), \ldots, (M+1)$ with

$$\mathcal{L}_k^{\delta^*}(\lfloor y \rfloor_{m_k}^{M}) = \begin{cases} [\lfloor y \rfloor_{m_k}^{M} - \delta^*, \lfloor y \rfloor_{m_k}^{M} + \delta^*], & \text{if } m_k > -(M+1) \\ (-\infty, \lfloor y \rfloor_{m_k}^{M} - 2\delta^*], & m_k = -(M+1) \end{cases} \tag{39}$$

Note that the rightmost term in Eq. (38) is to control for the fact that the weak convergence of F_0 to the indicator is not uniform when y_{ik} is close to $\lfloor y \rfloor_{m_k}^{M}$. Consider the bound

$$|F^{(M,Q,t,u,v)}(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{w}_i) - F^{*(M,Q,t,u)}(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{w}_i)|$$
$$\leq \sum_{\boldsymbol{m}\in\mathcal{M}} \sum_{\boldsymbol{q}\in Q} \sum_{\boldsymbol{d}\in D^{+}} \gamma_{\boldsymbol{q}}^{(t)}(\boldsymbol{x}_i)\xi_{\boldsymbol{d}}^{(u)}(\boldsymbol{w}_i) H(\mathcal{Y}_{\boldsymbol{m}}^{M}|\boldsymbol{x}_{\boldsymbol{q}}^{*Q},\boldsymbol{\theta}_{\boldsymbol{d}}^{*}) |F_0(\boldsymbol{y}_i;\boldsymbol{\psi}_{\boldsymbol{m}}^{M(v)}) - 1\{\boldsymbol{y}_i \geq \lfloor \boldsymbol{y} \rfloor_{\boldsymbol{m}}^{M}\}|$$
$$\leq \left\{ \max_{\boldsymbol{q}\in Q;\boldsymbol{d}\in D^{+}} \sum_{\boldsymbol{m}\in\mathcal{M}} \sum_{k=1}^{K} H(\mathcal{Y}_{\boldsymbol{m}}^{M}|\boldsymbol{x}_{\boldsymbol{q}}^{*Q},\boldsymbol{\theta}_{\boldsymbol{d}}^{*}) 1\{y_{ik} \in \mathcal{L}_k^{\delta^*}(\lfloor y \rfloor_{m_k}^{M})\} \right\} + \epsilon_6$$
$$= \sum_{k=1}^{K} \left\{ \max_{\boldsymbol{q}\in Q;\boldsymbol{d}\in D^{+}} \sum_{m_k=-(M+1),\ldots,(M+1)} H_k(\mathcal{Y}_{k,m_k}|\boldsymbol{x}_{\boldsymbol{q}}^{*Q},\boldsymbol{\theta}_{\boldsymbol{d}}^{*}) 1\{y_{ik} \in \mathcal{L}_k^{\delta^*}(\lfloor y \rfloor_{m_k}^{M})\} \right\} + \epsilon_6. \tag{40}$$

Since $\mathcal{L}_k^{\delta^*}(\lfloor y \rfloor_{m_k}^{M})$ is non-overlapping for $m_k = -(M+1), \ldots, (M+1)$, only one term in the summation of Eq. (40) is non-zero. Since H_k is a continuous distribution, for any $\epsilon_7 > 0$, we have $H_k(\mathcal{Y}_{k,m_k}|\boldsymbol{x}_{\boldsymbol{q}}^{*Q},\boldsymbol{\theta}_{\boldsymbol{d}}^{*}) \leq \epsilon_7$ for any $\boldsymbol{q} \in Q$, $\boldsymbol{d} \in D^{+}$ and $m_k = -(M+1), \ldots, (M+1)$ given that M is sufficiently large. Finally, using the same proof idea as Eq. (37), we have

$$|\tilde{F}^{(M,Q,t,u,v)}(\boldsymbol{y}|\boldsymbol{x}) - \tilde{F}^{*(M,Q,t,u)}(\boldsymbol{y}|\boldsymbol{x})| \leq N(K\epsilon_7 + \epsilon_6). \tag{41}$$

In summary, based on Eqs. (27), (29) (37) and (41), Theorem 1 holds because for sufficiently large M, Q, t and v, the following inequality holds uniformly for each $\boldsymbol{x}_i$ falling into a compact covariates space:

$$|\tilde{F}^{(M,Q,t,u\to\infty,v)}(\boldsymbol{y}|\boldsymbol{x}) - \tilde{H}(\boldsymbol{y}|\boldsymbol{x})| \leq (N\epsilon_2 + \epsilon_1) + \mathcal{O}_1(\epsilon_1) + N(2\epsilon_4 + \epsilon_5 + 2\epsilon_3) + N(K\epsilon_7 + \epsilon_6), \tag{42}$$

where ϵ_1 to ϵ_7 can be chosen to be arbitrarily small, and any parameters chosen in Steps 1 to 4 are independent of N, $\boldsymbol{S} = (S_1, \ldots, S_L)$ and $\boldsymbol{c}(\cdot)$.

Appendix B. Proof of Theorem 2

Other than the notational differences, the proof ideas of Theorems 1 and 2 are substantially similar. Precisely, the 4-step framework used to prove Theorem 1 also applies to prove Theorem 2. As a result, we only present a sketch proof of Theorem 2, with an emphasis on the key differences of proof techniques between the two theorems. Unless specified otherwise, the notations adopted in this proof section is the same as those defined in Appendix A. Analogous to Eqs. (20) and (21), in Step 1 we examine an approximation bound between the following two equations:

$$\tilde{H}(\boldsymbol{y}|\boldsymbol{x}) = \int_{\tilde{\boldsymbol{\theta}}} \left[\prod_{i\in\mathcal{I}} H(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}_i) \right] dG(\tilde{\boldsymbol{\theta}}), \tag{43}$$

and

$$\tilde{H}^{*}(\boldsymbol{y}|\boldsymbol{x}) = \sum_{\tilde{\boldsymbol{d}}\in\tilde{D}} \prod_{i\in\mathcal{I}} H(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}_{\boldsymbol{d}^{(i_L)}}^{*}) G(\tilde{\boldsymbol{\Theta}}_{\tilde{\boldsymbol{d}}}). \tag{44}$$

Similar to Appendix A.1, we choose compact spaces $\{\bar{\boldsymbol{\Theta}}_l \in \boldsymbol{\Theta}\}_{l=1,\ldots,L}$ such that $\mathbb{P}(\cap_{l=1}^{L} \cap_{i_l\in\mathcal{I}_l} \{\boldsymbol{\theta}_{i_l} \in \bar{\boldsymbol{\Theta}}_l\}) \geq 1 - NL\epsilon_1$, where $\mathcal{I}_l = \{\boldsymbol{i}_l : i_1 = 1, \ldots, N_0; \ldots; i_l = 1, \ldots, N_{i_{l-1}}\}$. Also, partition granular subspaces $\{\boldsymbol{\Theta}_{l,d_l}\}_{d_l=1,\ldots,D_l}$ of $\bar{\boldsymbol{\Theta}}_l$ and define the interior points $\{\boldsymbol{\theta}_{l,d_l}^{*}\}_{d_l=1,\ldots,D_l}$ analogous to Appendix A.1. Define $\tilde{D} = \prod_{l=1}^{L} \prod_{i_l\in\mathcal{I}_l} \mathcal{D}_l^{(i_l)}$ with $\mathcal{D}_l^{(i_l)} = \{1, \ldots, D_l\}$, $\tilde{\boldsymbol{d}} = \{d_l^{(i_l)}\}_{l=1,\ldots,L;i_l\in\mathcal{I}_l}$ with $d_l^{(i_l)} \in \mathcal{D}_l^{(i_l)}$, $\boldsymbol{d}^{(i_L)} = \{d_l^{(i_l)}\}_{l=1,\ldots,L}$, and $\tilde{\boldsymbol{\Theta}}_{\tilde{\boldsymbol{d}}} = \prod_{l=1}^{L} \prod_{i_l\in\mathcal{I}_l} \boldsymbol{\Theta}_{l,d_l^{(i_l)}}$. Then, using the same idea as Eqs. (25) to (27), we obtain an arbitrarily small approximation error bound between Eqs. (43) and (44).

In Step 2, we define the following function analogous to Eq. (22) and derive its error bound for approximating Eq. (44):

$$\tilde{F}^{**(u)}(\boldsymbol{y}|\boldsymbol{x}) = \int \prod_{i\in\mathcal{I}} F^{**(u)}(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{w}_i) d\Phi(\boldsymbol{w}) \tag{45}$$

with $F^{**(u)}(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{w}_i) = \sum_{\boldsymbol{d}\in D^{+}} \xi_{\boldsymbol{d}}^{(u)}(\boldsymbol{w}_i) H(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta}_{\boldsymbol{d}}^{*})$. Here, we choose $\xi_{\boldsymbol{d}}^{(u)}(\boldsymbol{w}_i)$ in a different way as Eq. (28), which is crucial to cater for the dependencies of random effects across levels, as follows:

$$\xi_{\boldsymbol{d}}^{(u)}(\boldsymbol{w}_i) = \exp\{\sum_{l=1}^{L} u^{1/l}(\tilde{\beta}_{d_l,0} + \tilde{\beta}_{d_l,1} w_{il})\} / \sum_{\boldsymbol{d}'\in D^{+}} \exp\{\sum_{l=1}^{L} u^{1/l}(\tilde{\beta}_{d'_l,0} + \tilde{\beta}_{d'_l,1} w_{il})\}, \tag{46}$$

where $D = \prod_{l=1}^{L}\{1, \ldots, D_l\}$ and $D^{+} = \prod_{l=1}^{L}\{0, 1, \ldots, D_l + 1\}$ are exactly the same as those defined in Appendix A.2, and $\boldsymbol{d}_l = (d_1, \ldots, d_l)$ and $\boldsymbol{d}'_l = (d'_1, \ldots, d'_l)$ with $\boldsymbol{d} = \boldsymbol{d}_L$ (and $\boldsymbol{d}' = \boldsymbol{d}'_L$).

In contrast to Appendix A.2, we construct intervals $\mathcal{W}_{l,\boldsymbol{d}_l}$ such that $\cup_{d_l=0}^{D_l+1} \mathcal{W}_{l,\boldsymbol{d}_l} = \mathbb{R}$ for any $\boldsymbol{d}_{l-1} \in \prod_{l'=1}^{l-1}\{1, \ldots, D_{l'}\}$ and $\boldsymbol{\Phi}_l(\mathcal{W}_{l,\boldsymbol{d}_l}) = G_l(\boldsymbol{\Theta}_{l,d_l}|\boldsymbol{\Theta}_{1,d_1}, \ldots, \boldsymbol{\Theta}_{l-1,d_{l-1}})$, which represents the probability of level-l random effect belongs to $\boldsymbol{\Theta}_{l,d_l}$ conditioned on the corresponding upper level random effects fall into $(\boldsymbol{\Theta}_{1,d_1}, \ldots, \boldsymbol{\Theta}_{l-1,d_{l-1}})$. The following lemma is analogous to Lemma 2:

Lemma 5. *There exists parameters* $\{(\tilde{\beta}_{d_l,0}, \tilde{\beta}_{d_l,1})\}_{\boldsymbol{d}\in D^{+};l=1,\ldots,L}$ *of* $\xi_{\boldsymbol{d}}^{(u)}(\boldsymbol{w}_i)$ *such that* $\xi_{\boldsymbol{d}}^{(u)}(\boldsymbol{w}_i) \xrightarrow{u\to\infty} \prod_{l=1}^{L} 1_{w_{il}}^{*}(\mathcal{W}_{l,\boldsymbol{d}_l})$ *for every* $\boldsymbol{d} \in D^{+}$.

Proof. Similar to the proof of Lemma 2, we choose suitable parameters such that $\tilde{\beta}_{d_l,0} + \tilde{\beta}_{d_l,1} w_{il} > \max_{d'_l\neq d_l} \tilde{\beta}_{d^{*}_l,0} + \tilde{\beta}_{d^{*}_l,1} w_{il}$ if and only if $w_{il} \in \mathcal{W}_{l,\boldsymbol{d}_l}^{*}$ for every $\boldsymbol{d} \in D^{+}$ and $l = 1, \ldots, L$, where $\mathcal{W}_{l,\boldsymbol{d}_l}^{*}$ is the interior of $\mathcal{W}_{l,\boldsymbol{d}_l}$ and $\boldsymbol{d}_l^{*} = (d_1, \ldots, d_{l-1}, d'_l)$. Observe the expression $\sum_{l=1}^{L} u^{1/l}(\tilde{\beta}_{d_l,0} + \tilde{\beta}_{d_l,1} w_{il})$ in Eq. (46) where the term corresponding to a higher level factor dominates when u is large. Therefore, for sufficient large u, we have $\sum_{l=1}^{L} u^{1/l}(\tilde{\beta}_{d_l,0} + \tilde{\beta}_{d_l,1} w_{il}) > \max_{\boldsymbol{d}\neq\boldsymbol{d}'} \sum_{l=1}^{L} u^{1/l}(\tilde{\beta}_{d'_l,0} + \tilde{\beta}_{d'_l,1} w_{il})$ for $\boldsymbol{w}_i$ satisfying $w_{il} \in \mathcal{W}_{l,\boldsymbol{d}_l}^{*}$ for every $l = 1, \ldots, L$. Following the same logic as Lemma 2, the result follows. □

The approximation bound between Eqs. (44) and (45) can be obtained using the same logic as that outlined by Eq. (29), where we further note that $\boldsymbol{\Phi}(\mathcal{W}_{\tilde{\boldsymbol{d}}}) = \prod_{l=1}^{L} \prod_{i_l\in\mathcal{I}_l} \boldsymbol{\Phi}_l(\mathcal{W}_{l,\boldsymbol{d}_l^{(i_l)}}) = \prod_{l=1}^{L} \prod_{i_l\in\mathcal{I}_l} G_l(\boldsymbol{\Theta}_{l,\boldsymbol{d}_l^{(i_l)}}|\boldsymbol{\Theta}_{1,\boldsymbol{d}_1^{(i_1)}}, \ldots, \boldsymbol{\Theta}_{l-1,\boldsymbol{d}_{l-1}^{(i_{l-1})}}) = G(\tilde{\boldsymbol{\Theta}}_{\tilde{\boldsymbol{d}}})$ with $\mathcal{W}_{\tilde{\boldsymbol{d}}} = \prod_{l=1}^{L} \prod_{i_l\in\mathcal{I}_l} \mathcal{W}_{l,\boldsymbol{d}_l^{(i_l)}}$.

Step 3 and 4 involve evaluations of the following expressions in analogous to Eqs. (23) and (24):

$$\tilde{F}^{*(M,Q,t,u)}(\boldsymbol{y}|\boldsymbol{x}) = \int \prod_{i\in\mathcal{I}} F^{*(M,Q,t,u)}(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{w}_i) d\Phi(\boldsymbol{w}) \tag{47}$$

with $F^{*(M,Q,t,u)}(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{w}_i) = \sum_{m\in\mathcal{M}}\sum_{q\in\mathcal{Q}}\sum_{d\in\mathcal{D}^+}\pi_j^{(t,u)}(\boldsymbol{x}_i,\boldsymbol{w}_i;\tilde{\boldsymbol{\alpha}}_q^Q,\tilde{\boldsymbol{\beta}}_d)1\{\boldsymbol{y}_i \geq \lfloor \boldsymbol{y} \rfloor_m^M\}$,

$$\tilde{F}^{(M,Q,t,u,v)}(\boldsymbol{y}|\boldsymbol{x}) = \int \prod_{i\in\mathcal{I}} F^{(M,Q,t,u,v)}(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{w}_i)d\boldsymbol{\Phi}(\boldsymbol{w}) \tag{48}$$

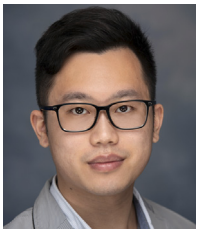
with $F^{(M,Q,t,u,v)}(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{w}_i) = \sum_{m\in\mathcal{M}}\sum_{q\in\mathcal{Q}}\sum_{d\in\mathcal{D}^+}\pi_j^{(t,u)}(\boldsymbol{x}_i,\boldsymbol{w}_i;\tilde{\boldsymbol{\alpha}}_q^Q,\tilde{\boldsymbol{\beta}}_d)F_0(\boldsymbol{y}_i;\boldsymbol{\psi}_m^{M(v)})$. The derivation techniques here are exactly the same as those presented in Appendices A.3 and A.4, so this part of the proof is omitted.

Data availability

No data was used for the research described in the article.

References

[1] R.A. Jacobs, M.I. Jordan, S.J. Nowlan, G.E. Hinton, Adaptive mixtures of local experts, Neural Comput. 3 (1) (1991) 79–87.
[2] M.I. Jordan, R.A. Jacobs, Hierarchical mixtures of experts and the EM algorithm, Neural Comput. 6 (2) (1994) 181–214.
[3] G.J. McLachlan, D. Peel, Finite Mixture Models, John Wiley & Sons, 2000.
[4] S.E. Yuksel, J.N. Wilson, P.D. Gader, Twenty years of mixture of experts, IEEE Trans. Neural Netw. Learn. Syst. 23 (8) (2012) 1177–1193.
[5] S. Masoudnia, R. Ebrahimpour, Mixture of experts: A literature survey, Artif. Intell. Rev. 42 (2) (2014) 275–293.
[6] H.D. Nguyen, F. Chamroukhi, Practical and theoretical aspects of mixture-of-experts modeling: An overview, Wiley Interdiscip. Reviews: Data Min. Knowl. Discov. 8 (4) (2018) e1246.
[7] L. Xu, M.I. Jordan, G.E. Hinton, An alternative model for mixtures of experts, Adv. Neural Inf. Process. Syst. (1995) 633–640.
[8] S. Ingrassia, S.C. Minotti, G. Vittadini, Local statistical modeling via a cluster-weighted approach with elliptical distributions, J. Classification 29 (3) (2012) 363–401.
[9] J. Geweke, M. Keane, Smoothly mixing regressions, J. Econometrics 138 (1) (2007) 252–290.
[10] C. Riquelme, J. Puigcerver, B. Mustafa, M. Neumann, R. Jenatton, A. Susano Pinto, D. Keysers, N. Houlsby, Scaling vision with sparse mixture of experts, Adv. Neural Inf. Process. Syst. 34 (2021) 8583–8595.
[11] H. Hazimeh, Z. Zhao, A. Chowdhery, M. Sathiamoorthy, Y. Chen, R. Mazumder, L. Hong, E. Chi, Dselect-k: Differentiable selection in the mixture of experts with applications to multi-task learning, Adv. Neural Inf. Process. Syst. 34 (2021) 29335–29347.
[12] M.I. Jordan, R.A. Jacobs, Hierarchies of adaptive experts, in: Advances in Neural Information Processing Systems, 1992, pp. 985–992.
[13] H.D. Nguyen, G.J. McLachlan, Laplace mixture of linear experts, Comput. Statist. Data Anal. 93 (2016) 177–191.
[14] F. Chamroukhi, Robust mixture of experts modeling using the t distribution, Neural Netw. 79 (2016) 20–36.
[15] F. Chamroukhi, Skew t mixture of experts, Neurocomputing 266 (2017) 390–408.
[16] T.C. Fung, A.L. Badescu, X.S. Lin, A new class of severity regression models with an application to IBNR prediction, N. Am. Actuar. J. 25 (2) (2021) 206–231.
[17] B. Grun, F. Leisch, FlexMix version 2: Finite mixtures with concomitant variables and varying and constant parameters, J. Stat. Softw. 28 (4) (2008) 1–35.
[18] T.C. Fung, A.L. Badescu, X.S. Lin, A class of mixture of experts models for general insurance: Application to correlated claim frequencies, ASTIN Bull.: J. IAA 49 (3) (2019) 647–688.
[19] A.J. Zeevi, R. Meir, V. Maiorov, Error bounds for functional approximation and estimation using mixtures of experts, IEEE Trans. Inform. Theory 44 (3) (1998) 1010–1025.
[20] W. Jiang, M.A. Tanner, On the approximation rate of hierarchical mixtures-of-experts for generalized linear models, Neural Comput. 11 (5) (1999) 1183–1198.
[21] H.D. Nguyen, L.R. Lloyd-Jones, G.J. McLachlan, A universal approximation theorem for mixture-of-experts models, Neural Comput. 28 (12) (2016) 2585–2593.
[22] H.D. Nguyen, F. Chamroukhi, F. Forbes, Approximation results regarding the multiple-output Gaussian gated mixture of linear experts model, Neurocomputing 366 (2019) 208–214.
[23] W. Jiang, M.A. Tanner, Hierarchical mixtures-of-experts for exponential family regression models: Approximation and maximum likelihood estimation, Ann. Statist. (1999) 987–1011.
[24] E.F. Mendes, W. Jiang, On convergence rates of mixtures of polynomial experts, Neural Comput. 24 (11) (2012) 3025–3051.
[25] A. Norets, et al., Approximation of conditional densities by smooth mixtures of regressions, Ann. Statist. 38 (3) (2010) 1733–1766.
[26] A. Norets, J. Pelenis, Posterior consistency in conditional density estimation by covariate dependent mixtures, Econometric Theory 30 (3) (2014) 606–646.
[27] H.D. Nguyen, T. Nguyen, F. Chamroukhi, G.J. McLachlan, Approximations of conditional probability density functions in lebesgue spaces via mixture of experts models, J. Stat. Distrib. Appl. 8 (1) (2021) 1–15.
[28] H. Tijms, Stochastic Models: An Algorithm Approach, John Wiley, 1994.
[29] L. Breuer, D. Baum, An Introduction to Queueing Theory and Matrix-Analytic Methods, Springer Science & Business Media, 2005.
[30] T.C. Fung, A.L. Badescu, X.S. Lin, A class of mixture of experts models for general insurance: Theoretical developments, Insurance Math. Econom. 89 (2019) 111–127.
[31] H. Goldstein, Multilevel Statistical Models, vol. 922, John Wiley & Sons, 2011.
[32] M. Aitkin, N. Longford, Statistical modelling issues in school effectiveness studies, J. R. Stat. Soc. Ser. A (General) 149 (1) (1986) 1–26.
[33] H. Goldstein, Multilevel mixed linear model analysis using iterative generalized least squares, Biometrika 73 (1) (1986) 43–56.
[34] E.W. Frees, J.-S. Kim, Multilevel model prediction, Psychometrika 71 (1) (2006) 79–104.
[35] G. Molenberghs, G. Verbeke, C.G. Demétrio, A.M. Vieira, A family of generalized linear models for repeated measures with normal and conjugate random effects, Statist. Sci. 25 (3) (2010) 325–347.
[36] J.-P. Boucher, M. Denuit, Fixed versus random effects in Poisson regression models for claim counts: A case study with motor insurance, ASTIN Bull. J. IAA 36 (1) (2006) 285–301.
[37] C. McGilchrist, Estimation in generalized mixed models, J. R. Stat. Soc. Ser. B Stat. Methodol. 56 (1) (1994) 61–69.
[38] M. Davidian, A.R. Gallant, The nonlinear mixed effects model with a smooth random effects density, Biometrika 80 (3) (1993) 475–488.
[39] T.G. Gregoire, O. Schabenberger, A non-linear mixed-effects model to predict cumulative bole volume of standing trees, J. Appl. Stat. 23 (2–3) (1996) 257–272.
[40] B. Bakker, T. Heskes, Task clustering and gating for bayesian multitask learning, J. Mach. Learn. Res. 4 (May) (2003) 83–99.
[41] S.-K. Ng, G. McLachlan, K.K. Yau, A.H. Lee, Modelling the distribution of ischaemic stroke-specific survival time using an EM-based mixture approach with random effects adjustment, Stat. Med. 23 (17) (2004) 2729–2744.
[42] S.-K. Ng, G.J. McLachlan, K. Wang, L. Ben-Tovim Jones, S.-W. Ng, A mixture model with random-effects components for clustering correlated gene-expression profiles, Bioinformatics 22 (14) (2006) 1745–1752.
[43] K.K. Yau, A.H. Lee, A.S. Ng, Finite mixture regression model with random effects: Application to neonatal hospital length of stay, Comput. Statist. Data Anal. 41 (3–4) (2003) 359–366.
[44] S.-K. Ng, G.J. McLachlan, Extension of mixture-of-experts networks for binary classification of hierarchical data, Artif. Intell. Med. 41 (1) (2007) 57–67.
[45] S.-K. Ng, G.J. McLachlan, Mixture models for clustering multilevel growth trajectories, Comput. Statist. Data Anal. 71 (2014) 43–51.
[46] S.C. Tseung, I.W. Chan, T.C. Fung, A.L. Badescu, X.S. Lin, Improving risk classification and ratemaking using mixture-of-experts models with random effects, J. Risk Insurance 90 (3) (2023) 789–820.



Tsz Chai Fung is an Assistant Professor in the Maurice R. Greenberg School of Risk Science, Georgia State University (GSU). Prior to joining GSU, he worked as a Postdoctoral Researcher at ETH Zurich during 2020-2021 and earned a Ph.D. Statistics degree at the University of Toronto in 2020. His current research interests include probabilistic neural network, universal approximation theory, statistical modeling, and inference techniques with applications in business and economics problems, including actuarial science and risk management.

Spark C. Tseung is a Ph.D. student in actuarial science at the Department of Statistical Sciences, University of Toronto (U of T). He is interested in applying novel statistical and machine learning methods to classical actuarial problems such as pricing and claims reserving. Prior to joining U of T, he obtained a bachelor's degree in actuarial science from the University of Hong Kong.