A non-asymptotic risk bound 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

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

TrungTin Nguyen[1][0000−0001−8433−5980],
Dung Ngoc Nguyen[2][0000−0003−2471−6292],
Hien Duy Nguyen[3][0000−0002−9958−432X], and
Faicel Chamroukhi[4][0000−0002−5894−3103]

[1] Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, Inria Grenoble Rhone-Alpes, 655 av. de l'Europe, 38335 Montbonnot, France.
trung-tin.nguyen@inria.fr
[2] Department of Statistical Sciences, University of Padova, Italy.
ngocdung.nguyen@unipd.it
[3] School of Mathematics and Physics, University of Queensland, Australia.
h.nguyen7@uq.edu.au
[4] IRT SystemX, Palaiseau, France.
faicel.chamroukhi@irt-systemx.fr

Abstract. We are motivated by the problem of identifying potentially nonlinear regression relationships between high-dimensional outputs and high-dimensional inputs of heterogeneous data. This requires regression, clustering, and model selection, simultaneously. In this framework, we apply the mixture of experts models which are among the most popular ensemble learning techniques developed in the field of neural networks. In particular, we consider a more general case of mixture of experts models characterized by multiple Gaussian experts whose means are polynomials of the input variables and whose covariance matrices have block-diagonal structures. More especially, each expert is weighted by a gating network that is a softmax function of a polynomial of the input variables. These models require several hyper-parameters, including the number of mixture components, the complexity of the softmax gating networks and Gaussian mean experts, and the hidden block-diagonal structures of the covariance matrices. We provide a non-asymptotic theory for model selection of such complex hyper-parameters using the slope heuristic approach in a penalized maximum likelihood estimation framework. Specifically, we establish a non-asymptotic risk bound on the penalized maximum likelihood estimation, which takes the form of an oracle inequality, given lower bound assumptions on the penalty function.

Keywords: Dimensionality reduction · Low rank estimation · Mixture of experts · Finite mixture regression · Non-asymptotic model selection · Oracle inequality · Variable selection.

1 Introduction

Mixture of experts (MoE) models, introduced by Jacobs et al. [16] are widely applied to decompose the prediction model through a combination of gating models and expert models, both of which depend on the input variables. These flexible models are specific instances of conditional computation [3], where different model experts are responsible for different regions of the input space. Thus, by applying only a subset of parameters to each example, MoE can increase model capacity while keeping training and inference costs roughly constant. For reviews on this topic, we refer to [19, 25]. Furthermore, they have gained popularity due to universal approximation properties in various special cases, including mixture models [30, 28], mixture of regression models [15], and fully-parameterized mixture of experts models [26, 27]. In high-dimensional multivariate multiple regression for heterogeneous data, we refer to outputs $\mathbf{Y} \in \mathcal{Y} \subset \mathbb{R}^Q$ as target or response variables, and inputs $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^P$ as explanatory or predictor variables, where Q and P are both much larger than the sample size. Additionally, hidden interactions may exist in the graphical structure between response variables. In such cases, regression, clustering, and model selection need to be performed simultaneously. Consequently, we employ MoE models to identify potential nonlinear relationships between output and input variables in the high-dimensional heterogeneous data. We assume that $\mathbf{Y}$, conditional on $\mathbf{X}$, follows a distribution with the true but unknown probability density function $s_0(\cdot \mid \mathbf{X} = \mathbf{x})$. Motivated by universal approximation theorems for MoE models, s_0 can be estimated by

$$\begin{aligned} s_{\boldsymbol{\psi}_K}(\mathbf{y} \mid \mathbf{x}) &= \sum_{k=1}^{K} g_k(\mathbf{w}(\mathbf{x}))\phi(\mathbf{y}, \mathbf{v}_k(\mathbf{x}), \boldsymbol{\Sigma}_k(B_k)), \quad \text{with} \\ g_k(\mathbf{w}(\mathbf{x})) &= \frac{\exp(w_k(\mathbf{x}))}{\sum_{l=1}^{K} \exp(w_l(\mathbf{x}))}, \quad \text{for } k = 1, \ldots, K. \end{aligned} \tag{1}$$

Here, on each cluster $k \in \{1, \ldots, K\}$, g_k is called a softmax gating network corresponding to the weight functions, $\mathbf{w}(\mathbf{x}) = (w_1(\mathbf{x}), \ldots, w_K(\mathbf{x}))$, of $\mathbf{x}$, and $\phi(\cdot, \mathbf{v}_k(\mathbf{x}), \boldsymbol{\Sigma}_k(B_k))$ is a Gaussian expert with the mean function $\mathbf{v}_k(\mathbf{x})$ and covariance matrix $\boldsymbol{\Sigma}_k(B_k)$ depending on the block-diagonal structure B_k. We call $s_{\boldsymbol{\psi}_K}(\mathbf{y} \mid \mathbf{x})$, defined as in (1), the *softmax-gated block-diagonal MoE* (SGaBloME) models with unknown functional parameters $\boldsymbol{\psi}_K = (w_k, \mathbf{v}_k, \boldsymbol{\Sigma}_k(B_k))_{k \in \{1,\ldots,K\}}$. Furthermore, when the weights and the means of the SGaBloME model $s_{\boldsymbol{\psi}_K}$ are the functions depending on polynomials of the input variables $\mathbf{x}$ which are specified, for $k \in \{1, \ldots, K\}$, respectively, as

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

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

Here, $\boldsymbol{\omega}_0 = (\omega_{k0})_{k\in\{1,\dots,K\}}$, $\boldsymbol{\omega} = (\boldsymbol{\omega}_{k1}, \dots, \boldsymbol{\omega}_{kD_W})_{k\in\{1,\dots,K\}}$ and $\boldsymbol{\upsilon}_0 = (\boldsymbol{\upsilon}_{k0})_{k\in\{1,\dots,K\}}$, $\boldsymbol{\Upsilon} = (\boldsymbol{\Upsilon}_{k1}, \dots, \boldsymbol{\Upsilon}_{kD_V})_{k\in\{1,\dots,K\}}$ are K-tuples of unknown coefficients with the maximum degrees D_W and D_V of polynomials for the weight and mean functions, respectively, and $\mathbf{x}^d = (x_1^d, \dots, x_P^d)$ is a vector of all components of $\mathbf{x}$ with power d. Then we call an SGaBloME model $s_{\boldsymbol{\psi}_K}$ defined by (1) with the weights and mean experts specified as in (2)-(3) the polynomial SGaBloME model.

Motivation for block-diagonal covariance matrices. It is worth mentioning that the block-diagonal covariance matrices $\boldsymbol{\Sigma}(\mathbf{B}) = \big(\boldsymbol{\Sigma}_k(B_k)\big)_{k\in\{1,\dots,K\}}$ depend on the block structures $\mathbf{B} = (B_k)_{k\in\{1,\dots,K\}}$ that are the partitions of the outputs' index set $\{1, \dots, Q\}$ for each cluster. This structure is not only a trade-off between the model complexity and sparsity but is also motivated by some real-world applications, where one wishes to perform prediction on data sets with heterogeneous observations and graph-structured hidden interactions between the outputs. A relevant example is the gene expression data where, subject to phenotypic response, genes interact with only a few other genes, there are small modules of correlated genes, see e.g. [14] for more details.

Motivation for polynomial regression. To solve the high-dimensional regression problem, some authors applied SGaBloME models with certain simplifying assumptions. More specifically, Devijver [13] focused on a mixture of Gaussian linear regression models where the gating networks do not depend on the input variables. On the other hand, Chamroukhi et al. [7] considered MoE for multiple regression models with the univariate output variable, however, the weights and means are linear functions of the inputs and thus the capacity of MoE models is limited. In fact, in the context of convolutional neural networks, Chen et al. [8] have empirically found that the mixture of linear experts performs better than a single expert, but is still significantly worse than the mixture of non-linear experts. Within this framework, we are motivated to integrate nonlinearities into SGaBloME models by defining the weights and mean experts as *linear combinations of bounded functions* (LinBo) whose coefficients belong to a compact set. Such a general setting may include the polynomial basis with a bounded input domain, the suitable re-normalized wavelet dictionaries, or the Fourier basis on an interval. If the dimensions of the inputs and outputs are not too large, it is not necessary to select relevant variables and/or use rank sparse models. Then we can work on the softmax-gated MoE models with the linear combinations of bounded functions for weight and mean functions as in [24]. However, to deal with high-dimensional data and simplify the interpretation of sparsity, we consider a special case of LinBo-SGaBloME models to explore the presence of nonlinearities that is the class of polynomial SGaBloME models defined by (1)-(3). On the convergence rates of polynomial SGaBloME models, we refer to [23] for a discussion of the optimal convergence rate of an MoE model where each expert is associated with a polynomial regression model.

Model selection for polynomial SGaBloME models. The estimation of SGaBloME models can be performed by using a well-known expectation-maximization (EM)

algorithm [11], which obtains the global convergence in regression mixture models [18]. However, it crucially requires data-driven hyper-parameter choices, including the number of mixture components, the degree of complexity of each softmax gating network and each Gaussian expert mean function, and the hidden block-diagonal structures of the covariance matrices. Hyper-parameter choices from the data-driven learning algorithms belong to the class of model selection problems that select the model with the lowest risk from the data. Typically, penalization is one of the main strategies proposed for model selection that minimizes the sum of the empirical risk with a term of penalty so that the model can be fitted to data while avoiding the overfitting problem.

Related works. Typically, model selection for MoE models is performed using the asymptotic criteria [31, 4], whose uses in small samples are limited. Birgé et al. [5] proposed a novel approach, called *slope heuristic*, supported by a non-asymptotic oracle inequality via a general model selection theorem, see [2] and the references therein for recent reviews. This method leads to an optimal data-driven choice of multiplicative constants for penalties. In fact, oracle inequalities for the least absolute shrinkage and selection operator (Lasso) [32] and general penalized maximum likelihood estimators were established in the spirit of the methods based on concentration inequalities developed by [20]. These results include work on the simplified assumptions of MoE models such as high-dimensional Gaussian graphical models [14], Gaussian mixture model selection [21], and finite mixture regression models [12, 13], or softmax-gated MoE models with linear combinations of bounded functions for weight and mean functions without consideration of variable selection in the high-dimensional setting [24].

1.1 Main contributions

In this work, we established an oracle inequality for model selection, as shown in Theorem 1, under lower bound assumptions on penalty terms. This allows us to obtain non-asymptotic risk bounds in the form of weak oracle inequalities allowing the numbers of predictor and response variables that grow or are even much larger than the sample size. More concretely, the constructed oracle inequality shows that the performance of our penalized maximum likelihood estimations is comparable to that of oracle models with sufficiently large constant multiples of the penalties. The forms of these constants are only known up to multiplicative constants and are proportional to the dimensions of the models. Moreover, the flexibility of polynomial SGaBloME models requires the hyper-parameters comprising 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. Therefore, the aforementioned theoretical justifications for the penalty shapes motivate the use of the heuristic slope criterion to select these hyper-parameters of the models under consideration.

Notations. For any matrix $\mathbf{A}$ with the elements A_{ij}, we denote $|||\mathbf{A}|||_\infty = \max_{i,j} |A_{ij}|$ the max-norm, and $\mathbf{A}_{\cdot j}$ the j^{th} column, of $\mathbf{A}$. Furthermore, the

smallest and largest eigenvalues of $\mathbf{A}$ are denoted by $\text{eig}(\mathbf{A})$ and $\text{Eig}(\mathbf{A})$, respectively. The notation $\mathbf{A} \succ 0$ indicates that $\mathbf{A}$ is positive definite. For any vector $\mathbf{a}$, we denote $\|\mathbf{a}\|_p$ the l_p-norm of $\mathbf{a}$ for $0 < p \leq \infty$. We call $\dim(\mathcal{S})$ the total number of parameters to be estimated or the dimension of a parametric model $\mathcal{S}$. If S is a finite set, we denote $\text{card}(S)$ the cardinality, $\mathcal{P}(S)$ the set of all subsets, and $\mathcal{B}(S)$ the set of all partitions, of S. The set of all natural numbers without zero is denoted by $\mathbb{N}^*$. For any $K \in \mathbb{N}^*$, the notation $[K]$ corresponds to the set $\{1, \ldots, K\}$. Finally, we refer to $a \wedge b$ as $\min\{a, b\}$ for $a, b \in \mathbb{R}$.

Paper organization. The rest of the paper is organized as follows. Section 2 is devoted to the construction of a collection of polynomial SGaBloME models for high-dimensional heterogeneous data. In Section 3, we state the main theoretical results of oracle inequality for the penalized maximum likelihood estimations under some conditions on the parameter space and input domain of the models. Finally, Section 4 contains concluding remarks and future directions.

2 Collection of polynomial SGaBloME models

For high-dimensional data, it is necessary to work with parsimonious models by combining two well-known approaches: *selection of relevant variables* and *rank sparse models*. Within this framework, the collection of polynomial SGaBloME models is then constructed.

2.1 Variable selection via selecting relevant variables

In this section, we introduce the index sets for the input and output variables so that they are related to each other. This facilitates the variable selection of the models in a highly dimensional framework. In particular, for every $p \in [P], q \in [Q]$, we call a couple (X_p, Y_q) *irrelevant* if the elements $(\boldsymbol{\Upsilon}_{kd})_{qp} = 0$ and $(\boldsymbol{\omega}_{kl})_p = 0$ for all cluster $k \in [K]$ and degrees $d \in [D_V]$, $l \in [D_W]$. Therefore, the variables (X_p, Y_q) are *relevant* if they are not irrelevant. Formally, 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 thus the set of indices of relevant couples with $J \in \mathcal{P}([P] \times [Q])$. In addition, we also denote $J_{in} = \{p \in [P] : \exists q \in [Q], (p, q) \in J\}$ the set of indices of input variables that are relevant to the outputs so that $J_{in} \subseteq [P]$.

We notice that, for every cluster $k \in [K]$ and degree $d \in [D_V]$, all entries of $\boldsymbol{\Upsilon}_{kd}$ 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}. Hence, the matrix $\boldsymbol{\Upsilon}_{kd}$ will have $Q \times \text{card}\,(J_{in})$ coefficients to be estimated, which are smaller than $Q \times P$ when all variables are considered. The number of parameters in the regression matrices is therefore considerably reduced when the cardinality of J_{in} is much smaller than the number of input variables P. The subsets J or J_{in} can be constructed by the Lasso [32] and has been extended to deal with multiple multivariate regression models for column sparsity using the Group-Lasso [33].

2.2 Variable selection via rank sparse models

Anderson et al. [1] introduced rank sparse models in the regression framework which is 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 polynomial SGaBloME models, we assume that for every cluster $k \in [K]$ and degree $d \in [D_V]$, the matrix $\boldsymbol{\Upsilon}_{kd}$ has the associated rank R_{kd} and therefore it is completely determined by $R_{kd} \times (P - (Q - R_{kd}))$ coefficients, which can be less than $Q \times P$. Combined with the selection of relevant variables method, we denote a rank matrix by $\mathbf{R} = (R_{kd})_{k\in[K],d\in[D_V]}$ with the element $R_{kd} \in [\text{card}\,(J_{in}) \wedge Q]$ for each $k \in [K], d \in [D_V]$.

2.3 Collection of polynomial SGaBloME models

So far, each polynomial SGaBloME model defined by (1)-(3) can be characterized by the set $\mathbf{m} = (K, D_W, D_V, \mathbf{B}, J, \mathbf{R})$ where K is the number of clusters, D_W and D_V are the maximum degrees of polynomials of the weight and mean functions, respectively, $\mathbf{B}$ is the set of the block-diagonal structures of the covariance matrices, J is the set of relevant variables, and $\mathbf{R}$ is the rank matrix of coefficient matrices. Let $\mathcal{S}_{\mathbf{m}}$ be a class of (conditional) densities of polynomial SGaBloME models with respect to $\mathbf{m}$, which is specified as

$$
\begin{aligned}
\mathcal{S}_{\mathbf{m}} = \Big\{ & s_{\boldsymbol{\psi}_{\mathbf{m}}} \equiv s_{\boldsymbol{\psi}_K} \text{defined by (1)-(3) with} \\
& \boldsymbol{\psi}_{\mathbf{m}} = (\boldsymbol{\omega}_0, \boldsymbol{\omega}, \boldsymbol{\upsilon}_0, \boldsymbol{\Upsilon}, \boldsymbol{\Sigma}(\mathbf{B})) \in \boldsymbol{\Psi}_{\mathbf{m}}, \\
& \boldsymbol{\Psi}_{\mathbf{m}} = \mathbb{R}^K \times \mathbf{W}_J^{K\times D_W} \times \mathbb{R}^{K\times Q} \times \mathbf{V}_{J,\mathbf{R}}^{K\times D_V} \times \boldsymbol{\Omega}_{\mathbf{B}}^K, \\
& \mathbf{W}_J = \{\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_P) \in \mathbb{R}^P : \ \alpha_j = 0, \ \forall j \in ([P] \setminus J_{in})\}, \\
& \mathbf{V}_{J,\mathbf{R}} = \{\mathbf{A} \in \mathbb{R}^{Q\times P} : \mathbf{A}_{\cdot j} = \mathbf{0}, \ \forall j \in [P] \setminus J_{in} \text{ and } \text{rank}(\mathbf{A}) = R \in \mathbf{R}\} \\
& \boldsymbol{\Omega}_{\mathbf{B}} = \{\boldsymbol{\Sigma}(B) \in \mathbb{R}^{Q\times Q} : \boldsymbol{\Sigma}(B) \succ 0 \text{ and } \boldsymbol{\Sigma}(B) \text{ depends on block } B \in \mathbf{B}\}\Big\},
\end{aligned}
\tag{4}
$$

for every $\mathbf{m} \in \mathbb{N}^* \times \mathbb{N}^* \times \mathbb{N}^* \times \mathcal{B}([Q])^K \times \mathcal{P}([P] \times [Q]) \times [\text{card}\,(J_{in}) \wedge Q]^{K\times D_V}$. The collection of polynomial SGaBloME models defined in (4) is generally large and therefore not feasible in practice. Therefore, we restrict the set of (K, D_W, D_V) to a finite set of $(\mathcal{K}, \mathcal{D}_W, \mathcal{D}_V)$ where, for $K^*, D_W^*, D_V^* \in \mathbb{N}^*$, $\mathcal{K} = [K^*]$, $\mathcal{D}_W = [D_W^*]$, $\mathcal{D}_V = [D_V^*]$. Accordingly, the collection of polynomial SGaBloME models on the deterministic set of hyper-parameters can be defined as

$$
\begin{aligned}
\mathcal{S} = \Big\{ & \mathcal{S}_{\mathbf{m}} \text{ defined by (4) such that } \mathbf{m} \in \mathcal{M}, \\
& \mathcal{M} = \mathcal{K} \times \mathcal{D}_W \times \mathcal{D}_V \times \mathcal{B}([Q])^K \times \mathcal{P}\left([P] \times [Q]\right) \times [\text{card}\,(J_{in}) \wedge Q]^{K\times D_V} \Big\}.
\end{aligned}
\tag{5}
$$

Furthermore, because the block structures are specified by the partitions of the index set $\{1, \ldots, Q\}$, the number of such structures follows the so-called Bell number, which grows exponentially even for a moderate number of variables Q and clusters K. Therefore, it is infeasible to consider an exhaustive exploration

of the combination of all the partitions to detect the block structures for covariance matrices. Motivated by the recent work of [14], for a set of thresholds $\mathcal{E}$ and on 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 $\mathcal{B}([Q])$, where $\mathcal{B}_{k,\epsilon}$ is the partition of the output variables corresponding to the block-diagonal structure of the adjacency matrix $\mathbf{E}_{k,\epsilon}$ based on the thresholded absolute values of the sample covariance matrix $\mathbf{S}_k$. More formally, for each $\epsilon \in \mathcal{E}$, $(\mathbf{E}_{k,\epsilon})_{qq'} = 1$ if $|\,(\mathbf{S}_k)_{qq'}\,| > \epsilon$, otherwise it is equal to 0 for $q, q' \in [Q]$. In fact, Mazumder et al. [22] have shown that the class of block-diagonal structures detected by the graphical Lasso algorithm is identical to the block-diagonal structures detected by the thresholding of the sample covariance matrices, which supports our motivation for this restriction.

For the set of relevant variables, we focus on a random subset $\mathcal{J}$ of $\mathcal{P}([P] \times [Q])$ with the controlled size of $\mathcal{J}$ required in the high-dimension case. Accordingly, the number of possible vectors of ranks is reduced by working on a random subset of $[\mathrm{card}\,(J_{in}) \wedge Q]^{K\times D_V}$, which is denoted by $\mathcal{R}_{(K,J,D_V)}$ depending on $J \in \mathcal{J}$ with the dimension of $K \in \mathcal{K}$ and $D_V \in \mathcal{D}_V$. As a result, the collection of polynomial SGaBloME models based on a random sub-collection of hyper-parameters can be specified as

$$\widetilde{\mathcal{S}} = \Big\{\mathcal{S}_{\mathbf{m}} \text{ defined by (4) such that } \mathbf{m} \in \widetilde{\mathcal{M}}, \qquad (6)$$
$$\widetilde{\mathcal{M}} = \mathcal{K} \times \mathcal{D}_W \times \mathcal{D}_V \times (\mathcal{B}_{k,\mathcal{E}})_{k\in[K]} \times \mathcal{J} \times \mathcal{R}_{(K,J,D_V)}\Big\}.$$

3 Main theoretical results

In this section, we begin by introducing conditions on the parameter space of the models and give an overview of loss functions that are useful for comparing two (conditional) probability density functions. A general principle of penalized maximum likelihood estimation is also derived. Next, we show a finite sample oracle inequality used to ensure that if we penalize the log-likelihood in an approximate approach, we are able to select a model that is as good as the oracle.

3.1 Boundedness conditions on the parameter space

By motivation of integrating the nonlinearities into SGaBloME models discussed in Section 1, we consider the class of linear combinations of bounded functions for the weights and mean experts whose coefficients belong to compact sets with a bounded input domain. More specifically, we let $\big(\mathbf{X}_{[N]}, \mathbf{Y}_{[N]}\big) = \big((\mathbf{X}_1, \mathbf{Y}_1), \ldots, (\mathbf{X}_N, \mathbf{Y}_N)\big)$ be N pairs of real-valued random variables $(\mathbf{X}, \mathbf{Y})$ where the covariates $\mathbf{X}$ are assumed to belong to a hypercube, that is $\mathcal{X} = [0, 1]^P$. Then, there exist the constants $C_{\boldsymbol{\omega}}, C_{\boldsymbol{\Upsilon}}, c_{\boldsymbol{\Sigma}}, C_{\boldsymbol{\Sigma}} > 0$ such that, for every $k \in [K]$,

$$\|\boldsymbol{\omega}_{kd}\|_\infty \leq C_{\boldsymbol{\omega}}, \quad \|\!|\boldsymbol{\Upsilon}_{kl}|\!\|_\infty \leq C_{\boldsymbol{\Upsilon}}, \quad \text{for every } d \in [D_W],\ l \in [D_V], \qquad (7)$$

moreover, the eigenvalues of the block-diagonal covariances of the Gaussian experts lie on a positive interval, that is

$$0 < c_{\boldsymbol{\Sigma}} \leq \operatorname{eig}\left(\boldsymbol{\Sigma}_k\left(B_k\right)\right) \leq \operatorname{Eig}\left(\boldsymbol{\Sigma}_k\left(B_k\right)\right) \leq C_{\boldsymbol{\Sigma}}. \tag{8}$$

This setting can be applied to the case of polynomial functions for the weights of the softmax gates and the means of the Gaussian experts as we described in (2)-(3). More generally, the oracle inequality provided by Theorem 1 still holds for monomials of weights, allowing for the interaction between different inputs.

3.2 Loss function

To evaluate the maximum likelihood estimate, the Kullback-Leibler (KL) divergence is the most natural loss function, which is generally defined by

$$KL(s,t) = \begin{cases} \int_{\mathbb{R}^D} \ln\left(\frac{s(x)}{t(x)}\right) s(x)dyx & \text{if } sdx \text{ is absolutely continuous w.r.t. } tdx, \\ +\infty & \text{otherwise,} \end{cases}$$

where $s(\cdot)$ and $t(\cdot)$ are two density functions. In our work, we will apply the *tensorized KL divergence* to capture the structure of the density functions conditional on the random variables $\mathbf{X}$, that is

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

Another case of the tensorized KL divergence is the *tensorized Jensen-KL divergence* [9], which is given, for any $\rho \in (0,1)$, by

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

A relationship between the tensorized KL and the tensorized Jensen-KL divergence can be found in [10, Proposition 1].

3.3 Penalized maximum likelihood estimation (PMLE)

In the context of maximum likelihood estimation, given a collection $S_{\mathbf{m}}$, we aim to estimate s_0 by the conditional density $\widehat{s}_{\mathbf{m}}$ that minimizes the negative log-likelihood (NLL) as

$$\widehat{s}_{\mathbf{m}} = \underset{s_{\mathbf{m}} \in S_{\mathbf{m}}}{\arg\min} \sum_{n=1}^{N} -\ln\left[s_{\mathbf{m}}\left(\mathbf{Y}_n \mid \mathbf{X}_n\right)\right].$$

It is important to us to look for almost minimizer of this quantity and thereby define an η-log-likelihood minimizer (LLM) that satisfies

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

where the error term $\eta > 0$ is added to avoid any existence issue such as the infimum may not be reached. See [6, Chapter 2], [10, 9, 24] for more details of this literature. However, this approach underestimates the risk of the estimation and leads to the selection of overly complex models. Therefore, a trade-off between good data fit and model complexity can be found by adding an appropriate penalty term $\text{pen}(\mathbf{m})$. More concretely, for a given choice of $\text{pen}(\mathbf{m})$, the *selected model* $\mathcal{S}_{\widehat{\mathbf{m}}}$ is chosen as the one whose index $\widehat{\mathbf{m}}$ is a η'-minimizer of the sum of the NLL and penalty function, that is

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

for $\eta' > 0$. We then call $\widehat{s}_{\widehat{\mathbf{m}}}$ the η'-PMLE that depends on both error terms η and η'. From now on, the term *selected model* or *best data-driven model* is used to indicate the model that satisfies (10).

3.4 Oracle inequality

In this section, we provide the construction of an oracle inequality that guarantees a non-asymptotic theory for model selection in high-dimensional polynomial SGaBloME models.

Theorem 1. *Let* $\left(\mathbf{X}_{[N]}, \mathbf{Y}_{[N]}\right)$ *be a random sample of* $(\mathbf{X}, \mathbf{Y})$ *where* $\mathbf{Y}|\mathbf{X}$ *arises from the unknown conditional density* s_0*. For every* $\mathbf{m} = (K, D_W, D_V, \mathbf{B}, J, \mathbf{R}) \in \mathcal{M}$*, the model* $\mathcal{S}_{\mathbf{m}}$ *can be specified 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*

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

Furthermore, we construct a random sub-collection $\widetilde{\mathcal{S}}$ *of* $\mathcal{S}$ *such that every model of* $\widetilde{\mathcal{S}}$ *depends on the sets of* $\widetilde{\mathcal{M}} \subset \mathcal{M}$ *as in* (5)-(6)*. Then, there is a constant* C *such that, for any* $\rho \in (0, 1)$ *and* $C_1 > 1$*, there are two constants* κ *and* C_2 *depending only on* ρ *and* C_1 *such that, for every* $\mathbf{m} \in \mathcal{M}$*,* $\xi_{\mathbf{m}} \in \mathbb{R}^+$*,* $\Xi = \sum_{\mathbf{m}\in\mathcal{M}} e^{-\xi_{\mathbf{m}}} < \infty$ *and*

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

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

$$\begin{aligned}\mathbb{E}_{\mathbf{X}_{[N]},\mathbf{Y}_{[N]}}\left[\text{JKL}_{\rho}^{\otimes \text{N}}\left(s_0, \widehat{s}_{\widehat{\mathbf{m}}}\right)\right] \leq C_1 \mathbb{E}_{\mathbf{X}_{[N]},\mathbf{Y}_{[N]}}\left[\inf_{\mathbf{m}\in\widetilde{\mathcal{M}}}\left(\inf_{s\in\mathcal{S}_{\mathbf{m}}} \text{KL}^{\otimes \text{N}}\left(s_0, s\right) + 2\frac{pen(\mathbf{m})}{N}\right)\right] \\ + C_2(1 \vee \tau)\frac{\Xi^2}{N} + \frac{\eta' + \eta}{N}.\end{aligned} \tag{12}$$

Remarks. Theorem 1 guarantees that a penalized criterion leads to a good model selection and that the penalty is only known up to multiplicative constants 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. We notice that (11) is not a strong assumption and is satisfied in the case s_0 is bounded with 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 degrees, or enough clusters due to the universal approximation of MoE models. This results in the term on the right of (12) being small, for $\mathcal{D}_W, \mathcal{D}_V$ and $\mathcal{K}$ well chosen. It should be emphasized that Theorem 1 extends the main result of [24], 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 [17, 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 [29].

Main challenges on the proof of Theorem 1. To prove Theorem 1 it is inspired by [24] for handling LinBo-SGaME models, however, our method 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 the 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 [24], which uses a model selection theorem for a deterministic collection of models from [10, 9], we need to find a way to use the model selection theorem for MLE among a random sub-collection (cf. [12, Theorem 5.1] and [14, Theorem 7.3]). We refer readers to Sections S-1 and S-2 in the supplementary materials for a sketch of proof and detailed proof of Theorem 1.

4 Conclusion and perspectives

We have studied PMLEs for polynomial SGaBloME models 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 of the penalty hold. The future direction is to empirically evaluate our oracle inequality and to extend the current oracle inequality to more general settings where Gaussian experts are replaced by elliptic distributions.

Bibliography

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

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

[3] Bengio, Y.: Deep Learning of Representations: Looking Forward. In: Statistical Language and Speech Processing. pp. 1–37. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)

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

[5] Birgé, L., Massart, P.: Minimal penalties for Gaussian model selection. Probability Theory and Related Fields **138**(1), 33–73 (2007)

[6] Borwein, J.M., Zhu, Q.J.: Techniques of Variational Analysis. Springer (2004)

[7] Chamroukhi, F., Huynh, B.T.: Regularized maximum-likelihood estimation of mixture-of-experts for regression and clustering. In: 2018 International Joint Conference on Neural Networks (IJCNN). pp. 1–8 (2018)

[8] Chen, Z., Deng, Y., Wu, Y., Gu, Q., Li, Y.: Towards Understanding the Mixture-of-Experts Layer in Deep Learning. In: NeurIPS (2022)

[9] Cohen, S.X., Le Pennec, E.: Partition-based conditional density estimation. ESAIM: Probability and Statistics **17**, 672–697 (2013)

[10] Cohen, S., Le Pennec, E.: Conditional density estimation by penalized likelihood model selection and applications. Technical report, INRIA (2011)

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

[12] Devijver, E.: Finite mixture regression: a sparse variable selection by model selection for clustering. Electronic Journal of Statistics **9**(2), 2642–2674 (2015)

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

[14] Devijver, E., Gallopin, M.: Block-diagonal covariance selection for high-dimensional Gaussian graphical models. Journal of the American Statistical Association **113**(521), 306–314 (2018)

[15] Ho, N., Yang, C.Y., Jordan, M.I.: Convergence Rates for Gaussian Mixtures of Experts. Journal of Machine Learning Research **23**(323), 1–81 (2022)

[16] Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive mixtures of local experts. Neural Computation **3**(1), 79–87 (1991)

[17] Khalili, A.: New estimation and feature selection methods in mixture-of-experts models. Canadian Journal of Statistics **38**(4), 519–539 (2010)
[18] Kwon, J., Qian, W., Caramanis, C., Chen, Y., Davis, D.: Global convergence of the EM algorithm for mixtures of two component linear regression. In: COLT. vol. 99, pp. 2055–2110. PMLR (2019)
[19] Masoudnia, S., Ebrahimpour, R.: Mixture of experts: a literature survey. Artificial Intelligence Review **42**(2), 275–293 (2014)
[20] Massart, P.: Concentration Inequalities and Model Selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003. Springer (2007)
[21] Maugis, C., Michel, B.: A non asymptotic penalized criterion for Gaussian mixture model selection. ESAIM: Probability and Statistics **15**, 41–68 (2011)
[22] Mazumder, R., Hastie, T.: Exact covariance thresholding into connected components for large-scale graphical lasso. The Journal of Machine Learning Research **13**(1), 781–794 (2012)
[23] Mendes, E.F., Jiang, W.: On convergence rates of mixtures of polynomial experts. Neural Computation **24**(11), 3025–3051 (2012)
[24] Montuelle, L., Le Pennec, E., et al.: Mixture of Gaussian regressions model with logistic weights, a penalized maximum likelihood approach. Electronic Journal of Statistics **8**(1), 1661–1695 (2014)
[25] Nguyen, H.D., Chamroukhi, F.: Practical and theoretical aspects of mixture-of-experts modeling: An overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **8**(4), e1246 (2018)
[26] Nguyen, H.D., Chamroukhi, F., Forbes, F.: Approximation results regarding the multiple-output Gaussian gated mixture of linear experts model. Neurocomputing **366**, 208–214 (2019)
[27] Nguyen, H.D., Nguyen, T., Chamroukhi, F., McLachlan, G.J.: Approximations of conditional probability density functions in Lebesgue spaces via mixture of experts models. Journal of Statistical Distributions and Applications **8**(1), 13 (2021)
[28] Nguyen, T., Chamroukhi, F., Nguyen, H.D., McLachlan, G.J.: Approximation of probability density functions via location-scale finite mixtures in Lebesgue spaces. Communications in Statistics - Theory and Methods pp. 1–12 (2022)
[29] Nguyen, T., Nguyen, H.D., Chamroukhi, F., McLachlan, G.J.: An l_1-oracle inequality for the Lasso in mixture-of-experts regression models. arXiv preprint arXiv:2009.10622 (2020)
[30] Nguyen, T., Nguyen, H.D., Chamroukhi, F., McLachlan, G.J.: Approximation by finite mixtures of continuous density functions that vanish at infinity. Cogent Mathematics & Statistics **7**(1), 1750861 (2020)
[31] Schwarz, G., et al.: Estimating the dimension of a model. The Annals of Statistics **6**(2), 461–464 (1978)
[32] Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B **58**(1), 267–288 (1996)
[33] Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B **68**(1), 49–67 (2006)

Supplementary Materials for "A non-asymptotic risk bound for model selection in a high-dimensional mixture of experts via joint rank and variable selection"

TrungTin Nguyen[1][0000−0001−8433−5980],
Dung Ngoc Nguyen[2][0000−0003−2471−6292],
Hien Duy Nguyen[3][0000−0002−9958−432X], and
Faicel Chamroukhi[4][0000−0002−5894−3103]

[1] Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, Inria Grenoble Rhone-Alpes, 655 av. de l'Europe, 38335 Montbonnot, France.
trung-tin.nguyen@inria.fr
[2] Department of Statistical Sciences, University of Padova, Italy.
ngocdung.nguyen@unipd.it
[3] School of Mathematics and Physics, University of Queensland, Australia.
h.nguyen7@uq.edu.au
[4] IRT SystemX, Palaiseau, France.
faicel.chamroukhi@irt-systemx.fr

In this supplementary material, we provide the proof sketch and the detailed proof of Theorem 1 in Section S-1 and Section S-2 respectively. We then provide proofs for the remaining technical results in Section S-3.

S-1 Proof sketch of Theorem 1

It is worth noting that to deal with random sub-collection, we need to use a general model selection theorem for MLE under a random sub-collection (cf. [3, Theorem 5.1] or [5, Theorem 7.3]). This is the extension of [2, Theorem 2], which dealt with conditional density estimation but not random sub-collection, and [11, Theorem 7.11], which only works for density estimation. We then explain how we use Theorem S-1 to obtain the oracle inequality, Theorem 1. To do this, our model collection must satisfy some regularity assumptions, which are proved in Section S-3. 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 [13, 4]. For the second, we extend the recent novel result on block-diagonal covariance matrices in [5] for Gaussian mixture models from [8, 12].

General model selection theorem for MLE among a random sub-collection. First, we impose a structural assumption on each model indexed

by $\mathbf{m} \in \mathcal{M}$ regarding the bracketing entropy, defined by (S-2), 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

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

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,

$$\mathcal{N}_{[\cdot],d}(\delta, S) := \min\left\{ n \in \mathbb{N}^\star : \exists t_1^-, t_1^+, \ldots, t_n^-, t_n^+ \text{ s.t. } d(t_k^-, t_k^+) \leq \delta, S \subset \bigcup_{k=1}^{n} \left[t_k^-, t_k^+\right] \right\}, \tag{S-2}$$

where the term $s \in \left[t_k^-, t_k^+\right]$ 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 1 (H).

Assumption 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 $\mathcal{S}_{\mathbf{m}}(\widetilde{s}, \sigma) = \{s_{\mathbf{m}} \in \mathcal{S}_{\mathbf{m}} : \mathrm{d}^{\otimes \mathrm{n}}(\widetilde{s}, s_{\mathbf{m}}) \leq \sigma\}$. 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. [15, 7, 10]). 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 $\mathcal{S}_{\mathbf{m}}(\widetilde{s}, \sigma)$.

For technical reasons, a separability assumption, always satisfied in the setting of this paper, is also required. Assumption 2 (Sep) is a mild condition, which is classical in empirical process theory [15, 7] and allows us to work with a countable subset.

Assumption 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 $\iota(\mathcal{Y} \setminus \mathcal{Y}'_{\mathbf{m}}) = 0$, where ι denotes Lebesgue measure, such that for every $t \in \mathcal{S}_{\mathbf{m}}$, there exists some sequence $(t_k)_{k\in\mathbb{N}^\star}$ 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 [11**?**].

Assumption 3 (K) *There is a family* $(\xi_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$ *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 [3, Theorem 5.1] for the model selection theorem for MLE under a random sub-collection.

Theorem S-1 *Let* $(\mathbf{X}_n, \mathbf{Y}_n)_{n\in[N]}$ *be the observations coming from an unknown conditional density* s_0. *Let the model collection* $\mathcal{S} = (\mathcal{S}_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$ *be an at most countable collection of conditional density sets. Assume that Assumption 1 (H), Assumption 2 (Sep), and Assumption 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*

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

and let $\tau > 0$, *such that*

$$\bar{s}_{\mathbf{m}} \geq e^{-\tau} s_0. \tag{S-3}$$

Next, we introduce $(\mathcal{S}_{\mathbf{m}})_{\mathbf{m}\in\widetilde{\mathcal{M}}}$ *a random sub-collection of* $(\mathcal{S}_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$ *and consider the collection* $(\widehat{s}_{\mathbf{m}})_{\mathbf{m}\in\widetilde{\mathcal{M}}}$ *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 1, the* η'*-PMLE* $\widehat{s}_{\widehat{\mathbf{m}}}$, *defined in (10) on the subset* $\widetilde{\mathcal{M}}$ *instead of* $\mathcal{M}$, *satisfies*

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

Strategy for the proof of Theorem 1 We will briefly show how Theorem S-1 can be used to prove Theorem 1. All we need to do is check that Assumption 3 (K), Assumption 2 (Sep) and Assumption 1 (H) hold for every $\mathbf{m} \in \mathcal{M}$. According to the result of [3, Section 5.3], Assumption 2 (Sep) holds if we consider Gaussian densities, and the assumption defined by (S-3) is true if we further assume that the true conditional density s_0 is bounded and compactly supported. Furthermore, since we have restricted to a 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 3 (K) is satisfied. Therefore, the remaining most difficult step of the proof for Assumption 1 (H) is presented in Section S-2. All technical results are moved to Section S-3.

S-2 Proof of Theorem 1

Note that the definition of model complexity in Assumption 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 S-1, which is established in [2, Proposition 2].

Proposition S-1 *If we have*

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

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

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

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

Then, we claim that Proposition S-1 implies Assumption 1 (H) because of the fact that

$$\mathcal{H}_{[.],\mathrm{d}^{\otimes \mathrm{n}}}(\delta, \mathcal{S}_{\mathbf{m}}) \leq \dim(\mathcal{S}_{\mathbf{m}}) \left(C_{\mathbf{m}} + \ln\left(\frac{1}{\delta}\right)\right), \tag{S-4}$$

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

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

$$\mathcal{S}_{\mathbf{m}} = \Big\{ s_{\boldsymbol{\psi}_{\mathbf{m}}} \equiv s_{\boldsymbol{\psi}_K} \in \mathcal{S} : \boldsymbol{\psi}_{\mathbf{m}} = (\boldsymbol{\omega}_0, \boldsymbol{\omega}, \boldsymbol{\upsilon}_0, \boldsymbol{\Upsilon}, \boldsymbol{\Sigma}(\mathbf{B})) \in \boldsymbol{\Psi}_{\mathbf{m}},$$
$$\boldsymbol{\Psi}_{\mathbf{m}} = \mathbb{R}^K \times \mathbf{W}_J^{K \times D_W} \times \mathbb{R}^{K \times Q} \times \mathbf{V}_{J,\mathbf{R}}^{K \times D_V} \times \boldsymbol{\Omega}_{\mathbf{B}}^K \Big\}. \tag{S-5}$$

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 [13]. 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 proof of 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:

$$\mathbf{W}_{K,D_W} = \{0\} \otimes \mathbf{W}^{K-1}, \quad \mathbf{W} = \left\{ \mathcal{X} \ni \mathbf{x} \mapsto \sum_{\alpha \in \mathcal{A}} \boldsymbol{\omega}_\alpha \mathbf{x}^{\boldsymbol{\alpha}} \in \mathbb{R} : \max_{\boldsymbol{\alpha} \in \mathcal{A}} |\omega_{\boldsymbol{\alpha}}| \leq C_{\boldsymbol{\omega}} \right\}. \tag{S-6}$$

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.,* [6, Chapter 2], the cardinality of the set $\mathcal{A}$, denoted by $\operatorname{card}(\mathcal{A})$, 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. Note that we call a couple (X_p, Y_q) *irrelevant* if the elements $(\boldsymbol{\Upsilon}_{kd})_{qp} = 0$ 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:

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

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

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

Next (S-4) can be obtained by first decomposing the entropy term between the softmax gating functions and the Gaussian experts via Lemma S-1, which is immediately obtained from [13, Lemma 6], an extension of the results in [8, Theorem 2], [9], [2, Lemma 7] and [1].

Lemma S-1 *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

$$d_{\|\sup\|_\infty}\left(\left(\mathbf{s}_k\right)_{k\in[K]}, \left(\mathbf{t}_k\right)_{k\in[K]} \right) = \max_{k\in[K]} \sup_{\mathbf{x}\in\mathcal{X}} \left\| \mathbf{s}_k(\mathbf{x}) - \mathbf{t}_k(\mathbf{x}) \right\|_2, \tag{S-7}$$

for arbitrary K-tuples of the functions $\left(\mathbf{s}_k\right)_{k\in[K]}$ and $\left(\mathbf{t}_k\right)_{k\in[K]}$. 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 S-2, which is proved in Section S-3.1.

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

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

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 S-3 allows us to construct the Gaussian brackets to handle with the entropy metric for Gaussian experts, which is established in Section S-3.2.

Lemma S-3 *For all* $\delta \in (0, \sqrt{2}]$,

$$\mathcal{H}_{[\cdot],d_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}}}\left(\frac{\delta}{2}, \mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}\right) \leq \dim\left(\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}\right)\left(C_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}} + \ln\left(\frac{1}{\delta}\right)\right). \tag{S-9}$$

Finally, (S-4) is proved via Lemmas S-1 to S-3. 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

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

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

S-3 Lemma proofs

S-3.1 Proof of Lemma S-2

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

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

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

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

S-3.2 Proof of Lemma S-3

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 S-3 is directly implied from Proposition 2 and arguments from Appendix B.2.3 of [13]. 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 [12, 13, 3, 4, 5].

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

$$\begin{aligned}
\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})} &= \Big\{\mathcal{X}\times\mathcal{Y} \ni (\mathbf{x},\mathbf{y}) \mapsto \phi\left(\mathbf{y}; \mathbf{v}_{(D_V,J,\mathbf{R}_k)}(\mathbf{x}), \boldsymbol{\Sigma}_k(B_k)\right) : \\
&\qquad \mathbf{v}_{(D_V,J,\mathbf{R}_k)} \in \mathbf{V}_{(D_V,J,\mathbf{R}_k)}, \boldsymbol{\Sigma}_k(B_k) \in \boldsymbol{\Omega}_{B_k}\Big\}, \\
d^2_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}}\left(\phi_k^+, \phi_k^-\right) &= \mathbb{E}_{\mathbf{X}_{[N]}}\left[\frac{1}{N}\sum_{n=1}^N \mathrm{d}^2\left(\phi_k^+(\cdot|\mathbf{X}_n), \phi_k^-(\cdot|\mathbf{X}_n)\right)\right]. \qquad \text{(S-10)}
\end{aligned}$$

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 $\operatorname{card}\left(d_k^{[g]}\right)$ 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: $\boldsymbol{\Omega}_{\mathbf{B}}^{K} = (\boldsymbol{\Omega}_{B_k})_{k\in[K]}$, for every $k \in [K]$, for every $k \in [K]$,

$$\boldsymbol{\Omega}_{\mathbf{B}}^{K} = \left\{ \boldsymbol{\Sigma}_k\left(B_k\right) \in \mathcal{S}_Q^{++} \left| \begin{array}{l} \boldsymbol{\Sigma}_k\left(B_k\right) = \mathbf{P}_k \begin{pmatrix} \boldsymbol{\Sigma}_k^{[1]} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_k^{[2]} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{\Sigma}_k^{[G_k]} \end{pmatrix} \mathbf{P}_k^{-1}, \\ \boldsymbol{\Sigma}_k^{[g]} \in \mathcal{S}^{++}_{\mathrm{card}\left(d_k^{[g]}\right)}, \forall g \in [G_k] \end{array} \right\}. \tag{S-11}$$

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 S-4, established in S-3.2.

Lemma S-4 *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 S-3 is implied immediately via Lemma S-4 and the following important Lemma S-5, which is proved in S-3.2.

Lemma S-5 *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*

$$\mathcal{H}_{[\cdot],d_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}}}\left(\frac{\delta}{2},\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}\right) \leq \dim\left(\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}\right)\left(C_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}} + \ln\left(\frac{1}{\delta}\right)\right). \tag{S-12}$$

To this end, by combining the previous two Lemmas S-4 and S-5, 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,

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

Proof of Lemma S-4 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 (S-2), 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 S-4 is followed by the definition of a minimal $\delta/2$-bracket covering number for $\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}$.

Proof of Lemma S-5 We need to bound the bracketing entropy in (S-12). To do this, we need to construct an extension to the multidimensional Gaussian mixture of [8], 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 $\boldsymbol{\Sigma}_k(B_k) \in \boldsymbol{\Omega}_{B_k}, k \in [K]$, we denote by $\mathrm{Adj}\left(\boldsymbol{\Sigma}_k(B_k)\right)$ the adjacency matrix associated to the covariance matrix $\boldsymbol{\Sigma}_k(B_k)$. 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 [5] to handle the block-diagonal covariance matrices $\boldsymbol{\Sigma}_k\left(B_k\right)$, 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

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

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 [5] 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

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

$$\operatorname{card}\left(\tilde{S}_{B_k}^{\text{disc}}(\epsilon)\right) \leq \left(\left\lfloor\frac{2C_{\boldsymbol{\Sigma}}}{\epsilon}\right\rfloor\frac{Q\left(Q-1\right)}{2D_{B_k}}\right)^{D_{B_k}}, \tag{S-13}$$

$$D_{B_k}=\dim\left(\boldsymbol{\Omega}_{B_k}\right)=\sum_{g=1}^{G_k}\frac{\operatorname{card}\left(b_k^{(g)}\right)\left(\operatorname{card}\left(b_k^{(g)}\right)-1\right)}{2}, \tag{S-14}$$

where

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

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

$$\left\|\boldsymbol{\Sigma}_k\left(B_k\right)-\widetilde{\boldsymbol{\Sigma}}_k\left(B_k\right)\right\|_2^2\leq\epsilon^2. \tag{S-15}$$

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 $\delta_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}>0$, we claim that the set

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

is an $\delta_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}$-brackets set over $\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}$ where the constant $\alpha>0$ and function $\mathcal{X}\ni\mathbf{x}\mapsto\widetilde{\mathbf{v}}_{(D_V,J,\mathbf{R}_k)}\left(\mathbf{x}\right)$ and its corresponding space $G_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}\left(\delta_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}\right)$ will be specified later. Indeed, we consider any function $\mathcal{X}\times\mathcal{Y}\ni(\mathbf{x},\mathbf{y})\mapsto f(\mathbf{x},\mathbf{y})=\phi\left(\mathbf{y};\mathbf{v}_{(D_V,J,\mathbf{R}_k)}(\mathbf{x}),\boldsymbol{\Sigma}_k\left(B_k\right)\right)$ that belongs to $\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}$, where $\mathbf{v}_{(D_V,J,\mathbf{R}_k)}\in\mathbf{V}_{(D_V,J,\mathbf{R}_k)}$ and $\boldsymbol{\Sigma}_k\left(B_k\right)\in\boldsymbol{\Omega}_{B_k}$. According to (S-15), 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 $\delta_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}} > 0$, any $\mathbf{v}_{(D_V,J,\mathbf{R}_k)} \in \mathbf{V}_{(D_V,J,\mathbf{R}_k)}$, there exist a minimal covering of δ_k-bracket $G_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}\left(\delta_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}\right)$ 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

$$\sup_{\mathbf{x}\in\mathcal{X}} \left\|\widetilde{\mathbf{v}}_{(D_V,J,\mathbf{R}_k)}(\mathbf{x}) - \mathbf{v}_{(D_V,J,\mathbf{R}_k)}(\mathbf{x})\right\|_2^2 \leq \delta^2_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}, \tag{S-16}$$

$$\operatorname{card}\left(G_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}\left(\delta_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}\right)\right) \leq \left(\frac{\exp\left(C_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}\right)}{\delta_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}}\right)^{\dim\left(\mathbf{V}_{(D_V,J,\mathbf{R}_k)}\right)}. \tag{S-17}$$

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 (S-16) and (S-17) are satisfied. Note that for each $k \in [K], d \in [D_V]$, it holds that

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

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, (S-16) is immediately followed if we now choose $\widetilde{\boldsymbol{\upsilon}}_{k0}$ and $\widetilde{\boldsymbol{\Upsilon}}_{kd}^{R_{kd}}$ such that

$$\sqrt{Q}\left\|\boldsymbol{\upsilon}_{k0} - \widetilde{\boldsymbol{\upsilon}}_{k0}\right\|_\infty \leq \frac{\delta_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}}{2}, \tag{S-18}$$

$$\left\|\left\|\boldsymbol{\Upsilon}_{kd}^{R_{kd}} - \widetilde{\boldsymbol{\Upsilon}}_{kd}^{R_{kd}}\right\|\right\|_\infty \leq \frac{\delta_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}}{2D_V P\sqrt{Q}}. \tag{S-19}$$

Let us now see how to construct $\widetilde{\boldsymbol{\upsilon}}_{k0}$ to get (S-18). This task can be accomplished if for all $k \in [K]$, $q \in [Q]$, we set

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

Next, let us now see how to construct $\widetilde{\boldsymbol{\Upsilon}}_{kd}^{R_{kd}}$ to get (S-19). The boundedness assumption in (7) implies that

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

Therefore, (S-19) 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 [14, 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, (S-17) is proved by the following entropy controlling

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

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

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

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 S-6.

Lemma S-6 (Proposition C.1 from [12]) *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 $\mathbf{y}\in\mathbb{R}^Q$,*

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

Then, Lemma S-7 allows us to fulfill the assumptions of Lemma S-6.

Lemma S-7 (Similar to Lemma B.8 from [12]) *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 $\mathbf{y}\in\mathbb{R}^Q$,*

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

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

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

and

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

By using Lemma S-6 and the same argument as in the proof of Lemma B.9 from [12], given $0 < \epsilon < c_{\boldsymbol{\Sigma}}/3$, where ϵ is chosen later, and $\alpha = 3\epsilon/c_{\boldsymbol{\Sigma}}$, we obtain

$$\max\left\{\frac{l(\mathbf{x},\mathbf{y})}{f(\mathbf{x},\mathbf{y})}, \frac{f(\mathbf{x},\mathbf{y})}{u(\mathbf{x},\mathbf{y})}\right\} \leq \left(1+2\alpha\right)^{-\frac{Q}{2}}\exp\left(\frac{\left\|\mathbf{v}_{(D_V,J,\mathbf{R}_k)}(\mathbf{x}) - \widetilde{\mathbf{v}}_{(D_V,J,\mathbf{R}_k)}(\mathbf{x})\right\|_2^2}{2\epsilon}\right). \tag{S-20}$$

Because $\ln\left(\cdot\right)$ is a non-decreasing function, $\ln\left(1+2\alpha\right) \geq \alpha, \forall\alpha \in [0,1]$. Combined with (S-16) where $\delta^2_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}} = Q\alpha\epsilon$, we conclude that

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

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 S-8.

Lemma S-8 (Proposition C.3 from [12]) *Let $\phi\left(\cdot;\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1\right)$ and $\phi\left(\cdot;\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2\right)$ be two Gaussian densities with full rank covariance. It holds that*

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

Therefore, using the fact that $\cosh(t) = \frac{e^{-t}+e^t}{2}$, Lemma S-8 leads to, for all $\mathbf{x} \in \mathcal{X}$,

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

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 $\alpha = \frac{3\epsilon}{c_{\boldsymbol{\Sigma}}}, \epsilon = \frac{\delta c_{\boldsymbol{\Sigma}}}{6\sqrt{6}Q}$, $\forall \delta \in (0,1], Q \in \mathbb{N}^\star, c_{\boldsymbol{\Sigma}} > 0$, which appears in (S-20) and satisfies $\alpha = \frac{\delta}{2\sqrt{6}Q}$ and $0 < \epsilon < \frac{c_{\boldsymbol{\Sigma}}}{3}$. Indeed, studying functions g and h yields

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

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 $t = Q\ln(1+2\alpha)$, $Q \in \mathbb{N}^\star$, $\alpha = \frac{\delta}{2\sqrt{6}Q}$, $\delta \in (0,1]$. 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 $\tilde{S}^{\mathrm{disc}}_{B_k}(\epsilon)$ 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 $\epsilon = \frac{\delta c_{\boldsymbol{\Sigma}}}{6\sqrt{6Q}}$, $\alpha = \frac{3\epsilon}{c_{\boldsymbol{\Sigma}}} = \frac{\delta}{2\sqrt{6}Q}$, and $\delta^2_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}} = Q\alpha\epsilon = Q\frac{\delta}{2\sqrt{6}Q}\frac{\delta c_{\boldsymbol{\Sigma}}}{6\sqrt{6}Q} = \frac{\delta^2 c_{\boldsymbol{\Sigma}}}{72Q}$, it holds that

$$
\begin{aligned}
&\mathcal{N}_{[\cdot],d_{\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}}}\left(\frac{\delta}{2},\mathcal{G}_{(K,D_V,\mathbf{B},J,\mathbf{R})}\right)\\
&\leq \operatorname{card}\left(\tilde{S}^{\mathrm{disc}}_{B_k}(\epsilon)\right)\times \operatorname{card}\left(G_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}\left(\delta_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}\right)\right)\\
&\leq \left(\left\lfloor\frac{2C_{\boldsymbol{\Sigma}}}{\epsilon}\right\rfloor\frac{Q\left(Q-1\right)}{2D_{B_k}}\right)^{D_{B_k}}\left(\frac{\exp\left(C_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}\right)}{\delta_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}}\right)^{\dim\left(\mathbf{V}_{(D_V,J,\mathbf{R}_k)}\right)} \qquad \text{(using (S-14) and (S-17))}\\
&\leq \left(\frac{2C_{\boldsymbol{\Sigma}}6\sqrt{6}Q}{\delta c_{\boldsymbol{\Sigma}}}\frac{Q\left(Q-1\right)}{2D_{B_k}}\right)^{D_{B_k}}\left(\frac{6\sqrt{2Q}\exp\left(C_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}\right)}{\delta\sqrt{c_{\boldsymbol{\Sigma}}}}\right)^{\dim\left(\mathbf{V}_{(D_V,J,\mathbf{R}_k)}\right)}\\
&= \left(\frac{6\sqrt{6}C_{\boldsymbol{\Sigma}}Q^2\left(Q-1\right)}{c_{\boldsymbol{\Sigma}}D_{B_k}}\right)^{D_{B_k}}\left(\frac{6\sqrt{2Q}\exp\left(C_{\mathbf{V}_{(D_V,J,\mathbf{R}_k)}}\right)}{\sqrt{c_{\boldsymbol{\Sigma}}}}\right)^{\dim\left(\mathbf{V}_{(D_V,J,\mathbf{R}_k)}\right)}\left(\frac{1}{\delta}\right)^{D_{B_k}+\dim\left(\mathbf{V}_{(D_V,J,\mathbf{R}_k)}\right)}.
\end{aligned}
$$

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

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

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

Bibliography

[1] Cohen, S.X., Le Pennec, E.: Partition-based conditional density estimation. ESAIM: Probability and Statistics **17**, 672–697 (2013)

[2] Cohen, S., Le Pennec, E.: Conditional density estimation by penalized likelihood model selection and applications. Technical report, INRIA (2011)

[3] Devijver, E.: Finite mixture regression: a sparse variable selection by model selection for clustering. Electronic Journal of Statistics **9**(2), 2642–2674 (2015)

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

[5] Devijver, E., Gallopin, M.: Block-diagonal covariance selection for high-dimensional Gaussian graphical models. Journal of the American Statistical Association **113**(521), 306–314 (2018)

[6] Feller, W.: An introduction to probability theory and its applications, vol. 1. John Wiley (1957)

[7] Van de Geer, S.: Empirical Processes in M-estimation, vol. 6. Cambridge University Press (2000)

[8] Genovese, C.R., Wasserman, L.: Rates of convergence for the Gaussian mixture sieve. The Annals of Statistics **28**(4), 1105–1127 (2000)

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

[10] Kosorok, M.R.: Introduction to empirical processes and semiparametric inference. Springer Science & Business Media (2007)

[11] Massart, P.: Concentration Inequalities and Model Selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003. Springer (2007)

[12] Maugis, C., Michel, B.: A non asymptotic penalized criterion for Gaussian mixture model selection. ESAIM: Probability and Statistics **15**, 41–68 (2011)

[13] Montuelle, L., Le Pennec, E., et al.: Mixture of Gaussian regressions model with logistic weights, a penalized maximum likelihood approach. Electronic Journal of Statistics **8**(1), 1661–1695 (2014)

[14] Strang, G.: Linear algebra and learning from data. Wellesley-Cambridge Press Cambridge (2019)

[15] Van Der Vaart, A., Wellner, J.: Weak convergence and empirical processes: With applications to statistics. Springer **58**, 59 (1996)