HAL
open science

Bayesian nonparametric mixture of experts for high-dimensional inverse problems

Trungtin Nguyen, Florence Forbes, Julyan Arbel, Hien Duy Nguyen

To cite this version:

Trungtin Nguyen, Florence Forbes, Julyan Arbel, Hien Duy Nguyen. Bayesian nonparametric mixture of experts for high-dimensional inverse problems. 2023. hal-04015203v2

HAL Id: hal-04015203

https://hal.science/hal-04015203v2

Preprint submitted on 1 Nov 2023 (v2), last revised 17 Jun 2024 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Bayesian nonparametric mixture of experts for high-dimensional inverse problems

TrungTin Nguyen[a], Florence Forbes[a], Julyan Arbel[a], and Hien Duy Nguyen[b,c]

[a]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France; [b]Department of Mathematics and Physical Science, La Trobe University, Melbourne Australia; [c]Institute of Mathematics for Industry, Kyushu University, Fukuoka Japan

ARTICLE HISTORY
Compiled November 1, 2023

Word Limits

7115 words for the main text only.
11191 words for the main text and appendices.

Abstract
Large classes of problems can be formulated as inverse problems, where the goal is to find parameter values that best explain some observed measures. The relationship between parameters and observations is typically highly non-linear, with high-dimensional observations and correlated multidimensional parameters. To deal with these constraints via inverse regression strategies, we consider the Gaussian Local Linear Mapping (GLLiM) model, a special instance of mixture of expert models. We propose a general scheme to design a Bayesian nonparametric GLLiM model to avoid any commitment to an arbitrary number of experts. A tractable estimation algorithm is designed using variational Bayesian expectation-maximization. We establish posterior consistency for the number of mixture components after the merge-truncate-merge algorithm post-processing. Illustrations on simulated data show good results in terms of recovering the true number of experts and the mean regression function.

KEYWORDS
Bayesian nonparametrics; mixture of experts; high-dimensional statistics; inverse problems; Gaussian locally-linear mapping models; linear cluster-weighted models; variational inference; clustering; regression; model selection.

Contents

1 Introduction **4**

2 High-dimensional BNP-GLLiM model **6**
2.1 Inverse regression framework 6
2.2 High-dimensional regression via GLLiM model 7
2.3 Construction of BNP-GLLiM model 9
2.3.1 Bayesian nonparametric priors 9
2.3.2 Hierarchical representation of BNP-GLLiM model 11

3 Variational inference for BNP-GLLiM **11**
3.1 VB-E steps 13
3.2 VB-M steps 14
3.3 Evidence lower-bound (ELBO) 16

4 Predictive conditional density **17**
4.1 Joint density 17
4.2 Inverse conditional density 18
4.3 Forward conditional density 19

5 Nonparametric Bayesian model selection **19**
5.1 Posterior contraction rate in Bayesian infinite mixtures 19
5.2 Merge-Truncate-Merge (MTM) algorithm for BNP-GLLiM 21

6 Numerical experiments **23**
6.1 Data generation 23
6.2 Model selection, clustering and regression tasks via MTM-BNP-GLLiM 24

7 Perspectives **27**

A Details of VBEM for BNP-GLLiM model **33**
A.1 Variational Bayesian expectation-maximization principle 33
A.2 VB-E-$\boldsymbol{\tau}$ step from Section 3.1 35
A.3 VB-E-(α, σ) step from Section 3.1 36
A.4 VB-E-$\mathbf{Z}$ step from Section 3.1 38
A.4.1 Updating $\boldsymbol{\Sigma}_k$ 40
A.5 VB-M-$(\mathbf{c}, \boldsymbol{\Gamma})$ step from Section 3.2 40
A.6 VB-M-$(\mathbf{A}, \mathbf{b}, \boldsymbol{\Sigma})$ step from Section 3.2 41
A.6.1 Updating $\mathbf{b}_k$ 42
A.6.2 Updating $\mathbf{A}_k$ 42

B Technical proofs **45**
B.1 Proof of Lemma 5.2 45
B.2 Proof of Proposition 3.1 46
B.3 Proof of Theorem 4.1 50
B.4 Proof of Theorem 4.2 53
B.5 Proof of Theorem 4.3 53

C BNP-GLLiM2: a model with an hyperprior on the gating parameters **56**
C.1 VBEM for BNP-GLLiM2 56

C.1.1 VB-E-$\mathcal{Z}$ step 57
C.1.2 VB-E-$\boldsymbol{\theta}^*$ step 58
C.1.3 VB-M-$\boldsymbol{\rho}$ step 59
C.1.4 ELBO for BNP-GLLiM2 59
C.2 Predictive conditional density for BNP-GLLiM2 60
C.2.1 Joint density 60
C.2.2 Inverse conditional density 60
C.2.3 Forward conditional density 61
C.3 Proof of Proposition C.1 62
C.4 Proof of Theorem C.2 65
C.5 Proof of Theorem C.3 70
C.6 Proof of Theorem C.4 71

1. Introduction

Mixture of experts models (MoE). Many problems can be formulated as inverse problems, where the goal is to find parameter values that best explain observations. Typical constraints in practice are that the relationships between parameters and observations are highly non-linear, with high-dimensional observations and multi-dimensionally correlated parameters. To handle these constraints, we consider probabilistic mixtures of locally linear models, namely the Gaussian locally linear mapping (GLLiM) approach in [18], which includes special affine instances of the mixture of experts (MoE) model [94].

Note that these MoEs are generalizations of neural network architectures proposed by [42] and [44]. Further, these flexible models also generalize the classical mixture models (MM) and mixture of regression (MoR) models [54]. Their flexibility comes from the fact that they allow the mixture weights (or the gating functions) to depend on the explanatory variables, together with the component densities (or the experts). In the context of regression, MoE models with Gaussian experts and softmax or normalized Gaussian gating functions (as in GLLiM) are the most popular choices. These models are powerful tools for modelling complex nonlinear relationships between outputs (responses) and inputs (predictors) that arise from different subpopulations. The popularity of these conditional mixture density models is largely due to their universal approximation properties [64,66,67,78] as well as their good convergence rate, see, *e.g.,* [20,39, for MoR] and [43,59,60,61,62, for MoE]. It is worth noting that these results improve the approximation capabilities and convergence rates of unconditional MMs, as discussed in [31,37,38,72,75,77,82,87]. At a high level, universal approximation theorems state that given a large enough number of components, MM and MoE models can approximate a large class of unconditional and conditional probability density functions (cPDF), respectively, to any degree of accuracy. See, *e.g.,* [15,21,25,52,63,68,96], for further detailed reviews of practical and theoretical aspects of MoE models in statistics and in diverse domains such as natural language processing, computer vision, and speech recognition.

Model selection in MoE models. Although universal approximation allows us to conclude that, given a sufficient number of components, a finite MoE can approximate any other cPDF to an arbitrary degree of accuracy, it is not clear how to choose a large enough number of components for realistic problems. This motivates a careful study of interesting and important model selection problems for MM and MoE models, which have attracted much attention in statistics and machine learning over the last 50 years, see, *e.g.,* [12] for a recent comprehensive review.

When selecting the best data-driven number of components for MoE models, there are several approaches to controlling and accounting for model complexity. Typically, model selection is performed using an information criterion, such as the Akaike information criterion (AIC; [1]), the Bayesian information criterion (BIC; [86], BIC-GLLiM; [28,29]) or the BIC-like approximation of integrated classification likelihood (ICL-BIC; [7]). However, an important limitation of these criteria is that they are only asymptotically valid. This means that there are no finite sample guarantees when using AIC, ICL-BIC or BIC to choose between different levels of complexity. Therefore, their use in small sample settings is ad hoc.

To overcome such difficulties, and to partially support a recent approach called the slope heuristic ([8], also see [3] for a recent review), the authors of [70,71,73,74,76] es-

tablished non-asymptotic risk bounds in the form of weak oracle inequalities, provided that lower bounds on the penalties hold, in high-dimensional regression scenarios for a variety of MoE models, including GLLiM. Another approach, from [65], is based on the closed testing principle and leads to a sequential testing procedure that allows for confidence statements to be made regarding the order of a finite mixture model. These works lead to an optimal data-driven choice of the number of components in finite-sample settings. However, all previous approaches require that a range of models with different values be trained and compared, which can be a computational bottleneck in a high-dimensional framework.

Recently, [45] proposed computationally efficient variational inference approaches to architecture selection in high-dimensional deep Gaussian mixture models using overfitted mixtures, see, *e.g.,* [27,84], where unnecessary components are dropped in the estimation. However, in this work, we are interested in the more general context of the Bayesian nonparametric (BNP) approach, see, *e.g.,* [33,36], where it is not necessary to know the upper bound of the true number of components as in Bayesian overfitted MM. This is one motivation for the BNP priors that are under consideration here for GLLiM.

Dirichlet process mixture models (DP-MM) and Pitman-Yor process mixture models (PYP-MM) are among the most popular BNP models, particularly suitable for density estimation and probabilistic clustering. However, the posterior of DP-MM or PYP-MM is inconsistent for the number of components if the true number of components is finite and the concentration parameter is known and fixed, see, *e.g.,* [55]. This is because a BNP prior such as DP or PYP places zero probability on mixing measures with a finite number of supporting atoms. An interesting recent result in [4] is that consistency for the number of components can be achieved if we impose some assumptions on the prior for the concentration parameter of the DP-MM.

It appears that BNP-GLLiM tends to produce many small extraneous components around the true clusters in our numerical experiment Section 6. This makes it difficult to use them to infer the true number of components when this becomes a quantity of interest [34,51]. This encourages the use of a novel, simple MTM post-processing algorithm described in [35]. This post-processing consistently estimates the number of components for any general Bayesian prior, even without knowing its exact structure, as long as the posterior for that prior contracts to the true mixing distribution at a known rate.

Contributions. To deal with high-dimensional and large-scale data, our novel contribution is to propose a general scheme for designing a tractable BNP-GLLiM model. This scheme avoids any commitment to an arbitrary number of components. In particular, it allows a single training run to automatically select the optimal data-driven number of components with respect to the trade-off between model complexity and goodness-of-fit. BNP mixture models (BNP-MM), which boil down to special cases of the BNP-GLLiM model, have been extensively studied in the literature [2,16,22,23,48,50,51,57], but this does not cover extensions to MoE models such as our proposed BNP-GLLiM. This motivates us to design a tractable, theoretically guaranteed estimation algorithm for BNP-GLLiM using variational Bayesian expectation maximization (VBEM). See *e.g.,* [11] for a review of variational inference.

We establish theoretical properties such as posterior consistency for recovering the true number of components in BNP-GLLiM using the post-processing merge-truncate-merge (MTM) algorithm. Finally, our illustrations on simulated data show good results

in terms of recovering the true number of components and mean regression functions. It is worth emphasising that, for the first time, we provide evidence that MTM consistency holds not only for the MM results of [35], but also in the more general context of MoE models for cPDFs.

Notations. Throughout this paper, $\{1, \ldots, D\}$ is abbreviated as $[D]$ for $D \in \mathbb{N}^\star$, where $\mathbb{N}^\star$ denotes the positive natural numbers. The notation $\equiv$ refers to a definition. It is used to simplify the notation or expression. For a parametric model S, $\dim(S)$ refers to its dimension, *i.e.,* the total number of parameters to be estimated. Furthermore, $[\cdot;\cdot]$ denotes vertical vector concatenation. A coupling between $\boldsymbol{\pi} \equiv (\pi_k)_{k\in[K]}$ and $\boldsymbol{\pi}^0 \equiv (\pi_l^0)_{l\in[K_0]}$ is a joint distribution $\mathbf{Q}$ on $[K] \times [K_0]$, which is expressed as a matrix $\mathbf{Q} = (q_{kl})_{k\in[K],l\in[K_0]} \in [0,1]^{K\times K_0}$ with marginal probabilities $\sum_{k=1}^{K} q_{kl} = \pi_l^0$ and $\sum_{l=1}^{K_0} q_{kl} = \pi_k$ for any $k \in [K]$ and $l \in [K_0]$. We use $\mathcal{Q}(\boldsymbol{\pi}, \boldsymbol{\pi}^0)$ to denote the space of all such couplings. Regarding the space of mixing measures, let $\mathcal{E}_K \equiv \mathcal{E}_K(\boldsymbol{\Theta})$ and $\mathcal{O}_K \equiv \mathcal{O}_K(\boldsymbol{\Theta})$ respectively denote the space of all mixing measures with exactly and at most K support points, all in some parameter space $\boldsymbol{\Theta}$. Additionally, denote $\mathcal{G} \equiv \mathcal{G}(\boldsymbol{\Theta}) = \bigcup_{K\in\mathbb{N}_+} \mathcal{E}_K$ the set of all discrete measures with finite supports on $\boldsymbol{\Theta}$. Moreover, $\overline{\mathcal{G}}(\boldsymbol{\Theta})$ denotes the space of all discrete measures (including those with countably infinite supports) on $\boldsymbol{\Theta}$. Finally, $\mathcal{P}(\boldsymbol{\Theta})$ stands for the space of all probability measures on $\boldsymbol{\Theta}$. Throughout, we use the following colour rule for observations and parameters: observations are represented in green; latent, random or unknown parameters in red; and (fixed) hyperparameters in blue.

Outline. The paper is organized as follows. In Section 2, we first discuss how to construct the BNP-GLLiM model. A VBEM algorithm and the corresponding ELBO are described in Section 3, and predictive cPDFs in Section 4. Next, Section 5 shows how we can integrate the Merge-Truncate-Merge (MTM) post-processing procedure and prove consistency for the MTM output. This is useful to perform regression, clustering and model selection simultaneously. We experimentally evaluate our new results on simulated datasets in Section 6. Some perspectives are provided in Section 7. All details of VBEM for the BNP-GLLiM model and technical proofs not included in the main paper are relegated to Appendix A and Appendix B respectively. Appendix C presents a more general and tractable model with a hyperprior on the gating parameters, called BNP-GLLiM2.

2. High-dimensional BNP-GLLiM model

2.1. Inverse regression framework

We are interested in estimating the law of a low-dimensional random variable $\mathbf{X} = (\mathbf{X}_l)_{l\in[L]}$ conditionally on a high-dimensional $\mathbf{Y} = (\mathbf{Y}_d)_{d\in[D]}$, where typically $D \gg L$. We follow an inverse regression framework as in *e.g.* [18,47]. Therefore, in training, the low-dimensional variable $\mathbf{X}$ plays the role of the regressor, while the response $\mathbf{Y}$ is a function of $\mathbf{X}$, possibly corrupted by noise through inverse cPDF $p(\mathbf{Y} \mid \mathbf{X}; \boldsymbol{\psi})$, where $\boldsymbol{\psi}$ is an inverse parameter. The low dimension of the regressor $\mathbf{X}$ allows to drastically reduce the number of parameters to be estimated. In addition, the forward parameter

$\boldsymbol{\psi}^*$ and cPDF $p(\mathbf{X} \mid \mathbf{Y}; \boldsymbol{\psi}^*)$ are tractable after estimating the inverse parameter $\boldsymbol{\psi}$. Therefore, this density can be used to predict the low-dimensional response $\mathbf{x}$ of a high-dimensional test point $\mathbf{y}$. This inverse-then-forward regression strategy justifies the unconventional notation: $\mathbf{Y}$ for the high-dimensional input and $\mathbf{X}$ for the low-dimensional response.

Here and subsequently, we refer to the low-dimensional data sample as $\mathcal{X} \equiv \{\mathbf{x}_n\}_{n\in[N]} \subset \left(\mathbb{R}^L\right)^N$, the high-dimensional data sample as $\mathcal{Y} \equiv \{\mathbf{y}_n\}_{n\in[N]} \subset \left(\mathbb{R}^D\right)^N$. We further denote the realisations by $(\mathbf{x}, \mathbf{y})$, which are generated from two random variables $\mathbf{X}$ and $\mathbf{Y}$.

The following assumptions will be required for the training phase throughout the paper. We assume that the covariates $\mathbf{X}$ are independent but not necessarily identically distributed. The assumptions on the responses $\mathbf{Y}$ are stronger: conditional on $\mathcal{X}$, the $\mathcal{Y}$, are independent, and each $\mathbf{Y}$ follows a law with true (but unknown) inverse cPDF $s_0\left(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}\right)$, which is approximated via GLLiM. In Section 2.2 we first have a more detailed recall of GLLiM model. Then in Section 2.3 we construct the BNP-GLLiM model.

2.2. High-dimensional regression via GLLiM model

The GLLiM models, as originally introduced in [18], are used to capture the non-linear relationship between the response and the set of covariates from a high-dimensional heterogeneous data. More specifically, Deleforge et al. [18] overcame the difficulty of high-to-low regression by tackling the problem the other way round, *i.e.,* low-to-high. This means that the roles of input and response variables are swapped so that the low-dimensional variable $\mathbf{X}$ becomes the regressor as in Section 2.1. GLLiM then relies on a piecewise linear model in the following way. The high-dimensional response $\mathbf{Y}$ is approximated by the local affine mappings K:

$$\mathbf{Y} = \sum_{k=1}^{K} \mathbb{I}\left(Z = k\right)\left(\mathbf{A}_k \mathbf{X} + \mathbf{b}_k + \mathbf{E}_k\right). \tag{1}$$

Here $\mathbb{I}$ is an indicator function and Z is a latent variable that captures a cluster relationship, such that $Z = k$ if $\mathbf{Y}$ comes from cluster $k \in [K]$. Matrices $\mathbf{A}_k \in \mathbb{R}^{D\times L}$ and vectors $\mathbf{b}_k \in \mathbb{R}^D$ define cluster-specific affine transformations. In addition, $\mathbf{E}_k$ are error terms that capture both the reconstruction error due to the local affine approximations as well as the observation noise in $\mathbb{R}^D$.

Following the usual assumption that $\mathbf{E}_k$ is a zero-mean Gaussian variable with a covariance matrix $\boldsymbol{\Sigma}_k \in \mathbb{R}^{D\times D}$, it follows that

$$p\left(\mathbf{y} \mid \mathbf{x}, Z = k; \boldsymbol{\psi}\right) = \mathcal{N}_D\left(\mathbf{y} \mid \mathbf{A}_k \mathbf{x} + \mathbf{b}_k, \boldsymbol{\Sigma}_k\right), \tag{2}$$

where $\boldsymbol{\psi}$ is the vector of model parameters, $\mathcal{N}_D\left(\mathbf{y}; \mathbf{A}_k \mathbf{x} + \mathbf{b}_k, \boldsymbol{\Sigma}_k\right)$ is the Gaussian cPDF of dimension D. In order to enforce the affine transformations to be local, $\mathbf{X}$ is defined as a mixture of K Gaussian components as follows:

$$p\left(\mathbf{x} \mid Z = k; \boldsymbol{\psi}\right) = \mathcal{N}_L\left(\mathbf{x} \mid \mathbf{c}_k, \boldsymbol{\Gamma}_k\right), \quad p\left(Z = k; \boldsymbol{\psi}\right) = \pi_k, \tag{3}$$

where $\mathbf{c}_k \in \mathbb{R}^L$, $\mathbf{\Gamma}_k \in \mathbb{R}^{L\times L}$, and

$$\boldsymbol{\pi} = (\pi_k)_{k\in[K]} \in \left\{ (\pi_k)_{k\in[K]} \in \left(\mathbb{R}^+\right)^K, \sum_{k=1}^K \pi_k = 1 \right\} \equiv \mathbb{S}_{K-1},$$

where $\mathbb{S}_{K-1}$ is the $K-1$ dimensional probability simplex. Then, according to the formulas for conditional multivariate Gaussian variables and the hierarchical decomposition

$$\begin{aligned} p\left(\mathbf{y}, \mathbf{x}; \boldsymbol{\psi}\right) &= \sum_{k=1}^K p\left(\mathbf{y} \mid \mathbf{x}, Z = k; \boldsymbol{\psi}\right) p\left(\mathbf{x} \mid Z = k; \boldsymbol{\psi}\right) p\left(Z = k; \boldsymbol{\psi}\right), \\ &= \sum_{k=1}^K \pi_k \mathcal{N}_L\left(\mathbf{x} \mid \mathbf{c}_k, \mathbf{\Gamma}_k\right) \mathcal{N}_D\left(\mathbf{y} \mid \mathbf{A}_k\mathbf{x} + \mathbf{b}_k, \mathbf{\Sigma}_k\right), \end{aligned}$$

we obtain the following inverse conditional density,

$$p\left(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\psi}\right) = \sum_{k=1}^K \frac{\pi_k \mathcal{N}_L\left(\mathbf{x} \mid \mathbf{c}_k, \mathbf{\Gamma}_k\right)}{\sum_{l=1}^K \pi_l \mathcal{N}_L\left(\mathbf{x} \mid \mathbf{c}_l, \mathbf{\Gamma}_l\right)} \mathcal{N}_D\left(\mathbf{y} \mid \mathbf{A}_k\mathbf{x} + \mathbf{b}_k, \mathbf{\Sigma}_k\right), \tag{4}$$

where $\boldsymbol{\psi} = (\pi_k, \mathbf{c}_k, \mathbf{\Gamma}_k, \mathbf{A}_k, \mathbf{b}_k, \mathbf{\Sigma}_k)_{k\in[K]} \in \mathbf{\Psi}$.

Without assuming anything about the structure of the parameters, the dimension of GLLiM is given by

$$\dim\left(\mathbf{\Psi}\right) = K\left(1 + D(L+1) + \frac{D(D+1)}{2} + \frac{L(L+1)}{2} + L\right) - 1.$$

It is worth noting that $\dim\left(\mathbf{\Psi}\right)$ can be very large compared to the sample size (see, *e.g.,* [18] for real data sets) whenever D is large and $D \gg L$. Furthermore, under the assumption that the K transformations are affine, it is more realistic to make the assumption on the residual covariance matrices $\mathbf{\Sigma}_k$ of the error vectors $\mathbf{E}_k$ rather than on $\mathbf{\Gamma}_k$ (*cf.,* [18, Section 3]). This justifies the use of the inverse regression trick from [18], which leads to a drastic reduction in the number of parameters to be estimated. For instance, $\mathbf{E}_k$ can be modelled with equal isotropic Gaussian noise, so we have $\mathbf{\Sigma}_k = \tilde{\sigma}^2\,\mathrm{I}_D, \forall k \in [K]$, with some positive $\tilde{\sigma}^2$. The number of parameters to be estimated, *i.e.* the size of $\boldsymbol{\psi}$, is then $K + K\left(L + L(L+1)/2 + DL + D\right)$. For example, it is $30,060$ if $K = 10, L = 2, D = 1000$. However, if a high-to-low regression is estimated directly instead, the size of the parameter vector will be $K + K(D + LD + D(D+1)/2 + L)$, which is $5,035,030$ in the previous example.

More specifically, in (2), the roles of input and response variables should be reversed so that $\mathbf{Y}$ becomes the covariate and $\mathbf{X}$ plays the role of the multivariate response. Therefore, based on a similar argument to the previous hierarchical Gaussian MM (GMM), its corresponding forward conditional density from $\mathbb{R}^D$ to $\mathbb{R}^L$ is defined as follows

$$p\left(\mathbf{x} \mid \mathbf{y}; \boldsymbol{\psi}^*\right) = \sum_{k=1}^K \frac{\pi_k^* \mathcal{N}_D\left(\mathbf{y} \mid \mathbf{c}_k^*, \mathbf{\Gamma}_k^*\right)}{\sum_{l=1}^K \pi_l^* \mathcal{N}_D\left(\mathbf{y} \mid \mathbf{c}_l^*, \mathbf{\Gamma}_l^*\right)} \mathcal{N}_L\left(\mathbf{x} \mid \mathbf{A}_k^*\mathbf{y} + \mathbf{b}_k^*, \mathbf{\Sigma}_k^*\right). \tag{5}$$

A useful feature of GLLiM models is described in the following Lemma 2.1, established for multivariate Gaussian and Student components in [18,80] and which can be straightforwardly extended to Gaussian scale mixtures and elliptical distributions [40,74].

Lemma 2.1. *The parameter $\boldsymbol{\psi}^*$ in the inverse cPDF, defined in (5), can then be deduced from $\boldsymbol{\psi}$ in (4) via the following one-to-one correspondence:*

$$\begin{pmatrix} \pi_k \\ \mathbf{c}_k \\ \boldsymbol{\Gamma}_k \\ \mathbf{A}_k \\ \mathbf{b}_k \\ \boldsymbol{\Sigma}_k \end{pmatrix}_{k\in[K]} \mapsto \begin{pmatrix} \pi_k^* \\ \mathbf{c}_k^* \\ \boldsymbol{\Gamma}_k^* \\ \mathbf{A}_k^* \\ \mathbf{b}_k^* \\ \boldsymbol{\Sigma}_k^* \end{pmatrix}_{k\in[K]} = \begin{pmatrix} \pi_k \\ \mathbf{A}_k\mathbf{c}_k + \mathbf{b}_k \\ \boldsymbol{\Sigma}_k + \mathbf{A}_k\boldsymbol{\Gamma}_k\mathbf{A}_k^\top \\ \boldsymbol{\Sigma}_k^*\mathbf{A}_k^\top\boldsymbol{\Sigma}_k^{-1} \\ \boldsymbol{\Sigma}_k^*(\boldsymbol{\Gamma}_k^{-1}\mathbf{c}_k - \mathbf{A}_k^\top\boldsymbol{\Sigma}_k^{-1}\mathbf{b}_k) \\ \left(\boldsymbol{\Gamma}_k^{-1} + \mathbf{A}_k^\top\boldsymbol{\Sigma}_k^{-1}\mathbf{A}_k\right)^{-1} \end{pmatrix}_{k\in[K]} . \tag{6}$$

2.3. Construction of BNP-GLLiM model

2.3.1. Bayesian nonparametric priors

Stick-breaking construction of Dirichlet process. Note that the Dirichlet process (DP) [26] is a central BNP prior and is the infinite-dimensional generalization of the Dirichlet distribution. Therefore, for the sake of completeness, let us first recall the definition of the DP. A DP on the space $\mathcal{G}$ is defined as a random process characterized by a concentration parameter α and a base distribution G_0, denoted by $G \sim \mathrm{DP}(\alpha, G_0)$, such that for any finite partition $\{A_1, \ldots, A_p\}$ of $\mathcal{G}$, the random vector $(G(A_1), \ldots, G(A_p))$ is Dirichlet distributed:

$$(G(A_1), \ldots, G(A_p)) \sim \mathrm{Dir}(\alpha G_0(A_1), \ldots, \alpha G_0(A_p)) . \tag{7}$$

A DP prior G can be constructed by three methods: the Blackwell–Macqueen urn scheme, the Chinese restaurant method, and the stick-breaking construction. We will use the stick-breaking construction of DP (SBDP):

$$\boldsymbol{\theta}_k^0 \mid G_0 \overset{\text{iid}}{\sim} G_0, \quad k \in \mathbb{N}^\star, \tag{8}$$

$$\tau_k \mid \alpha \overset{\text{iid}}{\sim} \mathrm{Beta}(\tau_k \mid 1, \alpha), \quad k \in \mathbb{N}^\star, \tag{9}$$

$$\pi_k(\boldsymbol{\tau}) = \tau_k \prod_{l=1}^{k-1} (1 - \tau_l), \quad k \in \mathbb{N}^\star, \tag{10}$$

$$G = \sum_{k=1}^{\infty} \pi_k(\boldsymbol{\tau}) \delta_{\boldsymbol{\theta}_k^0} \sim \mathrm{DP}(\alpha, G_0) . \tag{11}$$

Pitman–Yor process. As a generalized version of the Dirichlet process, in the Pitman–Yor process (PYP) [81], the τ_k's are independent ($\overset{\text{ind}}{\sim}$) but not identically

distributed. More specifically,

$$\boldsymbol{\theta}_k^0 \mid G_0 \overset{\text{iid}}{\sim} G_0, \quad k \in \mathbb{N}^\star, \tag{12}$$

$$\tau_k \mid \alpha, \sigma \overset{\text{ind}}{\sim} \text{Beta}(1-\sigma, \alpha + k\sigma), \quad k \in \mathbb{N}^\star, \tag{13}$$

$$\pi_k(\boldsymbol{\tau}) = \tau_k \prod_{l=1}^{k-1} (1-\tau_l), \quad k \in \mathbb{N}^\star, \tag{14}$$

$$G = \sum_{k=1}^{\infty} \pi_k(\boldsymbol{\tau}) \delta_{\boldsymbol{\theta}_k^0} \sim \text{PYP}(\alpha, \sigma, G_0). \tag{15}$$

Here $\sigma \in (0,1)$ is a discount parameter and α is a concentration parameter $\alpha > \sigma$. The PYP is a two-parameter generalisation of the DP that allows one to control the tail behaviour when modelling data with either exponential or power-law tails [41,81]. The PYP reduces to a DP when $\sigma = 0$. More general stick-breaking representations are possible, *e.g.,* Gibbs-type priors [17,33] or homogeneous normalised random measures with independent increments [24]. The PYP has a power-law behaviour for the number of clusters. This can make it more suitable for a number of applications. In other words, the number of clusters grows as $\mathcal{O}(N^\sigma)$ for PYP, while growing more slowly as $\mathcal{O}(\log N)$ for DP.

Since the hyperparameters α and σ can have a significant effect on the growth of the number of clusters with data sample size, it is possible to specify priors for them. For the DP case obtained with $\sigma = 0$, it is suggested in [10] to use a gamma prior, $\alpha \sim \text{Gam}(s_1, s_2)$, where the hyperparameters s_1 and s_2 can be estimated or fixed. A natural question is whether one can also find a tractable prior for the discount parameter σ. Following the work of [50], we use the following prior that satisfies the constraints $\sigma \in (0,1)$ and $\alpha > -\sigma$,

$$p(\alpha, \sigma \mid s_1, s_2, a) = p(\alpha \mid \sigma; s_1, s_2)\, p(\sigma \mid a), \tag{16}$$

where $p(\alpha \mid \sigma; s_1, s_2)$ is a shifted gamma distribution $\mathcal{SG}(\alpha \mid \sigma; s_1, s_2)$ and $p(\sigma, a)$ is a distribution depending on some parameter a which is not specified at the moment but which can typically be assumed to be a uniform distribution on the interval $(0,1)$. Such a shifted gamma distribution is the distribution of a variable $U - \sigma$, where σ is considered fixed and U follows a gamma distribution $\text{Gam}(s_1, s_2)$. The PDF of this shifted gamma distribution is obtained from the standard gamma distribution as $\mathcal{SG}(\alpha \mid \sigma; s_1, s_2) = \text{Gam}(\alpha + \sigma \mid s_1, s_2)$.

2.3.2. Hierarchical representation of BNP-GLLiM model

We propose the following hierarchical representation to generate a data point $(\mathbf{y}_n, \mathbf{x}_n)$ within our BNP-GLLiM model:

1. BNP prior $G \sim$ **BNP** (α, σ, G_0)**:**

$$(\alpha, \sigma) \mid s_1, s_2, a \sim \mathcal{SG}\left(\alpha \mid \sigma; s_1, s_2\right) p(\sigma \mid a) \equiv \mathrm{Gam}\left(\alpha + \sigma \mid s_1, s_2\right) p(\sigma \mid a), \tag{17}$$

$$\tau_k \mid \alpha, \sigma \overset{\text{ind}}{\sim} \mathrm{Beta}\left(\tau_k \mid 1 - \sigma, \alpha + k\sigma\right), \quad k \in \mathbb{N}^\star, \tag{18}$$

$$\text{Define } \pi_k(\boldsymbol{\tau}) = \tau_k \prod_{l=1}^{k-1} (1 - \tau_l), \quad k \in \mathbb{N}^\star, \tag{19}$$

$$\boldsymbol{\theta}_k^\star \mid G_0 \overset{\text{iid}}{\sim} G_0, \quad \text{where } \boldsymbol{\theta}_k^\star \equiv (\mathbf{c}_k, \boldsymbol{\Gamma}_k, \mathbf{A}_k, \mathbf{b}_k, \boldsymbol{\Sigma}_k), \quad k \in \mathbb{N}^\star, \tag{20}$$

$$\text{and define } G = \sum_{k=1}^{\infty} \pi_k(\boldsymbol{\tau}) \delta_{\boldsymbol{\theta}_k^\star} \tag{21}$$

2. BNP-GLLiM model, for each $n \in [N]$, $\mathbf{y}_n \sim$ **BNP-GLLiM**(s_1, s_2, a, G)

$$\boldsymbol{\theta}_n \mid G \overset{\text{iid}}{\sim} G, \tag{22}$$

$$\text{If } \boldsymbol{\theta}_n = \boldsymbol{\theta}_k^\star, \text{ set } z_n = k, \tag{23}$$

$$\mathbf{x}_n \mid z_n, \mathbf{c}, \boldsymbol{\Gamma} \overset{\text{iid}}{\sim} \mathcal{N}_L\left(\mathbf{x}_n \mid \mathbf{c}_{z_n}, \boldsymbol{\Gamma}_{z_n}\right), \quad (\mathbf{c}, \boldsymbol{\Gamma}) \equiv (\mathbf{c}_k, \boldsymbol{\Gamma}_k)_{k \in \mathbb{N}^\star}, \tag{24}$$

$$\mathbf{y}_n \mid \mathbf{x}_n, z_n, \mathbf{A}, \mathbf{b}, \boldsymbol{\Sigma} \overset{\text{iid}}{\sim} \mathcal{N}_D\left(\mathbf{y}_n \mid \mathbf{A}_{z_n}\mathbf{x}_n + \mathbf{b}_{z_n}, \boldsymbol{\Sigma}_{z_n}\right), \quad (\mathbf{A}, \mathbf{b}, \boldsymbol{\Sigma}) \equiv (\mathbf{A}_k, \mathbf{b}_k, \boldsymbol{\Sigma}_k)_{k \in \mathbb{N}^\star}. \tag{25}$$

Hierarchical representation of BNP-MM. BNP-MM, including DP-MM and PYP-MM, see, *e.g.*, [22,50], has the following hierarchical representation to generate a data point $\mathbf{x}_n$ as a special case of BNP-GLLiM:

1. BNP prior: (26)

$$G = \sum_{k=1}^{\infty} \pi_k(\boldsymbol{\tau}) \delta_{\boldsymbol{\theta}_k^\star} \sim \mathrm{BNP}\left(\alpha, \sigma, G_0\right), \quad \boldsymbol{\theta}_k^\star = (\boldsymbol{\mu}_k, \mathbf{V}_k), \tag{27}$$

2. BNP-MM: for each $n \in [N]$,

$$\boldsymbol{\theta}_n \mid G \overset{\text{iid}}{\sim} G, \tag{28}$$

$$\text{If } \boldsymbol{\theta}_n = \boldsymbol{\theta}_k^\star, \text{ set } z_n = k, \tag{29}$$

$$\mathbf{x}_n \mid z_n, \boldsymbol{\theta}^\star \overset{\text{iid}}{\sim} \mathcal{N}_L\left(\mathbf{x}_n \mid \boldsymbol{\theta}_{z_n}^\star\right), \quad \boldsymbol{\theta}^\star \equiv (\boldsymbol{\theta}_k^\star)_{k \in [K]}. \tag{30}$$

3. Variational inference for BNP-GLLiM

A brief summary of the variational Bayesian expectation maximization (VBEM) and associated notations is provided in Appendix A.1. The set of hyperparameters is $\boldsymbol{\phi} = (s_1, s_2, a)$, while $\boldsymbol{\Theta} = (\boldsymbol{\tau}, \alpha, \sigma, \boldsymbol{\theta}^\star)$ are specific to the BNP part, namely to the stick-breaking construction of PYP. The joint distribution of the observed data $\mathcal{X}, \mathcal{Y}$ and

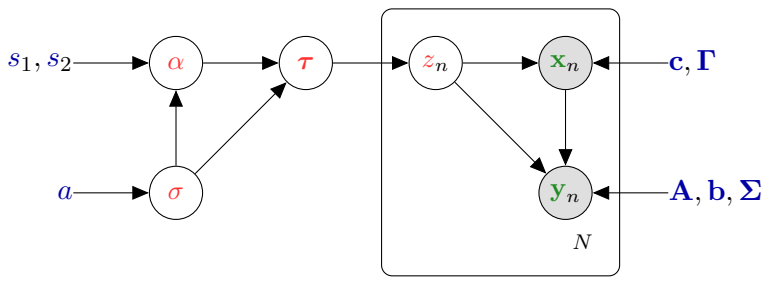


Figure 1. Graphical representation of BNP-GLLiM: the plate denotes N iid observations, white-filled circles correspond to unobserved (latent) variables and random or unknown parameters represented in red, while grey-filled circles correspond to observed variables represented in green. Hyperparameters are represented in blue.

of all the latent variables can be expressed in a hierarchical manner as follows

$$\begin{aligned} p(\mathcal{Y}, \mathcal{X}, \mathcal{Z}, \boldsymbol{\Theta}; \boldsymbol{\phi}) &= \prod_{n=1}^{N} p\left(\mathbf{y}_n \mid \mathbf{x}_n, z_n, \boldsymbol{\Theta}; \boldsymbol{\phi}\right) p\left(\mathbf{x}_n \mid z_n, \boldsymbol{\Theta}; \boldsymbol{\phi}\right) p(\mathcal{Z} \mid \boldsymbol{\Theta}; \boldsymbol{\phi}) p\left(\boldsymbol{\Theta}; \boldsymbol{\phi}\right) \\ &= \prod_{n=1}^{N} p\left(\mathbf{y}_n \mid \mathbf{x}_n, z_n; \mathbf{A}, \mathbf{b}, \boldsymbol{\Sigma}\right) p\left(\mathbf{x}_n \mid z_n, \mathbf{c}, \boldsymbol{\Gamma}\right) p(\mathcal{Z} \mid \boldsymbol{\tau}) \\ &\quad \prod_{k \in \mathbb{N}^\star} p\left(\tau_k \mid \alpha, \sigma\right) p\left(\alpha, \sigma \mid s_1, s_2, a\right) p(\boldsymbol{\theta}^\star | G_0). \end{aligned} \tag{31}$$

In most variational approximations, the posterior for the stick-breaking variables is approximated in a factorized form (mean-field approximation). Following the same approach, by factorizing the latent variables and the parameters, we choose the following variational distribution

$$q(\mathcal{Z}, \boldsymbol{\Theta}) = q_{\mathcal{Z}}(\mathcal{Z}) q_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}).$$

In particular, the intractable posterior on $\mathcal{Z}$ is approximated as $q_{\mathcal{Z}}(\mathcal{Z})$ that factorizes so as to handle intractability, namely

$$q_{\mathcal{Z}}(\mathcal{Z}) = \prod_{n=1}^{N} q_{z_n}\left(z_n\right). \tag{32}$$

Then the infinite state space for each z_j is dealt with by choosing a truncation of the state space to a maximum label $K \in \mathbb{N}^\star$, see, *e.g.*, [10,92]. In practice, this consists of assuming that the variational distributions q_{z_n} for $n \in [N]$, satisfy $q_{z_n}(k) = 0$ for $k > K$ and that the variational distribution on $\boldsymbol{\tau}$ also factorizes as $q_{\boldsymbol{\tau}}(\boldsymbol{\tau}) = \prod_{k=1}^{K-1} q_{\tau_k}\left(\tau_k\right)$ with the additional condition that $\tau_K = 1$. Thus the truncated variational posterior of parameters $\boldsymbol{\Theta}$ is given by

$$q_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}) = q_{\alpha,\sigma}(\alpha, \sigma) \prod_{k=1}^{K-1} q_{\tau_k}\left(\tau_k\right) \prod_{k=1}^{K} q_{\boldsymbol{\theta}_k^\star}(\boldsymbol{\theta}_k^\star). \tag{33}$$

In practice, for tractability reasons, we have to restrict to $p(\boldsymbol{\theta}_k^\star | G_0) \propto 1$ and $q_{\boldsymbol{\theta}_k^\star}(\boldsymbol{\theta}_k^\star)$ to Dirac distributions, $q_{\boldsymbol{\theta}_k^\star} = \delta_{\boldsymbol{\theta}_k^\star}$, which is equivalent to treating the $\boldsymbol{\theta}_k^\star$ as fixed unknown

hyperparameters, as illustrated graphically in Figure 1. These forms of $q_{\mathcal{Z}}$ and $q_{\boldsymbol{\Theta}}$ lead to our three VB-E steps and four VB-M steps, summarized below and with more detail in Appendix A. Set the initial value of $\boldsymbol{\phi}$ to $\boldsymbol{\phi}^{(0)}$. Then repeat the following steps iteratively. The iteration index is omitted in the update formulas for simplicity. Note that a more complex version with a normal-inverse-Wishart (NIW) distribution on the gating parameters $(\mathbf{c}_k, \boldsymbol{\Gamma}_k)$, referred to as BNP-GLLiM2, is presented in Appendix C.1.

3.1. VB-E steps

VB-E-$\boldsymbol{\tau}$ step. The VB-E-$\boldsymbol{\tau}$ step corresponds to a variational approximation in the exponential family case and results in a posterior from the same family as the prior, see, *e.g.,* [9, Section 10.4].

More precisely, to achieve this, we use (18), (23), (31), and are only interested in the functional dependence of the right-hand side of (A6) on the variable τ_k. Thus, any terms that do not depend on τ_k can be included in the additive normalization constant. Then, given for $k \in [K-1]$ that $N_k = \sum_{n=1}^{N} q_{z_n}(k)$ corresponds to the weight of the cluster k, see more details in Appendix A.2, it holds that

$$q_{\tau_k}(\tau_k) = \text{Beta}\left(\tau_k \mid \widehat{\gamma}_{k,1}, \widehat{\gamma}_{k,2}\right), \text{ where,} \tag{34}$$

$$\widehat{\gamma}_{k,1} = 1 - \mathbb{E}_{q_{\alpha,\sigma}}[\sigma] + N_k, \quad \widehat{\gamma}_{k,2} = \mathbb{E}_{q_{\alpha,\sigma}}[\alpha] + k\mathbb{E}_{q_{\alpha,\sigma}}[\sigma] + \sum_{l=k+1}^{K} N_l. \tag{35}$$

VB-E-$(\boldsymbol{\alpha}, \boldsymbol{\sigma})$ step. The (α, σ) variational posterior is more complex, but has a simple form in the DP case ($\sigma = 0$). Specifically, we have to compute

$$\widehat{s}_1 = s_1 + K - 1, \quad \widehat{s}_2 = s_2 - \sum_{k=1}^{K-1} \psi\left(\widehat{\gamma}_{k,2}\right) - \psi\left(\widehat{\gamma}_{k,1} + \widehat{\gamma}_{k,2}\right), \tag{36}$$

where $\psi(\cdot)$ is the digamma function defined by $\psi(\gamma) = \frac{d}{d\gamma}\log\Gamma(\gamma) = \frac{\Gamma'(\gamma)}{\Gamma(\gamma)}$.

When $\sigma = 0$ then $q_{\alpha,\sigma} \equiv q_{\alpha,0}$ is a gamma distribution $\text{Gam}\left(\widehat{s}_1, \widehat{s}_2\right)$ with $\mathbb{E}_{q_{\alpha,\sigma}}[\alpha] = \frac{\widehat{s}_1}{\widehat{s}_2}$. Otherwise (PYP case), $q_{\alpha,\sigma}$ is only identified up to a normalizing constant but the required $\mathbb{E}_{q_{\alpha,\sigma}}[\alpha]$ and $\mathbb{E}_{q_{\alpha,\sigma}}[\sigma]$ can be computed by importance sampling, see Appendix A.3 for more details.

We next consider the derivation of the update equation for the factor $q_{\mathcal{Z}}(\mathcal{Z})$.

VB-E-$\mathcal{Z}$ step. By using the mean-field approximation (32) and the truncation (see Appendix A.4), for all $n \in [N]$ and for all $k \in [K]$, this step consists in computing,

$$q_{z_n}(k) = \frac{\rho_{nk}}{\sum_{l=1}^{K} \rho_{nl}}. \tag{37}$$

Here, we define $\log \rho_{nk}$ by

$$
\begin{aligned}
&-\frac{1}{2}\Big\{ D\log(2\pi) + \log\left|\widehat{\boldsymbol{\Sigma}}_k\right| + (\mathbf{y}_n - \widehat{\mathbf{A}}_k\mathbf{x}_n - \widehat{\mathbf{b}}_k)^\top \widehat{\boldsymbol{\Sigma}}_k^{-1}(\mathbf{y}_n - \widehat{\mathbf{A}}_k\mathbf{x}_n - \widehat{\mathbf{b}}_k) \\
&\quad + L\log(2\pi) + \log\left|\widehat{\boldsymbol{\Gamma}}_k\right| + (\mathbf{x}_n - \widehat{\mathbf{c}}_k)^\top \widehat{\boldsymbol{\Gamma}}_k^{-1}(\mathbf{x}_n - \widehat{\mathbf{c}}_k) \Big\} \\
&\quad + \psi(\widehat{\gamma}_{k,1}) - \psi(\widehat{\gamma}_{k,1} + \widehat{\gamma}_{k,2}) + \sum_{l=1}^{k-1}\left[\psi(\widehat{\gamma}_{l,2}) - \psi(\widehat{\gamma}_{l,1} + \widehat{\gamma}_{l,2})\right].
\end{aligned} \tag{38}
$$

Note that in the above formula, symbols $\left(\widehat{\mathbf{c}}_k, \widehat{\boldsymbol{\Gamma}}_k, \widehat{\mathbf{A}}_k, \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right)$ are the hyperparameters more specifically defined in the following Section 3.2.

It is important to note that (37) provides assignment probabilities $q_{z_n}(k)$ rather than intermediate commitments to hard assignments of z_n. However, the hard assignments can be postponed to the end if desired to obtain a segmentation by the following MAP estimation:

$$
\widehat{z}_n = \operatorname*{argmax}_{k\in[K]} q_{z_n}(k). \tag{39}
$$

3.2. VB-M steps

The maximization step consists of updating the hyperparameters $\boldsymbol{\phi} = \left(s_1, s_2, a, \boldsymbol{\theta}^\star_{[K]}\right)$, where $\boldsymbol{\theta}^\star_{[K]} = (\mathbf{c}_k, \boldsymbol{\Gamma}_k, \mathbf{A}_k, \mathbf{b}_k, \boldsymbol{\Sigma}_k)_{k\in[K]}$, by maximizing the free energy, if they are not set heuristically:

$$
\boldsymbol{\phi}^{(r)} = \operatorname*{argmax}_{\boldsymbol{\phi}} \mathbb{E}_{q^{(r)}_{\mathcal{Z}} q^{(r)}_{\boldsymbol{\tau}} q^{(r)}_{\alpha,\sigma}}\left[\log p\left(\mathcal{Y}, \mathcal{X}, \mathcal{Z}, \boldsymbol{\tau}, \alpha, \sigma; \boldsymbol{\phi}\right)\right]. \tag{40}
$$

The VB-M-step can therefore be divided into 4 independent sub-steps, as listed below. From the conditional independence of (s_1, s_2, a) and $(\mathcal{Y}, \mathcal{X}, \mathcal{Z})$ given $(\boldsymbol{\tau}, \alpha, \sigma)$, the solution for the VB-M-(s_1, s_2) (in the DP case) step is straightforward. Only the M-(s_1, s_2, a) (in the PYP case) and M-$\boldsymbol{\theta}^\star_{[K]}$ steps are more involved.

VB-M-(s_1, s_2, a) step. This step is straightforward in the DP case ($\sigma = 0$). It can be expressed easily using the fact that both the prior and the variational posterior are Gamma distributions, and using the cross-entropy properties,

$$
(s_1, s_2)^{(r)} = \operatorname*{argmax}_{(s_1,s_2)} \mathbb{E}_{q^{(r)}_{\alpha,0}}\left[\log p(\alpha \mid s_1, s_2)\right] = (\widehat{s}_1^{(r)}, \widehat{s}_2^{(r)}), \tag{41}
$$

where $(\widehat{s}_1^{(r)}, \widehat{s}_2^{(r)})$ is given in (36). We can also solve this step numerically using importance sampling in the more general case of PYP ($\sigma \neq 0$). For more details, see [50, Appendix A.7].

VB-M-($\mathbf{c}, \mathbf{\Gamma}$) step. This step divides into K sub-steps that involve the following optimizations

$$\left(\widehat{\mathbf{c}}_k, \widehat{\mathbf{\Gamma}}_k\right) \equiv \left(\widehat{\mathbf{c}}_k^{(r)}, \widehat{\mathbf{\Gamma}}_k^{(r)}\right) = \underset{(\mathbf{c}_k, \mathbf{\Gamma}_k)}{\operatorname{argmax}} \ \mathbb{E}_{q_{\mathcal{Z}}^{(r)}} \left[\log p\left(\mathcal{X} \mid \mathcal{Z}; \mathbf{c}_k, \mathbf{\Gamma}_k\right)\right]. \tag{42}$$

We can then update the Gaussian gating parameters as follows:

$$\widehat{\mathbf{c}}_k = \frac{1}{N_k} \sum_{n=1}^{N} q_{z_n}(k) \mathbf{x}_n, \tag{43}$$

$$\widehat{\mathbf{\Gamma}}_k = \frac{1}{N_k} \sum_{n=1}^{N} q_{z_n}(k) \left(\mathbf{x}_n - \widehat{\mathbf{c}}_k\right) \left(\mathbf{x}_n - \widehat{\mathbf{c}}_k\right)^\top. \tag{44}$$

The technical details will be left to the Appendix A.5.

VB-M-($\mathbf{A}, \mathbf{b}, \mathbf{\Sigma}$) step. Using the same idea, this step is divided into K sub-steps, which include the following optimization problems

$$\left(\widehat{\mathbf{A}}_k, \widehat{\mathbf{b}}_k, \widehat{\mathbf{\Sigma}}_k\right) \equiv \left(\widehat{\mathbf{A}}_k^{(r)}, \widehat{\mathbf{b}}_k^{(r)}, \widehat{\mathbf{\Sigma}}_k^{(r)}\right) = \underset{(\mathbf{A}_k, \mathbf{b}_k, \mathbf{\Sigma}_k)}{\operatorname{argmax}} \ \mathbb{E}_{q_{\mathcal{Z}}^{(r)}} \left[\log p\left(\mathcal{Y} \mid \mathcal{X}, \mathcal{Z}; \mathbf{A}_k, \mathbf{b}_k, \mathbf{\Sigma}_k\right)\right]. \tag{45}$$

It requires to define the following quantities,

$$\begin{aligned}
\bar{\mathbf{x}}_k &= \frac{1}{N_k} \sum_{n=1}^{N} q_{z_n}(k) \ \mathbf{x}_n, \\
\bar{\mathbf{y}}_k &= \frac{1}{N_k} \sum_{n=1}^{N} q_{z_n}(k) \ \mathbf{y}_n, \\
\boldsymbol{X}_k &= \frac{1}{\sqrt{N_k}} \left(\sqrt{q_{z_1}(k)}(\mathbf{x}_1 - \bar{\mathbf{x}}_k), \ldots, \sqrt{q_{z_N}(k)}(\mathbf{x}_N - \bar{\mathbf{x}}_k)\right), \\
\boldsymbol{Y}_k &= \frac{1}{\sqrt{N_k}} \left(\sqrt{q_{z_1}(k)}(\mathbf{y}_1 - \bar{\mathbf{y}}_k), \ldots, \sqrt{q_{z_N}(k)}(\mathbf{y}_N - \bar{\mathbf{y}}_k)\right).
\end{aligned}$$

Then we can update the parameters for the Gaussian experts, see Appendix A.6, for more details, as follows:

$$\begin{aligned}
\widehat{\mathbf{A}}_k &= \boldsymbol{Y}_k \boldsymbol{X}_k^\top (\boldsymbol{X}_k \boldsymbol{X}_k^\top)^{-1}, \\
\widehat{\mathbf{b}}_k &= \frac{1}{N_k} \sum_{n=1}^{N} q_{z_n}(k) \ (\mathbf{y}_n - \widehat{\mathbf{A}}_k \mathbf{x}_n), \\
\widehat{\mathbf{\Sigma}}_k &= \frac{1}{N_k} \sum_{n=1}^{N} q_{z_n}(k) \ (\mathbf{y}_n - \widehat{\mathbf{A}}_k \mathbf{x}_n - \widehat{\mathbf{b}}_k)(\mathbf{y}_n - \widehat{\mathbf{A}}_k \mathbf{x}_n - \widehat{\mathbf{b}}_k)^\top.
\end{aligned}$$

3.3. Evidence lower-bound (ELBO)

Evaluating the ELBO in (A4) allows us not only to monitor the bound during the re-estimation to test for convergence but also to check both the mathematical expressions for the solutions and their software implementation. Indeed, the value of this bound (A4) at each step of the iterative re-estimation procedure should not decrease [88], in particular, see recent results for Bayesian nonparametric mixture models in [22, Appendix A].

Recall that $\widehat{\boldsymbol{\phi}} = \left(\widehat{s}_1, \widehat{s}_2, \widehat{a}, \left(\widehat{\mathbf{c}}_k, \widehat{\boldsymbol{\Gamma}}_k, \widehat{\mathbf{A}}_k, \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right)_{k\in\mathbb{N}^\star}\right)$. Here, in order to keep the notation uncluttered, we will sometimes omit the subscripts on the expectation operators because each expectation is taken with respect to all of the random variables in its argument, and the hat superscript $\widehat{\cdot}$ on the hyperparameters $\widehat{\boldsymbol{\phi}}$ of q distribution.

If $\sigma \neq 0$ and there is enough training data, the ELBO can be evaluated via the fact that the integral reduces to a point evaluation at the posterior mean of each parameter, see, *e.g.*, [49,69,83,93,95]. When $\sigma = 0$, we can analytically compute the ELBO in the BNP-GLLiM via Proposition 3.1 which is proved in Appendix B.2.

Proposition 3.1. *When $\sigma = 0$, the ELBO in the BNP-GLLiM is analytically derived as follows:*

$$\mathcal{F}\left(q_{\mathcal{Z}}, q_{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\phi}}\right) = \mathbb{E}\left[\log p(\mathcal{Y} \mid \mathcal{X}, \mathcal{Z}, \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] + \mathbb{E}\left[\log p(\mathcal{X} \mid \mathcal{Z}, \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] + \mathbb{E}\left[\log p(\mathcal{Z} \mid \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] + \mathbb{E}\left[\log p(\boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] - \mathbb{E}\left[\log q(\mathcal{Z})\right] - \mathbb{E}\left[\log q(\boldsymbol{\Theta})\right]. \tag{46}$$

Here, we have the following update formulas:

$$\mathbb{E}\left[\log p(\mathcal{Y} \mid \mathcal{X}, \mathcal{Z}, \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] = \sum_{n=1}^{N}\sum_{k=1}^{K} q_{z_n}(k) \log \mathcal{N}_D\left(\mathbf{y}_n \mid \widehat{\mathbf{A}}_k \mathbf{x}_n + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right) \tag{47}$$

$$\mathbb{E}\left[\log p(\mathcal{X} \mid \mathcal{Z}, \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] = \sum_{n=1}^{N}\sum_{k=1}^{K} q_{z_n}(k) \log \mathcal{N}_L\left(\mathbf{x}_n \mid \widehat{\boldsymbol{c}}_k, \widehat{\boldsymbol{\Gamma}}_k\right), \tag{48}$$

$$\mathbb{E}\left[\log p(\mathcal{Z} \mid \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] = \sum_{k=1}^{K} N_k \left[\psi\left(\widehat{\gamma}_{k,1}\right) - \psi\left(\widehat{\gamma}_{k,1} + \widehat{\gamma}_{k,2}\right) + \sum_{l=1}^{k-1}\left[\psi\left(\widehat{\gamma}_{l,2}\right) - \psi\left(\widehat{\gamma}_{l,1} + \widehat{\gamma}_{l,2}\right)\right]\right], \tag{49}$$

$$\mathbb{E}\left[\log p(\boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] = \sum_{k=1}^{K-1} \mathbb{E}\left[\log p\left(\tau_k \mid \alpha\right)\right] + \mathbb{E}\left[\log p\left(\alpha \mid \widehat{s}_1, \widehat{s}_2\right)\right], \tag{50}$$

$$\mathbb{E}\left[\log p\left(\tau_k \mid \alpha\right)\right] = \frac{\widehat{s}_1 - \widehat{s}_2}{\widehat{s}_2}\left[\psi\left(\widehat{\gamma}_{k,2}\right) - \psi\left(\widehat{\gamma}_{k,1} + \widehat{\gamma}_{k,2}\right)\right] + \psi\left(\widehat{s}_1\right) - \log(\widehat{s}_2),$$

$$\mathbb{E}\left[\log p\left(\alpha \mid \widehat{s}_1, \widehat{s}_2\right)\right] = -\log\Gamma\left(\widehat{s}_1\right) + \left(\widehat{s}_1 - 1\right)\psi\left(\widehat{s}_1\right) + \log\left(\widehat{s}_2\right) - \widehat{s}_1,$$

$$\mathbb{E}\left[\log q(\mathcal{Z})\right] = \sum_{n=1}^{N}\sum_{k=1}^{K} q_{z_n}(k)\log q_{z_n}(k)\,, \tag{51}$$

$$\mathbb{E}\left[\log q(\boldsymbol{\Theta})\right] = \mathbb{E}\left[\log q_{\alpha,0}(\alpha)\right] + \sum_{k=1}^{K-1}\mathbb{E}\left[\log q_{\tau_k}(\tau_k)\right], \tag{52}$$

$$\mathbb{E}\left[\log q_{\alpha,0}(\alpha)\right] = -\log\Gamma\left(\widehat{s}_1\right) + \left(\widehat{s}_1 - 1\right)\psi\left(\widehat{s}_1\right) + \log(\widehat{s}_2) - \widehat{s}_1, \tag{53}$$

$$\mathbb{E}\left[\log q_{\tau_k}(\tau_k)\right] = \sum_{l=1}^{2}\left(\widehat{\gamma}_{k,l} - 1\right)\left\{\psi\left(\widehat{\gamma}_{k,l}\right) - \psi\left(\widehat{\gamma}_{k,1} + \widehat{\gamma}_{k,2}\right)\right\} + \log\frac{\Gamma\left(\widehat{\gamma}_{k,1} + \widehat{\gamma}_{k,2}\right)}{\Gamma\left(\widehat{\gamma}_{k,1}\right)\Gamma\left(\widehat{\gamma}_{k,2}\right)}. \tag{54}$$

Note that if the free energy is computed at the end of each VBEM iteration, as in Section 3.2, we have $\mathbb{E}\left[\log q_{\alpha,0}(\alpha)\right] = \mathbb{E}\left[\log p\left(\alpha \mid \widehat{s}_1, \widehat{s}_2\right)\right]$.

4. Predictive conditional density

The most popular uses of BNP-GLLiM with discrete random probability measures, such as the one displayed in (25), relate to conditional density estimation and data clustering. More precisely, we will be interested in the predictive conditional density for a new value $(\widehat{\mathbf{y}}, \widehat{\mathbf{x}})$ of the observed variables. Note that associated with these observations will be a corresponding latent variable $\widehat{\mathbf{z}}$.

If $\sigma \neq 0$, we can use the following remark. Suppose there is enough training data. Then the posterior distribution of all parameters is usually highly peaked. This leads to the second approximation, where the integral reduces to a point evaluation at the posterior mean of each parameter, see, *e.g.,* [49,69,83,93,95].

When $\sigma = 0$, we can analytically approximate such densities via several following lemmas. In the following Theorems 4.1 to 4.3, the notation "$\approx$" means that we approximate the desired densities of the BNP-GLLiM by a mixture of Gaussians using factorized variational approximation posteriors and a truncation of the number of components.

4.1. Joint density

We first show how to compute the joint density $p(\widehat{\mathbf{y}}, \widehat{\mathbf{x}}, \mathcal{X}, \mathcal{Y})$ via Theorem 4.1, which is proved in Appendix B.3.

Theorem 4.1. *Via defining* $\widehat{\mathbf{w}} \equiv [\widehat{\mathbf{x}}; \widehat{\mathbf{y}}]$*, we obtain*

$$\boldsymbol{\mu}_k \equiv \mathbb{E}\left[\widehat{\mathbf{w}}\right] = \begin{pmatrix} \widehat{\mathbf{c}}_k \\ \widehat{\mathbf{A}}_k\widehat{\mathbf{c}}_k + \widehat{\mathbf{b}}_k \end{pmatrix}, \quad \mathbf{V}_k = cov\left[\widehat{\mathbf{w}}\right] = \begin{pmatrix} \widehat{\boldsymbol{\Gamma}}_k & \widehat{\boldsymbol{\Gamma}}_k\widehat{\mathbf{A}}_k^\top \\ \widehat{\mathbf{A}}_k\widehat{\boldsymbol{\Gamma}}_k & \widehat{\boldsymbol{\Sigma}}_k + \widehat{\mathbf{A}}_k\widehat{\boldsymbol{\Gamma}}_k\widehat{\mathbf{A}}_k^\top \end{pmatrix}, \tag{55}$$

and

$$p\left(\widehat{\mathbf{y}}, \widehat{\mathbf{x}}, \mathcal{X}, \mathcal{Y}\right) \approx \sum_{k=1}^{K} \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k(\boldsymbol{\tau})\right] \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{c}}_k, \widehat{\boldsymbol{\Gamma}}_k\right) \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{A}}_k \widehat{\mathbf{x}} + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right)$$
$$\equiv \sum_{k=1}^{K} \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k\left(\boldsymbol{\tau}\right)\right] \mathcal{N}_{L+D}\left(\widehat{\mathbf{w}} \mid \boldsymbol{\mu}_k, \mathbf{V}_k\right). \tag{56}$$

Furthermore, it holds that

$$\mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k\left(\boldsymbol{\tau}\right)\right] = \mathbb{E}_{q_{\tau_k}}\left[\tau_k\right] \prod_{l=1}^{k-1} \mathbb{E}_{q_{\tau_l}}\left[1-\tau_l\right]. \tag{57}$$

Here, we have for any $k \in [K]$,

$$\mathbb{E}_{q_{\tau_k}}\left[\tau_k\right] = \frac{\widehat{\gamma}_{k,1}}{\widehat{\gamma}_{k,1}+\widehat{\gamma}_{k,2}}, \quad \mathbb{E}_{q_{\tau_k}}\left[1-\tau_k\right] = 1 - \mathbb{E}_{q_{\tau_k}}\left[\tau_k\right] = \frac{\widehat{\gamma}_{k,2}}{\widehat{\gamma}_{k,1}+\widehat{\gamma}_{k,2}},$$
$$\widehat{\gamma}_{k,1} = 1 - \mathbb{E}_{q_{\alpha,\sigma}}\left[\sigma\right] + N_k, \quad \widehat{\gamma}_{k,2} = \mathbb{E}_{q_{\alpha,\sigma}}\left[\alpha\right] + k\mathbb{E}_{q_{\alpha,\sigma}}\left[\sigma\right] + \sum_{l=k+1}^{K} N_l, N_k = \sum_{n=1}^{N} q_{z_n}(k).$$

4.2. Inverse conditional density

We then show how to approximate the inverse conditional density $p(\widehat{\mathbf{y}} \mid \widehat{\mathbf{x}}, \mathcal{X}, \mathcal{Y})$. This predictive density in BNP-GLLiM is approximated by a GLLiM via Theorem 4.2 with the proof in Appendix B.4.

Theorem 4.2. *We approximate the inverse conditional density of the BNP-GLLiM as follows:*

$$p\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{x}}, \mathcal{X}, \mathcal{Y}\right) \approx \sum_{k=1}^{K} g_k\left(\widehat{\mathbf{x}} \mid \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\phi}}, \mathcal{X}, \mathcal{Y}\right) \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{A}}_k \widehat{\mathbf{x}} + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right). \tag{58}$$

Here, the gating posteriors belong to a $K-1$ dimensional probability simplex and are given by

$$g_k\left(\widehat{\mathbf{x}} \mid \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\phi}}, \mathcal{X}, \mathcal{Y}\right) = \frac{\mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k\left(\boldsymbol{\tau}\right)\right] \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{c}}_k, \widehat{\boldsymbol{\Gamma}}_k\right)}{\sum_{l=1}^{K} \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_l\left(\boldsymbol{\tau}\right)\right] \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{c}}_l, \widehat{\boldsymbol{\Gamma}}_l\right)}, \quad k \in [K], \tag{59}$$

For prediction, we also have

$$\mathbb{E}\left[\widehat{\mathbf{y}} \mid \widehat{\mathbf{x}}, \mathcal{X}, \mathcal{Y}\right] \approx \sum_{k=1}^{K} g_k\left(\widehat{\mathbf{x}} \mid \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\phi}}, \mathcal{X}, \mathcal{Y}\right) \left[\widehat{\mathbf{A}}_k \widehat{\mathbf{x}} + \widehat{\mathbf{b}}_k\right]. \tag{60}$$

4.3. Forward conditional density

To deal with high-dimensional regression data, namely high-to-low regression, given the inverse conditional density $p(\widehat{\mathbf{y}} \mid \widehat{\mathbf{x}}, \mathcal{X}, \mathcal{Y})$, we want to approximate the following forward conditional density via Theorem 4.3, whose proof is provided in Appendix B.5.

Theorem 4.3. *We approximate the forward conditional density of the BNP-GLLiM by*

$$p\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{y}}, \mathcal{X}, \mathcal{Y}\right) \approx \sum_{k=1}^{K} g_k\left(\widehat{\mathbf{y}} \mid \widehat{\boldsymbol{\Theta}}^*, \widehat{\boldsymbol{\phi}}^*, \mathcal{X}, \mathcal{Y}\right) \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{A}}_k^* \widehat{\mathbf{y}} + \widehat{\mathbf{b}}_k^*, \widehat{\boldsymbol{\Sigma}}_k^*\right), \tag{61}$$

which is again a GLLiM. Here, we defined

$$\begin{aligned}
g_k\left(\widehat{\mathbf{y}} \mid \widehat{\boldsymbol{\Theta}}^*, \widehat{\boldsymbol{\phi}}^*, \mathcal{X}, \mathcal{Y}\right) &= \frac{\mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k\left(\boldsymbol{\tau}\right)\right] \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{c}}_k^*, \widehat{\boldsymbol{\Gamma}}_k^*\right)}{\sum_{l=1}^{K} \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_l\left(\boldsymbol{\tau}\right)\right] \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{c}}_l^*, \widehat{\boldsymbol{\Gamma}}_l^*\right)}, \\
\widehat{\boldsymbol{\Sigma}}_k^* &= \left(\widehat{\boldsymbol{\Gamma}}_k^{-1} + \widehat{\mathbf{A}}_k^\top \widehat{\boldsymbol{\Sigma}}_k^{-1} \widehat{\mathbf{A}}_k\right)^{-1}, \\
\widehat{\mathbf{A}}_k^* &= \widehat{\boldsymbol{\Sigma}}_k^* \widehat{\mathbf{A}}_k^\top \widehat{\boldsymbol{\Sigma}}_k^{-1}, \\
\widehat{\mathbf{b}}_k^* &= \widehat{\boldsymbol{\Sigma}}_k^* \left[\widehat{\boldsymbol{\Gamma}}_k^{-1} \widehat{\mathbf{c}}_k - \widehat{\mathbf{A}}_k^\top \widehat{\boldsymbol{\Sigma}}_k^{-1} \widehat{\mathbf{b}}_k\right], \\
\widehat{\mathbf{c}}_k^* &= \widehat{\mathbf{A}}_k \widehat{\mathbf{c}}_k + \widehat{\mathbf{b}}_k, \\
\widehat{\boldsymbol{\Gamma}}_k^* &= \widehat{\boldsymbol{\Sigma}}_k + \widehat{\mathbf{A}}_k \widehat{\boldsymbol{\Gamma}}_k \widehat{\mathbf{A}}_k^\top.
\end{aligned} \tag{62}$$

Furthermore, we approximate the expectation and covariance matrix of $\widehat{\mathbf{x}} \mid \widehat{\mathbf{y}}, \mathcal{X}, \mathcal{Y}$ as follows:

$$\begin{aligned}
\mathbb{E}\left[\widehat{\mathbf{x}} \mid \widehat{\mathbf{y}}, \mathcal{X}, \mathcal{Y}\right] &\approx \sum_{k=1}^{K} g_k\left(\widehat{\mathbf{y}} \mid \widehat{\boldsymbol{\Theta}}^*, \widehat{\boldsymbol{\phi}}^*, \mathcal{X}, \mathcal{Y}\right)\left(\widehat{\mathbf{A}}_k^* \widehat{\mathbf{y}} + \widehat{\mathbf{b}}_k^*\right), \\
var\left[\widehat{\mathbf{x}} \mid \widehat{\mathbf{y}}, \mathcal{X}, \mathcal{Y}\right] &\approx \sum_{k=1}^{K} g_k\left(\widehat{\mathbf{y}} \mid \widehat{\boldsymbol{\Theta}}^*, \widehat{\boldsymbol{\phi}}^*, \mathcal{X}, \mathcal{Y}\right)\left[\widehat{\boldsymbol{\Sigma}}_k^* + \left(\widehat{\mathbf{A}}_k^* \widehat{\mathbf{y}} + \widehat{\mathbf{b}}_k^*\right)\left(\widehat{\mathbf{A}}_k^* \widehat{\mathbf{y}} + \widehat{\mathbf{b}}_k^*\right)^\top\right] \\
&\quad - \mathbb{E}\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{y}}, \mathcal{X}, \mathcal{Y}\right) \mathbb{E}\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{y}}, \mathcal{X}, \mathcal{Y}\right)^\top.
\end{aligned} \tag{63}$$

5. Nonparametric Bayesian model selection

5.1. Posterior contraction rate in Bayesian infinite mixtures

Problem setup. We first recall the GMM where we have iid samples $(\mathbf{W}_n)_{n\in[N]} \equiv \mathcal{W}$ coming from a true but unknown distribution P_{G_0} with given PDF

$$p_{G_0} \equiv \int \mathcal{N}(\mathbf{w} \mid \boldsymbol{\theta}) dG_0(\boldsymbol{\theta}) = \sum_{k=1}^{K_0} \pi_k^0 \mathcal{N}(\mathbf{w} \mid \boldsymbol{\theta}_k^0), \quad \boldsymbol{\theta}_k^0 \equiv \left(\boldsymbol{\mu}_k^0, \mathbf{V}_k^0\right), \tag{64}$$

where $G_0 = \sum_{k=1}^{K_0} \pi_k^0 \delta_{\boldsymbol{\theta}_k^0}$ is a true but unknown mixing distribution with exactly K_0 number of support points, for some unknown K_0. Furthermore, $\boldsymbol{\Theta}$ is a chosen parameter space, to which we empirically believe that the true parameters belong. In a well-specified setting, all support points of G_0 reside in $\boldsymbol{\Theta}$, but this may not be the case in a misspecified setting. In this section, we assume that the GMM is well-specified, *i.e.,* the data are i.i.d. samples from the mixture density p_{G_0}, where mixing measure G_0 has K_0 support atoms in compact parameter space $\boldsymbol{\Theta}$.

A Bayesian modeller places a prior distribution Π on a suitable subspace of $\overline{\mathcal{G}}(\boldsymbol{\Theta})$. Then, the posterior distribution over G is given by:

$$\Pi\left(G \in B \mid \mathcal{W}\right) \equiv \frac{\int_B \prod_{n=1}^N p_G\left(\mathbf{W}_n\right) \mathrm{d}\Pi(G)}{\int_{\overline{\mathcal{G}}(\boldsymbol{\Theta})} \prod_{n=1}^N p_G\left(\mathbf{W}_n\right) \mathrm{d}\Pi(G)} \tag{65}$$

Here, the GMM p_G is defined in (64) with $K \leq \infty$ unknown number of support points. We are interested in the posterior contraction behaviour of G toward G_0, in addition to recovering the true number of components K_0.

We next recall the notion of Wasserstein distance for mixing measures that prove useful in the next sections.

Wasserstein distance for MM. It is useful to analyze the identifiability and convergence of parameter estimation in mixture models using the notion of Wasserstein distance, as in [38,77]. This distance can be defined as the optimal cost of moving masses in the transformation from one probability measure to another [90,91].

Definition 5.1. Suppose $\boldsymbol{\Theta}$ is equipped with a metric d. The W_r Wasserstein distance for two discrete measures $G = \sum_{k=1}^{K} \pi_k \delta_{\boldsymbol{\theta}_k}$ and $G_0 = \sum_{k=1}^{K_0} \pi_k^0 \delta_{\boldsymbol{\theta}_k^0}$ is

$$W_r\left(G, G_0\right) = \inf_{\mathbf{Q} \in \mathcal{Q}(\boldsymbol{\pi}, \boldsymbol{\pi}^0)} \left[\sum_{i,j} q_{ij} d^r\left(\boldsymbol{\theta}_i, \boldsymbol{\theta}_i^0\right)\right]^{1/r}. \tag{66}$$

It should be emphasized that if a sequence of probability measures $G_N \in \mathcal{O}_{K_0}$ converges to $G_0 \in \mathcal{E}_{K_0}$ under the W_r metric at a rate $\omega_N = o(1)$ for some $r \geq 1$, then there exists a subsequence of G_N such that the set of atoms of G_N converges to the K_0 atoms of G_0, up to a permutation of the atoms, at the same rate ω_N.

Posterior contraction rate in infinite mixtures. With a similar idea as in [35], our starting point is the availability of a mixing measure sample G that is drawn from the posterior distribution $\Pi\left(G \mid \mathcal{W}\right)$, where $\mathcal{W}$ are iid samples of the mixing density p_{G_0}. Under certain conditions on the kernel density, it can be established that for some Wasserstein metric W_r,

$$\Pi\left(G \in \overline{\mathcal{G}}(\boldsymbol{\Theta}) : W_r(G, G_0) \leq \delta\omega_N \mid \mathcal{W}\right) \xrightarrow{N \to \infty} 1 \text{ in } P_{G_0}\text{-probability.} \tag{67}$$

for *all* constant $\delta > 0$, while $\omega_N = o(1)$ is a vanishing rate. Thus ω_N can be assumed to be (slightly) slower than the actual rate of posterior contraction of the mixture

measure. We can also write that ω_N is a rate such that, under the posterior distribution $\Pi(G \mid \mathcal{W})$, $W_r(G, G_0) = o_{P_{G_0}}(\omega_N)$. See [30,38,77] for concrete examples of posterior contraction rates in infinite and (overfitted) finite mixtures.

5.2. Merge-Truncate-Merge (MTM) algorithm for BNP-GLLiM

Link between GLLiM and joint GMM. We start by noting that a GLLiM model on $(\mathbf{X}, \mathbf{Y})$, see (4), with unconstrained parameters $\boldsymbol{\psi} = (\pi_k, \mathbf{c}_k, \boldsymbol{\Gamma}_k, \mathbf{A}_k, \mathbf{b}_k, \boldsymbol{\Sigma}_k)_{k\in[K]}$, is equivalent to a GMM on the joint variable $[\mathbf{X}; \mathbf{Y}]$ with unrestricted parameters, via Lemma 5.2, which is briefly proved in Appendix B.1.

Lemma 5.2. *A GLLiM model on $(\mathbf{X}, \mathbf{Y})$ with unconstrained parameters $\boldsymbol{\psi} = (\pi_k, \mathbf{c}_k, \boldsymbol{\Gamma}_k, \mathbf{A}_k, \mathbf{b}_k, \boldsymbol{\Sigma}_k)_{k\in[K]}$, defined in (4), is equivalent to a GMM on the joint variable $[\mathbf{X}; \mathbf{Y}] \equiv \mathbf{W}$ with unconstrained parameters $\boldsymbol{\nu} = \{\boldsymbol{\mu}_k, \mathbf{V}_k, \rho_k\}_{k=1}^K$,* i.e.,

$$p(\mathbf{w} \mid \boldsymbol{\psi}) = \sum_{k=1}^{K} \rho_k \mathcal{N}_{L+D}(\mathbf{w} \mid \boldsymbol{\mu}_k, \mathbf{V}_k). \tag{68}$$

The parameter $\boldsymbol{\psi}$ can be expressed as a function of $\boldsymbol{\nu}$ by:

$$\begin{aligned}
\pi_k &= \rho_k, \\
\mathbf{c}_k &= \boldsymbol{\mu}_k^{\mathbf{x}}, \\
\boldsymbol{\Gamma}_k &= \mathbf{V}_k^{\mathbf{xx}}, \\
\mathbf{A}_k &= \mathbf{V}_k^{\mathbf{xy}\top} (\mathbf{V}_k^{\mathbf{xx}})^{-1}, \\
\mathbf{b}_k &= \boldsymbol{\mu}_k^{\mathbf{y}} - \mathbf{V}_k^{\mathbf{xy}\top} (\mathbf{V}_k^{\mathbf{xx}})^{-1} \boldsymbol{\mu}_k^{\mathbf{x}}, \\
\boldsymbol{\Sigma}_k &= \mathbf{V}_k^{\mathbf{yy}} - \mathbf{V}_k^{\mathbf{xy}\top} (\mathbf{V}_k^{\mathbf{xx}})^{-1} \mathbf{V}_k^{\mathbf{xy}}.
\end{aligned} \tag{69}$$

Here, we have defined

$$\begin{aligned}
\boldsymbol{\mu}_k &= \begin{bmatrix} \boldsymbol{\mu}_k^{\mathbf{x}} \\ \boldsymbol{\mu}_k^{\mathbf{y}} \end{bmatrix}, \\
\mathbf{V}_k &= \begin{bmatrix} \mathbf{V}_k^{\mathbf{xx}} & \mathbf{V}_k^{\mathbf{xy}} \\ \mathbf{V}_k^{\mathbf{yx}} & \mathbf{V}_k^{\mathbf{yy}} \end{bmatrix}.
\end{aligned}$$

Note that the symmetry $\mathbf{V}_k^\top = \mathbf{V}_k$ of the covariance matrix implies that $\mathbf{V}_k^{\mathbf{xx}}$ and $\mathbf{V}_k^{\mathbf{yy}}$ are symmetric, while $\mathbf{V}_k^{\mathbf{xy}\top} = \mathbf{V}_k^{\mathbf{yx}}$. The parameter vector $\boldsymbol{\nu}$ can be expressed as a function of $\boldsymbol{\psi}$ by:

$$\begin{aligned}
\rho_k &= \pi_k, \\
\boldsymbol{\mu}_k &= \begin{bmatrix} \mathbf{c}_k \\ \mathbf{A}_k \mathbf{c}_k + \mathbf{b}_k \end{bmatrix}, \\
\mathbf{V}_k &= \begin{bmatrix} \boldsymbol{\Gamma}_k & (\mathbf{A}_k \boldsymbol{\Gamma}_k)^\top \\ \mathbf{A}_k \boldsymbol{\Gamma}_k & \boldsymbol{\Sigma}_k + \mathbf{A}_k \boldsymbol{\Gamma}_k \mathbf{A}_k^\top \end{bmatrix}.
\end{aligned} \tag{70}$$

Merge-Truncate-Merge (MTM) algorithm. The detailed and main idea of the pseudocode of the MTM Algorithm 1 for BNP joint GMM follows the novel MTM algorithm for BNP-MM of [35, Algorithm 1]. Broadly speaking, there are two main phases. The first stage is a probabilistic procedure for the merging of atoms that may be in close proximity. The second stage consists of a deterministic procedure for the truncation of extraneous atoms and their appropriate merging with the remaining atoms in a systematic way. More details can be found in [35].

Algorithm 1 MTM Algorithm for BNP joint GMM

Require: Posterior sample $G = \sum_{k=1}^{K} \pi_k \delta_{\boldsymbol{\theta}_k}$, posterior contraction rate ω_N from from (67), and a tunning parameter c.

Ensure: Discrete measure $\widetilde{G}$ and its number of supporting atoms $\widetilde{K}$.

Stage 1: Merge procedure:

1: Reorder atoms $\{\boldsymbol{\theta}_k\}_{k\in[K]}$ by simple random sampling without replacement with corresponding weights $\{\pi_1, \pi_2, \dots\}$.
let $\tau_1, \tau_2, \dots$ denote the new indices, and set $\mathcal{E} = \{\tau_j\}_j$ as the existing set of atoms.

2: Sequentially for each index $\tau_j \in \mathcal{E}$, if there exists an index $\tau_i < \tau_j$ such that $d\left(\boldsymbol{\theta}_{\tau_i}, \boldsymbol{\theta}_{\tau_j}\right) \leq \omega_N$, then:
update $\pi_{\tau_i} = \pi_{\tau_i} + \pi_{\tau_j}$, and remove τ_j from $\mathcal{E}$.

3: Collect $G' = \sum_{j:\ \tau_j \in \mathcal{E}} \pi_{\tau_j} \delta_{\boldsymbol{\theta}_{\tau_j}}$.
write G' as $\sum_{k=1}^{K} q_k \delta_{\boldsymbol{\phi}_k}$ so that $q_1 \geq q_2 \geq \dots$.

Stage 2: Truncate-Merge procedure:

4: Set $\mathcal{A} = \{i : q_i > (c\omega_N)^r\}$, $\mathcal{N} = \{i : q_i \leq (c\omega_N)^r\}$.

5: For each index $i \in \mathcal{A}$, if there is $j \in \mathcal{A}$ such that $j < i$ and $q_i \|\boldsymbol{\phi}_i - \boldsymbol{\phi}_j\|^r \leq (c\omega_N)^r$, then
remove i from $\mathcal{A}$ and add it to $\mathcal{N}$.

6: For each $i \in \mathcal{N}$, find atom $\boldsymbol{\phi}_j$ among $j \in \mathcal{A}$ that is nearest to $\boldsymbol{\phi}_i$
update $q_j = q_j + q_i$.

7: Return $\widetilde{G} = \sum_{j\in\mathcal{A}} q_j \delta_{\boldsymbol{\phi}_j}$ and $\widetilde{K} = |\mathcal{A}|$.

As a consequence, we obtain the following theoretical guarantee, Theorem 5.3, for the outcome of Algorithm 1.

Theorem 5.3 (MTM consistency for BNP joint GMM)**.** *Let G be a posterior sample from the posterior distribution of any Bayesian procedure, namely, $\Pi(G \mid \mathcal{W})$ according to which the upper bound (67) holds for all $\delta > 0$. Let $\widetilde{G}$ and $\widetilde{K}$ be the outcome of Algorithm 1 applied to G, for an arbitrary constant $c > 0$. Then the following hold*

(a) $\Pi\left(\widetilde{K} = K_0 \mid \mathcal{W}\right) \xrightarrow{N\to\infty} 1$ *in P_{G_0}-probability.*

(b) For all $\delta > 0$, $\Pi\left(G \in \overline{\mathcal{G}}(\boldsymbol{\Theta}) : W_r(\widetilde{G}, G_0) \leq \delta\omega_N \mid \mathcal{W}\right) \xrightarrow{N\to\infty} 1$ *in P_{G_0}-probability.*

Proof of Theorem 5.3. Lemma 5.2 implies that BNP joint GMM and BNP-GLLiM are considered equivalent with respect to the number of components in the model selection problem. Therefore, using the results from [35, Theorem 3.2] and Lemma 5.2, it follows that the result of the MTM Algorithm 1 for BNP joint GMM is a consistent estimate of both the number of components and the mixing measure. The latter also admits the upper bound of the posterior contraction rate ω_N, which leads to the desired

Theorem 5.3. □

Remark 1. Regarding the theorem, above, we provide the following comments on posterior consistency for the number of components in BNP-GLLiM after the MTM algorithm post-processing.

(i) As a complementary result to [35], the aim of this paper is to study the practical viability of MTM Algorithm 1 and Theorem 5.3 in the context of high-dimensional BNP-GLLiM. In order to do this, we first need to specify the metric d in $\boldsymbol{\Theta}$, *e.g.,*

$$d\left(\boldsymbol{\theta}_{\tau_i}, \boldsymbol{\theta}_{\tau_j}\right) = \left\|\boldsymbol{\mu}_{\tau_i} - \boldsymbol{\mu}_{\tau_j}\right\| + \left\|\mathbf{V}_{\tau_i} - \mathbf{V}_{\tau_j}\right\|. \tag{71}$$

Here, $\|\cdot\|$ denotes either the l_2-norm elements in $\mathbb{R}^{L+D}$ or the entrywise l_2-norm for matrices in $\mathbb{R}^{(L+D)\times(L+D)}$.

(ii) In practice, one may not have a mixing measure G sampled from the posterior $\Pi\left(\cdot \mid \mathcal{W}\right)$, but rather a sample of G itself. In particular, to deal with large data sets, we need to use VBEM. Therefore, in BNP-GLLiM, instead, we only obtain a sample F_N from the variational posterior $G_V = \sum_{k=1}^{K} \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k(\boldsymbol{\tau})\right] \delta_{\boldsymbol{\theta}_k}$. Here, $\mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k(\boldsymbol{\tau})\right]$ and $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \mathbf{V}_k)$ are defined in (57) and (55), respectively. However, as long as F_N is sufficiently close to G in the sense that $W_r(F_N, G) \lesssim W_r(G, G_0)$, we can still apply the MTM algorithm to F_N, instead. This requires an extension of the above Theorem 5.3 to cover this scenario and verify this approximation condition, which we leave for future work.

6. Numerical experiments

6.1. Data generation

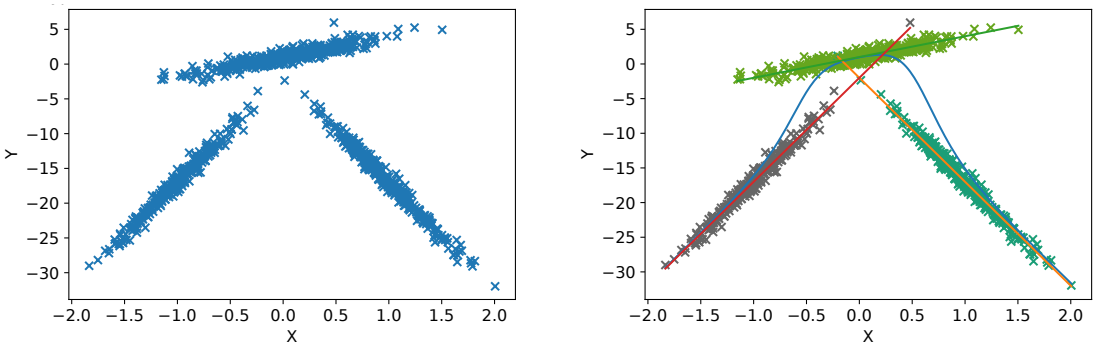
We illustrate our theoretical results on simulated datasets in a more general setting for the BNP approach compared to those considered by [14,56,74]. More specifically, we consider the following true inverse cPDF from GLLiM model as follows

$$s_0\left(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\psi}^0\right) = \sum_{k=1}^{K_0} \frac{\pi_k^0 \mathcal{N}_L\left(\mathbf{x} \mid \mathbf{c}_k^0, \boldsymbol{\Gamma}_k^0\right)}{\sum_{l=1}^{K} \pi_l^0 \mathcal{N}_L\left(\mathbf{x} \mid \mathbf{c}_l^0, \boldsymbol{\Gamma}_l^0\right)} \mathcal{N}_D\left(\mathbf{y} \mid \mathbf{A}_k^0 \mathbf{x} + \mathbf{b}_k^0, \boldsymbol{\Sigma}_k^0\right).$$

Here $K_0 = 3$, $L = D = 1$, and $\boldsymbol{\psi}^0 = \left(\boldsymbol{\pi}^0, \mathbf{c}^0, \boldsymbol{\Gamma}^0, \mathbf{A}^0, \mathbf{b}^0, \boldsymbol{\Sigma}^0\right)$, where

$$\boldsymbol{\pi}^0 = (0.3, 0.4, 0.3), \quad \mathbf{c}^0 = (1, 0.05, -1), \quad \boldsymbol{\Gamma}^0 = (0.1, 0.2, 0.1),$$
$$\mathbf{A}^0 = (-15, 3, 15), \quad \mathbf{b}^0 = (-2, 1, -2), \quad \boldsymbol{\Sigma}^0 = (0.5, 0.3, 0.5).$$

Figure 2a shows typical $N = 1000$ realizations of the true inverse cPDF from GLLiM, representing a $\boldsymbol{\pi}$-shape simulation with three clusters without labels.



(a) Typical realizations without labels.

(b) True conditional expectations with labels.

Figure 2. Typical 1000 realizations of GLLiM's true inverse cPDF with its true conditional expectations.

6.2. Model selection, clustering and regression tasks via MTM-BNP-GLLiM

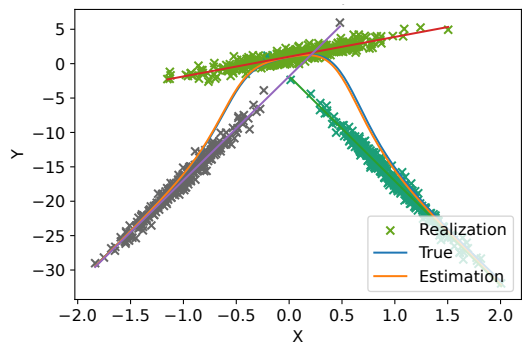
Our goal is to evaluate the inverse and forward cPDF, as well as the conditional means, to investigate the empirical performance of our MTM-BNP-GLLiM in the previous simulation. In Figure 3 it is clear that with the help of MTM Algorithm 1, MTM-BNP-GLLiM can perform regression, clustering and model selection well simultaneously. Without the MTM procedure, BNP-GLLiM performs poorly in model selection, clustering and cPDF estimation, except for conditional expectations as shown in Figure 4.

Next, we illustrate the performance of the MTM algorithm when applied to the variational posterior from BNP-GLLiM. More specifically, the samples in our 100 trials are drawn from $G_V = \sum_{k=1}^{K} \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k(\boldsymbol{\tau})\right] \delta_{\boldsymbol{\theta}_k}$, wherere $\mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k(\boldsymbol{\tau})\right]$ and $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \mathbf{V}_k)$ are defined in (57) and (55) respectively.

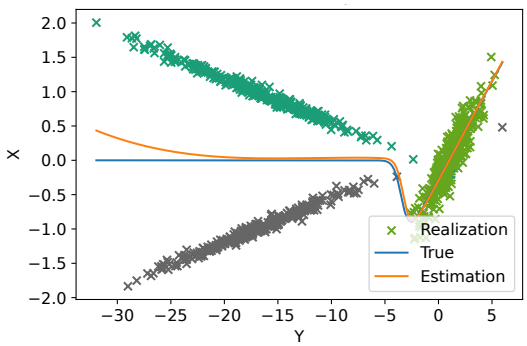
We know that for some constant $\widetilde{C}$, which depends on the covariance matrix $\mathbf{V}_k^0$, the location parameters $\boldsymbol{\mu}_k^0$ and the weights π_k^0, the contraction rate of mixing measures under the location Gaussian DP-MM is $\widetilde{C}\,(\log(N))^{-1/2}$ with respect to the W_2-norm. Similar to [35], our first attempt to choose w_N to satisfy (67) is $w_N = \left(\frac{\log(\log(N))}{\log(N)}\right)^{1/2}$. In fact, we can choose any w_N, as long as $\frac{w_N}{\log(N)^{-1/2}} \to \infty$, in order for w_N to satisfy (67).

Since we only work with finite sample N, it is not expected that the posterior probability for $K_{\text{MTM}} = K_0$ is close to 1 and the input c to Algorithm 1 should be chosen so that $\frac{\widetilde{C}}{(\log(\log(N)))^{1/2}} \leq c$. Furthermore, based on [35, Equation (26)] with a useful lower bound on the posterior mass the mode, for any $1 > \epsilon > 0$, $(1-\epsilon)\left(1 - \sum_{k=1}^{3} \frac{c_0^{r/2}}{\pi_k^0}\right)$, we hope to identify K_0 via the posterior mode with a reasonable estimate. To guarantee $K = K_0$ consistently using the posterior mode safety, we have to choose $c < c_0$, with c_0 satisfying

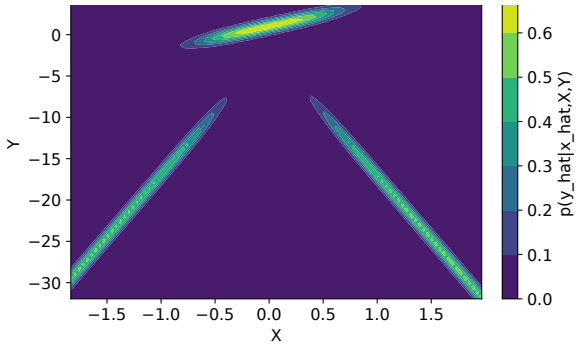
$$(1-\epsilon)\left(1 - \sum_{k=1}^{3} \frac{c_0^{r/2}}{\pi_k^0}\right) > \frac{1}{2} \Leftrightarrow \frac{1-2\epsilon}{2(1-\epsilon)}\left(\sum_{k=1}^{3} \frac{1}{\pi_k^0}\right)^{-1} > c_0^{r/2} > c^{r/2}. \tag{72}$$



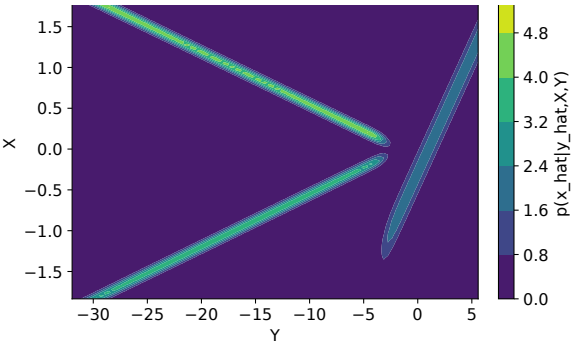
(a) True and estimated inverse means.



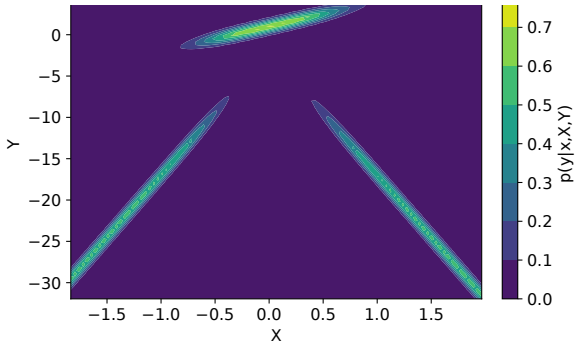
(b) True and estimated forward means.



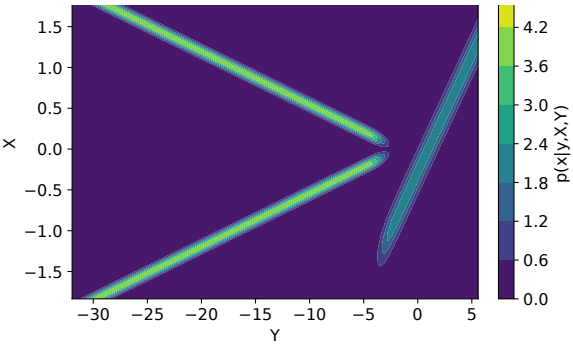
(c) Contour of estimated inverse cPDF.



(d) Contour of estimated forward cPDF.

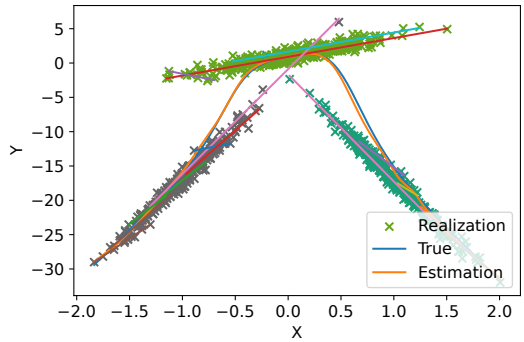


(e) Contour of true inverse cPDF.

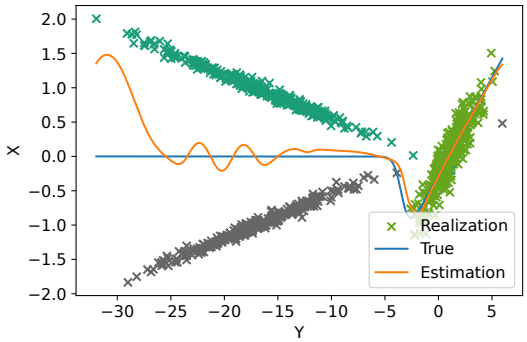


(f) Contour of true forward cPDF.

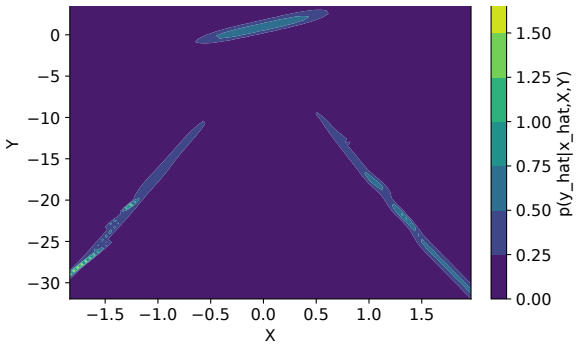
Figure 3. True and estimated inverse and forward cPDF of GLLiM with a good number of components ($K_{\mathrm{MTM}} = 3$) using MTM algorithm for post-processing in MTM-BNP-GLLiM.



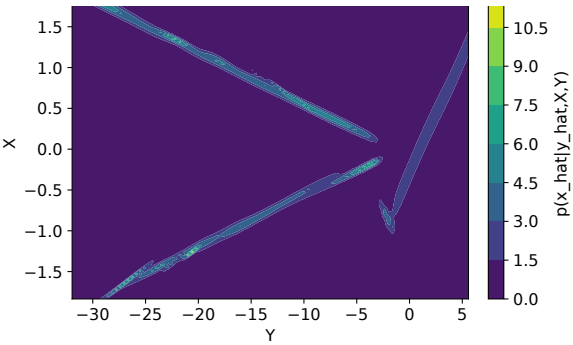
(a) True and estimated inverse means.



(b) True and estimated forward means.



(c) Contour of estimated inverse cPDF.



(d) Contour of estimated forward cPDF.

Figure 4. True and estimated inverse and forward means and CPDFs of GLLiM without MTM algorithm for post-processing in BNP-GLLiM with truncated number of components $K = 20$.

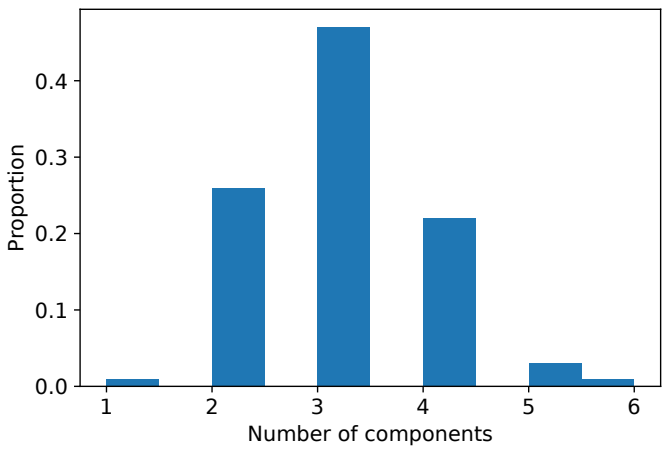


Figure 5. Histogram of K_{MTM} using variational posterior sample with 100 trials and $c = 0.45$.

Therefore, we can choose

$$\left[\frac{1-2\epsilon}{4(1-\epsilon)}\left(\sum_{k=1}^{3}\frac{1}{\pi_k^0}\right)^{-1}\right]^{2/r} = c_0 > c, \text{ for all } \frac{1}{2} > \epsilon > 0. \tag{73}$$

In particular, it is unrealistic to obtain the exact computation of the upper bound c_0 and the lower bound $\frac{\widetilde{C}}{(\log(\log(N)))^{1/2}}$. However, a reasonable estimate may be possible. To this end, we followed the same setting as in [35] to simplify such a complicated procedure by considering a large range of c, and show that there is a range where we can robustly identify the true number of components via the posterior mode. Here, $c = 0.45$ leads to a quite good posterior mode in our numerical experiments, see more in Figure 5.

Although we do not have a theoretical result for the convergence rate of the variational posterior of BNP-GLLiM to the true data generating process, Figure 5 seems to suggest that MTM-BNP-GLLiM gave a comparable good result to the location Gaussian DP-MM in the simulation studies in [35].

7. Perspectives

As indicated in Remark 1, there is a crucial need to formally establish general conditions on the prior, the likelihood and the variational class to characterise the convergence rate of the variational posterior of BNP-GLLiM to the true data generating process. Using the very similar "prior mass and testing" conditions as in [32], we believe that an interesting but challenging extension of the work on variational posterior unconditional distributions for MM models [97] and on adaptive Bayesian estimation for MM and MoE but for true posterior distribution [46,79,87] can help shed some light and answer this important question.

Furthermore, it is important to establish an extensional convergence property of our VBEM algorithm for BNP-GLLiM. This property is only known for GMM from [89]. A potential improvement of the VBEM algorithm developed for BNP-GLLiM can be achieved by combining it with MCMC, taking advantage of both inference approaches

as in [85]. Then, it is also of interest to investigate the performance of BNP-GLLiM on high-dimensional real-world datasets. Finally, as mentioned in Section 6.2, the selection of a good data-driven tuning parameter c as the same idea from the slope heuristic of [8] is crucial for the success of the MTM procedure for any BNP model. We leave these interesting but challenging questions for future research.

References

[1] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723. (Cited on page 4.)

[2] Arbel, J., Kon Kam King, G., Lijoi, A., Nieto-Barajas, L., & Prünster, I. (2021). BNPdensity: Bayesian nonparametric mixture modelling in R. *Australian & New Zealand Journal of Statistics*, 63(3), 542–564. (Cited on page 5.)

[3] Arlot, S. (2019). Minimal penalties and the slope heuristics: a survey. *Journal de la Société Française de Statistique*, 160(3), 1–106. (Cited on page 4.)

[4] Ascolani, F., Lijoi, A., Rebaudo, G., & Zanella, G. (2022). Clustering consistency with Dirichlet process mixtures. *Biometrika*, 110(2), 551–558. (Cited on page 5.)

[5] Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference*. PhD thesis, University of London, University College London (United Kingdom). (Cited on page 34.)

[6] Beal, M. J. & Ghahramani, Z. (2003). The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. *Bayesian statistics*, 7(453-464), 210. (Cited on page 34.)

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

[8] Birgé, L. & Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probability Theory and Related Fields*, 138(1), 33–73. Publisher: Springer. (Cited on pages 4 and 28.)

[9] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag. (Cited on pages 13, 34, 51, 54, 58, 67, and 71.)

[10] Blei, D. M. & Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian analysis*, 1(1), 121–143. ISBN: 1936-0975 Publisher: International Society for Bayesian Analysis. (Cited on pages 10, 12, 50, 52, and 65.)

[11] Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518), 859–877. Publisher: Taylor & Francis. (Cited on page 5.)

[12] Celeux, G., Frühwirth-Schnatter, S., & Robert, C. P. (2019). Model selection for mixture models–perspectives and strategies. In *Handbook of mixture analysis* (pp. 117–154). Chapman and Hall/CRC. (Cited on page 4.)

[13] Chaari, L., Vincent, T., Forbes, F., Dojat, M., & Ciuciu, P. (2013). Fast joint detection-estimation of evoked brain activity in event-related fMRI using a variational approach. *IEEE transactions on Medical Imaging*, 32(5), 821–837. ISBN: 0278-0062 Publisher: IEEE. (Cited on page 34.)

[14] Chamroukhi, F., Samé, A., Govaert, G., & Aknin, P. (2010). A hidden process regression model for functional data description. Application to curve discrimination. *Neurocomputing*, 73(7-9), 1210–1221. Publisher: Elsevier. (Cited on page 23.)

[15] Chen, Z., Deng, Y., Wu, Y., Gu, Q., & Li, Y. (2022). Towards Understanding the Mixture-of-Experts Layer in Deep Learning. In A. H. Oh, A. Agarwal, D. Belgrave, & K. Cho (Eds.), *Advances in Neural Information Processing Systems*. (Cited on page 4.)

[16] Corradin, R., Canale, A., & Nipoti, B. (2021). BNPmix: An R Package for Bayesian Nonparametric Modeling via Pitman-Yor Mixtures. *Journal of Statistical Software*, 100(15), 1 – 33. (Cited on page 5.)

[17] De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., & Ruggiero, M. (2015). Are

Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE transactions on pattern analysis and machine intelligence*, 37(2), 212–229. ISBN: 0162-8828 Publisher: IEEE. (Cited on page 10.)

[18] Deleforge, A., Forbes, F., & Horaud, R. (2015). High-dimensional regression with gaussian mixtures and partially-latent response variables. *Statistics and Computing*, 25(5), 893–911. (Cited on pages 4, 6, 7, 8, 9, 45, and 46.)

[19] Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1), 1–22. Publisher: John Wiley & Sons, Ltd. (Cited on page 33.)

[20] Do, D., Do, L., & Nguyen, X. (2022). Strong identifiability and parameter learning in regression with heterogeneous response. *arXiv preprint arXiv:2212.04091*. (Cited on page 4.)

[21] Do, T. G., Le, H. K., Nguyen, T., Pham, Q., Nguyen, B. T., Doan, T.-N., Liu, C., Ramasamy, S., Li, X., & HOI, S. (2023). HyperRouter: Towards Efficient Training and Inference of Sparse Mixture of Experts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* Singapore: Association for Computational Linguistics. (Cited on page 4.)

[22] Durand, J.-B., Forbes, F., Phan, C., Truong, L., Nguyen, H., & Dama, F. (2022). Bayesian non-parametric spatial prior for traffic crash risk mapping: A case study of Victoria, Australia. *Australian & New Zealand Journal of Statistics*, 64(2), 171–204. (Cited on pages 5, 11, and 16.)

[23] Escobar, M. D. & West, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, 90(430), 577–588. Publisher: Taylor & Francis. (Cited on page 5.)

[24] Favaro, S., Lijoi, A., Nava, C., Nipoti, B., Pruenster, I., & Teh, Y. W. (2016). On the stick-breaking representation for homogeneous NRMIs. *Bayesian Analysis*, 11(3), 697–724. ISBN: 1936-0975 Publisher: International Society for Bayesian Analysis. (Cited on page 10.)

[25] Fedus, W., Dean, J., & Zoph, B. (2022). A review of sparse expert models in deep learning. *arXiv preprint arXiv:2209.01667*. (Cited on page 4.)

[26] Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2), 209 – 230. Publisher: Institute of Mathematical Statistics. (Cited on page 9.)

[27] Forbes, F., Arnaud, A., Lemasson, B., & Barbier, E. (2019). Component elimination strategies to fit mixtures of multiple scale distributions. In *RSSDS 2019 - Research School on Statistics and Data Science*, volume 1150 of *Communications in Computer and Information Science* (pp. 81–95). Melbourne, Australia: Springer. (Cited on page 5.)

[28] Forbes, F., Nguyen, H. D., Nguyen, T., & Arbel, J. (2022a). Mixture of expert posterior surrogates for approximate Bayesian computation. In *JDS 2022 - 53èmes Journées de Statistique de la Société Française de Statistique (SFdS)* Lyon, France. (Cited on page 4.)

[29] Forbes, F., Nguyen, H. D., Nguyen, T., & Arbel, J. (2022b). Summary statistics and discrepancy measures for approximate Bayesian computation via surrogate posteriors. *Statistics and Computing*, 32(5), 85. (Cited on page 4.)

[30] Gao, F. & Vaart, A. v. d. (2016). Posterior contraction rates for deconvolution of Dirichlet-Laplace mixtures. *Electronic Journal of Statistics*, 10(1), 608 – 627. Publisher: Institute of Mathematical Statistics and Bernoulli Society. (Cited on page 21.)

[31] Genovese, C. R. & Wasserman, L. (2000). Rates of convergence for the Gaussian mixture sieve. *The Annals of Statistics*, 28(4), 1105 – 1127. Publisher: Institute of Mathematical Statistics. (Cited on page 4.)

[32] Ghosal, S., Ghosh, J. K., & Vaart, A. W. v. d. (2000). Convergence rates of posterior distributions. *The Annals of Statistics*, 28(2), 500 – 531. Publisher: Institute of Mathematical Statistics. (Cited on page 27.)

[33] Ghosal, S. & Van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press. (Cited on pages 5, 10, and 52.)

[34] Green, P. J. & Richardson, S. (2001). Modelling Heterogeneity With and Without the Dirichlet Process. *Scandinavian Journal of Statistics*, 28(2), 355–375. (Cited on page 5.)

[35] Guha, A., Ho, N., & Nguyen, X. (2021). On posterior contraction of parameters and interpretability in Bayesian mixture modeling. *Bernoulli*, 27(4), 2159 – 2188. Publisher: Bernoulli Society for Mathematical Statistics and Probability. (Cited on pages 5, 6, 20, 22, 23, 24, and 27.)

[36] Hjort, N. L., Holmes, C., Müller, P., & Walker, S. G., Eds. (2010). *Bayesian Nonparametrics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. (Cited on page 5.)

[37] Ho, N. & Nguyen, X. (2016a). Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *The Annals of Statistics*, 44(6), 2726 – 2755. Publisher: Institute of Mathematical Statistics and Bernoulli Society. (Cited on page 4.)

[38] Ho, N. & Nguyen, X. (2016b). On strong identifiability and convergence rates of parameter estimation in finite mixtures. *Electronic Journal of Statistics*, 10(1), 271–307. Publisher: The Institute of Mathematical Statistics and the Bernoulli Society. (Cited on pages 4, 20, and 21.)

[39] Ho, N., Yang, C.-Y., & Jordan, M. I. (2022). Convergence Rates for Gaussian Mixtures of Experts. *Journal of Machine Learning Research*, 23(323), 1–81. (Cited on page 4.)

[40] Ingrassia, S., Minotti, S. C., & Vittadini, G. (2012). Local Statistical Modeling via a Cluster-Weighted Approach with Elliptical Distributions. *Journal of Classification*, 29(3), 363–401. (Cited on page 9.)

[41] Ishwaran, H. & James, L. F. (2001). Gibbs Sampling Methods for Stick-Breaking Priors. *Journal of the American Statistical Association*, 96(453), 161–173. Publisher: Taylor & Francis. (Cited on page 10.)

[42] Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural computation*, 3(1), 79–87. Publisher: MIT Press. (Cited on page 4.)

[43] Jiang, W. & Tanner, M. A. (1999). Hierarchical mixtures-of-experts for exponential family regression models: approximation and maximum likelihood estimation. *Annals of Statistics*, (pp. 987–1011). Publisher: JSTOR. (Cited on page 4.)

[44] Jordan, M. I. & Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural computation*, 6(2), 181–214. Publisher: MIT Press. (Cited on page 4.)

[45] Kock, L., Klein, N., & Nott, D. J. (2022). Variational inference and sparsity in high-dimensional deep Gaussian mixture models. *Statistics and Computing*, 32(5), 70. (Cited on page 5.)

[46] Kruijer, W., Rousseau, J., & Vaart, A. v. d. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics*, 4(none), 1225 – 1257. Publisher: Institute of Mathematical Statistics and Bernoulli Society. (Cited on page 27.)

[47] Li, K.-C. (1991). Sliced Inverse Regression for Dimension Reduction. *Journal of the American Statistical Association*, 86(414), 316–327. (Cited on page 6.)

[48] Li, N., Li, W., Jiang, Y., & Xia, S.-T. (2022). Deep Dirichlet process mixture models. In J. Cussens & K. Zhang (Eds.), *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research* (pp. 1138–1147).: PMLR. (Cited on page 5.)

[49] Luo, C. & Sun, S. (2017). Variational Mixtures of Gaussian Processes for Classification. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17* (pp. 4603–4609). (Cited on pages 16 and 17.)

[50] Lü, H., Arbel, J., & Forbes, F. (2020). Bayesian nonparametric priors for hidden Markov random fields. *Statistics and Computing*, 30(4), 1015–1035. (Cited on pages 5, 10, 11, and 14.)

[51] Maceachern, S. N. & Müller, P. (1998). Estimating Mixture of Dirichlet Process Models. *Journal of Computational and Graphical Statistics*, 7(2), 223–238. Publisher: Taylor & Francis. (Cited on page 5.)

[52] Masoudnia, S. & Ebrahimpour, R. (2014). Mixture of experts: a literature survey. *Artificial Intelligence Review*, 42(2), 275–293. (Cited on page 4.)

[53] McLachlan, G. J. & Krishnan, T. (1997). *The EM algorithm and extensions*. New York, USA: Wiley. (Cited on page 33.)

[54] McLachlan, G. J. & Peel, D. (2000). *Finite Mixture Models*. John Wiley & Sons. (Cited

on page 4.)

[55] Miller, J. W. & Harrison, M. T. (2014). Inconsistency of Pitman-Yor Process Mixtures for the Number of Components. *Journal of Machine Learning Research*, 15(96), 3333–3370. (Cited on page 5.)

[56] Montuelle, L. & Le Pennec, E. (2014). Mixture of Gaussian regressions model with logistic weights, a penalized maximum likelihood approach. *Electronic Journal of Statistics*, 8(1), 1661–1695. Publisher: The Institute of Mathematical Statistics and the Bernoulli Society. (Cited on page 23.)

[57] Neal, R. M. (2000). Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, 9(2), 249–265. Publisher: Taylor & Francis. (Cited on page 5.)

[58] Neal, R. M. & Hinton, G. E. (1998). A View of the EM Algorithm that Justifies Incremental, Sparse, and other Variants. In M. I. Jordan (Ed.), *Learning in Graphical Models* (pp. 355–368). Dordrecht: Springer Netherlands. (Cited on page 33.)

[59] Nguyen, H., Akbarian, P., Nguyen, T., & Ho, N. (2023a). A General Theory for Softmax Gating Multinomial Logistic Mixture of Experts. *arXiv preprint arXiv:2310.14188*. (Cited on page 4.)

[60] Nguyen, H., Nguyen, T., & Ho, N. (2023b). Demystifying Softmax Gating in Gaussian Mixture of Experts. In *Advances in Neural Information Processing Systems*. (Cited on page 4.)

[61] Nguyen, H., Nguyen, T., Nguyen, K., & Ho, N. (2023c). Towards Convergence Rates for Parameter Estimation in Gaussian-gated Mixture of Experts. *arXiv preprint arXiv:2305.07572*. (Cited on page 4.)

[62] Nguyen, H., Nguyen, T., Nguyen, K., & Ho, N. (2023d). Towards convergence rates for parameter estimation in Gaussian-gated mixture of experts. *arxiv preprint arxiv 2305.07572*. (Cited on page 4.)

[63] Nguyen, H. D. & Chamroukhi, F. (2018). Practical and theoretical aspects of mixture-of-experts modeling: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4), e1246. Publisher: Wiley Online Library. (Cited on page 4.)

[64] Nguyen, H. D., Chamroukhi, F., & Forbes, F. (2019). Approximation results regarding the multiple-output Gaussian gated mixture of linear experts model. *Neurocomputing*, 366, 208–214. (Cited on page 4.)

[65] Nguyen, H. D., Fryer, D., & McLachlan, G. J. (2022a). Order selection with confidence for finite mixture models. *Journal of the Korean Statistical Society*. (Cited on page 5.)

[66] Nguyen, H. D., Lloyd-Jones, L. R., & McLachlan, G. J. (2016). A universal approximation theorem for mixture-of-experts models. *Neural computation*, 28(12), 2585–2593. Publisher: MIT Press. (Cited on page 4.)

[67] Nguyen, H. D., Nguyen, T., Chamroukhi, F., & McLachlan, G. J. (2021a). Approximations of conditional probability density functions in Lebesgue spaces via mixture of experts models. *Journal of Statistical Distributions and Applications*, 8(1), 13. (Cited on page 4.)

[68] Nguyen, T. (2021). *Model Selection and Approximation in High-dimensional Mixtures of Experts Models: from Theory to Practice*. PhD Thesis, Normandie Université. (Cited on page 4.)

[69] Nguyen, T. & Bonilla, E. (2014). Fast Allocation of Gaussian Process Experts. In E. P. Xing & T. Jebara (Eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research* (pp. 145–153). Bejing, China: PMLR. Issue: 1. (Cited on pages 16 and 17.)

[70] Nguyen, T., Chamroukhi, F., Nguyen, H. D., & Forbes, F. (2021b). Non-asymptotic model selection in block-diagonal mixture of polynomial experts models. *Preprint. arXiv:2104.08959*. arXiv: 2104.08959. (Cited on page 4.)

[71] Nguyen, T., Chamroukhi, F., Nguyen, H. D., & Forbes, F. (2022b). Model selection by penalization in mixture of experts models with a non-asymptotic approach. In *JDS 2022 - 53èmes Journées de Statistique de la Société Française de Statistique (SFdS)* Lyon, France. (Cited on page 4.)

[72] Nguyen, T., Chamroukhi, F., Nguyen, H. D., & McLachlan, G. J. (2023e). Approxima-

tion of probability density functions via location-scale finite mixtures in Lebesgue spaces. *Communications in Statistics - Theory and Methods*, 52(14), 5048–5059. (Cited on page 4.)

[73] Nguyen, T., Nguyen, D. N., Nguyen, H. D., & Chamroukhi, F. (2023f). A non-asymptotic theory for model selection in high-dimensional mixture of experts via joint rank and variable selection. In *AJCAI Australasian Joint Conference on Artificial Intelligence 2023*. (Cited on page 4.)

[74] Nguyen, T., Nguyen, H. D., Chamroukhi, F., & Forbes, F. (2022c). A non-asymptotic approach for model selection via penalization in high-dimensional mixture of experts models. *Electronic Journal of Statistics*, 16(2), 4742 – 4822. Publisher: Institute of Mathematical Statistics and Bernoulli Society. (Cited on pages 4, 9, and 23.)

[75] Nguyen, T., Nguyen, H. D., Chamroukhi, F., & McLachlan, G. J. (2020). Approximation by finite mixtures of continuous density functions that vanish at infinity. *Cogent Mathematics & Statistics*, 7(1), 1750861. Publisher: Cogent OA. (Cited on page 4.)

[76] Nguyen, T., Nguyen, H. D., Chamroukhi, F., & McLachlan, G. J. (2023g). Non-asymptotic oracle inequalities for the Lasso in high-dimensional mixture of experts. *arXiv:2009.10622*. (Cited on page 4.)

[77] Nguyen, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics*, 41(1), 370–400. Publisher: Institute of Mathematical Statistics. (Cited on pages 4, 20, and 21.)

[78] Norets, A. (2010). Approximation of conditional densities by smooth mixtures of regressions. *The Annals of Statistics*, 38(3), 1733 – 1766. Publisher: Institute of Mathematical Statistics. (Cited on page 4.)

[79] Norets, A. & Pati, D. (2017). Adaptive Bayesian estimation of conditional densities. *Econometric Theory*, 33(4), 980–1012. Publisher: Cambridge University Press. (Cited on page 27.)

[80] Perthame, E., Forbes, F., & Deleforge, A. (2018). Inverse regression approach to robust nonlinear high-to-low dimensional mapping. *Journal of Multivariate Analysis*, 163(C), 1–14. (Cited on page 9.)

[81] Pitman, J. & Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2), 855–900. ISBN: 0091-1798 Publisher: JSTOR. (Cited on pages 9 and 10.)

[82] Rakhlin, A., Panchenko, D., & Mukherjee, S. (2005). Risk bounds for mixture density estimation. *ESAIM: PS*, 9, 220–229. (Cited on page 4.)

[83] Ross, J. & Dy, J. (2013). Nonparametric Mixture of Gaussian Processes with Constraints. In S. Dasgupta & D. McAllester (Eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research* (pp. 1346–1354). Atlanta, Georgia, USA: PMLR. Issue: 3. (Cited on pages 16 and 17.)

[84] Rousseau, J. & Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5), 689–710. (Cited on page 5.)

[85] Ruiz, F. & Titsias, M. (2019). A Contrastive Divergence for Combining Variational Inference and MCMC. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research* (pp. 5537–5545).: PMLR. (Cited on page 28.)

[86] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464. Publisher: Institute of Mathematical Statistics. (Cited on page 4.)

[87] Shen, W., Tokdar, S. T., & Ghosal, S. (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika*, 100(3), 623–640. Publisher: Oxford University Press. (Cited on pages 4 and 27.)

[88] Svensén, M. & Bishop, C. M. (2005). Robust Bayesian mixture modelling. *Neurocomputing*, 64, 235–252. (Cited on page 16.)

[89] Titterington, D. M. & Wang, B. (2006). Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. *Bayesian Analysis*, 1(3), 625 – 650. Publisher: International Society for Bayesian Analysis. (Cited on page 27.)

[90] Villani, C. (2003). *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society. (Cited on page 20.)
[91] Villani, C. (2009). *Optimal transport: old and new*, volume 338. Springer. (Cited on page 20.)
[92] Wang, C., Paisley, J., & Blei, D. M. (2011). Online Variational Inference for the Hierarchical Dirichlet Process. In G. Gordon, D. Dunson, & M. Dudík (Eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research* (pp. 752–760). Fort Lauderdale, FL, USA: PMLR. (Cited on page 12.)
[93] Wu, D. & Ma, J. (2018). A Two-Layer Mixture Model of Gaussian Process Functional Regressions and Its MCMC EM Algorithm. *IEEE Transactions on Neural Networks and Learning Systems*, 29(10), 4894–4904. (Cited on pages 16 and 17.)
[94] Xu, L., Jordan, M., & Hinton, G. E. (1995). An Alternative Model for Mixtures of Experts. In G. Tesauro, D. Touretzky, & T. Leen (Eds.), *Advances in Neural Information Processing Systems*, volume 7: MIT Press. (Cited on page 4.)
[95] Yuan, C. & Neubauer, C. (2008). Variational Mixture of Gaussian Process Experts. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in Neural Information Processing Systems*, volume 21: Curran Associates, Inc. (Cited on pages 16 and 17.)
[96] Yuksel, S. E., Wilson, J. N., & Gader, P. D. (2012). Twenty Years of Mixture of Experts. *IEEE Transactions on Neural Networks and Learning Systems*, 23(8), 1177–1193. (Cited on page 4.)
[97] Zhang, F. & Gao, C. (2020). Convergence rates of variational posterior distributions. *The Annals of Statistics*, 48(4), 2180 – 2207. Publisher: Institute of Mathematical Statistics. (Cited on page 27.)

Appendix A. Details of VBEM for BNP-GLLiM model

A.1. Variational Bayesian expectation-maximization principle

The clustering task consists mainly of estimating the unknown labels $\mathcal{Z} = (z_n)_{n\in[N]}$ from the observed data $(\mathcal{Y}, \mathcal{X}) = (\mathbf{y}_n, \mathbf{x}_n)_{n\in[N]}$, whose joint distribution $p(\mathcal{Y}, \mathcal{X}, \mathcal{Z} \mid \boldsymbol{\Theta}; \boldsymbol{\phi})$ is determined by a set of parameters denoted by $\boldsymbol{\Theta}$ and often by additional hyperparameters $\boldsymbol{\phi}$.

The expectation-maximization (EM) algorithm [19,53] is a generative technique for maximum likelihood estimation (MLE) in the presence of unobserved latent variables or missing data. An EM iteration consists of two steps usually referred to as the E-step in which the expectation of the so-called complete log-likelihood is computed and the M-step in which this expectation is maximized over $\boldsymbol{\Theta}$. An equivalent way to define EM is the following. As discussed in [58], EM can be viewed as an alternating maximization procedure of a function $\mathcal{F}_0$ defined, for any probability distribution $q_{\mathcal{Z}}$ over labels $\mathcal{Z}$, by

$$\mathcal{F}_0(q_{\mathcal{Z}}, \boldsymbol{\Theta}, \boldsymbol{\phi}) = \sum_{\mathcal{Z}} q_{\mathcal{Z}}(\mathcal{Z}) \log p(\mathcal{Y}, \mathcal{X}, \mathcal{Z} \mid \boldsymbol{\Theta}, \boldsymbol{\phi}) - \mathbb{E}_{q_{\mathcal{Z}}}[\log q_{\mathcal{Z}}(\mathcal{Z})] \tag{A1}$$

$$= \mathbb{E}_{q_{\mathcal{Z}}}\left[\log \frac{p(\mathcal{Y}, \mathcal{X}, \mathcal{Z} \mid \boldsymbol{\Theta}, \boldsymbol{\phi})}{q_{\mathcal{Z}}(\mathcal{Z})}\right] \tag{A2}$$

where $-\mathbb{E}_{q_{\mathcal{Z}}}[\log q_{\mathcal{Z}}(\mathcal{Z})]$ is the entropy of $q_{\mathcal{Z}}$ and $\mathbb{E}_q[\cdot]$ is the expectation with respect to q. The function $\mathcal{F}_0$ depends on the observations $(\mathcal{Y}, \mathcal{X})$, which are fixed throughout and are therefore omitted from the notation.

When prior knowledge on the parameters is available, an alternative approach con-

sists of replacing the MLE by a maximum a posteriori (MAP) estimation of $\boldsymbol{\Theta}$ using the prior knowledge encoded in a distribution $p(\boldsymbol{\Theta})$. More precisely, the MLE of $\boldsymbol{\Theta}$ is then replaced by a point estimation $\widehat{\boldsymbol{\Theta}} = \operatorname{argmax}_{\boldsymbol{\Theta}\in\underline{\boldsymbol{\Theta}}} p(\boldsymbol{\Theta} \mid \mathcal{Y}, \mathcal{X})$. In this paper, instead of considering only point estimation of $\boldsymbol{\Theta}$, we carry out a fully Bayesian approach. That is, we integrate out $\boldsymbol{\Theta}$ as follows

$$p(\mathcal{Z} \mid \mathcal{Y}, \mathcal{X}) = \int_{\boldsymbol{\Theta}} p(\mathcal{Z} \mid \mathcal{Y}, \mathcal{X}, \boldsymbol{\Theta}) p(\boldsymbol{\Theta} \mid \mathcal{Y}, \mathcal{X}) d\boldsymbol{\Theta}. \tag{A3}$$

This integration requires the computation of the marginal point distribution $p(\boldsymbol{\Theta} \mid \mathcal{Y}, \mathcal{X})$, which is usually not available in closed-form. As an alternative to costly simulation-based methods (*e.g.,* Markov chain Monte Carlo (MCMC)), an EM-like procedure using variational approximation can provide approximations of the marginal posterior distributions $p(\boldsymbol{\Theta} \mid \mathcal{Y}, \mathcal{X})$ and $p(\mathcal{Z} \mid \mathcal{Y}, \mathcal{X})$. This approach is referred to as VBEM for variational Bayesian EM, was introduced by [6].

To deal with the BNP-GLLiM model, we need to use the VBEM with hyperparameter optimization of [5, Figure 2.5 and Algorithm 5.3]. Let $q_{\mathcal{Z}}$ and $q_{\boldsymbol{\Theta}}$ denote the distributions over $\mathcal{Z}$ and $\boldsymbol{\Theta}$ respectively, which will serve as approximations to the true posteriors. More specifically, in the Bayesian setting, the intractable posterior $p(\mathcal{Z}, \boldsymbol{\Theta} \mid \mathcal{Y}, \mathcal{X}; \boldsymbol{\phi})$ is approximated by the variational posterior $q(\mathcal{Z}, \boldsymbol{\Theta}) = q_{\mathcal{Z}}(\mathcal{Z}) q_{\boldsymbol{\Theta}}(\boldsymbol{\Theta})$.

Similar to standard EM, VBEM maximizes the following evidence lower bound (often abbreviated ELBO, and sometimes called the variational lower bound or negative variational free energy), defined for arbitrary $q_{\mathcal{Z}}$ and $q_{\boldsymbol{\Theta}}$ distributions by

$$\begin{aligned}\mathcal{F}\left(q_{\mathcal{Z}}, q_{\boldsymbol{\Theta}}, \boldsymbol{\phi}\right) &= \mathbb{E}_{q_{\mathcal{Z}} q_{\boldsymbol{\Theta}}}\left[\log \frac{p(\mathcal{Y}, \mathcal{X}, \mathcal{Z}, \boldsymbol{\Theta}; \boldsymbol{\phi})}{q_{\mathcal{Z}}(\mathcal{Z}) q_{\boldsymbol{\Theta}}(\boldsymbol{\Theta})}\right] \\ &= \log p(\mathcal{Y}, \mathcal{X} \mid \boldsymbol{\phi}) - \mathrm{KL}\left(q_{\mathcal{Z}} q_{\boldsymbol{\Theta}} \| p(\mathcal{Z}, \boldsymbol{\Theta} \mid \mathcal{Y}, \mathcal{X}, \boldsymbol{\phi})\right) \leq \log p(\mathcal{Y}, \mathcal{X} \mid \boldsymbol{\phi}),\end{aligned} \tag{A4}$$

alternatively over $q_{\mathcal{Z}}, q_{\boldsymbol{\Theta}}$ and $\boldsymbol{\phi}$. Here, KL stands for Kullback-Leibler divergence. It is worth noting that adding a prior on $\boldsymbol{\Theta}$ is formally equivalent to considering $\boldsymbol{\Theta}$ as missing variables, while the hyperparameters $\boldsymbol{\phi}$ play the role of the parameters of interest in MLE.

The alternate maximization over $\mathcal{F}$ leads to the VBEM algorithm, which can be decomposed into three steps. It is easy to show, using the KL divergence properties, that the maximization over $q_{\mathcal{Z}}$ and $q_{\boldsymbol{\Theta}}$ leads to the following E-steps, see, *e.g.,* [13, Appendix A], [5, Theorem 2.1] and [9, Section 10.1.1], which is essentially coordinate ascent in the function space of variational distributions. Furthermore, the following update rules for E-steps converge to a local maximum of $\mathcal{F}\left(q_{\mathcal{Z}}, q_{\boldsymbol{\Theta}}, \boldsymbol{\phi}\right)$. At the rth iteration, using current values $\boldsymbol{\phi}^{(r-1)}$ and $q_{\boldsymbol{\Theta}}^{(r-1)}$, we get the following updating,

$$\textbf{VB-E-}\mathcal{Z}\textbf{:}\ q_{\mathcal{Z}}^{(r)}(\mathcal{Z}) \propto \exp \mathbb{E}_{q_{\boldsymbol{\Theta}}^{(r-1)}}\left[\log p(\mathcal{Y}, \mathcal{X}, \mathcal{Z}, \boldsymbol{\Theta}; \boldsymbol{\phi}^{(r-1)})\right], \tag{A5}$$

$$\textbf{VB-E-}\boldsymbol{\Theta}\textbf{:}\ q_{\boldsymbol{\Theta}}^{(r)}(\boldsymbol{\Theta}) \propto \exp \mathbb{E}_{q_{\mathcal{Z}}^{(r)}}\left[\log p(\mathcal{Y}, \mathcal{X}, \mathcal{Z}, \boldsymbol{\Theta}; \boldsymbol{\phi}^{(r-1)})\right], \tag{A6}$$

$$\textbf{VB-M-}\boldsymbol{\phi}\textbf{:}\ \boldsymbol{\phi}^{(r)} \propto \operatorname*{argmax}_{\boldsymbol{\phi}} \mathbb{E}_{q_{\mathcal{Z}}^{(r)} q_{\boldsymbol{\Theta}}^{(r)}}\left[\log p(\mathcal{Y}, \mathcal{X}, \mathcal{Z}, \boldsymbol{\Theta}; \boldsymbol{\phi})\right]. \tag{A7}$$

In practice, we can decide which parameters to treat as genuine parameters $\boldsymbol{\Theta}$ or as hyperparameters $\boldsymbol{\phi}$, depending on whether some prior knowledge is available for only a subset of the parameters, or whether the model has hyperparameters $\boldsymbol{\phi}$ for

which no prior information is available. Furthermore, for complex models, $q_{\boldsymbol{\Theta}}$ and $q_{\mathcal{Z}}$ may need to be further restricted to simpler forms, such as factorised forms, to ensure tractable VB-E steps. This is illustrated in the next Appendix C.1 for the BNP-GLLiM inference.

A.2. VB-E-$\boldsymbol{\tau}$ step from Section 3.1

To achieve results from Section 3.1, we make use of (18), (23), (C2), and are only interested in the functional dependence of the right-hand side of (A6) on the variable τ_k. Thus any terms that do not depend on τ_k can be absorbed into the additive normalization constant, giving

$$\begin{aligned} q_{\tau_k}(\tau_k) &= \exp\left\{\mathbb{E}_{q_{\alpha,\sigma}}\left[\log p\left(\tau_k \mid \alpha, \sigma\right)\right] + \sum_{n=1}^{N} \mathbb{E}_{q_{z_n} q_{\tau_{\setminus\{k\}}}}\left[\log \pi_{z_n}(\boldsymbol{\tau})\right]\right\} + \text{constant} \\ &\propto \exp\left\{ -\mathbb{E}_{q_{\alpha,\sigma}}[\sigma] \log \tau_k + \left[\mathbb{E}_{q_{\alpha,\sigma}}[\alpha] + k\mathbb{E}_{q_{\alpha,\sigma}}[\sigma] - 1\right] \log\left(1-\tau_k\right)\right. \\ &\quad \left. + \sum_{n=1}^{N} q_{z_n}(k) \log \tau_k + \sum_{n=1}^{N} \sum_{j=k+1}^{K} q_{z_n}(j) \log\left(1-\tau_k\right)\right\} \\ &= \text{Beta}\left(\tau_k \mid \widehat{\gamma}_{k,1}, \widehat{\gamma}_{k,2}\right). \end{aligned} \tag{A8}$$

Here,

$$\widehat{\gamma}_{k,1} = 1 - \mathbb{E}_{q_{\alpha,\sigma}}[\sigma] + \sum_{n=1}^{N} q_{z_n}(k) = 1 - \mathbb{E}_{q_\sigma}[\sigma] + \sum_{n=1}^{N} q_{z_n}(k), \tag{A9}$$

$$\widehat{\gamma}_{k,2} = \mathbb{E}_{q_{\alpha,\sigma}}[\alpha] + k\mathbb{E}_{q_{\alpha,\sigma}}[\sigma] + \sum_{n=1}^{N} \sum_{j=k+1}^{K} q_{z_n}(j) = \mathbb{E}_{q_{\alpha,\sigma}}[\alpha] + k\mathbb{E}_{q_{\alpha,\sigma}}[\sigma] + \sum_{j=k+1}^{K} \sum_{n=1}^{N} q_{z_n}(j). \tag{A10}$$

Furthermore, we used the fact that

$$\begin{aligned} \log p\left(\tau_k \mid \alpha, \sigma\right) &= \log\left[\text{Beta}\left(\tau_k \mid 1-\sigma, \alpha + k\sigma\right)\right] = \log\left[\frac{\Gamma(1-\sigma+k\sigma)}{\Gamma(1-\sigma)\Gamma(\alpha+k\sigma)} \tau_k^{1-\sigma-1}(1-\tau_k)^{\alpha+k\sigma-1}\right], \\ \log \pi_{z_n}(\boldsymbol{\tau}) &= \log\left[\tau_{z_n} \prod_{l=1}^{z_n-1}(1-\tau_l)\right] = \log \tau_{z_n} + \sum_{l=1}^{z_n-1} \log\left(1-\tau_l\right), \\ q_{\tau_{\setminus\{k\}}}(\tau_l) &= \prod_{i=1, i\neq k}^{K-1} q_{\tau_i}(\tau_l), \quad d\tau_{\setminus\{k\}} = \prod_{i=1, i\neq k}^{K-1} d\tau_i. \end{aligned} \tag{A11}$$

Finally, we have for $k \in [K]$, let $N_k = \sum_{n=1}^{N} q_{z_n}(k)$ correspond to the weight of

cluster k, then

$$q_{\tau_k}(\tau_k) = \text{Beta}(\tau_k \mid \widehat{\gamma}_{k,1}, \widehat{\gamma}_{k,2}), \tag{A12}$$

$$\widehat{\gamma}_{k,1} = 1 - \mathbb{E}_{q_{\alpha,\sigma}}[\sigma] + N_k, \widehat{\gamma}_{k,2} = \mathbb{E}_{q_{\alpha,\sigma}}[\alpha] + k\mathbb{E}_{q_{\alpha,\sigma}}[\sigma] + \sum_{l=k+1}^{K} N_l. \tag{A13}$$

A.3. VB-E-(α, σ) step from Section 3.1

In the PY case, to achieve results from Section 3.1, we make use of (18), (17), (C2), (C3), and are only interested in the functional dependence of the right-hand side of (A6) to the variables (α, σ). Thus, any terms that do not depend on (α, σ) can be included in the additive normalization constant. This results in $q_{\alpha,\sigma}(\alpha, \sigma)$ being proportional to

$$\begin{aligned}
&\widetilde{q}_{\alpha,\sigma}(\alpha, \sigma) \\
&= p(\alpha, \sigma \mid s_1, s_2, a) \exp\left\{\mathbb{E}_{\prod_{k=1}^{K-1} q_{\tau_k}}\left[\log \prod_{k=1}^{K-1} p(\tau_k \mid \alpha, \sigma)\right]\right\} \\
&= p(\alpha, \sigma \mid s_1, s_2, a) \prod_{k=1}^{K-1} \frac{\Gamma(1-\sigma+\alpha+k\sigma)}{\Gamma(1-\sigma)\Gamma(\alpha+k\sigma)} \\
&\quad \times \exp\left\{\sum_{k=1}^{K-1} \mathbb{E}_{q_{\tau_k}}\left[-\sigma \log \tau_k + (\alpha - 1 + k\sigma) \log(1-\tau_k)\right]\right\} \\
&= p(\alpha, \sigma \mid s_1, s_2, a) \frac{1}{\Gamma(1-\sigma)^{K-1}} \prod_{k=1}^{K-1} \frac{[\alpha + (k-1)\sigma]\,\Gamma(\alpha + (k-1)\sigma)}{\Gamma(\alpha + k\sigma)} \\
&\quad \times \exp\left\{\sum_{k=1}^{K-1} \mathbb{E}_{q_{\tau_k}}\left[-\sigma(\log \tau_k - k \log(1-\tau_k)) + (\alpha - 1)\log(1-\tau_k)\right]\right\} \\
&= p(\alpha, \sigma \mid s_1, s_2, a) \frac{1}{\Gamma(1-\sigma)^{K-1}} \prod_{k=1}^{K-1} [\alpha + (k-1)\sigma] \frac{\prod_{l=0}^{K-2} \Gamma(\alpha + l\sigma)}{\prod_{k=1}^{K-1} \Gamma(\alpha + k\sigma)} \\
&\quad \times \exp\left\{\sum_{k=1}^{K-1} \mathbb{E}_{q_{\tau_k}}\left[-\sigma(\log \tau_k - k \log(1-\tau_k)) + (\alpha - 1)\log(1-\tau_k)\right]\right\} \\
&= p(\alpha, \sigma \mid s_1, s_2, a) \frac{\Gamma(\alpha)}{\Gamma(1-\sigma)^{K-1}\Gamma(\alpha + (K-1)\sigma)} \prod_{k=1}^{K-1} [\alpha + (k-1)\sigma] \\
&\quad \times \exp\left\{-\sigma\left[\sum_{k=1}^{K-1} \mathbb{E}_{q_{\tau_k}}[\log \tau_k] - \sum_{k=1}^{K-1} k\mathbb{E}_{q_{\tau_k}}[\log(1-\tau_k)]\right] + (\alpha - 1)\sum_{k=1}^{K-1} \mathbb{E}_{q_{\tau_k}}[\log(1-\tau_k)]\right\},
\end{aligned} \tag{A14}$$

where we used the fact that $\Gamma(x+1) = x\Gamma(x)$. Except in the DP-GLLiM case, *i.e.*, $\sigma = 0$, the normalizing constant, $\left(\int \widetilde{q}_{\alpha,\sigma}(\alpha, \sigma) d(\alpha, \sigma)\right)^{-1}$, for $\widetilde{q}_{\alpha,\sigma}$ is not tractable. However, to perform VBEM in Appendix C.1, we do not need the full $q_{\alpha,\sigma}$ distribution, but only the means $\mathbb{E}_{q_{\alpha,\sigma}}[\alpha]$ and $\mathbb{E}_{q_{\alpha,\sigma}}[\sigma]$. One solution, therefore, is to use importance sampling

or MCMC to compute these expectations by means of Monte Carlo sums. Via utilizing the prior on (α, σ) given in (17), it holds that

$$
\begin{aligned}
&\widetilde{q}_{\alpha,\sigma}(\alpha,\sigma) \\
&= \frac{1}{\Gamma(s_1)} {s_2}^{s_1} (\alpha+\sigma)^{s_1-1} \exp\{-s_2(\alpha+\sigma)\} \frac{\Gamma(\alpha)p(\sigma \mid a)}{\Gamma(1-\sigma)^{K-1}\Gamma(\alpha+(K-1)\sigma)} \prod_{k=1}^{K-1} [\alpha+(k-1)\sigma] \\
&\quad \times \exp\left\{-\sigma\left[\sum_{k=1}^{K-1} \mathbb{E}_{q_{\tau_k}}[\log \tau_k] - \sum_{k=1}^{K-1} k\mathbb{E}_{q_{\tau_k}}[\log(1-\tau_k)]\right] + (\alpha-1)\sum_{k=1}^{K-1} \mathbb{E}_{q_{\tau_k}}[\log(1-\tau_k)]\right\} \\
&= \frac{1}{\Gamma(s_1)} {s_2}^{s_1} (\alpha+\sigma)^{s_1-1} \exp\left\{-\left[s_2 - \sum_{k=1}^{K-1} \mathbb{E}_{q_{\tau_k}}[\log(1-\tau_k)]\right](\alpha+\sigma)\right\} \\
&\quad \times e^{-\sigma\xi} \exp\left\{-\sum_{k=1}^{K-1} \mathbb{E}_{q_{\tau_k}}[\log(1-\tau_k)]\right\} \frac{\Gamma(\alpha)p(\sigma \mid a)}{\Gamma(1-\sigma)^{K-1}\Gamma(\alpha+(K-1)\sigma)} \prod_{k=1}^{K-1} [\alpha+(k-1)\sigma] \\
&= \frac{1}{\Gamma(s_1)} (\alpha+\sigma)^{s_1-1} \exp\left\{-\left[s_2 - \sum_{k=1}^{K-1} \mathbb{E}_{q_{\tau_k}}[\log(1-\tau_k)]\right](\alpha+\sigma)\right\} \left[s_2 - \sum_{k=1}^{K-1} \mathbb{E}_{q_{\tau_k}}[\log(1-\tau_k)]\right]^{s_1-s_1} \\
&\quad \times e^{-\sigma\xi} \exp\left\{-\sum_{k=1}^{K-1} \mathbb{E}_{q_{\tau_k}}[\log(1-\tau_k)]\right\} \frac{\Gamma(\alpha)p(\sigma \mid a)}{\Gamma(1-\sigma)^{K-1}\Gamma(\alpha+(K-1)\sigma)} \prod_{k=1}^{K-1} [\alpha+(k-1)\sigma]\, {s_2}^{s_1} \\
&\propto \mathrm{Gam}(\alpha+\sigma \mid \widehat{s}_1, \widehat{s}_2)\, e^{-\sigma\xi} \frac{\Gamma(\alpha)p(\sigma \mid a)}{\Gamma(1-\sigma)^{K-1}\Gamma(\alpha+(K-1)\sigma)} \prod_{k=1}^{K-1} \frac{\alpha+(k-1)\sigma}{\alpha+\sigma}. \qquad \text{(A15)}
\end{aligned}
$$

Here, given that $\psi(\cdot)$ is the digamma function defined by $\psi(z) = \frac{d}{dz}\log\Gamma(z) = \frac{\Gamma'(z)}{\Gamma(z)}$, we have

$$
\xi = \sum_{k=1}^{K-1} \mathbb{E}_{q_{\tau_k}}[\log \tau_k] - \sum_{k=1}^{K-1} (k-1)\mathbb{E}_{q_{\tau_k}}[\log(1-\tau_k)], \qquad \text{(A16)}
$$

$$
\mathbb{E}_{q_{\tau_k}}[\log \tau_k] = \psi(\widehat{\gamma}_{k,1}) - \psi(\widehat{\gamma}_{k,1} + \widehat{\gamma}_{k,2}), \qquad \text{(A17)}
$$

$$
\mathbb{E}_{q_{\tau_k}}[\log(1-\tau_k)] = \psi(\widehat{\gamma}_{k,2}) - \psi(\widehat{\gamma}_{k,1} + \widehat{\gamma}_{k,2}), \qquad \text{(A18)}
$$

$$
\widehat{s}_1 = s_1 + K - 1, \quad \widehat{s}_2 = s_2 - \sum_{k=1}^{K-1} [\psi(\widehat{\gamma}_{k,2}) - \psi(\widehat{\gamma}_{k,1} + \widehat{\gamma}_{k,2})]. \qquad \text{(A19)}
$$

We propose to use the following important distribution $\nu(\alpha,\sigma) = \mathrm{Gam}(\alpha+\sigma \mid \widehat{s}_1, \widehat{s}_2)\, p(\sigma \mid a)$ where $p(\sigma \mid a)$ is the uniform distribution on $[0,1]$, denoted as $\mathcal{U}_{[0,1]}(\sigma)$. Then we obtain an expression for the importance weights,

$$
W(\alpha,\sigma) = \frac{\widetilde{q}_{\alpha,\sigma}(\alpha,\sigma)}{\nu(\alpha,\sigma)} = e^{-\sigma\xi} \frac{\Gamma(\alpha)}{\Gamma(1-\sigma)^{K-1}\Gamma(\alpha+(K-1)\sigma)} \prod_{k=1}^{K-1} \frac{\alpha+(k-1)\sigma}{\alpha+\sigma}. \qquad \text{(A20)}
$$

The importance sampling scheme then consists of

- For $m \in [M]$, first simulate independently σ_m from $\mathcal{U}_{[0,1]}(\sigma)$ and then simulate conditionally α_m with the σ_m-shifted gamma $\mathcal{G}(\sigma_m, \widehat{s}_1, \widehat{s}_2)$. This later simulation

is easily obtained by simulating a standard $\gamma\left(\alpha'_m \mid \widehat{s}_1, \widehat{s}_2\right)$ and then subtracting σ_m from the result.

- Compute the importance weights $w_m = W(\alpha_m, \alpha_m)$.
- Approximate the means

$$\mathbb{E}_{q_{\alpha,\sigma}}[\alpha] \simeq \sum_{m=1}^{M} \frac{w_m}{\sum_{i=1}^{M} w_i} \alpha_m, \quad \mathbb{E}_{q_{\alpha,\sigma}}[\sigma] \simeq \sum_{m=1}^{M} \frac{w_m}{\sum_{i=1}^{M} w_i} \sigma_m. \tag{A21}$$

Note that this complication is due to the PY. In the DP-GLLiM case, by substituting $\sigma = 0$ in (A15), the E step is much simpler, as it reduces to computing the approximate posterior expectation of α, namely

$$\mathbb{E}_{q_{\alpha,0}}[\alpha] = \frac{\widehat{s}_1}{\widehat{s}_2}, \quad q_{\alpha,0} = \mathrm{Gam}\left(\alpha \mid \widehat{s}_1, \widehat{s}_2\right). \tag{A22}$$

A.4. VB-E-Z step from Section 3.1

In some situations, it is useful to use a 1-of-K binary vector $\mathbf{z}_n$ to represent the latent variable z_n for each observation $(\mathbf{y}_n, \mathbf{x}_n)$. To be more precise, we introduce a K-dimensional binary random variable $\mathbf{z}_n = (z_{nk})_{k\in[K]}, K \leq \infty$, with a 1-of-$K$ representation in which a particular element z_{nk} is equal to 1, *i.e.*, $z_n = k$, and all other elements are equal to 0. The values of z_{nk} thus satisfy $z_{nk} \in \{0, 1\}$ and $\sum_{k\in[K]} z_{nk} = 1, \forall n \in \mathbb{N}^\star$. If there is no confusion, we also denote $\mathcal{Z}$ as the latent matrix $\mathcal{Z} = (z_{nk})_{n\in[N],k\in[K]}$. It is worth to mentioning that when using a 1-of-K representation of $\mathbf{z}_n$, we can also write down marginal the conditional distributions of $\mathcal{X}$ and $\mathcal{Y} \mid \mathcal{X}$, corresponding to (24) and (25), respectively, in the form

$$p\left(\mathcal{X} \mid \mathcal{Z}, \boldsymbol{c}, \boldsymbol{\Gamma}\right) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}_L\left(\mathbf{x}_n \mid \boldsymbol{c}_k, \boldsymbol{\Gamma}_k\right)^{z_{nk}}, \tag{A23}$$

$$p\left(\mathbf{x}_n \mid \boldsymbol{c}, \boldsymbol{\Gamma}\right) = \sum_{k=1}^{K} p_{z_n}(k) \mathcal{N}_L\left(\mathbf{x}_n \mid \boldsymbol{c}_k, \boldsymbol{\Gamma}_k\right), \quad p_{z_n}(k) \equiv p(z_n = k) = \pi_k(\boldsymbol{\tau}), \tag{A24}$$

$$p\left(\mathcal{Y} \mid \mathcal{X}, \mathcal{Z}; \mathbf{A}, \mathbf{b}, \boldsymbol{\Sigma}\right) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}_D\left(\mathbf{y}_n \mid \mathbf{A}_k \mathbf{x}_n + \mathbf{b}_k, \boldsymbol{\Sigma}_k\right)^{z_{nk}}, \tag{A25}$$

$$p\left(\mathbf{y}_n \mid \mathbf{x}_n; \mathbf{A}, \mathbf{b}, \boldsymbol{\Sigma}\right) = \sum_{k=1}^{K} \frac{p_{z_n}(k) \mathcal{N}_L\left(\mathbf{x}_n \mid \boldsymbol{c}_k, \boldsymbol{\Gamma}_k\right)}{\sum_{l=1}^{K} p_{z_n}(l) \mathcal{N}_L\left(\mathbf{x}_n \mid \boldsymbol{c}_l, \boldsymbol{\Gamma}_l\right)} \mathcal{N}_D\left(\mathbf{y}_n \mid \mathbf{A}_k \mathbf{x}_n + \mathbf{b}_k, \boldsymbol{\Sigma}_k\right). \tag{A26}$$

The observations $\mathcal{X}$ and $\mathcal{Y}$ are therefore iid and generated from the same GMM (A24) and infinite GLLiM (A25), respectively. Similarly, (23) can be written down in the form

$$p\left(\mathcal{Z} \mid \boldsymbol{\tau}\right) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k\left(\boldsymbol{\tau}\right)^{z_{nk}}. \tag{A27}$$

By using the decomposition (31), the representation (A23), (A25) and absorbing any terms that are independent on $\mathcal{Z}$ into the additive normalization constant, we obtain

$$
\begin{aligned}
&\log q_{\mathcal{Z}}(\mathcal{Z}) \equiv \log q_{\mathcal{Z}}^{(r)}(\mathcal{Z}) \\
&= \mathbb{E}_{q_\Theta}\left[\log p\left(\mathcal{Y} \mid \mathcal{X}, \mathcal{Z}; \widehat{\mathbf{A}}, \widehat{\mathbf{b}}, \widehat{\mathbf{\Sigma}}\right) p\left(\mathcal{X} \mid \mathcal{Z}; \widehat{\mathbf{c}}, \widehat{\mathbf{\Gamma}}\right) p\left(\mathcal{Z} \mid \boldsymbol{\tau}\right)\right] + \text{constant} \\
&\propto \mathbb{E}_{q_\Theta}\left[\log p\left(\mathcal{Y} \mid \mathcal{X}, \mathcal{Z}; \widehat{\mathbf{A}}, \widehat{\mathbf{b}}, \widehat{\mathbf{\Sigma}}\right)\right] + \mathbb{E}_{q_\Theta}\left[\log p\left(\mathcal{X} \mid \mathcal{Z}; \widehat{\mathbf{c}}, \widehat{\mathbf{\Gamma}}\right)\right] + \mathbb{E}_{q_\Theta}\left[\log p\left(\mathcal{Z} \mid \boldsymbol{\tau}\right)\right] \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\mathbb{E}_{q_\Theta}\left[\log \mathcal{N}_D\left(\mathbf{y}_n \mid \widehat{\mathbf{A}}_k\mathbf{x}_n + \widehat{\mathbf{b}}_k, \widehat{\mathbf{\Sigma}}_k\right)\right] + \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\mathbb{E}_{q_\Theta}\left[\log \mathcal{N}_L\left(\mathbf{x}_n \mid \widehat{\mathbf{c}}_k, \widehat{\mathbf{\Gamma}}_k\right)\right] \\
&\quad + \mathbb{E}_{q_\Theta}\left[\sum_{n=1}^{N}\log\left(\pi_{z_n}\left(\boldsymbol{\tau}\right)\right)\right] \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\log \mathcal{N}_D\left(\mathbf{y}_n \mid \widehat{\mathbf{A}}_k\mathbf{x}_n + \widehat{\mathbf{b}}_k, \widehat{\mathbf{\Sigma}}_k\right) + \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\log \mathcal{N}_L\left(\mathbf{x}_n \mid \widehat{\mathbf{c}}_k, \widehat{\mathbf{\Gamma}}_k\right) \\
&\quad + \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\mathbb{E}_{q_\tau}\left[\log\left(\pi_k\left(\boldsymbol{\tau}\right)\right)\right] = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\log \rho_{nk}.
\end{aligned}
\tag{A28}
$$

Here, we used the fact that

$$
\begin{aligned}
\log \rho_{nk} &= \log \mathcal{N}_D\left(\mathbf{y}_n \mid \widehat{\mathbf{A}}_k\mathbf{x}_n + \widehat{\mathbf{b}}_k, \widehat{\mathbf{\Sigma}}_k\right) + \log \mathcal{N}_L\left(\mathbf{x}_n \mid \widehat{\mathbf{c}}_k, \widehat{\mathbf{\Gamma}}_k\right) + \mathbb{E}_{q_\tau}\left[\log\left(\pi_k\left(\boldsymbol{\tau}\right)\right)\right] \\
&= -\frac{D}{2}\log(2\pi) - \frac{1}{2}\log\left|\widehat{\mathbf{\Sigma}}_k\right| - \frac{1}{2}\left(\mathbf{y}_n - \widehat{\mathbf{A}}_k\mathbf{x}_n - \widehat{\mathbf{b}}_k\right)^\top \widehat{\mathbf{\Sigma}}_k^{-1}\left(\mathbf{y}_n - \widehat{\mathbf{A}}_k\mathbf{x}_n - \widehat{\mathbf{b}}_k\right) \\
&\quad - \frac{L}{2}\log(2\pi) - \frac{1}{2}\log\left|\widehat{\mathbf{\Gamma}}_k\right| - \frac{1}{2}\left(\mathbf{x}_n - \widehat{\mathbf{c}}_k\right)^\top \widehat{\mathbf{\Gamma}}_k^{-1}\left(\mathbf{x}_n - \widehat{\mathbf{c}}_k\right) \\
&\quad + \psi(\widehat{\gamma}_{k,1}) - \psi(\widehat{\gamma}_{k,1} + \widehat{\gamma}_{k,2}) + \sum_{l=1}^{k-1}\psi(\widehat{\gamma}_{l,2}) - \psi(\widehat{\gamma}_{l,1} + \widehat{\gamma}_{l,2}).
\end{aligned}
\tag{A29}
$$

By taking exponential of both sides and taking into account the normalized constant, it holds that

$$
q_{\mathcal{Z}}(\mathcal{Z}) = \frac{1}{\sum_{l=1}^{K}\rho_{nl}}\prod_{n=1}^{N}\prod_{k=1}^{K}\rho_{nk}^{z_{nk}} = \prod_{n=1}^{N}\prod_{k=1}^{K} r_{nk}^{z_{nk}}, \quad r_{nk} = \frac{\rho_{nk}}{\sum_{l=1}^{K}\rho_{nl}}. \tag{A30}
$$

Note also that $z_n = k$ if and only if the latent matrix $\mathcal{Z}$ reduces to a sparse matrix $\mathcal{Z}_{nk}$ which has only one position different from 0, namely $z_{nk} = 1$. This leads to the following simplified notation:

$$
\log q_{z_n}(k) \equiv \log q_{z_n}(z_n = k) \equiv \log q_{\mathcal{Z}}(\mathcal{Z}_{nk}) = r_{nk}. \tag{A31}
$$

A.4.1. Updating $\mathbf{\Sigma}_k$

By using matrix derivatives, the derivative of the log likelihood with respect to $\mathbf{\Sigma}_k^{-1}$ is given by

$$
\begin{aligned}
&\frac{\partial}{\partial \mathbf{\Sigma}_k^{-1}} f_1\left(\widehat{\mathbf{A}}_k, \mathbf{b}_k, \mathbf{\Sigma}_k^{-1}\right) \\
&= -\frac{1}{2}\sum_{n=1}^{N} q_{z_n}(k) \frac{\partial}{\partial \mathbf{\Sigma}_k^{-1}}\left[-\log\left|\mathbf{\Sigma}_k^{-1}\right| + \operatorname{Tr}\left[\mathbf{\Sigma}_k^{-1\top}\left(\mathbf{y}_n - \mathbf{A}_k\mathbf{x}_n - \mathbf{b}_k\right)\left(\mathbf{y}_n - \mathbf{A}_k\mathbf{x}_n - \mathbf{b}_k\right)^\top\right]\right] \\
&= -\frac{1}{2}\sum_{n=1}^{N} q_{z_n}(k)\left[-\mathbf{\Sigma}_k + \left(\mathbf{y}_n - \mathbf{A}_k\mathbf{x}_n - \mathbf{b}_k\right)\left(\mathbf{y}_n - \mathbf{A}_k\mathbf{x}_n - \mathbf{b}_k\right)^\top\right] \\
&= \frac{N_k}{2}\mathbf{\Sigma}_k - \frac{1}{2}\sum_{n=1}^{N} q_{z_n}(k)\left(\mathbf{y}_n - \mathbf{A}_k\mathbf{x}_n - \mathbf{b}_k\right)\left(\mathbf{y}_n - \mathbf{A}_k\mathbf{x}_n - \mathbf{b}_k\right)^\top . \qquad \text{(A32)}
\end{aligned}
$$

Finally, setting to zero yields

$$
\widehat{\mathbf{\Sigma}}_k = \frac{1}{N_k}\sum_{n=1}^{N} q_{z_n}(k)\left(\mathbf{y}_n - \widehat{\mathbf{A}}_k\mathbf{x}_n - \widehat{\mathbf{b}}_k\right)\left(\mathbf{y}_n - \widehat{\mathbf{A}}_k\mathbf{x}_n - \widehat{\mathbf{b}}_k\right)^\top . \qquad \text{(A33)}
$$

A.5. VB-M-$(\mathbf{c}, \mathbf{\Gamma})$ step from Section 3.2

This step divides into K sub-steps that involve the following optimizations

$$
\left(\widehat{\mathbf{c}}_k, \widehat{\mathbf{\Gamma}}_k\right) \equiv \left(\widehat{\mathbf{c}}_k^{(r)}, \widehat{\mathbf{\Gamma}}_k^{(r)}\right) = \underset{(\mathbf{c}_k, \mathbf{\Gamma}_k)}{\operatorname{argmax}}\, \mathbb{E}_{q_{\mathcal{Z}}^{(r)}}\left[\log p\left(\mathcal{X} \mid \mathcal{Z}; \mathbf{c}_k, \mathbf{\Gamma}_k\right)\right] . \qquad \text{(A34)}
$$

By definition, we have

$$
\begin{aligned}
&\mathbb{E}_{q_{\mathcal{Z}}^{(r)}}\left[\log\left(p\left(\mathcal{X} \mid \mathcal{Z}; \mathbf{c}_k, \mathbf{\Gamma}_k\right)\right)\right] \\
&= \mathbb{E}_{q_{\mathcal{Z}}^{(r)}}\left[\log \prod_{n=1}^{N} \mathcal{N}_L\left(\mathbf{x}_n \mid \boldsymbol{c}_k, \mathbf{\Gamma}_k\right)^{z_{nk}}\right] \\
&= \sum_{n=1}^{N} \mathbb{E}_{q_{\mathcal{Z}}^{(r)}}\left[z_{nk} \log \mathcal{N}_L\left(\mathbf{x}_n \mid \boldsymbol{c}_k, \mathbf{\Gamma}_k\right)\right] \\
&= \sum_{n=1}^{N} \mathbb{E}_{q_{\mathcal{Z}}^{(r)}}\left[z_{nk}\right] \log \mathcal{N}_L\left(\mathbf{x}_n \mid \boldsymbol{c}_k, \mathbf{\Gamma}_k\right) \\
&= \sum_{n=1}^{N} q_{z_n}(k) \log \mathcal{N}_L\left(\mathbf{x}_n \mid \boldsymbol{c}_k, \mathbf{\Gamma}_k\right) \\
&= \sum_{n=1}^{N} q_{z_n}(k)\left[-\frac{L}{2}\log(2\pi) - \frac{1}{2}\log\left|\mathbf{\Gamma}_k\right| - \frac{1}{2}\left(\mathbf{x}_n - \mathbf{c}_k\right)^\top \mathbf{\Gamma}_k^{-1}\left(\mathbf{x}_n - \mathbf{c}_k\right)\right] \\
&\equiv f_2\left(\mathbf{c}_k, \mathbf{\Gamma}_k\right) . \qquad \text{(A35)}
\end{aligned}
$$

We aim to solve the following optimization

$$\left(\widehat{\mathbf{c}}_k, \widehat{\mathbf{\Gamma}}_k\right) = \underset{(\mathbf{c}_k, \mathbf{\Gamma}_k)}{\operatorname{argmax}} f_2\left(\mathbf{c}_k, \mathbf{\Gamma}_k\right). \tag{A36}$$

Similarly with Appendices A.4.1 and A.6.1, we obtain the following update:

$$\begin{aligned}
\widehat{\mathbf{c}}_k &= \frac{1}{N_k} \sum_{n=1}^{N} q_{z_n}(k) \mathbf{x}_n, \\
\widehat{\mathbf{\Gamma}}_k &= \frac{1}{N_k} \sum_{n=1}^{N} q_{z_n}(k) \left(\mathbf{x}_n - \widehat{\mathbf{c}}_k\right) \left(\mathbf{x}_n - \widehat{\mathbf{c}}_k\right)^\top.
\end{aligned} \tag{A37}$$

A.6. VB-M-$(\mathbf{A}, \mathbf{b}, \mathbf{\Sigma})$ step from Section 3.2

By definition, we have

$$\begin{aligned}
&\mathbb{E}_{q_{\mathcal{Z}}^{(r)}}\left[\log p\left(\mathcal{Y} \mid \mathcal{X}, \mathcal{Z}; \mathbf{A}_k, \mathbf{b}_k, \mathbf{\Sigma}_k\right)\right] \\
&= \mathbb{E}_{q_{\mathcal{Z}}^{(r)}}\left[\log \prod_{n=1}^{N} \mathcal{N}_D\left(\mathbf{y}_n \mid \mathbf{A}_k \mathbf{x}_n + \mathbf{b}_k, \mathbf{\Sigma}_k\right)^{z_{nk}}\right] \\
&= \sum_{n=1}^{N} \mathbb{E}_{q_{\mathcal{Z}}^{(r)}}\left[z_{nk} \log \mathcal{N}_D\left(\mathbf{y}_n \mid \mathbf{A}_k \mathbf{x}_n + \mathbf{b}_k, \mathbf{\Sigma}_k\right)\right] \\
&= \sum_{n=1}^{N} \mathbb{E}_{q_{\mathcal{Z}}^{(r)}}\left[z_{nk}\right] \log \mathcal{N}_D\left(\mathbf{y}_n \mid \mathbf{A}_k \mathbf{x}_n + \mathbf{b}_k, \mathbf{\Sigma}_k\right) \\
&= \sum_{n=1}^{N} q_{z_n}(k) \log \mathcal{N}_D\left(\mathbf{y}_n \mid \mathbf{A}_k \mathbf{x}_n + \mathbf{b}_k, \mathbf{\Sigma}_k\right) \\
&= \sum_{n=1}^{N} q_{z_n}(k) \left[-\frac{D}{2}\log(2\pi) - \frac{1}{2}\log\left|\mathbf{\Sigma}_k\right| - \frac{1}{2}\left(\mathbf{y}_n - \mathbf{A}_k \mathbf{x}_n - \mathbf{b}_k\right)^\top \mathbf{\Sigma}_k^{-1}\left(\mathbf{y}_n - \mathbf{A}_k \mathbf{x}_n - \mathbf{b}_k\right)\right] \\
&\equiv f_1\left(\mathbf{A}_k, \mathbf{b}_k, \mathbf{\Sigma}_k\right).
\end{aligned} \tag{A38}$$

We aim to solve the following optimization

$$\left(\widehat{\mathbf{A}}_k, \widehat{\mathbf{b}}_k, \widehat{\mathbf{\Sigma}}_k\right) = \underset{(\mathbf{A}_k, \mathbf{b}_k, \mathbf{\Sigma}_k)}{\operatorname{argmax}} f_1\left(\mathbf{A}_k, \mathbf{b}_k, \mathbf{\Sigma}_k\right). \tag{A39}$$

A.6.1. Updating $\mathbf{b}_k$

The derivative of the log likelihood with respect to $\mathbf{b}_k$ is given by

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{b}_k} f_1\left(\mathbf{A}_k, \mathbf{b}_k, \boldsymbol{\Sigma}_k\right) &= -\frac{1}{2}\sum_{n=1}^{N} q_{z_n}(k)\frac{\partial}{\partial \mathbf{b}_k}\left[\left(\mathbf{y}_n - \mathbf{A}_k\mathbf{x}_n - \mathbf{b}_k\right)^\top \boldsymbol{\Sigma}_k^{-1}\left(\mathbf{y}_n - \mathbf{A}_k\mathbf{x}_n - \mathbf{b}_k\right)\right] \\
&= -\sum_{n=1}^{N} q_{z_n}(k)\boldsymbol{\Sigma}_k^{-1}\left(\mathbf{y}_n - \mathbf{A}_k\mathbf{x}_n - \mathbf{b}_k\right)\frac{\partial}{\partial \mathbf{b}_k}\left(\mathbf{y}_n - \mathbf{A}_k\mathbf{x}_n - \mathbf{b}_k\right) \\
&= \sum_{n=1}^{N} q_{z_n}(k)\boldsymbol{\Sigma}_k^{-1}\left(\mathbf{y}_n - \mathbf{A}_k\mathbf{x}_n - \mathbf{b}_k\right).
\end{aligned}$$

Setting this derivative to zero, we obtain the solution for VB-M-$\mathbf{b}$ step given by

$$\begin{aligned}
&\sum_{n=1}^{N} q_{z_n}(k)\boldsymbol{\Sigma}_k^{-1}\left(\mathbf{y}_n - \mathbf{A}_k\mathbf{x}_n - \mathbf{b}_k\right) = 0 \\
&\Leftrightarrow \sum_{n=1}^{N} q_{z_n}(k)\boldsymbol{\Sigma}_k^{-1}\left(\mathbf{y}_n - \mathbf{A}_k\mathbf{x}_n\right) - \sum_{n=1}^{N} q_{z_n}(k)\boldsymbol{\Sigma}_k^{-1}\mathbf{b}_k = 0 \\
&\Leftrightarrow \sum_{n=1}^{N} q_{z_n}(k)\left(\mathbf{y}_n - \mathbf{A}_k\mathbf{x}_n\right) - \sum_{n=1}^{N} q_{z_n}(k)\mathbf{b}_k = 0 \,(\text{left multiplying by } \boldsymbol{\Sigma}_k) \\
&\Leftrightarrow \mathbf{b}_k = \frac{1}{\sum_{n=1}^{N} q_{z_n}(k)}\sum_{n=1}^{N} q_{z_n}(k)\left(\mathbf{y}_n - \mathbf{A}_k\mathbf{x}_n\right) \equiv \frac{1}{N_k}\sum_{n=1}^{N} q_{z_n}(k)\left(\mathbf{y}_n - \mathbf{A}_k\mathbf{x}_n\right). \quad \text{(A40)}
\end{aligned}$$

A.6.2. Updating $\mathbf{A}_k$

The derivative of the log likelihood with respect to $\mathbf{A}_k$ is given by

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{A}_k} f_1\left(\mathbf{A}_k, \mathbf{b}_k, \boldsymbol{\Sigma}_k\right) &= -\frac{1}{2}\sum_{n=1}^{N} q_{z_n}(k)\frac{\partial}{\partial \mathbf{A}_k}\left[\left(\mathbf{y}_n - \mathbf{A}_k\mathbf{x}_n - \mathbf{b}_k\right)^\top \boldsymbol{\Sigma}_k^{-1}\left(\mathbf{y}_n - \mathbf{A}_k\mathbf{x}_n - \mathbf{b}_k\right)\right] \\
&= -\sum_{n=1}^{N} q_{z_n}(k)\boldsymbol{\Sigma}_k^{-1}\left(\mathbf{y}_n - \mathbf{A}_k\mathbf{x}_n - \mathbf{b}_k\right)\frac{\partial}{\partial \mathbf{A}_k}\left(\mathbf{y}_n - \mathbf{A}_k\mathbf{x}_n - \mathbf{b}_k\right) \\
&= \sum_{n=1}^{N} q_{z_n}(k)\boldsymbol{\Sigma}_k^{-1}\left(\mathbf{y}_n - \mathbf{A}_k\mathbf{x}_n - \mathbf{b}_k\right)\mathbf{x}_n^\top.
\end{aligned}$$

Then, we set this derivative w.r.t. $\mathbf{A}_k$ equal to zero, giving

$$
\begin{aligned}
&\sum_{n=1}^{N} q_{z_n}(k)\boldsymbol{\Sigma}_k^{-1}\left(\mathbf{y}_n - \mathbf{A}_k\mathbf{x}_n - \mathbf{b}_k\right)\mathbf{x}_n^\top = 0 \\
&\Leftrightarrow \sum_{n=1}^{N} q_{z_n}(k)\left(\mathbf{y}_n - \mathbf{A}_k\mathbf{x}_n - \mathbf{b}_k\right)\mathbf{x}_n^\top = 0 \,(\text{left multiplying by } \boldsymbol{\Sigma}_k) \\
&\Leftrightarrow \sum_{n=1}^{N} q_{z_n}(k)\mathbf{y}_n\mathbf{x}_n^\top - \sum_{n=1}^{N} q_{z_n}(k)\mathbf{A}_k\mathbf{x}_n\mathbf{x}_n^\top - \sum_{n=1}^{N} q_{z_n}(k)\mathbf{b}_k\mathbf{x}_n^\top = 0 \\
&\Leftrightarrow \sum_{n=1}^{N} q_{z_n}(k)\mathbf{y}_n\mathbf{x}_n^\top - \mathbf{A}_k\sum_{n=1}^{N} q_{z_n}(k)\mathbf{x}_n\mathbf{x}_n^\top - \mathbf{b}_k\sum_{n=1}^{N} q_{z_n}(k)\mathbf{x}_n^\top = 0 \\
&\Leftrightarrow \mathbf{A}_k\sum_{n=1}^{N} q_{z_n}(k)\mathbf{x}_n\mathbf{x}_n^\top \\
&\quad = \sum_{n=1}^{N} q_{z_n}(k)\mathbf{y}_n\mathbf{x}_n^\top - \frac{1}{N_k}\sum_{n=1}^{N} q_{z_n}(k)\left(\mathbf{y}_n - \mathbf{A}_k\mathbf{x}_n\right)\sum_{n=1}^{N} q_{z_n}(k)\mathbf{x}_n^\top \,(\text{using (A40) for } \mathbf{b}_k) \\
&\Leftrightarrow \mathbf{A}_k\sum_{n=1}^{N} q_{z_n}(k)\mathbf{x}_n\left(\mathbf{x}_n^\top - \frac{1}{N_k}\sum_{n=1}^{N} q_{z_n}(k)\mathbf{x}_n^\top\right) = \sum_{n=1}^{N} q_{z_n}(k)\left(\mathbf{y}_n - \frac{1}{N_k}\sum_{n=1}^{N} q_{z_n}(k)\mathbf{y}_n\right)\mathbf{x}_n^\top \\
&\Leftrightarrow N_k\mathbf{A}_k\boldsymbol{X}_k\boldsymbol{X}_k^\top = N_k\boldsymbol{Y}_k\boldsymbol{X}_k^\top \Leftrightarrow \mathbf{A}_k = \boldsymbol{Y}_k\boldsymbol{X}_k^\top\left(\boldsymbol{X}_k\boldsymbol{X}_k^\top\right)^{-1}.
\end{aligned}
$$

Here, the last equality is obtained by firstly define the following quantities,

$$
\begin{aligned}
\bar{\mathbf{x}}_k &= \frac{1}{N_k}\sum_{n=1}^{N} q_{z_n}(k)\mathbf{x}_n, \\
\bar{\mathbf{y}}_k &= \frac{1}{N_k}\sum_{n=1}^{N} q_{z_n}(k)\mathbf{y}_n, \\
\boldsymbol{X}_k &= \frac{1}{\sqrt{N_k}}\left(\sqrt{q_{z_1}(k)}(\mathbf{x}_1 - \bar{\mathbf{x}}_k), \ldots, \sqrt{q_{z_N}(k)}(\mathbf{x}_N - \bar{\mathbf{x}}_k)\right)_{L\times N}, \\
\boldsymbol{Y}_k &= \frac{1}{\sqrt{N_k}}\left(\sqrt{q_{z_1}(k)}(\mathbf{y}_1 - \bar{\mathbf{y}}_k), \ldots, \sqrt{q_{z_N}(k)}(\mathbf{y}_N - \bar{\mathbf{y}}_k)\right)_{D\times N}.
\end{aligned}
$$

Then, we used the fact that

$$
\begin{aligned}
&\sum_{n=1}^{N} q_{z_n}(k)\mathbf{x}_n \left(\mathbf{x}_n^\top - \frac{1}{N_k}\sum_{n=1}^{N} q_{z_n}(k)\mathbf{x}_n^\top\right) \\
&= \sum_{n=1}^{N} q_{z_n}(k)\mathbf{x}_n \left(\mathbf{x}_n^\top - \bar{\mathbf{x}}_k^\top\right) \\
&= \sum_{n=1}^{N} q_{z_n}(k)\left(\mathbf{x}_n - \bar{\mathbf{x}}_k\right)\left(\mathbf{x}_n - \bar{\mathbf{x}}_k\right)^\top + \sum_{n=1}^{N} q_{z_n}(k)\bar{\mathbf{x}}_k\left(\mathbf{x}_n - \bar{\mathbf{x}}_k\right)^\top \\
&= \sum_{n=1}^{N} q_{z_n}(k)\left(\mathbf{x}_n - \bar{\mathbf{x}}_k\right)\left(\mathbf{x}_n - \bar{\mathbf{x}}_k\right)^\top \\
&= N_k \boldsymbol{X}_k \boldsymbol{X}_k^\top,
\end{aligned}
\tag{A41}
$$

and

$$
\begin{aligned}
&\sum_{n=1}^{N} q_{z_n}(k)\left(\mathbf{y}_n - \frac{1}{N_k}\sum_{n=1}^{N} q_{z_n}(k)\mathbf{y}_n\right)\mathbf{x}_n^\top \\
&= \sum_{n=1}^{N} q_{z_n}(k)\left(\mathbf{y}_n - \bar{\mathbf{y}}_k\right)\mathbf{x}_n^\top \\
&= \sum_{n=1}^{N} q_{z_n}(k)\left(\mathbf{y}_n - \bar{\mathbf{y}}_k\right)\left(\mathbf{x}_n - \bar{\mathbf{x}}_k\right)^\top + \sum_{n=1}^{N} q_{z_n}(k)\left(\mathbf{y}_n - \bar{\mathbf{y}}_k\right)\bar{\mathbf{x}}_k^\top \\
&= \sum_{n=1}^{N} q_{z_n}(k)\left(\mathbf{y}_n - \bar{\mathbf{y}}_k\right)\left(\mathbf{x}_n - \bar{\mathbf{x}}_k\right)^\top \\
&= N_k \boldsymbol{Y}_k \boldsymbol{X}_k^\top.
\end{aligned}
\tag{A42}
$$

Here, we used the fact that

$$
\begin{aligned}
&\sum_{n=1}^{N} q_{z_n}(k)\bar{\mathbf{x}}_k\left(\mathbf{x}_n - \bar{\mathbf{x}}_k\right)^\top = \sum_{n=1}^{N} q_{z_n}(k)\bar{\mathbf{x}}_k\mathbf{x}_n^\top - \bar{\mathbf{x}}_k N_k \bar{\mathbf{x}}_k^\top = \sum_{n=1}^{N} q_{z_n}(k)\bar{\mathbf{x}}_k\mathbf{x}_n^\top - \sum_{n=1}^{N} q_{z_n}(k)\bar{\mathbf{x}}_k\mathbf{x}_n^\top = 0, \\
&\sum_{n=1}^{N} q_{z_n}(k)\left(\mathbf{y}_n - \bar{\mathbf{y}}_k\right)\bar{\mathbf{x}}_k^\top = \sum_{n=1}^{N} q_{z_n}(k)\mathbf{y}_n\bar{\mathbf{x}}_k^\top - N_k\bar{\mathbf{y}}_k\bar{\mathbf{x}}_k^\top = \sum_{n=1}^{N} q_{z_n}(k)\mathbf{y}_n\bar{\mathbf{x}}_k^\top - \sum_{n=1}^{N} q_{z_n}(k)\mathbf{y}_n\bar{\mathbf{x}}_k^\top = 0,
\end{aligned}
\tag{A43}
$$

and for each $i, j \in [L]$, it holds that

$$\begin{aligned}
&\left[\sum_{n=1}^{N} q_{z_n}(k)\left(\mathbf{x}_n-\bar{\mathbf{x}}_k\right)\left(\mathbf{x}_n-\bar{\mathbf{x}}_k\right)^{\top}\right]_{ij}=\sum_{n=1}^{N} q_{z_n}(k)\left[\left(\mathbf{x}_n-\bar{\mathbf{x}}_k\right)\left(\mathbf{x}_n-\bar{\mathbf{x}}_k\right)^{\top}\right]_{ij} \\
&=\sum_{n=1}^{N} q_{z_n}(k)\left[\left(\mathbf{x}_n-\bar{\mathbf{x}}_k\right)\right]_{i1}\left[\left(\mathbf{x}_n-\bar{\mathbf{x}}_k\right)^{\top}\right]_{1j} \\
&\equiv\sum_{n=1}^{N} q_{z_n}(k)\left[\left(\mathbf{x}_n-\bar{\mathbf{x}}_k\right)\right]_{i}\left[\left(\mathbf{x}_n-\bar{\mathbf{x}}_k\right)\right]_{j} \\
&=\left[\left(\sqrt{q_{z_1}(k)}(\mathbf{x}_1-\bar{\mathbf{x}}_k),\ldots,\sqrt{q_{z_N}(k)}(\mathbf{x}_N-\bar{\mathbf{x}}_k)\right)\right]_{i\cdot}\left[\left(\sqrt{q_{z_1}(k)}(\mathbf{x}_1-\bar{\mathbf{x}}_k),\ldots,\sqrt{q_{z_N}(k)}(\mathbf{x}_N-\bar{\mathbf{x}}_k)\right)\right]_{\cdot j}^{\top} \\
&=N_k\left[\boldsymbol{X}_k\boldsymbol{X}_k^{\top}\right]_{ij}. \qquad \text{(A44)}
\end{aligned}$$

Appendix B. Technical proofs

B.1. Proof of Lemma 5.2

For the sake of completeness, we highlighted the main techniques using for the proof of Lemma 5.2 in this paper, see [18, Proposition 1] for more details. We first want to prove (69). Using the partition of a joint Gaussian with $\mathbf{x}_b \equiv \mathbf{x}, \boldsymbol{\mu}_b = \boldsymbol{\mu}_k^{\mathbf{x}}, \boldsymbol{\Sigma}_{bb} \equiv \mathbf{V}_k^{\mathbf{xx}}, \mathbf{x}_a \equiv \mathbf{y}, \boldsymbol{\mu}_a \equiv \boldsymbol{\mu}_k^{\mathbf{y}}, \boldsymbol{\Sigma}_{aa} \equiv \mathbf{V}_k^{\mathbf{yy}}$, we obtain

$$\begin{aligned}
p\left(\mathbf{x}_a \mid \mathbf{x}_b\right)=\mathcal{N}\left(\mathbf{x}_a \mid \boldsymbol{\mu}_{a|b}, \boldsymbol{\Gamma}_{aa}^{-1}\right), \quad \boldsymbol{\mu}_{a|b}=\boldsymbol{\mu}_a-\boldsymbol{\Gamma}_{aa}^{-1}\boldsymbol{\Gamma}_{ab}\left(\boldsymbol{x}_b-\boldsymbol{\mu}_b\right)=\boldsymbol{\mu}_a-\boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\left(\boldsymbol{x}_b-\boldsymbol{\mu}_b\right), \\
p\left(\boldsymbol{x}_b\right)=\mathcal{N}\left(\boldsymbol{x}_b \mid \boldsymbol{\mu}_b, \boldsymbol{\Sigma}_{bb}\right). \qquad \text{(B1)}
\end{aligned}$$

Recall that

$$\begin{aligned}
p\left(\mathbf{y} \mid \mathbf{x}, Z=k; \boldsymbol{\psi}\right)=\mathcal{N}_D\left(\mathbf{y} \mid \mathbf{A}_k\mathbf{x}+\mathbf{b}_k, \boldsymbol{\Sigma}_k\right), \\
p\left(\mathbf{x} \mid Z=k; \boldsymbol{\psi}\right)=\mathcal{N}_L\left(\mathbf{x} \mid \mathbf{c}_k, \boldsymbol{\Gamma}_k\right), p\left(Z=k; \boldsymbol{\psi}\right)=\pi_k. \qquad \text{(B2)}
\end{aligned}$$

By identifying the parameters of (B1) and (B2), it holds that

$$\begin{aligned}
\pi_k &= \rho_k \\
\mathbf{c}_k &= \boldsymbol{\mu}_k^{\mathbf{x}}, \\
\boldsymbol{\Gamma}_k &= \mathbf{V}_k^{\mathbf{xx}}, \\
\mathbf{A}_k &= -\boldsymbol{\Gamma}_{aa}^{-1}\boldsymbol{\Gamma}_{ab}=\mathbf{V}_k^{\mathbf{xy}^{\top}}\left(\mathbf{V}_k^{\mathbf{xx}}\right)^{-1}, \\
\mathbf{b}_k &= \boldsymbol{\mu}_a+\boldsymbol{\Gamma}_{aa}^{-1}\boldsymbol{\Gamma}_{ab}\boldsymbol{\mu}_b=\boldsymbol{\mu}_k^{\mathbf{y}}-\mathbf{V}_k^{\mathbf{xy}^{\top}}\left(\mathbf{V}_k^{\mathbf{xx}}\right)^{-1}\boldsymbol{\mu}_k^{\mathbf{x}}, \qquad \text{(B3)} \\
\boldsymbol{\Sigma}_k &= \boldsymbol{\Gamma}_{aa}^{-1}=\boldsymbol{\Sigma}_{aa}-\boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}=\mathbf{V}_k^{\mathbf{yy}}-\mathbf{V}_k^{\mathbf{xy}^{\top}}\left(\mathbf{V}_k^{\mathbf{xx}}\right)^{-1}\mathbf{V}_k^{\mathbf{xy}}. \qquad \text{(B4)}
\end{aligned}$$

The following decomposition of the joint probability distribution will be used:

$$
\begin{aligned}
p\left(\mathbf{w} \mid \boldsymbol{\psi}\right) &= \sum_{k=1}^{K} p\left(\mathbf{y} \mid \mathbf{x}, Z=k; \boldsymbol{\psi}\right) p\left(\mathbf{X}=\mathbf{x} \mid Z=k; \boldsymbol{\psi}\right) p\left(Z=k; \boldsymbol{\psi}\right) \\
&= \sum_{k=1}^{K} \pi_k \mathcal{N}_D\left(\mathbf{y} \mid \mathbf{A}_k \mathbf{x}+\mathbf{b}_k, \boldsymbol{\Sigma}_k\right) \mathcal{N}_L\left(\mathbf{x} \mid \mathbf{c}_k, \boldsymbol{\Gamma}_k\right) \\
&\equiv \sum_{k=1}^{K} \rho_k \mathcal{N}_{L+D}\left(\mathbf{w} \mid \boldsymbol{\mu}_k, \mathbf{V}_k\right). \qquad \text{(B5)}
\end{aligned}
$$

By using result for the joint Gaussian, see *e.g.,* (B24), we obtain the desired result (70).

Finally, Lemma 5.2 is proved via using the following two statements [18, Lemmas 1 and 2]:

(i) For any $\rho_k \in \mathbb{R}, \boldsymbol{\mu}_k \in \mathbb{R}^{L+D}$, and $\mathbf{V}_k \in \mathcal{S}_+^{L+D}$, there is a set of parameters $\mathbf{c}_k \in \mathbb{R}^L, \lambda_k \in \mathcal{S}_+^L, \pi_k \in \mathbb{R}, \mathbf{A}_k \in \mathbb{R}^{D\times L}, \mathbf{b}_k \in \mathbb{R}^D, \boldsymbol{\Sigma}_k \in \mathcal{S}_+^D$ such that (69) holds.

(ii) Reciprocally, for any $\mathbf{c}_k \in \mathbb{R}^L, \boldsymbol{\Lambda}_k \in \mathcal{S}_+^L, \pi_k \in \mathbb{R}, \mathbf{A}_k \in \mathbb{R}^{D\times L}, \mathbf{b}_k \in \mathbb{R}^D, \boldsymbol{\Sigma}_k \in \mathcal{S}_+^D$, there is a set of parameters $\rho_k \in \mathbb{R}, \boldsymbol{\mu}_k \in \mathbb{R}^{L+D}$ and $\mathbf{V}_k \in \mathcal{S}_+^{L+D}$ such that (70) holds.

B.2. Proof of Proposition 3.1

Using the sum and product rules for both discrete and continuous variables, the ELBO in BNP-GLLiM (A4) is given by

$$
\begin{aligned}
\mathcal{F}\left(q_{\mathcal{Z}}, q_{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\phi}}\right) &= \mathbb{E}_{q_{\mathcal{Z}} q_{\boldsymbol{\Theta}}}\left[\log \frac{p(\mathcal{Y}, \mathcal{X}, \mathcal{Z}, \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})}{q(\mathcal{Z}) q(\boldsymbol{\Theta})}\right] \equiv \mathbb{E}\left[\log \frac{p(\mathcal{Y}, \mathcal{X}, \mathcal{Z}, \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})}{q(\mathcal{Z}) q(\boldsymbol{\Theta})}\right] \\
&= \sum_{\mathcal{Z}} \int \int \int q(\mathcal{Z}) q(\boldsymbol{\Theta}) \log\left[\frac{p(\mathcal{Y}, \mathcal{X}, \mathcal{Z}, \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})}{q(\mathcal{Z}) q(\boldsymbol{\Theta})}\right] d\mathcal{Z} d\boldsymbol{\Theta} \\
&= \mathbb{E}\left[\log p(\mathcal{Y} \mid \mathcal{X}, \mathcal{Z}, \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] + \mathbb{E}\left[\log p(\mathcal{X} \mid \mathcal{Z}, \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] + \mathbb{E}\left[\log p(\mathcal{Z} \mid \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] \\
&\quad + \mathbb{E}\left[\log p(\boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] - \mathbb{E}\left[\log q(\mathcal{Z})\right] - \mathbb{E}\left[\log q(\boldsymbol{\Theta})\right]. \qquad \text{(B6)}
\end{aligned}
$$

Next, we evaluate the various terms in the ELBO (B6).

Proof of (47)

Via the mean field approximation and the truncation, we have the following computations.

$$\begin{aligned}
\mathbb{E}\left[\log p(\mathcal{Y} \mid \mathcal{X}, \mathcal{Z}, \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] &= \mathbb{E}\left[\log \prod_{n=1}^{N} p\left(\mathbf{y}_n \mid \mathbf{x}_n, z_n, \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}}\right)\right] \\
&= \mathbb{E}\left[\log \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}_D\left(\mathbf{y}_n \mid \widehat{\mathbf{A}}_k \mathbf{x}_n + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right)^{z_{nk}}\right] \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{E}\left[z_{nk} \log \mathcal{N}_D\left(\mathbf{y}_n \mid \widehat{\mathbf{A}}_k \mathbf{x}_n + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right)\right] \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{E}_{q_{\mathcal{Z}}}\left[z_{nk}\right] \log \mathcal{N}_D\left(\mathbf{y}_n \mid \widehat{\mathbf{A}}_k \mathbf{x}_n + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right) \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} q_{z_n}(k) \log \mathcal{N}_D\left(\mathbf{y}_n \mid \widehat{\mathbf{A}}_k \mathbf{x}_n + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right), \qquad \text{(B7)}
\end{aligned}$$

where

$$\begin{aligned}
\log \mathcal{N}_D\left(\mathbf{y}_n \mid \widehat{\mathbf{A}}_k \mathbf{x}_n + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right) = &-\frac{D}{2}\log(2\pi) - \frac{1}{2}\log\left|\widehat{\boldsymbol{\Sigma}}_k\right| \\
&- (\mathbf{y}_n - \widehat{\mathbf{A}}_k \mathbf{x}_n - \widehat{\mathbf{b}}_k)^\top \widehat{\boldsymbol{\Sigma}}_k^{-1} (\mathbf{y}_n - \widehat{\mathbf{A}}_k \mathbf{x}_n - \widehat{\mathbf{b}}_k). \qquad \text{(B8)}
\end{aligned}$$

Proof of (48)

Similarly, we obtain

$$\begin{aligned}
\mathbb{E}\left[\log p(\mathcal{X} \mid \mathcal{Z}, \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] &= \mathbb{E}\left[\log \prod_{n=1}^{N} p\left(\mathbf{x}_n \mid z_n, \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}}\right)\right] \\
&= \mathbb{E}\left[\log \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}_L\left(\mathbf{x}_n \mid \widehat{\boldsymbol{c}}_k, \widehat{\boldsymbol{\Gamma}}_k\right)^{z_{nk}}\right] \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{E}\left[z_{nk} \log \mathcal{N}_L\left(\mathbf{x}_n \mid \widehat{\boldsymbol{c}}_k, \widehat{\boldsymbol{\Gamma}}_k\right)\right] \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{E}_{q_{\mathcal{Z}}}\left[z_{nk}\right] \log \mathcal{N}_L\left(\mathbf{x}_n \mid \widehat{\boldsymbol{c}}_k, \widehat{\boldsymbol{\Gamma}}_k\right) \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} q_{z_n}(k) \log \mathcal{N}_L\left(\mathbf{x}_n \mid \widehat{\boldsymbol{c}}_k, \widehat{\boldsymbol{\Gamma}}_k\right), \qquad \text{(B9)}
\end{aligned}$$

where

$$\log \mathcal{N}_L\left(\mathbf{x}_n \mid \widehat{\boldsymbol{c}}_k, \widehat{\boldsymbol{\Gamma}}_k\right) = -\frac{L}{2}\log(2\pi) - \frac{1}{2}\log\left|\widehat{\boldsymbol{\Gamma}}_k\right| - \frac{1}{2}\left(\mathbf{x}_n - \widehat{\boldsymbol{c}}_k\right)^\top \widehat{\boldsymbol{\Gamma}}_k^{-1}\left(\mathbf{x}_n - \widehat{\boldsymbol{c}}_k\right). \qquad \text{(B10)}$$

Proof of (49)

Via calculation, it follows the expressions of the following quantities,

$$\begin{aligned}\mathbb{E}_{q_{\tau_k}}\left[\log\left(\tau_k\right)\right] &= \psi\left(\widehat{\gamma}_{k,1}\right) - \psi\left(\widehat{\gamma}_{k,1}+\widehat{\gamma}_{k,2}\right),\\ \mathbb{E}_{q_{\tau_k}}\left[\log\left(1-\tau_k\right)\right] &= \psi\left(\widehat{\gamma}_{k,2}\right) - \psi\left(\widehat{\gamma}_{k,1}+\widehat{\gamma}_{k,2}\right).\end{aligned} \tag{B11}$$

Via using (B11), it holds that

$$\begin{aligned}\mathbb{E}\left[\log p(\mathcal{Z}\mid\boldsymbol{\Theta};\widehat{\boldsymbol{\phi}})\right] &= \mathbb{E}\left[\log\prod_{n=1}^{N}\prod_{k=1}^{K}\left[\pi_k(\boldsymbol{\tau})\right]^{z_{nk}}\right]\\ &= \sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}_{q_{\mathcal{Z}}}\left[z_{nk}\right]\mathbb{E}_{q_{\boldsymbol{\Theta}}}\left[\log\left[\tau_k\prod_{l=1}^{k-1}\left(1-\tau_l\right)\right]\right]\\ &= \sum_{k=1}^{K}\sum_{n=1}^{N}q_{z_{nk}}\left[\mathbb{E}_{q_{\tau_k}}\left[\log\tau_k\right]+\sum_{l=1}^{k-1}\mathbb{E}_{q_{\tau_l}}\left[\log\left(1-\tau_l\right)\right]\right]\\ &= \sum_{k=1}^{K}N_k\left[\mathbb{E}_{q_{\tau_k}}\left[\log\tau_k\right]+\sum_{l=1}^{k-1}\mathbb{E}_{q_{\tau_l}}\left[\log\left(1-\tau_l\right)\right]\right]\\ &= \sum_{k=1}^{K}N_k\left[\psi\left(\widehat{\gamma}_{k,1}\right)-\psi\left(\widehat{\gamma}_{k,1}+\widehat{\gamma}_{k,2}\right)+\sum_{l=1}^{k-1}\left[\psi\left(\widehat{\gamma}_{l,2}\right)-\psi\left(\widehat{\gamma}_{l,1}+\widehat{\gamma}_{l,2}\right)\right]\right].\end{aligned} \tag{B12}$$

Proof of (50)

Given a chosen truncated value $K\in\mathbb{N}^\star$, it holds that

$$\mathbb{E}_{q_{\boldsymbol{\Theta}}}\left[\log p(\boldsymbol{\Theta};\widehat{\boldsymbol{\phi}})\right] = \sum_{k=1}^{K-1}\mathbb{E}_{q_{\boldsymbol{\Theta}}}\left[\log p\left(\tau_k\mid\alpha,\sigma\right)\right]+\mathbb{E}_{q_{\boldsymbol{\Theta}}}\left[\log p\left(\alpha,\sigma\mid\widehat{s}_1,\widehat{s}_2,\widehat{a}\right)\right].$$

Here, we have

$$\begin{aligned}\mathbb{E}_{q_{\boldsymbol{\Theta}}}\left[\log p\left(\tau_k\mid\alpha,\sigma\right)\right] &= \mathbb{E}_{q_{\boldsymbol{\Theta}}}\left[\log\mathrm{Beta}\left(\tau_k\mid 1-\sigma,\alpha+k\sigma\right)\right]\\ &= \mathbb{E}_{q_{\boldsymbol{\Theta}}}\left[\log\tau_k^{-\sigma}(1-\tau_k)^{\alpha+k\sigma-1}+\log C(\alpha,\sigma)\right],\\ &= -\mathbb{E}_{q_{\alpha,\sigma}}\left[\sigma\right]\mathbb{E}_{q_{\tau_k}}\left[\log\tau_k\right]+\mathbb{E}_{q_{\alpha,\sigma}}\left[\alpha+k\sigma-1\right]\mathbb{E}_{q_{\tau_k}}\left[\log(1-\tau_k)\right]\\ &\quad+\mathbb{E}_{q_{\alpha,\sigma}}\left[\log C(\alpha,\sigma)\right].\end{aligned}$$

where we have defined

$$C(\alpha,\sigma)=\frac{\Gamma(1-\sigma+\alpha+k\sigma)}{\Gamma(1-\sigma)\Gamma(\alpha+k\sigma)}.$$

Next, for the sake of simplicity, we use for σ a uniform prior $\mathcal{U}_{[0,1]}(\sigma)$ so that pa-

rameter a does not have to be taken into account. Then it holds that

$$\begin{aligned}\mathbb{E}_{q_\Theta}\left[\log p\left(\alpha,\sigma \mid \widehat{s}_1,\widehat{s}_2\right)\right] &= \mathbb{E}_{q_{\alpha,\sigma}}\left[\log \operatorname{Gam}\left(\alpha+\sigma \mid \widehat{s}_1,\widehat{s}_2\right)\right] + \mathbb{E}_{q_{\alpha,\sigma}}\left[\log \mathcal{U}_{[0,1]}(\sigma)\right] \\ &= \log\left[\frac{1}{\Gamma\left(\widehat{s}_1\right)}\widehat{s}_2^{\widehat{s}_1}\right] + \left(\widehat{s}_1-1\right)\mathbb{E}_{q_{\alpha,\sigma}}\left[\log\left(\alpha+\sigma\right)\right] - \widehat{s}_2\mathbb{E}_{q_{\alpha,\sigma}}\left[\alpha+\sigma\right] \\ &= \log\left[\frac{1}{\Gamma\left(\widehat{s}_1\right)}\widehat{s}_2^{\widehat{s}_1}\right] + \left(\widehat{s}_1-1\right)\mathbb{E}_{q_{\alpha,\sigma}}\left[\log\left(\alpha+\sigma\right)\right] - \widehat{s}_2\mathbb{E}_{q_{\alpha,\sigma}}\left[\alpha+\sigma\right].\end{aligned}$$

When $\sigma \neq 0$, the normalizing constant for $q_{\alpha,\sigma}(\alpha,\sigma)$ is not tractable. Nevertheless, to compute the ELBO, we do not need the full $q_{\alpha,\sigma}$ distribution but only the means $\mathbb{E}_{q_{\alpha,\sigma}}[\sigma]$, $\mathbb{E}_{q_{\alpha,\sigma}}[\alpha+k\sigma-1]$, $\mathbb{E}_{q_{\alpha,\sigma}}[\log C(\alpha,\sigma)]$, $\mathbb{E}_{q_{\alpha,\sigma}}[\log(\alpha+\sigma)]$ and $\mathbb{E}_{q_{\alpha,\sigma}}[\alpha+\sigma]$. One solution is therefore to use importance sampling or MCMC to compute these expectations via Monte Carlo sums.

When $\sigma = 0$, using integration by parts, it holds that $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$ and hence $C(\alpha,\sigma) \equiv C(\alpha) = \alpha$. Furthermore, the posterior $q_{\alpha,\sigma} \equiv q_\alpha$ is again a gamma distribution $\operatorname{Gam}\left(\alpha \mid \widehat{s}_1,\widehat{s}_2\right)$ with $\mathbb{E}_{q_{\alpha,\sigma}}[\alpha] \equiv \mathbb{E}_{q_\alpha}[\alpha] = \frac{\widehat{s}_1}{\widehat{s}_2}$ and $\mathbb{E}_{q_{\alpha,\sigma}}[\log\alpha] \equiv \mathbb{E}_{q_\alpha}[\log\alpha] = \psi\left(\widehat{s}_1\right) - \log\left(\widehat{s}_2\right)$. Therefore, we have the following tractable formulas:

$$\begin{aligned}\mathbb{E}_{q_\Theta}\left[\log p\left(\tau_k \mid \alpha,\sigma\right)\right] &\equiv \mathbb{E}_{q_\Theta}\left[\log p\left(\tau_k \mid \alpha\right)\right] \\ &= \left[\mathbb{E}_{q_{\alpha,0}}[\alpha]-1\right]\mathbb{E}_{q_{\tau_k}}\left[\log(1-\tau_k)\right] + \mathbb{E}_{q_{\alpha,0}}\left[\log\alpha\right], \\ &= \frac{\widehat{s}_1-\widehat{s}_2}{\widehat{s}_2}\left[\psi\left(\widehat{\gamma}_{k,2}\right) - \psi\left(\widehat{\gamma}_{k,1}+\widehat{\gamma}_{k,2}\right)\right] + \psi\left(\widehat{s}_1\right) - \log(\widehat{s}_2),\end{aligned} \tag{B13}$$

$$\begin{aligned}\mathbb{E}_{q_\Theta}\left[\log p\left(\alpha,\sigma \mid \widehat{s}_1,\widehat{s}_2\right)\right] &\equiv \mathbb{E}_{q_\Theta}\left[\log p\left(\alpha \mid \widehat{s}_1,\widehat{s}_2\right)\right] \\ &= \log\left[\frac{1}{\Gamma\left(\widehat{s}_1\right)}\widehat{s}_2^{\widehat{s}_1}\right] + \left(\widehat{s}_1-1\right)\left[\psi\left(\widehat{s}_1\right) - \log\left(\widehat{s}_2\right)\right] - \widehat{s}_1.\end{aligned} \tag{B14}$$

Proof of (51)

Due to the mean-field approximation (32) and truncation, this step is analytically computed as follows:

$$\begin{aligned}\mathbb{E}_{q_\mathcal{Z}}\left[\log q(\mathcal{Z})\right] &= \mathbb{E}_{q_\mathcal{Z}}\left[\log\prod_{n=1}^{N} q_{z_n}\left(z_n\right)\right] = \mathbb{E}_{q_\mathcal{Z}}\left[\log\prod_{n=1}^{N}\prod_{k=1}^{K} q_{z_n}\left(k\right)^{z_{nk}}\right] \\ &= \sum_{n=1}^{N}\sum_{k=1}^{K}\log q_{z_n}\left(k\right)\mathbb{E}_{q_\mathcal{Z}}\left[z_{nk}\right] = \sum_{n=1}^{N}\sum_{k=1}^{K} q_{z_n}\left(k\right)\log q_{z_n}\left(k\right).\end{aligned} \tag{B15}$$

Proof of (52)

We have

$$\mathbb{E}\left[\log q(\boldsymbol{\Theta})\right] = \mathbb{E}\left[\log q_{\alpha,\sigma}(\alpha,\sigma)\right] + \sum_{k=1}^{K-1}\mathbb{E}\left[\log q_{\tau_k}\left(\tau_k\right)\right]. \tag{B16}$$

Note that these terms involving expectations of the logs of the q distributions simply represent the negative entropies of those distributions.

Since $q_{\alpha,\sigma}(\alpha,\sigma)$ is not tractable, when $\sigma \neq 0$, we cannot calculate analytically $\mathbb{E}\left[\log q_{\alpha,\sigma}(\alpha,\sigma)\right]$. Furthermore, it is also difficult to approximate it using MCMC or importance sampling.

When $\sigma = 0$, the posterior $q_{\alpha,\sigma} \equiv q_\alpha$ is again a gamma distribution $\mathrm{Gam}\left(\alpha \mid \widehat{s}_1, \widehat{s}_2\right)$ with

$$\begin{aligned}\mathbb{E}\left[\log q_{\alpha,0}(\alpha)\right] &\equiv \mathbb{E}\left[\log \mathrm{Gam}\left(\alpha \mid \widehat{s}_1, \widehat{s}_2\right)\right] \\ &= -\mathrm{H}\left[\mathrm{Gam}\left(\alpha \mid \widehat{s}_1, \widehat{s}_2\right)\right] \\ &= -\log\Gamma\left(\widehat{s}_1\right) + \left(\widehat{s}_1 - 1\right)\psi\left(\widehat{s}_1\right) + \log(\widehat{s}_2) - \widehat{s}_1. \end{aligned} \tag{B17}$$

Since we had $q_{\tau_k}\left(\tau_k\right) = \mathrm{Beta}\left(\tau_k \mid \widehat{\gamma}_{k,1}, \widehat{\gamma}_{k,2}\right)$, its differential entropy is given by

$$\begin{aligned}\mathbb{E}\left[\log q_{\tau_k}\left(\tau_k\right)\right] &= -\mathrm{H}\left[\mathrm{Beta}\left(\tau_k \mid \widehat{\gamma}_{k,1}, \widehat{\gamma}_{k,2}\right)\right] \\ &= \sum_{l=1}^{2}\left(\widehat{\gamma}_{k,l} - 1\right)\left\{\psi\left(\widehat{\gamma}_{k,l}\right) - \psi\left(\widehat{\gamma}_{k,1} + \widehat{\gamma}_{k,2}\right)\right\} + \log\frac{\Gamma\left(\widehat{\gamma}_{k,1} + \widehat{\gamma}_{k,2}\right)}{\Gamma\left(\widehat{\gamma}_{k,1}\right)\Gamma\left(\widehat{\gamma}_{k,2}\right)}. \end{aligned} \tag{B18}$$

B.3. Proof of Theorem 4.1

Recall that we defined $\boldsymbol{\Theta} = (\boldsymbol{\tau}, \alpha, \sigma)$. Then,

$$\begin{aligned}p\left(\widehat{\mathbf{y}}, \widehat{\mathbf{x}}, \mathcal{X}, \mathcal{Y}\right) &= \sum_{\widehat{\mathbf{z}}}\int p\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{x}}, \widehat{\mathbf{z}}, \boldsymbol{\Theta}, \mathcal{X}, \mathcal{Y}\right) p\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{z}}, \boldsymbol{\Theta}, \mathcal{X}, \mathcal{Y}\right) p\left(\widehat{\mathbf{z}} \mid \boldsymbol{\Theta}, \mathcal{X}, \mathcal{Y}\right) p\left(\boldsymbol{\Theta} \mid \mathcal{X}, \mathcal{Y}\right) d\boldsymbol{\Theta} \\ &= \sum_{\widehat{\mathbf{z}}}\int p\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{x}}, \widehat{\mathbf{z}}; \boldsymbol{A}, \boldsymbol{b}, \boldsymbol{\Sigma}\right) p\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{z}}, \boldsymbol{c}, \boldsymbol{\Gamma}\right) p\left(\widehat{\mathbf{z}} \mid \boldsymbol{\tau}; \beta\right) p\left(\boldsymbol{\Theta} \mid \mathcal{X}, \mathcal{Y}\right) d\boldsymbol{\Theta} \equiv D_1. \end{aligned} \tag{B19}$$

Note that in (B19), $p\left(\boldsymbol{\Theta} \mid \mathcal{X}, \mathcal{Y}\right)$ is in fact the (unknown) true posterior distribution of the parameters given a sample $(\mathcal{X}, \mathcal{Y})$. Because the integrations w.r.t. true posterior distribution are intractable, we approximate the predictive conditional density by replacing the true posterior distribution $p\left(\boldsymbol{\Theta} \mid \mathcal{X}, \mathcal{Y}\right)$ with its truncated variational posterior of parameters $\boldsymbol{\Theta}$ given by

$$q_\Theta(\boldsymbol{\Theta} \mid \mathcal{X}, \mathcal{Y}) = q_{\alpha,\sigma}(\alpha, \sigma \mid \mathcal{X}, \mathcal{Y}) \prod_{k=1}^{K-1} q_{\tau_k}\left(\tau_k \mid \mathcal{X}, \mathcal{Y}\right). \tag{B20}$$

Recall that the infinite state space for each z_j is dealt with by choosing a truncation of the state space to a maximum label K [10]. In practice, this consists of assuming that the variational distributions q_{z_n} for $n \in [N]$, satisfy $q_{z_n}(k) = 0$ for $k > K$ and that the variational distribution on τ also factorizes as $q_{\boldsymbol{\tau}}(\boldsymbol{\tau}) = \prod_{k=1}^{K-1} q_{\tau_k}\left(\tau_k\right)$ with the additional condition that $\tau_K = 1$. In particular, here we choose $K = \widehat{K}$ where $\widehat{K}$ is estimated from some suitable procedures.

For simplicity, we consider the case when $\beta = 0, \sigma = 0$. Then we have

$$
\begin{aligned}
D_1 &\approx \sum_{\widehat{\mathbf{z}}} \int p\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{x}}, \widehat{\mathbf{z}}; \widehat{\boldsymbol{A}}, \widehat{\boldsymbol{b}}, \widehat{\boldsymbol{\Sigma}}\right) p\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{z}}, \widehat{\boldsymbol{c}}, \widehat{\boldsymbol{\Gamma}}\right) p\left(\widehat{\mathbf{z}} \mid \boldsymbol{\tau}\right) q_{\Theta}(\boldsymbol{\Theta} \mid \mathcal{X}, \mathcal{Y}) d\boldsymbol{\Theta} \\
&= \sum_{k=1}^{\infty} \int \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{A}}_k \widehat{\mathbf{x}} + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right) \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{c}}_k, \widehat{\boldsymbol{\Gamma}}_k\right) \pi_k(\boldsymbol{\tau}) q_{\Theta}(\boldsymbol{\Theta} \mid \mathcal{X}, \mathcal{Y}) d\boldsymbol{\Theta} \\
&\approx \sum_{k=1}^{K} \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{c}}_k, \widehat{\boldsymbol{\Gamma}}_k\right) \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{A}}_k \widehat{\mathbf{x}} + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right) \int \pi_k(\boldsymbol{\tau}) q_{\Theta}(\boldsymbol{\Theta} \mid \mathcal{X}, \mathcal{Y}) d\boldsymbol{\Theta} \\
&= \sum_{k=1}^{K} \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{c}}_k, \widehat{\boldsymbol{\Gamma}}_k\right) \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{A}}_k \widehat{\mathbf{x}} + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right) \int \pi_k(\boldsymbol{\tau}) q_{\boldsymbol{\tau}}(\boldsymbol{\tau} \mid \mathcal{X}, \mathcal{Y}) d\boldsymbol{\tau} \underbrace{\int q_{\alpha,0}(\alpha \mid \mathcal{X}, \mathcal{Y}) d\alpha}_{=1} \\
&= \sum_{k=1}^{K} \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k(\boldsymbol{\tau})\right] \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{c}}_k, \widehat{\boldsymbol{\Gamma}}_k\right) \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{A}}_k \widehat{\mathbf{x}} + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right) \\
&\equiv \sum_{k=1}^{K} \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k\left(\boldsymbol{\tau}\right)\right] \mathcal{N}_{L+D}\left(\widehat{\mathbf{w}} \mid \mathbb{E}\left[\mathbf{w}\right], \operatorname{cov}\left[\mathbf{w}\right]\right) \\
&= \sum_{k=1}^{K} \mathbb{E}_{q_{\tau_k}}\left[\tau_k\right] \prod_{l=1}^{k-1} \mathbb{E}_{q_{\tau_l}}\left[1 - \tau_l\right] \mathcal{N}_{L+D}\left(\widehat{\mathbf{w}} \mid \mathbb{E}\left[\mathbf{w}\right], \operatorname{cov}\left[\mathbf{w}\right]\right).
\end{aligned}
$$

Here, by defining $\widehat{\mathbf{w}} \equiv [\widehat{\mathbf{x}}; \widehat{\mathbf{y}}]$, we used the fact that,

$$
\mathbb{E}\left[\mathbf{w}\right] = \begin{pmatrix} \widehat{\mathbf{c}}_k \\ \widehat{\mathbf{A}}_k \widehat{\mathbf{c}}_k + \widehat{\mathbf{b}}_k \end{pmatrix}, \quad \operatorname{cov}\left[\mathbf{w}\right] = \begin{pmatrix} \widehat{\boldsymbol{\Gamma}}_k & \widehat{\boldsymbol{\Gamma}}_k \widehat{\mathbf{A}}_k^{\top} \\ \widehat{\mathbf{A}}_k \widehat{\boldsymbol{\Gamma}}_k & \widehat{\boldsymbol{\Sigma}}_k + \widehat{\mathbf{A}}_k \widehat{\boldsymbol{\Gamma}}_k \widehat{\mathbf{A}}_k^{\top} \end{pmatrix}. \tag{B21}
$$

Indeed, we made use of the following result for the joint Gaussian, see, *e.g.*, [9, Eq. (2.115), page 93]. Given a marginal Gaussian distribution for $\mathbf{x}$ and a conditional Gaussian distribution for $\mathbf{y}$ given $\mathbf{x}$ in the form

$$
p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Gamma}^{-1}\right), \tag{B22}
$$

$$
p(\mathbf{y} \mid \mathbf{x}) = \mathcal{N}\left(\mathbf{y} \mid \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}\right), \tag{B23}
$$

then the joint distribution of $\mathbf{w} \equiv [\mathbf{x}; \mathbf{y}]$ is given by

$$
\begin{aligned}
p(\mathbf{w}) &= \mathcal{N}\left(\mathbf{w} \mid \mathbb{E}\left[\mathbf{w}\right], \operatorname{cov}\left[\mathbf{w}\right]\right), \text{ where} \\
\operatorname{cov}\left[\mathbf{w}\right] &= \begin{pmatrix} \boldsymbol{\Gamma}^{-1} & \boldsymbol{\Gamma}^{-1}\mathbf{A}^{\top} \\ \mathbf{A}\boldsymbol{\Gamma}^{-1} & \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Gamma}^{-1}\mathbf{A}^{\top} \end{pmatrix}, \quad \mathbb{E}\left[\mathbf{w}\right] = \begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{A}\boldsymbol{\mu} + \mathbf{b} \end{pmatrix}.
\end{aligned} \tag{B24}
$$

In our situation, the desired result is obtained via using $\mathbf{y} \equiv \widehat{\mathbf{y}}, \mathbf{A} \equiv \widehat{\mathbf{A}}_k, \mathbf{b} \equiv \widehat{\mathbf{b}}_k, \mathbf{L}^{-1} = \widehat{\boldsymbol{\Sigma}}_k$, $\mathbf{x} \equiv \widehat{\mathbf{x}}, \boldsymbol{\mu} \equiv \widehat{\mathbf{c}}_k, \boldsymbol{\Gamma}^{-1} \equiv \widehat{\boldsymbol{\Gamma}}_k$.

Furthermore, we also used the fact that

$$\begin{aligned}
\mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k\left(\boldsymbol{\tau}\right)\right] &= \int \tau_k q_{\tau_k}\left(\tau_k \mid \mathcal{X}, \mathcal{Y}\right) d\tau_k \int \prod_{l=1}^{k-1}\left(1-\tau_l\right) \prod_{j=1, j\neq k}^{K-1} q_{\tau_j}\left(\tau_j \mid \mathcal{X}, \mathcal{Y}\right) \prod_{j=1, j\neq k}^{K} d\tau_j \\
&= \mathbb{E}_{q_{\tau_k}}\left[\tau_k\right] \int \prod_{l=1}^{k-1}\left(1-\tau_l\right) \prod_{j=1}^{k-1} q_{\tau_j}\left(\tau_j \mid \mathcal{X}, \mathcal{Y}\right) \underbrace{\int \prod_{j=k+1}^{K-1} q_{\tau_j}\left(\tau_j \mid \mathcal{X}, \mathcal{Y}\right) \prod_{j=k+1}^{K-1} d\tau_j}_{=1} \prod_{j=1}^{k-1} d\tau_j \\
&= \mathbb{E}_{q_{\tau_k}}\left[\tau_k\right] \prod_{l=1}^{k-1} \int \left(1-\tau_l\right) q_{\tau_l}\left(\tau_l \mid \mathcal{X}, \mathcal{Y}\right) d\tau_l \\
&= \mathbb{E}_{q_{\tau_k}}\left[\tau_k\right] \prod_{l=1}^{k-1} \mathbb{E}_{q_{\tau_l}}\left[1-\tau_l\right]. \qquad \text{(B25)}
\end{aligned}$$

Next, we aim to prove that

$$\sum_{k=1}^{K} \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k\left(\boldsymbol{\tau}\right)\right] = 1. \qquad \text{(B26)}$$

Indeed, recall that we have defined

$$\tau_k \mid \alpha, \sigma \overset{\text{ind}}{\sim} \text{Beta}\left(\tau_k \mid 1-\sigma, \alpha + k\sigma\right), \quad k \in \mathbb{N}^\star, \qquad \text{(B27)}$$

$$\pi_k(\boldsymbol{\tau}) = \tau_k \prod_{l=1}^{k-1}\left(1-\tau_l\right), \quad k \in \mathbb{N}^\star, \qquad \text{(B28)}$$

$$p\left(\mathcal{Z} \mid \boldsymbol{\tau}\right) \propto \prod_{n=1}^{N} \pi_{z_n}\left(\boldsymbol{\tau}\right), \qquad \text{(B29)}$$

and to deal with the infinite state space for each z_j, we considered a truncation of the state space to a maximum label $K \equiv K_{\max}, K_{\max} \in \mathbb{N}^\star$ [10]. In practice, this consists of assuming that the variational distributions q_{z_n} for $n \in [N]$, satisfy $q_{z_n}(k) = 0$ for $k > K$ and that the variational distribution on τ also factorizes as $q_{\boldsymbol{\tau}}(\boldsymbol{\tau}) = \prod_{k=1}^{K-1} q_{\tau_k}\left(\tau_k\right)$ with the additional condition that $\tau_K = 1$. Based on the proof from [33, Lemma 3.4], it holds that a necessary and sufficient condition to guarantee that these π_k's sum to 1 almost surely, *i.e.*,

$$\sum_{k=1}^{\infty} \pi_k(\boldsymbol{\tau}) = \sum_{k=1}^{\infty} \tau_k \prod_{l=1}^{k-1}\left(1-\tau_l\right) = 1, \qquad \text{(B30)}$$

is that the expectation $\mathbb{E}\left[\prod_{l=1}^{k-1}\left(1-\tau_l\right)\right]$ tends to 0 as k tends to ∞. In particular, if $\tau_1, \tau_2, \ldots$ are iid, *e.g.*, when $\sigma = 0$, it suffices that $p\left(\tau_1 > 0\right) > 0$. Then

$$1 = \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\sum_{k=1}^{\infty} \pi_k(\boldsymbol{\tau})\right] = \sum_{k=1}^{\infty} \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k(\boldsymbol{\tau})\right] = \sum_{k=1}^{K} \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k(\boldsymbol{\tau})\right].$$

B.4. Proof of Theorem 4.2

From the product rule of probability, we see that this conditional distribution can be evaluated from the joint and marginal distributions. Furthermore, by integrating out $\widehat{\mathbf{z}}$ and $\boldsymbol{\Theta}$, the predictive conditional density is then given by

$$p\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{x}}, \mathcal{X}, \mathcal{Y}\right) = \frac{p\left(\widehat{\mathbf{y}}, \widehat{\mathbf{x}} \mid \mathcal{X}, \mathcal{Y}\right)}{p\left(\widehat{\mathbf{x}} \mid \mathcal{X}, \mathcal{Y}\right)} = \frac{\sum_{\widehat{\mathbf{z}}} \int p\left(\widehat{\mathbf{y}}, \widehat{\mathbf{x}}, \widehat{\mathbf{z}}, \boldsymbol{\Theta} \mid \mathcal{X}, \mathcal{Y}\right) d\widehat{\mathbf{z}} d\boldsymbol{\Theta}}{\sum_{\widehat{\mathbf{z}}} \int p\left(\widehat{\mathbf{x}}, \widehat{\mathbf{z}}, \boldsymbol{\Theta} \mid \mathcal{X}, \mathcal{Y}\right) d\widehat{\mathbf{z}} d\boldsymbol{\Theta}} \equiv \frac{D_1}{D_2}. \tag{B31}$$

Next, with a similar step as in the proof of Theorem 4.1, we also obtain

$$D_2 \approx \sum_{k=1}^{K} \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k\left(\boldsymbol{\tau}\right)\right] \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{c}}_k, \widehat{\boldsymbol{\Gamma}}_k\right). \tag{B32}$$

Therefore, we obtain

$$\begin{aligned}
&p\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{x}}, \mathcal{X}, \mathcal{Y}\right) \\
&\approx \frac{\sum_{k=1}^{K} \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k\left(\boldsymbol{\tau}\right)\right] \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{c}}_k, \widehat{\boldsymbol{\Gamma}}_k\right) \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{A}}_k \widehat{\mathbf{x}} + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right)}{\sum_{k=1}^{K} \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k\left(\boldsymbol{\tau}\right)\right] \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{c}}_k, \widehat{\boldsymbol{\Gamma}}_k\right)} \\
&= \sum_{k=1}^{K} \frac{\mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k\left(\boldsymbol{\tau}\right)\right] \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{c}}_k, \widehat{\boldsymbol{\Gamma}}_k\right)}{\sum_{k=1}^{K} \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k\left(\boldsymbol{\tau}\right)\right] \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{c}}_k, \widehat{\boldsymbol{\Gamma}}_k\right)} \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{A}}_k \widehat{\mathbf{x}} + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right) \\
&\equiv \sum_{k=1}^{K} g_k\left(\widehat{\mathbf{x}} \mid \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\phi}}, \mathcal{X}, \mathcal{Y}\right) \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{A}}_k \widehat{\mathbf{x}} + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right),
\end{aligned} \tag{B33}$$

which is a mixture of Gaussian experts since we have

$$g_k\left(\widehat{\mathbf{x}} \mid \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\phi}}, \mathcal{X}, \mathcal{Y}\right) = \frac{\mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k\left(\boldsymbol{\tau}\right)\right] \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{c}}_k, \widehat{\boldsymbol{\Gamma}}_k\right)}{\sum_{k=1}^{K} \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k\left(\boldsymbol{\tau}\right)\right] \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{c}}_k, \widehat{\boldsymbol{\Gamma}}_k\right)}, \quad k \in [K], \tag{B34}$$

belongs to a $K-1$ dimensional probability simplex.

B.5. Proof of Theorem 4.3

To deal with high-dimensional regression data, namely high-to-low regression, given the inverse conditional density $p(\widehat{\mathbf{y}} \mid \widehat{\mathbf{x}}, \mathcal{X}, \mathcal{Y})$, we want to compute the following forward conditional density

$$p(\widehat{\mathbf{x}} \mid \widehat{\mathbf{y}}, \mathcal{X}, \mathcal{Y}) = \frac{p(\widehat{\mathbf{x}}, \widehat{\mathbf{y}} \mid \mathcal{X}, \mathcal{Y})}{p(\widehat{\mathbf{y}} \mid \mathcal{X}, \mathcal{Y})} = \frac{p(\widehat{\mathbf{x}}, \widehat{\mathbf{y}} \mid \mathcal{X}, \mathcal{Y})}{\int_{\widehat{\mathbf{x}}} p(\widehat{\mathbf{x}}, \widehat{\mathbf{y}} \mid \mathcal{X}, \mathcal{Y}) d\widehat{\mathbf{x}}} = \frac{D_1}{\int_{\widehat{\mathbf{x}}} D_1(\widehat{\mathbf{x}}) d\widehat{\mathbf{x}}} \equiv \frac{D_1}{D_3}. \tag{B35}$$

Then, we have to compute or approximate D_3. Using Theorem 4.1, we obtain

$$
\begin{aligned}
D_3 &\approx \sum_{k=1}^{K} \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k\left(\boldsymbol{\tau}\right)\right] \int \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{c}}_k, \widehat{\boldsymbol{\Gamma}}_k\right) \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{A}}_k \widehat{\mathbf{x}} + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right) d\widehat{\mathbf{x}} \\
&= \sum_{k=1}^{K} \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k\left(\boldsymbol{\tau}\right)\right] \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{A}}_k \widehat{\mathbf{c}}_k + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k + \widehat{\mathbf{A}}_k \widehat{\boldsymbol{\Gamma}}_k \widehat{\mathbf{A}}_k^{\top}\right).
\end{aligned}
$$

Indeed, we made use of the following results for marginal and conditional Gaussians, see, *e.g.,* [9, Eq. (2.115), page 93]. Given a marginal Gaussian distribution for $\mathbf{x}$ and a conditional Gaussian distribution for $\mathbf{y}$ given $\mathbf{x}$ in the form

$$
p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Gamma}^{-1}\right), \tag{B36}
$$

$$
p(\mathbf{y} \mid \mathbf{x}) = \mathcal{N}\left(\mathbf{y} \mid \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}\right), \tag{B37}
$$

then the marginal distribution of $\mathbf{y}$ and the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ are given by

$$
p(\mathbf{y}) = \int p(\mathbf{y} \mid \mathbf{x}) p(\mathbf{x}) d\mathbf{x} = \mathcal{N}\left(\mathbf{y} \mid \mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Gamma}^{-1}\mathbf{A}^{\top}\right), \tag{B38}
$$

$$
p(\mathbf{x} \mid \mathbf{y}) = \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\Sigma}\left[\mathbf{A}^{\top}\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Gamma}\boldsymbol{\mu}\right], \boldsymbol{\Sigma}\right), \boldsymbol{\Sigma} = \left(\boldsymbol{\Gamma} + \mathbf{A}^{\top}\mathbf{L}\mathbf{A}\right)^{-1}. \tag{B39}
$$

In our situation, the desired result is obtained via using $\mathbf{y} \equiv \widehat{\mathbf{y}}, \mathbf{A} \equiv \widehat{\mathbf{A}}_k, \mathbf{b} \equiv \widehat{\mathbf{b}}_k, \mathbf{L}^{-1} = \widehat{\boldsymbol{\Sigma}}_k$, $\mathbf{x} \equiv \widehat{\mathbf{x}}, \boldsymbol{\mu} \equiv \widehat{\mathbf{c}}_k, \boldsymbol{\Gamma}^{-1} \equiv \widehat{\boldsymbol{\Gamma}}_k$.

Finally, we obtain

$$
\begin{aligned}
&p\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{y}}, \mathcal{X}, \mathcal{Y}\right) \\
&\approx \sum_{k=1}^{K} \frac{\mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k\left(\boldsymbol{\tau}\right)\right] \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{A}}_k \widehat{\mathbf{x}} + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right) \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{c}}_k, \widehat{\boldsymbol{\Gamma}}_k\right)}{\sum_{k=1}^{K} \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k\left(\boldsymbol{\tau}\right)\right] \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{A}}_k \widehat{\mathbf{c}}_k + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k + \widehat{\mathbf{A}}_k \widehat{\boldsymbol{\Gamma}}_k \widehat{\mathbf{A}}_k^{\top}\right)} \\
&= \sum_{k=1}^{K} g_k\left(\widehat{\mathbf{y}} \mid \widehat{\boldsymbol{\Theta}}^*, \widehat{\boldsymbol{\phi}}^*, \mathcal{X}, \mathcal{Y}\right) \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{A}}_k^* \widehat{\mathbf{y}} + \widehat{\mathbf{b}}_k^*, \widehat{\boldsymbol{\Sigma}}_k^*\right),
\end{aligned} \tag{B40}
$$

where

$$
g_k\left(\widehat{\mathbf{y}} \mid \widehat{\boldsymbol{\Theta}}^*, \widehat{\boldsymbol{\phi}}^*, \mathcal{X}, \mathcal{Y}\right) = \frac{\mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k\left(\boldsymbol{\tau}\right)\right] \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{c}}_k^*, \widehat{\boldsymbol{\Gamma}}_k^*\right)}{\sum_{k=1}^{K} \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k\left(\boldsymbol{\tau}\right)\right] \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{c}}_k^*, \widehat{\boldsymbol{\Gamma}}_k^*\right)}. \tag{B41}
$$

Here, we used the fact that $p(\widehat{\mathbf{y}}, \widehat{\mathbf{x}} \mid \widehat{z} = k) = p(\widehat{\mathbf{x}} \mid \widehat{\mathbf{y}}, \widehat{z} = k) p(\widehat{\mathbf{y}} \mid \widehat{z} = k)$, namely,

$$
\begin{aligned}
&\mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{A}}_k \widehat{\mathbf{x}} + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right) \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{c}}_k, \widehat{\boldsymbol{\Gamma}}_k\right) \\
&= \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\boldsymbol{\Sigma}}_k^*\left[\widehat{\mathbf{A}}_k^{\top} \widehat{\boldsymbol{\Sigma}}_k^{-1}(\widehat{\mathbf{y}} - \widehat{\mathbf{b}}_k) + \widehat{\boldsymbol{\Gamma}}_k^{-1} \widehat{\mathbf{c}}_k\right], \widehat{\boldsymbol{\Sigma}}_k^*\right) \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{A}}_k \widehat{\mathbf{c}}_k + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k + \widehat{\mathbf{A}}_k \widehat{\boldsymbol{\Gamma}}_k \widehat{\mathbf{A}}_k^{\top}\right) \\
&= \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{A}}_k^* \widehat{\mathbf{y}} + \widehat{\mathbf{b}}_k^*, \widehat{\boldsymbol{\Sigma}}_k^*\right) \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{c}}_k^*, \widehat{\boldsymbol{\Gamma}}_k^*\right),
\end{aligned}
$$

with

$$
\begin{aligned}
\widehat{\boldsymbol{\Sigma}}_k^* &= \left(\widehat{\boldsymbol{\Gamma}}_k^{-1} + \widehat{\mathbf{A}}_k^\top \widehat{\boldsymbol{\Sigma}}_k^{-1} \widehat{\mathbf{A}}_k\right)^{-1}, \\
\widehat{\mathbf{A}}_k^* &= \widehat{\boldsymbol{\Sigma}}_k^* \widehat{\mathbf{A}}_k^\top \widehat{\boldsymbol{\Sigma}}_k^{-1}, \\
\widehat{\mathbf{b}}_k^* &= \widehat{\boldsymbol{\Sigma}}_k^* \left[\widehat{\boldsymbol{\Gamma}}_k^{-1} \widehat{\mathbf{c}}_k - \widehat{\mathbf{A}}_k^\top \widehat{\boldsymbol{\Sigma}}_k^{-1} \widehat{\mathbf{b}}_k\right], \\
\widehat{\mathbf{c}}_k^* &= \widehat{\mathbf{A}}_k \widehat{\mathbf{c}}_k + \widehat{\mathbf{b}}_k, \\
\widehat{\boldsymbol{\Gamma}}_k^* &= \widehat{\boldsymbol{\Sigma}}_k + \widehat{\mathbf{A}}_k \widehat{\boldsymbol{\Gamma}}_k \widehat{\mathbf{A}}_k^\top.
\end{aligned} \tag{B42}
$$

When required, it is straightforward to approximate the expectation and covariance matrix of $\widehat{\mathbf{x}} \mid \widehat{\mathbf{y}}, \mathcal{X}, \mathcal{Y}$ as follows:

$$
\begin{aligned}
\mathbb{E}\left[\widehat{\mathbf{x}} \mid \widehat{\mathbf{y}}, \mathcal{X}, \mathcal{Y}\right] &\approx \int \left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{y}}, \mathcal{X}, \mathcal{Y}\right) \sum_{k=1}^{K} g_k\left(\widehat{\mathbf{y}} \mid \widehat{\boldsymbol{\Theta}}^*, \widehat{\boldsymbol{\phi}}^*, \mathcal{X}, \mathcal{Y}\right) \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{A}}_k^* \widehat{\mathbf{y}} + \widehat{\mathbf{b}}_k^*, \widehat{\boldsymbol{\Sigma}}_k^*\right) d\widehat{\mathbf{x}} \\
&= \sum_{k=1}^{K} g_k\left(\widehat{\mathbf{y}} \mid \widehat{\boldsymbol{\Theta}}^*, \widehat{\boldsymbol{\phi}}^*, \mathcal{X}, \mathcal{Y}\right) \int \left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{y}}, \mathcal{X}, \mathcal{Y}\right) \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{A}}_k^* \widehat{\mathbf{y}} + \widehat{\mathbf{b}}_k^*, \widehat{\boldsymbol{\Sigma}}_k^*\right) d\widehat{\mathbf{x}} \\
&= \sum_{k=1}^{K} g_k\left(\widehat{\mathbf{y}} \mid \widehat{\boldsymbol{\Theta}}^*, \widehat{\boldsymbol{\phi}}^*, \mathcal{X}, \mathcal{Y}\right) \left(\widehat{\mathbf{A}}_k^* \widehat{\mathbf{y}} + \widehat{\mathbf{b}}_k^*\right), \\
\operatorname{var}\left[\widehat{\mathbf{x}} \mid \widehat{\mathbf{y}}, \mathcal{X}, \mathcal{Y}\right] &= \mathbb{E}\left[\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{y}}, \mathcal{X}, \mathcal{Y}\right)\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{y}}, \mathcal{X}, \mathcal{Y}\right)^\top\right] - \mathbb{E}\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{y}}, \mathcal{X}, \mathcal{Y}\right) \mathbb{E}\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{y}}, \mathcal{X}, \mathcal{Y}\right)^\top \\
&\approx \int \left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{y}}, \mathcal{X}, \mathcal{Y}\right)\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{y}}, \mathcal{X}, \mathcal{Y}\right)^\top \sum_{k=1}^{K} g_k\left(\widehat{\mathbf{y}} \mid \widehat{\boldsymbol{\Theta}}^*, \widehat{\boldsymbol{\phi}}^*, \mathcal{X}, \mathcal{Y}\right) \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{A}}_k^* \widehat{\mathbf{y}} + \widehat{\mathbf{b}}_k^*, \widehat{\boldsymbol{\Sigma}}_k^*\right) d\widehat{\mathbf{x}} \\
&\quad - \mathbb{E}\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{y}}, \mathcal{X}, \mathcal{Y}\right) \mathbb{E}\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{y}}, \mathcal{X}, \mathcal{Y}\right)^\top \\
&\approx \sum_{k=1}^{K} g_k\left(\widehat{\mathbf{y}} \mid \widehat{\boldsymbol{\Theta}}^*, \widehat{\boldsymbol{\phi}}^*, \mathcal{X}, \mathcal{Y}\right) \int \left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{y}}, \mathcal{X}, \mathcal{Y}\right)\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{y}}, \mathcal{X}, \mathcal{Y}\right)^\top \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{A}}_k^* \widehat{\mathbf{y}} + \widehat{\mathbf{b}}_k^*, \widehat{\boldsymbol{\Sigma}}_k^*\right) d\widehat{\mathbf{x}} \\
&\quad - \mathbb{E}\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{y}}, \mathcal{X}, \mathcal{Y}\right) \mathbb{E}\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{y}}, \mathcal{X}, \mathcal{Y}\right)^\top \\
&\approx \sum_{k=1}^{K} g_k\left(\widehat{\mathbf{y}} \mid \widehat{\boldsymbol{\Theta}}^*, \widehat{\boldsymbol{\phi}}^*, \mathcal{X}, \mathcal{Y}\right) \left[\widehat{\boldsymbol{\Sigma}}_k^* + \left(\widehat{\mathbf{A}}_k^* \widehat{\mathbf{y}} + \widehat{\mathbf{b}}_k^*\right)\left(\widehat{\mathbf{A}}_k^* \widehat{\mathbf{y}} + \widehat{\mathbf{b}}_k^*\right)^\top\right] \\
&\quad - \mathbb{E}\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{y}}, \mathcal{X}, \mathcal{Y}\right) \mathbb{E}\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{y}}, \mathcal{X}, \mathcal{Y}\right)^\top,
\end{aligned}
$$

where we used the following definitions

$$
\operatorname{cov}\left(\mathbf{X}, \mathbf{Y}\right) = \mathbb{E}\left(\mathbf{X}\mathbf{Y}^\top\right) - \mathbb{E}\left(\mathbf{X}\right) \mathbb{E}\left(\mathbf{Y}\right)^\top, \operatorname{var}\left(\mathbf{X}\right) = \operatorname{cov}\left(\mathbf{X}, \mathbf{X}\right).
$$

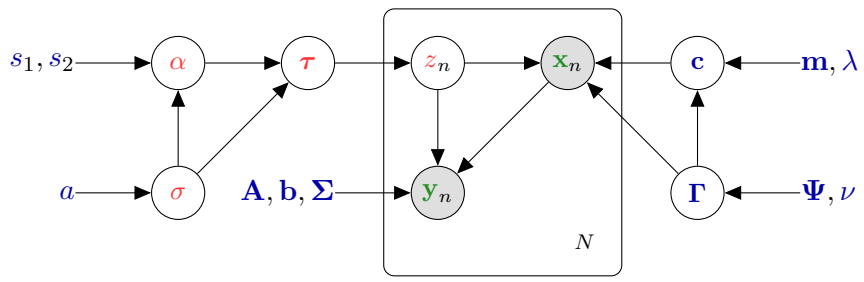


Figure C1. Graphical representation of BNP-GLLiM2: the plate denotes N iid observations, white-filled circles correspond to unobserved (latent) variables and random or unknown parameters represented in red, while grey-filled circles correspond to observed variables represented in green. Hyperparameters are represented in blue.

Appendix C. BNP-GLLiM2: a model with an hyperprior on the gating parameters

C.1. VBEM for BNP-GLLiM2

A more general BNP-GLLiM model, referred to as BNP-GLLiM2 can be considered by specifying a prior on the gating parameters $(\mathbf{c}_k, \boldsymbol{\Gamma}_k)$ as a normal-inverse-Wishart (NIW) distribution parameterized by $\boldsymbol{\rho}_k = (\mathbf{m}_k, \lambda_k, \boldsymbol{\Psi}_k, \nu_k)$ with a PDF

$$p\left(\mathbf{c}_k, \boldsymbol{\Gamma}_k \mid \boldsymbol{\rho}_k\right) \equiv \mathcal{NIW}\left(\mathbf{c}_k, \boldsymbol{\Gamma}_k \mid \boldsymbol{\rho}_k\right) = \mathcal{N}\left(\mathbf{c}_k \mid \mathbf{m}_k, \lambda_k^{-1}\boldsymbol{\Gamma}_k\right) \mathcal{IW}\left(\boldsymbol{\Gamma}_k \mid \boldsymbol{\Psi}_k, \nu_k\right). \quad \text{(C1)}$$

The assumptions on the other parameters are not changed, so that hyperparameters and parameters are now as follows:

$$\boldsymbol{\phi} = \left(s_1, s_2, a, (\boldsymbol{\rho}_k, \mathbf{A}_k, \mathbf{b}_k, \boldsymbol{\Sigma}_k)_{k\in\mathbb{N}^\star}\right), \text{ while } \boldsymbol{\Theta} = \left(\boldsymbol{\tau}, \alpha, \sigma, \boldsymbol{\theta}^*\right), \boldsymbol{\theta}^* = (\boldsymbol{\theta}_k^*)_{k\in\mathbb{N}^\star} \equiv (\mathbf{c}_k, \boldsymbol{\Gamma}_k)_{k\in\mathbb{N}^\star}.$$

BNP-GLLiM2 can be represented graphically as in Figure C1. The joint distribution of the observed data $\mathcal{X}, \mathcal{Y}$ and all latent variables can be expressed hierarchically as

$$\begin{aligned} p(\mathcal{Y}, \mathcal{X}, \mathcal{Z}, \boldsymbol{\Theta}; \boldsymbol{\phi}) &= \prod_{n=1}^{N} p\left(\mathbf{y}_n \mid \mathbf{x}_n, z_n, \boldsymbol{\Theta}; \boldsymbol{\phi}\right) p\left(\mathbf{x}_n \mid z_n, \boldsymbol{\Theta}; \boldsymbol{\phi}\right) p(\mathcal{Z} \mid \boldsymbol{\Theta}; \boldsymbol{\phi}) p\left(\boldsymbol{\Theta}; \boldsymbol{\phi}\right) \\ &= \prod_{n=1}^{N} p\left(\mathbf{y}_n \mid \mathbf{x}_n, z_n; \mathbf{A}, \mathbf{b}, \boldsymbol{\Sigma}\right) p\left(\mathbf{x}_n \mid z_n, \mathbf{c}, \boldsymbol{\Gamma}\right) p(\mathcal{Z} \mid \boldsymbol{\tau}) \\ &\quad \prod_{k\in\mathbb{N}^\star} p\left(\tau_k \mid \alpha, \sigma\right) p\left(\alpha, \sigma \mid s_1, s_2, a\right) \prod_{k\in\mathbb{N}^\star} p\left(\mathbf{c}_k, \boldsymbol{\Gamma}_k; \boldsymbol{\rho}_k\right). \end{aligned} \quad \text{(C2)}$$

Following the same idea as in Section 3, we only consider the truncated variational posterior of parameters $\boldsymbol{\Theta}$ as follows

$$q_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}) = q_{\alpha,\sigma}(\alpha, \sigma) \prod_{k=1}^{K-1} q_{\tau_k}\left(\tau_k\right) \prod_{k=1}^{K} q_{\boldsymbol{\theta}_k^*}\left(\boldsymbol{\theta}_k^*\right). \quad \text{(C3)}$$

These forms of $q_{\mathcal{Z}}$ and $q_{\boldsymbol{\Theta}}$ lead to our four VB-E steps and three VB-M steps, summarized below with details in Appendix A. Set the initial value of $\boldsymbol{\phi}$ to $\boldsymbol{\phi}^{(0)}$.

Then, repeat iteratively the following steps. The iteration index is omitted in the update formulas for simplicity.

VB-E steps

Note that the VB-E-$\boldsymbol{\tau}$, VB-E-(α, σ) steps are the same as in Section 3. We only highlight the modified steps as follows.

We first consider the derivation of the update equation for the factor $q_{\mathcal{Z}}(\mathcal{Z})$.

C.1.1. VB-E-$\mathcal{Z}$ step

By using the mean-field approximation (32) and the truncation, see Appendix A.4 for more details, for all $n \in [N]$ and for $k \in [K]$, this step consists in computing,

$$q_{z_n}(k) = \frac{\rho_{nk}}{\sum_{l=1}^{K} \rho_{nl}}. \tag{C4}$$

Here, given $\mathcal{N}_n$ represents the neighbors of n, we define $\log \rho_{nk}$ by

$$\begin{aligned} &-\frac{1}{2}\Bigg\{\log\left|\widehat{\boldsymbol{\Sigma}}_k\right| + (\mathbf{y}_n - \widehat{\mathbf{A}}_k\mathbf{x}_n - \widehat{\mathbf{b}}_k)^\top \widehat{\boldsymbol{\Sigma}}_k^{-1}(\mathbf{y}_n - \widehat{\mathbf{A}}_k\mathbf{x}_n - \widehat{\mathbf{b}}_k) \\ &\quad + \log\left|\frac{\widehat{\boldsymbol{\Psi}}_k}{2}\right| - \sum_{l=1}^{L}\psi\left(\frac{\widehat{\nu}_k + (1-l)}{2}\right) + \widehat{\nu}_k(\mathbf{x}_n - \widehat{\mathbf{m}}_k)^\top \widehat{\boldsymbol{\Psi}}_k^{-1}(\mathbf{x}_n - \widehat{\mathbf{m}}_k) + \frac{L}{\widehat{\lambda}_k}\Bigg\} \\ &\quad + \psi(\widehat{\gamma}_{k,1}) - \psi(\widehat{\gamma}_{k,1} + \widehat{\gamma}_{k,2}) + \sum_{l=1}^{k-1}\left[\psi(\widehat{\gamma}_{l,2}) - \psi(\widehat{\gamma}_{l,1} + \widehat{\gamma}_{l,2})\right]. \end{aligned} \tag{C5}$$

Note that in the above formula, symbols $\left(\widehat{\mathbf{m}}_k, \widehat{\lambda}_k, \widehat{\boldsymbol{\Psi}}_k, \widehat{\nu}_k\right)$ and $\left(\widehat{\mathbf{A}}_k, \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right)$ are the hyperparameters more specifically defined in the following Appendix C.1.2 and Section 3.2.

Proof of (C5)**.** With respect to the VBEM for BNP-GLLiM2 model from Appendix C.1.1, it is almost similar to the previous step in Appendix A.4, except that we have to take into account the randomness of $\mathbf{c}$ and $\boldsymbol{\Gamma}$. Namely, we have

$$\begin{aligned} &q_{z_n}(z_n) \\ &\propto \exp \mathbb{E}_{q_{\boldsymbol{\Theta}}}\left[\log\left(p\left(\mathbf{y}_n \mid \mathbf{x}_n, z_n; \widehat{\mathbf{A}}_{z_n}, \widehat{\mathbf{b}}_{z_n}, \widehat{\boldsymbol{\Sigma}}_{z_n}\right) p\left(\mathbf{x}_n \mid z_n, \mathbf{c}_{z_n}, \boldsymbol{\Gamma}_{z_n}\right) p(\mathbf{z} \mid \boldsymbol{\tau})\right)\right] \\ &= \exp\left\{\log p\left(\mathbf{y}_n \mid \mathbf{x}_n, z_n; \widehat{\mathbf{A}}_{z_n}, \widehat{\mathbf{b}}_{z_n}, \widehat{\boldsymbol{\Sigma}}_{z_n}\right) + \mathbb{E}_{q_{\boldsymbol{\theta}^*_{z_n}}}\left[\log p\left(\mathbf{x}_n \mid z_n, \mathbf{c}_{z_n}, \boldsymbol{\Gamma}_{z_n}\right)\right] + \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\log \pi_{z_n}(\boldsymbol{\tau})\right]\right\}. \end{aligned} \tag{C6}$$

Here, for $z_n = k$, it holds that

$$\begin{aligned} &\mathbb{E}_{q_{\boldsymbol{\theta}^*_{z_n}}}\left[\log p\left(\mathbf{x}_n \mid z_n, \widehat{\mathbf{c}}_{z_n}, \widehat{\boldsymbol{\Gamma}}_{z_n}\right)\right] = \mathbb{E}_{q_{\boldsymbol{\theta}^*_{z_n}}}\left[\log \mathcal{N}_L\left(\mathbf{x}_n \mid \widehat{\mathbf{c}}_k, \widehat{\boldsymbol{\Gamma}}_k\right)\right] \\ &= -\frac{L}{2}\log(2\pi) - \frac{1}{2}\mathbb{E}_{q_{\boldsymbol{\Gamma}_k}}\left[\log\left|\widehat{\boldsymbol{\Gamma}}_k\right|\right] - \frac{1}{2}\mathbb{E}_{q_{\boldsymbol{\theta}^*_{z_n}}}\left[\left(\mathbf{x}_n - \widehat{\mathbf{c}}_k\right)^\top \widehat{\boldsymbol{\Gamma}}_k^{-1}\left(\mathbf{x}_n - \widehat{\mathbf{c}}_k\right)\right], \end{aligned}$$

where we used the fact that

$$\mathbb{E}_{q_{\boldsymbol{\Gamma}_k}}\left[\log\left|\widehat{\boldsymbol{\Gamma}}_k\right|\right] = \log\left|\frac{\widehat{\boldsymbol{\Psi}}_k}{2}\right| - \sum_{l=1}^{L}\psi\left(\frac{\widehat{\nu}_k + (1-l)}{2}\right),$$

$$\mathbb{E}_{q_{\boldsymbol{\theta}^*_{z_n}}}\left[(\mathbf{x}_n - \widehat{\mathbf{c}}_k)^\top \widehat{\boldsymbol{\Gamma}}_k^{-1}(\mathbf{x}_n - \widehat{\mathbf{c}}_k)\right] = \widehat{\nu}_k(\mathbf{x}_n - \widehat{\mathbf{m}}_k)^\top \widehat{\boldsymbol{\Psi}}_k^{-1}(\mathbf{x}_n - \widehat{\mathbf{m}}_k) + \frac{L}{\widehat{\lambda}_k}.$$

Plugging in all of the above expression back into (C6) yields the desired results in (C5). □

C.1.2. VB-E-$\boldsymbol{\theta}^*$ step

This step is divided into K parts where the computation is similar to that in standard Bayesian GMM with a choice of conjugate prior. Hence, for each $k \leq K$, the variational posterior is a Normal-inverse-Wishart distribution defined as

$$q_{\boldsymbol{\theta}^*_k}(\mathbf{c}_k, \boldsymbol{\Gamma}_k) = \mathcal{NIW}(\mathbf{c}_k, \boldsymbol{\Gamma}_k \mid \widehat{\mathbf{m}}_k, \widehat{\lambda}_k, \widehat{\boldsymbol{\Psi}}_k, \widehat{\nu}_k). \tag{C7}$$

Here, the hyperparameters are updated as follows (see, *e.g.*, [9, Section 10.2.1]):

$$\widehat{\lambda}_k = \lambda_k + N_k, \quad \widehat{\nu}_k = \nu_k + N_k, N_k = \sum_{n=1}^{N} q_{z_n}(k) \tag{C8}$$

$$\widehat{\boldsymbol{\Psi}}_k = \boldsymbol{\Psi}_k + N_k \boldsymbol{S}_k + \frac{\lambda_k N_k}{\lambda_k + N_k}(\mathbf{m}_k - \bar{\mathbf{c}}_k)(\mathbf{m}_k - \bar{\mathbf{c}}_k)^\top, \tag{C9}$$

$$\widehat{\mathbf{m}}_k = \frac{\lambda_k \mathbf{m}_k + N_k \bar{\mathbf{c}}_k}{\lambda_k + N_k} = \frac{\lambda_k \mathbf{m}_k + N_k \bar{\mathbf{c}}_k}{\widehat{\lambda}_k}, \tag{C10}$$

$$\bar{\mathbf{c}}_k = \frac{1}{N_k}\sum_{n=1}^{N} q_{z_n}(k)\mathbf{x}_n, \tag{C11}$$

$$\boldsymbol{S}_k = \frac{1}{N_k}\sum_{n=1}^{N} q_{z_n}(k)(\mathbf{x}_n - \bar{\mathbf{c}}_k)(\mathbf{x}_n - \bar{\mathbf{c}}_k)^\top. \tag{C12}$$

VB-M steps

The maximization step consists of updating the hyperparameters $\boldsymbol{\phi} = \left(s_1, s_2, a, (\boldsymbol{\rho}_k, \mathbf{A}_k, \mathbf{b}_k, \boldsymbol{\Sigma}_k)_{k\in[K]}\right)$, where $\boldsymbol{\rho}_k = (\mathbf{m}_k, \lambda_k, \boldsymbol{\Psi}_k, \nu_k), k \in [K]$, by maximizing the free energy, if they are not set heuristically:

$$\boldsymbol{\phi}^{(r)} = \underset{\boldsymbol{\phi}}{\operatorname{argmax}}\ \mathbb{E}_{q^{(r)}_{\mathcal{Z}} q^{(r)}_{\boldsymbol{\tau}} q^{(r)}_{\alpha,\sigma} q^{(r)}_{\boldsymbol{\theta}^*}}\left[\log p\left(\mathcal{Y}, \mathcal{X}, \mathcal{Z}, \boldsymbol{\tau}, \alpha, \sigma, \boldsymbol{\theta}^*; \boldsymbol{\phi}\right)\right]. \tag{C13}$$

The VB-M-step can therefore be divided into four independent sub-steps as listed below. From the conditional independence of $(s_1, s_2, a, \boldsymbol{\rho})$ and $(\mathcal{Y}, \mathcal{X}, \mathcal{Z})$ given $(\boldsymbol{\tau}, \alpha, \sigma, \boldsymbol{\theta}^*)$, the solutions for the VB-M-(s_1, s_2) (in the DP case) and VB-M-$\boldsymbol{\rho}$ steps are straightforward. Only the M-(s_1, s_2, a) step (in the PYP case) and $(\mathbf{A}_k, \mathbf{b}_k, \boldsymbol{\Sigma}_k)_{k\in[K]}$ are more involved.

Note that the VB-M-(s_1, s_2, a), VB-M-$(\mathbf{A}, \mathbf{b}, \boldsymbol{\Sigma})$ steps are the same as in Section 3. We only highlight the modified step below.

C.1.3. VB-M-$\boldsymbol{\rho}$ step

This step divides into K sub-steps that involve again cross-entropies,

$$\rho_k^{(r)} = \underset{\rho}{\operatorname{argmax}}\ \mathbb{E}_{q_{\theta_k^*}^{(r)}}\left[\log p(\mathbf{c}_k, \boldsymbol{\Gamma}_k; \rho_k)\right] \ = \ \widehat{\rho}_k^{(r)} \tag{C14}$$

where $\widehat{\rho}_k^{(r)} = (\widehat{\lambda}_k^{(r)}, \widehat{\nu}_k^{(r)}, \widehat{\boldsymbol{\Psi}}_k^{(r)}, \widehat{\mathbf{m}}_k^{(r)})$ is given in (C8).

C.1.4. ELBO for BNP-GLLiM2

Proposition C.1. *When $\sigma = 0$, the ELBO in BNP-GLLiM2 is determined analytically as follows:*

$$\begin{aligned}\mathcal{F}\left(q_{\mathcal{Z}}, q_{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\phi}}\right) &= \mathbb{E}\left[\log p(\mathcal{Y} \mid \mathcal{X}, \mathcal{Z}, \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] + \mathbb{E}\left[\log p(\mathcal{X} \mid \mathcal{Z}, \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] + \mathbb{E}\left[\log p(\mathcal{Z} \mid \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] \\ &\quad + \mathbb{E}\left[\log p(\boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] - \mathbb{E}\left[\log q(\mathcal{Z})\right] - \mathbb{E}\left[\log q(\boldsymbol{\Theta})\right].\end{aligned} \tag{C15}$$

Here, we have the following update formulas:

$$\mathbb{E}\left[\log p(\mathcal{Y} \mid \mathcal{X}, \mathcal{Z}, \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] = \sum_{n=1}^{N}\sum_{k=1}^{K} q_{z_n}(k) \log \mathcal{N}_D\left(\mathbf{y}_n \mid \widehat{\mathbf{A}}_k \mathbf{x}_n + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right), \tag{C16}$$

$$\begin{aligned}&\mathbb{E}\left[\log p(\mathcal{X} \mid \mathcal{Z}, \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] \\ &= \frac{1}{2}\sum_{k=1}^{K} N_k \left[\log \widetilde{\boldsymbol{\Gamma}}_k - L\log(2\pi) - L\widehat{\lambda}_k^{-1} - \widehat{\nu}_k\, Tr\left(\mathbf{S}_k \widehat{\boldsymbol{\Psi}}_k^{-1}\right) - \widehat{\nu}_k\left(\bar{\mathbf{x}}_k - \widehat{\mathbf{m}}_k\right)\widehat{\boldsymbol{\Psi}}_k^{-1}\left(\bar{\mathbf{x}}_k - \widehat{\mathbf{m}}_k\right)\right],\end{aligned} \tag{C17}$$

$$\mathbb{E}\left[\log p(\mathcal{Z} \mid \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] = \sum_{k=1}^{K} N_k \left[\psi\left(\widehat{\gamma}_{k,1}\right) - \psi\left(\widehat{\gamma}_{k,1} + \widehat{\gamma}_{k,2}\right) + \sum_{l=1}^{k-1}\left[\psi\left(\widehat{\gamma}_{l,2}\right) - \psi\left(\widehat{\gamma}_{l,1} + \widehat{\gamma}_{l,2}\right)\right]\right], \tag{C18}$$

$$\mathbb{E}\left[\log p(\boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] = \sum_{k=1}^{K-1}\mathbb{E}\left[\log p\left(\tau_k \mid \alpha\right)\right] + \mathbb{E}\left[\log p\left(\alpha \mid \widehat{s}_1, \widehat{s}_2\right)\right] + \sum_{k=1}^{K}\mathbb{E}\left[\log p\left(\mathbf{c}_k, \boldsymbol{\Gamma}_k; \widehat{\boldsymbol{\rho}}_k\right)\right], \tag{C19}$$

$$\mathbb{E}\left[\log p\left(\tau_k \mid \alpha\right)\right] = \frac{\widehat{s}_1 - \widehat{s}_2}{\widehat{s}_2}\left[\psi\left(\widehat{\gamma}_{k,2}\right) - \psi\left(\widehat{\gamma}_{k,1} + \widehat{\gamma}_{k,2}\right)\right] + \psi\left(\widehat{s}_1\right) - \log(\widehat{s}_2),$$

$$\mathbb{E}\left[\log p\left(\alpha \mid \widehat{s}_1, \widehat{s}_2\right)\right] = -\log\Gamma\left(\widehat{s}_1\right) + \left(\widehat{s}_1 - 1\right)\psi\left(\widehat{s}_1\right) + \log(\widehat{s}_2) - \widehat{s}_1,$$

$$\mathbb{E}\left[\log p\left(\mathbf{c}_k, \boldsymbol{\Gamma}_k; \widehat{\boldsymbol{\rho}}_k\right)\right] = \frac{1}{2}L\log\left(\frac{\widehat{\lambda}_k}{2\pi}\right) - \frac{L}{2} - \frac{L}{2}\widehat{\nu}_k + \log B\left(\widehat{\boldsymbol{\Psi}}_k^{-1}, \widehat{\nu}_k\right) - \frac{\widehat{\nu}_k - L}{2}\log\widetilde{\boldsymbol{\Gamma}}_k,$$

$$\log\widetilde{\boldsymbol{\Gamma}}_k = \sum_{l=1}^{L}\psi\left(\frac{\widehat{\nu}_k + 1 - l}{2}\right) + L\log 2 + \log\left|\widehat{\boldsymbol{\Psi}}_k\right|,$$

$$\mathbb{E}\left[\log q(\mathcal{Z})\right] = \sum_{n=1}^{N}\sum_{k=1}^{K} q_{z_n}\left(k\right)\log q_{z_n}\left(k\right), \tag{C20}$$

$$\mathbb{E}\left[\log q(\boldsymbol{\Theta})\right] = \mathbb{E}\left[\log q_{\alpha,0}(\alpha)\right] + \sum_{k=1}^{K-1}\mathbb{E}\left[\log q_{\tau_k}\left(\tau_k\right)\right] + \sum_{k=1}^{K}\mathbb{E}\left[\log q_{\mathbf{c}_k,\boldsymbol{\Gamma}_k}\left(\mathbf{c}_k, \boldsymbol{\Gamma}_k\right)\right], \tag{C21}$$

$$\mathbb{E}\left[\log q_{\alpha,0}(\alpha)\right] = -\log\Gamma\left(\widehat{s}_1\right) + \left(\widehat{s}_1 - 1\right)\psi\left(\widehat{s}_1\right) + \log(\widehat{s}_2) - \widehat{s}_1,$$

$$\mathbb{E}\left[\log q_{\tau_k}\left(\tau_k\right)\right] = \sum_{l=1}^{2}\left(\widehat{\gamma}_{k,l} - 1\right)\left\{\psi\left(\widehat{\gamma}_{k,l}\right) - \psi\left(\widehat{\gamma}_{k,1} + \widehat{\gamma}_{k,2}\right)\right\} + \log\frac{\Gamma\left(\widehat{\gamma}_{k,1} + \widehat{\gamma}_{k,2}\right)}{\Gamma\left(\widehat{\gamma}_{k,1}\right)\Gamma\left(\widehat{\gamma}_{k,2}\right)},$$

$$\mathbb{E}\left[\log q_{\mathbf{c}_k,\boldsymbol{\Gamma}_k}\left(\mathbf{c}_k, \boldsymbol{\Gamma}_k\right)\right] = \frac{L}{2}\log\frac{\widehat{\lambda}_k}{2\pi} - \frac{L}{2} + \log B\left(\widehat{\boldsymbol{\Psi}}_k^{-1}, \widehat{\nu}_k\right) - \frac{\widehat{\nu}_k - L}{2}\log\widetilde{\boldsymbol{\Gamma}}_k - \frac{\widehat{\nu}_k L}{2}. \tag{C22}$$

C.2. Predictive conditional density for BNP-GLLiM2

C.2.1. Joint density

We first show how to compute the joint density $p(\widehat{\mathbf{y}}, \widehat{\mathbf{x}}, \mathcal{X}, \mathcal{Y})$ via Theorem C.2, which is proved in Appendix C.4.

Theorem C.2. *We approximate the joint density of BNP-GLLiM2 by a mixture of product between Gaussian and Student's t-distributions as follows:*

$$p\left(\widehat{\mathbf{y}}, \widehat{\mathbf{x}}, \mathcal{X}, \mathcal{Y}\right) \approx \sum_{k=1}^{K}\mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k\left(\boldsymbol{\tau}\right)\right] St\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{m}}_k, \mathbf{L}_k, \widehat{\nu}_k + 1 - L\right)\mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{A}}_k\widehat{\mathbf{x}} + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right). \tag{C23}$$

Here, the positive semidefinite shape matrices of Student's t-distributions are given by

$$\mathbf{L}_k = \frac{\left(\widehat{\nu}_k + 1 - L\right)\widehat{\lambda}_k}{1 + \widehat{\lambda}_k}\widehat{\boldsymbol{\Psi}}_k. \tag{C24}$$

C.2.2. Inverse conditional density

We then show how to approximate the inverse conditional density $p(\widehat{\mathbf{y}} \mid \widehat{\mathbf{x}}, \mathcal{X}, \mathcal{Y})$. This predictive density in BNP-GLLiM2 is approximated by a MoE via Theorem C.3 with the proof in Appendix C.5.

Theorem C.3. *We approximate the inverse conditional density of BNP-GLLiM2 by a MoE as follows:*

$$p\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{x}}, \mathcal{X}, \mathcal{Y}\right) \approx \sum_{k=1}^{K} g_k\left(\widehat{\mathbf{x}} \mid \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\phi}}, \mathcal{X}, \mathcal{Y}\right) \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{A}}_k \widehat{\mathbf{x}} + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right). \tag{C25}$$

Here, the gating posteriors are defined as

$$g_k\left(\widehat{\mathbf{x}} \mid \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\phi}}, \mathcal{X}, \mathcal{Y}\right) = \frac{\mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k\left(\boldsymbol{\tau}\right)\right] St\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{m}}_k, \mathbf{L}_k, \widehat{\nu}_k + 1 - L\right)}{\sum_{l=1}^{K} \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_l\left(\boldsymbol{\tau}\right)\right] St\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{m}}_l, \mathbf{L}_l, \widehat{\nu}_l + 1 - L\right)}, \quad k \in [K]. \tag{C26}$$

Furthermore, for any $k \in [K]$, it holds that

$$\mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k\left(\boldsymbol{\tau}\right)\right] = \mathbb{E}_{q_{\tau_k}}\left[\tau_k\right] \prod_{l=1}^{k-1} \mathbb{E}_{q_{\tau_l}}\left[1 - \tau_l\right],$$

$$\mathbb{E}_{q_{\tau_k}}\left[\tau_k\right] = \frac{\widehat{\gamma}_{k,1}}{\widehat{\gamma}_{k,1} + \widehat{\gamma}_{k,2}}, \quad \mathbb{E}_{q_{\tau_k}}\left[1 - \tau_k\right] = 1 - \mathbb{E}_{q_{\tau_k}}\left[\tau_k\right] = \frac{\widehat{\gamma}_{k,2}}{\widehat{\gamma}_{k,1} + \widehat{\gamma}_{k,2}},$$

$$\widehat{\gamma}_{k,1} = 1 - \mathbb{E}_{q_{\alpha,\sigma}}\left[\sigma\right] + N_k, \quad \widehat{\gamma}_{k,2} = \mathbb{E}_{q_{\alpha,\sigma}}\left[\alpha\right] + k\mathbb{E}_{q_{\alpha,\sigma}}\left[\sigma\right] + \sum_{l=k+1}^{K} N_l, N_k = \sum_{n=1}^{n} q_{z_n}(k).$$

The prediction task is carried out via the following approximation

$$\mathbb{E}\left[\widehat{\mathbf{y}} \mid \widehat{\mathbf{x}}, \mathcal{X}, \mathcal{Y}\right] \approx \sum_{k=1}^{K} g_k\left(\widehat{\mathbf{x}} \mid \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\phi}}, \mathcal{X}, \mathcal{Y}\right) \left[\widehat{\mathbf{A}}_k \widehat{\mathbf{x}} + \widehat{\mathbf{b}}_k\right]. \tag{C27}$$

C.2.3. Forward conditional density

To deal with high-dimensional regression data, namely high-to-low regression, given the inverse conditional density $p(\widehat{\mathbf{y}} \mid \widehat{\mathbf{x}}, \mathcal{X}, \mathcal{Y})$, we want to approximate the following forward conditional density via Theorem C.4, whose proof is provided in Appendix C.6.

Theorem C.4. *It holds that*

$$p(\widehat{\mathbf{x}} \mid \widehat{\mathbf{y}}, \mathcal{X}, \mathcal{Y}) \approx \sum_{i=1}^{I} \sum_{k=1}^{K} g_{ki}\left(\widehat{\mathbf{y}} \mid \widehat{\boldsymbol{\Theta}}^*, \widehat{\boldsymbol{\phi}}^*, \mathcal{X}, \mathcal{Y}\right) \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{A}}_k^*\left(\eta_i\right) \widehat{\mathbf{y}} + \widehat{\mathbf{b}}_k^*\left(\eta_i\right), \widehat{\boldsymbol{\Sigma}}_k^*\left(\eta_i\right)\right),$$

which is a mixture of Gaussian experts, where, for all $k \in [K], i \in [I]$,

$$
\begin{aligned}
g_{ki}\left(\widehat{\mathbf{y}} \mid \widehat{\boldsymbol{\Theta}}^*, \widehat{\boldsymbol{\phi}}^*, \mathcal{X}, \mathcal{Y}\right) &= \frac{\mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k\left(\boldsymbol{\tau}\right)\right] \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{c}}_k^*, \widehat{\boldsymbol{\Gamma}}_k^*\left(\eta_i\right)\right) Gam\left(\eta_i \mid \frac{\widehat{\nu}_k+1-L}{2}, \frac{\widehat{\nu}_k+1-L}{2}\right)}{\sum_{i=1}^{I}\sum_{l=1}^{K} \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_l\left(\boldsymbol{\tau}\right)\right] \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{c}}_l^*, \widehat{\boldsymbol{\Gamma}}_l^*\left(\eta_i\right)\right) Gam\left(\eta_i \mid \frac{\widehat{\nu}_l+1-L}{2}, \frac{\widehat{\nu}_l+1-L}{2}\right)}, \\
\widehat{\boldsymbol{\Sigma}}_k^*\left(\eta_i\right) &= \left(\eta_i \mathbf{L}_k + \widehat{\mathbf{A}}_k^\top \widehat{\boldsymbol{\Sigma}}_k^{-1} \widehat{\mathbf{A}}_k\right)^{-1}, \\
\widehat{\mathbf{A}}_k^*\left(\eta_i\right) &= \widehat{\boldsymbol{\Sigma}}_k^*\left(\eta_i\right) \widehat{\mathbf{A}}_k^\top \widehat{\boldsymbol{\Sigma}}_k^{-1}, \\
\widehat{\mathbf{b}}_k^*\left(\eta_i\right) &= \widehat{\boldsymbol{\Sigma}}_k^*\left(\eta_i\right) \left[\eta \mathbf{L}_k \widehat{\mathbf{m}}_k - \widehat{\mathbf{A}}_k^\top \widehat{\boldsymbol{\Sigma}}_k^{-1} \widehat{\mathbf{b}}_k\right], \\
\widehat{\mathbf{c}}_k^* &= \widehat{\mathbf{A}}_k \widehat{\mathbf{m}}_k + \widehat{\mathbf{b}}_k, \\
\widehat{\boldsymbol{\Gamma}}_k^*\left(\eta_i\right) &= \widehat{\boldsymbol{\Sigma}}_k + \widehat{\mathbf{A}}_k \left(\eta_i \mathbf{L}_k\right)^{-1} \widehat{\mathbf{A}}_k^\top.
\end{aligned} \tag{C28}
$$

Here, $\eta_i, i \in [I]$, *are chosen via discretizing* η*-space,* $[0, U_\eta]$, *into a grid,* e.g., *uniform. Note that for simplicity, we evaluate the integrand as a Riemann integral with a truncated value* $0 < U_\eta < \infty$ *and a number of point* $I \in \mathbb{N}^\star$ *for approximating the integration but we can use any scheme to approximate such* 1*-dimensional integration.*

C.3. Proof of Proposition C.1

Using the sum and product rules for both discrete and continuous variables, the ELBO in BNP-GLLiM (A4) is given by

$$
\begin{aligned}
\mathcal{F}\left(q_{\mathcal{Z}}, q_{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\phi}}\right) &= \mathbb{E}_{q_{\mathcal{Z}} q_{\boldsymbol{\Theta}}}\left[\log \frac{p(\mathcal{Y}, \mathcal{X}, \mathcal{Z}, \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})}{q(\mathcal{Z}) q_{\boldsymbol{\Theta}}(\boldsymbol{\Theta})}\right] \equiv \mathbb{E}\left[\log \frac{p(\mathcal{Y}, \mathcal{X}, \mathcal{Z}, \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})}{q(\mathcal{Z}) q(\boldsymbol{\Theta})}\right] \\
&= \sum_{\mathcal{Z}} \int\int\int q(\mathcal{Z}) q(\boldsymbol{\Theta}) \log\left[\frac{p(\mathcal{Y}, \mathcal{X}, \mathcal{Z}, \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})}{q(\mathcal{Z}) q(\boldsymbol{\Theta})}\right] d\mathcal{Z} d\boldsymbol{\Theta} \\
&= \mathbb{E}\left[\log p(\mathcal{Y} \mid \mathcal{X}, \mathcal{Z}, \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] + \mathbb{E}\left[\log p(\mathcal{X} \mid \mathcal{Z}, \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] \\
&\quad + \mathbb{E}\left[\log p(\mathcal{Z} \mid \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] + \mathbb{E}\left[\log p(\boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] - \mathbb{E}\left[\log q(\mathcal{Z})\right] - \mathbb{E}\left[\log q(\boldsymbol{\Theta})\right].
\end{aligned} \tag{C29}
$$

Note that the proof of (C16), (C18), (C20) are the same as in the proof of Proposition 3.1.

Proof of (C17)

$$\mathbb{E}\left[\log p(\mathcal{X} \mid \mathcal{Z}, \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] = \mathbb{E}\left[\log \prod_{n=1}^{N} p\left(\mathbf{x}_n \mid z_n, \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}}\right)\right] = \mathbb{E}\left[\log \prod_{n=1}^{N}\prod_{k=1}^{K} \mathcal{N}_L\left(\mathbf{x}_n \mid \boldsymbol{c}_k, \boldsymbol{\Gamma}_k\right)^{z_{nk}}\right]$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} \mathbb{E}\left[z_{nk} \log \mathcal{N}_L\left(\mathbf{x}_n \mid \boldsymbol{c}_k, \boldsymbol{\Gamma}_k\right)\right]$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} \mathbb{E}_{q_{\mathcal{Z}}}\left[z_{nk}\right] \mathbb{E}_{q_{\boldsymbol{c}_k, \boldsymbol{\Gamma}_k}}\left[\log \mathcal{N}_L\left(\mathbf{x}_n \mid \boldsymbol{c}_k, \boldsymbol{\Gamma}_k\right)\right]$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} q_{z_n}(k)\left[-\frac{L}{2}\log(2\pi) - \frac{1}{2}\mathbb{E}\left[\log|\boldsymbol{\Gamma}_k|\right] - \frac{1}{2}\mathbb{E}\left[\left(\mathbf{x}_n - \boldsymbol{c}_k\right)^\top \boldsymbol{\Gamma}_k^{-1}\left(\mathbf{x}_n - \boldsymbol{c}_k\right)\right]\right]$$

$$= \sum_{k=1}^{K}\sum_{n=1}^{N} q_{z_n}(k)\left[-\frac{L}{2}\log(2\pi) - \frac{1}{2}\mathbb{E}_{q_{\boldsymbol{\Gamma}_k}}\left[\log|\boldsymbol{\Gamma}_k|\right] - \frac{1}{2}\mathbb{E}_{q_{\boldsymbol{c}_k, \boldsymbol{\Gamma}_k}}\left[\left(\mathbf{x}_n - \boldsymbol{c}_k\right)^\top \boldsymbol{\Gamma}_k^{-1}\left(\mathbf{x}_n - \boldsymbol{c}_k\right)\right]\right] \text{ (Lemma C.5)}$$

$$= \frac{1}{2}\sum_{k=1}^{K}\sum_{n=1}^{N} q_{z_n}(k)\left[-\log\widetilde{\boldsymbol{\Gamma}}_k - L\log(2\pi) - L\widehat{\lambda}_k^{-1} - \widehat{\nu}_k\left(\mathbf{x}_n - \widehat{\mathbf{m}}_k\right)\widehat{\boldsymbol{\Psi}}_k^{-1}\left(\mathbf{x}_n - \widehat{\mathbf{m}}_k\right)\right]$$

$$= \frac{1}{2}\sum_{k=1}^{K} N_k\left[\log\widetilde{\boldsymbol{\Gamma}}_k - L\log(2\pi) - L\widehat{\lambda}_k^{-1}\right] - \frac{1}{2}\sum_{k=1}^{K}\sum_{n=1}^{N} q_{z_n}(k)\left[\widehat{\nu}_k\left(\mathbf{x}_n - \widehat{\mathbf{m}}_k\right)\widehat{\boldsymbol{\Psi}}_k^{-1}\left(\mathbf{x}_n - \widehat{\mathbf{m}}_k\right)\right]$$

$$= \frac{1}{2}\sum_{k=1}^{K} N_k\left[\log\widetilde{\boldsymbol{\Gamma}}_k - L\log(2\pi) - L\widehat{\lambda}_k^{-1} - \widehat{\nu}_k \mathrm{Tr}\left(\mathbf{S}_k\widehat{\boldsymbol{\Psi}}_k^{-1}\right) - \widehat{\nu}_k\left(\bar{\mathbf{x}}_k - \widehat{\mathbf{m}}_k\right)\widehat{\boldsymbol{\Psi}}_k^{-1}\left(\bar{\mathbf{x}}_k - \widehat{\mathbf{m}}_k\right)\right]$$

$$\text{(using (C33) from Lemma C.5)}. \tag{C30}$$

To obtain (C30), we have to use the following Lemma C.5.

Lemma C.5. *We can compute the expectations w.r.t. the variational distributions of the parameters as follows:*

$$\log\widetilde{\boldsymbol{\Gamma}}_k \equiv \mathbb{E}_{q_{\boldsymbol{\Gamma}_k}}\left[\log|\boldsymbol{\Gamma}_k|\right] = \sum_{l=1}^{L}\psi\left(\frac{\widehat{\nu}_k + 1 - l}{2}\right) + L\log 2 + \log\left|\widehat{\boldsymbol{\Psi}}_k\right|, \tag{C31}$$

$$\mathbb{E}_{q_{\boldsymbol{c}_k, \boldsymbol{\Gamma}_k}}\left[\left(\mathbf{x}_n - \boldsymbol{c}_k\right)^\top \boldsymbol{\Gamma}_k^{-1}\left(\mathbf{x}_n - \boldsymbol{c}_k\right)\right] = L\widehat{\lambda}_k^{-1} + \widehat{\nu}_k\left(\mathbf{x}_n - \widehat{\mathbf{m}}_k\right)\widehat{\boldsymbol{\Psi}}_k^{-1}\left(\mathbf{x}_n - \widehat{\mathbf{m}}_k\right). \tag{C32}$$

Furthermore, for each $k \in [K]$, it holds that

$$\sum_{n=1}^{N} q_{z_n}(k)\left[\widehat{\nu}_k\left(\mathbf{x}_n - \widehat{\mathbf{m}}_k\right)\widehat{\boldsymbol{\Psi}}_k^{-1}\left(\mathbf{x}_n - \widehat{\mathbf{m}}_k\right)\right] = N_k\left[\widehat{\nu}_k Tr\left(\mathbf{S}_k\widehat{\boldsymbol{\Psi}}_k^{-1}\right) + \widehat{\nu}_k\left(\bar{\mathbf{x}}_k - \widehat{\mathbf{m}}_k\right)\widehat{\boldsymbol{\Psi}}_k^{-1}\left(\bar{\mathbf{x}}_k - \widehat{\mathbf{m}}_k\right)\right]. \tag{C33}$$

Proof of (C19)

Given a chosen truncated value $K \in \mathbb{N}^\star$, it holds that

$$\mathbb{E}_{q_\Theta}\left[\log p(\boldsymbol{\Theta};\widehat{\boldsymbol{\phi}})\right] = \sum_{k=1}^{K-1} \mathbb{E}_{q_\Theta}\left[\log p\left(\tau_k \mid \alpha,\sigma\right)\right] + \mathbb{E}_{q_\Theta}\left[\log p\left(\alpha,\sigma \mid \widehat{s}_1,\widehat{s}_2,\widehat{a}\right)\right] + \sum_{k=1}^{K} \mathbb{E}_{q_\Theta}\left[\log p\left(\mathbf{c}_k,\boldsymbol{\Gamma}_k;\widehat{\boldsymbol{\rho}}_k\right)\right].$$

Note that $\mathbb{E}_{q_\Theta}\left[\log p\left(\tau_k \mid \alpha,\sigma\right)\right]$ and $\mathbb{E}_{q_\Theta}\left[\log p\left(\alpha,\sigma \mid \widehat{s}_1,\widehat{s}_2,\widehat{a}\right)\right]$ are calculated in the same way as in Proposition 3.1.

Finally, we have to compute the remaining term

$$\begin{aligned}
&\mathbb{E}_{q_\Theta}\left[\log p\left(\mathbf{c}_k,\boldsymbol{\Gamma}_k;\widehat{\boldsymbol{\rho}}_k\right)\right] \\
&= \mathbb{E}_{q_{\mathbf{c}_k,\boldsymbol{\Gamma}_k}}\left[-\frac{L}{2}\log(2\pi) - \frac{1}{2}\log \widehat{\lambda}_k^{-L}\left|\boldsymbol{\Gamma}_k\right| - \frac{1}{2}\left(\boldsymbol{c}_k - \widehat{\mathbf{m}}_k\right)^\top \left(\widehat{\lambda}_k^{-1}\boldsymbol{\Gamma}_k\right)^{-1}\left(\boldsymbol{c}_k - \widehat{\mathbf{m}}_k\right)\right] \\
&\quad + \mathbb{E}_{q_{\boldsymbol{\Gamma}_k}}\left[\log \mathcal{W}\left(\boldsymbol{\Gamma}_k^{-1} \mid \widehat{\boldsymbol{\Psi}}_k^{-1},\widehat{\nu}_k\right)\right] \\
&= -\frac{1}{2}L\log(2\pi) + \frac{1}{2}L\log\widehat{\lambda}_k - \frac{1}{2}\mathbb{E}_{q_{\boldsymbol{\Gamma}_k}}\left[\log\left|\boldsymbol{\Gamma}_k\right|\right] - \frac{1}{2}\widehat{\lambda}_k\mathbb{E}_{q_{\mathbf{c}_k,\boldsymbol{\Gamma}_k}}\left[\left(\boldsymbol{c}_k - \widehat{\mathbf{m}}_k\right)^\top \boldsymbol{\Gamma}_k^{-1}\left(\boldsymbol{c}_k - \widehat{\mathbf{m}}_k\right)\right] \\
&\quad + \mathbb{E}_{q_{\boldsymbol{\Gamma}_k}}\left[\log B\left(\widehat{\boldsymbol{\Psi}}_k^{-1},\widehat{\nu}_k\right) + \frac{\widehat{\nu}_k - L - 1}{2}\log\left|\boldsymbol{\Gamma}_k^{-1}\right| - \frac{1}{2}\mathrm{Tr}\left(\widehat{\boldsymbol{\Psi}}_k\boldsymbol{\Gamma}_k^{-1}\right)\right] \\
&= \frac{1}{2}L\log\left(\frac{\widehat{\lambda}_k}{2\pi}\right) - \frac{1}{2}\mathbb{E}_{q_{\boldsymbol{\Gamma}_k}}\left[\log\left|\boldsymbol{\Gamma}_k\right|\right] - \frac{1}{2}\widehat{\lambda}_k\mathbb{E}_{q_{\mathbf{c}_k,\boldsymbol{\Gamma}_k}}\left[\left(\boldsymbol{c}_k - \widehat{\mathbf{m}}_k\right)^\top \boldsymbol{\Gamma}_k^{-1}\left(\boldsymbol{c}_k - \widehat{\mathbf{m}}_k\right)\right] \\
&\quad + \log B\left(\widehat{\boldsymbol{\Psi}}_k^{-1},\widehat{\nu}_k\right) - \frac{\widehat{\nu}_k - L - 1}{2}\mathbb{E}_{q_{\boldsymbol{\Gamma}_k}}\left[\log\left|\boldsymbol{\Gamma}_k\right|\right] - \frac{1}{2}\mathrm{Tr}\left(\widehat{\boldsymbol{\Psi}}_k\mathbb{E}_{q_{\boldsymbol{\Gamma}_k}}\left[\boldsymbol{\Gamma}_k^{-1}\right)\right] \\
&= \frac{1}{2}L\log\left(\frac{\widehat{\lambda}_k}{2\pi}\right) - \frac{1}{2}\widehat{\lambda}_k\left[L\widehat{\lambda}_k^{-1} + \widehat{\nu}_k\left(\widehat{\mathbf{m}}_k - \widehat{\mathbf{m}}_k\right)^\top \widehat{\boldsymbol{\Psi}}_k^{-1}\left(\widehat{\mathbf{m}}_k - \widehat{\mathbf{m}}_k\right)\right] \\
&\quad + \log B\left(\widehat{\boldsymbol{\Psi}}_k^{-1},\widehat{\nu}_k\right) - \frac{\widehat{\nu}_k - L}{2}\log\widetilde{\boldsymbol{\Gamma}}_k - \frac{1}{2}\widehat{\nu}_k\mathrm{Tr}\left(\widehat{\boldsymbol{\Psi}}_k\widehat{\boldsymbol{\Psi}}_k^{-1}\right) \text{ (using Lemma C.5)} \\
&= \frac{1}{2}L\log\left(\frac{\widehat{\lambda}_k}{2\pi}\right) - \frac{L}{2} - \frac{L}{2}\widehat{\nu}_k + \log B\left(\widehat{\boldsymbol{\Psi}}_k^{-1},\widehat{\nu}_k\right) - \frac{\widehat{\nu}_k - L}{2}\log\widetilde{\boldsymbol{\Gamma}}_k,
\end{aligned}$$

where

$$\log\widetilde{\boldsymbol{\Gamma}}_k \equiv \mathbb{E}_{q_{\boldsymbol{\Gamma}_k}}\left[\log\left|\boldsymbol{\Gamma}_k\right|\right] = \sum_{l=1}^{L}\psi\left(\frac{\widehat{\nu}_k + 1 - l}{2}\right) + L\log 2 + \log\left|\widehat{\boldsymbol{\Psi}}_k\right|. \tag{C34}$$

Proof of (C21)

We have

$$\mathbb{E}\left[\log q(\boldsymbol{\Theta})\right] = \mathbb{E}\left[\log q_{\alpha,\sigma}(\alpha,\sigma)\right] + \sum_{k=1}^{K-1}\mathbb{E}\left[\log q_{\tau_k}\left(\tau_k\right)\right] + \sum_{k=1}^{K}\mathbb{E}\left[\log q_{\mathbf{c}_k,\boldsymbol{\Gamma}_k}\left(\mathbf{c}_k,\boldsymbol{\Gamma}_k\right)\right]. \tag{C35}$$

Note that these terms involving expectations of the logs of the q distributions simply represent the negative entropies of those distributions. In particular, the first two terms are calculated in the same way as in Proposition 3.1.

Similarly, we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\log q_{\mathbf{c}_k,\mathbf{\Gamma}_k}\left(\mathbf{c}_k,\mathbf{\Gamma}_k\right)\right] \\
&= \mathbb{E}\left[\log \mathcal{N}_L\left(\mathbf{c}_k \mid \widehat{\mathbf{m}}_k, \widehat{\lambda}_k^{-1}\mathbf{\Gamma}_k\right)\right] + \mathbb{E}\left[\log \mathcal{W}\left(\mathbf{\Gamma}_k^{-1} \mid \widehat{\mathbf{\Psi}}_k^{-1}, \widehat{\nu}_k\right)\right] \\
&= -\mathrm{H}\left[\mathcal{N}_L\left(\mathbf{c}_k \mid \widehat{\mathbf{m}}_k, \widehat{\lambda}_k^{-1}\mathbf{\Gamma}_k\right)\right] - \mathrm{H}\left[\mathcal{W}\left(\mathbf{\Gamma}_k^{-1} \mid \widehat{\mathbf{\Psi}}_k^{-1}, \widehat{\nu}_k\right)\right] \\
&= \frac{L}{2}\log\frac{\widehat{\lambda}_k}{2\pi} + \frac{1}{2}\mathbb{E}\left[\log|\mathbf{\Gamma}_k|\right] - \frac{L}{2} + \log B\left(\widehat{\mathbf{\Psi}}_k^{-1}, \widehat{\nu}_k\right) - \frac{\widehat{\nu}_k - L - 1}{2}\mathbb{E}\left[\log|\mathbf{\Gamma}_k|\right] - \frac{\widehat{\nu}_k L}{2} \\
&= \frac{L}{2}\log\frac{\widehat{\lambda}_k}{2\pi} - \frac{L}{2} + \log B\left(\widehat{\mathbf{\Psi}}_k^{-1}, \widehat{\nu}_k\right) - \frac{\widehat{\nu}_k - L}{2}\log\widetilde{\mathbf{\Gamma}}_k - \frac{\widehat{\nu}_k L}{2}. \qquad \text{(C36)}
\end{aligned}
$$

C.4. Proof of Theorem C.2

Recall that we defined $\mathbf{\Theta} = (\boldsymbol{\tau}, \alpha, \sigma, \boldsymbol{\theta}^*)$, $\boldsymbol{\theta}^* = (\boldsymbol{\theta}_k^*)_{k\in\mathbb{N}^\star} := (\mathbf{c}_k, \mathbf{\Gamma}_k)_{k\in\mathbb{N}^\star}$. Then,

$$
\begin{aligned}
p\left(\widehat{\mathbf{y}}, \widehat{\mathbf{x}}, \mathcal{X}, \mathcal{Y}\right) &= \sum_{\widehat{\mathbf{z}}} \int p\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{x}}, \widehat{\mathbf{z}}, \mathbf{\Theta}, \mathcal{X}, \mathcal{Y}\right) p\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{z}}, \mathbf{\Theta}, \mathcal{X}, \mathcal{Y}\right) p\left(\widehat{\mathbf{z}} \mid \mathbf{\Theta}, \mathcal{X}, \mathcal{Y}\right) p\left(\mathbf{\Theta} \mid \mathcal{X}, \mathcal{Y}\right) d\mathbf{\Theta} \\
&= \sum_{\widehat{\mathbf{z}}} \int p\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{x}}, \widehat{\mathbf{z}}; \boldsymbol{A}, \boldsymbol{b}, \mathbf{\Sigma}\right) p\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{z}}, \boldsymbol{c}, \mathbf{\Gamma}\right) p\left(\widehat{\mathbf{z}} \mid \boldsymbol{\tau}; \beta\right) p\left(\mathbf{\Theta} \mid \mathcal{X}, \mathcal{Y}\right) d\mathbf{\Theta} \equiv T_1.
\end{aligned}
\qquad \text{(C37)}
$$

Note that in (C37), $p\left(\mathbf{\Theta} \mid \mathcal{X}, \mathcal{Y}\right)$ is in fact the (unknown) true posterior distribution of the parameters given a sample $(\mathcal{X}, \mathcal{Y})$. Because the integrations w.r.t. true posterior distribution are intractable, we approximate the predictive conditional density by replacing the true posterior distribution $p\left(\mathbf{\Theta} \mid \mathcal{X}, \mathcal{Y}\right)$ with its truncated variational posterior of parameters $\mathbf{\Theta}$ given by

$$
q_\Theta(\mathbf{\Theta} \mid \mathcal{X}, \mathcal{Y}) = q_{\alpha,\sigma}(\alpha, \sigma \mid \mathcal{X}, \mathcal{Y}) \prod_{k=1}^{K-1} q_{\tau_k}\left(\tau_k \mid \mathcal{X}, \mathcal{Y}\right) \prod_{k=1}^{K} q_{\boldsymbol{\theta}_k^*}\left(\boldsymbol{\theta}_k^* \mid \mathcal{X}, \mathcal{Y}\right). \qquad \text{(C38)}
$$

Recall that the infinite state space for each z_j is dealt with by choosing a truncation of the state space to a maximum label K [10]. In practice, this consists of assuming that the variational distributions q_{z_n} for $n \in [N]$, satisfy $q_{z_n}(k) = 0$ for $k > K$ and that the variational distribution on τ also factorizes as $q_{\boldsymbol{\tau}}(\boldsymbol{\tau}) = \prod_{k=1}^{K-1} q_{\tau_k}\left(\tau_k\right)$ with the additional condition that $\tau_K = 1$. In particular, here we choose $K = \widehat{K}$ where $\widehat{K}$ is estimated from some suitable procedures.

For simplicity, we consider the case when $\beta = 0, \sigma = 0$. Then, we obtain

$$
\begin{aligned}
T_1 &\approx \sum_{\widehat{\mathbf{z}}} \int p\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{x}}, \widehat{\mathbf{z}}; \boldsymbol{A}, \boldsymbol{b}, \boldsymbol{\Sigma}\right) p\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{z}}, \boldsymbol{c}, \boldsymbol{\Gamma}\right) p\left(\widehat{\mathbf{z}} \mid \boldsymbol{\tau}\right) q_{\Theta}(\boldsymbol{\Theta} \mid \mathcal{X}, \mathcal{Y}) d\boldsymbol{\Theta} \\
&= \sum_{k=1}^{\infty} \int \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{A}}_k \widehat{\mathbf{x}} + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right) \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \mathbf{c}_k, \boldsymbol{\Gamma}_k\right) \pi_k(\boldsymbol{\tau}) q_{\Theta}(\boldsymbol{\Theta} \mid \mathcal{X}, \mathcal{Y}) d\boldsymbol{\Theta} \\
&\approx \sum_{k=1}^{K} \int \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{A}}_k \widehat{\mathbf{x}} + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right) \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \mathbf{c}_k, \boldsymbol{\Gamma}_k\right) \int \pi_k(\boldsymbol{\tau}) q_{\boldsymbol{\tau}}\left(\boldsymbol{\tau} \mid \mathcal{X}, \mathcal{Y}\right) d\boldsymbol{\tau} \\
&\qquad \times \underbrace{\int q_{\alpha,0}(\alpha \mid \mathcal{X}, \mathcal{Y}) d\alpha}_{=1} \prod_{k=1}^{K} q_{\boldsymbol{\theta}_k^*}\left(\boldsymbol{c}_k, \boldsymbol{\Gamma}_k \mid \mathcal{X}, \mathcal{Y}\right) d\boldsymbol{c} d\boldsymbol{\Gamma} \\
&= \sum_{k=1}^{K} \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k\left(\boldsymbol{\tau}\right)\right] \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{A}}_k \widehat{\mathbf{x}} + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right) \int \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \mathbf{c}_k, \boldsymbol{\Gamma}_k\right) q_{\boldsymbol{\theta}_k^*}\left(\boldsymbol{c}_k, \boldsymbol{\Gamma}_k \mid \mathcal{X}, \mathcal{Y}\right) d\boldsymbol{c}_k d\boldsymbol{\Gamma}_k \\
&\qquad \times \underbrace{\int \prod_{j=1, j \neq k}^{K} q_{\boldsymbol{\theta}_j^*}\left(\boldsymbol{c}_j, \boldsymbol{\Gamma}_j \mid \mathcal{X}, \mathcal{Y}\right) \prod_{j=1, j \neq k}^{K} d\boldsymbol{c}_j d\boldsymbol{\Gamma}_j}_{=1} \\
&= \sum_{k=1}^{K} \mathbb{E}_{q_{\boldsymbol{\tau}_k}}\left[\pi_k\left(\boldsymbol{\tau}\right)\right] \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{A}}_k \widehat{\mathbf{x}} + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right) \underbrace{\int \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \mathbf{c}_k, \boldsymbol{\Gamma}_k\right) q_{\boldsymbol{\theta}_k^*}\left(\boldsymbol{c}_k, \boldsymbol{\Gamma}_k \mid \mathcal{X}, \mathcal{Y}\right) d\boldsymbol{c}_k d\boldsymbol{\Gamma}_k}_{=\mathrm{St}(\widehat{\mathbf{x}} \mid \widehat{\mathbf{m}}_k, \mathbf{L}_k, \widehat{\nu}_k + 1 - L)(\text{Lemma C.6})} \\
&= \sum_{k=1}^{K} \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k\left(\boldsymbol{\tau}\right)\right] \mathrm{St}\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{m}}_k, \mathbf{L}_k, \widehat{\nu}_k + 1 - L\right) \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{A}}_k \widehat{\mathbf{x}} + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right). \qquad \text{(C39)}
\end{aligned}
$$

Here we used the following Lemma C.6

Lemma C.6. *For each $k \in [K]$, it holds that*

$$
\int \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \mathbf{c}_k, \boldsymbol{\Gamma}_k\right) q_{\boldsymbol{\theta}_k^*}\left(\boldsymbol{c}_k, \boldsymbol{\Gamma}_k \mid \mathcal{X}, \mathcal{Y}\right) d\boldsymbol{c}_k d\boldsymbol{\Gamma}_k = St\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{m}}_k, \mathbf{L}_k, \widehat{\nu}_k + 1 - L\right). \qquad \text{(C40)}
$$

Proof of Theorem C.6

By definition, we obtain

$$
\begin{aligned}
&\int\int \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \boldsymbol{c}_k, \boldsymbol{\Gamma}_k^{-1}\right) q(\boldsymbol{\pi} \mid \mathcal{X}) q\left(\mathbf{c}_k, \boldsymbol{\Gamma}_k \mid \mathcal{X}\right) d\boldsymbol{c}_k d\boldsymbol{\Gamma}_k \\
&= \int\int \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \boldsymbol{c}_k, \boldsymbol{\Gamma}_k^{-1}\right) \mathcal{N}_L\left(\mathbf{c}_k \mid \widehat{\mathbf{m}}_k, \left(\widehat{\lambda}_k \boldsymbol{\Gamma}_k\right)^{-1}, \mathcal{X}\right) \mathcal{W}\left(\boldsymbol{\Gamma}_k \mid \widehat{\boldsymbol{\Psi}}_k, \widehat{\nu}_k, \mathcal{X}\right) d\boldsymbol{c}_k d\boldsymbol{\Gamma}_k \\
&= \int \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \boldsymbol{c}_k, \boldsymbol{\Gamma}_k^{-1}\right) \mathcal{N}_L\left(\mathbf{c}_k \mid \widehat{\mathbf{m}}_k, \left(\widehat{\lambda}_k \boldsymbol{\Gamma}_k\right)^{-1}, \mathcal{X}\right) \mathcal{W}\left(\boldsymbol{\Gamma}_k \mid \widehat{\boldsymbol{\Psi}}_k, \widehat{\nu}_k, \mathcal{X}\right) d\boldsymbol{c}_k d\boldsymbol{\Gamma}_k \\
&= \int \left[\int \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \boldsymbol{c}_k, \boldsymbol{\Gamma}_k^{-1}\right) \mathcal{N}_L\left(\mathbf{c}_k \mid \widehat{\mathbf{m}}_k, \left(\widehat{\lambda}_k \boldsymbol{\Gamma}_k\right)^{-1}, \mathcal{X}\right) d\boldsymbol{c}_k\right] \mathcal{W}\left(\boldsymbol{\Gamma}_k \mid \widehat{\boldsymbol{\Psi}}_k, \widehat{\nu}_k, \mathcal{X}\right) d\boldsymbol{\Gamma}_k \\
&= \int \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{m}}_k, \left(1+\widehat{\lambda}_k^{-1}\right) \boldsymbol{\Gamma}_k^{-1}, \mathcal{X}\right) \mathcal{W}\left(\boldsymbol{\Gamma}_k \mid \widehat{\boldsymbol{\Psi}}_k, \widehat{\nu}_k, \mathcal{X}\right) d\boldsymbol{\Gamma}_k. \qquad \text{(C41)}
\end{aligned}
$$

When the size of the data set is large, *i.e.,* $N \to \infty$, this predictive distribution (C41) becomes a mixture of Gaussians with component means $\widehat{\mathbf{m}}_k$ and precisions $\mathbf{L}_k$. In particular, we made use of the following results for marginal and conditional Gaussians, see, *e.g.,* [9, Eq. (2.115), page 93]. Given a marginal Gaussian distribution for $\mathbf{x}$ and a conditional Gaussian distribution for $\mathbf{y}$ given $\mathbf{x}$ in the form

$$
p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Gamma}^{-1}\right), \qquad \text{(C42)}
$$

$$
p(\mathbf{y} \mid \mathbf{x}) = \mathcal{N}\left(\mathbf{y} \mid \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}\right), \qquad \text{(C43)}
$$

then the marginal distribution of $\mathbf{y}$ and the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ are given by

$$
p(\mathbf{y}) = \mathcal{N}\left(\mathbf{y} \mid \mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Gamma}^{-1}\mathbf{A}^\top\right), \qquad \text{(C44)}
$$

$$
p(\mathbf{x} \mid \mathbf{y}) = \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\Sigma}\left\{\mathbf{A}^\top \mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Gamma}\boldsymbol{\mu}\right\}, \boldsymbol{\Sigma}\right), \qquad \text{(C45)}
$$

where

$$
\boldsymbol{\Sigma} = \left(\boldsymbol{\Gamma} + \mathbf{A}^\top \mathbf{L}\mathbf{A}\right)^{-1}. \qquad \text{(C46)}
$$

In our situation, via using $\mathbf{y} \equiv \widehat{\mathbf{x}}, \mathbf{x} \equiv \mathbf{c}_k, \mathbf{A} \equiv \mathbf{I}, \mathbf{b} \equiv \mathbf{0}, \mathbf{L}^{-1} = \boldsymbol{\Gamma}_k^{-1}, \boldsymbol{\mu} \equiv \widehat{\mathbf{m}}_k, \boldsymbol{\Gamma}^{-1} \equiv \left(\widehat{\lambda}_k \boldsymbol{\Gamma}_k\right)^{-1}$, we obtain

$$
\begin{aligned}
p\left(\widehat{\mathbf{x}} | \boldsymbol{\Gamma}_k^{-1}, \mathcal{X}\right) &= \int \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \boldsymbol{c}_k, \boldsymbol{\Gamma}_k^{-1}\right) \mathcal{N}_L\left(\mathbf{c}_k \mid \widehat{\mathbf{m}}_k, \left(\widehat{\lambda}_k \boldsymbol{\Gamma}_k\right)^{-1}, \mathcal{X}\right) d\boldsymbol{c}_k \\
&= \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{m}}_k, \boldsymbol{\Gamma}_k^{-1} + \left(\widehat{\lambda}_k \boldsymbol{\Gamma}_k\right)^{-1}, \mathcal{X}\right) \\
&= \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{m}}_k, \left(\frac{1+\widehat{\lambda}_k}{\widehat{\lambda}_k}\right) \boldsymbol{\Gamma}_k^{-1}, \mathcal{X}\right). \qquad \text{(C47)}
\end{aligned}
$$

Notice that the Wishart distribution is a conjugate prior for the Gaussian distribution with known mean and unknown precision. Therefore, it holds that the product of

$$\mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{m}}_k, \left(1+\widehat{\lambda}_k^{-1}\right)\boldsymbol{\Gamma}_k^{-1}, \mathcal{X}\right)\mathcal{W}\left(\boldsymbol{\Gamma}_k \mid \widehat{\boldsymbol{\Psi}}_k, \widehat{\nu}_k, \mathcal{X}\right)$$

is again a Wishart distribution without normalized. This can be verified by focusing on the dependency on $\boldsymbol{\Gamma}_k$. More precisely, by using the trace trick of quadratic form, $(\widehat{\mathbf{x}}-\widehat{\mathbf{m}}_k)^\top \boldsymbol{\Gamma}_k (\widehat{\mathbf{x}}-\widehat{\mathbf{m}}_k) = \mathrm{Tr}\left((\widehat{\mathbf{x}}-\widehat{\mathbf{m}}_k)(\widehat{\mathbf{x}}-\widehat{\mathbf{m}}_k)^\top \boldsymbol{\Gamma}_k\right)$, we obtain

$$\begin{aligned}
&\mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{m}}_k, \left(1+\widehat{\lambda}_k^{-1}\right)\boldsymbol{\Gamma}_k^{-1}, \mathcal{X}\right)\mathcal{W}\left(\boldsymbol{\Gamma}_k \mid \widehat{\boldsymbol{\Psi}}_k, \widehat{\nu}_k, \mathcal{X}\right)\\
&= \underbrace{\frac{B(\widehat{\boldsymbol{\Psi}}_k, \widehat{\nu}_k)}{\left(2\pi\left(1+\widehat{\lambda}_k^{-1}\right)\right)^{L/2}}}_{:=C\left(\widehat{\boldsymbol{\Psi}}_k, \widehat{\nu}_k, \widehat{\lambda}_k\right)} |\boldsymbol{\Gamma}_k|^{1/2+(\widehat{\nu}_k-L-1)/2}\\
&\quad\times \exp\left\{-\frac{1}{2\left(1+\widehat{\lambda}_k^{-1}\right)}(\widehat{\mathbf{x}}-\widehat{\mathbf{m}}_k)^\top \boldsymbol{\Gamma}_k (\widehat{\mathbf{x}}-\widehat{\mathbf{m}}_k) - \frac{1}{2}\mathrm{Tr}\left(\widehat{\boldsymbol{\Psi}}_k^{-1}\boldsymbol{\Gamma}_k\right)\right\}\\
&= C\left(\widehat{\boldsymbol{\Psi}}_k, \widehat{\nu}_k, \widehat{\lambda}_k\right)|\boldsymbol{\Gamma}_k|^{(\widehat{\nu}_k+1-L-1)/2}\exp\left\{-\frac{1}{2}\mathrm{Tr}\left(\left(1+\widehat{\lambda}_k^{-1}\right)^{-1}(\widehat{\mathbf{x}}-\widehat{\mathbf{m}}_k)(\widehat{\mathbf{x}}-\widehat{\mathbf{m}}_k)^\top \boldsymbol{\Gamma}_k + \widehat{\boldsymbol{\Psi}}_k^{-1}\boldsymbol{\Gamma}_k\right)\right\}\\
&= C\left(\widehat{\boldsymbol{\Psi}}_k, \widehat{\nu}_k, \widehat{\lambda}_k\right)|\boldsymbol{\Gamma}_k|^{(\widehat{\nu}_k+1-L-1)/2}\exp\left\{-\frac{1}{2}\mathrm{Tr}\left\{\left[\left(1+\widehat{\lambda}_k^{-1}\right)^{-1}(\widehat{\mathbf{x}}-\widehat{\mathbf{m}}_k)(\widehat{\mathbf{x}}-\widehat{\mathbf{m}}_k)^\top + \widehat{\boldsymbol{\Psi}}_k^{-1}\right]\boldsymbol{\Gamma}_k\right\}\right\}\\
&= \frac{C\left(\widehat{\boldsymbol{\Psi}}_k, \widehat{\nu}_k, \widehat{\lambda}_k\right)}{B\left(\widehat{\boldsymbol{\Psi}}_k^*, \widehat{\nu}_k^*\right)}\mathcal{W}\left(\boldsymbol{\Gamma}_k \mid \boldsymbol{\Psi}_k^*, \widehat{\nu}_k^*\right). \qquad \text{(C48)}
\end{aligned}$$

Here, $\widehat{\nu}_k^* = \widehat{\nu}_k + 1$, and

$$\begin{aligned}
\boldsymbol{\Psi}_k^* &= \left[\left(1+\widehat{\lambda}_k^{-1}\right)^{-1}(\widehat{\mathbf{x}}-\widehat{\mathbf{m}}_k)(\widehat{\mathbf{x}}-\widehat{\mathbf{m}}_k)^\top + \widehat{\boldsymbol{\Psi}}_k^{-1}\right]^{-1},\\
\left|\widehat{\boldsymbol{\Psi}}_k^*\right|^{(\widehat{\nu}_k+1)/2} &= \left|\left(1+\widehat{\lambda}_k^{-1}\right)^{-1}(\widehat{\mathbf{x}}-\widehat{\mathbf{m}}_k)(\widehat{\mathbf{x}}-\widehat{\mathbf{m}}_k)^\top + \widehat{\boldsymbol{\Psi}}_k^{-1}\right|^{-(\widehat{\nu}_k+1)/2}\\
&= \left|\widehat{\boldsymbol{\Psi}}_k^{-1}\left[\left(1+\widehat{\lambda}_k^{-1}\right)^{-1}\widehat{\boldsymbol{\Psi}}_k(\widehat{\mathbf{x}}-\widehat{\mathbf{m}}_k)(\widehat{\mathbf{x}}-\widehat{\mathbf{m}}_k)^\top + \mathbf{I}\right]\right|^{-(\widehat{\nu}_k+1)/2}\\
&= \left|\widehat{\boldsymbol{\Psi}}_k\right|^{(\widehat{\nu}_k+1)/2}\left|\left(1+\widehat{\lambda}_k^{-1}\right)^{-1}\widehat{\boldsymbol{\Psi}}_k(\widehat{\mathbf{x}}-\widehat{\mathbf{m}}_k)(\widehat{\mathbf{x}}-\widehat{\mathbf{m}}_k)^\top + \mathbf{I}\right|^{-(\widehat{\nu}_k+1)/2}\\
&= \left|\widehat{\boldsymbol{\Psi}}_k\right|^{(\widehat{\nu}_k+1)/2}\left[1+\left(1+\widehat{\lambda}_k^{-1}\right)^{-1}(\widehat{\mathbf{x}}-\widehat{\mathbf{m}}_k)^\top\widehat{\boldsymbol{\Psi}}_k(\widehat{\mathbf{x}}-\widehat{\mathbf{m}}_k)\right]^{-(\widehat{\nu}_k+1)/2}. \qquad \text{(C49)}
\end{aligned}$$

Via using the normalization constant we have

$$
\begin{aligned}
&\int \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{m}}_k, \left(1+\widehat{\lambda}_k^{-1}\right)\mathbf{\Gamma}_k^{-1}, \mathcal{X}\right)\mathcal{W}\left(\mathbf{\Gamma}_k \mid \widehat{\mathbf{\Psi}}_k, \widehat{\nu}_k, \mathcal{X}\right) d\mathbf{\Gamma}_k \\
&= \frac{C\left(\widehat{\mathbf{\Psi}}_k, \widehat{\nu}_k, \widehat{\lambda}_k\right)}{B\left(\widehat{\mathbf{\Psi}}_k^*, \widehat{\nu}_k^*\right)} \underbrace{\int \mathcal{W}\left(\mathbf{\Gamma}_k \mid \mathbf{\Psi}_k^*, \widehat{\nu}_k^*\right) d\mathbf{\Gamma}_k}_{=1} = \frac{C\left(\widehat{\mathbf{\Psi}}_k, \widehat{\nu}_k, \widehat{\lambda}_k\right)}{B\left(\widehat{\mathbf{\Psi}}_k^*, \widehat{\nu}_k^*\right)} = \frac{B(\widehat{\mathbf{\Psi}}_k, \widehat{\nu}_k)}{\left(2\pi\left(1+\widehat{\lambda}_k^{-1}\right)\right)^{L/2}} \frac{1}{B\left(\widehat{\mathbf{\Psi}}_k^*, \widehat{\nu}_k^*\right)} \\
&= \frac{1}{\left(2\pi\left(1+\widehat{\lambda}_k^{-1}\right)\right)^{L/2}} \frac{\left|\widehat{\mathbf{\Psi}}_k\right|^{-\widehat{\nu}_k/2}\left(2^{\widehat{\nu}_k L/2}\pi^{L(L-1)/4}\prod_{l=1}^{L}\Gamma\left(\frac{\widehat{\nu}_k+1-l}{2}\right)\right)^{-1}}{\left|\widehat{\mathbf{\Psi}}_k^*\right|^{-(\widehat{\nu}_k+1)/2}\left(2^{(\widehat{\nu}_k+1)L/2}\pi^{L(L-1)/4}\prod_{l=1}^{L}\Gamma\left(\frac{(\widehat{\nu}_k+1)+1-l}{2}\right)\right)^{-1}} \\
&= \frac{1}{\left(\pi\left(1+\widehat{\lambda}_k^{-1}\right)\right)^{L/2}} \frac{\left|\widehat{\mathbf{\Psi}}_k\right|^{-\widehat{\nu}_k/2}\Gamma\left(\frac{\widehat{\nu}_k+1}{2}\right)\Gamma\left(\frac{\widehat{\nu}_k}{2}\right)\ldots\Gamma\left(\frac{\widehat{\nu}_k+2-L}{2}\right)}{\left|\widehat{\mathbf{\Psi}}_k^*\right|^{-(\widehat{\nu}_k+1)/2}\Gamma\left(\frac{\widehat{\nu}_k}{2}\right)\Gamma\left(\frac{\widehat{\nu}_k-1}{2}\right)\ldots\Gamma\left(\frac{\widehat{\nu}_k+2-L}{2}\right)\Gamma\left(\frac{\widehat{\nu}_k+1-L}{2}\right)} \\
&= \frac{\Gamma\left(\frac{\widehat{\nu}_k+1}{2}\right)}{\Gamma\left(\frac{\widehat{\nu}_k+1-L}{2}\right)\pi^{L/2}} \frac{\left|\widehat{\mathbf{\Psi}}_k\right|^{-\widehat{\nu}_k/2}}{\left(1+\widehat{\lambda}_k^{-1}\right)^{L/2}} \left|\widehat{\mathbf{\Psi}}_k\right|^{(\widehat{\nu}_k+1)/2}\left[1+\left(1+\widehat{\lambda}_k^{-1}\right)^{-1}\left(\widehat{\mathbf{x}}-\widehat{\mathbf{m}}_k\right)^\top\widehat{\mathbf{\Psi}}_k\left(\widehat{\mathbf{x}}-\widehat{\mathbf{m}}_k\right)\right]^{-(\widehat{\nu}_k+1)/2} \\
&= \frac{\Gamma\left(\frac{\widehat{\nu}_k+1}{2}\right)}{\Gamma\left(\frac{\widehat{\nu}_k+1-L}{2}\right)\pi^{L/2}} \frac{\left|\widehat{\mathbf{\Psi}}_k\right|^{1/2}}{\left(1+\widehat{\lambda}_k^{-1}\right)^{L/2}} \left[1+\left(1+\widehat{\lambda}_k^{-1}\right)^{-1}\left(\widehat{\mathbf{x}}-\widehat{\mathbf{m}}_k\right)^\top\widehat{\mathbf{\Psi}}_k\left(\widehat{\mathbf{x}}-\widehat{\mathbf{m}}_k\right)\right]^{-(\widehat{\nu}_k+1)/2} \\
&= \mathrm{St}\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{m}}_k, \mathbf{L}_k, \widehat{\nu}_k+1-L\right). \qquad \text{(C50)}
\end{aligned}
$$

Here,

$$
\mathbf{L}_k = \frac{\left(\widehat{\nu}_k+1-L\right)\widehat{\lambda}_k}{1+\widehat{\lambda}_k}\widehat{\mathbf{\Psi}}_k, \qquad \text{(C51)}
$$

and Δ^2 is the squared Mahalanobis distance defined by

$$
\Delta^2 = \left(\widehat{\mathbf{x}}-\widehat{\mathbf{m}}_k\right)^\top \mathbf{L}_k\left(\widehat{\mathbf{x}}-\widehat{\mathbf{m}}_k\right). \qquad \text{(C52)}
$$

Then, the last equality holds since we have

$$\begin{aligned}
\mathrm{St}\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{m}}_k, \mathbf{L}_k, \widehat{\nu}_k+1-L\right) &= \frac{\Gamma\left(\frac{\widehat{\nu}_k+1-L}{2}+\frac{L}{2}\right)}{\Gamma\left(\frac{\widehat{\nu}_k+1-L}{2}\right)\pi^{L/2}} \frac{\left|\mathbf{L}_k\right|^{1/2}}{\left(\widehat{\nu}_k+1-L\right)^{L/2}}\left[1+\frac{\Delta^2}{\widehat{\nu}_k+1-L}\right]^{-(\widehat{\nu}_k+1-L)/2-L/2} \\
&= \frac{\Gamma\left(\frac{\widehat{\nu}_k+1}{2}\right)}{\Gamma\left(\frac{\widehat{\nu}_k+1-L}{2}\right)\pi^{L/2}} \frac{\left(\widehat{\nu}_k+1-L\right)^{L/2}\left|\widehat{\mathbf{\Psi}}_k\right|^{1/2}}{\left(\widehat{\nu}_k+1-L\right)^{L/2}\left(1+\widehat{\lambda}_k^{-1}\right)^{L/2}}\left[1+\frac{\Delta^2}{\widehat{\nu}_k+1-L}\right]^{-(\widehat{\nu}_k+1-L)/2-L/2} \\
&= \frac{\Gamma\left(\frac{\widehat{\nu}_k+1}{2}\right)}{\Gamma\left(\frac{\widehat{\nu}_k+1-L}{2}\right)\pi^{L/2}} \frac{\left|\widehat{\mathbf{\Psi}}_k\right|^{1/2}}{\left(1+\widehat{\lambda}_k^{-1}\right)^{L/2}}\left[1+\frac{\left(\widehat{\nu}_k+1-L\right)\widehat{\lambda}_k}{\left(1+\widehat{\lambda}_k\right)}\frac{\left(\widehat{\mathbf{x}}-\widehat{\mathbf{m}}_k\right)^\top\widehat{\mathbf{\Psi}}_k\left(\widehat{\mathbf{x}}-\widehat{\mathbf{m}}_k\right)}{\widehat{\nu}_k+1-L}\right]^{-(\widehat{\nu}_k+1)/2} \\
&= \frac{\Gamma\left(\frac{\widehat{\nu}_k+1}{2}\right)}{\Gamma\left(\frac{\widehat{\nu}_k+1-L}{2}\right)\pi^{L/2}} \frac{\left|\widehat{\mathbf{\Psi}}_k\right|^{1/2}}{\left(1+\widehat{\lambda}_k^{-1}\right)^{L/2}}\left[1+\frac{\left(\widehat{\mathbf{x}}-\widehat{\mathbf{m}}_k\right)^\top\widehat{\mathbf{\Psi}}_k\left(\widehat{\mathbf{x}}-\widehat{\mathbf{m}}_k\right)}{\left(1+\widehat{\lambda}_k^{-1}\right)}\right]^{-(\widehat{\nu}_k+1)/2}. \qquad \text{(C53)}
\end{aligned}$$

C.5. Proof of Theorem C.3

From the product rule of probability, we see that this conditional distribution can be evaluated from the joint and marginal distributions. Furthermore, by integrating out $\widehat{\mathbf{z}}$ and $\mathbf{\Theta}$, the predictive conditional density is then given by

$$p\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{x}}, \mathcal{X}, \mathcal{Y}\right) = \frac{p\left(\widehat{\mathbf{y}}, \widehat{\mathbf{x}} \mid \mathcal{X}, \mathcal{Y}\right)}{p\left(\widehat{\mathbf{x}} \mid \mathcal{X}, \mathcal{Y}\right)} = \frac{\sum_{\widehat{\mathbf{z}}}\int p\left(\widehat{\mathbf{y}}, \widehat{\mathbf{x}}, \widehat{\mathbf{z}}, \mathbf{\Theta} \mid \mathcal{X}, \mathcal{Y}\right) d\widehat{\mathbf{z}} d\mathbf{\Theta}}{\sum_{\widehat{\mathbf{z}}}\int p\left(\widehat{\mathbf{x}}, \widehat{\mathbf{z}}, \mathbf{\Theta} \mid \mathcal{X}, \mathcal{Y}\right) d\widehat{\mathbf{z}} d\mathbf{\Theta}} \equiv \frac{T_1}{T_2}. \qquad \text{(C54)}$$

Next, with a similar step as in the proof of Theorem C.2, we also obtain

$$T_2 = \sum_{k=1}^{K} \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k\left(\boldsymbol{\tau}\right)\right] \mathrm{St}\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{m}}_k, \mathbf{L}_k, \widehat{\nu}_k+1-L\right). \qquad \text{(C55)}$$

Therefore, we obtain

$$\begin{aligned}
&p\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{x}}, \mathcal{X}, \mathcal{Y}\right) \\
&\approx \frac{\sum_{k=1}^{K} \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k\left(\boldsymbol{\tau}\right)\right] \mathrm{St}\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{m}}_k, \mathbf{L}_k, \widehat{\nu}_k+1-L\right) \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{A}}_k\widehat{\mathbf{x}}+\widehat{\mathbf{b}}_k, \widehat{\mathbf{\Sigma}}_k\right)}{\sum_{k=1}^{K} \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k\left(\boldsymbol{\tau}\right)\right] \mathrm{St}\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{m}}_k, \mathbf{L}_k, \widehat{\nu}_k+1-L\right)} \\
&= \sum_{k=1}^{K} \frac{\mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k\left(\boldsymbol{\tau}\right)\right] \mathrm{St}\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{m}}_k, \mathbf{L}_k, \widehat{\nu}_k+1-L\right)}{\sum_{k=1}^{K} \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k\left(\boldsymbol{\tau}\right)\right] \mathrm{St}\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{m}}_k, \mathbf{L}_k, \widehat{\nu}_k+1-L\right)} \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{A}}_k\widehat{\mathbf{x}}+\widehat{\mathbf{b}}_k, \widehat{\mathbf{\Sigma}}_k\right) \\
&\equiv \sum_{k=1}^{K} g_k\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{\Theta}}, \widehat{\boldsymbol{\phi}}, \mathcal{X}, \mathcal{Y}\right) \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{A}}_k\widehat{\mathbf{x}}+\widehat{\mathbf{b}}_k, \widehat{\mathbf{\Sigma}}_k\right), \qquad \text{(C56)}
\end{aligned}$$

which is a mixture of Gaussian experts since we have

$$g_k\left(\widehat{\mathbf{x}} \mid \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\phi}}, \mathcal{X}, \mathcal{Y}\right) = \frac{\mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k\left(\boldsymbol{\tau}\right)\right] \mathrm{St}\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{m}}_k, \mathbf{L}_k, \widehat{\nu}_k + 1 - L\right)}{\sum_{k=1}^{K} \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k\left(\boldsymbol{\tau}\right)\right] \mathrm{St}\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{m}}_k, \mathbf{L}_k, \widehat{\nu}_k + 1 - L\right)}, \quad k \in [K], \tag{C57}$$

belongs to a $K-1$ dimensional probability simplex.

C.6. Proof of Theorem C.4

To deal with high-dimensional regression data, namely high-to-low regression, given the inverse conditional density $p(\widehat{\mathbf{y}} \mid \widehat{\mathbf{x}}, \mathcal{X}, \mathcal{Y})$, we want to compute the following forward conditional density

$$p(\widehat{\mathbf{x}} \mid \widehat{\mathbf{y}}, \mathcal{X}, \mathcal{Y}) = \frac{p(\widehat{\mathbf{x}}, \widehat{\mathbf{y}} \mid \mathcal{X}, \mathcal{Y})}{p(\widehat{\mathbf{y}} \mid \mathcal{X}, \mathcal{Y})} = \frac{p(\widehat{\mathbf{x}}, \widehat{\mathbf{y}} \mid \mathcal{X}, \mathcal{Y})}{\int_{\widehat{\mathbf{x}}} p(\widehat{\mathbf{x}}, \widehat{\mathbf{y}} \mid \mathcal{X}, \mathcal{Y}) d\widehat{\mathbf{x}}} = \frac{T_1}{\int_{\widehat{\mathbf{x}}} T_1(\widehat{\mathbf{x}}) d\widehat{\mathbf{x}}} \equiv \frac{T_1}{T_3}. \tag{C58}$$

Then, we have to compute or numerically approximate D_3. Using Theorem C.2 and definition of Student's t-distribution, we obtain

$$T_3 = \sum_{k=1}^{K} \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k\left(\boldsymbol{\tau}\right)\right] D_k. \tag{C59}$$

Then, by definition of Student's t-distribution, it holds that

$$\begin{aligned}
D_k &= \int \mathrm{St}\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{m}}_k, \mathbf{L}_k, \widehat{\nu}_k + 1 - L\right) \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{A}}_k \widehat{\mathbf{x}} + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right) d\widehat{\mathbf{x}} \\
&= \int \int_0^{\infty} \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{m}}_k, (\eta \mathbf{L}_k)^{-1}\right) \mathrm{Gam}\left(\eta \mid \frac{\widehat{\nu}_k + 1 - L}{2}, \frac{\widehat{\nu}_k + 1 - L}{2}\right) \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{A}}_k \widehat{\mathbf{x}} + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right) d\eta d\widehat{\mathbf{x}} \\
&= \int_0^{\infty} \int \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{A}}_k \widehat{\mathbf{x}} + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right) \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{m}}_k, (\eta \mathbf{L}_k)^{-1}\right) d\widehat{\mathbf{x}} \mathrm{Gam}\left(\eta \mid \frac{\widehat{\nu}_k + 1 - L}{2}, \frac{\widehat{\nu}_k + 1 - L}{2}\right) d\eta \\
&= \int_0^{\infty} \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{A}}_k \widehat{\mathbf{m}}_k + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k + \eta^{-1} \widehat{\mathbf{A}}_k \mathbf{L}_k^{-1} \widehat{\mathbf{A}}_k^{\top}\right) \mathrm{Gam}\left(\eta \mid \frac{\widehat{\nu}_k + 1 - L}{2}, \frac{\widehat{\nu}_k + 1 - L}{2}\right) d\eta.
\end{aligned}$$

Furthermore, we used the fact that

$$\int \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{A}}_k \widehat{\mathbf{x}} + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right) \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{m}}_k, (\eta \mathbf{L}_k)^{-1}\right) d\widehat{\mathbf{x}} = \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{A}}_k \widehat{\mathbf{m}}_k + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k + \widehat{\mathbf{A}}_k \left(\eta \mathbf{L}_k\right)^{-1} \widehat{\mathbf{A}}_k^{\top}\right).$$

Indeed, we made use of the following results for marginal and conditional Gaussians, see, *e.g.,* [9, Eq. (2.115), page 93]. Given a marginal Gaussian distribution for $\mathbf{x}$ and a conditional Gaussian distribution for $\mathbf{y}$ given $\mathbf{x}$ in the form

$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Gamma}^{-1}\right), \tag{C60}$$

$$p(\mathbf{y} \mid \mathbf{x}) = \mathcal{N}\left(\mathbf{y} \mid \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}\right), \tag{C61}$$

then the marginal distribution of $\mathbf{y}$ and the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ are

given by

$$p(\mathbf{y}) = \int p(\mathbf{y} \mid \mathbf{x}) p(\mathbf{x}) d\mathbf{x} = \mathcal{N}\left(\mathbf{y} \mid \mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Gamma}^{-1}\mathbf{A}^\top\right), \tag{C62}$$

$$p(\mathbf{x} \mid \mathbf{y}) = \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\Sigma}\left\{\mathbf{A}^\top \mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Gamma}\boldsymbol{\mu}\right\}, \boldsymbol{\Sigma}\right), \boldsymbol{\Sigma} = \left(\boldsymbol{\Gamma} + \mathbf{A}^\top \mathbf{L}\mathbf{A}\right)^{-1}. \tag{C63}$$

In our situation, the desired result is obtained via using $\mathbf{y} \equiv \widehat{\mathbf{y}}, \mathbf{A} \equiv \widehat{\mathbf{A}}_k, \mathbf{b} \equiv \widehat{\mathbf{b}}_k, \mathbf{L}^{-1} = \widehat{\boldsymbol{\Sigma}}_k$, $\mathbf{x} \equiv \widehat{\mathbf{x}}, \boldsymbol{\mu} \equiv \widehat{\mathbf{m}}_k, \boldsymbol{\Gamma}^{-1} \equiv (\eta \mathbf{L}_k)^{-1}$.

Therefore, we obtain

$$\begin{aligned}
&p(\widehat{\mathbf{x}} \mid \widehat{\mathbf{y}}, \mathcal{X}, \mathcal{Y}) \\
&\approx \sum_{k=1}^{K} \frac{\mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k(\boldsymbol{\tau})\right] \mathrm{St}\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{m}}_k, \mathbf{L}_k, \widehat{\nu}_k + 1 - L\right) \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{A}}_k \widehat{\mathbf{x}} + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right)}{\sum_{k=1}^{K} \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k(\boldsymbol{\tau})\right] \int \mathrm{St}\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{m}}_k, \mathbf{L}_k, \widehat{\nu}_k + 1 - L\right) \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{A}}_k \widehat{\mathbf{x}} + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right) d\widehat{\mathbf{x}}} \\
&= \sum_{k=1}^{K} \frac{\mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k(\boldsymbol{\tau})\right] \int_0^\infty \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{A}}_k \widehat{\mathbf{x}} + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right) \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{m}}_k, (\eta \mathbf{L}_k)^{-1}\right) \mathrm{Gam}\left(\eta \mid \frac{\widehat{\nu}_k + 1 - L}{2}, \frac{\widehat{\nu}_k + 1 - L}{2}\right) d\eta}{\sum_{k=1}^{K} \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k(\boldsymbol{\tau})\right] \int_0^\infty \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{A}}_k \widehat{\mathbf{m}}_k + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k + \eta^{-1} \widehat{\mathbf{A}}_k \mathbf{L}_k^{-1} \widehat{\mathbf{A}}_k^\top\right) \mathrm{Gam}\left(\eta \mid \frac{\widehat{\nu}_k + 1 - L}{2}, \frac{\widehat{\nu}_k + 1 - L}{2}\right) d\eta} \\
&= \sum_{k=1}^{K} \frac{\mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k(\boldsymbol{\tau})\right] \int_0^\infty \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{c}}_k^*, \widehat{\boldsymbol{\Gamma}}_k^*(\eta)\right) \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{A}}_k^*(\eta) \widehat{\mathbf{y}} + \widehat{\mathbf{b}}_k^*(\eta), \widehat{\boldsymbol{\Sigma}}_k^*(\eta)\right) \mathrm{Gam}\left(\eta \mid \frac{\widehat{\nu}_k + 1 - L}{2}, \frac{\widehat{\nu}_k + 1 - L}{2}\right) d\eta}{\sum_{k=1}^{K} \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k(\boldsymbol{\tau})\right] \int_0^\infty \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{c}}_k^*, \widehat{\boldsymbol{\Gamma}}_k^*(\eta)\right) \mathrm{Gam}\left(\eta \mid \frac{\widehat{\nu}_k + 1 - L}{2}, \frac{\widehat{\nu}_k + 1 - L}{2}\right) d\eta} \\
&\approx \sum_{i=1}^{I} \sum_{k=1}^{K} g_{ki}\left(\widehat{\mathbf{y}} \mid \widehat{\boldsymbol{\Theta}}^*, \widehat{\boldsymbol{\phi}}^*, \mathcal{X}, \mathcal{Y}\right) \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{A}}_k^*(\eta_i) \widehat{\mathbf{y}} + \widehat{\mathbf{b}}_k^*(\eta_i), \widehat{\boldsymbol{\Sigma}}_k^*(\eta_i)\right),
\end{aligned}$$

where, for all $k \in [K], i \in [I]$,

$$g_{ki}\left(\widehat{\mathbf{y}} \mid \widehat{\boldsymbol{\Theta}}^*, \widehat{\boldsymbol{\phi}}^*, \mathcal{X}, \mathcal{Y}\right) = \frac{\mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k(\boldsymbol{\tau})\right] \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{c}}_k^*, \widehat{\boldsymbol{\Gamma}}_k^*(\eta_i)\right) \mathrm{Gam}\left(\eta_i \mid \frac{\widehat{\nu}_k + 1 - L}{2}, \frac{\widehat{\nu}_k + 1 - L}{2}\right)}{\sum_{i=1}^{I} \sum_{k=1}^{K} \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k(\boldsymbol{\tau})\right] \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{c}}_k^*, \widehat{\boldsymbol{\Gamma}}_k^*(\eta_i)\right) \mathrm{Gam}\left(\eta_i \mid \frac{\widehat{\nu}_k + 1 - L}{2}, \frac{\widehat{\nu}_k + 1 - L}{2}\right)}.$$

Here, we used the fact that $p(\widehat{\mathbf{y}}, \widehat{\mathbf{x}} \mid \widehat{z} = k) = p(\widehat{\mathbf{x}} \mid \widehat{\mathbf{y}}, \widehat{z} = k) p(\widehat{\mathbf{y}} \mid \widehat{z} = k)$, namely,

$$\begin{aligned}
&\mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{A}}_k \widehat{\mathbf{x}} + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right) \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{m}}_k, (\eta \mathbf{L}_k)^{-1}\right) \\
&= \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\boldsymbol{\Sigma}}_k^*\left[\widehat{\mathbf{A}}_k^\top \widehat{\boldsymbol{\Sigma}}_k^{-1}(\widehat{\mathbf{y}} - \widehat{\mathbf{b}}_k) + \eta \mathbf{L}_k \widehat{\mathbf{m}}_k\right], \widehat{\boldsymbol{\Sigma}}_k^*\right) \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{A}}_k \widehat{\mathbf{m}}_k + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k + \widehat{\mathbf{A}}_k (\eta \mathbf{L}_k)^{-1} \widehat{\mathbf{A}}_k^\top\right) \\
&= \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{A}}_k^*(\eta) \widehat{\mathbf{y}} + \widehat{\mathbf{b}}_k^*(\eta), \widehat{\boldsymbol{\Sigma}}_k^*(\eta)\right) \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{c}}_k^*, \widehat{\boldsymbol{\Gamma}}_k^*(\eta)\right),
\end{aligned}$$

where

$$
\begin{aligned}
\widehat{\boldsymbol{\Sigma}}_k^*(\eta) &= \left(\eta \mathbf{L}_k + \widehat{\mathbf{A}}_k^\top \widehat{\boldsymbol{\Sigma}}_k^{-1} \widehat{\mathbf{A}}_k\right)^{-1}, \\
\widehat{\mathbf{A}}_k^*(\eta) &= \widehat{\boldsymbol{\Sigma}}_k^*(\eta) \widehat{\mathbf{A}}_k^\top \widehat{\boldsymbol{\Sigma}}_k^{-1}, \\
\widehat{\mathbf{b}}_k^*(\eta) &= \widehat{\boldsymbol{\Sigma}}_k^*(\eta) \left[\eta \mathbf{L}_k \widehat{\mathbf{m}}_k - \widehat{\mathbf{A}}_k^\top \widehat{\boldsymbol{\Sigma}}_k^{-1} \widehat{\mathbf{b}}_k\right], \\
\widehat{\mathbf{c}}_k^* &= \widehat{\mathbf{A}}_k \widehat{\mathbf{m}}_k + \widehat{\mathbf{b}}_k, \\
\widehat{\boldsymbol{\Gamma}}_k^*(\eta) &= \widehat{\boldsymbol{\Sigma}}_k + \widehat{\mathbf{A}}_k \left(\eta \mathbf{L}_k\right)^{-1} \widehat{\mathbf{A}}_k^\top.
\end{aligned}
$$

The last approximation is deduced by using the fact that one simplistic strategy for evaluating integration would be to discretize η-space (1-dimensional) into a uniform grid and to evaluate the integrand as a Riemann integral with a truncated value $0 < U_\eta < \infty$ and a number of point $I \in \mathbb{N}^\star$ for approximating the integration.