A non-asymptotic theory for model selection in a high-dimensional mixture of experts via joint rank and variable selection

Trungtin Nguyen, Dung Ngoc Nguyen, Hien Duy Nguyen, Faicel Chamroukhi

1–45, 2023

A non-asymptotic theory for model selection in high-dimensional mixture of experts via joint rank and variable selection

TrungTin Nguyen TRUNG-TIN.NGUYEN@INRIA.FR
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, Inria Grenoble Rhone-Alpes, 655 av. de l'Europe, 38335 Montbonnot, France.

Dung Ngoc Nguyen NGOCDUNG.NGUYEN@UNIPD.IT
Department of Statistical Sciences, University of Padova, Italy.

Hien D Nguyen H.NGUYEN7@UQ.EDU.AU
School of Mathematics and Physics, University of Queensland, St. Lucia, Brisbane, Australia.

Faicel Chamroukhi FAICEL.CHAMROUKHI@IRT-SYSTEMX.FR
IRT SystemX, Palaiseau, France.

Abstract

Mixture of experts (MoE) models are among the most popular and interesting combination techniques, with great potential for improving the performance of machine learning and statistical learning systems. We are the first to consider a polynomial softmax-gated block-diagonal mixture of experts (PSGaBloME) model for the identification of potentially nonlinear regression relationships for complex and high-dimensional heterogeneous data, where the number of explanatory and response variables can be much larger than the sample size and possibly hidden graph-structured interactions exist. These PSGaBloME models are characterized by several hyperparameters, including the number of mixture components, the complexity of softmax gating networks and Gaussian mean experts, and the hidden block-diagonal structures of covariance matrices. We contribute a non-asymptotic theory for model selection of such complex hyperparameters with the help of the slope heuristic approach in a penalized maximum likelihood estimation (PMLE) framework. In particular, we establish a non-asymptotic risk bound on the PMLE, which takes the form of an oracle inequality, given lower bound assumptions on the penalty function. Furthermore, we propose two Lasso–MLE–rank procedures, based on a new generalized expectation–maximization algorithm, to tackle the estimation problem of the collection of PSGaBloME models.

Keywords: Mixture of experts, mixture of regressions, dimensionality reduction, low rank estimation, non-asymptotic theory, concentration inequality, oracle inequality, variable selection.

1. Introduction

In this work, our primary objective is to identify potential complex non-linear relationships between high-dimensional heterogeneous outputs (also referred to as target or response variables) and inputs (also termed explanatory or predictor variables), where the number of explanatory and response variables can be far greater than the sample size, and where there are possibly hidden interactions in the graphical structure. This involves performing regression, clustering, and model selection, simultaneously. Mixture of experts (MoE) models, introduced by Jacobs et al. (1991); Jordan and Jacobs (1994), are extremely well suited for the task, described. Indeed, these flexible models decompose the prediction model by a combination of the gating models and expert models, both depending on the input variables. Furthermore, the MoE is a specific instance of conditional computation Bengio (2013), where different model experts are responsible for different regions of the input space. Thus,

by applying only a subset of parameters to each example, the MoE can increase model capacity while keeping training and inference costs roughly constant. In particular, it has been popularized for its universal approximation properties in the context of mixture models (MM) (Genovese and Wasserman, 2000; Rakhlin et al., 2005; Nguyen, 2013; Ho and Nguyen, 2016a,b; Nguyen et al., 2020b, 2022), mixture of regressions (MoR) models (Do et al., 2022; Ho et al., 2022), and more generally (Jiang and Tanner, 1999b; Norets, 2010; Nguyen et al., 2016, 2019, 2021). The reader is referred to Yuksel et al. (2012); Masoudnia and Ebrahimpour (2014); Nguyen and Chamroukhi (2018); Nguyen (2021) for reviews on this topic.

To the best of our knowledge, we are the first to propose one of the most general MoE models for high-dimensional multivariate multiple regression, when the number of explanatory variables can be much larger than the sample size, and when there are possibly hidden graphically structured interactions between responses.

1.1. MoE models for heterogeneous complex data

Definition of MoE models. We are interested in estimating the law of a multivariate random variable $\mathbf{Y} \in \mathcal{Y} \subset \mathbb{R}^Q$, conditionally on $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^P$. Here and subsequently, given any $N \in \mathbb{N}^\star$, $\left(\mathbf{X}_{[N]}, \mathbf{Y}_{[N]}\right) \equiv (\mathbf{X}_n, \mathbf{Y}_n)_{n\in[N]}$, $[N] := \{1, \ldots, N\}$, denotes a random sample, and $\mathbf{x}$ and $\mathbf{y}$ stand for the observed values of the random variables $\mathbf{X}$ and $\mathbf{Y}$, respectively. The following assumptions will be needed throughout the paper. We assume that the covariates $\mathbf{X}$ are independent but not necessarily identically distributed. The assumptions on the responses $\mathbf{Y}$ are stronger. Namely, conditional on $\mathbf{X}_{[N]}$, $\mathbf{Y}_{[N]}$ are independent, and each $\mathbf{Y}$ follows a law with true (but unknown) PDF $s_0\left(\cdot \mid \mathbf{X} = \mathbf{x}\right)$. Motivated by universal approximation theorems for MoEs, we propose the *softmax-gated block-diagonal MoE* (SGaBloME) models to estimate s_0:

$$s_{\boldsymbol{\psi}_K}(\mathbf{y} \mid \mathbf{x}) = \sum_{k=1}^{K} g_k(\mathbf{w}(\mathbf{x}))\phi(\mathbf{y}, \mathbf{v}_k(\mathbf{x}), \boldsymbol{\Sigma}_k(B_k)), \quad g_k(\mathbf{w}(\mathbf{x})) = \frac{\exp(w_k(\mathbf{x}))}{\sum_{l=1}^{K}\exp(w_l(\mathbf{x}))}, \tag{1}$$

with unknown functional parameters $\boldsymbol{\psi}_K = (w_k, \mathbf{v}_k, \boldsymbol{\Sigma}_k(B_k))_{k\in[K]}$. Here we call g_k a soft-max gating network corresponding to the weight, $\mathbf{w} = (w_1, \ldots, w_K)$, functions of $\mathbf{x}$, and $\phi(\cdot)$ a Gaussian expert with the mean function $\mathbf{v}_k$, of $\mathbf{x}$ on cluster k-th. In particular, we define $\mathcal{S}$ a set of *polynomial SGaBloME (PSGaBloME)* models with $s_{\boldsymbol{\psi}_K}$ satisfying (1) and

$$w_k(x) = \omega_{k0} + \sum_{d=1}^{D_W} \boldsymbol{\omega}_{kd}^T \mathbf{x}^d, \;\; \text{with } \omega_{k0} \in \mathbb{R}, \; \boldsymbol{\omega}_{kd} \in \mathbb{R}^P, \tag{2}$$

$$\mathbf{v}_k(x) = \boldsymbol{\upsilon}_{k0} + \sum_{d=1}^{D_V} \boldsymbol{\Upsilon}_{kd} \mathbf{x}^d, \;\; \text{with } \boldsymbol{\upsilon}_{k0} \in \mathbb{R}^Q, \; \boldsymbol{\Upsilon}_{kd} \in \mathbb{R}^{Q\times P}. \tag{3}$$

Here, $\mathbf{x}^d$ is a vector of all components of $\mathbf{x}$ with power d, D_W and D_V are the degrees of the weights and means of $\mathbf{x}$, respectively.

Motivation for block-diagonal covariance matrices. It is worth mentioning that the block-diagonal covariance matrix $\boldsymbol{\Sigma}_k(B_k)$ depends on the block structure $\mathbf{B} = (B_k)_{k\in[K]}$ partitioning the output index set $[Q]$ for each cluster. This structure is not only to trade-off between complexity and sparsity, but is also motivated by some real applications, where one wants to perform prediction on data sets with heterogeneous observations and hidden graph-structured interactions between

outputs. For example, for gene expression datasets where, conditional on the phenotypic response, genes only interact with a few other genes, *i.e.,* there are small modules of correlated genes (see *e.g.,* Devijver et al., 2017; Devijver and Gallopin, 2018 for more details). Furthermore, Broto et al. (2022) estimated a high-dimensional block-diagonal covariance matrix of Gaussian variables for estimating Shapley effects. While Andrade et al. (2020) proposed a robust Bayesian model selection for variable clustering with the Gaussian graphical models.

Motivation for polynomial regression. With some restrictions on SGaBloME models, some authors have developed methods to deal with high-dimensional regression problems. Notably, Devijver (2017a) focused on a mixture of Gaussian linear regression (MoGLR) model, where the gating networks g_k are independent of the inputs for all subpopulations $k \in [K]$. Whereas Chamroukhi and Huynh (2018, 2019) considered MoE for multiple regression models, but only with the univariate target variable. In particular, in (1) they all consider only linear functions for w_k and $\mathbf{v}_k$, which limits the capacity of MoE models. For example, in the context of convolutional neural networks (CNNs), Chen et al. (2022) has empirically found that the mixture of linear experts outperforms the single expert, but is still significantly worse than the mixture of non-linear experts. This motivates us to integrate nonlinearities into (1) by defining w_k and $\mathbf{v}_k$ as *linear combinations of bounded functions (LinBo)*. We call this model *LinBoSGaBloME*, it contains two special cases *LinBoSGaME* and *SGaME* models without considering block-diagonal covariance matrices, where SGaME is an affine instance. For a comprehensive classification and nomenclature of MoE models with soft-max gating networks, please refer to Figure 1. If P and Q are not too large, we do not need to select relevant variables and/or use rank sparse models. Then we can work on LinBoSGaME as in Montuelle and Le Pennec (2014). However, to deal with high-dimensional data and to simplify the interpretation of sparsity, we propose to use the PSGaBloME model. Note that this is one of the simplest models among the class of LinBoSGaBloME to explore the presence of nonlinearities. For example, Punzo (2014) considered polynomial Gaussian cluster-weighted MoE models, which outperform the polynomial Gaussian MoR and allow for possible nonlinear dependencies in the expert components by considering a polynomial regression. For polynomial MoR, Fang et al. (2022) considered a likelihood ratio test for determining if there is a higher-degree polynomial term in one of the components, leading to a better model compared to linear MoR. In particular, on the convergence rates of PSGaBloME models, we refer to Mendes and Jiang (2012) to discuss the optimal convergence rate on a MoE structure where K experts experts, where each expert is associated with a polynomial regression model of order D_V.

Model selection problem for PSGaBloME. On the one hand, estimation in SGaBloME models can be performed using a well-known expectation maximisation (EM) algorithm (Dempster et al., 1977; McLachlan and Krishnan, 1997), which enjoys global convergence in the context of MoR (Kwon et al., 2019; Klusowski et al., 2019). However, it crucially depends on and requires data-driven hyper-parameter choices, including the number of mixture components (or clusters), the degree of complexity of each soft-max gating network and each Gaussian expert mean function, and the hidden block-diagonal structures of the covariance matrices. It is worth noting that the hyper-parameter choice of data-driven learning algorithms belongs to the class of model selection problems that have received much attention in statistics and machine learning over the last 50 years (Akaike, 1974; Mallows, 1973; Burnham and Anderson, 2002; Massart, 2007; Arlot, 2019). More precisely, given a set of models, how do we select the one with the lowest possible risk from the data? It should be noted that penalisation is one of the main strategies proposed for model selection.

The idea is to select the estimator that minimizes the sum of the empirical risk and some penalty terms corresponding to the fit of the model to the data, while avoiding overfitting.

1.2. Related works

Generally, model selection for MoE models is commonly carried out using the Akaike information criterion (AIC) (Akaike, 1974), the Bayesian information criterion (BIC) (Schwarz, 1978) or the BIC-like approximation of the integrated classification likelihood (ICL-BIC) (Biernacki et al., 2000). Nevertheless, an important limitation of these criteria is that they are only asymptotically valid. In other words, there is no finite sample guarantee when using AIC, BIC or ICL-BIC to choose between different levels of complexity. Therefore, their use in small samples is ad hoc. To address such difficulties, a novel approach called the slope heuristic, supported by a non-asymptotic oracle inequality via a general model selection theorem, was proposed in Birgé and Massart (2007). This method leads to an optimal data-driven choice of multiplicative constants for penalties. See Baudry et al. (2012); Arlot (2019) and the references therein for recent reviews and practical issues related to the slope heuristic.

A number of oracle inequalities for the least absolute shrinkage and selection operator (Lasso) (Tibshirani, 1996) and general penalized maximum likelihood estimators (PMLE) were established in the spirit of the methods based on concentration inequalities developed in Massart (2007); Massart and Meynet (2011); Cohen and Le Pennec (2011, 2013). These results include work on high-dimensional Gaussian graphical models (Devijver and Gallopin, 2018), Gaussian mixture model selection (Maugis and Michel, 2011b,a; Maugis-Rabusseau and Michel, 2013), finite mixture regression models (Meynet, 2013; Devijver, 2015b,a, 2017b,a) and LinBoSGaME models outside the high dimensional setting (Montuelle and Le Pennec, 2014).

1.3. Main contributions

In this work, we establish an important oracle inequality, as shown in Theorem 1, which provides non-asymptotic risk bounds, taking the form of a weak oracle inequalities, under lower-bound assumptions on penalty terms. Our non-asymptotic risk bounds allow the number of observations N to be fixed, while the dimensionality and cardinality of the models, characterised by the number of covariates and the size of the response, are allowed to grow with N and can be much larger than N, unlike traditional criteria such as AIC, BIC or ICL-BIC, which are based on asymptotic theory or Bayesian approaches.

Notably, our oracle inequality shows that the Jensen–Kullback–Leibler loss performance of our PMLEs is comparable to that of oracle models, when we choose sufficiently large constant multiples of the penalties. The shapes of these constants are only known up to multiplicative constants and are proportional to the dimensions of the models. The aforementioned theoretical justifications for the shapes of the penalties are the motivation for using the slope heuristic criterion to select several hyperparameters. These comprise the number of mixture components, the degree of polynomial mean functions, and the potential hidden block-diagonal structures of the covariance matrices of the multivariate output.

Specifically, our oracle inequality, and its corresponding Lasso l_2–MLE and Lasso l_2–rank procedures, help to partially answer the following two important questions in the area of high-dimensional MoE models: (1) what is the number of mixture components K to choose, given the sample size N; and (2) is it better to use a few complex experts or a combination of many simple

experts, given the total number of parameters? We point out that such problems are also addressed in the work of (Mendes and Jiang, 2012, Proposition 1), where the authors provide some qualitative guidance and only propose a practical method for choosing K and D_V, using a complexity penalty or cross-validation. Their approach is also unregularised and is therefore not suitable for the high-dimensional setting.

Notations. Throughout this paper, $\{1, \dots, P\}$ is abbreviated as $[P]$ for $P \in \mathbb{N}^*$. For a matrix $\mathbf{A} = (A_{i,j}) \in \mathbb{R}^{P \times Q}$ with the elements $A_{i,j}$, we denote $|||\mathbf{A}|||_\infty = \max_{i \in [P], j \in [Q]} |A_{i,j}|$ its max-norm. Furthermore, $m(\mathbf{A})$ and $M(\mathbf{A})$ are denoted by the smallest and largest eigenvalues of $\mathbf{A}$, respectively. Similarly, for a vector $\mathbf{a} \in \mathbb{R}^P$, $\|\mathbf{a}\|_p$ denotes the l_p-norm of $\mathbf{a}$ for $0 < p \leq \infty$. For a parametric model S, $\dim(S)$ refers to its dimension, *i.e.,* the total number of parameters to be estimated. If S is a finite set, we denote $\mathrm{card}(S)$ the cardinality, $\mathcal{P}(S)$ the set of all subsets, and $\mathcal{B}(S)$ the set of all partitions, of S. Finally, we refer to $a \wedge b$ as $\min\{a, b\}$ for $a, b \in \mathbb{R}$.

Paper organization In Section 2 we discuss the construction of a collection of SGaBloME models. Then, in Section 3, our main theoretical result is given: a PMLE inequality for SGaBloME models. Two practical procedures are then discussed in Section 4: Lasso l_2–MLE and Lasso l_2–Rank for handling high-dimensional data, and their generalized EM algorithms can be found in Section 5. All technical proofs and detailed algorithms not included in the main paper are relegated to the Appendices A to D.

2. Collection of PSGaBloME models for high-dimensional data

When working with high-dimensional complex data, it is necessary to work with parsimonious models by combining two well-known approaches: *selecting relevant variables* and *ranking sparse models*.

2.1. Variable selection via selecting relevant variables

In this section, we focus on the set of indices of $(\mathbf{X}, \mathbf{Y}) \equiv (X_p, Y_q)_{p \in [P], q \in [Q]}$ so that they are relevant via the notion of irrelevant indices. We call a couple (X_p, Y_q) *irrelevant* if the elements $(\boldsymbol{\Upsilon}_{kd})_{q,p} = 0$ and $(\omega_{kl})_p = 0$ for all $k \in [K]$, $d \in [D_V]$, $l \in [D_W]$. Hence, (X_p, Y_q) is *relevant* if they are not irrelevant. We denote $I = \{(p, q) \in [P] \times [Q] : (X_p, Y_q) \text{ is irrelevant}\}$ the set of indices of irrelevant couples, and the complement of I, called $J = ([P] \times [Q]) \setminus I$, is the set of indices of relevant couples with $J \in \mathcal{P}([P \times Q])$. Furthermore, we call $J_{in} = \{p \in [P] : \exists q \in [Q], (p, q) \in J\}$ the set of indices of relevant input variables.

Let us notice that, for all $k \in [K], d \in [D_V]$, all the entries in the matrix $\boldsymbol{\Upsilon}_{kd}$ belonging to columns indexed by $[P] \setminus J_{in}$ equal to 0, in other words, $\boldsymbol{\Upsilon}_{kd}$ has the relevant columns indexed by J_{in}. Therefore, since $J_{in} \subseteq \{1, \dots, P\}$ by definition, $\boldsymbol{\Upsilon}_{kd}$ will have $Q \times \mathrm{card}\,(J_{in})$ regression coefficients needed to be estimated, which are smaller than $Q \times P$ when all variables are considered. This leads to the number of parameters in regression matrices being then drastically reduced when $\mathrm{card}\,(J_{in}) \ll P$.

The subset J can be constructed by the Lasso estimator, originally established by Tibshirani (1996), is a classical choice and has been extended to deal with multiple multivariate regression models for column sparsity using the Group-Lasso estimator (Yuan and Lin, 2006).

2.2. Variable selection via rank sparse models

Anderson et al. (1998) introduced rank sparse models in the regression framework as follows: if regression matrices have low rank, or at least can be well approximated by low-rank matrices, then the corresponding regression models are said to be rank sparse. In the PSGaBloME model, we assume that for $k \in [K]$, $d \in [D_V]$, the matrix $\boldsymbol{\Upsilon}_{kd}$ has rank R_{kd} and is therefore completely determined by $R_{kd}\left(P-(Q-R_{kd})\right)$ coefficients, which can be less than QP. Combined with the selection of relevant variables, we denote a rank matrix by $\mathbf{R} = (R_{kd})_{k\in[K],d\in[D_V]}$ with the elements $R_{kd} \in [\mathrm{card}\,(J_{in}) \wedge Q]$ for each $k \in [K], d \in [D_V]$.

2.3. Collection of PSGaBloME models

In fact, the class of conditional densities for PSGaBloME models, with relevant variables and rank sparse models, is characterized by the sextuplets $\mathbf{m} = (K, D_W, D_V, \mathbf{B}, J, \mathbf{R})$. That includes the number of clusters, $K \in \mathbb{N}^*$, degrees of polynomials of weights, $D_W \in \mathbb{N}^*$, and of means, $D_V \in \mathbb{N}^*$, the partitions of output indices, $\mathbf{B} \in \mathcal{B}([Q])^K$, the set of relevant indices of variables, $J \in \mathcal{P}([P] \times [Q])$, and the ranks of coefficient matrices, $\mathbf{R}$. For convenience, for any K, D_W, D_V, we denote $\boldsymbol{\omega}_0 = (\omega_{k0})_{k\in[K]}, \boldsymbol{\omega} = (\boldsymbol{\omega}_{kd})_{k\in[K],d\in[D_W]}$, $\boldsymbol{\upsilon}_0 = (\boldsymbol{\upsilon}_{k0})_{k\in[K]}$, $\boldsymbol{\Upsilon} = (\boldsymbol{\Upsilon}_{kd})_{k\in[K],d\in[D_V]}$, $\boldsymbol{\Sigma}(\mathbf{B}) = (\boldsymbol{\Sigma}_k(B_k))_{k\in[K]}$. Then, more precisely, the class of conditional densities of PSGaBloME models with respect to $\mathbf{m}$, $\mathcal{S}_\mathbf{m}$, can be specified as

$$\mathcal{S}_\mathbf{m} = \Big\{ s_{\boldsymbol{\psi}_\mathbf{m}} \equiv s_{\boldsymbol{\psi}_K} \in \mathcal{S} : \boldsymbol{\psi}_\mathbf{m} = (\boldsymbol{\omega}_0, \boldsymbol{\omega}, \boldsymbol{\upsilon}_0, \boldsymbol{\Upsilon}, \boldsymbol{\Sigma}(\mathbf{B})) \in \boldsymbol{\Psi}_\mathbf{m}, \\ \boldsymbol{\Psi}_\mathbf{m} = \mathbb{R}^K \times \mathbf{W}_J^{K\times D_W} \times \mathbb{R}^{K\times Q} \times \mathbf{V}_{J,\mathbf{R}}^{K\times D_V} \times \boldsymbol{\Omega}_\mathbf{B}^K \Big\}. \tag{4}$$

Here, $\mathbf{W}_J$ is the set of vectors in $\mathbb{R}^P$ restricted to the set of indices of relevant input variables J_{in}, $\mathbf{V}_{J,\mathbf{R}}$ the set of matrices with relevant columns indexed by J_{in} and ranks $\mathbf{R}$, and $\boldsymbol{\Omega}_\mathbf{B}$ the set of positive definite block-diagonal matrices depending on partitions $\mathbf{B}$. Note that the collection of PSGaBloME models defined in (4) is generally large and therefore not tractable in practice. This motivates us to restrict (K, D_W, D_V) to the finite sets $\mathcal{K} = [K^*]$, $\mathcal{D}_W = [D_W^*]$, $\mathcal{D}_V = [D_V^*]$ with $K^*, D_W^*, D_V^* \in \mathbb{N}^*$. Furthermore, we focus on a (potentially random) subcollection $\mathcal{J}$ of $\mathcal{P}([P] \times [Q])$ with the controlled size being required in high-dimension case. Moreover, the number of possible vectors of ranks considered is reduced by working on a subset (potentially random) $\mathcal{R}_{(K,J,D_V)}$ of $[\mathrm{card}\,(J_{in}) \wedge Q]^{KD_V}$. Furthermore, we recall that $\mathbf{B}$ is selected among a list of structures $\mathcal{B}([Q])^K$. It is worth mentioning that $\mathrm{card}\,(\mathcal{B}([Q]))^K$, i.e. the power of Bell number, is very large even for a moderate number of variables Q and number of clusters K. This prevents us to consider an exhaustive exploration of the set $\mathcal{B}([Q])^K$. Motivated by the recent work from Devijver and Gallopin (2018), for each cluster $k \in [K]$, we restrict our attention to the sub-collection $\mathcal{B}_{k,\mathcal{E}} = (\mathcal{B}_{k,\epsilon})_{\epsilon\in\mathcal{E}}$ of $\mathcal{B}([Q])$. Here $\mathcal{B}_{k,\epsilon}$ is the partition of the output variables corresponding to the block-diagonal structure of the adjacency matrix $\mathbf{E}_{k,\epsilon} = \left(\mathbb{I}\left\{\left|(\mathbf{S}_k)_{q,q'}\right| > \epsilon\right\}\right)_{q,q'\in[Q]}$, which is based on the thresholded absolute value ϵ of the sample covariance matrix $\mathbf{S}_k$ in each cluster $k \in [K]$. It is important to point out that the class of block-diagonal structures detected by the graphical Lasso algorithm when the regularization parameter varies is identical to the block-diagonal structures $\mathcal{B}_{k,\mathcal{E}}$ detected by the thresholding of the sample covariance for each cluster $k \in [K]$ (Mazumder and Hastie, 2012).

Finally, our collection of PSGaBloME models based on the deterministic and random subcollections are defined, respectively, as, for $\mathcal{S}_{\mathbf{m}}$ specified in (4),

$$\mathcal{S} = \{\mathcal{S}_{\mathbf{m}} : \mathbf{m} \in \mathcal{M}\}, \mathcal{M} = \mathcal{K} \times \mathcal{D}_W \times \mathcal{D}_V \times \mathcal{B}([Q])^K \times \mathcal{P}([P] \times [Q]) \times [\mathrm{card}(J_{in}) \wedge Q]^{KD_V},$$
$$\widetilde{\mathcal{S}} = \left\{\mathcal{S}_{\mathbf{m}} : \mathbf{m} \in \widetilde{\mathcal{M}}\right\}, \widetilde{\mathcal{M}} = \mathcal{K} \times \mathcal{D}_W \times \mathcal{D}_V \times (\mathcal{B}_{k,\mathcal{E}})_{k\in[K]} \times \mathcal{J} \times \mathcal{R}_{(K,J,D_V)}. \tag{5}$$

3. Main theoretical result

3.1. Boundedness conditions on the parameter space

In order to establish our oracle inequality, Theorem 1, we assume that $\mathcal{X}$ is a bounded set in $\mathbb{R}^P$ and make explicit some classical boundedness conditions on the parameter space. We further assume that the covariates $\mathbf{X}$ belong to an hypercube, *e.g.,* $\mathcal{X} = [0,1]^P$, for the simplicity of notation. We then assume that the coefficients from the weights of the softmax gating functions and the means of the Gaussian experts belong to a compact set. Furthermore, the eigenvalues of the block-diagonal covariances of the Gaussian experts lie on a positive interval. More precisely, there exist some for constants $C_{\boldsymbol{\omega}}, C_{\boldsymbol{\Upsilon}}, c_{\boldsymbol{\Sigma}}, C_{\boldsymbol{\Sigma}} > 0$ such that

$$\|\boldsymbol{\omega}_{kd}\|_\infty \leq C_{\boldsymbol{\omega}}, \quad \|\|\boldsymbol{\Upsilon}_{kd}\|\|_\infty \leq C_{\boldsymbol{\Upsilon}}, \quad 0 < c_{\boldsymbol{\Sigma}} \leq m(\boldsymbol{\Sigma}_k(B_k)) \leq M(\boldsymbol{\Sigma}_k(B_k)) \leq C_{\boldsymbol{\Sigma}}. \tag{6}$$

Note that this is a fairly general and mild assumption. In particular, just to simplify the interpretation of sparsity in the high-dimensional setting, the weights of softmax gating functions and the means of Gaussian experts are considered to be polynomial functions of the explanatory variables. However, our Theorem 1 still holds when the weights of softmax gating functions are monomial, allowing for the interaction between different input variables.

To establish our oracle inequality, we need to introduce loss functions, which are useful for comparing two conditional probability density functions. A general principle of penalised maximum likelihood estimation (PMLE) is also derived.

3.2. Loss functions and penalized maximum likelihood estimator

In the maximum likelihood approach, the Kullback-Leibler (KL) divergence is the most natural loss function, defined for two densities s and t. However, to capture the structure of the conditional PDFs and the random covariates $\mathbf{X}_{[N]}$, we instead consider the *tensorized KL* (TKL) divergence:

$$\mathrm{KL}^{\otimes \mathrm{n}}(s,t) = \mathbb{E}_{\mathbf{X}_{[N]}}\left[\frac{1}{N}\sum_{n=1}^{N}\mathrm{KL}(s(\cdot \mid \mathbf{X}_n), t(\cdot \mid \mathbf{X}_n))\right]. \tag{7}$$

Furthermore, given any $\rho \in (0,1)$, a *tensorized Jensen-KL* (TJKL) divergence is given by

$$\mathrm{JKL}_\rho^{\otimes \mathrm{n}}(s,t) = \mathbb{E}_{\mathbf{X}_{[N]}}\left[\frac{1}{N}\sum_{n=1}^{N}\frac{1}{\rho}\mathrm{KL}(s(\cdot \mid \mathbf{X}_n), (1-\rho)s(\cdot \mid \mathbf{X}_n) + \rho t(\cdot \mid \mathbf{X}_n))\right]. \tag{8}$$

In the context of MLE, given the collection of conditional PDFs $\mathcal{S}_{\mathbf{m}}$, we aim to estimate s_0 by the conditional PDF $\widehat{s}_{\mathbf{m}}$ that minimizes the negative log-likelihood (NLL). We should work with

almost minimizer of this quantity and define an η-log-likelihood minimizer (LLM) as any $\widehat{s}_{\mathbf{m}}$ that satisfies:

$$\sum_{n=1}^{N} -\ln\left[\widehat{s}_{\mathbf{m}}\left(\mathbf{Y}_n \mid \mathbf{X}_n\right)\right] \leq \inf_{s_{\mathbf{m}} \in \mathcal{S}_{\mathbf{m}}} \sum_{n=1}^{N} -\ln\left[s_{\mathbf{m}}\left(\mathbf{Y}_n \mid \mathbf{X}_n\right)\right] + \eta. \tag{9}$$

However, this MLE underestimates the risk of the estimate and leads to the selection of models that are too complex. In the context of the PMLE, it is hoped that by adding an appropriate penalty $\text{pen}(\mathbf{m})$, a trade-off can be made between good data fit and model complexity. For a given choice of $\text{pen}(\mathbf{m})$, the *selected model* $S_{\widehat{\mathbf{m}}}$ is chosen as the one whose index is an η'-almost minimizer of the sum of the NLL and this penalty:

$$\sum_{n=1}^{N} -\ln\left[\widehat{s}_{\widehat{\mathbf{m}}}\left(\mathbf{Y}_n \mid \mathbf{X}_n\right)\right] + \text{pen}\left(\widehat{\mathbf{m}}\right) \leq \inf_{\mathbf{m} \in \mathcal{M}} \left\{ \sum_{n=1}^{N} -\ln\left[\widehat{s}_{\mathbf{m}}\left(\mathbf{Y}_n \mid \mathbf{X}_n\right)\right] + \text{pen}(\mathbf{m}) \right\} + \eta'. \tag{10}$$

Note that $\widehat{s}_{\widehat{\mathbf{m}}}$ is then called the η'-PMLE and depends on the error terms η and η'. These error terms are necessary to avoid any existence problem, *e.g.,* the infimum may not be reached. Roughly speaking, the Ekeland variational principle states that for any extended-valued lower semicontinuous function which is bounded below, one can add a small perturbation to ensure the existence of the minimum, see, *e.g.,* (Borwein and Zhu, 2004, Chapter 2). From hereon in, the term *selected model* or *best data-driven model* is used to indicate that it satisfies the definition in (10).

3.3. Oracle inequality

We state our first main contribution, Theorem 1, an oracle inequality that guarantees a non-asymptotic theory for model selection in high-dimensional PSGaBloME, which is proved in Appendix A.

Theorem 1 *Let $\left(\mathbf{X}_{[N]}, \mathbf{Y}_{[N]}\right)$ be the random sample arising from the unknown conditional density s_0. For each $\mathbf{m} = (K, D_W, D_V, \mathbf{B}, J, \mathbf{R}) \in \mathcal{M}$, let $\mathcal{S}_{\mathbf{m}}$ be define by (4). Assume that there exists $\tau > 0$ and $\epsilon_{KL} > 0$ such that, for all $\mathbf{m} \in \mathcal{M}$, one can find $\bar{s}_{\mathbf{m}} \in \mathcal{S}_{\mathbf{m}}$ such that*

$$\text{KL}^{\otimes \text{n}}\left(s_0, \bar{s}_{\mathbf{m}}\right) \leq \inf_{t \in \mathcal{S}_{\mathbf{m}}} \text{KL}^{\otimes \text{n}}\left(s_0, t\right) + \frac{\epsilon_{KL}}{N}, \text{ and } \bar{s}_{\mathbf{m}} \geq e^{-\tau} s_0. \tag{11}$$

Furthermore, we construct a random subcollection $\left(\mathcal{S}_{\mathbf{m}}\right)_{\mathbf{m} \in \widetilde{\mathcal{M}}}$ of $\left(\mathcal{S}_{\mathbf{m}}\right)_{\mathbf{m} \in \mathcal{M}}$ as in (5). Then, there is a constant C such that for any $\rho \in (0,1)$, and any $C_1 > 1$, there are two constants κ and C_2 depending only on ρ and C_1 such that, for every index $\mathbf{m} \in \mathcal{M}$, $\xi_{\mathbf{m}} \in \mathbb{R}^+$, $\Xi = \sum_{\mathbf{m} \in \mathcal{M}} e^{-\xi_{\mathbf{m}}} < \infty$ and

$$\textit{pen}(\mathbf{m}) \geq \kappa\left[(C + \ln N)\dim(\mathcal{S}_{\mathbf{m}}) + (1 \vee \tau)\xi_{\mathbf{m}}\right],$$

the η'-PMLE $\widehat{s}_{\widehat{\mathbf{m}}}$, defined in (10) on the subset $\widetilde{\mathcal{M}} \subset \mathcal{M}$, satisfies

$$\mathbb{E}_{\mathbf{X}_{[N]}, \mathbf{Y}_{[N]}}\left[\text{JKL}_{\rho}^{\otimes \text{n}}\left(s_0, \widehat{s}_{\widehat{\mathbf{m}}}\right)\right] \leq C_1 \mathbb{E}_{\mathbf{X}_{[N]}, \mathbf{Y}_{[N]}}\left[\inf_{\mathbf{m} \in \widetilde{\mathcal{M}}}\left(\inf_{t \in \mathcal{S}_{\mathbf{m}}} \text{KL}^{\otimes \text{n}}\left(s_0, t\right) + 2\frac{\textit{pen}(\mathbf{m})}{N}\right)\right] + C_2(1 \vee \tau)\frac{\Xi^2}{N} + \frac{\eta' + \eta}{N}.$$

Remark 2 *Theorem 1 guarantees that a penalized criterion leads to a good model selection and that the penalty is only known up to multiplicative constants,* e.g., κ*, and is proportional to the dimensions of the models* $\dim(\mathcal{S}_{\mathbf{m}})$*. In particular, in the small and finite sample setting, these multiplicative constants can be calibrated using the slope heuristic approach. Note that* (11) *is not a strong assumption and is satisfied, for example, if* s_0 *is bounded, with a compact support. This oracle inequality compares the performance of our PMLE with the best model in the collection. However, Theorem 1 allows us to approximate well a rich class of conditional PDFs if we use polynomials of weights and Gaussian expert means of sufficient degree, or enough clusters due to the universal approximation of MoE models. This results in the term on the right being small, for* $\mathcal{D}_W, \mathcal{D}_V$ *and* $\mathcal{K}$ *well chosen.*

It should be emphasised that Theorem 1 extends the main result of Montuelle and Le Pennec (2014), which is only valid for a full collection of LinBoSGaME models in the low-dimensional setting. Furthermore, in the context of MoE models, our non-asymptotic oracle inequality for SGaME models in Theorem 1 can be seen as a complementary result to a classical asymptotic theory (Khalili, 2010, Theorems 1, 2, and 3), and an l_1 *oracle inequality that focuses on the properties of the Lasso estimator rather than the model selection procedure (Nguyen et al., 2020a).*

Main challenges on the proof of Theorem 1. Our idea to prove Theorem 1 is inspired by Montuelle and Le Pennec (2014) for handling LinBoSGaME models, but our strategy and most of the technical details differ. This is because their approach is not directly applicable to our high-dimensional SGaBloME models, due to restrictions on relevant predictor variables and rank reduction, and Gaussian experts with block-diagonal covariance matrices. In particular, the main difficulty in proving our oracle inequality lies in bounding the bracketing entropy for our collections of SGaBloME models. This requires several regularity assumptions, which are not easy to verify due to the complexity of SGaBloME models and technical reasons. Therefore, our proofs require the development of several new ideas. Furthermore, unlike Montuelle and Le Pennec (2014), which uses a model selection theorem for a deterministic collection of models from (Cohen and Le Pennec, 2011, 2013), we need to find a way to use the model selection theorem for MLE among a random subcollection (cf. (Devijver, 2015a, Theorem 5.1) and (Devijver and Gallopin, 2018, Theorem 7.3)). The main reason is that our model collection constructed by our Lasso+l_2–MLE and Lasso+l_2–rank procedures in Section 4 is usually random. In particular, our oracle inequality in Theorem 1 still holds for any random subcollection of $\mathcal{M}$ constructed by some suitable tools for PSGaBloME. This reinforces our original contributions concerning the control of bracketing entropy of SGaBloME models.

4. Practical procedures: Lasso+l_2–MLE and Lasso+l_2–Rank

In order to detect the block-diagonal structures, the relevant variables and the rank sparse models in the multiple multivariate regression for high dimensional heterogeneous data using our SGaBloME models, we need to extend the results from Khalili (2010); Stadler et al. (2010) to the multivariate response $\mathbf{Y}$, and the results from linear MoR of Devijver (2015a, 2017a,b) to polynomial SGaBloME with arbitrary degrees of weights and means. This leads to our Lasso+l_2-MLE and Lasso+l_2-Rank procedures. The former takes advantage of MLE, while the latter exploits the matrix structure of the Cholesky decomposition through low-rank estimation. Both procedures involve three main steps. First, for fixed (K, D_W, D_V), we construct a collection of models with relevant variables indexed

by J, where J is constructed by the l_1 and l_2 penalized functions in terms of weights and means. Second, we refit the estimates by MLE for Lasso+l_2-MLE, and by MLE under a rank constraint for Lasso+l_2-Rank, on the restricted set of relevant indices obtained from the first step. Finally, Theorem 1 motivates the selection of the best data-driven model using a non-asymptotic approach slope heuristic.

Identifiability of MoE models. Accounting for the identifiability of MoE models from Jiang and Tanner (1999a); Hennig (2000), we parameterize the gating parameters via the constraints, w.l.g., $\boldsymbol{\omega}_K = (\boldsymbol{\omega}_{K0}, \ldots, \boldsymbol{\omega}_{KD_W}) = \mathbf{0}$, $\boldsymbol{\omega} = (\boldsymbol{\omega}_k)_{k\in[K-1]}$, s.t., $g_K(\mathbf{x}, \boldsymbol{\omega}) \equiv g_K(\mathbf{w}(\mathbf{x}, \boldsymbol{\omega})) = 1 - \sum_{k=1}^{K-1} g_k(\mathbf{w}(\mathbf{x}, \boldsymbol{\omega}))$ with $g_k(\mathbf{w}(\mathbf{x}, \boldsymbol{\omega})) = \exp(w_k(\mathbf{x}; \boldsymbol{\omega}_k)) / \left[1 + \sum_{i=1}^{K-1} \exp(w_i(\mathbf{x}; \boldsymbol{\omega}_i))\right] \equiv g_k(\mathbf{x}, \boldsymbol{\omega})$, for all $k \in [K-1]$.

4.1. Model collection construction

To reduce the complexity for practical procedures, we assume that $\boldsymbol{\Sigma}_k$ is a diagonal matrix for all $k \in [K]$. However, the support of Theorem 1 for using the slope heuristic in Section 4.3 still holds for any block-diagonal structures to which our procedure can be extended. For fixed $K \in \mathcal{K}$, $D_W \in \mathcal{D}_W$ and $D_V \in \mathcal{D}_V$, the Lasso $+l_2$–PMLEs for SGaBloME models can be computed as follows:

$$\widehat{\boldsymbol{\psi}}^{\text{Lasso}+l_2}(\boldsymbol{\lambda}) = \underset{\boldsymbol{\psi}\in\boldsymbol{\Psi}_{(K,D_W,D_V,J,\mathbf{R})}}{\arg\min} \left\{ -\frac{1}{N}\sum_{n=1}^{N} \ln\left(s_{\boldsymbol{\psi}}(\mathbf{y}_n|\mathbf{x}_n)\right) + \text{pen}_{\boldsymbol{\lambda}}(\boldsymbol{\psi}) \right\}, \text{ with} \tag{12}$$

$$\text{pen}_{\boldsymbol{\lambda}}(\boldsymbol{\psi}) = \sum_{k=1}^{K-1}\sum_{d=1}^{D_W} \lambda_{kd}^{[1]} \|\boldsymbol{\omega}_{kd}\|_1 + \sum_{k=1}^{K}\sum_{d=1}^{D_V} \lambda_{kd}^{[2]} \|\mathbf{Q}_k \boldsymbol{\Upsilon}_{kd}\|_1 + \frac{\lambda^{[3]}}{2} \sum_{k=1}^{K-1}\sum_{d=1}^{D_W} \|\boldsymbol{\omega}_{kd}\|_2^2, \tag{13}$$

where $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1^{[1]}, \ldots, \boldsymbol{\lambda}_{K-1}^{[1]}, \boldsymbol{\lambda}_1^{[2]}, \ldots, \boldsymbol{\lambda}_K^{[2]}, \lambda^{[3]})$ is the vector of non-negative regularization parameters with $\boldsymbol{\lambda}_k^{[1]} = (\lambda_{k1}^{[1]}, \ldots, \lambda_{kD_W}^{[1]})$, $\boldsymbol{\lambda}_k^{[2]} = (\lambda_{k1}^{[2]}, \ldots, \lambda_{kD_V}^{[2]})$, and $\mathbf{Q}_k^\top \mathbf{Q}_k = \boldsymbol{\Sigma}_k^{-1}$ is the Cholesky decomposition of $\boldsymbol{\Sigma}_k$, for every $k \in [K]$. Note that the first two terms of (13) are the usual Lasso penalization, while the l_2 penalty function for the weights in the last term is added to avoid wildly large positive or negative estimates of the regression coefficients corresponding to the mixing proportions. This behaviour can be observed in logistic and multinomial regression models when the number of features is potentially large and highly correlated (*e.g.,* Park and Hastie (2008); Bunea (2008)). For given $\boldsymbol{\lambda}$, we can then apply a generalized expectation maximization (EM) algorithm (Dempster et al., 1977; McLachlan and Krishnan, 1997) for (12)–(13) to find the index set of relevant variables $J_{(K,D_W,D_V,\boldsymbol{\lambda})}$, see Section 5. In particular, we proposed new methods by extending the works of Jordan and Jacobs (1994), Khalili (2010), Chamroukhi and Huynh (2018, 2019); Huynh and Chamroukhi (2019) to multivariate responses.

4.2. Refitting

The Lasso$+l_2$-MLE procedure. The MLE can be approximated as

$$\widehat{s}^{(K,D_W,D_V,J)} = \underset{s\in\mathcal{S}_{(K,D_W,D_V,J)}}{\arg\min} \left\{ -\frac{1}{N}\sum_{n=1}^{N} \ln\left(s(\mathbf{y}_n|\mathbf{x}_n)\right) \right\}, \tag{14}$$

by using an EM algorithm for each model $(K, D_W, D_V, J) \in \mathcal{K} \times \mathcal{D}_W \times \mathcal{D}_V \times \mathcal{J}$.

The Lasso+l_2-Rank procedure. We use the generalized EM algorithm to estimate the parameters by MLE under a rank constraint $\mathbf{R}$ on the restricted set of relevant columns J.

4.3. Model selection

The third step is devoted to model selection. We follow the framework from Devijver (2017b, Section 3) to select the refitted model rather than selecting the regularization parameter. Instead of using the asymptotic criteria, we use the slope heuristic, which is a data-driven non-asymptotic criterion for selecting the best model among a collection of models.

5. Generalized EM algorithm for the Lasso $+l_2$ estimator

More often, it is difficult to obtain MLE estimators directly from the likelihood, especially of SGaBloME models. However, in the EM framework, to alleviate this, the data are augmented by imputing, for each incomplete observed data vector $(\mathbf{x}_n, \mathbf{y}_n)$, a latent but unobserved variable that indicates the allocation of the observed data in the context of the mixture model. More formally, for each $n \in [N]$, let $\mathbf{Z}_n = (Z_{nk})_{k\in[K]}$ be indicator binary-valued variables such that $Z_{nk} = 1$ if $(\mathbf{x}_n, \mathbf{y}_n)$ is generated from the k-th expert component, and $Z_{nk} = 0$ otherwise. Therefore, given $\mathbf{x}_n$, $\mathbf{Z}_n$ are IID variables followed by a multinomial distribution

$$\mathbf{Z}_n|\mathbf{x}_n \sim \text{Mult}\left(1, (g_k(\mathbf{x}_n, \boldsymbol{\omega}))_{k\in[K]}\right). \tag{15}$$

We use the generalized EM, or GEM, algorithm to address the problem of an intractable maximization step in original EM algorithm. Specifically, in our work, this iterative algorithm primarily consists of an expectation step (E-step), which computes the conditional expectation of the penalized complete-data log-likelihood (PCDLL) given the observed data, and a maximization step (M-step), which updates the parameters based on their changes in such a way as to increase their values, instead of aiming to maximize the conditional expectation in E-step as in the original method. After starting with initial values for parameters, it alternates between the E and M-steps until convergence, *e.g.,* when there is no longer significant change in the relative variation of the regularized log-likelihood. Specifically, the EM algorithm for solving (14) first requires the construction of the PCDLL based on the penalty function $\text{pen}_{\boldsymbol{\lambda}}(\boldsymbol{\psi})$ in (13) as follows:

$$\text{PL}_\text{c}(\boldsymbol{\psi}, \mathbf{Z}) = \text{L}_\text{c}(\boldsymbol{\psi}, \mathbf{Z}) - \text{pen}_{\boldsymbol{\lambda}}(\boldsymbol{\psi}), \tag{16}$$

where $\text{L}_\text{c}(\boldsymbol{\psi}, \mathbf{Z})$ the standard complete-data log-likelihood (CDLL), which is described as

$$\text{L}_\text{c}(\boldsymbol{\psi}, \mathbf{Z}) = \sum_{n=1}^{N}\sum_{k=1}^{K} Z_{nk} \ln\left[g_k(\mathbf{x}_n; \boldsymbol{\omega})\, \phi(\mathbf{y}_n; \mathbf{v}_k(\mathbf{x}_n; \boldsymbol{\Upsilon}_k), \boldsymbol{\Sigma}_k)\right]. \tag{17}$$

5.1. E-step

The E-step computes the conditional expectation of the PCDLL (16), given the observed data $(\mathbf{x}_n, \mathbf{y}_n)_{n\in[N]}$ under the parameter vector $\boldsymbol{\psi}^{(\text{t})}$ at t-th iteration of the algorithm as follows

$$\begin{aligned} O_{\text{pen}}\left(\boldsymbol{\psi}; \boldsymbol{\psi}^{(\text{t})}\right) &= \mathbb{E}\left[\text{L}_\text{c}(\boldsymbol{\psi}, \mathbf{Z}) \,|\, (\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_N, \mathbf{y}_N), \boldsymbol{\psi}^{(\text{t})}\right] - \text{pen}_{\boldsymbol{\lambda}}(\boldsymbol{\psi}) \\ &= \sum_{n=1}^{N}\sum_{k=1}^{K} \mathbb{E}\left[Z_{nk}|\mathbf{x}_n, \mathbf{y}_n, \boldsymbol{\psi}^{(\text{t})}\right] \ln\left[g_k(\mathbf{x}_n; \boldsymbol{\omega})\, \phi(\mathbf{y}_n; \mathbf{v}_k(\mathbf{x}_n; \boldsymbol{\Upsilon}_k), \boldsymbol{\Sigma}_k)\right] - \text{pen}_{\boldsymbol{\lambda}}(\boldsymbol{\psi}). \end{aligned}$$

For convenience, we denote $\tau_{nk}^{(\mathrm{t})} = \mathbb{E}\left[Z_{nk}|\mathbf{x}_n, \mathbf{y}_n, \boldsymbol{\psi}^{(\mathrm{t})}\right]$ and compute it as

$$\tau_{nk}^{(\mathrm{t})} = \frac{g_k\left(\mathbf{x}_n; \boldsymbol{\omega}^{(\mathrm{t})}\right) \phi\left(\mathbf{y}_n; \mathbf{v}_k(\mathbf{x}_n; \boldsymbol{\Upsilon}_k^{(\mathrm{t})}), \boldsymbol{\Sigma}_k^{(\mathrm{t})}\right)}{\sum_{l=1}^{K} g_l\left(\mathbf{x}_n; \boldsymbol{\omega}^{(\mathrm{t})}\right) \phi\left(\mathbf{y}_n; \mathbf{v}_l(\mathbf{x}_n; \boldsymbol{\Upsilon}_l^{(\mathrm{t})}), \boldsymbol{\Sigma}_l^{(\mathrm{t})}\right)}, \quad \text{for } n \in [N], k \in [K].$$

More importantly, it recognizes that $\tau_{nk}^{(\mathrm{t})}$ is the posterior probability that the data point $(\mathbf{x}_n, \mathbf{y}_n)$ belongs to the kth expert. This step therefore only requires the computation of the conditional component probabilities $\tau_{nk}^{(\mathrm{t})}$ for $n \in [N]$ for each of the K experts.

5.2. Generalized M-step

The generalized M-step aims to update the parameters via improving the value of $O_{\text{pen}}\left(\boldsymbol{\psi}; \boldsymbol{\psi}^{(\mathrm{t})}\right)$ w.r.t. $\boldsymbol{\psi}$, which can be decomposed into independent expressions for the gate and expert models:

$$O_{\text{pen}}\left(\boldsymbol{\psi}; \boldsymbol{\psi}^{(\mathrm{t})}\right) = O_{\text{pen}}\left(\boldsymbol{\omega}; \boldsymbol{\psi}^{(\mathrm{t})}\right) + O_{\text{pen}}\left(\boldsymbol{\Upsilon}, \boldsymbol{\Sigma}; \boldsymbol{\psi}^{(\mathrm{t})}\right), \text{ with} \tag{18}$$

$$O_{\text{pen}}\left(\boldsymbol{\omega}; \boldsymbol{\psi}^{(\mathrm{t})}\right) = \sum_{n=1}^{N}\sum_{k=1}^{K} \tau_{nk}^{(\mathrm{t})} \ln\left[g_k\left(\mathbf{x}_n; \boldsymbol{\omega}\right)\right] - \sum_{k=1}^{K-1}\sum_{d=1}^{D_W} \lambda_{kd}^{[1]} \left\|\boldsymbol{\omega}_{kd}\right\|_1 - \frac{\lambda^{[3]}}{2} \sum_{k=1}^{K-1}\sum_{d=1}^{D_W} \left\|\boldsymbol{\omega}_{kd}\right\|_2^2,$$

$$O_{\text{pen}}\left(\boldsymbol{\Upsilon}, \boldsymbol{\Sigma}; \boldsymbol{\psi}^{(\mathrm{t})}\right) = \sum_{n=1}^{N}\sum_{k=1}^{K} \tau_{nk}^{(\mathrm{t})} \ln\left[\phi\left(\mathbf{y}_n; \mathbf{v}_k(\mathbf{x}_n; \boldsymbol{\Upsilon}_k), \boldsymbol{\Sigma}_k\right)\right] - \sum_{k=1}^{K}\sum_{d=1}^{D_V} \lambda_{kd}^{[2]} \left\|\mathbf{Q}_k \boldsymbol{\Upsilon}_{kd}\right\|_1.$$

where $\mathbf{Q}_k^\top \mathbf{Q}_k = \boldsymbol{\Sigma}_k^{-1}$ for $k \in [K]$. In this way, to maximize $O_{\text{pen}}\left(\boldsymbol{\psi}; \boldsymbol{\psi}^{(\mathrm{t})}\right)$ with respect to model parameters $\boldsymbol{\psi}$ in (18), the M-step can be performed independently for gate and expert parameters, as in (Moerland, 1997; Peralta and Soto, 2014). Moreover, in our problem, each of these optimizations has an additional term given by the respective regularization term, which is similar to a regularized logistic regression in Lee et al. (2006).

The parameters $\boldsymbol{\omega}$ are therefore updated separately by maximizing the following function

$$\begin{aligned} O_{\text{pen}}\left(\boldsymbol{\omega}; \boldsymbol{\psi}^{(\mathrm{t})}\right) = &\sum_{n=1}^{N}\sum_{k=1}^{K-1} \tau_{nk}^{(\mathrm{t})} w_k\left(\mathbf{x}_n; \boldsymbol{\omega}_k\right) - \sum_{n=1}^{N} \ln\left[1 + \sum_{k=1}^{K-1} \exp\left(w_k\left(\mathbf{x}_n; \boldsymbol{\omega}_k\right)\right)\right] \\ &- \sum_{k=1}^{K-1}\sum_{d=1}^{D_W} \lambda_{kd}^{[1]} \left\|\boldsymbol{\omega}_{kd}\right\|_1 - \frac{\lambda^{[3]}}{2} \sum_{k=1}^{K-1}\sum_{d=1}^{D_W} \left\|\boldsymbol{\omega}_{kd}\right\|_2^2, \end{aligned} \tag{19}$$

keeping in mind that $w_k\left(\mathbf{x}_n; \boldsymbol{\omega}_k\right)$ is a polynomial specified by (1). For the polynomial mean Gaussian experts and multiple responses, we propose several approaches for maximizing (19) such a majorization–minimization (MM) algorithm and a coordinate ascent algorithm, see Appendix C for more details. These approaches have some advantages since they do not use any approximate for the penalty function, and have a separate structure that avoids matrix inversion. Finally, the update for $(\boldsymbol{\Upsilon}, \boldsymbol{\Sigma})$ from $O_{\text{pen}}\left(\boldsymbol{\Upsilon}, \boldsymbol{\Sigma}; \boldsymbol{\psi}^{(\mathrm{t})}\right)$ can be found in Appendix D.

6. Conclusion and perspectives

We have studied the PMLEs for PSGaBloME in high-dimensional heterogeneous data. Our main contribution is to establish a non-asymptotic risk bound in the form of an oracle inequality, provided that lower bounds on the penalty hold. By providing some non-asymptotic theoretical foundations for model selection techniques in this area, and proposing two Lasso–MLE–rank procedures based on a new generalized expectation–maximization algorithm to tackle the problem of estimating a collection of PSGaBloME models, our contributions will help to popularize PSGaBloME models as well as slope heuristics. Finally, an important future direction for our work is to perform our procedures as an open-source package and implement them so that we can evaluate their performance on synthetic and real data sets. Furthermore, it is interesting to extend the current oracle inequality, Theorem 1, to more general frameworks where Gaussian experts are replaced by the elliptic distributions.

References

H Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

C W Anderson, E A Stolz, and S Shamsunder. Multivariate autoregressive models for classification of spontaneous electroencephalographic signals during mental tasks. *IEEE Transactions on Biomedical Engineering*, 45(3):277–286, mar 1998. ISSN 1558-2531.

Daniel Andrade, Akiko Takeda, and Kenji Fukumizu. Robust Bayesian model selection for variable clustering with the Gaussian graphical model. *Statistics and Computing*, 30(2):351–376, March 2020. ISSN 1573-1375.

Sylvain Arlot. Minimal penalties and the slope heuristics: a survey. *Journal de la Société Française de Statistique*, 160(3):1–106, 2019.

Andrew R Barron, Cong Huang, Jonathan Li, and Xi Luo. The MDL principle, penalized likelihoods, and statistical risk. *Festschrift in Honor of Jorma Rissanen on the Occasion of his 75th Birthday*, pages 33–63, 2008.

Jean-Patrick Baudry, Cathy Maugis, and Bertrand Michel. Slope heuristics: overview and implementation. *Statistics and Computing*, 22(2):455–470, 2012. ISSN 1573-1375.

Yoshua Bengio. Deep Learning of Representations: Looking Forward. In Adrian-Horia Dediu, Carlos Martín-Vide, Ruslan Mitkov, and Bianca Truthe, editors, *Statistical Language and Speech Processing*, pages 1–37, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-39593-2.

C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719–725, 2000.

Lucien Birgé and Pascal Massart. Minimal penalties for Gaussian model selection. *Probability Theory and Related Fields*, 138(1):33–73, 2007.

Jonathan M Borwein and Qiji J Zhu. *Techniques of Variational Analysis*. Springer, 2004.

Baptiste Broto, François Bachoc, Laura Clouvel, and Jean-Marc Martinez. Block-Diagonal Covariance Estimation and Application to the Shapley Effects in Sensitivity Analysis. *SIAM/ASA Journal on Uncertainty Quantification*, 10(1):379–403, 2022.

Florentina Bunea. Honest variable selection in linear and logistic regression models via l_1 and l_1+l_2 penalization. *Electronic Journal of Statistics*, 2:1153–1194, 2008.

Kenneth P. Burnham and David R. Anderson. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer, New York, NY, 2002. ISBN 978-0-387-22456-5.

Faicel Chamroukhi and Bao Tuyen Huynh. Regularized maximum-likelihood estimation of mixture-of-experts for regression and clustering. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2018.

Faicel Chamroukhi and Bao Tuyen Huynh. Regularized maximum likelihood estimation and feature selection in mixtures-of-experts models. *Journal de la Société Française de Statistique*, 160(1): 57–85, 2019.

Zixiang Chen, Yihe Deng, Yue Wu, Quanquan Gu, and Yuanzhi Li. Towards understanding mixture of experts in deep learning. *arXiv preprint arXiv:2208.02813*, 2022.

S X Cohen and Erwan Le Pennec. Partition-based conditional density estimation. *ESAIM: Probability and Statistics*, 17:672–697, 2013. ISSN 1292-8100.

Serge Cohen and Erwan Le Pennec. Conditional density estimation by penalized likelihood model selection and applications. *Technical report, INRIA*, 2011.

Jan De Leeuw. Applications of convex analysis to multidimensional scaling. *Recent developments in statistics*, 1977.

A P Dempster, N M Laird, and D B Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, mar 1977. ISSN 00359246.

Emilie Devijver. Finite mixture regression: a sparse variable selection by model selection for clustering. *Electronic journal of statistics*, 9(2):2642–2674, 2015a.

Emilie Devijver. An l_1-oracle inequality for the Lasso in multivariate finite mixture of multivariate Gaussian regression models. *ESAIM: PS*, 19:649–670, 2015b.

Emilie Devijver. Joint rank and variable selection for parsimonious estimation in a high-dimensional finite mixture regression model. *Journal of Multivariate Analysis*, 157:1–13, 2017a.

Emilie Devijver. Model-based regression clustering for high-dimensional data: application to functional data. *Advances in Data Analysis and Classification*, 11(2):243–279, 2017b.

Emilie Devijver and Mélina Gallopin. Block-diagonal covariance selection for high-dimensional Gaussian graphical models. *Journal of the American Statistical Association*, 113(521):306–314, 2018. Publisher: Taylor & Francis.

Emilie Devijver, Mélina Gallopin, and Emeline Perthame. Nonlinear network-based quantitative trait prediction from transcriptomic data. *arXiv preprint arXiv:1701.07899*, 2017.

Dat Do, Linh Do, and XuanLong Nguyen. Strong identifiability and parameter learning in regression with heterogeneous response. *arXiv preprint arXiv:2212.04091*, 2022.

Xiaoqiong Fang, Andy W. Chen, and Derek S. Young. Predictors with measurement error in mixtures of polynomial regressions. *Computational Statistics*, May 2022. ISSN 1613-9658.

William Feller. *An introduction to probability theory and its applications, Vol. 1*. John Wiley, 1957.

Christopher R Genovese and Larry Wasserman. Rates of convergence for the Gaussian mixture sieve. *The Annals of Statistics*, 28(4):1105–1127, aug 2000.

Subhashis Ghosal and Aad W van der Vaart. Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *The Annals of Statistics*, 29(5): 1233–1263, oct 2001.

C Hennig. Identifiablity of Models for Clusterwise Linear Regression. *Journal of Classification*, 17 (2):273–296, 2000. ISSN 1432-1343.

Nhat Ho and XuanLong Nguyen. Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *The Annals of Statistics*, 44(6):2726 – 2755, 2016a. Publisher: Institute of Mathematical Statistics and Bernoulli Society.

Nhat Ho and XuanLong Nguyen. On strong identifiability and convergence rates of parameter estimation in finite mixtures. *Electronic Journal of Statistics*, 10(1):271–307, 2016b. Publisher: The Institute of Mathematical Statistics and the Bernoulli Society.

Nhat Ho, Chiao-Yu Yang, and Michael I Jordan. Convergence Rates for Gaussian Mixtures of Experts. *Journal of Machine Learning Research*, 2022.

David R Hunter and Kenneth Lange. Quantile Regression via an MM Algorithm. *Journal of Computational and Graphical Statistics*, 9(1):60–77, 2000.

David R Hunter and Kenneth Lange. A Tutorial on MM Algorithms. *The American Statistician*, 58 (1):30–37, 2004.

Bao Tuyen Huynh and Faicel Chamroukhi. Estimation and feature selection in mixtures of generalized linear experts models. *arXiv preprint arXiv:1907.06994*, 2019.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

W Jiang and M A Tanner. On the identifiability of mixtures-of-experts. *Neural Networks*, 12(9): 1253–1258, 1999a. ISSN 0893-6080.

Wenxin Jiang and Martin A Tanner. Hierarchical mixtures-of-experts for exponential family regression models: approximation and maximum likelihood estimation. *Annals of Statistics*, pages 987–1011, 1999b. Publisher: JSTOR.

Michael I Jordan and Robert A Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural computation*, 6(2):181–214, 1994. Publisher: MIT Press.

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

Jason M. Klusowski, Dana Yang, and W. D. Brinda. Estimating the Coefficients of a Mixture of Two Linear Regressions by Expectation Maximization. *IEEE Transactions on Information Theory*, 65 (6):3515–3524, 2019.

Michael R Kosorok. *Introduction to empirical processes and semiparametric inference*. Springer Science & Business Media, 2007.

Jeongyeol Kwon, Wei Qian, Constantine Caramanis, Yudong Chen, and Damek Davis. Global Convergence of the EM Algorithm for Mixtures of Two Component Linear Regression. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2055–2110. PMLR, June 2019.

Kenneth Lange. *MM Optimization Algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2016.

Su-In Lee, Honglak Lee, Pieter Abbeel, and Andrew Y Ng. Efficient L_1 regularized logistic regression. In *AAAI*, volume 6, pages 401–408, 2006.

Luke R Lloyd-Jones, Hien D Nguyen, and Geoffrey J McLachlan. A globally convergent algorithm for lasso-penalized mixture of linear regression models. *Computational Statistics & Data Analysis*, 119:19–38, 2018. Publisher: Elsevier.

C L Mallows. Some Comments on CP. *Technometrics*, 15(4):661–675, feb 1973. ISSN 00401706.

Saeed Masoudnia and Reza Ebrahimpour. Mixture of experts: a literature survey. *Artificial Intelligence Review*, 42(2):275–293, 2014. ISSN 1573-7462.

P Massart and C Meynet. The Lasso as an l_1-ball model selection procedure. *Electronic Journal of Statistics*, 5:669–687, 2011.

Pascal Massart. *Concentration Inequalities and Model Selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003*. Springer, 2007.

Cathy Maugis and Bertrand Michel. Data-driven penalty calibration: A case study for Gaussian mixture model selection. *ESAIM: PS*, 15:320–339, 2011a.

Cathy Maugis and Bertrand Michel. A non asymptotic penalized criterion for gaussian mixture model selection. *ESAIM: Probability and Statistics*, 15:41–68, 2011b.

Cathy Maugis-Rabusseau and Bertrand Michel. Adaptive density estimation for clustering with Gaussian mixtures. *ESAIM: Probability and Statistics*, 17:698–724, 2013.

Rahul Mazumder and Trevor Hastie. Exact covariance thresholding into connected components for large-scale graphical lasso. *The Journal of Machine Learning Research*, 13(1):781–794, 2012. ISSN 1532-4435.

Geoffrey J McLachlan and Thriyambakam Krishnan. *The EM Algorithm and Extensions*. Wiley, 1997.

Eduardo F Mendes and Wenxin Jiang. On convergence rates of mixtures of polynomial experts. *Neural computation*, 24(11):3025–3051, 2012.

C Meynet. An l_1-oracle inequality for the Lasso in finite mixture Gaussian regression models. *ESAIM: Probability and Statistics*, 17:650–671, 2013.

Perry Moerland. Some methods for training mixtures of experts. Technical report, IDIAP Research Institute, 1997.

Lucie Montuelle and Erwan Le Pennec. Mixture of gaussian regressions model with logistic weights, a penalized maximum likelihood approach. *Electronic Journal of Statistics*, 8(1):1661–1695, 2014.

Hien D Nguyen. An introduction to Majorization-Minimization algorithms for machine learning and statistical estimation. *WIREs Data Mining and Knowledge Discovery*, 7(2):e1198, 2017.

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

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

Hien D Nguyen, Faicel Chamroukhi, and Florence Forbes. Approximation results regarding the multiple-output Gaussian gated mixture of linear experts model. *Neurocomputing*, 366:208–214, 2019. ISSN 0925-2312.

Hien Duy Nguyen, TrungTin Nguyen, Faicel Chamroukhi, and Geoffrey John McLachlan. Approximations of conditional probability density functions in Lebesgue spaces via mixture of experts models. *Journal of Statistical Distributions and Applications*, 8(1):13, August 2021. ISSN 2195-5832.

TrungTin Nguyen. *Model Selection and Approximation in High-dimensional Mixtures of Experts Models: from Theory to Practice*. PhD Thesis, Normandie Université, December 2021.

TrungTin Nguyen, Hien D Nguyen, Faicel Chamroukhi, and Geoffrey J McLachlan. An l_1-oracle inequality for the Lasso in mixture-of-experts regression models. *arXiv preprint arXiv:2009.10622*, 2020a.

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

TrungTin Nguyen, Faicel Chamroukhi, Hien D. Nguyen, and Geoffrey J. McLachlan. Approximation of probability density functions via location-scale finite mixtures in Lebesgue spaces. *Communications in Statistics - Theory and Methods*, pages 1–12, May 2022. ISSN 0361-0926. Publisher: Taylor & Francis.

XuanLong Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics*, 41(1):370–400, 2013.

Andriy Norets. Approximation of conditional densities by smooth mixtures of regressions. *The Annals of Statistics*, 38(3):1733 – 1766, 2010. Publisher: Institute of Mathematical Statistics.

Mee Young Park and Trevor Hastie. Penalized logistic regression for detecting gene interactions. *Biostatistics*, 9(1):30–50, 2008.

Billy Peralta and Alvaro Soto. Embedded local feature selection within mixture of experts. *Information Sciences*, 269:176–187, 2014. ISSN 0020-0255.

Antonio Punzo. Flexible mixture modelling with the polynomial Gaussian cluster-weighted model. *Statistical Modelling*, 14(3):257–291, 2014.

Alexander Rakhlin, Dmitry Panchenko, and Sayan Mukherjee. Risk bounds for mixture density estimation. *ESAIM: PS*, 9:220–229, 2005.

Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

N Stadler, P Buhlmann, and S Van de Geer. l_1-penalization for mixture regression models. *TEST*, 19:209–256, 2010.

Gilbert Strang. *Linear algebra and learning from data*. Wellesley-Cambridge Press Cambridge, 2019. ISBN 0692196382.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

P Tseng. Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494, 2001. ISSN 1573-2878.

Paul Tseng. Coordinate ascent for maximizing nondifferentiable concave functions. *LIDS-P-1840. Technical Report.*, 1988.

Sara Van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

AW Van Der Vaart and JA Wellner. Weak convergence and empirical processes: With applications to statistics springer series in statistics. *Springer*, 58:59, 1996.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

S E Yuksel, J N Wilson, and P D Gader. Twenty Years of Mixture of Experts. *IEEE Transactions on Neural Networks and Learning Systems*, 23(8):1177–1193, 2012. ISSN 2162-2388 VO - 23.

Appendix A. Proof of Theorem 1

Sketch of the proof It is worth noting that because of the work on random subcollection, we need to use a model selection theorem for MLE under a random subcollection (*cf.,* Devijver (2015a, Theorem 5.1) or Devijver and Gallopin (2018, Theorem 7.3)). This is the extension of Cohen and Le Pennec (2011, Theorem 2), which dealt with conditional density estimation but not random subcollection, and Massart (2007, Theorem 7.11), which only works for density estimation. We then explain how we use Theorem 3 to obtain the oracle inequality, Theorem 1, in Appendix A.2. To do this, our model collection must satisfy some regularity assumptions, which are proved in Appendix B. The main difficulties in proving our oracle inequality lie in bounding the bracketing entropy of the weights and means restricted to relevant variables, as well as in rank sparse models, and in particular with block-diagonal covariance matrices for the SGaBloME model. To overcome the first problem, we extend and adapt the strategies of Montuelle and Le Pennec (2014); Devijver (2017a). For the second, we extend the recent novel result on block-diagonal covariance matrices in Devijver and Gallopin (2018) for Gaussian mixture models from Genovese and Wasserman (2000); Maugis and Michel (2011b).

A.1. Model selection theorem for MLE among a random subcollection

Before stating the general theorem, we need to make some necessary assumptions. We are working in a more general context here, with $(\mathbf{X}, \mathbf{Y}) \in \mathcal{X} \times \mathcal{Y}$, and $(\mathcal{S}_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$ defining a model collection indexed by $\mathcal{M}$.

First, we impose a structural assumption on each model indexed by $\mathbf{m} \in \mathcal{M}$ regarding the bracketing entropy, defined by (21), conditioned on the model $\mathcal{S}_{\mathbf{m}}$ w.r.t. a tensorized squared Hellinger (TSH) distance $\mathrm{d}^{2\otimes \mathrm{n}}$. In fact, this is an extension of the squared Hellinger distance $\mathrm{d}^{2\otimes \mathrm{n}}$, as follows:

$$\mathrm{d}^{2\otimes \mathrm{n}}(s,t) = \mathbb{E}_{\mathbf{X}_{[N]}}\left[\frac{1}{N}\sum_{n=1}^{N} \mathrm{d}^2\left(s\left(\cdot \mid \mathbf{X}_n\right), t\left(\cdot \mid \mathbf{X}_n\right)\right)\right]. \tag{20}$$

Recall that the bracketing entropy of a set S with respect to an arbitrary distance d, denoted by $\mathcal{H}_{[\cdot],d}((\delta, S))$, is defined as the logarithm of the minimal number $\mathcal{N}_{[\cdot],d}(\delta, S)$ of brackets $[t^-, t^+]$ covering S, such that $d(t^-, t^+) \leq \delta$. That is,

$$\mathcal{N}_{[\cdot],d}(\delta, S) := \min\left\{ n \in \mathbb{N}^\star : \exists t_1^-, t_1^+, \ldots, t_n^-, t_n^+ \text{ s.t } d(t_k^-, t_k^+) \leq \delta, S \subset \bigcup_{k=1}^{n} \left[t_k^-, t_k^+\right] \right\}, \tag{21}$$

where the bracket $s \in \left[t_k^-, t_k^+\right]$ is defined by $t_k^-(\mathbf{x}, \mathbf{y}) \leq s(\mathbf{x}, \mathbf{y}) \leq t_k^+(\mathbf{x}, \mathbf{y}), \forall (\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$. This leads to the following Assumption A.1 (H).

Assumption A.1 (H) *For every model $\mathcal{S}_{\mathbf{m}}$ in the collection $\mathcal{S}$, there is a non-decreasing function $\phi_{\mathbf{m}}$ such that $\delta \mapsto \frac{1}{\delta}\phi_{\mathbf{m}}(\delta)$ is non-increasing on $(0, \infty)$ and for every $\sigma \in \mathbb{R}^+$,*

$$\int_0^\sigma \sqrt{\mathcal{H}_{[\cdot],\mathrm{d}^{\otimes \mathrm{n}}}\left(\delta, \mathcal{S}_{\mathbf{m}}\left(\widetilde{s}, \sigma\right)\right)} d\delta \leq \phi_{\mathbf{m}}(\sigma),$$

where $\mathcal{S}_{\mathbf{m}}\left(\widetilde{s}, \sigma\right) = \{\mathcal{S}_{\mathbf{m}} \in \mathcal{S}_{\mathbf{m}} : \mathrm{d}^{\otimes \mathrm{n}}\left(\widetilde{s}, \mathcal{S}_{\mathbf{m}}\right) \leq \sigma\}$. The model complexity $\mathcal{D}_{\mathbf{m}}$ of $\mathcal{S}_{\mathbf{m}}$ is then defined as $N\sigma_{\mathbf{m}}^2$, where $\sigma_{\mathbf{m}}$ is the unique root of $\frac{1}{\sigma}\phi_{\mathbf{m}}(\sigma) = \sqrt{N}\sigma$.

This bracketing entropy integral, often call Dudley integral, plays an important role in empirical processes theory (*cf.,* Van Der Vaart and Wellner, 1996; Van de Geer, 2000; Kosorok, 2007). Observe that the model complexity does not depend on the bracketing entropies of the global models $\mathcal{S}_{\mathbf{m}}$, but rather on those of smaller localized sets $\mathcal{S}_{\mathbf{m}}\left(\widetilde{s},\sigma\right)$.

For technical reasons, a seperability assumption, always satisfied in the setting of this paper, is also required. Assumption A.2 (Sep) is a mild condition, which is classical in empirical process theory (Van Der Vaart and Wellner, 1996; Van de Geer, 2000) and allows us to work with a countable subset.

Assumption A.2 (Sep) *For every model $\mathcal{S}_{\mathbf{m}}$, there exists some countable subset $S'_{\mathbf{m}}$ of $\mathcal{S}_{\mathbf{m}}$ and a set $\mathcal{Y}'_{\mathbf{m}}$ with $\iota\left(\mathcal{Y}\setminus\mathcal{Y}'_{\mathbf{m}}\right)=0$, where ι denotes Lebesgue measure, such that for every $t\in\mathcal{S}_{\mathbf{m}}$, there exists some sequence $\left(t_k\right)_{k\in\mathbb{N}^\star}$ of elements of $S'_{\mathbf{m}}$, such that for every $\mathbf{x}\in\mathcal{X}$ and every $\mathbf{y}\in\mathcal{Y}'_{\mathbf{m}}, \ln\left(t_k\left(\mathbf{y}|\mathbf{x}\right)\right)\xrightarrow{k\to+\infty}\ln\left(t\left(\mathbf{y}|\mathbf{x}\right)\right)$.*

To control the complexity of our collection, we also need an information-theoretic assumption. We assume the existence of a Kraft-type inequality for the collection (Massart, 2007; Barron et al., 2008).

Assumption A.3 (K) *There is a family $\left(\xi_{\mathbf{m}}\right)_{\mathbf{m}\in\mathcal{M}}$ of non-negative numbers and a real number Ξ such that*

$$\sum_{\mathbf{m}\in\mathcal{M}} e^{-\xi_{\mathbf{m}}}\leq\Xi<+\infty.$$

We can now state the main result of (Devijver, 2015a, Theorem 5.1) for the model selection theorem for MLE under a random subcollection.

Theorem 3 *Let $\left(\mathbf{X}_n,\mathbf{Y}_n\right)_{n\in[N]}$ be the observations coming from an unknown conditional density s_0. Let the model collection $\mathcal{S}=\left(\mathcal{S}_{\mathbf{m}}\right)_{\mathbf{m}\in\mathcal{M}}$ be an at most countable collection of conditional density sets. Assume that Assumption A.1 (H), Assumption A.2 (Sep), and Assumption A.3 (K) hold for every $\mathbf{m}\in\mathcal{M}$. Let $\epsilon_{KL}>0$, and $\bar{s}_{\mathbf{m}}\in\mathcal{S}_{\mathbf{m}}$, such that*

$$\mathrm{KL}^{\otimes\mathrm{n}}\left(s_0,\bar{s}_{\mathbf{m}}\right)\leq\inf_{t\in\mathcal{S}_{\mathbf{m}}}\mathrm{KL}^{\otimes\mathrm{n}}\left(s_0,t\right)+\frac{\epsilon_{KL}}{N};$$

and let $\tau>0$, such that

$$\bar{s}_{\mathbf{m}}\geq e^{-\tau}s_0. \tag{22}$$

Next, we introduce $\left(\mathcal{S}_{\mathbf{m}}\right)_{\mathbf{m}\in\widetilde{\mathcal{M}}}$ a random subcollection of $\left(\mathcal{S}_{\mathbf{m}}\right)_{\mathbf{m}\in\mathcal{M}}$ and consider the collection $\left(\widehat{s}_{\mathbf{m}}\right)_{\mathbf{m}\in\widetilde{\mathcal{M}}}$ of η-LLMs defined in (9). Then, for any $\rho\in(0,1)$, and any $C_1>1$, there are two constants κ_0 and C_2 depending only on ρ and C_1, such that, for every index $\mathbf{m}\in\mathcal{M}$,

$$pen(\mathbf{m})\geq\kappa\left[\mathcal{D}_{\mathbf{m}}+(1\vee\tau)\xi_{\mathbf{m}}\right],\kappa>\kappa_0,$$

where the model complexity $\mathcal{D}_{\mathbf{m}}$ is defined in Assumption A.1, the η'-PMLE $\widehat{s}_{\widehat{\mathbf{m}}}$, defined in (10) on the subset $\widetilde{\mathcal{M}}$ instead of $\mathcal{M}$, satisfies

$$\begin{aligned}\mathbb{E}_{\mathbf{X}_{[N]},\mathbf{Y}_{[N]}}\left[\mathrm{JKL}_{\rho}^{\otimes\mathrm{n}}\left(s_0,\widehat{s}_{\widehat{\mathbf{m}}}\right)\right]\leq C_1\mathbb{E}_{\mathbf{X}_{[N]},\mathbf{Y}_{[N]}}\left[\inf_{\mathbf{m}\in\widetilde{\mathcal{M}}}\left(\inf_{t\in\mathcal{S}_{\mathbf{m}}}\mathrm{KL}^{\otimes\mathrm{n}}\left(s_0,t\right)+2\frac{pen(\mathbf{m})}{N}\right)\right]\\+C_2(1\vee\tau)\frac{\Xi^2}{N}+\frac{\eta'+\eta}{N}.\end{aligned}$$

In the next section, we show how Theorem 3 can be utilized to prove Theorem 1. In particular, the penalty can be chosen roughly proportional to the intrinsic dimension of the model, and thus of the order of the variance.

A.2. Detail of proof of Theorem 1

It should be stressed that all we need is to verify that Assumption A.3 (K), Assumption A.2 (Sep) and Assumption A.1 (H) hold for every $\mathbf{m} \in \mathcal{M}$. According to the result from Devijver (2015a, Section 5.3), Assumption A.2 (Sep) holds when we consider Gaussian densities and the assumption defined by (22) is true if we assume further that the true conditional density s_0 is bounded and compactly supported. Furthermore, since we restricted to finite collection of models, it is true that there exists a family $(\xi_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$ and $\Xi > 0$ such that Assumption A.3 (K) is satisfied. Therefore, the remaining most difficult step of the proof of concept for Assumption A.1 (H) is presented in Appendix A.2. All technical results are postponed to Appendix B.

Note that the definition of model complexity in Proposition Assumption A.1 (H) is related to a classical entropy dimension of a compact set w.r.t. a Hellinger type divergence $\mathrm{d}^{\otimes \mathrm{n}}$, thanks to the following Proposition 4, which is established in (Cohen and Le Pennec, 2011, Proposition 2).

Proposition 4 *If we have*

$$\mathcal{H}_{[.],\mathrm{d}^{\otimes \mathrm{n}}}\left(\delta, \mathcal{S}_{\mathbf{m}}\right) \leq \dim(\mathcal{S}_{\mathbf{m}}) \left(C_{\mathbf{m}} + \ln\left(\frac{1}{\delta}\right)\right), \text{ for any } \delta \in (0, \sqrt{2}], \text{ then the function}$$

$$\phi_{\mathbf{m}}\left(\delta\right) = \delta\sqrt{\dim\left(\mathcal{S}_{\mathbf{m}}\right)}\left(\sqrt{C_{\mathbf{m}}} + \sqrt{\pi} + \sqrt{\ln\left(\frac{1}{\min\left(\delta, 1\right)}\right)}\right)$$

satisfies Assumption A.1 (H). Furthermore, the unique solution $\delta_{\mathbf{m}}$ of $\frac{1}{\delta}\phi_{\mathbf{m}}\left(\delta\right) = \sqrt{N}\delta$ satisfies

$$N\delta_{\mathbf{m}}^2 \leq \dim(\mathcal{S}_{\mathbf{m}})\left(2\left(\sqrt{C_{\mathbf{m}}} + \sqrt{\pi}\right)^2 + \left(\ln\frac{N}{\left(\sqrt{C_{\mathbf{m}}} + \sqrt{\pi}\right)^2 \dim\left(\mathcal{S}_{\mathbf{m}}\right)}\right)_+\right).$$

Then, we claim that Proposition 4 implies Assumption A.1 (H) because of the fact that

$$\mathcal{H}_{[.],\mathrm{d}^{\otimes \mathrm{n}}}\left(\delta, \mathcal{S}_{\mathbf{m}}\right) \leq \dim(\mathcal{S}_{\mathbf{m}}) \left(C_{\mathbf{m}} + \ln\left(\frac{1}{\delta}\right)\right), \tag{23}$$

where $C_{\mathbf{m}}$ is a constant depending on the model.

Next, recall that the definition from (4) is defined as follows:

$$\begin{aligned}\mathcal{S}_{\mathbf{m}} = \Big\{ s_{\boldsymbol{\psi}_{\mathbf{m}}} \equiv s_{\boldsymbol{\psi}_K} \in \mathcal{S} : \boldsymbol{\psi}_{\mathbf{m}} = (\boldsymbol{\omega}_0, \boldsymbol{\omega}, \boldsymbol{\upsilon}_0, \boldsymbol{\Upsilon}, \boldsymbol{\Sigma}(\mathbf{B})) \in \boldsymbol{\Psi}_{\mathbf{m}}, \\ \boldsymbol{\Psi}_{\mathbf{m}} = \mathbb{R}^K \times \mathbf{W}_J^{K\times D_W} \times \mathbb{R}^{K\times Q} \times \mathbf{V}_{J,\mathbf{R}}^{K\times D_V} \times \boldsymbol{\Omega}_{\mathbf{B}}^K \Big\}.\end{aligned} \tag{24}$$

Here, $\mathbf{m} = (K, D_W, D_V, \mathbf{B}, J, \mathbf{R})$. $\mathbf{W}_J$ is the set of vectors restricted to the set of indices of relevant input variables J_{in}, $\mathbf{V}_{J,\mathbf{R}}$ the set of matrices with relevant columns indexed by J_{in} and ranks $\mathbf{R}$, and $\boldsymbol{\Omega}_{\mathbf{B}}$ the set of positive definite block-diagonal matrices depending on partitions $\mathbf{B}$.

If P and Q are not too large, we do not need to select relevant variables and/or use rank sparse models. We can then work on the structures for means and weights as in LinBoSGaME Montuelle

and Le Pennec (2014). However, to deal with high-dimensional data and to simplify the interpretation of sparsity, we propose to use monomials for weights and polynomial regression models for the soft-max gating functions and the means of Gaussian experts. It is worth mentioning that here we provide a more general result compared to the model defined as in (4). More precisely, we replace the polynomial constructions for the weighting functions with monomials that allow interactions between covariates as follows:

$$\mathbf{W}_{K,D_W} = \{0\} \otimes \mathbf{W}^{K-1}, \quad \mathbf{W} = \left\{ \mathcal{X} \ni \mathbf{x} \mapsto \sum_{\alpha \in \mathcal{A}} \boldsymbol{\omega}_\alpha \mathbf{x}^{\boldsymbol{\alpha}} \in \mathbb{R} : \max_{\boldsymbol{\alpha} \in \mathcal{A}} |\omega_{\boldsymbol{\alpha}}| \leq C_{\boldsymbol{\omega}} \right\}. \tag{25}$$

Here, note that the multi-index $\boldsymbol{\alpha} = (\alpha_p)_{p\in[P]}, \alpha_p \in \mathbb{N}^\star \bigcup \{0\} \equiv \mathbb{N}, \forall p \in [P]$, is an P-tuple of nonnegative integers that satisfies $\mathbf{x}^{\boldsymbol{\alpha}} = \prod_{p=1}^P x_p^{\alpha_p}$ and $|\boldsymbol{\alpha}| = \sum_{p=1}^P \alpha_p$. Then, for all $l \in [D_W]$, we define $\mathcal{A} = \bigcup_{l=0}^{D_W} \mathcal{A}_{|l|}$, $\mathcal{A}_{|l|} = \left\{ \boldsymbol{\alpha} = (\alpha_p)_{p\in[P]} \in \mathbb{N}^P, |\boldsymbol{\alpha}| = l \right\}$. The number $\boldsymbol{\alpha}$ is called the order or degree of monomials $\mathbf{x}^{\boldsymbol{\alpha}}$. By using the well-known stars and bars methods, *e.g.,* Feller (1957, Chapter 2), the cardinality of the set $\mathcal{A}$, denoted by $\mathrm{card}\,(\mathcal{A})$, equals $\binom{D_W+P}{P}$. Note that, for all $d \in [D_{\boldsymbol{\Upsilon}}]$, we define $\mathbf{x}^d$ as $\left(x_p^d\right)_{p\in[P]}$ for the means, which are often used for polynomial regression models. Here, $\mathcal{A}_J$ is the set of multi-index (vector) in $\mathbb{R}^P$ restricted to the set of indices of relevant input variables J_{in}, that is, $\mathcal{A}_J = \left\{ \boldsymbol{\alpha} = (\alpha_t)_{t\in[p]} \in \mathcal{A} : \alpha_j > 0, j \in J_{in} \right\}$. Furthermore, given a regressor $\mathbf{x}$, for all $l \in [D_W]$, $p \in [P]$, we define $\boldsymbol{\omega}_k^{(p,l)} = \left\{ \omega_{k\boldsymbol{\alpha}} \in \mathbb{R} : \boldsymbol{\alpha} = (\alpha_p)_{p\in[P]} \in \mathcal{A}_{|l|}, \alpha_p > 0 \right\}$. We then generalize the definition of relevant variables for monomials as follows. We call a couple (X_p, Y_q) *irrelevant* if the elements $(\boldsymbol{\Upsilon}_{kd})_{q,p} = 0$ and $\boldsymbol{\omega}_k^{(p,l)} = \mathbf{0}$ for all $k \in [K]$, $d \in [D_V]$, $l \in [D_W]$.

We also require some additional definitions of the following sets:

$$\begin{aligned}
\mathcal{P}_{(K,D_W,J)} = \Big\{ \mathcal{X} \ni \mathbf{x} \mapsto \left(g_k\left(\mathbf{w}(\mathbf{x})\right)\right)_{k\in[K]} : g_k\left(\mathbf{w}(\mathbf{x})\right) = \frac{\exp\left(w_k(\mathbf{x})\right)}{\sum_{l=1}^K \exp\left(w_l(\mathbf{x})\right)}, \\
\mathbf{w} = (w_k)_{k\in[K]} \in \mathbf{W}_{K,D_W,J} \Big\}, \\
\mathbf{W}_{(K,D_W,J)} = \{0\} \otimes \mathbf{W}_J^{K-1}, \quad \mathbf{V}_{(K,D_V,J,\mathbf{R})} = R^{K\times Q} \times \mathbf{V}_{J,\mathbf{R}}^{K\times D_V}, \\
\mathbf{W}_J = \left\{ \mathcal{X} \ni \mathbf{x} \mapsto w\left(\mathbf{x}\right) = \sum_{|\boldsymbol{\alpha}|=0}^{D_W} \omega_{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \mathcal{A}_J, \max_{\boldsymbol{\alpha}\in\mathcal{A}} |\omega_{\boldsymbol{\alpha}}| \leq C_{\boldsymbol{\omega}} \right\}, \\
\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})} = \Big\{ \mathcal{X} \times \mathcal{Y} \ni (\mathbf{x},\mathbf{y}) \mapsto \left(\phi\left(\mathbf{y}; \mathbf{v}_k(\mathbf{x}), \boldsymbol{\Sigma}_k\left(B_k\right)\right)\right)_{k\in[K]} : \\
\boldsymbol{v} \in \mathbf{V}_{(K,D_V,J,\mathbf{R})}, \boldsymbol{\Sigma}(\mathbf{B}) \in \boldsymbol{\Omega}_{\mathbf{B}}^K \Big\}.
\end{aligned}$$

We define the following distance over conditional densities:

$$\sup_{\mathbf{x}} d_{\mathbf{y}}(s,t) = \sup_{\mathbf{x}\in\mathcal{X}} d_{\mathbf{y}}(s,t), \text{ where } d_{\mathbf{y}}(s,t) = \left(\int_{\mathcal{Y}} \left(\sqrt{s(\mathbf{y} \mid \mathbf{x})} - \sqrt{t(\mathbf{y} \mid \mathbf{x})} \right)^2 d\mathbf{y} \right)^{1/2}.$$

This leads straightforwardly to $\mathrm{d}^{2\otimes \mathrm{n}}(s,t) \leq \sup_{\mathbf{x}} d_{\mathbf{y}}(s,t)$. Then, we also define

$$\sup_{\mathbf{x}} d_k\left(\mathbf{g}, \mathbf{g}'\right) = \sup_{\mathbf{x}\in\mathcal{X}} \left(\sum_{k=1}^K \left(\sqrt{g_k(\mathbf{x})} - \sqrt{g_k'(\mathbf{x})} \right)^2 \right)^{1/2},$$

for any gating functions $\mathbf{g} = (g_k)_{k\in[K]}$ and $\mathbf{g}' = (g'_k)_{k\in[K]}$. To this end, given any densities s and t over $\mathcal{X}$, the following distance, depending on $\mathbf{y}$, is constructed as follows:

$$\begin{aligned}\sup_{\mathbf{y}} \max_{k} d_{\mathbf{x}}(s,t) &= \sup_{\mathbf{y}\in\mathcal{Y}} \max_{k\in[K]} d_{\mathbf{x}}\left(s_k(\cdot,\mathbf{y}), t_k(\cdot,\mathbf{y})\right) \\ &= \sup_{\mathbf{y}\in\mathcal{Y}} \max_{k\in[K]} \left(\int_{\mathcal{X}} \left(\sqrt{s_k(\mathbf{x},\mathbf{y})} - \sqrt{t_k(\mathbf{x},\mathbf{y})}\right)^2 d\mathbf{x}\right)^{1/2}.\end{aligned}$$

Moreover, given any $\mathbf{g}^+, \mathbf{g}^- \in \mathcal{P}_{(K,D_W,J)}$ and $\phi^+, \phi^- \in \mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}$, let us define

$$\begin{aligned} d^2_{\mathcal{P}_{(K,D_W,J)}}\left(\mathbf{g}^+, \mathbf{g}^-\right) &= \mathbb{E}_{\mathbf{X}_{[N]}}\left[\frac{1}{N}\sum_{n=1}^{N} d_k^2\left(\mathbf{g}^+\left(\mathbf{X}_n\right), \mathbf{g}^-(\mathbf{X}_n)\right)\right], \\ d^2_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}}\left(\phi^+, \phi^-\right) &= \mathbb{E}_{\mathbf{X}_{[N]}}\left[\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} d_{\mathbf{y}}^2\left(\phi_k^+\left(\cdot|\mathbf{X}_n\right), \phi_k^-\left(\cdot|\mathbf{X}_n\right)\right)\right].\end{aligned}$$

Next (23) can be obtained by first decomposing the entropy term between the softmax gating functions and the Gaussian experts via Lemma 5, which is immediately obtained from Montuelle and Le Pennec (2014, Lemma 6), an extension of the results in Genovese and Wasserman (2000, Theorem 2), Ghosal and van der Vaart (2001), Cohen and Le Pennec (2011, Lemma 7) and Cohen and Le Pennec (2013).

Lemma 5 *For all $\delta \in (0, \sqrt{2}]$, it holds that*

$$\mathcal{H}_{[\cdot],\mathrm{d}^{\otimes \mathrm{n}}}\left(\delta, \mathcal{S}_{\mathbf{m}}\right) \leq \mathcal{H}_{[\cdot],d_{\mathcal{P}_{(K,D_W,J)}}}\left(\frac{\delta}{2}, \mathcal{P}_{(K,D_W,J)}\right) + \mathcal{H}_{[\cdot],d_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}}}\left(\frac{\delta}{2}, \mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}\right).$$

Then, we define the metric entropy of the set $\mathbf{W}_{(K,D_W,J)}$: $\mathcal{H}_{\mathrm{d}_{\|\sup\|_\infty}}(\delta, \mathbf{W}_{(K,D_W,J)})$, which measures the logarithm of the minimum number of spheres with radius at most δ, corresponding to the distance $d_{\|\sup\|_\infty}$ needed to cover $\mathbf{W}_{(K,D_W,J)}$, where

$$d_{\|\sup\|_\infty}\left(\left(\mathbf{s}_k\right)_{k\in[K]}, \left(\mathbf{t}_k\right)_{k\in[K]}\right) = \max_{k\in[K]} \sup_{\mathbf{x}\in\mathcal{X}} \left\|\mathbf{s}_k(\mathbf{x}) - \mathbf{t}_k(\mathbf{x})\right\|_2, \tag{26}$$

for arbitrary K-tuples of the functions $(\mathbf{s}_k)_{k\in[K]}$ and $(\mathbf{t}_k)_{k\in[K]}$. Here $\mathbf{s}_k, \mathbf{t}_k : \mathcal{X} \ni \mathbf{x} \mapsto \mathbf{s}_k(\mathbf{x}), \mathbf{t}_k(\mathbf{x}) \in \mathbb{R}^P, \forall k \in [K]$, and given $\mathbf{x} \in \mathcal{X}, k \in [K]$, $\left\|\mathbf{s}_k(\mathbf{x}) - \mathbf{t}_k(\mathbf{x})\right\|_2$ is the Euclidean distance in $\mathbb{R}^P$.

Based on this metric, one can first relate the bracketing entropy of $\mathcal{P}_{(K,D_W,J)}$ to $\mathcal{H}_{\mathrm{d}_{\|\sup\|_\infty}}(\delta, \mathbf{W}_{(K,D_W,J)})$, and then obtain the upper bound for its entropy via Lemma 6, which is proved in Appendix B.1.

Lemma 6 *For all $\delta \in (0, \sqrt{2}]$,*

$$\begin{aligned} H_{[\cdot],d_{\mathcal{P}_{(K,D_W,J)}}}\left(\frac{\delta}{2}, \mathcal{P}_{(K,D_W,J)}\right) &\leq H_{d_{\|\sup\|_\infty}}\left(\frac{3\sqrt{3}\delta}{8\sqrt{K-1}}, \mathbf{W}_{(K,D_W,J)}\right) \\ &\leq \dim\left(\mathbf{W}_{(K,D_W,J)}\right)\left(C_{\mathbf{W}_{(K,D_W,J)}} + \ln\left(\frac{8\sqrt{K-1}}{3\sqrt{3}\delta}\right)\right), \end{aligned} \tag{27}$$

where $\dim\left(\mathbf{W}_{(K,D_W,J)}\right) = (K-1)\,\mathrm{card}\left(\mathcal{A}_J\right)$, $\mathrm{card}\left(\mathcal{A}_J\right) = \binom{D_W + \mathrm{card}(J_{in})}{\mathrm{card}(J_{in})}$ *and* $C_{\mathbf{W}_{(K,D_W,J)}} = \ln\left(\sqrt{2} + \frac{C_{\boldsymbol{\omega}} D_W}{3\sqrt{3}}\right)$.

Lemma 7 allows us to construct the Gaussian brackets to handle with the entropy metric for Gaussian experts, which is established in Appendix B.2.

Lemma 7 *For all* $\delta \in (0, \sqrt{2}]$,

$$\mathcal{H}_{[\cdot],d_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}}}\left(\frac{\delta}{2}, \mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}\right) \leq \dim\left(\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}\right)\left(C_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}} + \ln\left(\frac{1}{\delta}\right)\right). \tag{28}$$

Finally, (23) is proved via Lemmas 5,6, and 7. Indeed, with the fact that $\dim(\mathcal{S}_{\mathbf{m}}) = \dim\left(\mathbf{W}_{(K,D_W,J)}\right) + \dim\left(\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}\right)$, it follows that

$$\begin{aligned}
&\mathcal{H}_{[\cdot],\mathrm{d}^{\otimes \mathrm{n}}}(\delta, \mathcal{S}_{\mathbf{m}}) \leq H_{[\cdot],d_{\mathcal{P}_{(K,D_W,J)}}}\left(\frac{\delta}{2}, \mathcal{P}_{(K,D_W,J)}\right) + \mathcal{H}_{[\cdot],d_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}}}\left(\frac{\delta}{2}, \mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}\right) \\
&\leq \dim\left(\mathbf{W}_{(K,D_W,J)}\right)\left(C_{\mathbf{W}_{(K,D_W,J)}} + \ln\left(\frac{8\sqrt{K-1}}{3\sqrt{3}\delta}\right)\right) + \dim\left(\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}\right)\left(C_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}} + \ln\left(\frac{1}{\delta}\right)\right) \\
&=: \dim(\mathcal{S}_{\mathbf{m}})\left(C_{\mathbf{m}} + \ln\left(\frac{1}{\delta}\right)\right), \text{ where} \\
C_{\mathbf{m}} &= \frac{\dim\left(\mathbf{W}_{(K,D_W,J)}\right)}{\dim(\mathcal{S}_{\mathbf{m}})}\left(C_{\mathbf{W}_{(K,D_W,J)}} + \ln\left(\frac{8\sqrt{K-1}}{3\sqrt{3}}\right)\right) + \frac{\dim\left(\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}\right) C_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}}}{\dim\left(\mathcal{S}_{\mathbf{m}}\right)} \\
&\leq C_{\mathbf{W}_{(K,D_W,J)}} + \ln\left(\frac{8\sqrt{K_{\max}-1}}{3\sqrt{3}}\right) + C_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}} := \mathfrak{C}.
\end{aligned}$$

It is interesting that the constant $\mathfrak{C}$ does not depend on the dimension of the model $\mathbf{m}$ thanks to the hypothesis that $C_{\mathbf{W}_{(K,D_W,J)}}$ is common for every model $\mathbf{m}$ in the collection. Therefore, Proposition 4 implies that, given $C = 2\left(\sqrt{\mathfrak{C}} + \sqrt{\pi}\right)^2$, the model complexity $\mathcal{D}_{\mathbf{m}}$ satisfies

$$\begin{aligned}
\mathcal{D}_{\mathbf{m}} \equiv N\delta_{\mathbf{m}}^2 &\leq \dim(\mathcal{S}_{\mathbf{m}})\left(2\left(\sqrt{\mathfrak{C}} + \sqrt{\pi}\right)^2 + \left(\ln\frac{N}{\left(\sqrt{\mathfrak{C}} + \sqrt{\pi}\right)^2 \dim\left(\mathcal{S}_{\mathbf{m}}\right)}\right)_+\right) \\
&\leq \dim(\mathcal{S}_{\mathbf{m}})\left(C + \ln N\right).
\end{aligned}$$

To this end, Theorem 3 implies that to a collection of PSGaBloME models $\mathcal{S} = (\mathcal{S}_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$ with the penalty functions satisfies $\mathrm{pen}(\mathbf{m}) \geq \kappa\left[\dim(\mathcal{S}_{\mathbf{m}})\left(C + \ln N\right) + (1 \vee \tau)\xi_{\mathbf{m}}\right]$ with $\kappa > \kappa_0$ the oracle inequality of Theorem 1 holds.

Appendix B. Lemma proofs

B.1. Proof of Lemma 6

Following the same argument from the proof of (Montuelle and Le Pennec, 2014, Lemma 4), it holds that

$$H_{[\cdot],d_{\mathcal{P}_{(K,D_W,J)}}}\left(\frac{\delta}{2}, \mathcal{P}_{(K,D_W,J)}\right) \leq H_{d_{\|\sup\|_\infty}}\left(\frac{3\sqrt{3}\delta}{8\sqrt{K-1}}, \mathbf{W}_{(K,D_W,J)}\right).$$

Next, we need to find an upper bound of $H_{d_{\|\sup\|_\infty}}\left(\frac{3\sqrt{3}\delta}{8\sqrt{K-1}}, \mathbf{W}_{(K,D_W,J)}\right)$. Note that for all $\mathbf{w}, \mathbf{v} \in \mathbf{W}_{(K,D_W,J)}$, we obtain the following important inequality

$$d_{\|\sup\|_\infty}(\mathbf{w},\mathbf{v}) = \max_{k\in[K-1]} \sup_{\mathbf{x}\in\mathcal{X}} \left| \sum_{|\boldsymbol{\alpha}|=0}^{D_W} \omega_{k,\boldsymbol{\alpha}}^{\mathbf{w}} \mathbf{x}^{\boldsymbol{\alpha}} - \sum_{|\boldsymbol{\alpha}|=0}^{D_W} \omega_{k,\boldsymbol{\alpha}}^{\mathbf{v}} \mathbf{x}^{\boldsymbol{\alpha}} \right|$$
$$\leq \max_{k\in[K-1]} \sum_{|\boldsymbol{\alpha}|=0}^{D_W} \left|\omega_{k,\boldsymbol{\alpha}}^{\mathbf{w}} - \omega_{k,\boldsymbol{\alpha}}^{\mathbf{v}}\right| \sup_{\mathbf{x}\in\mathcal{X}} \mathbf{x}^{\boldsymbol{\alpha}} \leq \operatorname{card}(\mathcal{A}_J) \max_{k\in[K-1],\boldsymbol{\alpha}\in\mathcal{A}_J} \left|\omega_{k,\boldsymbol{\alpha}}^{\mathbf{w}} - \omega_{k,\boldsymbol{\alpha}}^{\mathbf{v}} \mathbf{x}^{\boldsymbol{\alpha}}\right|.$$

Therefore, given the fact that $\operatorname{card}(\mathcal{A}_J) = \binom{D_W+\operatorname{card}(J_{in})}{\operatorname{card}(J_{in})}$, for all $\delta \in (0, \sqrt{2}]$, it holds that

$$\begin{aligned}
H_{[\cdot],d_{\mathcal{P}_{(K,D_W,J)}}}\left(\frac{\delta}{2}, \mathcal{P}_{(K,D_W,J)}\right) &\leq H_{d_{\|\sup\|_\infty}}\left(\frac{3\sqrt{3}\delta}{8\sqrt{K-1}}, \mathbf{W}_{(K,D_W,J)}\right) \\
&\leq H_{\|\cdot\|_\infty}\left(\frac{3\sqrt{3}\delta}{8\sqrt{K-1}\operatorname{card}(\mathcal{A}_J)}, \left\{\boldsymbol{\omega} \in \mathbb{R}^{(K-1)\operatorname{card}(\mathcal{A}_J)} : \|\boldsymbol{\omega}\|_\infty \leq C_{\boldsymbol{\omega}}\right\}\right) \\
&\leq (K-1)\operatorname{card}(\mathcal{A}_J) \ln\left(1 + \frac{8\sqrt{K-1}C_{\boldsymbol{\omega}}\operatorname{card}(\mathcal{A}_J)}{3\sqrt{3}\delta}\right) \\
&= (K-1)\operatorname{card}(\mathcal{A}_J)\left[\ln\left(\sqrt{2} + \frac{C_{\boldsymbol{\omega}}\operatorname{card}(\mathcal{A}_J)}{3\sqrt{3}}\right) + \ln\left(\frac{8\sqrt{K-1}}{3\sqrt{3}\delta}\right)\right] \\
&= \dim\left(\mathbf{W}_{(K,D_W,J)}\right)\left(C_{\mathbf{W}_{(K,D_W,J)}} + \ln\left(\frac{8\sqrt{K-1}}{3\sqrt{3}\delta}\right)\right).
\end{aligned}$$

B.2. Proof of Lemma 7

It is worth noting that without restriction on relevant variables, rank sparse models on the means and structures on covariance matrices of Gaussian experts from the collection $\mathcal{M}$, the upper bound of the bracketing entropy of Gaussian experts from Lemma 7 is directly implied from Proposition 2 and arguments from Appendix B.2.3 of Montuelle and Le Pennec (2014). However, in order to overcome the much more challenging problems with random subcollection based on relevant variables, rank sparse models on the means and block-diagonal covariance matrices, we have to reply on a much more constructive bracketing entropy in the spirits of works developed in Maugis and Michel (2011b); Montuelle and Le Pennec (2014); Devijver (2015a, 2017a); Devijver and Gallopin (2018).

Given any $k \in [K]$, we first define the following set and its corresponding distance:

$$\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})} = \Big\{\mathcal{X}\times\mathcal{Y} \ni (\mathbf{x},\mathbf{y}) \mapsto \phi\left(\mathbf{y}; \mathbf{v}_{(D_V,J,\mathbf{R}_k)}(\mathbf{x}), \boldsymbol{\Sigma}_k(B_k)\right) : \mathbf{v}_{(D_V,J,\mathbf{R}_k)} \in \mathbf{V}_{(D_V,J,\mathbf{R}_k)}, \boldsymbol{\Sigma}_k(B_k) \in \boldsymbol{\Omega}_{B_k}\Big\}, \tag{29}$$

$$d^2_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}}\left(\phi_k^+, \phi_k^-\right) = \mathbb{E}_{\mathbf{X}_{[N]}}\left[\frac{1}{N}\sum_{n=1}^N \mathrm{d}^2\left(\phi_k^+(\cdot|\mathbf{X}_n), \phi_k^-(\cdot|\mathbf{X}_n)\right)\right].$$

We need to specific block-diagonal structures for $\boldsymbol{\Sigma}_k(B_k)$. To be more precise, for $k \in [K]$, we decompose $\boldsymbol{\Sigma}_k(B_k)$ into G_k blocks, $G_k \in \mathbb{N}^\star$, and we denote by $d_k^{[g]}$ the set of variables into the

gth group, for $g \in [G_k]$, and by $\text{card}\left(d_k^{[g]}\right)$ the number of variables in the corresponding set. Then, we define $B_k = \left(d_k^{[g]}\right)_{g\in[G_k]}$ to be a block structure for the cluster k, and $\mathbf{B} = (B_k)_{k\in[K]}$ to be the output indexes into each group for each cluster. In this way, to construct the block-diagonal covariance matrices, up to a permutation, we make the following definition: $\mathbf{\Omega}_{\mathbf{B}}^K = (\mathbf{\Omega}_{B_k})_{k\in[K]}$, for every $k \in [K]$, for every $k \in [K]$,

$$\mathbf{\Omega}_{\mathbf{B}}^K = \left\{ \mathbf{\Sigma}_k\left(B_k\right) \in \mathcal{S}_Q^{++} \;\middle|\; \begin{array}{l} \mathbf{\Sigma}_k\left(B_k\right) = \mathbf{P}_k \begin{pmatrix} \mathbf{\Sigma}_k^{[1]} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_k^{[2]} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{\Sigma}_k^{[G_k]} \end{pmatrix} \mathbf{P}_k^{-1}, \\ \mathbf{\Sigma}_k^{[g]} \in \mathcal{S}^{++}_{\text{card}\left(d_k^{[g]}\right)}, \forall g \in [G_k] \end{array} \right\}. \tag{30}$$

Here, $\mathbf{P}_k$ corresponds to the permutation leading to a block-diagonal matrix in cluster k. It is worth pointing out that outside the blocks, all coefficients of the matrix are zeros and we also authorize reordering of the blocks: *e.g.,* $\{(1,3);(2,4)\}$ is identical to $\{(2,4);(1,3)\}$, and the permutation inside blocks: *e.g.,* the partition of 4 variables into blocks $\{(1,3);(2,4)\}$ is the same as the partition $\{(3,1);(4,2)\}$.

Then, it follows that $\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})} = \prod_{k=1}^K \mathcal{G}_{(D_V,\mathbf{B}_k,J,\mathbf{R}_k)}$, where $\prod$ stands for the Cartesian product, and Lemma 8, established in B.2.1.

Lemma 8 *Given* $\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})} = \prod_{k=1}^K \mathcal{G}_{(D_V,\mathbf{B}_k,J,\mathbf{R}_k)}$, *it holds that*

$$\mathcal{H}_{[\cdot],d_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}}}\left(\frac{\delta}{2},\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}\right) \leq \sum_{k=1}^K \mathcal{H}_{[\cdot],d_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}}}\left(\frac{\delta}{2\sqrt{K}},\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}\right).$$

Next, we claim that Lemma 7 is implied immediately via Lemma 8 and the following important Lemma 9, which is proved in B.2.2.

Lemma 9 *For all* $\delta \in (0,\sqrt{2}]$ *and* $k \in [K]$, *there exists a constant* $C_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}}$ *such that*

$$\mathcal{H}_{[\cdot],d_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}}}\left(\frac{\delta}{2},\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}\right) \leq \dim\left(\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}\right)\left(C_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}} + \ln\left(\frac{1}{\delta}\right)\right). \tag{31}$$

To this end, by combining the previous two Lemmas 8 and 9, we have

$$\begin{aligned} &\mathcal{H}_{[\cdot],d_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}}}\left(\frac{\delta}{2},\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}\right) \\ &\leq \sum_{k=1}^K \dim\left(\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}\right)\left(C_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}} + \ln\left(\sqrt{K}\right) + \ln\left(\frac{1}{\delta}\right)\right) \\ &= \dim\left(\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}\right)\left(C_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}} + \ln\left(\frac{1}{\delta}\right)\right). \end{aligned}$$

Here,

$$\begin{aligned}
\dim\left(\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}\right) &= \sum_{k=1}^{K} \dim\left(\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}\right), \\
\dim\left(\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}\right) &= \dim\left(\mathbf{V}_{(D_V,J,\mathbf{R}_k)}\right) + D_{B_k}, \\
C_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}} &= \sum_{k=1}^{K} C_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}} + \ln\left(\sqrt{K}\right), \\
D_{B_k} &= \dim\left(\mathbf{\Omega}_{B_k}\right) = \sum_{g=1}^{G_k} \frac{\operatorname{card}\left(b_k^{(g)}\right)\left(\operatorname{card}\left(b_k^{(g)}\right)+1\right)}{2}.
\end{aligned}$$

B.2.1. PROOF OF LEMMA 8

It is sufficient to verify that

$$\mathcal{N}_{[\cdot],d_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}}}\left(\frac{\delta}{2},\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}\right) \leq \prod_{k=1}^{K} \mathcal{N}_{[\cdot],d_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}}}\left(\frac{\delta}{2\sqrt{K}},\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}\right).$$

By (21), for each $k \in [K]$, let $\left\{\left[\phi_k^{l,-},\phi_k^{l,+}\right]\right\}_{1\leq l\leq \mathcal{N}_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}}}$ be a minimal covering of δ_k-bracket for $d_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}}$ of $\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}$ with cardinality $\mathcal{N}_{[\cdot],d_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}}}\left(\delta_k,\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}\right) =: \mathcal{N}_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}}$. By definition, we have

$$\forall l \in \left[\mathcal{N}_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}}\right], d_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}}\left(\phi_k^{l,-},\phi_k^{l,+}\right) \leq \delta_k.$$

This leads to the set $\left\{\prod_{k=1}^{K}\left[\phi_k^{l,-},\phi_k^{l,+}\right]\right\}_{1\leq l\leq \mathcal{N}_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}}}$ is a covering of $\delta/2$-bracket for $d_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}}$ of $\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}$ with cardinality $\prod_{k=1}^{K}\mathcal{N}_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}}$. Indeed, let any $\phi = (\phi_k)_{k\in[K]} \in \mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}$. Consequently, for each $k \in [K], \phi_k \in \mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}$, and there exists $l(k) \in \left[\mathcal{N}_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}}\right]$, such that

$$\phi_k^{l(k),-} \leq \phi_k \leq \phi_k^{l(k),+}, d^2_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}}\left(\phi_k^{l(k),+},\phi_k^{l(k),-}\right) \leq \left(\delta_k\right)^2.$$

Then, it follows that $\phi \in [\phi^-,\phi^+] \in \left\{\prod_{k=1}^{K}\left[\phi_k^{l,-},\phi_k^{l,+}\right]\right\}_{1\leq l\leq \mathcal{N}_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}}}$, with $\phi^- = \left(\phi_k^{l(k),-}\right)_{k\in[K]}$, $\phi^+ = \left(\phi_k^{l(k),+}\right)_{k\in[K]}$, which leads to $\left\{\prod_{k=1}^{K}\left[\phi_k^{l,-},\phi_k^{l,+}\right]\right\}_{1\leq l\leq \mathcal{N}_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}}}$ is a bracket covering of $\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}$.

Now, we want to verify that the size of this bracket is $\delta/2$ via choosing $\delta_k = \frac{\delta}{2\sqrt{K}}, \forall k \in [K]$. It holds that

$$\begin{aligned} d^2_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}}\left(\phi^-,\phi^+\right) &= \mathbb{E}_{\mathbf{X}_{[N]}}\left[\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\mathrm{d}^2\left(\phi_k^{l(k),-}\left(\cdot|\mathbf{X}_n\right),\phi_k^{l(k),+}\left(\cdot|\mathbf{X}_n\right)\right)\right] \\ &= \sum_{k=1}^{K}\mathbb{E}_{\mathbf{X}_{[N]}}\left[\frac{1}{N}\sum_{n=1}^{N}\mathrm{d}^2\left(\phi_k^{l(k),-}\left(\cdot|\mathbf{X}_n\right),\phi_k^{l(k),+}\left(\cdot|\mathbf{X}_n\right)\right)\right] \\ &= \sum_{k=1}^{K} d^2_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}}\left(\phi_k^{l(k),-},\phi_k^{l(k),+}\right) \leq K\left(\frac{\delta}{2\sqrt{K}}\right)^2 = \left(\frac{\delta}{2}\right)^2. \end{aligned}$$

Finally, Lemma 8 is followed by the definition of a minimal $\delta/2$-bracket covering number for $\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}$.

B.2.2. PROOF OF LEMMA 9

We need to bound the bracketing entropy in (31). To do this, we need to construct an extension to the multidimensional Gaussian mixture of Genovese and Wasserman (2000), defining a net over the parameter space of Gaussian experts. Next, we aim to construct a bracket covering of $\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}$ according to the tensorized Hellinger distance, $d_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}}$ based on Gaussian dilatations.

Step 1: Construction of a net for the block-diagonal covariance matrices. Firstly, for a given matrix $\mathbf{\Sigma}_k(B_k) \in \mathbf{\Omega}_{B_k}, k \in [K]$, we denote by $\mathrm{Adj}\left(\mathbf{\Sigma}_k(B_k)\right)$ the adjacency matrix associated to the covariance matrix $\mathbf{\Sigma}_k(B_k)$. Note that this matrix of size Q^2 can be defined by a vector of concatenated upper triangular vectors. We are going to make use of the result from Devijver and Gallopin (2018) to handle the block-diagonal covariance matrices $\mathbf{\Sigma}_k\left(B_k\right)$, via its corresponding adjacency matrix. To do this, we need to construct a discrete space for $\{0,1\}^{Q(Q-1)/2}$, which is a one-to-one correspondence (bijection) with

$$\mathcal{A}_{B_k} = \left\{\mathbf{A}_{B_k} \in \mathcal{S}_Q\left(\{0,1\}\right) : \exists \mathbf{\Sigma}_k\left(B_k\right) \in \mathbf{\Omega}_{B_k} \text{ s.t } \mathrm{Adj}\left(\mathbf{\Sigma}_k\left(B_k\right)\right) = \mathbf{A}_{B_k}\right\},$$

where $\mathcal{S}_Q\left(\{0,1\}\right)$ is the set of symmetric matrices of size Q taking values on $\{0,1\}$.

Then, we want to deduce a discretization of the set of covariance matrices. Let h denotes Hamming distance on $\{0,1\}^{Q(Q-1)/2}$ defined by

$$d(z,z') = \sum_{n=1}^{N}\mathbb{I}\left\{z \neq z'\right\}, \text{ for all } z, z' \in \{0,1\}^{Q(Q-1)/2}.$$

Let $\{0,1\}^{Q(Q-1)/2}_{B_k}$ be the subset of $\{0,1\}^{Q(Q-1)/2}$ of vectors for which the corresponding graph has structure $B_k = \left(b_k^{(g)}\right)_{g\in[G_k]}$. Then, given any $\epsilon > 0$, Corollary 1 and Proposition 2 from Supplementary Material A of Devijver and Gallopin (2018) lead to that there exists some subset $\mathcal{R}$

of $\{0,1\}^{Q(Q-1)/2}$, as well as its equivalent $\mathcal{A}_{B_k}^{\text{disc}}$ for adjacency matrices satisfy

$$\left\|\boldsymbol{\Sigma}_k\left(B_k\right)-\widetilde{\boldsymbol{\Sigma}}_k\left(B_k\right)\right\|_2^2 \leq \frac{D_{B_k}}{2} \wedge \epsilon^2, \forall\left(\boldsymbol{\Sigma}_k\left(B_k\right), \widetilde{\boldsymbol{\Sigma}}_k\left(B_k\right)\right) \in\left(\tilde{S}_{B_k}^{\text{disc}}(\epsilon)\right)^2 \text{ s.t. } \boldsymbol{\Sigma}_k\left(B_k\right) \neq \widetilde{\boldsymbol{\Sigma}}_k\left(B_k\right),$$

$$\operatorname{card}\left(\tilde{S}_{B_k}^{\text{disc}}(\epsilon)\right) \leq\left(\left\lfloor\frac{2 C_{\boldsymbol{\Sigma}}}{\epsilon}\right\rfloor \frac{Q(Q-1)}{2 D_{B_k}}\right)^{D_{B_k}}, \tag{32}$$

$$D_{B_k}=\dim\left(\boldsymbol{\Omega}_{B_k}\right)=\sum_{g=1}^{G_k} \frac{\operatorname{card}\left(b_k^{(g)}\right)\left(\operatorname{card}\left(b_k^{(g)}\right)-1\right)}{2}, \text{ where} \tag{33}$$

$$\tilde{S}_{B_k}^{\text{disc}}(\epsilon)=\left\{\boldsymbol{\Sigma}_k\left(B_k\right) \in \mathcal{S}_Q^{++}(\mathbb{R}) : \operatorname{Adj}\left(\boldsymbol{\Sigma}_k\left(B_k\right)\right) \in \mathcal{A}_{B_k}^{\text{disc}},\right.$$
$$\left.\left(\boldsymbol{\Sigma}_k\left(B_k\right)\right)_{i,j}=\sigma_{i,j} \epsilon, \sigma_{i,j} \in\left[\frac{-C_{\boldsymbol{\Sigma}}}{\epsilon}, \frac{C_{\boldsymbol{\Sigma}}}{\epsilon}\right] \bigcap \mathbb{Z}\right\}.$$

Therefore, by choosing $\epsilon^2 \leq \frac{D_{B_k}}{2}$, given $\boldsymbol{\Sigma}_k\left(B_k\right) \in \boldsymbol{\Omega}_{B_k}$, there exists $\widetilde{\boldsymbol{\Sigma}}_k\left(B_k\right) \in \tilde{S}_{B_k}^{\text{disc}}(\epsilon)$, such that

$$\left\|\boldsymbol{\Sigma}_k\left(B_k\right)-\widetilde{\boldsymbol{\Sigma}}_k\left(B_k\right)\right\|_2^2 \leq \epsilon^2. \tag{34}$$

Based on $\widetilde{\boldsymbol{\Sigma}}_k\left(B_k\right)$, we can construct the following bracket covering of $\mathcal{G}_{(K, D_V, \mathbf{B}, J, \mathbf{R})}$ via defining suitable nets for the means of Gaussian experts. More precisely, given any $\delta_{\mathbf{V}_{\left(D_V, J, \mathbf{R}_k\right)}}>0$, we claim that the set

$$\left\{[l, u] \;\middle|\; \begin{array}{l} l(\mathbf{x}, \mathbf{y})=(1+2 \alpha)^{-D_V} \phi\left(\mathbf{y} ; \widetilde{\mathbf{v}}_{\left(D_V, J, \mathbf{R}_k\right)}(\mathbf{x}),(1+\alpha)^{-1} \widetilde{\boldsymbol{\Sigma}}_k\left(B_k\right)\right), \\ u(\mathbf{x}, \mathbf{y})=(1+2 \alpha)^{D_V} \phi\left(\mathbf{y} ; \widetilde{\mathbf{v}}_{\left(D_V, J, \mathbf{R}_k\right)}(\mathbf{x}),(1+\alpha) \widetilde{\boldsymbol{\Sigma}}_k\left(B_k\right)\right), \\ \widetilde{\mathbf{v}}_{\left(D_V, J, \mathbf{R}_k\right)} \in G_{\mathbf{V}_{\left(D_V, J, \mathbf{R}_k\right)}}\left(\delta_{\mathbf{V}_{\left(D_V, J, \mathbf{R}_k\right)}}\right), \widetilde{\boldsymbol{\Sigma}}_k\left(B_k\right) \in \tilde{S}_{B_k}^{\text{disc}}(\epsilon) \end{array} \right\},$$

is an $\delta_{\mathbf{V}_{\left(D_V, J, \mathbf{R}_k\right)}}$-brackets set over $\mathcal{G}_{(K, D_V, \mathbf{B}, J, \mathbf{R})}$ where the constant $\alpha>0$ and function $\mathcal{X} \ni \mathbf{x} \mapsto \widetilde{\mathbf{v}}_{\left(D_V, J, \mathbf{R}_k\right)}(\mathbf{x})$ and its corresponding space $G_{\mathbf{V}_{\left(D_V, J, \mathbf{R}_k\right)}}\left(\delta_{\mathbf{V}_{\left(D_V, J, \mathbf{R}_k\right)}}\right)$ will be specified later. Indeed, we consider any function $\mathcal{X} \times \mathcal{Y} \ni(\mathbf{x}, \mathbf{y}) \mapsto f(\mathbf{x}, \mathbf{y})=\phi\left(\mathbf{y} ; \mathbf{v}_{\left(D_V, J, \mathbf{R}_k\right)}(\mathbf{x}), \boldsymbol{\Sigma}_k\left(B_k\right)\right)$ that belongs to $\mathcal{G}_{(K, D_V, \mathbf{B}, J, \mathbf{R})}$, where $\mathbf{v}_{\left(D_V, J, \mathbf{R}_k\right)} \in \mathbf{V}_{\left(D_V, J, \mathbf{R}_k\right)}$ and $\boldsymbol{\Sigma}_k\left(B_k\right) \in \boldsymbol{\Omega}_{B_k}$. According to (34), there exists $\widetilde{\boldsymbol{\Sigma}}_k\left(B_k\right) \in \tilde{S}_{B_k}^{\text{disc}}(\epsilon)$ such that

$$\left\|\boldsymbol{\Sigma}_k\left(B_k\right)-\widetilde{\boldsymbol{\Sigma}}_k\left(B_k\right)\right\|_2^2 \leq \epsilon^2.$$

Step 2: Construction of a net for the mean functions. We claim that given any $\delta_{\mathbf{V}_{\left(D_V, J, \mathbf{R}_k\right)}}>0$, any $\mathbf{v}_{\left(D_V, J, \mathbf{R}_k\right)} \in \mathbf{V}_{\left(D_V, J, \mathbf{R}_k\right)}$, there exist a minimal covering of δ_k-bracket $G_{\mathbf{V}_{\left(D_V, J, \mathbf{R}_k\right)}}\left(\delta_{\mathbf{V}_{\left(D_V, J, \mathbf{R}_k\right)}}\right)$

and a function $\widetilde{\mathbf{v}}_{(D_V,J,\mathbf{R}_k)} \in G_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}\left(\delta_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}\right)$ such that

$$\sup_{\mathbf{x}\in\mathcal{X}} \left\|\widetilde{\mathbf{v}}_{(D_V,J,\mathbf{R}_k)}(\mathbf{x}) - \mathbf{v}_{(D_V,J,\mathbf{R}_k)}(\mathbf{x})\right\|_2^2 \leq \delta^2_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}, \tag{35}$$

$$\operatorname{card}\left(G_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}\left(\delta_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}\right)\right) \leq \left(\frac{\exp\left(C_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}\right)}{\delta_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}}\right)^{\dim\left(\mathbf{V}_{(D_V,J,\mathbf{R}_k)}\right)}. \tag{36}$$

To accomplish this, we use the singular value decomposition of $\boldsymbol{\Upsilon}_{kd}^{R_{kd}} = \sum_{r=1}^{R_{kd}} (\sigma_{kd})_r (\mathbf{u}_{kd})_{\bullet,r} (\mathbf{v}_{kd}^\top)_{r,\bullet}$, $k \in [K], d \in [D_V]$, with $(\sigma_{kd})_r, r \in [R_{kd}]$, denote the singular values of $\boldsymbol{\Upsilon}_{kd}^{R_{kd}}$, with corresponding orthogonal unit vectors $((\mathbf{u}_{kd})_{\bullet,r})_{r\in[R_{kd}]}$ and $\left((\mathbf{v}_{kd}^\top)_{r,\bullet}\right)_{r\in[R_{kd}]}$. Then, we construct $\widetilde{\mathbf{v}}_{(D_V,J,\mathbf{R}_k)}(\mathbf{x}) = \widetilde{\boldsymbol{\Upsilon}}_{k0} + \sum_{d=1}^{D_V} \widetilde{\boldsymbol{\Upsilon}}_{kd}^{R_{kd}} \mathbf{x}^d$, where $\widetilde{\boldsymbol{\upsilon}}_{k0}$ and $\widetilde{\boldsymbol{\Upsilon}}_{kd}^{R_{kd}} = \sum_{r=1}^{R_{kd}} (\widetilde{\sigma}_{kd})_r (\widetilde{\mathbf{u}}_{kd})_{\bullet,r} (\widetilde{\mathbf{v}}_{kd}^\top)_{r,\bullet}$, $k \in [K], d \in [D_V]$, are determined so that (35) and (36) are satisfied. Note that for each $k \in [K], d \in [D_V]$, it holds that

$$\begin{aligned}
\left\|\widetilde{\mathbf{v}}_{(D_V,J,\mathbf{R}_k)}(\mathbf{x}) - \mathbf{v}_{(D_V,J,\mathbf{R}_k)}(\mathbf{x})\right\|_2 &= \left\|\widetilde{\boldsymbol{\upsilon}}_{k0} - \boldsymbol{\upsilon}_{k0} + \sum_{d=1}^{D_V}\left(\widetilde{\boldsymbol{\Upsilon}}_{kd}^{R_{kd}} - \boldsymbol{\Upsilon}_{kd}^{R_{kd}}\right)\mathbf{x}^d\right\|_2 \\
&\leq \left\|\widetilde{\boldsymbol{\upsilon}}_{k0} - \boldsymbol{\upsilon}_{k0}\right\|_2 + \sum_{d=1}^{D_V}\left\|\left(\widetilde{\boldsymbol{\Upsilon}}_{kd}^{R_{kd}} - \boldsymbol{\Upsilon}_{kd}^{R_{kd}}\right)\mathbf{x}^d\right\|_2 \\
&\leq \sqrt{Q}\left\|\widetilde{\boldsymbol{\upsilon}}_{k0} - \boldsymbol{\upsilon}_{k0}\right\|_\infty + P\sqrt{Q}\sum_{d=1}^{D_V}\left\|\widetilde{\boldsymbol{\Upsilon}}_{kd}^{R_{kd}} - \boldsymbol{\Upsilon}_{kd}^{R_{kd}}\right\|_\infty \left\|\mathbf{x}^d\right\|_\infty \\
&\leq \sqrt{Q}\left\|\widetilde{\boldsymbol{\upsilon}}_{k0} - \boldsymbol{\upsilon}_{k0}\right\|_\infty + P\sqrt{Q}\sum_{d=1}^{D_V}\left\|\widetilde{\boldsymbol{\Upsilon}}_{kd}^{R_{kd}} - \boldsymbol{\Upsilon}_{kd}^{R_{kd}}\right\|_\infty,
\end{aligned}$$

where we used the fact that for all $d \in [D_V]$, $\mathbf{x} \in \mathcal{X}$, $\left\|\mathbf{x}^d\right\|_\infty \leq 1$ as $\mathcal{X} = [0,1]^P$. Thus, (35) is immediately followed if we now choose $\widetilde{\boldsymbol{\upsilon}}_{k0}$ and $\widetilde{\boldsymbol{\Upsilon}}_{kd}^{R_{kd}}$ such that

$$\sqrt{Q}\left\|\boldsymbol{\upsilon}_{k0} - \widetilde{\boldsymbol{\upsilon}}_{k0}\right\|_\infty \leq \frac{\delta_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}}{2}, \tag{37}$$

$$\left\|\boldsymbol{\Upsilon}_{kd}^{R_{kd}} - \widetilde{\boldsymbol{\Upsilon}}_{kd}^{R_{kd}}\right\|_\infty \leq \frac{\delta_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}}{2D_V P\sqrt{Q}}. \tag{38}$$

Let us now see how to construct $\widetilde{\boldsymbol{\upsilon}}_{k0}$ to get (37). This task can be accomplished if for all $k \in [K]$, $q \in [Q]$, we set

$$\begin{aligned}
B &= \mathbb{Z} \cap \left[\left\lfloor -A_{\mathbf{u},\mathbf{v}}\frac{2\sqrt{Q}}{\delta_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}}\right\rfloor, \left\lfloor A_{\mathbf{u},\mathbf{v}}\frac{2\sqrt{Q}}{\delta_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}}\right\rfloor\right], \\
(\widetilde{\boldsymbol{\upsilon}}_{k0})_q &= \arg\min_{b\in B}\left|(\boldsymbol{\upsilon}_{k0})_q - \frac{\delta_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}}{2\sqrt{Q}}b\right|.
\end{aligned}$$

Next, let us now see how to construct $\widetilde{\boldsymbol{\Upsilon}}_{kd}^{R_{kd}}$ to get (38). The boundedness assumption in (6) implies that

$$
\begin{aligned}
\left|\left|\left|\boldsymbol{\Upsilon}_{kd}^{R_{kd}} - \widetilde{\boldsymbol{\Upsilon}}_{kd}^{R_{kd}}\right|\right|\right|_{\infty} &= \max_{q\in[Q],p\in[P]} \left|\sum_{r=1}^{R_{kd}} \left[(\sigma_{kd})_r (\mathbf{u}_{kd})_{q,r} (\mathbf{v}_{kd}^{\top})_{r,p} - (\widetilde{\sigma}_{kd})_r (\widetilde{\mathbf{u}}_{kd})_{q,r} (\widetilde{\mathbf{v}}_{kd}^{\top})_{r,p}\right]\right| \\
&= \max_{q\in[Q],p\in[P]} \left|\sum_{r=1}^{R_{kd}} \Bigg[\left((\sigma_{kd})_r - (\widetilde{\sigma}_{kd})_r\right) (\mathbf{u}_{kd})_{q,r} (\mathbf{v}_{kd}^{\top})_{r,p} \right. \\
&\qquad - (\widetilde{\sigma}_{kd})_r \left((\widetilde{\mathbf{u}}_{kd})_{q,r} - (\mathbf{u}_{kd})_{q,r}\right) (\widetilde{\mathbf{v}}_{kd}^{\top})_{r,p} \\
&\qquad \left. - (\widetilde{\sigma}_{kd})_r (\mathbf{u}_{kd})_{q,r} \left((\mathbf{v}_{kd}^{\top})_{r,p} - (\widetilde{\mathbf{v}}_{kd}^{\top})_{r,p}\right) \Bigg] \right| \\
&\leq \max_{r\in[R_{kd}]} |(\sigma_{kd})_r - (\widetilde{\sigma}_{kd})_r| \max_{q\in[Q],p\in[P]} \sum_{r=1}^{R_{kd}} \left|(\mathbf{u}_{kd})_{q,r} (\mathbf{v}_{kd}^{\top})_{r,p}\right| \\
&\quad + \max_{q\in[Q],r\in[R_{kd}]} |(\widetilde{\mathbf{u}}_{kd})_{q,r} - (\mathbf{u}_{kd})_{q,r}| \max_{p\in[P]} \sum_{r=1}^{R_{kd}} \left|(\widetilde{\sigma}_{kd})_r (\widetilde{\mathbf{v}}_{kd}^{\top})_{r,p}\right| \\
&\quad + \max_{r\in[R_{kd}],p\in[P]} \left|(\mathbf{v}_{kd}^{\top})_{r,p} - (\widetilde{\mathbf{v}}_{kd}^{\top})_{r,p}\right| \max_{q\in[Q]} \sum_{r=1}^{R_{kd}} |(\widetilde{\sigma}_{kd})_r (\mathbf{u}_{kd})_{q,r}| \\
&\leq R_{kd} A_{\mathbf{u},\mathbf{v}}^2 \max_{r\in[R_{kd}]} |(\sigma_{kd})_r - (\widetilde{\sigma}_{kd})_r| \\
&\quad + R_{kd} A_{\mathbf{u},\mathbf{v}} A_{\sigma} \Big(\max_{q\in[Q],r\in[R_{kd}]} |(\widetilde{\mathbf{u}}_{kd})_{q,r} - (\mathbf{u}_{kd})_{q,r}| \\
&\qquad + \max_{r\in[R_{kd}],p\in[P]} \left|(\mathbf{v}_{kd}^{\top})_{r,p} - (\widetilde{\mathbf{v}}_{kd}^{\top})_{r,p}\right| \Big) .
\end{aligned}
$$

Therefore, (38) is immediately implied if we now choose $(\widetilde{\sigma}_{kd})_r$, $(\widetilde{\mathbf{u}}_{kd})_{q,r}$ and $(\widetilde{\mathbf{v}}_{kd}^{\top})_{r,p}$ such that

$$
\begin{aligned}
\max_{r\in[R_{kd}]} |(\sigma_{kd})_r - (\widetilde{\sigma}_{kd})_r| &\leq \frac{\delta_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}}{6R_{kd}A_{\mathbf{u},\mathbf{v}}^2 D_V P\sqrt{Q}}, \\
\max_{q\in[Q],r\in[R_{kd}]} |(\widetilde{\mathbf{u}}_{kd})_{q,r} - (\mathbf{u}_{kd})_{q,r}| &\leq \frac{\delta_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}}{6R_{kd}A_{\mathbf{u},\mathbf{v}} A_{\sigma} D_V P\sqrt{Q}}, \\
\max_{r\in[R_{kd}],p\in[P]} \left|(\mathbf{v}_{kd}^{\top})_{r,p} - (\widetilde{\mathbf{v}}_{kd}^{\top})_{r,p}\right| &\leq \frac{\delta_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}}{6R_{kd}A_{\mathbf{u},\mathbf{v}} A_{\sigma} D_V P\sqrt{Q}}.
\end{aligned}
$$

This task can be accomplished as follows: for all $r \in [R_{kd}]$, $p \in [P]$, $q \in [Q]$, set

$$
\begin{aligned}
S &= \mathbb{Z} \cap \left[0, \left\lfloor A_\sigma \frac{6R_{kd}A_{\mathbf{u},\mathbf{v}}^2 D_V P\sqrt{Q}}{\delta_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}} \right\rfloor\right], \\
(\widetilde{\sigma}_{kd})_r &= \arg\min_{\zeta \in S} \left| (\sigma_{kd})_r - \frac{\delta_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}}{6R_{kd}A_{\mathbf{u},\mathbf{v}}^2 D_V P\sqrt{Q}} \zeta \right|, \\
U &= \mathbb{Z} \cap \left[\left\lfloor -A_{\mathbf{u},\mathbf{v}} \frac{6R_{kd}A_{\mathbf{u},\mathbf{v}} A_\sigma D_V P\sqrt{Q}}{\delta_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}} \right\rfloor, \left\lfloor A_{\mathbf{u},\mathbf{v}} \frac{6R_{kd}A_{\mathbf{u},\mathbf{v}} A_\sigma D_V P\sqrt{Q}}{\delta_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}} \right\rfloor \right], \\
(\widetilde{\mathbf{u}}_{kd})_{q,r} &= \arg\min_{\mu \in U} \left| (\mathbf{u}_{kd})_{q,r} - \frac{\delta_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}}{6R_{kd}A_{\mathbf{u},\mathbf{v}} A_\sigma D_V P\sqrt{Q}} \mu \right|, \\
(\widetilde{\mathbf{v}}_{kd}^\top)_{r,p} &= \arg\min_{\upsilon \in U} \left| (\mathbf{v}_{kd}^\top)_{r,p} - \frac{\delta_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}}{6R_{kd}A_{\mathbf{u},\mathbf{v}} A_\sigma D_V P\sqrt{Q}} \upsilon \right|.
\end{aligned}
$$

Note that, according to Strang (2019, I.8), we only need to determine the vectors $\left(((\widetilde{\mathbf{u}}_{kd})_{q,r})_{q\in[Q-r]}\right)_{r\in[R_{kd}]}$ and $\left(((\widetilde{\mathbf{v}}_{kd})_{r,p})_{j\in[\mathrm{card}(J_{in})-r]}\right)_{r\in[R_{kd}]}$ since the remaining elements of such vectors belong to the the nullspace of $\boldsymbol{\Upsilon}_{kd}^{R_{kd}}$ and $\boldsymbol{\Upsilon}_{kd}^{R_{kd}\top}$. The number of total free parameters in the previous two vectors are

$$
\begin{aligned}
\sum_{r=1}^{R_{kd}} (Q - r) &= R_{kd}\left(\frac{2Q - R_{kd} - 1}{2}\right), \\
\sum_{r=1}^{R_{kd}} (\mathrm{card}\,(J_{in}) - r) &= R_{kd}\left(\frac{2\,\mathrm{card}\,(J_{in}) - R_{kd} - 1}{2}\right).
\end{aligned}
$$

To this end, for all $k \in [K]$, $d \in [D_V]$, and $q \in [Q]$, we let

$$
(\widetilde{\boldsymbol{\Upsilon}}_{kd}^{R_{kd}})_{q,p} = \begin{cases} \sum_{r=1}^{R_{kd}} (\widetilde{\sigma}_{kd})_r (\widetilde{\mathbf{u}}_{kd})_{q,r} (\widetilde{\mathbf{v}}_{kd}^\top)_{r,p} & \text{if } p \in J_{in}, \\ 0 & \text{if } p \in [P] \setminus J_{in}. \end{cases}
$$

In particular, (36) is proved by the following entropy controlling

$$
\begin{aligned}
&\mathrm{card}\left(G_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}\left(\delta_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}\right)\right) \\
&\leq \left\lfloor \frac{4A_{\mathbf{u},\mathbf{v}}\sqrt{Q}}{\delta_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}} \right\rfloor^Q \prod_{d=1}^{D_V} \left\lfloor \frac{6R_{kd}A_\sigma A_{\mathbf{u},\mathbf{v}}^2 D_V P\sqrt{Q}}{\delta_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}} \right\rfloor^{R_{kd}} \left\lfloor \frac{12R_{kd}A_\sigma A_{\mathbf{u},\mathbf{v}}^2 D_V P\sqrt{Q}}{\delta_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}} \right\rfloor^{R_{kd}(q+\mathrm{card}(J_{in})-R_{kd}-1)} \\
&= \left\lfloor \frac{\exp\left(C_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}\right)}{\delta_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}} \right\rfloor^{\dim\left(\mathbf{V}_{(D_V,J,\mathbf{R}_k)}\right)}, \text{ where}
\end{aligned}
$$

$$
\dim\left(\mathbf{V}_{(D_V,J,\mathbf{R}_k)}\right) = Q + \sum_{d=1}^{D_V} R_{kd}\left(Q + \mathrm{card}\,(J_{in}) - R_{kd}\right),\ C_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}} = \frac{\ln\left(C_{(D_V,J,\mathbf{R}_k)}\right)}{\dim\left(\mathbf{V}_{(D_V,J,\mathbf{R}_k)}\right)},
$$

and $C_{(D_V,J,\mathbf{R}_k)} = \left\lfloor 4A_{\mathbf{u},\mathbf{v}}\sqrt{Q} \right\rfloor^Q \left\lfloor 12R_{kd}A_\sigma A_{\mathbf{u},\mathbf{v}}^2 D_V P\sqrt{Q} \right\rfloor^{\sum_{d=1}^{D_V} R_{kd}(Q+\mathrm{card}(J_{in})-R_{kd})} 2^{-\sum_{d=1}^{D_V} R_{kd}}$.

Step 3: Upper bound of the number of the bracketing entropy for $\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}$. Next, in order to evaluate the ratio of two Gaussian densities, we make use of Lemma 10.

Lemma 10 (Proposition C.1 from Maugis and Michel (2011b)) *Let $\phi(\cdot;\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$ and $\phi(\cdot;\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$ be two Gaussian densities. If $\boldsymbol{\Sigma}_2-\boldsymbol{\Sigma}_1$ is a positive definite matrix then for all $\mathbf{y}\in\mathbb{R}^Q$,*

$$\frac{\phi(\mathbf{y};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)}{\phi(\mathbf{y};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)}\leq\sqrt{\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}}\exp\left[\frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top(\boldsymbol{\Sigma}_2-\boldsymbol{\Sigma}_1)^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\right].$$

Then, Lemma 11 allows us to fulfill the assumptions of Lemma 10.

Lemma 11 (Similar to Lemma B.8 from Maugis and Michel (2011b)) *Assume that $0<\epsilon<c_{\boldsymbol{\Sigma}}^2/9$, and set $\alpha=3\sqrt{\epsilon}/c_{\boldsymbol{\Sigma}}$. Then, for every $k\in[K]$, $(1+\alpha)\widetilde{\boldsymbol{\Sigma}}_k(B_k)-\boldsymbol{\Sigma}_k(B_k)$ and $\boldsymbol{\Sigma}_k(B_k)-(1+\alpha)^{-1}\widetilde{\boldsymbol{\Sigma}}_k(B_k)$ are both positive definite matrices. Moreover, for all $\mathbf{y}\in\mathbb{R}^Q$,*

$$\mathbf{y}^\top\left[(1+\alpha)\widetilde{\boldsymbol{\Sigma}}_k(B_k)-\boldsymbol{\Sigma}_k(B_k)\right]\mathbf{y}\geq\epsilon\|\mathbf{y}\|_2^2,\quad\mathbf{y}^\top\left[\boldsymbol{\Sigma}_k(B_k)-(1+\alpha)^{-1}\widetilde{\boldsymbol{\Sigma}}_k(B_k)\right]\mathbf{y}\geq\epsilon\|\mathbf{y}\|_2^2.$$

Proof For all $\mathbf{y}\neq\mathbf{0}$, since $\sup_{\lambda\in\mathrm{vp}(\boldsymbol{\Sigma}_k(B_k)-\widetilde{\boldsymbol{\Sigma}}_k(B_k))}|\lambda|=\left\|\boldsymbol{\Sigma}_k(B_k)-\widetilde{\boldsymbol{\Sigma}}_k(B_k)\right\|_2\leq\epsilon$, $-\epsilon\geq-c_{\boldsymbol{\Sigma}}/3$, and $\alpha=3\epsilon/c_{\boldsymbol{\Sigma}}$, it follow that

$$\begin{aligned}\mathbf{y}^\top\left[(1+\alpha)\widetilde{\boldsymbol{\Sigma}}_k(B_k)-\boldsymbol{\Sigma}_k(B_k)\right]\mathbf{y}&=(1+\alpha)\mathbf{y}^\top\left[\widetilde{\boldsymbol{\Sigma}}_k(B_k)-\boldsymbol{\Sigma}_k(B_k)\right]\mathbf{y}+\alpha\mathbf{y}^\top\boldsymbol{\Sigma}_k(B_k)\mathbf{y}\\&\geq-(1+\alpha)\left\|\widetilde{\boldsymbol{\Sigma}}_k(B_k)-\boldsymbol{\Sigma}_k(B_k)\right\|_2\|\mathbf{y}\|_2^2+\alpha c_{\boldsymbol{\Sigma}}\|\mathbf{y}\|_2^2\\&\geq(\alpha c_{\boldsymbol{\Sigma}}-(1+\alpha)\epsilon)\|\mathbf{y}\|_2^2=(\alpha c_{\boldsymbol{\Sigma}}-\alpha\epsilon-\epsilon)\|\mathbf{y}\|_2^2\\&\geq\left(\frac{2}{3}\alpha c_{\boldsymbol{\Sigma}}-\epsilon\right)\|\mathbf{y}\|_2^2=\epsilon\|\mathbf{y}\|_2^2>0,\end{aligned}$$

and

$$\begin{aligned}&\mathbf{y}^\top\left[\boldsymbol{\Sigma}_k(B_k)-(1+\alpha)^{-1}\widetilde{\boldsymbol{\Sigma}}_k(B_k)\right]\mathbf{y}\\&=(1+\alpha)^{-1}\mathbf{y}^\top\left[\boldsymbol{\Sigma}_k(B_k)-\widetilde{\boldsymbol{\Sigma}}_k(B_k)\right]\mathbf{y}+\left(1-(1+\alpha)^{-1}\right)\mathbf{y}^\top\boldsymbol{\Sigma}_k(B_k)\mathbf{y}\\&\geq\left(\frac{\alpha c_{\boldsymbol{\Sigma}}-\epsilon}{1+\alpha}\right)\|\mathbf{y}\|_2^2=\frac{2\epsilon}{1+\alpha}\|\mathbf{y}\|_2^2\geq\epsilon\|\mathbf{y}\|_2^2>0.\end{aligned}$$

■

By using Lemma 10 and the same argument as in the proof of Lemma B.9 from Maugis and Michel (2011b), given $0<\epsilon<c_{\boldsymbol{\Sigma}}/3$, where ϵ is chosen later, and $\alpha=3\epsilon/c_{\boldsymbol{\Sigma}}$, we obtain

$$\max\left\{\frac{l(\mathbf{x},\mathbf{y})}{f(\mathbf{x},\mathbf{y})},\frac{f(\mathbf{x},\mathbf{y})}{u(\mathbf{x},\mathbf{y})}\right\}\leq(1+2\alpha)^{-\frac{Q}{2}}\exp\left(\frac{\left\|\mathbf{v}_{(D_V,J,\mathbf{R}_k)}(\mathbf{x})-\widetilde{\mathbf{v}}_{(D_V,J,\mathbf{R}_k)}(\mathbf{x})\right\|_2^2}{2\epsilon}\right).\quad(39)$$

Because $\ln(\cdot)$ is a non-decreasing function, $\ln(1+2\alpha) \geq \alpha, \forall \alpha \in [0,1]$. Combined with (35) where $\delta^2_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}} = Q\alpha\epsilon$, we conclude that

$$\max\left\{\ln\left(\frac{l(\mathbf{x},\mathbf{y})}{f(\mathbf{x},\mathbf{y})}\right), \ln\left(\frac{f(\mathbf{x},\mathbf{y})}{u(\mathbf{x},\mathbf{y})}\right)\right\} \leq -\frac{Q}{2}\ln(1+2\alpha) + \frac{\delta^2_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}}{2\epsilon} \leq -\frac{Q}{2}\alpha + \frac{\delta^2_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}}{2\epsilon} = 0.$$

This means that $l(\mathbf{x},\mathbf{y}) \leq f(\mathbf{x},\mathbf{y}) \leq u(\mathbf{x},\mathbf{y}), \forall (\mathbf{x},\mathbf{y}) \in \mathcal{X}\times\mathcal{Y}$. Hence, it remains to bound the size of bracket $[l,u]$ w.r.t. $d_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}}$.

To this end, we aim to verify that $d^2_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}}(l,u) \leq \frac{\delta}{2}$. To accomplish this, we make use of Lemma 12.

Lemma 12 (Proposition C.3 from Maugis and Michel (2011b)) *Let $\phi(\cdot;\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$ and $\phi(\cdot;\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$ be two Gaussian densities with full rank covariance. It holds that*

$$\begin{aligned}
&d^2\left(\phi(\cdot;\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1), \phi(\cdot;\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)\right) \\
&= 2\left\{1 - 2^{q/2}\left|\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2\right|^{-1/4}\left|\boldsymbol{\Sigma}_1^{-1}+\boldsymbol{\Sigma}_2^{-1}\right|^{-1/2}\exp\left[-\frac{1}{4}\left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right)^\top\left(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2\right)^{-1}\left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right)\right]\right\}.
\end{aligned}$$

Therefore, using the fact that $\cosh(t) = \frac{e^{-t}+e^t}{2}$, Lemma 12 leads to, for all $\mathbf{x} \in \mathcal{X}$,

$$\begin{aligned}
\mathrm{d}^2(l(\mathbf{x},\cdot), u(\mathbf{x},\cdot)) &= \int_{\mathcal{Y}}\left[l(\mathbf{x},\mathbf{y}) + u(\mathbf{x},\mathbf{y}) - 2\sqrt{l(\mathbf{x},\mathbf{y})u(\mathbf{x},\mathbf{y})}\right]d\mathbf{y} \\
&= (1+2\alpha)^{-Q} + (1+2\alpha)^{Q} - 2 \\
&+ \mathrm{d}^2\left(\phi\left(\cdot;\widetilde{\mathbf{v}}_{(D_V,J,\mathbf{R}_k)}(\mathbf{x}), (1+\alpha)^{-1}\widetilde{\boldsymbol{\Sigma}}_k(B_k)\right), \phi\left(\cdot;\widetilde{\mathbf{v}}_{(D_V,J,\mathbf{R}_k)}(\mathbf{x}), (1+\alpha)\widetilde{\boldsymbol{\Sigma}}_k(B_k)\right)\right) \\
&= 2\cosh\left[Q\ln(1+2\alpha)\right] - 2 \\
&\quad + 2\left[1 - 2^{Q/2}\left[(1+\alpha)^{-1} + (1+\alpha)\right]^{-Q/2}\left|\widetilde{\boldsymbol{\Sigma}}_k(B_k)\right|^{-1/2}\left|\widetilde{\boldsymbol{\Sigma}}_k(B_k)\right|^{1/2}\right] \\
&= 2\cosh\left[Q\ln(1+2\alpha)\right] - 2 + 2 - 2\left[\cosh\left(\ln(1+\alpha)\right)\right]^{-Q/2} \\
&= 2g\left(Q\ln(1+2\alpha)\right) + 2h\left(\ln(1+\alpha)\right),
\end{aligned}$$

where $g(t) = \cosh(t) - 1 = \frac{e^{-t}+e^t}{2} - 1$, and $h(t) = 1 - \cosh(t)^{-Q/2}$. The upper bounds of terms g and h separately imply that, for all $\mathbf{y} \in \mathcal{Y}$,

$$\mathrm{d}^2(l(\mathbf{x},\cdot), u(\mathbf{x},\cdot)) \leq 2\left(2\cosh\left(\frac{1}{\sqrt{6}}\right)\alpha^2Q^2 + \frac{1}{4}\alpha^2Q^2\right) \leq 6\alpha^2Q^2 = \frac{\delta^2}{4},$$

where we choose $\alpha = \frac{3\epsilon}{c_{\mathbf{\Sigma}}}, \epsilon = \frac{\delta c_{\mathbf{\Sigma}}}{6\sqrt{6Q}}$, $\forall \delta \in (0,1], Q \in \mathbb{N}^\star, c_{\mathbf{\Sigma}} > 0$, which appears in (39) and satisfies $\alpha = \frac{\delta}{2\sqrt{6Q}}$ and $0 < \epsilon < \frac{c_{\mathbf{\Sigma}}}{3}$. Indeed, studying functions g and h yields

$$
\begin{aligned}
&\mathbf{g}'(t) = \sinh(t), \mathbf{g}''(t) = \cosh(t) \leq \cosh(c), \forall t \in [0,c], c \in \mathbb{R}_+, \\
&h'(t) = \frac{Q}{2}\cosh(t)^{-Q/2-1}\sinh(t), \\
&h''(t) = \frac{Q}{2}\left(-\frac{Q}{2}-1\right)\cosh(t)^{-Q/2-2}\sinh^2(t) + \frac{Q}{2}\cosh(t)^{-Q/2} \\
&\quad = \frac{Q}{2}\left(1 - \left(\frac{Q}{2}+1\right)\left(\frac{\sinh(t)}{\cosh(t)}\right)^2\right)\cosh(t)^{-Q/2} \leq \frac{Q}{2},
\end{aligned}
$$

where we used the fact that $\cosh(t) \geq 1$. Then, since $g(0) = 0, \mathbf{g}'(0) = 0, h(0) = 0, h'(0) = 0$, by applying Taylor's Theorem, it is true that

$$
\begin{aligned}
&g(t) = g(t) - g(0) - \mathbf{g}'(0)t = R_{0,1}(t) \leq \cosh(c)\frac{t^2}{2}, \forall t \in [0,c], \\
&h(t) = h(t) - h(0) - h'(0)t = R_{0,1}(t) \leq \frac{Q}{2}\frac{t^2}{2} \leq \frac{Q^2}{2}\frac{t^2}{2}, \forall t \geq 0.
\end{aligned}
$$

We wish to find an upper bound for $t = Q\ln(1+2\alpha)$, $Q \in \mathbb{N}^\star$, $\alpha = \frac{\delta}{2\sqrt{6Q}}$, $\delta \in (0,1]$. Since $\ln(\cdot)$ is an increasing function, then we have

$$
t = Q\ln\left(1 + \frac{\delta}{\sqrt{6}Q}\right) \leq Q\ln\left(1 + \frac{1}{\sqrt{6}Q}\right) \leq Q\frac{1}{\sqrt{6}Q} = \frac{1}{\sqrt{6}}, \forall \delta \in (0,1],
$$

since $\ln\left(1 + \frac{1}{\sqrt{6}Q}\right) \leq \frac{1}{\sqrt{6}Q}, \forall Q \in \mathbb{N}^\star$. Then, since $\ln(1+2\alpha) \leq 2\alpha, \forall \alpha \geq 0$,

$$
\begin{aligned}
&g\left(Q\ln(1+2\alpha)\right) \leq \cosh\left(\frac{1}{\sqrt{6}}\right)\frac{\left(Q\ln(1+2\alpha)\right)^2}{2} \leq \cosh\left(\frac{1}{\sqrt{6}}\right)\frac{Q^2}{2}4\alpha^2, \\
&h\left(\ln(1+\alpha)\right) \leq \frac{Q^2}{2}\frac{\left(\ln(1+\alpha)\right)^2}{2} \leq \frac{Q^2\alpha^2}{4}.
\end{aligned}
$$

Next, note that the set of $\delta/2$-brackets $[l,u]$ over $\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}$ is totally defined by the parameter spaces $\tilde{S}^{\text{disc}}_{B_k}(\epsilon)$ and $G_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}\left(\delta_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}\right)$. This leads to an upper bound of the $\delta/2$-bracketing entropy of $\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}$ is evaluated from an upper bound of the two set cardinalities. Hence, given any $\delta > 0$, by choosing $\epsilon = \frac{\delta c_{\mathbf{\Sigma}}}{6\sqrt{6Q}}$, $\alpha = \frac{3\epsilon}{c_{\mathbf{\Sigma}}} = \frac{\delta}{2\sqrt{6Q}}$, and $\delta^2_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}} = Q\alpha\epsilon =$

$Q\frac{\delta}{2\sqrt{6}Q}\frac{\delta c_{\mathbf{\Sigma}}}{6\sqrt{6}Q} = \frac{\delta^2 c_{\mathbf{\Sigma}}}{72Q}$, it holds that

$$
\begin{aligned}
&\mathcal{N}_{[\cdot],d_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}}}\left(\frac{\delta}{2},\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}\right)\\
&\leq \operatorname{card}\left(\tilde{S}^{\text{disc}}_{B_k}(\epsilon)\right)\times\operatorname{card}\left(G_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}\left(\delta_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}\right)\right)\\
&\leq \left(\left\lfloor\frac{2C_{\mathbf{\Sigma}}}{\epsilon}\right\rfloor\frac{Q\left(Q-1\right)}{2D_{B_k}}\right)^{D_{B_k}}\left(\frac{\exp\left(C_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}\right)}{\delta_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}}\right)^{\dim\left(\mathbf{V}_{(D_V,J,\mathbf{R}_k)}\right)} \qquad \text{(using (33) and (36))}\\
&\leq \left(\frac{2C_{\mathbf{\Sigma}}6\sqrt{6}Q}{\delta c_{\mathbf{\Sigma}}}\frac{Q\left(Q-1\right)}{2D_{B_k}}\right)^{D_{B_k}}\left(\frac{6\sqrt{2Q}\exp\left(C_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}\right)}{\delta\sqrt{c_{\mathbf{\Sigma}}}}\right)^{\dim\left(\mathbf{V}_{(D_V,J,\mathbf{R}_k)}\right)}\\
&= \left(\frac{6\sqrt{6}C_{\mathbf{\Sigma}}Q^2\left(Q-1\right)}{c_{\mathbf{\Sigma}}D_{B_k}}\right)^{D_{B_k}}\left(\frac{6\sqrt{2Q}\exp\left(C_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}\right)}{\sqrt{c_{\mathbf{\Sigma}}}}\right)^{\dim\left(\mathbf{V}_{(D_V,J,\mathbf{R}_k)}\right)}\left(\frac{1}{\delta}\right)^{D_{B_k}+\dim\left(\mathbf{V}_{(D_V,J,\mathbf{R}_k)}\right)}.
\end{aligned}
$$

To this end, note that $\dim\left(\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}\right) = D_{B_k}+\dim\left(\mathbf{V}_{(D_V,J,\mathbf{R}_k)}\right)$, we obtain

$$
\begin{aligned}
&\mathcal{H}_{[\cdot],d_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}}}\left(\frac{\delta}{2},\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}\right)\\
&= \ln\left(\mathcal{N}_{[\cdot],d_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}}}\left(\frac{\delta}{2},\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}\right)\right)\\
&\leq D_{B_k}\ln\left(\frac{6\sqrt{6}C_{\mathbf{\Sigma}}Q^2\left(Q-1\right)}{c_{\mathbf{\Sigma}}D_{B_k}}\right)+\dim\left(\mathbf{V}_{(D_V,J,\mathbf{R}_k)}\right)\ln\left(\frac{6\sqrt{2Q}\exp\left(C_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}\right)}{\sqrt{c_{\mathbf{\Sigma}}}}\right)\\
&\quad+\left(D_{B_k}+\dim\left(\mathbf{V}_{(D_V,J,\mathbf{R}_k)}\right)\right)\ln\left(\frac{1}{\delta}\right)\\
&= \dim\left(\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}\right)\left(C_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}}+\ln\left(\frac{1}{\delta}\right)\right),
\end{aligned}
$$

where $C_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}} = \dfrac{D_{B_k}\ln\left(\frac{6\sqrt{6}C_{\mathbf{\Sigma}}Q^2(Q-1)}{c_{\mathbf{\Sigma}}D_{B_k}}\right)+\dim\left(\mathbf{V}_{(D_V,J,\mathbf{R}_k)}\right)\ln\left(\frac{6\sqrt{2Q}\exp\left(C_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}\right)}{\sqrt{c_{\mathbf{\Sigma}}}}\right)}{\dim\left(\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}\right)}$.

Appendix C. Update for the gating network

C.1. MM algorithm for updating the gating network

The task of determining the maximizers of (19) may be complicated by various factors that fall outside the scope of traditional optimization. Such factors include the lack of differentiability of the objective functions $O_{\text{pen}}\left(\boldsymbol{\omega};\boldsymbol{\psi}^{(\text{t})}\right)$ or difficulty in obtaining closed-form solutions to the first-order condition (FOC) equation $\nabla_{\boldsymbol{\omega}}O_{\text{pen}}\left(\boldsymbol{\omega};\boldsymbol{\psi}^{(\text{t})}\right)=\mathbf{0}$, where $\nabla_{\boldsymbol{\omega}}$ is the gradient operator with respect

to $\boldsymbol{\omega}$. To overcome such difficulties, De Leeuw (1977) presented an MM algorithm for multidimensional scaling contemporary with the classic Dempster et al. (1977) paper on EM algorithms, then Hunter and Lange (2000) proposed the MM algorithm framework to solve the quantile regression via iterative minimization of surrogate functions. MM algorithms are particularly attractive due to the monotonicity and thus stability of their objective sequences as well as the global convergence of their limits, in general settings. A comprehensive treatment of the theory and implementation of MM algorithms for various general problems can be found in Hunter and Lange (2004); Lange (2016); Nguyen (2017), in particular for Lasso-penalized mixture of linear regression models in Lloyd-Jones et al. (2018).

Definition 13 (Philosophy of the MM algorithm, *e.g.*, Hunter and Lange (2004); Nguyen (2017)) *Let $\theta^{(r)}$ a fixed value of the parameter θ, and let $G\left(\theta;\theta^{(r)}\right)$ represent a real-value function of θ whose form depends on $\theta^{(r)}$. The function $G\left(\theta;\theta^{(r)}\right)$ is said to minorize $F(\theta)$ at the point $\theta^{(r)}$ if and only if for all θ, it holds that*

$$F\left(\theta\right) \geq G\left(\theta;\theta^{(r)}\right), \quad F\left(\theta^{(r)}\right) \geq G\left(\theta;\theta^{(r)}\right). \tag{40}$$

In other words, the surface $\theta \mapsto G\left(\theta;\theta^{(r)}\right)$ lies below the surface $F(\theta)$ and is tangent to it at the point $\theta = \theta^{(r)}$. Suppose we wish to obtain

$$\hat{\theta} = \arg\max_{\theta\in\Theta} F(\theta), \tag{41}$$

for some difficulty to manipulate objective function F, where Θ is a subset of some Euclidean space. In the maximization step of the MM algorithm, we maximize the surrogate function $G\left(\theta;\theta^{(r)}\right)$, rather than the function $F(\theta)$ itself. Let $\theta^{(0)}$ be some initial value and $\theta^{(r)}$ be the r-th iterate. We say that $\theta^{(r+1)}$ is the $(r+1)$-th iterate of an MM algorithm if it satisfies

$$\theta^{(r+1)} = \arg\max_{\theta\in\Theta} G\left(\theta;\theta^{(r)}\right). \tag{42}$$

By Definition 13, we can deduce the monotonicity property of all MM algorithms. Indeed, we can show that the MM algorithm forces $F(\theta)$ uphill, because (42) and (40) imply that

$$F(\theta^{(r)}) = G\left(\theta^{(r)};\theta^{(r)}\right) \leq G\left(\theta^{(r+1)};\theta^{(r)}\right) \leq F(\theta^{(r)}). \tag{43}$$

If $G\left(\theta;\theta^{(r)}\right)$ is well constructed, then we can avoid matrix inversion when maximizing it. Next, we devise the surrogate function for $O_{\text{pen}}\left(\boldsymbol{\omega};\boldsymbol{\psi}^{(\text{t})}\right)$ via Lemma 14.

Lemma 14 *The objective function $O_{pen}\left(\boldsymbol{\omega};\boldsymbol{\psi}^{(\text{t})}\right)$ is minorized at $\boldsymbol{\omega}^{(r)}$ by*

$$\begin{aligned} G\left(\boldsymbol{\omega};\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(\text{t})}\right) = &\sum_{n=1}^{N}\sum_{k=1}^{K-1}\tau_{nk}^{(\text{t})} w_k\left(\mathbf{x}_n;\boldsymbol{\omega}_k\right) + H\left(\boldsymbol{\omega};\boldsymbol{\omega}^{(r)}\right) \\ &- \sum_{k=1}^{K-1}\sum_{d=1}^{D_W}\lambda_{kd}^{[1]}\left\|\boldsymbol{\omega}_{kd}\right\|_1 - \frac{\lambda^{[3]}}{2}\sum_{k=1}^{K-1}\sum_{d=1}^{D_W}\left\|\boldsymbol{\omega}_{kd}\right\|_2^2, \end{aligned} \tag{44}$$

where $w_k\left(\mathbf{x}_n;\boldsymbol{\omega}_k\right)$ *is specified in* (2) *and* $H\left(\boldsymbol{\omega};\boldsymbol{\omega}^{(r)}\right)$ *minorizes* $-\sum_{n=1}^{N}\ln\left[1+\sum_{k=1}^{K-1}\exp\left(w_k\left(\mathbf{x}_n;\boldsymbol{\omega}_k\right)\right)\right]$ *and is defined as follows*

$$\sum_{n=1}^{N}\left[-\sum_{k=1}^{K-1}\frac{g_k\left(\mathbf{x}_n;\boldsymbol{\omega}^{(r)}\right)\sum_{d=0}^{D_W}\sum_{p=1}^{P}\exp\left[\left(PD_W+1\right)\left(\omega_{kdp}-\omega_{kdp}^{(r)}\right)x_{np}^{d}\right]}{PD_W+1}-\ln C_n^{(r)}+\frac{C_n^{(r)}-1}{C_n^{(r)}}\right].$$

Proof Firstly, we claim that if $\omega>0$, then the function $-\ln\left(1+\omega\right)$ can be minorized by

$$-\ln\left(1+\omega^{(r)}\right)-\frac{\omega-\omega^{(r)}}{1+\omega^{(r)}},\ \text{at}\ \omega^{(r)}>0. \tag{45}$$

One of the virtues of applying inequality (45) in defining a surrogate function is that it eliminates the log terms w.r.t. model parameters. Then, by (45), $-\ln\left[1+\sum_{k=1}^{K-1}\exp\left(w_k\left(\mathbf{x}_n;\boldsymbol{\omega}_k\right)\right)\right]$ is minorized by

$$\begin{aligned}
&-\ln\left[1+\sum_{k=1}^{K-1}\exp\left(w_k^{(r)}(\mathbf{x}_n)\right)\right]-\frac{\sum_{k=1}^{K-1}\left[\exp\left(w_k\left(\mathbf{x}_n;\boldsymbol{\omega}_k\right)\right)-\exp\left(w_k^{(r)}(\mathbf{x}_n)\right)\right]}{1+\sum_{k=1}^{K-1}\exp\left(w_k^{(r)}(\mathbf{x}_n)\right)}\\
&=-\ln C_n^{(r)}-\sum_{k=1}^{K-1}\frac{\exp\left(w_k^{(r)}(\mathbf{x}_n)\right)\exp\left(w_k\left(\mathbf{x}_n;\boldsymbol{\omega}_k\right)-w_k^{(r)}(\mathbf{x}_n)\right)}{C_n^{(r)}}+\frac{C_n^{(r)}-1}{C_n^{(r)}}.
\end{aligned}$$

Here, $C_n^{(r)}=1+\sum_{k=1}^{K-1}\exp\left(w_k^{(r)}(\mathbf{x}_n)\right)$. Now we wish to apply the weighted arithmetic-geometric mean inequality to the exponential functions $\exp\left(w_k\left(\mathbf{x}_n;\boldsymbol{\omega}_k\right)-w_k^{(r)}(\mathbf{x}_n)\right)$ to separate parameters. This feature is critically important in high-dimensional problems because it reduces optimization over $\mathbf{x}_n$ in potential large p-dimension to a sequence of one-dimensional optimizations over each component $x_{np}, n\in[N], p\in[P]$.

In fact, by the weighted arithmetic-geometric mean inequality,

$$\begin{aligned}
\exp\left(w_k\left(\mathbf{x}_n;\boldsymbol{\omega}_k\right)-w_k^{(r)}(\mathbf{x}_n)\right)&=\exp\left(\omega_{k0}-\omega_{k0}^{(r)}+\sum_{d=1}^{D_W}\left(\boldsymbol{\omega}_{kd}-\boldsymbol{\omega}_{kd}^{(r)}\right)^{\top}\mathbf{x}_n^d\right)\\
&=\exp\left(\omega_{k0}-\omega_{k0}^{(r)}+\sum_{d=1}^{D_W}\sum_{p=1}^{P}\left(\omega_{kdp}-\omega_{kdp}^{(r)}\right)x_{np}^{d}\right)\\
&\leq\frac{\exp\left(PD_W+1\right)}{PD_W+1}\sum_{d=0}^{D_W}\sum_{p=1}^{P}\exp\left[\left(\omega_{kdp}-\omega_{kdp}^{(r)}\right)x_{np}^{d}\right],
\end{aligned} \tag{46}$$

where $\omega_{k0p}^{(r)}=\omega_{k0}^{(r)},\omega_{k0p}=\omega_{k0},x_{np}^{0}=1$, for all $p\in[P]$ and the equality holds when

$$\left(\omega_{k0},\left(\boldsymbol{\omega}_{kd}\right)_{d\in[D_W]}\right)=\left(\omega_{k0}^{(r)},\left(\boldsymbol{\omega}_{kd}^{(r)}\right)_{d\in[D_W]}\right).$$

Therefore, $-\sum_{n=1}^{N}\ln\left[1+\sum_{k=1}^{K-1}\exp\left(w_k\left(\mathbf{x}_n;\boldsymbol{\omega}_k\right)\right)\right]$ is minorized by $H\left(\boldsymbol{\omega};\boldsymbol{\omega}^{(r)}\right)$, defined as follows:

$$\sum_{n=1}^{N}\left[-\sum_{k=1}^{K-1}\frac{\exp\left(w_k^{(r)}(\mathbf{x}_n)\right)\sum_{d=0}^{D_W}\sum_{p=1}^{P}\exp\left[\left(PD_W+1\right)\left(\omega_{kdp}-\omega_{kdp}^{(r)}\right)x_{np}^{d}\right]}{C_n^{(r)}\left(PD_W+1\right)}-\ln C_n^{(r)}+\frac{C_n^{(r)}-1}{C_n^{(r)}}\right]$$
$$=\sum_{n=1}^{N}\left[-\sum_{k=1}^{K-1}\frac{g_k\left(\mathbf{x}_n;\boldsymbol{\omega}^{(r)}\right)\sum_{d=0}^{D_W}\sum_{p=1}^{P}\exp\left[\left(PD_W+1\right)\left(\omega_{kdp}-\omega_{kdp}^{(r)}\right)x_{np}^{d}\right]}{PD_W+1}-\ln C_n^{(r)}+\frac{C_n^{(r)}-1}{C_n^{(r)}}\right].$$

■

Lemma 14 allows us to maximize $O_{\text{pen}}\left(\boldsymbol{\omega};\boldsymbol{\psi}^{(\mathrm{t})}\right)$ via its surrogate function $G\left(\boldsymbol{\omega};\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(\mathrm{t})}\right)$, which benefits the elimination the log terms w.r.t. model parameters and avoiding matrix inversion in high-dimensional problems via separating of parameters. Next, we aim to decompose $G\left(\boldsymbol{\omega};\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(\mathrm{t})}\right)$ according to parameters as follows:

$$G\left(\boldsymbol{\omega};\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(\mathrm{t})}\right)=G\left(\omega_{k0};\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(\mathrm{t})}\right)+\sum_{k=1}^{K}\sum_{d=1}^{D_W}\sum_{p=1}^{P}G\left(\omega_{kdp};\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(\mathrm{t})}\right)+I(\boldsymbol{\omega}^{(r)}),\quad(47)$$

where $I(\boldsymbol{\omega}^{(r)})$ is a function of $\boldsymbol{\omega}^{(r)}$. For every $k\in[K-1], p\in[P], d\in\{0\}\cup[D_W]$, we have that

$$G\left(\omega_{k0};\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(\mathrm{t})}\right)=\sum_{n=1}^{N}\tau_{nk}^{(\mathrm{t})}\omega_{k0}-\sum_{n=1}^{N}\frac{g_k\left(\mathbf{x}_n;\boldsymbol{\omega}^{(r)}\right)\exp\left[\left(PD_W+1\right)\left(\omega_{k0}-\omega_{k0}^{(r)}\right)\right]}{PD_W+1},\quad(48)$$

and

$$G\left(\omega_{kdp};\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(\mathrm{t})}\right)=\sum_{n=1}^{N}\tau_{nk}^{(\mathrm{t})}x_{np}^{d}\omega_{kdp}-\sum_{n=1}^{N}\frac{g_k\left(\mathbf{x}_n;\boldsymbol{\omega}^{(r)}\right)\exp\left[\left(PD_W+1\right)x_{np}^{d}\left(\omega_{kdp}-\omega_{kdp}^{(r)}\right)\right]}{PD_W+1}$$
$$-\lambda_{kd}^{[1]}\left|\omega_{kdp}\right|-\frac{\lambda^{[3]}}{2}\omega_{kdp}^{2}.\quad(49)$$

Then, by maximizing (48), we can update the ω_{k0} via solving the first-order condition

$$\nabla_{\omega_{k0}}G\left(\omega_{k0};\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(\mathrm{t})}\right)=\sum_{n=1}^{N}\tau_{nk}^{(\mathrm{t})}-\left(PD_W+1\right)\exp\left[\left(PD_W+1\right)\left(\omega_{k0}-\omega_{k0}^{(r)}\right)\right]\frac{\sum_{n=1}^{N}g_k\left(\mathbf{x}_n;\boldsymbol{\omega}^{(r)}\right)}{PD_W+1}.$$

Then, by solving $\nabla_{\omega_{k0}}G\left(\omega_{k0};\boldsymbol{\omega}_k^{(r)},\boldsymbol{\psi}^{(\mathrm{t})}\right)=0$, we obtain that

$$\omega_{k0}^{(r+1)}=\omega_{k0}^{(r)}+\frac{1}{PD_W+1}\ln\left[\frac{\sum_{n=1}^{N}\tau_{nk}^{(\mathrm{t})}}{\sum_{n=1}^{N}g_k\left(\mathbf{x}_n;\boldsymbol{\omega}^{(r)}\right)}\right].\quad(50)$$

Remark that $G\left(\omega_{kdp};\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(\mathrm{t})}\right)$ is a concave and univariate function w.r.t. ω_{kdp}. Therefore, we can maximize it globally w.r.t. each coefficient ω_{kdp} separately and then avoid matrix inversion. Indeed, note that

$$\begin{aligned} G\left(\omega_{kdp};\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(\mathrm{t})}\right) &= U\left(\omega_{kdp};\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(\mathrm{t})}\right) - \lambda_{kd}^{[1]}\left|\omega_{kdp}\right| \\ &= \begin{cases} U\left(\omega_{kdp};\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(\mathrm{t})}\right) - \lambda_{kd}^{[1]}\omega_{kdp}, & \text{if } \omega_{kdp} > 0, \\ U\left(0;\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(\mathrm{t})}\right), & \text{if } \omega_{kdp} = 0, \\ U\left(\omega_{kdp};\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(\mathrm{t})}\right) + \lambda_{kd}^{[1]}\omega_{kdp}, & \text{if } \omega_{kdp} < 0, \end{cases} \end{aligned}$$

where

$$\begin{aligned} U\left(\omega_{kdp};\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(\mathrm{t})}\right) &= \sum_{n=1}^{N}\tau_{nk}^{(\mathrm{t})}x_{np}^{d}\omega_{kdp} \\ &\quad - \sum_{n=1}^{N}\frac{g_k\left(\mathbf{x}_n;\boldsymbol{\omega}^{(r)}\right)\exp\left[\left(PD_W+1\right)x_{np}^{d}\left(\omega_{kdp}-\omega_{kdp}^{(r)}\right)\right]}{PD_W+1} - \frac{\lambda^{[3]}}{2}\omega_{kdp}^{2}. \end{aligned}$$

Remark that $G\left(\cdot;\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(\mathrm{t})}\right)$ is a smooth concave function on both $\mathbb{R}^{+}$ and $\mathbb{R}^{-}$. We therefore can use a one-dimensional generalized Newton-Raphson (GNR) algorithm to find the global maximizers of these functions and compare with $G\left(0;\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(\mathrm{t})}\right)$ so that we have

$$\omega_{kdp}^{(r+1)} = \underset{\omega_{kdp}}{\arg\max}\, G\left(\omega_{kdp};\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(\mathrm{t})}\right). \tag{51}$$

After starting from an initial value $s=0$, $\omega_{kdp}^{(0)} = \omega_{kdp}^{(r)}$, at each iteration s of the GNR, according to the following updating rule:

$$\omega_{kdp}^{(s+1)} = \omega_{kdp}^{(s)} - \left(\frac{\partial^2 G\left(\omega_{kdp};\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(\mathrm{t})}\right)}{\partial^2\omega_{kdp}}\right)^{-1}\Bigg|_{\omega_{kdp}^{(s)}} \frac{\partial G\left(\omega_{kdp};\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(\mathrm{t})}\right)}{\partial\omega_{kdp}}\Bigg|_{\omega_{kdp}^{(s)}}. \tag{52}$$

Here, the scalar gradient and Hessian are respectively given by:

$$\begin{aligned} \frac{\partial G\left(\omega_{kdp};\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(\mathrm{t})}\right)}{\partial\omega_{kdp}} &= \begin{cases} \frac{\partial U\left(\omega_{kdp};\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(\mathrm{t})}\right)}{\partial\omega_{kdp}} - \lambda_{kd}^{[1]}, & \text{if } \omega_{kdp} > 0, \\ \frac{\partial U\left(\omega_{kdp};\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(\mathrm{t})}\right)}{\partial\omega_{kdp}} + \lambda_{kd}^{[1]}, & \text{if } \omega_{kdp} < 0, \end{cases} \\ \frac{\partial^2 G\left(\omega_{kdp};\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(\mathrm{t})}\right)}{\partial^2\omega_{kdp}} &= \frac{\partial^2 U\left(\omega_{kdp};\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(\mathrm{t})}\right)}{\partial^2\omega_{kdp}}, \qquad \text{if } \omega_{kdp} \neq 0. \end{aligned} \tag{53}$$

Note that we have

$$\begin{aligned} \frac{\partial U\left(\omega_{kdp};\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(\mathrm{t})}\right)}{\partial\omega_{kdp}} &= \sum_{n=1}^{N}\tau_{nk}^{(\mathrm{t})}x_{np}^{d} - \sum_{n=1}^{N}x_{np}^{d}g_k\left(\mathbf{x}_n;\boldsymbol{\omega}^{(r)}\right)\exp\left[\left(PD_W+1\right)x_{np}^{d}\left(\omega_{kdp}-\omega_{kdp}^{(r)}\right)\right] - \lambda^{[3]}\omega_{kdp}, \\ \frac{\partial^2 U\left(\omega_{kdp};\boldsymbol{\omega}^{(r)},\boldsymbol{\psi}^{(\mathrm{t})}\right)}{\partial^2\omega_{kdp}} &= -\left(PD_W+1\right)\sum_{n=1}^{N}x_{np}^{d2}g_k\left(\mathbf{x}_n;\boldsymbol{\omega}^{(r)}\right)\exp\left[\left(PD_W+1\right)x_{np}^{d}\left(\omega_{kdp}-\omega_{kdp}^{(r)}\right)\right] - \lambda^{[3]}. \end{aligned}$$

C.2. Coordinate ascent algorithm for updating the gating network

Motivated by Tseng (1988, 2001), we aim to use the coordinate ascent algorithm to update the parameters $\boldsymbol{\omega} = \left(\omega_{k0}, (\boldsymbol{\omega}_{kd})_{d\in[D_W]}\right)_{k\in[K]}$ of the gating networks. This is an iterative method so that we fix most elements of the vector of parameters at their values from the current iteration, and solve univariate optimization problems in a loop to the remaining elements. In this way, it allows us to solve more easily than the full problem based on the fact that each such task has lower dimensions in the optimization problems.

We first update to ω_{k0} by using a univariate Newton-Raphson algorithm. In particular, starting with initial values $\omega_{k0}^{(r=0)} = \omega_{k0}^{(\mathrm{t})}$ at t^{th} step in M-step, we attempt to construct a sequence of $\omega_{k0}^{(r>0)}$ that converges towards a maximizer of the objective function of ω_{k0}, that is,

$$\omega_{k0}^{(r)} = \arg\max_{\omega_{k0}} O_{\text{pen}}\left(\omega_{k0}; \boldsymbol{\psi}^{(\mathrm{t})}\right)$$

where

$$O_{\text{pen}}\left(\omega_{k0}; \boldsymbol{\psi}^{(\mathrm{t})}\right) = \sum_{n=1}^{N} \tau_{nk}^{(\mathrm{t})} \left(\omega_{k0} + \sum_{d=1}^{D_W} \boldsymbol{\omega}_{kd}^{\top} \mathbf{x}_n^d\right) - \sum_{n=1}^{N} \ln\left[1 + \sum_{k=1}^{K-1} \exp\left(\omega_{k0} + \sum_{d=1}^{D_W} \boldsymbol{\omega}_{kd}^{\top} \mathbf{x}_n^d\right)\right],$$

which is independent of the regularized part. Here, we apply the generalized Newton-Raphson (GNR) method for updating $\omega_{k0}^{(r)}$ at step r of the coordinate ascent algorithm based on the second-order Taylor series approximation of $O_{\text{pen}}\left(\omega_{k0} + h;\ \boldsymbol{\psi}^{(\mathrm{t})}\right)$ with h sufficiently close to zero. More specifically, at each r^{th} step, GNR algorithm iteratively improves the approximation of ω_{k0} by

$$\omega_{k0}^{(r,\, s+1)} = \omega_{k0}^{(r,\, s)} - \left(\frac{\partial^2 O_{\text{pen}}\left(\omega_{k0}; \boldsymbol{\psi}^{(\mathrm{t})}\right)}{\partial^2 \omega_{k0}}\right)^{-1} \Bigg|_{\omega_{k0}^{(s)}} \frac{\partial O_{\text{pen}}\left(\omega_{k0}; \boldsymbol{\psi}^{(\mathrm{t})}\right)}{\partial \omega_{k0}} \Bigg|_{\omega_{k0}^{(s)}},$$

with $\omega_{k0}^{(r,\, s=0)} = \omega_{k0}^{(r)}$. Essentially, it requires the expressions of the gradient and Hessian of $O_{\text{pen}}\left(\cdot, \boldsymbol{\psi}^{(\mathrm{t})}\right)$ with respect to ω_{k0} that can be computed respectively as follows

$$\frac{\partial O_{\text{pen}}\left(\omega_{k0}; \boldsymbol{\psi}^{(\mathrm{t})}\right)}{\partial \omega_{k0}} = \sum_{n=1}^{N} \tau_{nk}^{(\mathrm{t})} - \sum_{n=1}^{N} \frac{\exp\left(\omega_{k0} + \sum_{d=1}^{D_W} \boldsymbol{\omega}_{kd}^{\top} \mathbf{x}_n^d\right)}{C_n\left(\omega_{k0}\right)},$$

$$\frac{\partial^2 O_{\text{pen}}\left(\omega_{k0}; \boldsymbol{\psi}^{(\mathrm{t})}\right)}{\partial^2 \omega_{k0}} = -\sum_{n=1}^{N} \frac{\exp\left(\omega_{k0} + \sum_{d=1}^{D_W} \boldsymbol{\omega}_{kd}^{\top} \mathbf{x}_n^d\right)\left[C_n\left(\omega_{k0}\right) - \exp\left(\omega_{k0} + \sum_{d=1}^{D_W} \boldsymbol{\omega}_{kd}^{\top} \mathbf{x}_n^d\right)\right]}{C_n\left(\omega_{k0}\right)^2},$$

where $$C_n\left(\omega_{k0}\right) = 1 + \sum_{u=1}^{K-1} \exp\left(\omega_{u0} + \sum_{d=1}^{D_W} \boldsymbol{\omega}_{ud}^{\top} \mathbf{x}_n^d\right).$$

Similarly, for parameter vector $\boldsymbol{\omega}_{kd}, d \neq 0$, the coefficient ω_{kdp} can be updated at step r^{th} by

$$\omega_{kdp}^{(r)} = \arg\max_{\omega_{kdp}} O_{\text{pen}}\left(\boldsymbol{\omega}_{kd}; \boldsymbol{\psi}^{(\mathrm{t})}\right),$$

where

$$O_{\text{pen}}\left(\boldsymbol{\omega}_{kd};\boldsymbol{\psi}^{(\text{t})}\right)=U\left(\omega_{kdp};\boldsymbol{\psi}^{(\text{t})}\right)-\lambda_{kd}^{[1]}\left|\omega_{kdp}\right|=\begin{cases}U\left(\omega_{kdp};\boldsymbol{\psi}^{(\text{t})}\right)-\lambda_{kd}^{[1]}\omega_{kdp}, & \text{if } \omega_{kdp}>0,\\ U\left(0;\boldsymbol{\psi}^{(\text{t})}\right), & \text{if } \omega_{kdp}=0,\\ U\left(\omega_{kdp};\boldsymbol{\psi}^{(\text{t})}\right)+\lambda_{kd}^{[1]}\omega_{kdp}, & \text{if } \omega_{kdp}<0,\end{cases}$$

with

$$U\left(\omega_{kdp};\boldsymbol{\psi}^{(\text{t})}\right)=\sum_{n=1}^{N}\tau_{nk}^{(\text{t})}\left(\omega_{k0}+\sum_{d=1}^{D_W}\boldsymbol{\omega}_{kd}^{\top}\mathbf{x}_n^d\right)-\sum_{n=1}^{N}\ln\left[1+\sum_{k=1}^{K-1}\exp\left(\omega_{k0}+\sum_{d=1}^{D_W}\boldsymbol{\omega}_{kd}^{\top}\mathbf{x}_n^d\right)\right]-\frac{\lambda^{[3]}}{2}\omega_{kdp}^2.$$

With a similar approach, at each step r^{th} of the coordinate ascent algorithm, GNR approximately updates ω_{kdp} by

$$\omega_{kdp}^{(r,\,s+1)}=\omega_{kdp}^{(r,\,s)}-\left(\frac{\partial^2 O_{\text{pen}}\left(\boldsymbol{\omega}_{kd};\boldsymbol{\psi}^{(\text{t})}\right)}{\partial^2\omega_{kdp}}\right)^{-1}\Bigg|_{\omega_{kdp}^{(s)}}\frac{\partial O_{\text{pen}}\left(\boldsymbol{\omega}_{kd};\boldsymbol{\psi}^{(\text{t})}\right)}{\partial\omega_{kdp}}\Bigg|_{\omega_{kdp}^{(s)}},\tag{54}$$

with the initial value $\omega_{kdp}^{(r,\,s=0)}=\omega_{kdp}^{(r)}$. It is essential to require the expressions of the first and second orders gradient of $O_{\text{pen}}\left(\cdot;\boldsymbol{\psi}^{(\text{t})}\right)$ with respect to ω_{kdp}, in particular, that are

$$\begin{aligned}\frac{\partial O_{\text{pen}}\left(\boldsymbol{\omega}_{kd};\boldsymbol{\psi}^{(\text{t})}\right)}{\partial\omega_{kdp}}&=\begin{cases}\frac{\partial U\left(\omega_{kdp};\boldsymbol{\psi}^{(\text{t})}\right)}{\partial\omega_{kdp}}-\lambda_{kd}^{[1]}, & \text{if } \omega_{kdp}>0,\\ \frac{\partial U\left(\omega_{kdp};\boldsymbol{\psi}^{(\text{t})}\right)}{\partial\omega_{kdp}}+\lambda_{kd}^{[1]}, & \text{if } \omega_{kdp}<0,\end{cases}\\ \frac{\partial^2 O_{\text{pen}}\left(\boldsymbol{\omega}_{kd};\boldsymbol{\psi}^{(\text{t})}\right)}{\partial^2\omega_{kdp}}&=\frac{\partial^2 U\left(\omega_{kdp};\boldsymbol{\psi}^{(\text{t})}\right)}{\partial^2\omega_{kdp}},\ \text{if } \omega_{kdp}\neq 0.\end{aligned}\tag{55}$$

where

$$\begin{aligned}\frac{\partial U\left(\omega_{kdp};\boldsymbol{\psi}^{(\text{t})}\right)}{\partial\omega_{kdp}}&=\sum_{n=1}^{N}\tau_{nk}^{(\text{t})}x_{np}^d-\sum_{n=1}^{N}\frac{x_{np}^d\exp\left(\omega_{k0}+\sum_{d=1}^{D_W}\boldsymbol{\omega}_{kd}^{\top}\mathbf{x}_n^d\right)}{C_n\left(\omega_{kdp}\right)}-\lambda^{[3]}\omega_{kdp},\\ \frac{\partial^2 U\left(\omega_{kdp};\boldsymbol{\psi}^{(\text{t})}\right)}{\partial^2\omega_{kdp}}&=-\sum_{n=1}^{N}\frac{x_{np}^{d2}\exp\left(\omega_{k0}+\sum_{d=1}^{D_W}\boldsymbol{\omega}_{kd}^{\top}\mathbf{x}_n^d\right)\left[C_n\left(\omega_{kdp}\right)-\exp\left(\omega_{k0}+\sum_{d=1}^{D_W}\boldsymbol{\omega}_{kd}^{\top}\mathbf{x}_n^d\right)\right]}{C_n\left(\omega_{kdp}\right)^2}-\lambda^{[3]},\\ C_n\left(\omega_{kdp}\right)&=1+\sum_{u=1}^{K-1}\exp\left(\omega_{u0}+\sum_{d=1}^{D_W}\boldsymbol{\omega}_{ud}^{\top}\mathbf{x}_n^d\right).\end{aligned}$$

Appendix D. Updating the Gaussian expert networks

For the penalized MoE models with univariate response variables, performing the update for the Gaussian expert networks' parameters corresponds to solving K separated weighted Lasso problems (see Chamroukhi and Huynh (2019, Section 3.3.3) for more details). However, for multivariate

cases, generalizing the approach from Devijver (2017b), we propose a new method to deal with the following complex objective function

$$
\begin{aligned}
&O_{\text{pen}}\left(\boldsymbol{\Upsilon}, \boldsymbol{\Sigma}, \boldsymbol{\psi}^{(\mathrm{t})}\right) = \sum_{n=1}^{N}\sum_{k=1}^{K} \tau_{nk}^{(\mathrm{t})} \ln\left[\phi\left(\mathbf{y}_n; \mathbf{v}_k(\mathbf{x}_n; \boldsymbol{\Upsilon}_k), \boldsymbol{\Sigma}_k\right)\right] - \sum_{k=1}^{K}\sum_{d=1}^{D_V} \lambda_{kd}^{[2]} \left\|\boldsymbol{\Gamma}_{kd}\right\|_1 \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} \tau_{nk}^{(\mathrm{t})} \ln\left[\frac{1}{(2\pi)^{q/2}\det\left(\boldsymbol{\Sigma}_k\right)^{1/2}} \exp\left(-\frac{\left(\mathbf{y}_n - \mathbf{v}_k(\mathbf{x}_n; \boldsymbol{\Upsilon}_k)\right)^{\top} \boldsymbol{\Sigma}_k^{-1}\left(\mathbf{y}_n - \mathbf{v}_k(\mathbf{x}_n; \boldsymbol{\Upsilon}_k)\right)}{2}\right)\right] \\
&\quad - \sum_{k=1}^{K}\sum_{d=1}^{D_V} \lambda_{kd}^{[2]} \left\|\boldsymbol{\Gamma}_{kd}\right\|_1 \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} \tau_{nk}^{(\mathrm{t})} \ln\left[\frac{\det\left(\mathbf{Q}_k\right)}{(2\pi)^{q/2}} \exp\left(-\frac{\left(\mathbf{Q}_k\mathbf{y}_n - \sum_{d=0}^{D_V}\boldsymbol{\Gamma}_{kd}\mathbf{x}_n^d\right)^{\top}\left(\mathbf{Q}_k\mathbf{y}_n - \sum_{d=0}^{D_V}\boldsymbol{\Gamma}_{kd}\mathbf{x}_n^d\right)}{2}\right)\right] \\
&\quad - \sum_{k=1}^{K}\sum_{d=1}^{D_V} \lambda_{kd}^{[2]} \left\|\boldsymbol{\Gamma}_{kd}\right\|_1 \\
&= O_{\text{pen}}\left(\boldsymbol{\Gamma}, \mathbf{Q}; \boldsymbol{\psi}^{(\mathrm{t})}\right). \qquad (56)
\end{aligned}
$$

where $\boldsymbol{\Gamma}_{kd} = \mathbf{Q}_k\boldsymbol{\Upsilon}_{kd}$ with $\mathbf{Q}_k^{\top}\mathbf{Q}_k = \boldsymbol{\Sigma}_k^{-1}$ so that $\boldsymbol{\Gamma}_{k0} = \mathbf{Q}_k\boldsymbol{\upsilon}_{k0} \in \mathbb{R}^{Q\times 1}$, $\mathbf{x}_n^0 \equiv 1_{1\times 1}$. By (56), optimizing $O_{\text{pen}}\left(\boldsymbol{\Upsilon}, \boldsymbol{\Sigma}; \boldsymbol{\psi}^{(\mathrm{t})}\right)$ w.r.t. $(\boldsymbol{\Upsilon}, \boldsymbol{\Sigma})$ is equivalent to maximize the objective function $O_{\text{pen}}\left(\boldsymbol{\Gamma}, \mathbf{Q}; \boldsymbol{\psi}^{(\mathrm{t})}\right)$ w.r.t. $(\boldsymbol{\Gamma}, \mathbf{Q}) = (\boldsymbol{\Gamma}_k, \mathbf{Q}_k)_{k\in[K]} = (\boldsymbol{\Gamma}_{kd}, \mathbf{Q}_k)_{k\in[K], d\in(\{0\}\cup[D_V])}$.

Similar to Section C.2 for the gating network, we apply the block coordinate ascent algorithm to update $(\boldsymbol{\Gamma}, \mathbf{Q})$ of the expert networks. For all $n \in [N], k \in [K]$, let

$$
\left(\mathbf{y}_{nk}^{(\mathrm{t})}, \mathbf{x}_{nk}^{(\mathrm{t})}\right) = \sqrt{\tau_{nk}^{(\mathrm{t})}}\left(\mathbf{y}_n, \mathbf{x}_n\right) \in \mathbb{R}^Q \times \mathbb{R}^P, \quad \text{and} \quad N_k^{(\mathrm{t})} = \sum_{n=1}^{N} \tau_{nk}^{(\mathrm{t})}. \qquad (57)
$$

Then, $O_{\text{pen}}\left(\boldsymbol{\Gamma}, \mathbf{Q}; \boldsymbol{\psi}^{(\mathrm{t})}\right)$ can be decoupled for each components into k distinct optimization problems of the form

$$
\begin{aligned}
&O_{\text{pen}}\left(\boldsymbol{\Gamma}_k, \mathbf{Q}_k; \boldsymbol{\psi}^{(\mathrm{t})}\right) \\
&= \sum_{n=1}^{N} \tau_{nk}^{(\mathrm{t})} \sum_{q=1}^{Q} \ln\left(\left(\mathbf{Q}_k\right)_{q,q}\right) - \frac{1}{2}\sum_{n=1}^{N} \tau_{nk}^{(\mathrm{t})}\left(\mathbf{Q}_k\mathbf{y}_n - \sum_{d=0}^{D_V}\boldsymbol{\Gamma}_{kd}\mathbf{x}^d\right)^{\top}\left(\mathbf{Q}_k\mathbf{y}_n - \sum_{d=0}^{D_V}\boldsymbol{\Gamma}_{kd}\mathbf{x}^d\right) - \sum_{d=1}^{D_V}\lambda_{kd}^{[2]}\left\|\boldsymbol{\Gamma}_{kd}\right\|_1 \\
&= N_k^{(\mathrm{t})} \sum_{q=1}^{Q} \ln\left(\left(\mathbf{Q}_k\right)_{q,q}\right) - \frac{1}{2}\sum_{n=1}^{N}\left(\mathbf{Q}_k\mathbf{y}_{nk}^{(\mathrm{t})} - \sum_{d=0}^{D_V}\boldsymbol{\Gamma}_{kd}\mathbf{x}_{nk}^{(\mathrm{t})\,d}\right)^{\top}\left(\mathbf{Q}_k\mathbf{y}_{nk}^{(\mathrm{t})} - \sum_{d=0}^{D_V}\boldsymbol{\Gamma}_{kd}\mathbf{x}_{nk}^{(\mathrm{t})\,d}\right) - \sum_{d=1}^{D_V}\lambda_{kd}^{[2]}\left\|\boldsymbol{\Gamma}_{kd}\right\|_1 \\
&= N_k^{(\mathrm{t})} \sum_{q=1}^{Q} \ln\left(\left(\mathbf{Q}_k\right)_{q,q}\right) - \frac{1}{2}\sum_{n=1}^{N}\sum_{q=1}^{Q}\left(\left(\mathbf{Q}_k\right)_{q,q} y_{nkq}^{(\mathrm{t})} - \sqrt{\tau_{nk}^{(\mathrm{t})}}\boldsymbol{\Gamma}_{k0} - \sum_{d=1}^{D_V}\left(\boldsymbol{\Gamma}_{kd}\right)_{q,*}\mathbf{x}_{nk}^{(\mathrm{t})\,d}\right)^2 - \sum_{d=1}^{D_V}\lambda_{kd}^{[2]}\left\|\boldsymbol{\Gamma}_{kd}\right\|_1. \qquad (58)
\end{aligned}
$$

The last equality is based on the fact that $\mathbf{\Sigma}_k$ is the diagonal matrix for all cluster k by assumption, therefore, $\mathbf{Q}_k$ is also a diagonal matrix. With respect to $(\mathbf{Q}_k)_{q,q}$, the optimization of (58) is the closed-form solutions to the FOC equation $\nabla_{(\mathbf{Q}_k)_{q,q}} O_{\text{pen}}\left(\mathbf{\Gamma}_k, \mathbf{Q}_k; \boldsymbol{\psi}^{(\text{t})}\right) = 0$ which is

$$\begin{aligned}
&\frac{N_k^{(\text{t})}}{(\mathbf{Q}_k)_{q,q}} - \sum_{n=1}^{N} y_{nkq}^{(\text{t})} \left((\mathbf{Q}_k)_{q,q}\, y_{nkq}^{(\text{t})} - \sqrt{\tau_{nk}^{(\text{t})}} \mathbf{\Gamma}_{k0} - \sum_{d=1}^{D_V} (\mathbf{\Gamma}_{kd})_{q,*} \mathbf{x}_{nk}^{(\text{t})\,d} \right) = 0 \\
\iff &\sum_{n=1}^{N} y_{nkq}^{(\text{t})\,2} (\mathbf{Q}_k)_{q,q}^2 - \sum_{n=1}^{N} y_{nkq}^{(\text{t})} \left(\sqrt{\tau_{nk}^{(\text{t})}} \mathbf{\Gamma}_{k0} - \sum_{d=1}^{D_V} (\mathbf{\Gamma}_{kd})_{q,*} \mathbf{x}_{nk}^{(\text{t})\,d} \right) (\mathbf{Q}_k)_{q,q} - N_k^{(\text{t})} = 0, \quad (59)
\end{aligned}$$

which is a quadratic equation of $(\mathbf{Q}_k)_{q,q}$. Moreover, based on the fact that $(\mathbf{Q}_k)_{q,q} > 0$, we get that

$$(\mathbf{Q}_k)_{q,q} = \frac{\sum_{n=1}^{N} y_{nkq}^{(\text{t})} \left(\sqrt{\tau_{nk}^{(\text{t})}} \mathbf{\Gamma}_{k0} - \sum_{d=1}^{D_V} (\mathbf{\Gamma}_{kd})_{q,*} \mathbf{x}_{nk}^{(\text{t})\,d} \right) + \sqrt{\Delta_{k,q}}}{2 \sum_{n=1}^{N} y_{nkq}^{(\text{t})\,2}}, \quad (60)$$

where

$$\Delta_{k,q} = \left[\sum_{n=1}^{N} y_{nkq}^{(\text{t})} \left(\sqrt{\tau_{nk}^{(\text{t})}} \mathbf{\Gamma}_{k0} - \sum_{d=1}^{D_V} (\mathbf{\Gamma}_{kd})_{q,*} \mathbf{x}_{nk}^{(\text{t})\,d} \right) \right]^2 + 4 N_k^{(\text{t})} \sum_{n=1}^{N} y_{nkq}^{(\text{t})\,2}. \quad (61)$$

Similarly, with respect to $(\mathbf{\Gamma}_{kd})_{q,p}$, the optimization of (58) is the closed-form solutions to the FOC equation $\nabla_{(\mathbf{\Gamma}_{kd})_{q,p}} O_{\text{pen}}\left(\mathbf{\Gamma}_k, \mathbf{Q}_k; \boldsymbol{\psi}^{(\text{t})}\right) = 0$. More precisely, for $d = 0$, we have

$$\begin{aligned}
&\sum_{n=1}^{N} \sqrt{\tau_{nk}^{(\text{t})}} \sum_{q=1}^{Q} \left((\mathbf{Q}_k)_{q,q}\, y_{nkq}^{(\text{t})} - \sqrt{\tau_{nk}^{(\text{t})}} \mathbf{\Gamma}_{k0} - \sum_{d=1}^{D_V} (\mathbf{\Gamma}_{kd})_{q,*} \mathbf{x}_{nk}^{(\text{t})\,d} \right) = 0, \\
\iff &\sum_{n=1}^{N} \tau_{nk}^{(\text{t})} \sum_{q=1}^{Q} \mathbf{\Gamma}_{k0} = \sum_{n=1}^{N} \sqrt{\tau_{nk}^{(\text{t})}} \sum_{q=1}^{Q} \left((\mathbf{Q}_k)_{q,q}\, y_{nkq}^{(\text{t})} - \sum_{d=1}^{D_V} (\mathbf{\Gamma}_{kd})_{q,*} \mathbf{x}_{nk}^{(\text{t})\,d} \right) \\
\iff &\mathbf{\Gamma}_{k0} = \frac{\sum_{n=1}^{N} \sqrt{\tau_{nk}^{(\text{t})}} \sum_{q=1}^{Q} \left((\mathbf{Q}_k)_{q,q}\, y_{nkq}^{(\text{t})} - \sum_{d=1}^{D_V} (\mathbf{\Gamma}_{kd})_{q,*} \mathbf{x}_{nk}^{(\text{t})\,d} \right)}{Q \sum_{n=1}^{N} \tau_{nk}^{(\text{t})}}. \quad (62)
\end{aligned}$$

For every $d \in [D_V]$,

$$\begin{aligned}
&\sum_{n=1}^{N} \mathbf{x}_{nkp}^{(\text{t})\,d} \left((\mathbf{Q}_k)_{q,q}\, y_{nkq}^{(\text{t})} - \sqrt{\tau_{nk}^{(\text{t})}} \mathbf{\Gamma}_{k0} - \sum_{d=1}^{D_V} (\mathbf{\Gamma}_{kd})_{q,*} \mathbf{x}_{nk}^{(\text{t})\,d} \right) - \lambda_{kd}^{[2]} \operatorname{sign}\left((\mathbf{\Gamma}_{kd})_{q,p} \right) = 0, \\
\iff &\sum_{n=1}^{N} \left(\mathbf{x}_{nkp}^{(\text{t})\,d} \right)^2 (\mathbf{\Gamma}_{kd})_{q,p} = \sum_{n=1}^{N} \mathbf{x}_{nkp}^{(\text{t})\,d} \left((\mathbf{Q}_k)_{q,q}\, y_{nkq}^{(\text{t})} - \sqrt{\tau_{nk}^{(\text{t})}} \mathbf{\Gamma}_{k0} - \sum_{i=1, i\neq d}^{D_V} \sum_{j=1, j\neq p}^{P} (\mathbf{\Gamma}_{ki})_{q,j} \mathbf{x}_{nkj}^{(\text{t})\,i} \right) \\
&- \lambda_{kd}^{[2]} \operatorname{sign}\left((\mathbf{\Gamma}_{kd})_{q,p} \right). \quad (63)
\end{aligned}$$

Here, $\mathrm{sign}(\cdot)$ is the sign function. Therefore, we obtain

$$(\mathbf{\Gamma}_{kd})_{q,p} = \frac{(\widetilde{\mathbf{\Gamma}}_{kd}^{(\mathrm{t})})_{q,p} - \lambda_{kd}^{[2]}\mathrm{sign}\left((\mathbf{\Gamma}_{kd})_{q,p}\right)}{\sum_{n=1}^{N}\left(\mathbf{x}_{nkp}^{(\mathrm{t})\,d}\right)^2} = \begin{cases} \frac{(\widetilde{\mathbf{\Gamma}}_{kd}^{(\mathrm{t})})_{q,p} - \lambda_{kd}^{[2]}}{\sum_{n=1}^{N}\left(\mathbf{x}_{nkp}^{(\mathrm{t})\,d}\right)^2} & \text{if } (\widetilde{\mathbf{\Gamma}}_{kd}^{(\mathrm{t})})_{q,p} > \lambda_{kd}^{[2]}, \\ \frac{(\widetilde{\mathbf{\Gamma}}_{kd}^{(\mathrm{t})})_{q,p} + \lambda_{kd}^{[2]}}{\sum_{n=1}^{N}\left(\mathbf{x}_{nkp}^{(\mathrm{t})\,d}\right)^2} & \text{if } (\widetilde{\mathbf{\Gamma}}_{kd}^{(\mathrm{t})})_{q,p} < -\lambda_{kd}^{[2]}, \\ 0 & \text{if } -\lambda_{kd}^{[2]} \leq (\widetilde{\mathbf{\Gamma}}_{kd}^{(\mathrm{t})})_{q,p} \leq \lambda_{kd}^{[2]}. \end{cases} \tag{64}$$

where

$$(\widetilde{\mathbf{\Gamma}}_{kd}^{(\mathrm{t})})_{q,p} = \sum_{n=1}^{N}\mathbf{x}_{nkp}^{(\mathrm{t})\,d}\left((\mathbf{Q}_k)_{q,q}\,y_{nkq}^{(\mathrm{t})} - \sqrt{\tau_{nk}^{(\mathrm{t})}}\mathbf{\Gamma}_{k0} - \sum_{i=1,i\neq d}^{D_V}\sum_{j=1,j\neq p}^{P}(\mathbf{\Gamma}_{ki})_{q,j}\mathbf{x}_{nkj}^{(\mathrm{t})\,i}\right). \tag{65}$$

In summary, the updated formulas are as follows:

$$\begin{aligned}
\mathbf{\Sigma}_k^{(\mathrm{ite}+1)} &= \left[\mathbf{Q}_k^{(\mathrm{ite}+1)\,\top}\mathbf{Q}_k^{(\mathrm{ite}+1)}\right]^{-1}, \quad \mathbf{\Upsilon}_{kd}^{(\mathrm{ite}+1)} = \left[\mathbf{Q}_k^{(\mathrm{ite}+1)}\right]^{-1}\mathbf{\Gamma}_{kd}^{(\mathrm{ite}+1)}, \\
(\mathbf{Q}_k)_{q,q}^{(\mathrm{ite}+1)} &= \frac{\sum_{n=1}^{N} y_{nkq}^{(\mathrm{t})}\left(\sqrt{\tau_{nk}^{(\mathrm{t})}}\mathbf{\Gamma}_{k0}^{(\mathrm{t})} - \sum_{d=1}^{D_V}(\mathbf{\Gamma}_{kd})_{q,*}^{(\mathrm{t})}\mathbf{x}_{nk}^{(\mathrm{t})\,d}\right) + \sqrt{\Delta_{k,q}^{(\mathrm{t})}}}{2\sum_{n=1}^{N} y_{nkq}^{(\mathrm{t})\,2}}, \\
\mathbf{\Gamma}_{k0}^{(\mathrm{ite}+1)} &= \frac{\sum_{n=1}^{N}\sqrt{\tau_{nk}^{(\mathrm{t})}}\sum_{q=1}^{Q}\left((\mathbf{Q}_k)_{q,q}^{(\mathrm{t})}\,y_{nkq}^{(\mathrm{t})} - \sum_{d=1}^{D_V}(\mathbf{\Gamma}_{kd})_{q,*}^{(\mathrm{t})}\mathbf{x}_{nk}^{(\mathrm{t})\,d}\right)}{Q\sum_{n=1}^{N}\tau_{nk}^{(\mathrm{t})}} \\
(\mathbf{\Gamma}_{kd})_{q,p}^{(\mathrm{ite}+1)} &= \begin{cases} \frac{(\widetilde{\mathbf{\Gamma}}_{kd}^{(\mathrm{t})})_{q,p} - \lambda_{kd}^{[2]}}{\sum_{n=1}^{N}\left(\mathbf{x}_{nkp}^{(\mathrm{t})\,d}\right)^2} & \text{if } (\widetilde{\mathbf{\Gamma}}_{kd}^{(\mathrm{t})})_{q,p} > \lambda_{kd}^{[2]}, \\ \frac{(\widetilde{\mathbf{\Gamma}}_{kd}^{(\mathrm{t})})_{q,p} + \lambda_{kd}^{[2]}}{\sum_{n=1}^{N}\left(\mathbf{x}_{nkp}^{(\mathrm{t})\,d}\right)^2} & \text{if } (\widetilde{\mathbf{\Gamma}}_{kd}^{(\mathrm{t})})_{q,p} < -\lambda_{kd}^{[2]}, \\ 0 & \text{if } -\lambda_{kd}^{[2]} \leq (\widetilde{\mathbf{\Gamma}}_{kd}^{(\mathrm{t})})_{q,p} \leq \lambda_{kd}^{[2]}, \end{cases}
\end{aligned}$$

Here,

$$\begin{aligned}
\Delta_{k,q}^{(\mathrm{t})} &= \left[\sum_{n=1}^{N} y_{nkq}^{(\mathrm{t})}\left(\sqrt{\tau_{nk}^{(\mathrm{t})}}\mathbf{\Gamma}_{k0}^{(\mathrm{t})} - \sum_{d=1}^{D_V}(\mathbf{\Gamma}_{kd})_{q,*}^{(\mathrm{t})}\mathbf{x}_{nk}^{(\mathrm{t})\,d}\right)\right]^2 + 4N_k^{(\mathrm{t})}\sum_{n=1}^{N} y_{nkq}^{(\mathrm{t})\,2}, \\
(\widetilde{\mathbf{\Gamma}}_{kd}^{(\mathrm{t})})_{q,p} &= \sum_{n=1}^{N}\mathbf{x}_{nkp}^{(\mathrm{t})\,d}\left((\mathbf{Q}_k)_{q,q}^{(\mathrm{t})}\,y_{nkq}^{(\mathrm{t})} - \sqrt{\tau_{nk}^{(\mathrm{t})}}\mathbf{\Gamma}_{k0} - \sum_{i=1,i\neq d}^{D_V}\sum_{j=1,j\neq p}^{P}(\mathbf{\Gamma}_{ki})_{q,j}^{(\mathrm{t})}\mathbf{x}_{nkj}^{(\mathrm{t})\,i}\right).
\end{aligned}$$

Appendix E. A comprehensive classification and nomenclature of MoE models with softmax gating networks.

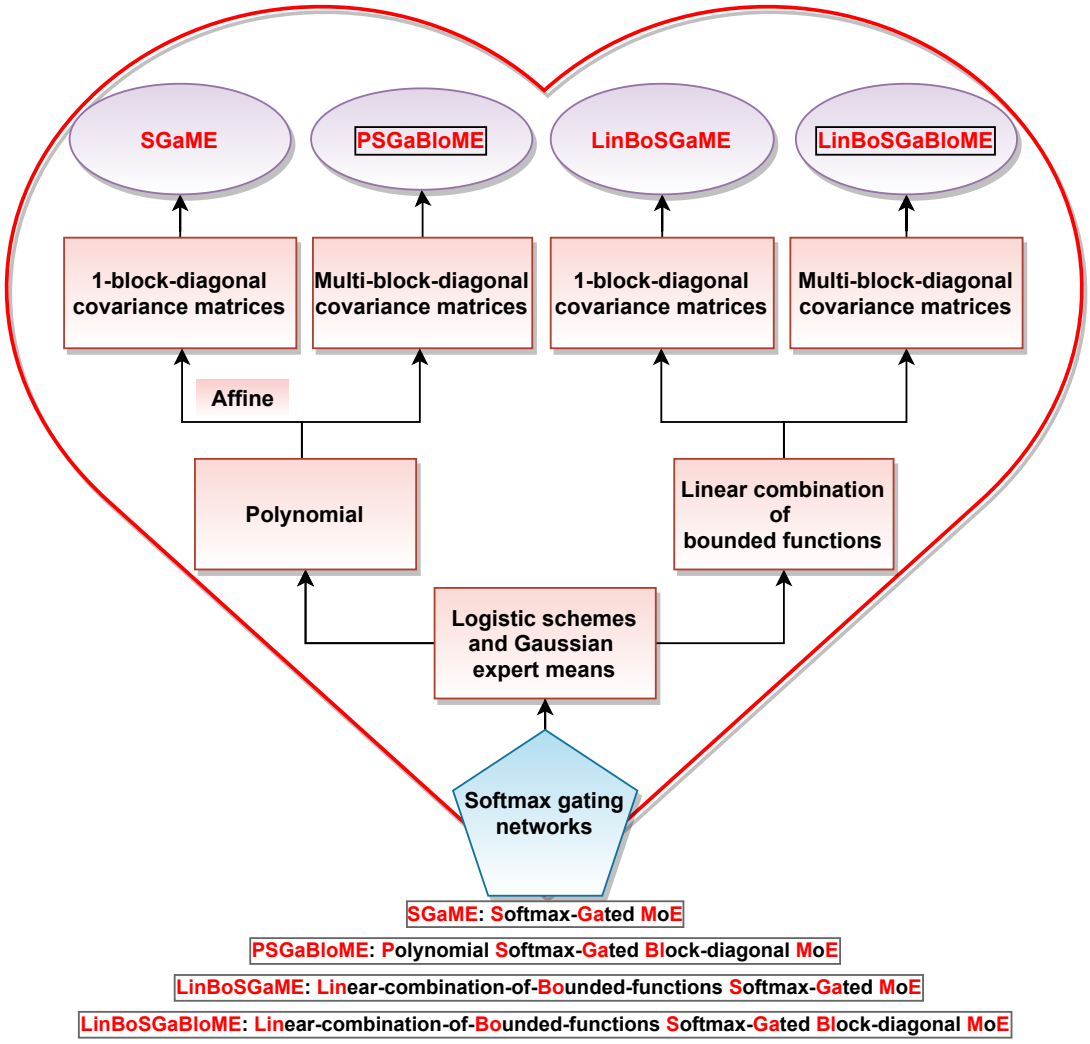

Figure 1: A comprehensive classification and nomenclature of MoE models with softmax gating networks.