Bayesian nonparametric mixture of experts for inverse problems

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

Bayesian nonparametric mixture of experts for inverse problems

TrungTin Nguyen[a,d], 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]School of Computing, Engineering and Mathematical Sciences, La Trobe University, Melbourne Australia; [c]Institute of Mathematics for Industry, Kyushu University, Fukuoka Japan; [d]School of Mathematics and Physics, The University of Queensland, St. Lucia, Australia.

ARTICLE HISTORY
Compiled June 17, 2024

Word Limits

11578 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 relatively 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-maximisation. 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 regression function.

KEYWORDS

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

1. Introduction

Inverse problems. Many problems can be formulated as inverse problems, where the goal is to find parameter values that best explain an observed phenomenon. Typical constraints in practice are that the relationships between parameters and observations are highly non-linear, with relatively high dimensional observations and correlated multivariate parameters. To handle these constraints, we consider probabilistic mixtures of locally linear models, namely the Gaussian locally linear mapping (GLLiM) approach of Deleforge et al. (2015), which includes affine instances of the mixture of experts (MoE) model (Xu et al., 1995) and the classical inverse linear regression model (Hoadley, 1970; Li, 1991), as special cases. We propose to address inverse problems in a Bayesian framework, making use of the availability of simulations from forward models of interest. The GLLiM approach has been used in many applications, *e.g.,* in medical imaging (Boux

et al., 2021), planetary science (Kugler et al., 2022; Nguyen et al., 2024), head-pose estimation problem in computer vision (Lathuilière et al., 2017) and quantitative trait prediction from biological data (Blein-Nicolas et al., 2024).

Mixture of experts models (MoE) are generalizations of neural network architectures proposed by Jacobs et al. (1991). Further, these flexible models also generalize the classical mixture models (MMs) and mixture of regression models (McLachlan and Peel, 2000). 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 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 (Norets, 2010; Nguyen et al., 2016, 2019, 2021a) as well as their good convergence rate, see, *e.g.,* for mixture of regression in Ho et al. (2022) and for MoE in Jiang and Tanner (1999); Nguyen et al. (2024b, 2023a, 2024a). It is worth noting that these results improve the approximation capabilities and convergence rates of unconditional MMs, as discussed in Genovese and Wasserman (2000); Rakhlin et al. (2005); Nguyen (2013); Shen et al. (2013); Ho and Nguyen (2016a,b); Nguyen et al. (2020, 2023b). At a high level, universal approximation theorems state that given a large enough number of components, MMs 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.,* Yuksel et al. (2012); Nguyen and Chamroukhi (2018); Do et al. (2023); Nguyen (2021); Chen et al. (2022), for further detailed reviews of practical and theoretical aspects of MoE models in statistics and in diverse domains, *e.g.,* natural language processing and computer vision.

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 MMs and MoE models, which have attracted much attention in statistics and machine learning over the last 50 years, see, *e.g.,* Celeux et al. (2019) 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; Akaike, 1974), the Bayesian information criterion (BIC; Schwarz, 1978, BIC-GLLiM; Forbes et al., 2022a,b) or the BIC-like approximation of integrated classification likelihood (ICL-BIC; Biernacki et al., 2000). However, an important limitation of these criteria is that they are only asymptotically valid (Westerhout et al., 2024). 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 the so-called slope heuristic approach (Birgé and Massart, 2007, also see Arlot, 2019 for a recent review), Nguyen et al. (2021b, 2022c,b, 2023d,c) recently established 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 Nguyen et al. (2022a), 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, Kock et al. (2022) proposed computationally efficient variational inference approaches for architecture selection in high-dimensional deep Gaussian mixture models using overfitted mixtures (see, *e.g.,* Rousseau and Mengersen, 2011; Forbes et al., 2019), where unnecessary components are dropped in the estimation. However, we are interested here in the more general context of the Bayesian nonparametric (BNP) approach, (see, *e.g.,* Hjort et al., 2010; Ghosal and Van der Vaart, 2017), where it is not necessary to know an upper bound of the true number of components as in Bayesian overfitted MM. This is one motivation for the BNP priors that are considered here for GLLiM.

Dirichlet process mixture models (DP-MMs) and Pitman-Yor process mixture models (PYP-MMs) are among the most popular BNP models, particularly suitable for density estimation and probabilistic clustering. However, the posterior of the DP-MMs or PYP-MMs are inconsistent for the number of components if the true number of components is finite and the concentration parameter is fixed (see, *e.g.,* Miller and Harrison, 2014, and Alamichel et al., 2024 for a review). 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 Ascolani et al. (2022) is that consistency for the number of components can be achieved if a prior is placed on the concentration parameter of the DP-MM, under some assumptions on this prior.

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 (Maceachern and Müller, 1998; Green and Richardson, 2001). This encourages the use of a novel, simple post-processing algorithm in the spirit of the Merge-Truncate-Merge (MTM) introduced by Guha et al. (2021). This post-processing procedure 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 address the challenges of highly nonlinear inverse problems with relatively high dimensional observations and correlated parameters, we propose a novel BNP-GLLiM model and inference procedure, which is computationally efficient and avoids any commitment to an arbitrary number of components. Although BNP mixture models (BNP-MM), which are special cases of the BNP-GLLiM model, have been extensively studied in the literature (Escobar and West, 1995; Maceachern and Müller, 1998; Neal, 2000; Arbel et al., 2021; Li et al., 2022; Durand et al., 2022), the extension to MoE models in an inverse regression framework has not been covered. In addition, 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 MMs results of Guha et al. (2021), but also in the more general context

of MoE models for cPDFs.

Notations. Throughout this paper, $\{1, \dots, D\}$ 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. 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. We recall the standard Bayesian nonparametric priors and variational Bayesian expectation-maximisation principle in Appendices A and B, respectively. All details of VBEM for the BNP-GLLiM model, evidence lower-bound, and technical proofs not included in the main paper are relegated to Appendices C to E, respectively. Appendix F presents a more general model with a hyperprior on the gating parameters, called BNP-GLLiM2.

2. BNP-GLLiM model

In Sections 2.1 and 2.2, we present the advantage of adopting a GLLiM model, which uses an inverse regression approach to estimate nonlinear high-to-low dimensional mappings. Such a strategy allows to greatly reduce the number of required parameters. To avoid any commitment to an arbitrary number of components, we then construct the BNP-GLLiM model in Section 2.3.

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.,* Li (1991); Deleforge et al. (2015). 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 $\mathcal{X} \equiv \{\mathbf{x}_n\}_{n\in[N]} \subset (\mathbb{R}^L)^N$, the high-dimensional data sample as $\mathcal{Y} \equiv \{\mathbf{y}_n\}_{n\in[N]} \subset (\mathbb{R}^D)^N$. We denote the observed values as $(\mathbf{x}, \mathbf{y})$, which are independently

and identically distributed (i.i.d.) samples from the random variables $(\mathbf{X}, \mathbf{Y})$.

2.2. *Nonlinear high-to-low dimensional mapping via GLLiM model*

The GLLiM models, as originally introduced in Deleforge et al. (2015), are used to capture the non-linear relationship between the response and the set of covariates from a high-to-low heterogeneous data. Specifically, Deleforge et al. (2015) 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:

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

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 $\mathbf{\Sigma}_k \in \mathbb{R}^{D\times D}$, it follows that

$$p(\mathbf{y} \mid \mathbf{x}, Z = k; \boldsymbol{\psi}) = \mathcal{N}_D(\mathbf{y} \mid \mathbf{A}_k\mathbf{x} + \mathbf{b}_k, \mathbf{\Sigma}_k), \tag{1}$$

where $\boldsymbol{\psi}$ is the vector of model parameters, $\mathcal{N}_D(\mathbf{y}; \mathbf{A}_k\mathbf{x} + \mathbf{b}_k, \mathbf{\Sigma}_k)$ 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:

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

where $\mathbf{c}_k \in \mathbb{R}^L$, $\mathbf{\Gamma}_k \in \mathbb{R}^{L\times L}$, and $\boldsymbol{\pi} = (\pi_k)_{k\in[K]}$ belongs to a probability simplex, defined as $\left\{(\pi_k)_{k\in[K]} \in (\mathbb{R}^+)^K, \sum_{k=1}^K \pi_k = 1\right\}$. Then, via the conditional property of Gaussian variables and hierarchical decomposition, given any $\boldsymbol{\psi} = (\pi_k, \mathbf{c}_k, \mathbf{\Gamma}_k, \mathbf{A}_k, \mathbf{b}_k, \mathbf{\Sigma}_k)_{k\in[K]} \in \mathbf{\Psi}$, we obtain the following inverse conditional density:

$$p(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\psi}) = \sum_{k=1}^{K} \frac{\pi_k \mathcal{N}_L(\mathbf{x} \mid \mathbf{c}_k, \mathbf{\Gamma}_k)}{\sum_{l=1}^K \pi_l \mathcal{N}_L(\mathbf{x} \mid \mathbf{c}_l, \mathbf{\Gamma}_l)} \mathcal{N}_D(\mathbf{y} \mid \mathbf{A}_k\mathbf{x} + \mathbf{b}_k, \mathbf{\Sigma}_k). \tag{2}$$

Using the inverse regression framework, in (1), 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 previous hierarchical one in (2), its corresponding forward conditional density from $\mathbb{R}^D$ to $\mathbb{R}^L$ is defined by

$$p(\mathbf{x} \mid \mathbf{y}; \boldsymbol{\psi}^*) = \sum_{k=1}^{K} \frac{\pi_k^* \mathcal{N}_D(\mathbf{y} \mid \mathbf{c}_k^*, \mathbf{\Gamma}_k^*)}{\sum_{l=1}^K \pi_l^* \mathcal{N}_D(\mathbf{y} \mid \mathbf{c}_l^*, \mathbf{\Gamma}_l^*)} \mathcal{N}_L(\mathbf{x} \mid \mathbf{A}_k^*\mathbf{y} + \mathbf{b}_k^*, \mathbf{\Sigma}_k^*). \tag{3}$$

A useful feature of GLLiM models is described in the following Lemma 2.1, established for multivariate Gaussian and Student components in Deleforge et al. (2015); Perthame et al. (2018) and which can be straightforwardly extended to Gaussian scale mixtures and elliptical distributions (Nguyen et al., 2022c; Ingrassia et al., 2012).

Lemma 2.1. *The parameter $\boldsymbol{\psi}^*$ in the forward cPDF, defined in (3), can then be deduced from $\boldsymbol{\psi}$ in (2) via the following one-to-one correspondence:*

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

Remark 1 (GLLiM models are computationally efficient). Without assuming anything about the structure of the parameters, the dimension of GLLiM is $\dim(\boldsymbol{\Psi}) = K(1 + D(L+1) + D(D+1)/2 + L(L+1)/2 + L) - 1$. It is worth noting that $\dim(\boldsymbol{\Psi})$ can be very large compared to the sample size (see, *e.g.,* Deleforge et al. 2015 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 $\boldsymbol{\Sigma}_k$ of the error vectors $\mathbf{E}_k$ rather than on $\boldsymbol{\Gamma}_k$ (*cf.,* Section 3 Deleforge et al., 2015). This justifies using the inverse regression trick from Deleforge et al. (2015), drastically reducing the number of parameters to be estimated. For instance, $\mathbf{E}_k$ can be modelled with equal isotropic Gaussian noise, so we have $\boldsymbol{\Sigma}_k = \tilde{\sigma}^2\,\mathrm{I}_D, \forall k \in [K]$, with some positive $\tilde{\sigma}^2$. The number of parameters to be estimated is then $K + K\,(L + L(L+1)/2 + DL + D)$. 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 $5,035,030 \gg 30,060$ in the previous example.

A notable recent illustration of the GLLiM good features is described in Blein-Nicolas et al. (2024), tackling the challenge of nonlinear quantitative trait prediction using biological data. In this work, the authors focused on predicting a small set of continuous quantitative traits ($L = 2$) from a large set of biomarkers ($D \approx 1000 \gg L$). Their inverse regression approach is not only computationally efficient, but also has the advantage of preserving all covariates while assuming a specific structure for the covariance matrix, which aids dimensionality reduction and improves prediction accuracy. This method operates by estimating an inverse model through several low-dimensional regressions and then inverting these estimators to solve the initial high-dimensional regression problem using Lemma 2.1.

2.3. Construction of BNP-GLLiM model

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

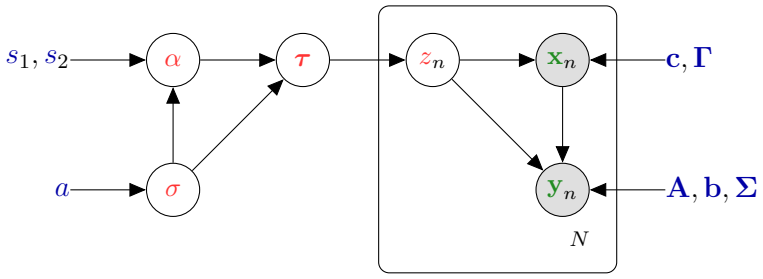


Figure 1. Graphical representation of BNP-GLLiM: the plate denotes N i.i.d. observations, white-filled circles correspond to latent variables and random or unknown parameters represented in red, while grey-filled circles correspond to observed variables represented in green. Hyperparameters are represented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

(1) BNP prior: $G \sim \text{BNP}\,(\alpha, \sigma, G_0)$.

$$(\alpha, \sigma) \mid s_1, s_2, a \sim \mathcal{SG}\,(\alpha \mid \sigma; s_1, s_2)\, p(\sigma \mid a) \equiv \text{Gam}\,(\alpha + \sigma \mid s_1, s_2)\, p(\sigma \mid a), \quad (4)$$

$$\tau_k \mid \alpha, \sigma \overset{\text{ind}}{\sim} \text{Beta}\,(\tau_k \mid 1 - \sigma, \alpha + k\sigma), \quad k \in \mathbb{N}^\star. \quad (5)$$

Define $\pi_k(\boldsymbol{\tau}) = \tau_k \prod_{l=1}^{k-1} (1 - \tau_l), k \in \mathbb{N}^\star$, $\boldsymbol{\theta}_k^\star \mid G_0 \overset{\text{iid}}{\sim} G_0$, where $\boldsymbol{\theta}_k^\star \equiv (\mathbf{c}_k, \mathbf{\Gamma}_k, \mathbf{A}_k, \mathbf{b}_k, \mathbf{\Sigma}_k), k \in \mathbb{N}^\star$, and finally define $G = \sum_{k=1}^{\infty} \pi_k(\boldsymbol{\tau}) \delta_{\boldsymbol{\theta}_k^\star}$.

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

$$\boldsymbol{\theta}_n \mid G \overset{\text{iid}}{\sim} G, \quad \text{if } \boldsymbol{\theta}_n = \boldsymbol{\theta}_k^\star, \text{ set } z_n = k; \quad (6)$$

$$\mathbf{x}_n \mid z_n, \mathbf{c}, \mathbf{\Gamma} \overset{\text{iid}}{\sim} \mathcal{N}_L\,(\mathbf{x}_n \mid \mathbf{c}_{z_n}, \mathbf{\Gamma}_{z_n}), \quad (\mathbf{c}, \mathbf{\Gamma}) \equiv (\mathbf{c}_k, \mathbf{\Gamma}_k)_{k \in \mathbb{N}^\star}; \quad (7)$$

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

3. Variational inference for BNP-GLLiM

For a brief summary of variational Bayesian expectation maximisation (VBEM) and its notation, see Appendix B. The task of conditional density estimation and clustering using the BNP-GLLiM model is mainly to estimate the unknown labels $\mathcal{Z} = (\mathbf{z}_n)_{n \in [N]}$ from the observed data $(\mathcal{Y}, \mathcal{X}) = (\mathbf{y}_n, \mathbf{x}_n)_{n \in [N]}$, whose joint distribution $p\,(\mathcal{Y}, \mathcal{X}, \mathcal{Z}, \mathbf{\Theta}; \boldsymbol{\phi})$ is determined by a set of BNP prior parameters $\mathbf{\Theta} = (\boldsymbol{\tau}, \alpha, \sigma, \boldsymbol{\theta}^\star)$, namely the stick-breaking construction of Pitman–Yor process (PYP) (Pitman and Yor, 1997) in Appendix A; and by additional hyperparameters $\boldsymbol{\phi} = (s_1, s_2, a)$. Then the desired joint distribution is given by:

$$p(\mathcal{Y}, \mathcal{X}, \mathcal{Z}, \mathbf{\Theta}; \boldsymbol{\phi}) = \prod_{n=1}^{N} p\,(\mathbf{y}_n \mid \mathbf{x}_n, z_n; \mathbf{A}, \mathbf{b}, \mathbf{\Sigma})\, p\,(\mathbf{x}_n \mid z_n, \mathbf{c}, \mathbf{\Gamma})\, p(\mathcal{Z} \mid \boldsymbol{\tau}) \times \prod_{k \in \mathbb{N}^\star} p\,(\tau_k \mid \alpha, \sigma)\, p\,(\alpha, \sigma \mid s_1, s_2, a)\, p(\boldsymbol{\theta}^\star | G_0). \quad (8)$$

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: $q(\mathcal{Z}, \mathbf{\Theta}) = q_{\mathcal{Z}}(\mathcal{Z}) q_{\mathbf{\Theta}}(\mathbf{\Theta})$. In particular, the intractable posterior

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

$$q_{\mathcal{Z}}(\mathcal{Z}) = \prod_{n=1}^{N} q_{z_n}(z_n). \tag{9}$$

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.,* Blei and Jordan (2006); Wang et al. (2011). 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 $q_{\boldsymbol{\tau}}(\boldsymbol{\tau}) = \prod_{k=1}^{K-1} q_{\tau_k}(\tau_k)$ with the additional condition that $\tau_K = 1$. Thus the truncated variational posterior of parameters $\boldsymbol{\Theta}$ is given by

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

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 C. 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 F.

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. More precisely, to achieve this, we use (5), (6), (8), and are only interested in the functional dependence of (B3) 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 C.1, it holds that

$$q_{\tau_k}(\tau_k) = \mathrm{Beta}(\tau_k \mid \widehat{\gamma}_{k,1}, \widehat{\gamma}_{k,2}), \text{ where,}$$
$$\widehat{\gamma}_{k,1} = 1 - \mathbb{E}_{q_{\alpha,\sigma}}[\sigma] + N_k, \quad \widehat{\gamma}_{k,2} = \mathbb{E}_{q_{\alpha,\sigma}}[\alpha] + k\mathbb{E}_{q_{\alpha,\sigma}}[\sigma] + \sum_{l=k+1}^{K} N_l.$$

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

$$\widehat{s}_1 = s_1 + K - 1, \quad \widehat{s}_2 = s_2 - \sum_{k=1}^{K-1} \psi(\widehat{\gamma}_{k,2}) - \psi(\widehat{\gamma}_{k,1} + \widehat{\gamma}_{k,2}), \tag{10}$$

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 $\mathrm{Gam}(\widehat{s}_1, \widehat{s}_2)$ with $\mathbb{E}_{q_{\alpha,\sigma}}[\alpha] = \widehat{s}_1/\widehat{s}_2$. 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 C.2 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 (9) and the truncation (see Appendix C.3), for all $n \in [N]$ and for all $k \in [K]$, this step consists in computing

$$q_{z_n}(k) = \frac{\rho_{nk}}{\sum_{l=1}^{K} \rho_{nl}}, \text{ where we define } \log \rho_{nk} \text{ as follows:} \tag{11}$$

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

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 Specifically defined in the following Section 3.2. It is important to note that (11) 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 $\widehat{z}_n = \operatorname{argmax}_{k\in[K]} q_{z_n}(k)$.

3.2. VB-M steps

The maximisation 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 as follows:

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

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,

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

where $(\widehat{s}_1^{(r)}, \widehat{s}_2^{(r)})$ is given in (10). We can also solve this step numerically using importance sampling in the more general case of PYP ($\sigma \neq 0$). For more details, see Appendix A.7 in Lü et al. (2020).

VB-M-$(\mathbf{c}, \mathbf{\Gamma})$ step. This step is divided into solving K optimisation problems:

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

We can then update the Gaussian gating parameters as follows:

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

The technical details will be left to the Appendix C.4.

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 optimisation problems

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

Given the following quantities:

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

We can update the parameters for the Gaussian experts as follows (cf. Appendix C.5):

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

3.3. Evidence lower-bound (ELBO)

Evaluating the ELBO in (B1) allows us to not only 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 (B1) at each step of the iterative re-estimation procedure should not decrease (Svensén and Bishop, 2005), in particular, see recent results for Bayesian nonparametric mixture models in Appendix A of Durand et al. (2022). Recall that $\widehat{\boldsymbol{\phi}} = (\widehat{s}_1, \widehat{s}_2, \widehat{a}, (\widehat{\mathbf{c}}_k, \widehat{\mathbf{\Gamma}}_k, \widehat{\mathbf{A}}_k, \widehat{\mathbf{b}}_k, \widehat{\mathbf{\Sigma}}_k)_{k \in \mathbb{N}^\star})$. 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 are 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.*, Yuan and Neubauer

(2008); Luo and Sun (2017); Wu and Ma (2018); Nguyen and Bonilla (2014). When $\sigma = 0$, we can analytically compute the ELBO in the BNP-GLLiM via Proposition 3.1 which is proved in Appendix E.2.

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

$$\mathcal{F}\left(q_{\mathcal{Z}}, q_{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\phi}}\right) = \mathbb{E}\left[\log p(\mathcal{Y} \mid \mathcal{X}, \mathcal{Z}, \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] + \mathbb{E}\left[\log p(\mathcal{X} \mid \mathcal{Z}, \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] + \mathbb{E}\left[\log p(\mathcal{Z} \mid \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] + \mathbb{E}\left[\log p(\boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] - \mathbb{E}\left[\log q(\mathcal{Z})\right] - \mathbb{E}\left[\log q(\boldsymbol{\Theta})\right], \tag{12}$$

where all the terms have a closed-form expression.

The closed-form expressions for the terms of the right-hand side of Equation (12) are provided in Appendix D. 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 (7), relate to conditional density estimation and data clustering. Specifically, we are interested in the predicted conditional density for a new value $(\widehat{\mathbf{y}}, \widehat{\mathbf{x}})$ of the observed variables. Note that there will be a corresponding latent variable $\widehat{\mathbf{z}}$ associated with these observations. If $\sigma \neq 0$, we can use the previous remark in Section 3.3, where the integral reduces to a point evaluation at the posterior mean of each parameter. When $\sigma = 0$, we can analytically approximate such densities via several following theorems. In 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 K.

4.1. Joint density

We first compute the joint density via Theorem 4.1, which is proved in Appendix E.3.

Theorem 4.1. *With $\widehat{\mathbf{w}} \equiv [\widehat{\mathbf{x}}; \widehat{\mathbf{y}}]$, we have*

$$\boldsymbol{\mu}_k \equiv \mathbb{E}\left[\widehat{\mathbf{w}}\right] = \begin{pmatrix} \widehat{\mathbf{c}}_k \\ \widehat{\mathbf{A}}_k \widehat{\mathbf{c}}_k + \widehat{\mathbf{b}}_k \end{pmatrix}, \quad \mathbf{V}_k = cov\left[\widehat{\mathbf{w}}\right] = \begin{pmatrix} \widehat{\boldsymbol{\Gamma}}_k & \widehat{\boldsymbol{\Gamma}}_k \widehat{\mathbf{A}}_k^\top \\ \widehat{\mathbf{A}}_k \widehat{\boldsymbol{\Gamma}}_k & \widehat{\boldsymbol{\Sigma}}_k + \widehat{\mathbf{A}}_k \widehat{\boldsymbol{\Gamma}}_k \widehat{\mathbf{A}}_k^\top \end{pmatrix}, \tag{13}$$

$$p\left(\widehat{\mathbf{y}}, \widehat{\mathbf{x}}, \mathcal{X}, \mathcal{Y}\right) \approx \sum_{k=1}^{K} \mathbb{E}_{q_{\tau_k}}\left[\tau_k\right] \prod_{l=1}^{k-1} \mathbb{E}_{q_{\tau_l}}\left[1 - \tau_l\right] \mathcal{N}_{L+D}\left(\widehat{\mathbf{w}} \mid \boldsymbol{\mu}_k, \mathbf{V}_k\right), \tag{14}$$

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

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 E.4.

Theorem 4.2. *We approximate the inverse conditional density $p(\widehat{\mathbf{y}} \mid \widehat{\mathbf{x}}, \mathcal{X}, \mathcal{Y})$ and its conditional expectation for prediction by:*

$$p\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{x}}, \mathcal{X}, \mathcal{Y}\right) \approx \sum_{k=1}^{K} g_{Lk}\left(\widehat{\mathbf{x}} \mid \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\phi}}, \mathcal{X}, \mathcal{Y}\right) \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{A}}_k \widehat{\mathbf{x}} + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right), \tag{15}$$

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

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

4.3. Forward conditional density

Given the inverse conditional density $p(\widehat{\mathbf{y}} \mid \widehat{\mathbf{x}}, \mathcal{X}, \mathcal{Y})$, we approximate the following forward conditional density via Theorem 4.3, whose proof is provided in Appendix E.5.

Theorem 4.3. *We approximate the forward conditional density and its corresponding conditional expectation and variance for prediction and uncertainty estimation by*

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

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

$$var\left[\widehat{\mathbf{x}} \mid \widehat{\mathbf{y}}, \mathcal{X}, \mathcal{Y}\right] \approx \sum_{k=1}^{K} g_{Dk}\left(\widehat{\mathbf{y}} \mid \widehat{\boldsymbol{\Theta}}^*, \widehat{\boldsymbol{\phi}}^*, \mathcal{X}, \mathcal{Y}\right)\left[\widehat{\boldsymbol{\Sigma}}_k^* + \left(\widehat{\mathbf{A}}_k^* \widehat{\mathbf{y}} + \widehat{\mathbf{b}}_k^*\right)\left(\widehat{\mathbf{A}}_k^* \widehat{\mathbf{y}} + \widehat{\mathbf{b}}_k^*\right)^\top\right]$$
$$- \mathbb{E}\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{y}}, \mathcal{X}, \mathcal{Y}\right) \mathbb{E}\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{y}}, \mathcal{X}, \mathcal{Y}\right)^\top, \text{ where}$$

$$g_{Dk}\left(\widehat{\mathbf{y}} \mid \widehat{\boldsymbol{\Theta}}^*, \widehat{\boldsymbol{\phi}}^*, \mathcal{X}, \mathcal{Y}\right) = \frac{\mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k(\boldsymbol{\tau})\right] \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{c}}_k^*, \widehat{\boldsymbol{\Gamma}}_k^*\right)}{\sum_{l=1}^{K} \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_l(\boldsymbol{\tau})\right] \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{c}}_l^*, \widehat{\boldsymbol{\Gamma}}_l^*\right)},$$

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

5. Bayesian nonparametric model selection

Notations. 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 $\sum_{k \in [K]} q_{kl} = \pi_l^0$ and $\sum_{l \in [K_0]} q_{kl} = \pi_k$, 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, with $\mathcal{G} \equiv \mathcal{G}(\boldsymbol{\Theta}) = \bigcup_{K \in \mathbb{N}_+} \mathcal{E}_K$ to denote 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 $\boldsymbol{\Theta}$.

5.1. Posterior contraction rate in Bayesian infinite mixtures

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

$$p_{G_0} \equiv \int \mathcal{N}(\mathbf{w} \mid \boldsymbol{\theta}) dG_0(\boldsymbol{\theta}) = \sum_{k\in[K_0]} \pi_k^0 \mathcal{N}(\mathbf{w} \mid \boldsymbol{\theta}_k^0), \quad \boldsymbol{\theta}_k^0 \equiv \left(\boldsymbol{\mu}_k^0, \mathbf{V}_k^0\right), \tag{16}$$

where $G_0 = \sum_{k\in[K_0]} \pi_k^0 \delta_{\boldsymbol{\theta}_k^0}$ is a true but unknown mixing distribution with exactly K_0 number of support points, where K_0 is also unknown. Furthermore, $\boldsymbol{\Theta}$ is a chosen parameter space, to which we 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:

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

Here, the GMM p_G is defined in (16) 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 Nguyen (2013); Ho and Nguyen (2016b). This distance can be defined as the optimal cost of moving masses in the transformation from one probability measure to another (Villani, 2003, 2009).

Definition 5.1. Suppose $\boldsymbol{\Theta}$ is equipped with a metric d. The Wasserstein distance W_r between two discrete measures $G = \sum_{k\in[K]} \pi_k \delta_{\boldsymbol{\theta}_k}$ and $G_0 = \sum_{l\in[K_0]} \pi_l^0 \delta_{\boldsymbol{\theta}_l^0}$ is given by

$$W_r\left(G, G_0\right) = \inf_{\mathbf{Q}\in\mathcal{Q}(\boldsymbol{\pi},\boldsymbol{\pi}^0)} \left[\sum_{k\in[K],l\in[K_0]} q_{kl} \left[d\left(\boldsymbol{\theta}_k, \boldsymbol{\theta}_l^0\right)\right]^r\right]^{1/r},$$

where couplings $\mathbf{Q}$ are defined at the beginning of this section. See Delon and Desolneux (2020) for more details.

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 Guha et al. (2021), our starting point is the availability of a mixing measure sample G that is drawn from the posterior distribution $\Pi(G \mid \mathcal{W})$, where $\mathcal{W}$ are i.i.d. 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,

$$\Pi\left(G \in \overline{\mathcal{G}}(\boldsymbol{\Theta}) : W_r(G, G_0) \leq \delta\omega_N \mid \mathcal{W}\right) \xrightarrow{N\to\infty} 1 \text{ in } P_{G_0}\text{-probability}, \tag{17}$$

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 Nguyen (2013); Gao and Vaart (2016); Ho and Nguyen (2016b) 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 (2), 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 E.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 (2), 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.,

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

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

$$\begin{aligned} &\pi_k = \rho_k, \quad \mathbf{c}_k = \boldsymbol{\mu}_k^{\mathbf{x}}, \quad \boldsymbol{\Gamma}_k = \mathbf{V}_k^{\mathbf{xx}}, \quad \mathbf{A}_k = \mathbf{V}_k^{\mathbf{xy}\top} (\mathbf{V}_k^{\mathbf{xx}})^{-1}, \\ &\mathbf{b}_k = \boldsymbol{\mu}_k^{\mathbf{y}} - \mathbf{V}_k^{\mathbf{xy}\top} (\mathbf{V}_k^{\mathbf{xx}})^{-1} \boldsymbol{\mu}_k^{\mathbf{x}}, \quad \boldsymbol{\Sigma}_k = \mathbf{V}_k^{\mathbf{yy}} - \mathbf{V}_k^{\mathbf{xy}\top} (\mathbf{V}_k^{\mathbf{xx}})^{-1} \mathbf{V}_k^{\mathbf{xy}}. \end{aligned} \tag{18}$$

Here, we have defined

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

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:*

$$\rho_k = \pi_k, \quad \boldsymbol{\mu}_k = \begin{bmatrix} \mathbf{c}_k \\ \mathbf{A}_k\mathbf{c}_k + \mathbf{b}_k \end{bmatrix}, \quad \mathbf{V}_k = \begin{bmatrix} \boldsymbol{\Gamma}_k & (\mathbf{A}_k\boldsymbol{\Gamma}_k)^\top \\ \mathbf{A}_k\boldsymbol{\Gamma}_k & \boldsymbol{\Sigma}_k + \mathbf{A}_k\boldsymbol{\Gamma}_k\mathbf{A}_k^\top \end{bmatrix}. \tag{19}$$

Merge-Truncate-Merge (MTM) algorithm is a post-processing procedure applied to a posterior sample of the mixing measure G in BNP-MM, essential for achieving posterior contraction rates under the Wasserstein metric (Guha et al., 2021). We propose in Algorithm 1 an algorithm for BNP joint GMM which follows the same steps as the original MTM algorithm for BNP-MM from Guha et al. (2021). This algorithm involves two main stages: the Merge procedure and the Truncate-Merge procedure. In the Merge stage, atoms are reordered by simple random sampling without replacement, ensuring random permutation. Sequentially, atoms are merged based on a distance threshold ω_N, updating weights and removing merged atoms from the set, resulting in a new measure G' with reordered weights. In the Truncate-Merge stage, atoms are divided into two sets based on a weight threshold $(c\omega_N)^r$. For each atom in the significant weight set, if another atom within the threshold distance exists, it is moved to the negligible weight set. Atoms in the negligible set are then merged with the nearest significant atom, resulting in the final measure $\widetilde{G}$ and the number of its supporting atoms $\widetilde{K}$. See Algorithm 1 for a pseudo-code description.

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 (17), 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 $d(\boldsymbol{\theta}_{\tau_i}, \boldsymbol{\theta}_{\tau_j}) \leq \omega_N$, then update $\pi_{\tau_i} = \pi_{\tau_i} + \pi_{\tau_j}$, and remove τ_j from $\mathcal{E}$.

3: Collect $G' = \sum_{j:\ \tau_j\in\mathcal{E}} \pi_{\tau_j}\delta_{\boldsymbol{\theta}_{\tau_j}}$. Then 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 $\boldsymbol{\phi}_i$, then 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 of our MTM algorithm, we obtain the theoretical guarantee of 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 (17) 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 Theorem 3.2 from Guha et al. (2021) 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 2. Regarding the above theorem, 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 Guha et al. (2021), the aim of this paper is to study the practical viability of MTM Algorithm 1 and Theorem 5.3 in the context of high-to-low dimensional inverse regression via BNP-GLLiM model. In order to do this, we first need to specify the metric d in $\boldsymbol{\Theta}$, *e.g.,* $d\left(\boldsymbol{\theta}_{\tau_i}, \boldsymbol{\theta}_{\tau_j}\right) = \left\|\boldsymbol{\mu}_{\tau_i} - \boldsymbol{\mu}_{\tau_j}\right\| + \left\|\mathbf{V}_{\tau_i} - \mathbf{V}_{\tau_j}\right\|$. 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 (14) and (13), 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

The code to reproduce our simulation study is publicly available[1] and all simulations below were performed in Python 3.9.13 on a standard Unix machine. For proof-of-concept numerical experiments, we consider only the simple data generating mechanism with $D = 1$ and $L = 1$ and demonstrate that BNP-GLLiM performs well in model selection, clustering and cPDF estimation with the MTM procedure. Real world examples are postponed to future work.

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 Chamroukhi et al. (2010); Montuelle and Le Pennec (2014); Nguyen et al. (2022c). Specifically, we consider the

[1] https://github.com/Trung-TinNGUYEN/BNP-GLLiM

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 $N = 1,000$ realisations of the true inverse cPDF from GLLiM, representing a π-shape simulation with three clusters without labels.

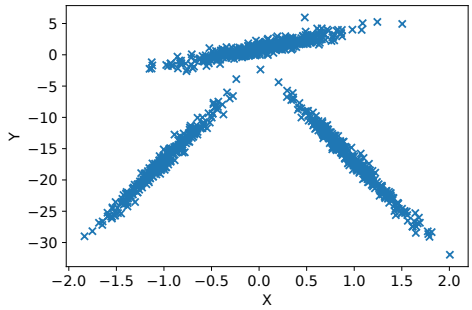
6.2. Model selection, clustering and regression tasks via MTM-BNP-GLLiM

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 2 it is clear that with the help of MTM Algorithm 1, MTM-BNP-GLLiM can simultaneously perform regression, clustering and model selection well. Without the MTM procedure, BNP-GLLiM performs poorly in model selection, clustering and cPDF estimation, except for conditional expectations as shown in Figure 3.

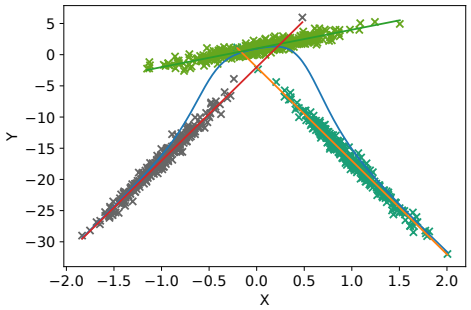
Next, we illustrate the performance of the MTM algorithm when applied to the variational posterior from BNP-GLLiM. Specifically, the samples in our 100 trials are drawn from $G_V = \sum_{k=1}^{K} \mathbb{E}_{q_\tau}\left[\pi_k(\boldsymbol{\tau})\right] \delta_{\boldsymbol{\theta}_k}$, wherere $\mathbb{E}_{q_\tau}\left[\pi_k(\boldsymbol{\tau})\right]$ and $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \mathbf{V}_k)$ are defined in (14) and (13), 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 $\widetilde{C}\left(\log(N)\right)^{-1/2}$ with respect to the W_2-norm. Similar to Guha et al. (2021), our first attempt to choose w_N to satisfy (17) 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 (17).

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 $\frac{\widetilde{C}}{\left(\log(\log(N))\right)^{1/2}} \leq c$. Furthermore, based on Equation (26) from Guha et al. (2021), with a useful lower bound on the posterior mass the mode, for any $1 > \epsilon > 0$, $(1-\epsilon)\left(1 - \sum_{k=1}^{3} \frac{c_0^{r/2}}{\pi_k^0}\right) > \frac{1}{2}$, 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

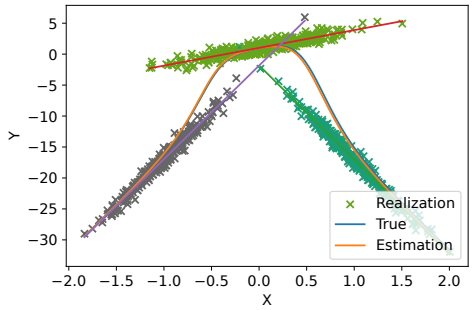
$$(1-\epsilon)\left(1 - \sum_{k=1}^{3} \frac{c_0^{r/2}}{\pi_k^0}\right) > \frac{1}{2} \Leftrightarrow \frac{1-2\epsilon}{2(1-\epsilon)}\left(\sum_{k=1}^{3} \frac{1}{\pi_k^0}\right)^{-1} > c_0^{r/2} > c^{r/2}.$$



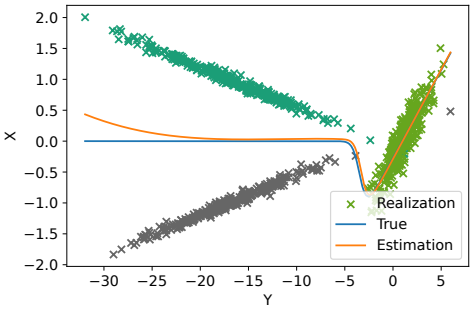
(a) Typical realisations without labels.



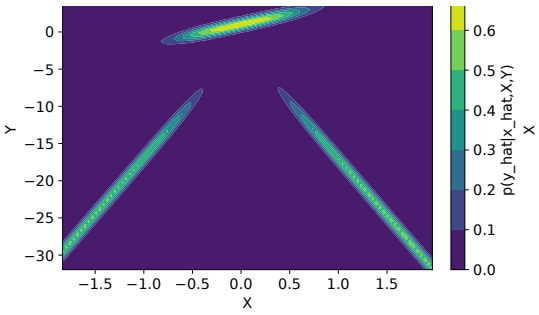
(b) True conditional expectations with labels.



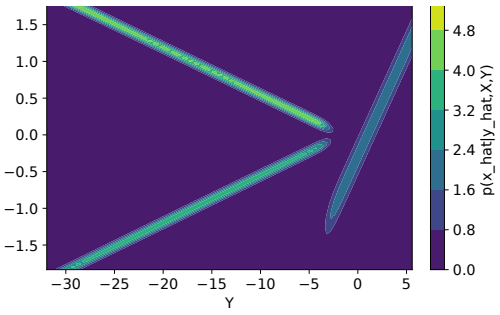
(c) True and estimated inverse means.



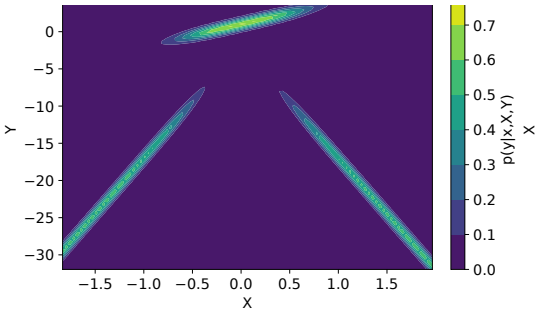
(d) True and estimated forward means.



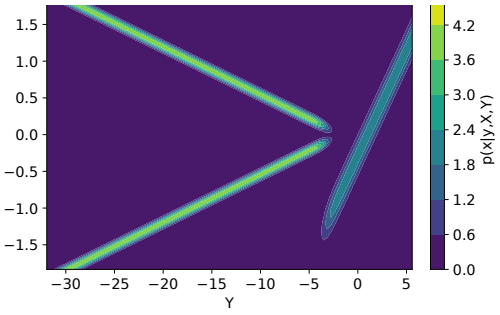
(e) Contour of estimated inverse cPDF.



(f) Contour of estimated forward cPDF.

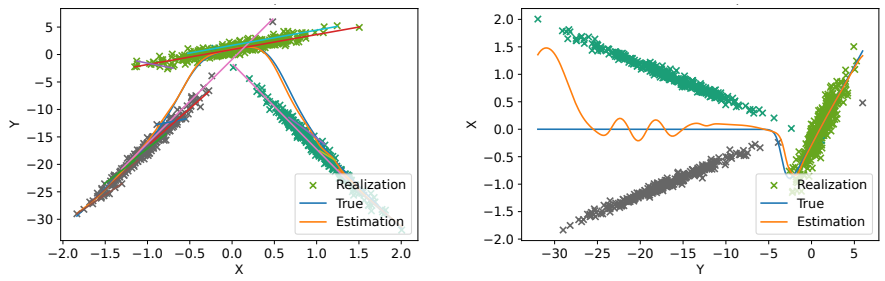


(g) Contour of true inverse cPDF.



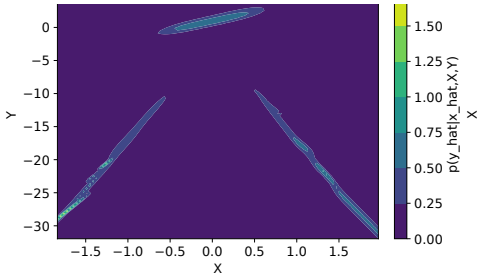
(h) Contour of true forward cPDF.

Figure 2. Top row: Typical 1,000 realisations of GLLiM's true inverse cPDF with its true conditional expectations. 2nd row and below: True and estimated inverse and forward cPDF of GLLiM with the true number of components ($K_{\mathrm{MTM}} = 3$) using MTM algorithm for post-processing in MTM-BNP-GLLiM.

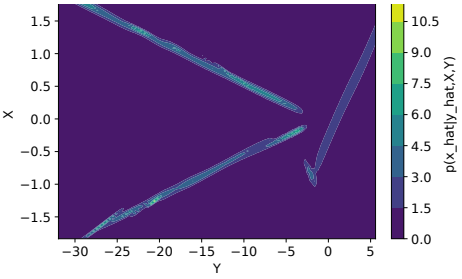


(a) True and estimated inverse means.

(b) True and estimated forward means.



(c) Contour of estimated inverse cPDF.



(d) Contour of estimated forward cPDF.

Figure 3. 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 $K = 20$.

Therefore, we can choose

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

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 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. Guha et al. (2021) also used the same setting in their experiments. Figure 4 indicates that $c = 0.45$ leads to a quite good posterior mode in our experiments.

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 4 seems to suggest that MTM-BNP-GLLiM gives a comparable good result to the location Gaussian DP-MM in the simulation studies in Guha et al. (2021).

7. Perspectives

To address the dimensionality issue when N is less than D, we could incorporate sparsity penalty terms as described in the GLLiM context in Chamroukhi et al. (2019). Alternatively, we could impose additional structural restrictions, such as block-diagonal covariance matrices as in Blein-Nicolas et al. (2024)), which extends the GLLiM method to account for hidden module-structured regulatory networks of predictors through block-diagonal structures. In particular, Blein-Nicolas et al. (2024) applied their method to

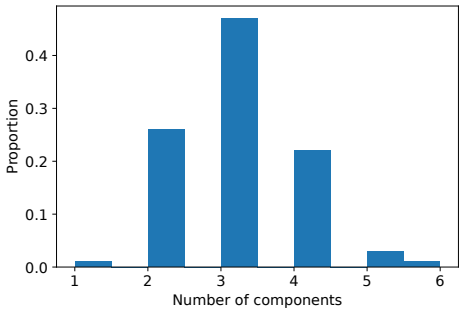


Figure 4. Histogram of K_{MTM} using variational posterior sample with 100 trials and $c = 0.45$.

the prediction of drought-related traits ($L = 2$) from protein abundance ($D = 973$) in $N = 233$ maize genotypes. We leave for future research the intriguing but challenging questions of how to integrate BNP priors into the GLLiM model and how to establish posterior convergence theory in high-dimensional settings where $D \gg N$.

As indicated in Remark 2, 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 similar "prior mass and testing" conditions as in Ghosal et al. (2000), we believe that an interesting but challenging extension of the work on variational posterior unconditional distributions for MMs (Zhang and Gao, 2020) and on adaptive Bayesian estimation for MMs and MoE models but for true posterior distribution (Kruijer et al., 2010; Shen et al., 2013; Norets and Pati, 2017) 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 Titterington and Wang (2006). 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 Ruiz and Titsias (2019). 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 Birgé and Massart (2007) is crucial for the success of the MTM procedure for any BNP model. We leave these interesting but challenging questions for future research.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723. 2

Alamichel, L., Bystrova, D., Arbel, J., and Kon Kam King, G. (2024). Bayesian mixture models (in)consistency for the number of clusters. *Scandinavian Journal of Statistics*. 3

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

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

Ascolani, F., Lijoi, A., Rebaudo, G., and Zanella, G. (2022). Clustering consistency with Dirichlet process mixtures. *Biometrika*, 110(2):551–558. 3

Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference*. PhD thesis, University of London, University College London (United Kingdom). 30

Beal, M. J. and Ghahramani, Z. (2003). The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. *Bayesian statistics*, 7(453-464):210. 30

Biernacki, C., Celeux, G., and 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. 2

Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probability Theory and Related Fields*, 138(1):33–73. 2, 20

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg. 30, 48, 51, 55, 63, 67

Blei, D. M. and Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian analysis*, 1(1):121–143. 8, 28, 47, 49, 61

Blein-Nicolas, M., Devijver, E., Gallopin, M., and Perthame, E. (2024). Nonlinear network-based quantitative trait prediction from biological data. *Journal of the Royal Statistical Society Series C: Applied Statistics*, page qlae012. 2, 6, 19

Boux, F., Forbes, F., Arbel, J., Lemasson, B., and Barbier, E. L. (2021). Bayesian inverse regression for vascular magnetic resonance fingerprinting. *IEEE Transactions on Medical Imaging*, 40(7):1827–1837. 1

Celeux, G., Frühwirth-Schnatter, S., and Robert, C. P. (2019). Model selection for mixture models–perspectives and strategies. In *Handbook of mixture analysis*, pages 117–154. Chapman and Hall/CRC. 2

Chaari, L., Vincent, T., Forbes, F., Dojat, M., and 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. 30

Chamroukhi, F., Lecocq, F., and Nguyen, H. D. (2019). Regularized Estimation and Feature Selection in Mixtures of Gaussian-Gated Experts Models. In Nguyen, H., editor, *Statistics and Data Science*, pages 42–56, Singapore. Springer Singapore. 19

Chamroukhi, F., Samé, A., Govaert, G., and Aknin, P. (2010). A hidden process regression model for functional data description. Application to curve discrimination. *Neurocomputing*, 73(7-9):1210–1221. 16

Chen, Z., Deng, Y., Wu, Y., Gu, Q., and Li, Y. (2022). Towards Understanding the Mixture-of-Experts Layer in Deep Learning. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*. 2

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., and 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. 28

Deleforge, A., Forbes, F., and Horaud, R. (2015). High-dimensional regression with gaussian mixtures and partially-latent response variables. *Statistics and Computing*, 25(5):893–911. 1, 4, 5, 6, 43

Delon, J. and Desolneux, A. (2020). A Wasserstein-type distance in the space of Gaussian mixture models. *SIAM Journal on Imaging Sciences*, 13(2):936–970. 13

Dempster, A. P., Laird, N. M., and 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. 29

Do, T. G., Le, H. K., Nguyen, T., Pham, Q., Nguyen, B. T., Doan, T.-N., Liu, C., Ramasamy, S., Li, X., and 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. 2

Durand, J.-B., Forbes, F., Phan, C., Truong, L., Nguyen, H., and 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. 3, 10, 29

Escobar, M. D. and West, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, 90(430):577–588. 3

Favaro, S., Lijoi, A., Nava, C., Nipoti, B., Pruenster, I., and Teh, Y. W. (2016). On the

stick-breaking representation for homogeneous NRMIs. *Bayesian Analysis*, 11(3):697–724. 28

Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209 – 230. 27

Forbes, F., Arnaud, A., Lemasson, B., and 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*, pages 81–95, Melbourne, Australia. Springer. 3

Forbes, F., Nguyen, H. D., Nguyen, T., and 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. 2

Forbes, F., Nguyen, H. D., Nguyen, T., and Arbel, J. (2022b). Summary statistics and discrepancy measures for approximate Bayesian computation via surrogate posteriors. *Statistics and Computing*, 32(5):85. 2

Gao, F. and Vaart, A. v. d. (2016). Posterior contraction rates for deconvolution of Dirichlet-Laplace mixtures. *Electronic Journal of Statistics*, 10(1):608 – 627. 14

Genovese, C. R. and Wasserman, L. (2000). Rates of convergence for the Gaussian mixture sieve. *The Annals of Statistics*, 28(4):1105 – 1127. 2

Ghosal, S., Ghosh, J. K., and Vaart, A. W. v. d. (2000). Convergence rates of posterior distributions. *The Annals of Statistics*, 28(2):500 – 531. 20

Ghosal, S. and Van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press. 3, 28, 49

Green, P. J. and Richardson, S. (2001). Modelling Heterogeneity With and Without the Dirichlet Process. *Scandinavian Journal of Statistics*, 28(2):355–375. 3

Guha, A., Ho, N., and Nguyen, X. (2021). On posterior contraction of parameters and interpretability in Bayesian mixture modeling. *Bernoulli*, 27(4):2159 – 2188. 3, 14, 15, 16, 17, 19

Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G., editors (2010). *Bayesian Nonparametrics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. 3

Ho, N. and Nguyen, X. (2016a). Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *The Annals of Statistics*, 44(6):2726 – 2755. 2

Ho, N. and Nguyen, X. (2016b). On strong identifiability and convergence rates of parameter estimation in finite mixtures. *Electronic Journal of Statistics*, 10(1):271–307. 2, 13, 14

Ho, N., Yang, C.-Y., and Jordan, M. I. (2022). Convergence Rates for Gaussian Mixtures of Experts. *Journal of Machine Learning Research*, 23(323):1–81. 2

Hoadley, B. (1970). A Bayesian Look at Inverse Linear Regression. *Journal of the American Statistical Association*, 65(329):356–369. 1

Ingrassia, S., Minotti, S. C., and Vittadini, G. (2012). Local Statistical Modeling via a Cluster-Weighted Approach with Elliptical Distributions. *Journal of Classification*, 29(3):363–401. 6

Ishwaran, H. and James, L. F. (2001). Gibbs Sampling Methods for Stick-Breaking Priors. *Journal of the American Statistical Association*, 96(453):161–173. 28

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87. 2

Jiang, W. and Tanner, M. A. (1999). Hierarchical mixtures-of-experts for exponential family regression models: approximation and maximum likelihood estimation. *Annals of Statistics*, pages 987–1011. 2

Kock, L., Klein, N., and Nott, D. J. (2022). Variational inference and sparsity in high-dimensional deep Gaussian mixture models. *Statistics and Computing*, 32(5):70. 3

Kruijer, W., Rousseau, J., and Vaart, A. v. d. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics*, 4(none):1225 – 1257. 20

Kugler, B., Forbes, F., and Douté, S. (2022). Fast Bayesian Inversion for high dimensional inverse problems. *Statistics and Computing*, 32(2). 2

Lathuilière, S., Juge, R., Mesejo, P., Muñoz-Salinas, R., and Horaud, R. (2017). Deep mixture of linear inverse regressions applied to head-pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4817–4825. 2

Li, K.-C. (1991). Sliced Inverse Regression for Dimension Reduction. *Journal of the American Statistical Association*, 86(414):316–327. 1, 4

Li, N., Li, W., Jiang, Y., and Xia, S.-T. (2022). Deep Dirichlet process mixture models. In Cussens, J. and Zhang, K., editors, *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 1138–1147. PMLR. 3

Luo, C. and Sun, S. (2017). Variational Mixtures of Gaussian Processes for Classification. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 4603–4609. 11

Lü, H., Arbel, J., and Forbes, F. (2020). Bayesian nonparametric priors for hidden Markov random fields. *Statistics and Computing*, 30(4):1015–1035. 9, 28, 29

Maceachern, S. N. and Müller, P. (1998). Estimating Mixture of Dirichlet Process Models. *Journal of Computational and Graphical Statistics*, 7(2):223–238. 3

McLachlan, G. J. and Krishnan, T. (1997). *The EM algorithm and extensions*. Wiley, New York, USA. 29

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. John Wiley & Sons. 2

Miller, J. W. and Harrison, M. T. (2014). Inconsistency of Pitman-Yor Process Mixtures for the Number of Components. *Journal of Machine Learning Research*, 15(96):3333–3370. 3

Montuelle, L. and 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. 16

Neal, R. M. (2000). Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, 9(2):249–265. 3

Neal, R. M. and Hinton, G. E. (1998). A View of the EM Algorithm that Justifies Incremental, Sparse, and other Variants. In Jordan, M. I., editor, *Learning in Graphical Models*, pages 355–368. Springer Netherlands, Dordrecht. 29

Nguyen, D. T., Jacquemoud, S., Lucas, A., Douté, S., Ferrari, C., Coustance, S., Marcq, S., and Meygret, A. (2024). Unveiling the Characteristics of the Lunar Surface by Massive Inversion of a Photometric Model. In *LPI Contributions*, volume 3040 of *LPI Contributions*, page 1998. 2

Nguyen, H., Akbarian, P., Nguyen, T., and Ho, N. (2024a). A General Theory for Softmax Gating Multinomial Logistic Mixture of Experts. In *Proceedings of The 41st International Conference on Machine Learning*. 2

Nguyen, H., Nguyen, T., and Ho, N. (2023a). Demystifying Softmax Gating in Gaussian Mixture of Experts. In *Advances in Neural Information Processing Systems*. 2

Nguyen, H., Nguyen, T., Nguyen, K., and Ho, N. (2024b). Towards Convergence Rates for Parameter Estimation in Gaussian-gated Mixture of Experts. In Dasgupta, S., Mandt, S., and Li, Y., editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 2683–2691. PMLR. 2

Nguyen, H. D. and 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. 2

Nguyen, H. D., Chamroukhi, F., and Forbes, F. (2019). Approximation results regarding the multiple-output Gaussian gated mixture of linear experts model. *Neurocomputing*, 366:208–214. 2

Nguyen, H. D., Fryer, D., and McLachlan, G. J. (2022a). Order selection with confidence for finite mixture models. *Journal of the Korean Statistical Society*. 3

Nguyen, H. D., Lloyd-Jones, L. R., and McLachlan, G. J. (2016). A universal approximation theorem for mixture-of-experts models. *Neural computation*, 28(12):2585–2593. 2

Nguyen, H. D., Nguyen, T., Chamroukhi, F., and 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. 2

Nguyen, T. (2021). *Model Selection and Approximation in High-dimensional Mixtures of Experts Models: from Theory to Practice*. PhD Thesis, Normandie Université. 2

Nguyen, T. and Bonilla, E. (2014). Fast Allocation of Gaussian Process Experts. In Xing, E. P. and Jebara, T., editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 145–153, Bejing, China. PMLR. 11

Nguyen, T., Chamroukhi, F., Nguyen, H. D., and Forbes, F. (2021b). Non-asymptotic model selection in block-diagonal mixture of polynomial experts models. *Preprint. arXiv:2104.08959*. 2

Nguyen, T., Chamroukhi, F., Nguyen, H. D., and 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. 2

Nguyen, T., Chamroukhi, F., Nguyen, H. D., and McLachlan, G. J. (2023b). Approximation of probability density functions via location-scale finite mixtures in Lebesgue spaces. *Communications in Statistics - Theory and Methods*, 52(14):5048–5059. 2

Nguyen, T., Nguyen, D. N., Nguyen, H. D., and Chamroukhi, F. (2023c). 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*. 2

Nguyen, T., Nguyen, H. D., Chamroukhi, F., and 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. 2, 6, 16

Nguyen, T., Nguyen, H. D., Chamroukhi, F., and McLachlan, G. J. (2020). Approximation by finite mixtures of continuous density functions that vanish at infinity. *Cogent Mathematics & Statistics*, 7(1):1750861. 2

Nguyen, T., Nguyen, H. D., Chamroukhi, F., and McLachlan, G. J. (2023d). Non-asymptotic oracle inequalities for the Lasso in high-dimensional mixture of experts. *arXiv:2009.10622*. 2

Nguyen, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics*, 41(1):370–400. 2, 13, 14

Norets, A. (2010). Approximation of conditional densities by smooth mixtures of regressions. *The Annals of Statistics*, 38(3):1733 – 1766. 2

Norets, A. and Pati, D. (2017). Adaptive Bayesian estimation of conditional densities. *Econometric Theory*, 33(4):980–1012. 20

Perthame, E., Forbes, F., and Deleforge, A. (2018). Inverse regression approach to robust nonlinear high-to-low dimensional mapping. *Journal of Multivariate Analysis*, 163(C):1–14. 6

Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855–900. 7, 28

Rakhlin, A., Panchenko, D., and Mukherjee, S. (2005). Risk bounds for mixture density estimation. *ESAIM: PS*, 9:220–229. 2

Rousseau, J. and 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. 3

Ruiz, F. and Titsias, M. (2019). A Contrastive Divergence for Combining Variational Inference and MCMC. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5537–5545. PMLR. 20

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464. 2

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650. 28

Shen, W., Tokdar, S. T., and Ghosal, S. (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika*, 100(3):623–640. 2, 20

Svensén, M. and Bishop, C. M. (2005). Robust Bayesian mixture modelling. *Neurocomputing*, 64:235–252. 10

Titterington, D. M. and 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. 20

Villani, C. (2003). *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society. 13

Villani, C. (2009). *Optimal transport: old and new*, volume 338. Springer. 13

Wang, C., Paisley, J., and Blei, D. M. (2011). Online Variational Inference for the Hierarchical Dirichlet Process. In Gordon, G., Dunson, D., and Dudík, M., editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 752–760, Fort Lauderdale, FL, USA. PMLR. 8

Westerhout, J., Nguyen, T., Guo, X., and Nguyen, H. D. (2024). On the Asymptotic Distribution of the Minimum Empirical Risk. In *Forty-first International Conference on Machine Learning*. 2

Wu, D. and 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. 11

Xu, L., Jordan, M., and Hinton, G. E. (1995). An Alternative Model for Mixtures of Experts. In Tesauro, G., Touretzky, D., and Leen, T., editors, *Advances in Neural Information Processing Systems*, volume 7. MIT Press. 1

Yuan, C. and Neubauer, C. (2008). Variational Mixture of Gaussian Process Experts. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc. 10

Yuksel, S. E., Wilson, J. N., and Gader, P. D. (2012). Twenty Years of Mixture of Experts. *IEEE Transactions on Neural Networks and Learning Systems*, 23(8):1177–1193. 2

Zhang, F. and Gao, C. (2020). Convergence rates of variational posterior distributions. *The Annals of Statistics*, 48(4):2180 – 2207. 20

Supplementary Materials for "Bayesian nonparametric mixture of experts for inverse problems"

In this supplementary material, we first recall the standard Bayesian nonparametric priors and variational Bayesian expectation-maximisation principle in Appendices A and B, respectively. Then, all specifications of the VBEM for the BNP-GLLiM model, the evidence lower bound and the technical proofs that are not included in the main paper are placed in Appendices C to E. . Finally, Appendix F proposes a more general model with a hyper prior on the gating parameters, referred to as BNP-GLLiM2.

Contents

Appendix A. Bayesian nonparametric priors

Stick-breaking construction of Dirichlet process. Note that the Dirichlet process (DP) (Ferguson, 1973) 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:

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

We use the stick-breaking construction of the DP G (SBDP), due to Sethuraman (1994):

$$\begin{aligned}
&\boldsymbol{\theta}_k^0 \mid G_0 \overset{\text{iid}}{\sim} G_0, \quad k \in \mathbb{N}^\star, \\
&\tau_k \mid \alpha \overset{\text{iid}}{\sim} \text{Beta}(\tau_k \mid 1, \alpha), \quad k \in \mathbb{N}^\star, \\
&\pi_k(\boldsymbol{\tau}) = \tau_k \prod_{l=1}^{k-1} (1 - \tau_l), \quad k \in \mathbb{N}^\star, \\
&G = \sum_{k=1}^{\infty} \pi_k(\boldsymbol{\tau}) \delta_{\boldsymbol{\theta}_k^0} \sim \text{DP}(\alpha, G_0).
\end{aligned}$$

Pitman–Yor process. As a generalized version of the Dirichlet process, in the Pitman–Yor process (PYP) (Pitman and Yor, 1997), the τ_k's are independent ($\overset{\text{ind}}{\sim}$) but not identically distributed. Specifically,

$$\begin{aligned}
&\boldsymbol{\theta}_k^0 \mid G_0 \overset{\text{iid}}{\sim} G_0, \quad k \in \mathbb{N}^\star, \\
&\tau_k \mid \alpha, \sigma \overset{\text{ind}}{\sim} \text{Beta}(1 - \sigma, \alpha + k\sigma), \quad k \in \mathbb{N}^\star, \\
&\pi_k(\boldsymbol{\tau}) = \tau_k \prod_{l=1}^{k-1} (1 - \tau_l), \quad k \in \mathbb{N}^\star, \\
&G = \sum_{k=1}^{\infty} \pi_k(\boldsymbol{\tau}) \delta_{\boldsymbol{\theta}_k^0} \sim \text{PYP}(\alpha, \sigma, G_0).
\end{aligned}$$

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 (Ishwaran and James, 2001; Pitman and Yor, 1997). The PYP reduces to a DP when $\sigma = 0$. More general stick-breaking representations are possible, *e.g.,* Gibbs-type priors (De Blasi et al., 2015; Ghosal and Van der Vaart, 2017) or homogeneous normalised random measures with independent increments (Favaro et al., 2016). 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 Blei and Jordan (2006) 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 Lü et al. (2020), we use the following prior that satisfies the constraints $\sigma \in (0, 1)$ and $\alpha > -\sigma$,

$$p(\alpha, \sigma \mid s_1, s_2, a) = p(\alpha \mid \sigma; s_1, s_2)\, p(\sigma \mid a),$$

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 $\mathcal{SG}(\alpha \mid \sigma; s_1, s_2) = \mathrm{Gam}(\alpha + \sigma \mid s_1, s_2)$.

Hierarchical representation of BNP-MM. BNP-MMs, including DP-MMs and PYP-MMs, see, *e.g.,* Lü et al. (2020); Durand et al. (2022), have the following hierarchical representation to generate a data point $\mathbf{x}_n$ as a special case of BNP-GLLiMs:

1. BNP prior:

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

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

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

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

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

Appendix B. Variational Bayesian expectation-maximisation principle

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

The expectation-maximisation (EM) algorithm (Dempster et al., 1977; McLachlan and Krishnan, 1997) 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 Neal and Hinton (1998), EM can be viewed as an alternating maximisation procedure of a function $\mathcal{F}_0$ defined, for any probability distribution $q_{\mathcal{Z}}$ over labels $\mathcal{Z}$, by

$$\begin{aligned}\mathcal{F}_0(q_{\mathcal{Z}}, \boldsymbol{\Theta}, \boldsymbol{\phi}) &= \sum_{\mathcal{Z}} q_{\mathcal{Z}}(\mathcal{Z}) \log p(\mathcal{Y}, \mathcal{X}, \mathcal{Z} \mid \boldsymbol{\Theta}, \boldsymbol{\phi}) - \mathbb{E}_{q_{\mathcal{Z}}}[\log q_{\mathcal{Z}}(\mathcal{Z})] \\ &= \mathbb{E}_{q_{\mathcal{Z}}}\left[\log \frac{p(\mathcal{Y}, \mathcal{X}, \mathcal{Z} \mid \boldsymbol{\Theta}, \boldsymbol{\phi})}{q_{\mathcal{Z}}(\mathcal{Z})}\right],\end{aligned}$$

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 consists 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 $\widehat{\boldsymbol{\Theta}} = \mathrm{argmax}_{\boldsymbol{\Theta} \in \boldsymbol{\Theta}} p(\boldsymbol{\Theta} \mid \mathcal{Y}, \mathcal{X})$. 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

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

This integration requires the computation of the density $p(\boldsymbol{\Theta} \mid \mathcal{Y}, \mathcal{X})$, 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, as introduced by Beal and Ghahramani (2003).

To deal with the BNP-GLLiM model, we need to use the VBEM with hyperparameter optimisation of Beal (2003, 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. 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

$$\begin{aligned} \mathcal{F}(q_{\mathcal{Z}}, q_{\boldsymbol{\Theta}}, \boldsymbol{\phi}) &= \mathbb{E}_{q_{\mathcal{Z}} q_{\boldsymbol{\Theta}}} \left[\log \frac{p(\mathcal{Y}, \mathcal{X}, \mathcal{Z}, \boldsymbol{\Theta}; \boldsymbol{\phi})}{q_{\mathcal{Z}}(\mathcal{Z}) q_{\boldsymbol{\Theta}}(\boldsymbol{\Theta})} \right] \\ &= \log p(\mathcal{Y}, \mathcal{X} \mid \boldsymbol{\phi}) - \mathrm{KL}\left(q_{\mathcal{Z}} q_{\boldsymbol{\Theta}} \| p(\mathcal{Z}, \boldsymbol{\Theta} \mid \mathcal{Y}, \mathcal{X}, \boldsymbol{\phi})\right) \leq \log p(\mathcal{Y}, \mathcal{X} \mid \boldsymbol{\phi}), \end{aligned} \tag{B1}$$

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 maximisation 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 maximisation over $q_{\mathcal{Z}}$ and $q_{\boldsymbol{\Theta}}$ leads to the following E-steps, see, *e.g.,* Chaari et al. (2013, Appendix A), Beal (2003, Theorem 2.1) and Bishop (2006, 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 $\mathcal{F}(q_{\mathcal{Z}}, q_{\boldsymbol{\Theta}}, \boldsymbol{\phi})$. At the rth iteration, using current values $\boldsymbol{\phi}^{(r-1)}$ and $q_{\boldsymbol{\Theta}}^{(r-1)}$, we get the following updating,

$$\textbf{VB-E-}\mathcal{Z}\textbf{:}\ q_{\mathcal{Z}}^{(r)}(\mathcal{Z}) \propto \exp \mathbb{E}_{q_{\boldsymbol{\Theta}}^{(r-1)}} \left[\log p(\mathcal{Y}, \mathcal{X}, \mathcal{Z}, \boldsymbol{\Theta}; \boldsymbol{\phi}^{(r-1)}) \right], \tag{B2}$$

$$\textbf{VB-E-}\boldsymbol{\Theta}\textbf{:}\ q_{\boldsymbol{\Theta}}^{(r)}(\boldsymbol{\Theta}) \propto \exp \mathbb{E}_{q_{\mathcal{Z}}^{(r)}} \left[\log p(\mathcal{Y}, \mathcal{X}, \mathcal{Z}, \boldsymbol{\Theta}; \boldsymbol{\phi}^{(r-1)}) \right], \tag{B3}$$

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

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 F.1 for the BNP-GLLiM inference.

Appendix C. Details of VBEM for BNP-GLLiM model

C.1. VB-E-τ step from Section 3.1

To achieve results from Section 3.1, we make use of (5), (6), (F1), and are only interested in the functional dependence of the right-hand side of (B3) on the variable τ_k. Thus, any terms that do not depend on τ_k can be absorbed into the additive normalization constant, giving

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

Here,

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

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

Furthermore, we used the fact that

$$\log p(\tau_k \mid \alpha,\sigma) = \log\left[\text{Beta}(\tau_k \mid 1-\sigma, \alpha+k\sigma)\right] = \log\left[\frac{\Gamma(1-\sigma+k\sigma)}{\Gamma(1-\sigma)\Gamma(\alpha+k\sigma)}\tau_k^{1-\sigma-1}(1-\tau_k)^{\alpha+k\sigma-1}\right],$$

$$\log \pi_{z_n}(\boldsymbol{\tau}) = \log\left[\tau_{z_n}\prod_{l=1}^{z_n-1}(1-\tau_l)\right] = \log\tau_{z_n} + \sum_{l=1}^{z_n-1}\log(1-\tau_l),$$

$$q_{\tau_{\setminus\{k\}}}(\tau_l) = \prod_{i=1, i\neq k}^{K-1} q_{\tau_i}(\tau_l), \quad d\tau_{\setminus\{k\}} = \prod_{i=1, i\neq k}^{K-1} d\tau_i.$$

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

$$\begin{aligned}
q_{\tau_k}(\tau_k) &= \text{Beta}(\tau_k \mid \widehat{\gamma}_{k,1}, \widehat{\gamma}_{k,2}), \\
\widehat{\gamma}_{k,1} &= 1 - \mathbb{E}_{q_{\alpha,\sigma}}[\sigma] + N_k, \widehat{\gamma}_{k,2} = \mathbb{E}_{q_{\alpha,\sigma}}[\alpha] + k\mathbb{E}_{q_{\alpha,\sigma}}[\sigma] + \sum_{l=k+1}^{K} N_l.
\end{aligned}$$

C.2. VB-E-(α, σ) step from Section 3.1

In the PY case, to achieve results from Section 3.1, we make use of (5), (4), (F1), (F2), and are only interested in the functional dependence of the right-hand side of (B3) 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

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

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 F.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 the prior

on (α, σ) given in (4), it holds that

$$
\begin{aligned}
&\widetilde{q}_{\alpha,\sigma}(\alpha,\sigma) \\
&= \frac{1}{\Gamma(s_1)} {s_2}^{s_1} (\alpha+\sigma)^{s_1-1} \exp\{-s_2(\alpha+\sigma)\} \frac{\Gamma(\alpha)p(\sigma \mid a)}{\Gamma(1-\sigma)^{K-1}\Gamma(\alpha+(K-1)\sigma)} \prod_{k=1}^{K-1} [\alpha+(k-1)\sigma] \\
&\quad \times \exp\left\{-\sigma\left[\sum_{k=1}^{K-1} \mathbb{E}_{q_{\tau_k}}[\log\tau_k] - \sum_{k=1}^{K-1} k\mathbb{E}_{q_{\tau_k}}[\log(1-\tau_k)]\right] + (\alpha-1)\sum_{k=1}^{K-1}\mathbb{E}_{q_{\tau_k}}[\log(1-\tau_k)]\right\} \\
&= \frac{1}{\Gamma(s_1)} {s_2}^{s_1} (\alpha+\sigma)^{s_1-1} \exp\left\{-\left[s_2 - \sum_{k=1}^{K-1}\mathbb{E}_{q_{\tau_k}}[\log(1-\tau_k)]\right](\alpha+\sigma)\right\} \\
&\quad \times e^{-\sigma\xi} \exp\left\{-\sum_{k=1}^{K-1}\mathbb{E}_{q_{\tau_k}}[\log(1-\tau_k)]\right\} \frac{\Gamma(\alpha)p(\sigma \mid a)}{\Gamma(1-\sigma)^{K-1}\Gamma(\alpha+(K-1)\sigma)} \prod_{k=1}^{K-1} [\alpha+(k-1)\sigma] \\
&= \frac{1}{\Gamma(s_1)} (\alpha+\sigma)^{s_1-1} \exp\left\{-\left[s_2 - \sum_{k=1}^{K-1}\mathbb{E}_{q_{\tau_k}}[\log(1-\tau_k)]\right](\alpha+\sigma)\right\} \left[s_2 - \sum_{k=1}^{K-1}\mathbb{E}_{q_{\tau_k}}[\log(1-\tau_k)]\right]^{s_1-s_1} \\
&\quad \times e^{-\sigma\xi} \exp\left\{-\sum_{k=1}^{K-1}\mathbb{E}_{q_{\tau_k}}[\log(1-\tau_k)]\right\} \frac{\Gamma(\alpha)p(\sigma \mid a)}{\Gamma(1-\sigma)^{K-1}\Gamma(\alpha+(K-1)\sigma)} \prod_{k=1}^{K-1} [\alpha+(k-1)\sigma]\, {s_2}^{s_1} \\
&\propto \mathrm{Gam}(\alpha+\sigma \mid \widehat{s}_1, \widehat{s}_2)\, e^{-\sigma\xi} \frac{\Gamma(\alpha)p(\sigma \mid a)}{\Gamma(1-\sigma)^{K-1}\Gamma(\alpha+(K-1)\sigma)} \prod_{k=1}^{K-1} \frac{\alpha+(k-1)\sigma}{\alpha+\sigma}. \qquad \text{(C1)}
\end{aligned}
$$

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

$$
\begin{aligned}
\xi &= \sum_{k=1}^{K-1}\mathbb{E}_{q_{\tau_k}}[\log\tau_k] - \sum_{k=1}^{K-1}(k-1)\mathbb{E}_{q_{\tau_k}}[\log(1-\tau_k)], \\
\mathbb{E}_{q_{\tau_k}}[\log\tau_k] &= \psi(\widehat{\gamma}_{k,1}) - \psi(\widehat{\gamma}_{k,1}+\widehat{\gamma}_{k,2}), \\
\mathbb{E}_{q_{\tau_k}}[\log(1-\tau_k)] &= \psi(\widehat{\gamma}_{k,2}) - \psi(\widehat{\gamma}_{k,1}+\widehat{\gamma}_{k,2}), \\
\widehat{s}_1 = s_1 + K - 1, \quad \widehat{s}_2 &= s_2 - \sum_{k=1}^{K-1}[\psi(\widehat{\gamma}_{k,2}) - \psi(\widehat{\gamma}_{k,1}+\widehat{\gamma}_{k,2})].
\end{aligned}
$$

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,

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

The importance sampling scheme then consists of the following steps

- 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 $\gamma(\alpha'_m \mid \widehat{s}_1, \widehat{s}_2)$ and then subtracting σ_m from the result.
- Compute the importance weights $w_m = W(\alpha_m, \alpha_m)$.

- Approximate the means

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

Note that this complication is due to the PY. In the DP-GLLiM case, by substituting $\sigma = 0$ in (C1), the E step is much simpler, as it reduces to computing the approximate posterior expectation of α, namely,

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

C.3. 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 (7), in the form

$$p(\mathcal{X} \mid \mathcal{Z}, \boldsymbol{c}, \boldsymbol{\Gamma}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}_L(\mathbf{x}_n \mid \boldsymbol{c}_k, \boldsymbol{\Gamma}_k)^{z_{nk}}, \tag{C2}$$

$$p(\mathbf{x}_n \mid \boldsymbol{c}, \boldsymbol{\Gamma}) = \sum_{k=1}^{K} p_{z_n}(k) \mathcal{N}_L(\mathbf{x}_n \mid \boldsymbol{c}_k, \boldsymbol{\Gamma}_k), \quad p_{z_n}(k) \equiv p(z_n = k) = \pi_k(\boldsymbol{\tau}), \tag{C3}$$

$$p(\mathcal{Y} \mid \mathcal{X}, \mathcal{Z}; \mathbf{A}, \mathbf{b}, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}_D(\mathbf{y}_n \mid \mathbf{A}_k \mathbf{x}_n + \mathbf{b}_k, \boldsymbol{\Sigma}_k)^{z_{nk}}, \tag{C4}$$

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

The observations $\mathcal{X}$ and $\mathcal{Y}$ are therefore i.i.d. and generated from the same GMM (C3) and infinite GLLiM (C4), respectively. Similarly, (6) can be written down in the form

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

By using the decomposition (8), the representation (C2), (C4) 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_{\boldsymbol{\Theta}}}\left[\log p\left(\mathcal{Y} \mid \mathcal{X}, \mathcal{Z}; \widehat{\mathbf{A}}, \widehat{\mathbf{b}}, \widehat{\boldsymbol{\Sigma}}\right) p\left(\mathcal{X} \mid \mathcal{Z}; \widehat{\mathbf{c}}, \widehat{\boldsymbol{\Gamma}}\right) p\left(\mathcal{Z} \mid \boldsymbol{\tau}\right)\right] + \text{constant} \\
&\propto \mathbb{E}_{q_{\boldsymbol{\Theta}}}\left[\log p\left(\mathcal{Y} \mid \mathcal{X}, \mathcal{Z}; \widehat{\mathbf{A}}, \widehat{\mathbf{b}}, \widehat{\boldsymbol{\Sigma}}\right)\right] + \mathbb{E}_{q_{\boldsymbol{\Theta}}}\left[\log p\left(\mathcal{X} \mid \mathcal{Z}; \widehat{\mathbf{c}}, \widehat{\boldsymbol{\Gamma}}\right)\right] + \mathbb{E}_{q_{\boldsymbol{\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_{\boldsymbol{\Theta}}}\left[\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} z_{nk}\mathbb{E}_{q_{\boldsymbol{\Theta}}}\left[\log \mathcal{N}_L\left(\mathbf{x}_n \mid \widehat{\mathbf{c}}_k, \widehat{\boldsymbol{\Gamma}}_k\right)\right] \\
&\quad + \mathbb{E}_{q_{\boldsymbol{\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{\boldsymbol{\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{\boldsymbol{\Gamma}}_k\right) \\
&\quad + \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\mathbb{E}_{q_{\boldsymbol{\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}
$$

Here, we used the fact that

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

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

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

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:

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

C.3.1. Updating $\boldsymbol{\Sigma}_k$

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

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

Finally, setting to zero yields

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

C.4. VB-M-$(\mathbf{c}, \boldsymbol{\Gamma})$ *step from Section 3.2*

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

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

By definition, we have

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

We aim to solve the following optimisation

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

Similarly with Appendices C.3.1 and C.5.1, we obtain the following update:

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

C.5. VB-M-$(\mathbf{A}, \mathbf{b}, \mathbf{\Sigma})$ step from Section 3.2

By definition, we have

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

We aim to solve the following optimisation

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

C.5.1. 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

$$\begin{aligned}&\sum_{n=1}^{N} q_{z_n}(k)\boldsymbol{\Sigma}_k^{-1}\left(\mathbf{y}_n - \mathbf{A}_k\mathbf{x}_n - \mathbf{b}_k\right) = 0 \\ &\Leftrightarrow \sum_{n=1}^{N} q_{z_n}(k)\boldsymbol{\Sigma}_k^{-1}\left(\mathbf{y}_n - \mathbf{A}_k\mathbf{x}_n\right) - \sum_{n=1}^{N} q_{z_n}(k)\boldsymbol{\Sigma}_k^{-1}\mathbf{b}_k = 0 \\ &\Leftrightarrow \sum_{n=1}^{N} q_{z_n}(k)\left(\mathbf{y}_n - \mathbf{A}_k\mathbf{x}_n\right) - \sum_{n=1}^{N} q_{z_n}(k)\mathbf{b}_k = 0 \,(\text{left multiplying by } \boldsymbol{\Sigma}_k) \\ &\Leftrightarrow \mathbf{b}_k = \frac{1}{\sum_{n=1}^{N} q_{z_n}(k)}\sum_{n=1}^{N} q_{z_n}(k)\left(\mathbf{y}_n - \mathbf{A}_k\mathbf{x}_n\right) \equiv \frac{1}{N_k}\sum_{n=1}^{N} q_{z_n}(k)\left(\mathbf{y}_n - \mathbf{A}_k\mathbf{x}_n\right). \qquad \text{(C5)}\end{aligned}$$

C.5.2. 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

$$
\begin{aligned}
&\sum_{n=1}^{N} q_{z_n}(k)\boldsymbol{\Sigma}_k^{-1}\left(\mathbf{y}_n - \mathbf{A}_k\mathbf{x}_n - \mathbf{b}_k\right)\mathbf{x}_n^\top = 0 \\
&\Leftrightarrow \sum_{n=1}^{N} q_{z_n}(k)\left(\mathbf{y}_n - \mathbf{A}_k\mathbf{x}_n - \mathbf{b}_k\right)\mathbf{x}_n^\top = 0 \,(\text{left multiplying by } \boldsymbol{\Sigma}_k) \\
&\Leftrightarrow \sum_{n=1}^{N} q_{z_n}(k)\mathbf{y}_n\mathbf{x}_n^\top - \sum_{n=1}^{N} q_{z_n}(k)\mathbf{A}_k\mathbf{x}_n\mathbf{x}_n^\top - \sum_{n=1}^{N} q_{z_n}(k)\mathbf{b}_k\mathbf{x}_n^\top = 0 \\
&\Leftrightarrow \sum_{n=1}^{N} q_{z_n}(k)\mathbf{y}_n\mathbf{x}_n^\top - \mathbf{A}_k\sum_{n=1}^{N} q_{z_n}(k)\mathbf{x}_n\mathbf{x}_n^\top - \mathbf{b}_k\sum_{n=1}^{N} q_{z_n}(k)\mathbf{x}_n^\top = 0 \\
&\Leftrightarrow \mathbf{A}_k\sum_{n=1}^{N} q_{z_n}(k)\mathbf{x}_n\mathbf{x}_n^\top \\
&\quad = \sum_{n=1}^{N} q_{z_n}(k)\mathbf{y}_n\mathbf{x}_n^\top - \frac{1}{N_k}\sum_{n=1}^{N} q_{z_n}(k)\left(\mathbf{y}_n - \mathbf{A}_k\mathbf{x}_n\right)\sum_{n=1}^{N} q_{z_n}(k)\mathbf{x}_n^\top \,(\text{using (C5) for } \mathbf{b}_k) \\
&\Leftrightarrow \mathbf{A}_k\sum_{n=1}^{N} q_{z_n}(k)\mathbf{x}_n\left(\mathbf{x}_n^\top - \frac{1}{N_k}\sum_{n=1}^{N} q_{z_n}(k)\mathbf{x}_n^\top\right) = \sum_{n=1}^{N} q_{z_n}(k)\left(\mathbf{y}_n - \frac{1}{N_k}\sum_{n=1}^{N} q_{z_n}(k)\mathbf{y}_n\right)\mathbf{x}_n^\top \\
&\Leftrightarrow N_k\mathbf{A}_k\boldsymbol{X}_k\boldsymbol{X}_k^\top = N_k\boldsymbol{Y}_k\boldsymbol{X}_k^\top \Leftrightarrow \mathbf{A}_k = \boldsymbol{Y}_k\boldsymbol{X}_k^\top\left(\boldsymbol{X}_k\boldsymbol{X}_k^\top\right)^{-1}.
\end{aligned}
$$

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

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

and

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

Here, we also use the equalities

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

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

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

Appendix D. Details on the ELBO

In this section, we provide the closed-form expressions for the ELBO stated in Proposition 3.1. Let us recal that when $\sigma = 0$, the ELBO in the BNP-GLLiM is derived as follows:

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

The terms of the right-hand side of the above equation have the following closed-form expressions:

$$\mathbb{E}\left[\log p(\mathcal{Y} \mid \mathcal{X}, \mathcal{Z}, \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] = \sum_{n=1}^{N}\sum_{k=1}^{K} q_{z_n}(k) \log \mathcal{N}_D\left(\mathbf{y}_n \mid \widehat{\mathbf{A}}_k \mathbf{x}_n + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right) \tag{D1}$$

$$\mathbb{E}\left[\log p(\mathcal{X} \mid \mathcal{Z}, \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] = \sum_{n=1}^{N}\sum_{k=1}^{K} q_{z_n}(k) \log \mathcal{N}_L\left(\mathbf{x}_n \mid \widehat{\boldsymbol{c}}_k, \widehat{\boldsymbol{\Gamma}}_k\right), \tag{D2}$$

$$\mathbb{E}\left[\log p(\mathcal{Z} \mid \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] = \sum_{k=1}^{K} N_k\left[\psi\left(\widehat{\gamma}_{k,1}\right) - \psi\left(\widehat{\gamma}_{k,1} + \widehat{\gamma}_{k,2}\right) + \sum_{l=1}^{k-1}\left[\psi\left(\widehat{\gamma}_{l,2}\right) - \psi\left(\widehat{\gamma}_{l,1} + \widehat{\gamma}_{l,2}\right)\right]\right], \tag{D3}$$

$$\mathbb{E}\left[\log p(\boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] = \sum_{k=1}^{K-1} \mathbb{E}\left[\log p\left(\tau_k \mid \alpha\right)\right] + \mathbb{E}\left[\log p\left(\alpha \mid \widehat{s}_1, \widehat{s}_2\right)\right], \tag{D4}$$

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

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

$$\mathbb{E}\left[\log q(\mathcal{Z})\right] = \sum_{n=1}^{N}\sum_{k=1}^{K} q_{z_n}(k)\log q_{z_n}(k)\,, \tag{D5}$$

$$\mathbb{E}\left[\log q(\boldsymbol{\Theta})\right] = \mathbb{E}\left[\log q_{\alpha,0}(\alpha)\right] + \sum_{k=1}^{K-1}\mathbb{E}\left[\log q_{\tau_k}(\tau_k)\right], \tag{D6}$$

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

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

Appendix E. Technical proofs

E.1. Proof of Lemma 5.2

We first want to prove (18). 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

$$p(\mathbf{x}_a \mid \mathbf{x}_b) = \mathcal{N}\left(\mathbf{x}_a \mid \boldsymbol{\mu}_{a|b}, \boldsymbol{\Gamma}_{aa}^{-1}\right), \quad \boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a - \boldsymbol{\Gamma}_{aa}^{-1}\boldsymbol{\Gamma}_{ab}(\boldsymbol{x}_b - \boldsymbol{\mu}_b) = \boldsymbol{\mu}_a - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\boldsymbol{x}_b - \boldsymbol{\mu}_b),$$
$$p(\boldsymbol{x}_b) = \mathcal{N}(\boldsymbol{x}_b \mid \boldsymbol{\mu}_b, \boldsymbol{\Sigma}_{bb}). \tag{E1}$$

Recall that

$$p(\mathbf{y} \mid \mathbf{x}, Z = k; \boldsymbol{\psi}) = \mathcal{N}_D(\mathbf{y} \mid \mathbf{A}_k\mathbf{x} + \mathbf{b}_k, \boldsymbol{\Sigma}_k),$$
$$p(\mathbf{x} \mid Z = k; \boldsymbol{\psi}) = \mathcal{N}_L(\mathbf{x} \mid \mathbf{c}_k, \boldsymbol{\Gamma}_k), p(Z = k; \boldsymbol{\psi}) = \pi_k. \tag{E2}$$

By identifying the parameters of (E1) and (E2), it holds that

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

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

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

By using result for the joint Gaussian, see *e.g.,* (E6), we obtain the desired result (19).

Finally, Lemma 5.2 is proved via using the following two statements (Deleforge et al., 2015, 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 (18) 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 (19) holds.

E.2. Proof of Proposition 3.1

Using the sum and product rules for both discrete and continuous variables, the ELBO in BNP-GLLiM (B1) is given by

$$\begin{aligned}
\mathcal{F}\left(q_{\mathcal{Z}}, q_{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\phi}}\right) &= \mathbb{E}_{q_{\mathcal{Z}}q_{\boldsymbol{\Theta}}}\left[\log \frac{p(\mathcal{Y},\mathcal{X},\mathcal{Z},\boldsymbol{\Theta};\widehat{\boldsymbol{\phi}})}{q(\mathcal{Z})q(\boldsymbol{\Theta})}\right] \equiv \mathbb{E}\left[\log \frac{p(\mathcal{Y},\mathcal{X},\mathcal{Z},\boldsymbol{\Theta};\widehat{\boldsymbol{\phi}})}{q(\mathcal{Z})q(\boldsymbol{\Theta})}\right] \\
&= \sum_{\mathcal{Z}} \int\int\int q(\mathcal{Z})q(\boldsymbol{\Theta}) \log\left[\frac{p(\mathcal{Y},\mathcal{X},\mathcal{Z},\boldsymbol{\Theta};\widehat{\boldsymbol{\phi}})}{q(\mathcal{Z})q(\boldsymbol{\Theta})}\right] d\mathcal{Z}d\boldsymbol{\Theta} \\
&= \mathbb{E}\left[\log p(\mathcal{Y} \mid \mathcal{X},\mathcal{Z},\boldsymbol{\Theta};\widehat{\boldsymbol{\phi}})\right] + \mathbb{E}\left[\log p(\mathcal{X} \mid \mathcal{Z},\boldsymbol{\Theta};\widehat{\boldsymbol{\phi}})\right] + \mathbb{E}\left[\log p(\mathcal{Z} \mid \boldsymbol{\Theta};\widehat{\boldsymbol{\phi}})\right] \\
&\quad + \mathbb{E}\left[\log p(\boldsymbol{\Theta};\widehat{\boldsymbol{\phi}})\right] - \mathbb{E}\left[\log q(\mathcal{Z})\right] - \mathbb{E}\left[\log q(\boldsymbol{\Theta})\right]. \qquad \text{(E3)}
\end{aligned}$$

Next, we evaluate the various terms in the ELBO (E3).

Proof of (D1)

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

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

where

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

Proof of (D2)

Similarly to the previous proof, we obtain

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

where

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

Proof of (D3)

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

$$\begin{aligned}
\mathbb{E}_{q_{\tau_k}}\left[\log\left(\tau_k\right)\right] &= \psi\left(\widehat{\gamma}_{k,1}\right) - \psi\left(\widehat{\gamma}_{k,1} + \widehat{\gamma}_{k,2}\right), \\
\mathbb{E}_{q_{\tau_k}}\left[\log\left(1 - \tau_k\right)\right] &= \psi\left(\widehat{\gamma}_{k,2}\right) - \psi\left(\widehat{\gamma}_{k,1} + \widehat{\gamma}_{k,2}\right).
\end{aligned} \tag{E4}$$

Via (E4), it holds that

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

Proof of (D4)

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, for σ, we use a uniform prior $\mathcal{U}_{[0,1]}(\sigma)$ so that parameter a does not have to be taken into account. Then it holds that

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

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 $\mathbb{E}_{q_{\alpha,\sigma}}\left[\sigma\right]$, $\mathbb{E}_{q_{\alpha,\sigma}}\left[\alpha+k\sigma-1\right]$, $\mathbb{E}_{q_{\alpha,\sigma}}\left[\log C(\alpha,\sigma)\right]$, $\mathbb{E}_{q_{\alpha,\sigma}}\left[\log\left(\alpha+\sigma\right)\right]$ and $\mathbb{E}_{q_{\alpha,\sigma}}\left[\alpha+\sigma\right]$. 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 $\mathrm{Gam}\left(\alpha \mid \widehat{s}_1, \widehat{s}_2\right)$ 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:

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

Proof of (D5)

Due to the mean-field approximation (9) and truncation, this step is analytically computed as follows:

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

Proof of (D6)

We have

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

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

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

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

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

E.3. Proof of Theorem 4.1

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

$$\begin{aligned}p(\widehat{\mathbf{y}}, \widehat{\mathbf{x}}, \mathcal{X}, \mathcal{Y}) &= \sum_{\widehat{\mathbf{z}}} \int p(\widehat{\mathbf{y}} \mid \widehat{\mathbf{x}}, \widehat{\mathbf{z}}, \boldsymbol{\Theta}, \mathcal{X}, \mathcal{Y})\, p(\widehat{\mathbf{x}} \mid \widehat{\mathbf{z}}, \boldsymbol{\Theta}, \mathcal{X}, \mathcal{Y})\, p(\widehat{\mathbf{z}} \mid \boldsymbol{\Theta}, \mathcal{X}, \mathcal{Y})\, p(\boldsymbol{\Theta} \mid \mathcal{X}, \mathcal{Y})\, d\boldsymbol{\Theta} \\ &= \sum_{\widehat{\mathbf{z}}} \int p(\widehat{\mathbf{y}} \mid \widehat{\mathbf{x}}, \widehat{\mathbf{z}}; \boldsymbol{A}, \boldsymbol{b}, \boldsymbol{\Sigma})\, p(\widehat{\mathbf{x}} \mid \widehat{\mathbf{z}}, \boldsymbol{c}, \boldsymbol{\Gamma})\, p(\widehat{\mathbf{z}} \mid \boldsymbol{\tau}; \beta)\, p(\boldsymbol{\Theta} \mid \mathcal{X}, \mathcal{Y})\, d\boldsymbol{\Theta} \equiv D_1.\end{aligned} \tag{E5}$$

Note that in (E5), $p(\boldsymbol{\Theta} \mid \mathcal{X}, \mathcal{Y})$ 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 $p(\boldsymbol{\Theta} \mid \mathcal{X}, \mathcal{Y})$ with its truncated variational posterior of parameters $\boldsymbol{\Theta}$ given by

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

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 (Blei and Jordan, 2006). 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 $q_{\boldsymbol{\tau}}(\boldsymbol{\tau}) = \prod_{k=1}^{K-1} q_{\tau_k}(\tau_k)$ 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

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

Indeed, we made use of the following result for the joint Gaussian, see, *e.g.,* Bishop (2006, 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

$$
\begin{aligned}
p(\mathbf{x}) &= \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Gamma}^{-1}\right), \\
p(\mathbf{y} \mid \mathbf{x}) &= \mathcal{N}\left(\mathbf{y} \mid \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}\right),
\end{aligned}
$$

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

$$
\begin{aligned}
p(\mathbf{w}) &= \mathcal{N}\left(\mathbf{w} \mid \mathbb{E}\left[\mathbf{w}\right], \operatorname{cov}\left[\mathbf{w}\right]\right), \text{ where} \\
\operatorname{cov}\left[\mathbf{w}\right] &= \begin{pmatrix} \boldsymbol{\Gamma}^{-1} & \boldsymbol{\Gamma}^{-1}\mathbf{A}^{\top} \\ \mathbf{A}\boldsymbol{\Gamma}^{-1} & \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Gamma}^{-1}\mathbf{A}^{\top} \end{pmatrix}, \quad \mathbb{E}\left[\mathbf{w}\right] = \begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{A}\boldsymbol{\mu} + \mathbf{b} \end{pmatrix}.
\end{aligned} \tag{E6}
$$

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

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

Next, we aim to prove that

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

Indeed, recall that we have defined

$$\begin{aligned}
&\tau_k \mid \alpha, \sigma \overset{\text{ind}}{\sim} \text{Beta}\left(\tau_k \mid 1-\sigma, \alpha+k\sigma\right), \quad k \in \mathbb{N}^\star, \\
&\pi_k(\boldsymbol{\tau}) = \tau_k \prod_{l=1}^{k-1}\left(1-\tau_l\right), \quad k \in \mathbb{N}^\star, \\
&p\left(\mathcal{Z} \mid \boldsymbol{\tau}\right) \propto \prod_{n=1}^{N} \pi_{z_n}\left(\boldsymbol{\tau}\right),
\end{aligned}$$

and to deal with the infinite state space for each z_j, we considered a truncation of the state space to a maximum label $K \equiv K_{\max}, K_{\max} \in \mathbb{N}^\star$ (Blei and Jordan, 2006). 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 Ghosal and Van der Vaart (2017, 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.*,

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

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 i.i.d., *e.g.*, when $\sigma = 0$, it suffices that $p\left(\tau_1 > 0\right) > 0$. Then

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

E.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

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

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

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

Therefore, we obtain

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

which is a mixture of Gaussian experts since we have

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

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

E.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

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

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.,* Bishop (2006, 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

$$\begin{aligned} p(\mathbf{x}) &= \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Gamma}^{-1}\right), \\ p(\mathbf{y} \mid \mathbf{x}) &= \mathcal{N}\left(\mathbf{y} \mid \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}\right), \end{aligned}$$

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

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

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

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

where

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

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

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

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_{Dk}\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_{Dk}\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_{Dk}\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_{Dk}\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_{Dk}\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_{Dk}\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

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

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

F.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)

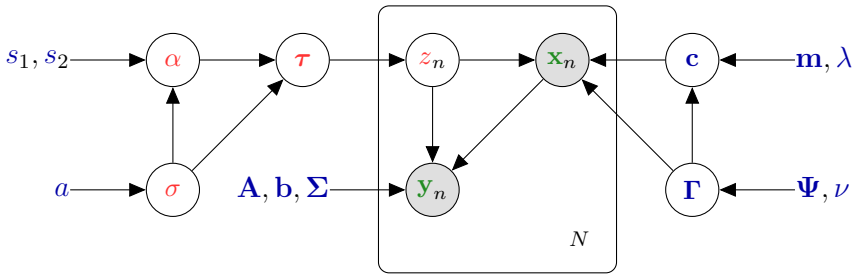


Figure F1. Graphical representation of BNP-GLLiM2: the plate denotes N i.i.d. observations, white-filled circles correspond to unobserved (latent) variables and random or unknown parameters represented in red, while grey-filled circles correspond to observed variables represented in green. Hyperparameters are represented in blue.

distribution parameterized by $\boldsymbol{\rho}_k = (\mathbf{m}_k, \lambda_k, \boldsymbol{\Psi}_k, \nu_k)$ with a PDF

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

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 F1. The joint distribution of the observed data $\mathcal{X}, \mathcal{Y}$ and all latent variables can be expressed hierarchically as

$$\begin{aligned} p(\mathcal{Y}, \mathcal{X}, \mathcal{Z}, \boldsymbol{\Theta}; \boldsymbol{\phi}) &= \prod_{n=1}^{N} p\left(\mathbf{y}_n \mid \mathbf{x}_n, z_n, \boldsymbol{\Theta}; \boldsymbol{\phi}\right) p\left(\mathbf{x}_n \mid z_n, \boldsymbol{\Theta}; \boldsymbol{\phi}\right) p(\mathcal{Z} \mid \boldsymbol{\Theta}; \boldsymbol{\phi}) p\left(\boldsymbol{\Theta}; \boldsymbol{\phi}\right) \\ &= \prod_{n=1}^{N} p\left(\mathbf{y}_n \mid \mathbf{x}_n, z_n; \mathbf{A}, \mathbf{b}, \boldsymbol{\Sigma}\right) p\left(\mathbf{x}_n \mid z_n, \mathbf{c}, \boldsymbol{\Gamma}\right) p(\mathcal{Z} \mid \boldsymbol{\tau}) \\ &\quad \prod_{k\in\mathbb{N}^\star} p\left(\tau_k \mid \alpha, \sigma\right) p\left(\alpha, \sigma \mid s_1, s_2, a\right) \prod_{k\in\mathbb{N}^\star} p\left(\mathbf{c}_k, \boldsymbol{\Gamma}_k; \boldsymbol{\rho}_k\right). \end{aligned} \tag{F1}$$

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

$$q_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}) = q_{\alpha,\sigma}(\alpha, \sigma) \prod_{k=1}^{K-1} q_{\tau_k}\left(\tau_k\right) \prod_{k=1}^{K} q_{\boldsymbol{\theta}_k^*}\left(\boldsymbol{\theta}_k^*\right). \tag{F2}$$

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 C. 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 $q_{\mathcal{Z}}(\mathcal{Z})$.

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

By using the mean-field approximation (9) and the truncation, see Appendix C.3 for more details, for all $n \in [N]$ and for $k \in [K]$, this step consists of computing

$$q_{z_n}(k) = \frac{\rho_{nk}}{\sum_{l=1}^{K} \rho_{nl}}. \tag{F3}$$

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

$$\begin{aligned} &-\frac{1}{2}\Bigg\{\log\left|\widehat{\mathbf{\Sigma}}_k\right| + (\mathbf{y}_n - \widehat{\mathbf{A}}_k\mathbf{x}_n - \widehat{\mathbf{b}}_k)^\top \widehat{\mathbf{\Sigma}}_k^{-1}(\mathbf{y}_n - \widehat{\mathbf{A}}_k\mathbf{x}_n - \widehat{\mathbf{b}}_k) \\ &+ \log\left|\frac{\widehat{\mathbf{\Psi}}_k}{2}\right| - \sum_{l=1}^{L}\psi\left(\frac{\widehat{\nu}_k + (1-l)}{2}\right) + \widehat{\nu}_k(\mathbf{x}_n - \widehat{\mathbf{m}}_k)^\top \widehat{\mathbf{\Psi}}_k^{-1}(\mathbf{x}_n - \widehat{\mathbf{m}}_k) + \frac{L}{\widehat{\lambda}_k}\Bigg\} \\ &+ \psi(\widehat{\gamma}_{k,1}) - \psi(\widehat{\gamma}_{k,1} + \widehat{\gamma}_{k,2}) + \sum_{l=1}^{k-1}\left[\psi(\widehat{\gamma}_{l,2}) - \psi(\widehat{\gamma}_{l,1} + \widehat{\gamma}_{l,2})\right]. \end{aligned} \tag{F4}$$

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

Proof of (F4)**.** With respect to the VBEM for BNP-GLLiM2 model from Appendix F.1.1, it is almost similar to the previous step in Appendix C.3, except that we have to take into account the randomness of $\mathbf{c}$ and $\mathbf{\Gamma}$. Namely, we have

$$\begin{aligned} &q_{z_n}(z_n) \\ &\propto \exp \mathbb{E}_{q_{\mathbf{\Theta}}}\left[\log\left(p\left(\mathbf{y}_n \mid \mathbf{x}_n, z_n; \widehat{\mathbf{A}}_{z_n}, \widehat{\mathbf{b}}_{z_n}, \widehat{\mathbf{\Sigma}}_{z_n}\right) p\left(\mathbf{x}_n \mid z_n, \mathbf{c}_{z_n}, \mathbf{\Gamma}_{z_n}\right) p(\mathbf{z} \mid \boldsymbol{\tau})\right)\right] \\ &= \exp\left\{\log p\left(\mathbf{y}_n \mid \mathbf{x}_n, z_n; \widehat{\mathbf{A}}_{z_n}, \widehat{\mathbf{b}}_{z_n}, \widehat{\mathbf{\Sigma}}_{z_n}\right) + \mathbb{E}_{q_{\boldsymbol{\theta}^*_{z_n}}}\left[\log p\left(\mathbf{x}_n \mid z_n, \mathbf{c}_{z_n}, \mathbf{\Gamma}_{z_n}\right)\right] + \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\log \pi_{z_n}(\boldsymbol{\tau})\right]\right\}. \end{aligned} \tag{F5}$$

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

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

where we used the fact that

$$\begin{aligned} \mathbb{E}_{q_{\mathbf{\Gamma}_k}}\left[\log\left|\widehat{\mathbf{\Gamma}}_k\right|\right] &= \log\left|\frac{\widehat{\mathbf{\Psi}}_k}{2}\right| - \sum_{l=1}^{L}\psi\left(\frac{\widehat{\nu}_k + (1-l)}{2}\right), \\ \mathbb{E}_{q_{\boldsymbol{\theta}^*_{z_n}}}\left[(\mathbf{x}_n - \widehat{\mathbf{c}}_k)^\top \widehat{\mathbf{\Gamma}}_k^{-1}(\mathbf{x}_n - \widehat{\mathbf{c}}_k)\right] &= \widehat{\nu}_k(\mathbf{x}_n - \widehat{\mathbf{m}}_k)^\top \widehat{\mathbf{\Psi}}_k^{-1}(\mathbf{x}_n - \widehat{\mathbf{m}}_k) + \frac{L}{\widehat{\lambda}_k}. \end{aligned}$$

Plugging in all of the above expression back into (F5) yields the desired results in (F4). □

F.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 density defined as

$$q_{\boldsymbol{\theta}_k^*}(\mathbf{c}_k, \boldsymbol{\Gamma}_k) = \mathcal{NIW}(\mathbf{c}_k, \boldsymbol{\Gamma}_k \mid \widehat{\mathbf{m}}_k, \widehat{\lambda}_k, \widehat{\boldsymbol{\Psi}}_k, \widehat{\nu}_k). \tag{F6}$$

Here, the hyperparameters are updated as follows (see, *e.g.*, Bishop (2006, Section 10.2.1)):

$$\begin{aligned}
\widehat{\lambda}_k &= \lambda_k + N_k, \quad \widehat{\nu}_k = \nu_k + N_k, N_k = \sum_{n=1}^{N} q_{z_n}(k) \\
\widehat{\boldsymbol{\Psi}}_k &= \boldsymbol{\Psi}_k + N_k \boldsymbol{S}_k + \frac{\lambda_k N_k}{\lambda_k + N_k} (\mathbf{m}_k - \bar{\boldsymbol{c}}_k)(\mathbf{m}_k - \bar{\boldsymbol{c}}_k)^\top, \\
\widehat{\mathbf{m}}_k &= \frac{\lambda_k \mathbf{m}_k + N_k \bar{\boldsymbol{c}}_k}{\lambda_k + N_k} = \frac{\lambda_k \mathbf{m}_k + N_k \bar{\boldsymbol{c}}_k}{\widehat{\lambda}_k}, \\
\bar{\boldsymbol{c}}_k &= \frac{1}{N_k} \sum_{n=1}^{N} q_{z_n}(k) \mathbf{x}_n, \\
\boldsymbol{S}_k &= \frac{1}{N_k} \sum_{n=1}^{N} q_{z_n}(k) (\mathbf{x}_n - \bar{\boldsymbol{c}}_k)(\mathbf{x}_n - \bar{\boldsymbol{c}}_k)^\top.
\end{aligned} \tag{F7}$$

VB-M steps

The maximisation 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:

$$\boldsymbol{\phi}^{(r)} = \underset{\boldsymbol{\phi}}{\operatorname{argmax}} \; \mathbb{E}_{q_{\mathcal{Z}}^{(r)} q_{\boldsymbol{\tau}}^{(r)} q_{\alpha,\sigma}^{(r)} q_{\boldsymbol{\theta}^*}^{(r)}} \left[\log p\left(\mathcal{Y}, \mathcal{X}, \mathcal{Z}, \boldsymbol{\tau}, \alpha, \sigma, \boldsymbol{\theta}^*; \boldsymbol{\phi}\right)\right]. \tag{F8}$$

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.

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

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

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

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 (F7).

F.1.4. ELBO for BNP-GLLiM2

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

$$\mathcal{F}\left(q_{\mathcal{Z}}, q_{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\phi}}\right) = \mathbb{E}\left[\log p(\mathcal{Y} \mid \mathcal{X}, \mathcal{Z}, \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] + \mathbb{E}\left[\log p(\mathcal{X} \mid \mathcal{Z}, \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] + \mathbb{E}\left[\log p(\mathcal{Z} \mid \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] + \mathbb{E}\left[\log p(\boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] - \mathbb{E}\left[\log q(\mathcal{Z})\right] - \mathbb{E}\left[\log q(\boldsymbol{\Theta})\right]. \tag{F9}$$

Here, we have the following update formulas:

$$\mathbb{E}\left[\log p(\mathcal{Y} \mid \mathcal{X}, \mathcal{Z}, \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] = \sum_{n=1}^{N}\sum_{k=1}^{K} q_{z_n}(k) \log \mathcal{N}_D\left(\mathbf{y}_n \mid \widehat{\mathbf{A}}_k \mathbf{x}_n + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right), \tag{F10}$$

$$\mathbb{E}\left[\log p(\mathcal{X} \mid \mathcal{Z}, \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] = \frac{1}{2}\sum_{k=1}^{K} N_k \left[\log \widetilde{\boldsymbol{\Gamma}}_k - L\log(2\pi) - L\widehat{\lambda}_k^{-1} - \widehat{\nu}_k \mathit{Tr}\left(\mathbf{S}_k \widehat{\boldsymbol{\Psi}}_k^{-1}\right) - \widehat{\nu}_k \left(\bar{\mathbf{x}}_k - \widehat{\mathbf{m}}_k\right) \widehat{\boldsymbol{\Psi}}_k^{-1} \left(\bar{\mathbf{x}}_k - \widehat{\mathbf{m}}_k\right)\right], \tag{F11}$$

$$\mathbb{E}\left[\log p(\mathcal{Z} \mid \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] = \sum_{k=1}^{K} N_k \left[\psi\left(\widehat{\gamma}_{k,1}\right) - \psi\left(\widehat{\gamma}_{k,1} + \widehat{\gamma}_{k,2}\right) + \sum_{l=1}^{k-1}\left[\psi\left(\widehat{\gamma}_{l,2}\right) - \psi\left(\widehat{\gamma}_{l,1} + \widehat{\gamma}_{l,2}\right)\right]\right], \tag{F12}$$

$$\mathbb{E}\left[\log p(\boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] = \sum_{k=1}^{K-1} \mathbb{E}\left[\log p\left(\tau_k \mid \alpha\right)\right] + \mathbb{E}\left[\log p\left(\alpha \mid \widehat{s}_1, \widehat{s}_2\right)\right] + \sum_{k=1}^{K} \mathbb{E}\left[\log p\left(\mathbf{c}_k, \boldsymbol{\Gamma}_k; \widehat{\boldsymbol{\rho}}_k\right)\right], \tag{F13}$$

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

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

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

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

$$\mathbb{E}\left[\log q(\mathcal{Z})\right] = \sum_{n=1}^{N}\sum_{k=1}^{K} q_{z_n}\left(k\right)\log q_{z_n}\left(k\right), \tag{F14}$$

$$\mathbb{E}\left[\log q(\boldsymbol{\Theta})\right] = \mathbb{E}\left[\log q_{\alpha,0}(\alpha)\right] + \sum_{k=1}^{K-1}\mathbb{E}\left[\log q_{\tau_k}\left(\tau_k\right)\right] + \sum_{k=1}^{K}\mathbb{E}\left[\log q_{\mathbf{c}_k,\boldsymbol{\Gamma}_k}\left(\mathbf{c}_k, \boldsymbol{\Gamma}_k\right)\right], \tag{F15}$$

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

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

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

F.2. Predictive conditional density for BNP-GLLiM2

F.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 F.2, which is proved in Appendix F.4.

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

$$p\left(\widehat{\mathbf{y}}, \widehat{\mathbf{x}}, \mathcal{X}, \mathcal{Y}\right) \approx \sum_{k=1}^{K} \mathbb{E}_{q_{\boldsymbol{\tau}}}\left[\pi_k\left(\boldsymbol{\tau}\right)\right] St\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{m}}_k, \mathbf{L}_k, \widehat{\nu}_k + 1 - L\right) \mathcal{N}_D\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{A}}_k \widehat{\mathbf{x}} + \widehat{\mathbf{b}}_k, \widehat{\boldsymbol{\Sigma}}_k\right). \tag{F16}$$

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

$$\mathbf{L}_k = \frac{\left(\widehat{\nu}_k + 1 - L\right) \widehat{\lambda}_k}{1 + \widehat{\lambda}_k} \widehat{\boldsymbol{\Psi}}_k. \tag{F17}$$

F.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 F.3 with the proof in Appendix F.5.

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

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

Here, the gating posteriors are defined as

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

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

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

F.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 F.4, whose proof is provided in Appendix F.6.

Theorem F.4. *It holds that*

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

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

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

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

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.*

F.3. Proof of Proposition F.1

Using the sum and product rules for both discrete and continuous variables, the ELBO in BNP-GLLiM (B1) is given by

$$\begin{aligned}
\mathcal{F}\left(q_{\mathcal{Z}}, q_{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\phi}}\right) &= \mathbb{E}_{q_{\mathcal{Z}} q_{\boldsymbol{\Theta}}}\left[\log \frac{p(\mathcal{Y}, \mathcal{X}, \mathcal{Z}, \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})}{q(\mathcal{Z}) q_{\boldsymbol{\Theta}}(\boldsymbol{\Theta})}\right] \equiv \mathbb{E}\left[\log \frac{p(\mathcal{Y}, \mathcal{X}, \mathcal{Z}, \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})}{q(\mathcal{Z}) q(\boldsymbol{\Theta})}\right] \\
&= \sum_{\mathcal{Z}} \int\int\int q(\mathcal{Z}) q(\boldsymbol{\Theta}) \log\left[\frac{p(\mathcal{Y}, \mathcal{X}, \mathcal{Z}, \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})}{q(\mathcal{Z}) q(\boldsymbol{\Theta})}\right] d\mathcal{Z} d\boldsymbol{\Theta} \\
&= \mathbb{E}\left[\log p(\mathcal{Y} \mid \mathcal{X}, \mathcal{Z}, \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] + \mathbb{E}\left[\log p(\mathcal{X} \mid \mathcal{Z}, \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] \\
&\quad + \mathbb{E}\left[\log p(\mathcal{Z} \mid \boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] + \mathbb{E}\left[\log p(\boldsymbol{\Theta}; \widehat{\boldsymbol{\phi}})\right] - \mathbb{E}\left[\log q(\mathcal{Z})\right] - \mathbb{E}\left[\log q(\boldsymbol{\Theta})\right].
\end{aligned} \tag{F18}$$

Note that the proofs of (F10), (F12), (F14) are the same as in the proof of Proposition 3.1.

Proof of (F11)

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

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

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

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

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

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

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

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

(using (F21) from Lemma F.5). (F19)

To obtain (F19), we have to use the following Lemma F.5.

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

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

$$\mathbb{E}_{q_{\boldsymbol{c}_k, \boldsymbol{\Gamma}_k}}\left[\left(\mathbf{x}_n - \boldsymbol{c}_k\right)^\top \boldsymbol{\Gamma}_k^{-1}\left(\mathbf{x}_n - \boldsymbol{c}_k\right)\right] = L\widehat{\lambda}_k^{-1} + \widehat{\nu}_k\left(\mathbf{x}_n - \widehat{\mathbf{m}}_k\right)\widehat{\boldsymbol{\Psi}}_k^{-1}\left(\mathbf{x}_n - \widehat{\mathbf{m}}_k\right). \quad \text{(F20)}$$

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

$$\sum_{n=1}^{N} q_{z_n}(k)\left[\widehat{\nu}_k\left(\mathbf{x}_n - \widehat{\mathbf{m}}_k\right)\widehat{\boldsymbol{\Psi}}_k^{-1}\left(\mathbf{x}_n - \widehat{\mathbf{m}}_k\right)\right] = N_k\left[\widehat{\nu}_k Tr\left(\mathbf{S}_k\widehat{\boldsymbol{\Psi}}_k^{-1}\right) + \widehat{\nu}_k\left(\bar{\mathbf{x}}_k - \widehat{\mathbf{m}}_k\right)\widehat{\boldsymbol{\Psi}}_k^{-1}\left(\bar{\mathbf{x}}_k - \widehat{\mathbf{m}}_k\right)\right]. \quad \text{(F21)}$$

Proof of (F13)

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

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

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 F.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

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

Proof of (F15)

We have

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

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

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

F.4. Proof of Theorem F.2

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

$$
\begin{aligned}
p\left(\widehat{\mathbf{y}}, \widehat{\mathbf{x}}, \mathcal{X}, \mathcal{Y}\right) &= \sum_{\widehat{\mathbf{z}}} \int p\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{x}}, \widehat{\mathbf{z}}, \mathbf{\Theta}, \mathcal{X}, \mathcal{Y}\right) p\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{z}}, \mathbf{\Theta}, \mathcal{X}, \mathcal{Y}\right) p\left(\widehat{\mathbf{z}} \mid \mathbf{\Theta}, \mathcal{X}, \mathcal{Y}\right) p\left(\mathbf{\Theta} \mid \mathcal{X}, \mathcal{Y}\right) d\mathbf{\Theta} \\
&= \sum_{\widehat{\mathbf{z}}} \int p\left(\widehat{\mathbf{y}} \mid \widehat{\mathbf{x}}, \widehat{\mathbf{z}}; \boldsymbol{A}, \boldsymbol{b}, \mathbf{\Sigma}\right) p\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{z}}, \boldsymbol{c}, \mathbf{\Gamma}\right) p\left(\widehat{\mathbf{z}} \mid \boldsymbol{\tau}; \beta\right) p\left(\mathbf{\Theta} \mid \mathcal{X}, \mathcal{Y}\right) d\mathbf{\Theta} \equiv T_1.
\end{aligned}
\tag{F22}
$$

Note that in (F22), $p\left(\mathbf{\Theta} \mid \mathcal{X}, \mathcal{Y}\right)$ 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

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

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 (Blei and Jordan, 2006). 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} \\
&\quad \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 \\
&\quad \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_{\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}\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{m}}_k, \mathbf{L}_k, \widehat{\nu}_k + 1 - L\right) \text{(Lemma F.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).
\end{aligned}
$$

Here we used the following Lemma F.6

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

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

Proof of Lemma F.6

By definition, we obtain

$$
\begin{aligned}
&\int \int \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \boldsymbol{c}_k, \boldsymbol{\Gamma}_k^{-1}\right) q(\boldsymbol{\pi} \mid \mathcal{X}) q\left(\mathbf{c}_k, \boldsymbol{\Gamma}_k \mid \mathcal{X}\right) d\boldsymbol{c}_k d\boldsymbol{\Gamma}_k \\
&= \int \int \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \boldsymbol{c}_k, \boldsymbol{\Gamma}_k^{-1}\right) \mathcal{N}_L\left(\mathbf{c}_k \mid \widehat{\mathbf{m}}_k, \left(\widehat{\lambda}_k \boldsymbol{\Gamma}_k\right)^{-1}, \mathcal{X}\right) \mathcal{W}\left(\boldsymbol{\Gamma}_k \mid \widehat{\boldsymbol{\Psi}}_k, \widehat{\nu}_k, \mathcal{X}\right) d\boldsymbol{c}_k d\boldsymbol{\Gamma}_k \\
&= \int \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \boldsymbol{c}_k, \boldsymbol{\Gamma}_k^{-1}\right) \mathcal{N}_L\left(\mathbf{c}_k \mid \widehat{\mathbf{m}}_k, \left(\widehat{\lambda}_k \boldsymbol{\Gamma}_k\right)^{-1}, \mathcal{X}\right) \mathcal{W}\left(\boldsymbol{\Gamma}_k \mid \widehat{\boldsymbol{\Psi}}_k, \widehat{\nu}_k, \mathcal{X}\right) d\boldsymbol{c}_k d\boldsymbol{\Gamma}_k \\
&= \int \left[\int \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \boldsymbol{c}_k, \boldsymbol{\Gamma}_k^{-1}\right) \mathcal{N}_L\left(\mathbf{c}_k \mid \widehat{\mathbf{m}}_k, \left(\widehat{\lambda}_k \boldsymbol{\Gamma}_k\right)^{-1}, \mathcal{X}\right) d\boldsymbol{c}_k\right] \mathcal{W}\left(\boldsymbol{\Gamma}_k \mid \widehat{\boldsymbol{\Psi}}_k, \widehat{\nu}_k, \mathcal{X}\right) d\boldsymbol{\Gamma}_k \\
&= \int \mathcal{N}_L\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{m}}_k, \left(1 + \widehat{\lambda}_k^{-1}\right) \boldsymbol{\Gamma}_k^{-1}, \mathcal{X}\right) \mathcal{W}\left(\boldsymbol{\Gamma}_k \mid \widehat{\boldsymbol{\Psi}}_k, \widehat{\nu}_k, \mathcal{X}\right) d\boldsymbol{\Gamma}_k. \qquad \text{(F23)}
\end{aligned}
$$

When the size of the data set is large, *i.e.,* $N \to \infty$, this predictive distribution (F23) 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.,* Bishop (2006, 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

$$
\begin{aligned}
p(\mathbf{x}) &= \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Gamma}^{-1}\right), \\
p(\mathbf{y} \mid \mathbf{x}) &= \mathcal{N}\left(\mathbf{y} \mid \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}\right),
\end{aligned}
$$

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

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

where

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

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

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

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

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

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

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

Via 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)\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 \\
&= \frac{C\left(\widehat{\boldsymbol{\Psi}}_k, \widehat{\nu}_k, \widehat{\lambda}_k\right)}{B\left(\widehat{\boldsymbol{\Psi}}_k^*, \widehat{\nu}_k^*\right)} \underbrace{\int \mathcal{W}\left(\boldsymbol{\Gamma}_k \mid \boldsymbol{\Psi}_k^*, \widehat{\nu}_k^*\right) d\boldsymbol{\Gamma}_k}_{=1} = \frac{C\left(\widehat{\boldsymbol{\Psi}}_k, \widehat{\nu}_k, \widehat{\lambda}_k\right)}{B\left(\widehat{\boldsymbol{\Psi}}_k^*, \widehat{\nu}_k^*\right)} = \frac{B(\widehat{\boldsymbol{\Psi}}_k, \widehat{\nu}_k)}{\left(2\pi\left(1+\widehat{\lambda}_k^{-1}\right)\right)^{L/2}} \frac{1}{B\left(\widehat{\boldsymbol{\Psi}}_k^*, \widehat{\nu}_k^*\right)} \\
&= \frac{1}{\left(2\pi\left(1+\widehat{\lambda}_k^{-1}\right)\right)^{L/2}} \frac{\left|\widehat{\boldsymbol{\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{\boldsymbol{\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{\boldsymbol{\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{\boldsymbol{\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{\boldsymbol{\Psi}}_k\right|^{-\widehat{\nu}_k/2}}{\left(1+\widehat{\lambda}_k^{-1}\right)^{L/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} \\
&= \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{\boldsymbol{\Psi}}_k\right|^{1/2}}{\left(1+\widehat{\lambda}_k^{-1}\right)^{L/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} \\
&= \mathrm{St}\left(\widehat{\mathbf{x}} \mid \widehat{\mathbf{m}}_k, \mathbf{L}_k, \widehat{\nu}_k+1-L\right).
\end{aligned}
$$

Here,

$$
\mathbf{L}_k = \frac{(\widehat{\nu}_k+1-L)\,\widehat{\lambda}_k}{1+\widehat{\lambda}_k} \widehat{\boldsymbol{\Psi}}_k,
$$

and Δ^2 is the squared Mahalanobis distance defined by

$$
\Delta^2 = (\widehat{\mathbf{x}}-\widehat{\mathbf{m}}_k)^\top \mathbf{L}_k (\widehat{\mathbf{x}}-\widehat{\mathbf{m}}_k).
$$

Then, the last equality holds since we have

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

F.5. Proof of Theorem F.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

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

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

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

Therefore

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

which is a mixture of Gaussian experts since we have

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

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

F.6. Proof of Theorem F.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

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

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

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

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.*, Bishop (2006, 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

$$\begin{aligned}
p(\mathbf{x}) &= \mathcal{N} \left(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Gamma}^{-1} \right), \\
p(\mathbf{y} \mid \mathbf{x}) &= \mathcal{N} \left(\mathbf{y} \mid \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1} \right),
\end{aligned}$$

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

$$\begin{aligned}
p(\mathbf{y}) &= \int p(\mathbf{y} \mid \mathbf{x}) p(\mathbf{x}) d\mathbf{x} = \mathcal{N} \left(\mathbf{y} \mid \mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Gamma}^{-1}\mathbf{A}^{\top} \right), \\
p(\mathbf{x} \mid \mathbf{y}) &= \mathcal{N} \left(\mathbf{x} \mid \boldsymbol{\Sigma} \left\{ \mathbf{A}^{\top}\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Gamma}\boldsymbol{\mu} \right\}, \boldsymbol{\Sigma} \right), \boldsymbol{\Sigma} = \left(\boldsymbol{\Gamma} + \mathbf{A}^{\top}\mathbf{L}\mathbf{A} \right)^{-1}.
\end{aligned}$$

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

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

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.