# Nonparametric Bayesian Multi-Facet Clustering for Longitudinal Data

**Luwei Wang**[1] **Kieran Richards**[1] **Sohan Seth**[1]

[1]School of Informatics, University of Edinburgh, UK

## Abstract

Complex real-world time series data are inherently multi-faceted, e.g., temporal data can be described by seasonality and trend. Popular clustering methods typically aggregate information from all facets, treating them collectively rather than individually. This aggregation may diminish the interpretability of clusters by obscuring the specific contributions of individual facets to the clustering outcome. This limitation can be addressed by multi-facet clustering that builds a separate clustering model for each facet simultaneously. In this paper, we explore Bayesian multi-facet clustering modelling for temporal data using nonparametric priors to select an appropriate number of clusters automatically and using variational inference to efficiently explore the parameter space. We apply this framework to nonlinear growth models and vector autoregressive models and observe their performance through simulation studies. We apply these models to real-world time series data from the English Longitudinal Study of Ageing (ELSA), highlighting its utility in identifying meaningful and interpretable clusters. These findings underscore the potential of the framework for advancing the analysis of multi-faceted longitudinal data in diverse fields. Code is available at GitHub.

## 1 INTRODUCTION

Clustering, a key task in unsupervised machine learning, partitions unlabelled datasets into subgroups based on similarity measures [Murphy, 2012]. Classical algorithms such as $k$-means [MacQueen, 1967], hierarchical clustering [Hastie et al., 2009], Gaussian mixture models [Fraley and Raftery, 2002] and DBSCAN [Ester et al., 1996] are widely applied to uncover hidden data structures across various fields [Xu and Wunsch, 2005]. In the context of longitudinal data, clustering is crucial for exploring shared dynamics over time, with applications in speech processing, medical diagnosis and social sciences [Wilpon and Rabiner, 1985, Warren Liao, 2005, Bulteel et al., 2016, Marshall et al., 2024].

Existing clustering methods typically identify a single partition of the data by accumulating contributions from all facets (we refer to this as *single-facet clustering*). However, the rise of high-dimensional data and complex data structures in many clustering applications may reveal multiple interesting clustering structures when focusing on different characteristics or facets of the data. For instance, in images of objects, two interesting facets might be the shape and the color of the objects. Similarly, in temporal data, two interesting facets might be the seasonality and the trend of the data. Falck et al. [2021] argued that focusing on a single facet, rather than considering multiple facets, is an arbitrary and incomplete approach to clustering high-dimensional datasets. In practice, heterogeneous samples are often more effectively clustered based on a subset of characteristics, with other characteristics being uninformative or redundant [Kundu and Lukemire, 2024].

Standard clustering using combined information from all facets for complex data structures, such as time series, highlights the limitation in interpretability. For instance, Marshall et al. [2024] employed a mixture of nonlinear growth models to cluster individual income trajectories into several groups. Here the average income value and the variation of income over time both contribute to the clustering, and it is not clear which facet is driving the inferred clustering outcome more. Instead, it might be more effective to cluster each facet separately to find clusters with respect to both average income and variation of income simultaneously, and an individual can be assigned to both a specific average income cluster and a variation over time cluster. Similarly, the mixture of vector autoregressive models proposed by Bulteel et al. [2016] can be potentially more interpretable if the multivariate time series are clustered separately based on their average values and their temporal dynamics.

*Proceedings of the 41st Conference on Uncertainty in Artificial Intelligence* (UAI 2025), PMLR 244:4454–4485.

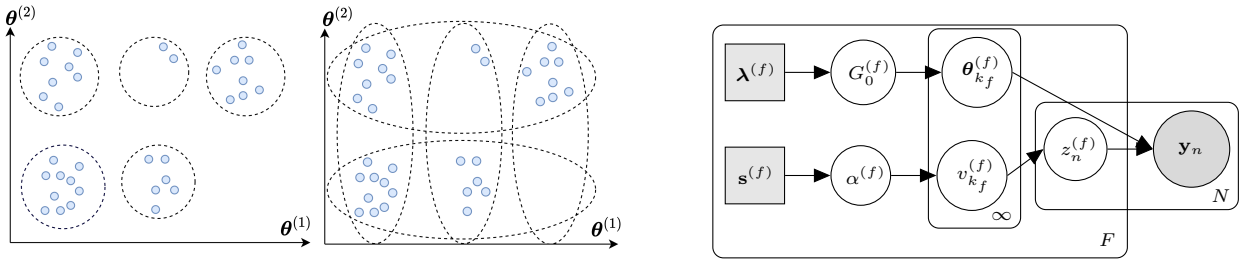



Figure 1: **left)** single-facet versus multi-facet clustering. **right)** nonparametric Bayesian multi-facet mixture model

Multi-facet clustering offers significant potential to address these challenges by constructing separate partitions for each facet. This approach not only ensures that items within specific clusters exhibit homogeneity but also facilitates the exploration of facets and their corresponding clusters, each representing distinct characteristics of the data [Falck et al., 2021]. Furthermore, this method requires a number of clusters that scale linearly with the number of facets, rather than exponentially. As a result, it reduces the total number of clusters required to represent high-dimensional data while explicitly capturing the unique characteristics driving the clustering process. For instance, consider the example illustrated in Figure 1 (left). Unlike single-facet clustering which identifies five clusters by considering both characteristics in aggregate, multi-facet clustering explicitly partitions the data across multiple facets. Specifically, it identifies $K_1 = 3$ clusters for Facet 1 and $K_2 = 2$ clusters for Facet 2, effectively summarizing six potential clusters into a more interpretable structure. It is important to emphasize that the multi-facet clustering is conceptually distinct from *multi-view/aspect clustering* approaches [Chao et al., 2021, Nayak and Luong, 2023]. Multi-view clustering aims to derive a single clustering solution that integrates information from multiple inputs (views) of the same sample cohort. In contrast, multi-facet clustering seeks to uncover multiple clustering solutions, each described by distinct characteristics/facets of a single input cohort.

**Related Work** The concept of multi-facet clustering aligns with the notion of learning *multiple clusterings* as highlighted by Gordon [1999]. Various methods have since been developed to address this issue by adapting conventional clustering approaches. For example, Friedman and Meulman [2004] proposed a distance-based clustering algorithm that automatically detects subgroups of objects clustering on different, possibly overlapping subsets of attribute variables. Galimberti and Soffritti [2007] introduced a two-step procedure with the first step identifying independent subsets of variables and the second step applying a model-based approach to identify cluster structures based on these subsets. A Bayesian method by Niu et al. [2012] introduced a probabilistic nonparametric Bayesian model to learn overlapping feature facets and clusters within each facet in a joint framework. Zong et al. [2024] proposed a similar model-based multi-facet clustering approach using a mixture of Gaussian mixture models, particularly suited for high-dimensional nonclusterable genes. These methods leverage feature selection techniques to identify relevant subsets of features as facets for clustering. Additionally, Falck et al. [2021] presented a deep learning approach extending the variational autoencoder (VAE), a feature-based method, to develop a multi-facet clustering algorithm. This model identifies facets by learning latent variables for each facet and simultaneously learns multiple clusterings in an end-to-end framework. Notably, all these models are tailored for *static feature data* and are not suitable for *temporal data*. A recent study by Kundu and Lukemire [2024] introduced a product Dirichlet process mixture model that employs Dirichlet process (DP) mixture priors on model parameters. Their approach primarily focused on applications to vector autoregressive models, relying on Markov chain Monte Carlo (MCMC) methods.

**Contributions** While nonparametric Bayesian approaches for multi-facet clustering using parameters of the model as facets have been explored in prior research [Kundu and Lukemire, 2024], they often rely on computationally intensive techniques such as MCMC sampling, which might limit their scalability and practical applicability for large datasets. **(A)** This study extends existing methods by incorporating a Variational Bayesian (VB) framework, enabling efficient and scalable inference while separately modelling key characteristics and identifying their corresponding clusters. **(B)** We implement the method for the nonlinear growth model (for the first time) to handle complex temporal data. **(C)** We apply the framework to novel real-world data from the English Longitudinal Study of Ageing (ELSA), showcasing its effectiveness in capturing meaningful and interpretable clusters of income trajectories. These contributions enhance the practical applicability of Bayesian multi-facet clustering in large-scale longitudinal data analysis.

## 2 MULTI-FACET MIXTURE MODEL

A standard mixture model with $K$ components is described as $\sum_{k=1}^{K} \pi_k p(\mathbf{y} \mid \boldsymbol{\theta}_k)$ where $\pi_k$ is the probability of the $k$-th mixture component and $\boldsymbol{\theta}_k$ is the parameter of the respective component. The multi-facet mixture model (MMM), is described as

$$\sum_{k_1=1}^{K_1} \cdots \sum_{k_F=1}^{K_F} \pi_{k_1}^{(1)} \cdots \pi_{k_F}^{(F)} p(\mathbf{y} \mid \boldsymbol{\theta}_{k_1}^{(1)}, \ldots, \boldsymbol{\theta}_{k_F}^{(F)})$$

where $f = 1, \ldots, F$ are $F$ facets of the mixture component and we assume independence among these facets apriori. Here each facet has its respective mixture components described by the probabilities $\boldsymbol{\pi}^{(f)}$ and parameters $\boldsymbol{\Theta}^{(f)}$. In MMM, for each sample $\mathbf{y}_n$, the cluster assignments for different facets are generated independently, and the sample is generated using respective parameters simultaneously i.e.,

$$\begin{aligned} z_n^{(f)} &\sim \mathrm{Cat}(\boldsymbol{\pi}^{(f)}) \qquad \forall\, f = 1, \ldots, F \\ \mathbf{y}_n &\sim p\left(\mathbf{y} \,\middle|\, \boldsymbol{\theta}_{z_n^{(1)}}^{(1)}, \ldots, \boldsymbol{\theta}_{z_n^{(F)}}^{(F)}\right). \end{aligned}$$

Each facet $\boldsymbol{\theta}^{(f)}$ is typically an exclusive partition of the entire parameter space, i.e., $\boldsymbol{\theta}^{(1)} \cup \ldots \cup \boldsymbol{\theta}^{(F)} = \boldsymbol{\theta}$ allowing the model to disentangle and cluster different aspects of the data; and the choice of partition, i.e., facet is guided by the user to provide flexibility and interpretability.

We use the Dirichlet process [Ferguson, 1973] as a nonparametric prior for the parameters $\boldsymbol{\theta}^{(f)}$ of each facet, i.e.,

$$\begin{aligned} G^{(f)} &\sim \mathrm{DP}(G_0^{(f)}, \alpha^{(f)}) \\ \boldsymbol{\theta}_n^{(f)} &\sim G^{(f)} \qquad \forall\, n = 1, \ldots, N \end{aligned}$$

$G^{(f)}$ is a random probability measure made up of discrete values (atoms) for $\boldsymbol{\theta}^{(f)}$. The Dirichlet process prior has two hyperparameters: the base distribution $G_0^{(f)}$ is the mean distribution of the Dirichlet process, commonly chosen to be a conjugate prior; and the concentration parameter $\alpha^{(f)}$, which controls how many distinct clusters are likely to form [Antoniak, 1974]. We use the stick-breaking construction [Sethuraman, 1994], i.e.,

$$\begin{aligned} v_{k_f}^{(f)} &\sim \mathrm{Beta}(1, \alpha^{(f)}) \\ \pi_{k_f}^{(f)} &= v_{k_f}^{(f)} \prod_{i=1}^{k_f - 1} (1 - v_i^{(f)}) \\ \boldsymbol{\theta}_{k_f}^{(f)} &\sim G_0^{(f)}(\boldsymbol{\theta}_{k_f}^{(f)} \mid \boldsymbol{\lambda}_{k_f}^{(f)}). \end{aligned}$$

Here $\boldsymbol{\pi}^{(f)}$ follows the Griffiths-Engen-McCloskey (GEM) distribution. $\boldsymbol{\lambda}_{k_f}^{(f)}$ denotes the parameters of base distribution in general. Given the influence of the concentration parameter on the growth of components within the data, we place conjugate $\mathrm{Gamma}$ priors on $\alpha^{(f)}$, as suggested by Blei and Jordan [2006].

$$\alpha^{(f)} \sim \mathrm{Gamma}(s_1^{(f)}, s_2^{(f)})$$

Thus, the parameters for nonparametric MMM are $\boldsymbol{\Theta} = \left\{\boldsymbol{\theta}_{k_f}^{(f)}, v_{k_f}^{(f)}, \alpha^{(f)}, \ldots\right\}$ alongside auxiliary variables $\mathbf{Z} = \{z_n^{(f)}, \ldots\}$. The plate diagram is shown in Figure 1 (right).

**Mean Field Variational Inference** We apply variational inference using the mean-field method [Blei and Jordan, 2006] to approximate the posterior distribution of the variables of interest. This approach leverages a coordinate ascent algorithm to optimize the evidence lower bound (ELBO). In comparison to the Gibbs sampler, variational inference demonstrates faster convergence, with its runtime largely unaffected by dimensionality Blei and Jordan [2006]. To ensure computational efficiency within the mean-field framework, we adopt fully factorized variational distributions, which assume no dependencies between unobserved variables. In addition, we consider a truncated stick-breaking representation [Blei and Jordan, 2006] to approximate the distribution of the infinite-dimensional random measure $G^{(f)}$. This approach involves setting a fixed truncation level $\ell$ and defining $q(v_\ell^{(f)} = 1) = 1$ for any facet parameter, ensuring that the mixture probabilities $\pi_k^{(f)}(\mathbf{v}^{(f)})$ are zero for $k > \ell$. We use variational distribution in the same family as the respective prior (see Equation C.7). The update rules for $\alpha^{(f)}$ and $\mathbf{v}^{(f)}$ do not depend on the choice of distribution $p$ while $z_n^{(f)}$ and $\boldsymbol{\theta}^{(f)}$ do (see Equation C.8).

**Multi-facet Nonlinear Growth Model** The nonlinear growth model captures complex growth dynamics [Suk et al., 2018], and spline functions [Ahlberg et al., 1967] have been a well-established method for modelling such nonlinearity. We assume that for each individual $n$, the trajectory $\mathbf{y}_n^{\text{obs}}$ is observed at locations $\mathbf{t}_n^{\text{obs}} \in [0, T]^{T_n^{\text{obs}}}$ where $T_n^{\text{obs}}$ is the number of observed locations for an individual. Given the individual cluster assignments $z_n^{(a)}$, $z_n^{(\boldsymbol{\beta})}$ and $z_n^{(\tau)}$ for each facet and corresponding cluster parameters, the likelihood of the $n$-th time series at time $\mathbf{t}_n^{\text{obs}}$ is described as

$$\begin{aligned} \mathbf{y}_n^{\text{obs}} \mid z_n^{(a)} = k_a, z_n^{(\boldsymbol{\beta})} = k_{\boldsymbol{\beta}}, z_n^{(\tau)} = k_\tau \quad \sim \\ \mathcal{N}_{T_n^{\text{obs}}}(a_{k_a} + \boldsymbol{\beta}_{k_{\boldsymbol{\beta}}} \mathcal{B}(\mathbf{t}_n^{\text{obs}}), \tau_{k_\tau} \mathbf{I}) \end{aligned} \tag{1}$$

where $\mathcal{B}(\mathbf{t}_n^{\text{obs}}) \in \mathbb{R}^{L \times T_n^{\text{obs}}}$ is the basis matrix generated by evaluating spline basis functions at locations $\mathbf{t}_n^{\text{obs}}$. We treat the intercept $a$, the coefficient row vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_L)$ and noise precision $\tau$ as three facets. We employ B-spline basis functions of order $p = 2$ with $M$ internal knots, and exclude one of the basis functions to explicitly include an intercept term $a$ (see Supplement D), thus, $L = M + p - 1$. We use the following priors over the parameters, and choose variational distributions in the same family,

$$\begin{aligned} a &\sim \mathcal{N}(\mu^{(a)}, \tau^{(a)}) \\ \boldsymbol{\beta} &\sim \mathcal{N}_L(\boldsymbol{\mu}^{(\boldsymbol{\beta})}, \tau^{(\boldsymbol{\beta})} \mathbf{I}) \\ \tau &\sim \mathrm{Gamma}(\lambda_1^{(\tau)}, \lambda_2^{(\tau)}). \end{aligned}$$

Figure 2 (left) shows the plate diagram of the multi-facet mixture of nonlinear growth models (NLG). Supplement B.1 provides additional model specifications and Supplement C.1 derives the update rules.

**Multi-facet Vector Autoregressive Model** The Vector Autoregressive (VAR) model [Lütkepohl, 2007] captures the linear dynamical relationships among multiple time series. In VAR model involving $D$ variables over $T$ time points, each variable is modelled as a linear transformation of its $P$ preceding values. Here we consider $P = 1$ since we are interested in short time series. Given the individual cluster assignments $z_n^{(a)}$, $z_n^{(\mathbf{B})}$ and $z_n^{(\tau)}$ for each facet and corresponding cluster parameters, the distribution of the $n$-th time series at time $t$ is:

$$\mathbf{y}_{nt} \,|\, \mathbf{y}_{n(t-1)}, z_n^{(a)} = k_a, z_n^{(\mathbf{B})} = k_{\mathbf{B}}, z_n^{(\tau)} = k_\tau \quad \sim \\ \mathcal{N}_D(\mathbf{a}_{k_a} + \mathbf{B}_{k_{\mathbf{B}}}(\mathbf{y}_{n(t-1)} - \mathbf{a}_{k_a}), \mathrm{diag}(\boldsymbol{\tau}_{k_\tau})). \quad (2)$$

We assume the outcome vector at the first time point $\mathbf{y}_{n0} \sim \mathcal{N}_D(\mathbf{a}_{k_a}, \mathrm{diag}(\boldsymbol{\tau}_{k_\tau}))$. We view the intercept vector $\mathbf{a} \in \mathbb{R}^D$, the coefficient matrix $\mathbf{B} \in \mathbb{R}^{D\times D}$ and the noise precision vector $\boldsymbol{\tau} \in \mathbb{R}_+^D$ as three facets. We use Yule-Walker representation [Ghosh et al., 2019, Lütkepohl, 2007] to assign the intercept $\mathbf{a}$ to the trajectory after the evolution equation, allowing the trajectories to be centered around $\mathbf{a}$. This can be extended to accommodate varying time lengths $T_n$ across individuals.

We use the following priors over the parameters, and choose variational distribution in the same family,

$$\mathbf{a} \sim \mathcal{N}_D(\boldsymbol{\mu}^{(a)}, \tau^{(a)}\mathbf{I})$$
$$\mathbf{B} \sim \mathcal{MN}_{D,D}(\mathbf{M}^{(\mathbf{B})}, \mathrm{diag}(\tau^{(\mathbf{B})}), \mathbf{I})$$
$$\tau_d \sim \mathrm{Gamma}(\lambda_1^{(\tau)}, \lambda_2^{(\tau)}) \text{ for } d = 1, \ldots, D$$

The right figure in Figure 2 shows the plate diagram of the nonparametric Bayesian multi-facet vector autoregressive model (VAR). Supplement B.2 provides additional model specification, and Supplement C.2 derives the update rules.

**Implementation Details** Bayesian mixture models often face challenges due to the high multimodality of posterior distributions [Mena and Walker, 2015, Stephens, 1996, Carreira-Perpiñán and Williams, 2003]. Therefore, we perform multiple optimization runs with diverse initializations in parallel and select the run with the highest ELBO as defined in Equation C.6. In addition, we incorporate cluster ordering and cluster pruning techniques during the learning process to enhance the algorithm's performance, following the methods demonstrated by Kurihara et al. [2007] and Lim and Wang [2018]. Cluster ordering involves rearranging clusters in descending order based on their estimated probabilities at each iteration. Cluster pruning discards clusters whose estimated probabilities fall below a specified threshold, dynamically reducing the number of active clusters during the learning process.

We observe that large clusters are often subdivided into smaller, similar clusters (see Supplement A.1). We adjust the hyperparameters of $\alpha^{(f)}$'s prior to have a mean smaller than 1, with a small variance to encourage the automatic merging of such clusters. We found that combining cluster pruning with a smaller prior mean for $\alpha^{(f)}$ helps mitigate the cluster splitting when sufficient iterations for convergence are allowed (see Table A.3). Intuitively, while a smaller prior mean for $\alpha^{(f)}$ can reduce the probability of forming redundant clusters, it typically requires many iterations to reach convergence. Cluster pruning accelerates this process by dynamically shrinking the truncation level $\ell$ during each iteration to approximate the optimal number of clusters.

## 3 SIMULATION STUDIES

**NLG** We use two datasets of different sizes: (NLG-S) a small dataset with $N = 2,400$ and (NLG-L) a large dataset with $N = 15,000$. Furthermore, we use two versions of the same dataset, namely, complete (C) (i.e., no missing values) and incomplete (I). For both cases, the number of time points is set to $T = 10$, while for the incomplete dataset, $50\%$ of the values are randomly removed. The ground truth number of facet clusters in the simulated large datasets is $K_a = 5,\ K_{\boldsymbol{\beta}} = 5$ and $K_\sigma = 5$. Table A.1 reports the resulting average relative $L_2$ errors and adjusted Rand indices (ARIs) for the simulated datasets. Visual representations of the estimations are provided in Supplement A.1.

We observed that the estimations for the intercept and coefficient facets are accurate, as indicated by the low relative $L_2$ errors across all datasets, with values consistently near $0$ for the intercept and ranging between $0.006$ and $0.011$ for the coefficient facet. For the noise parameter, while the estimation is precise (relative $L_2$ error between $0.006$ and $0.017$) in the complete or small datasets (NLG-S-C, NLG-S-I, NLG-L-C), it shows greater error (rel$L_2 = 0.046$ and ARI $< 0.5$) in the large incomplete dataset (NLG-L-I). In terms of ARI for the intercept and coefficient facets, the model achieves near-perfect clustering results (ARI $> 0.9$) under the complete or small datasets (NLG-S-C, NLG-S-I, NLG-L-C) but demonstrates less accurate results ($0.7 <$ ARI $< 0.9$) for the large incomplete dataset (NLG-L-I). Both results are expected since the substantial missingness might hinder the model's ability to effectively infer noise from the time series data points and degrade clustering performance. We found that most mis-clustered trajectories originate from clusters with large noise. This is reasonable as individual trajectories within high-noise clusters often deviate significantly from the mean, making them more susceptible to being mis-clustered into other groups.

A comparison of computational efficiency across inference methods is presented in Table A.5 in the Appendix. Our Variational Bayes method achieves runtimes between ADVI and MLE, offering a substantial speed-up over MCMC while

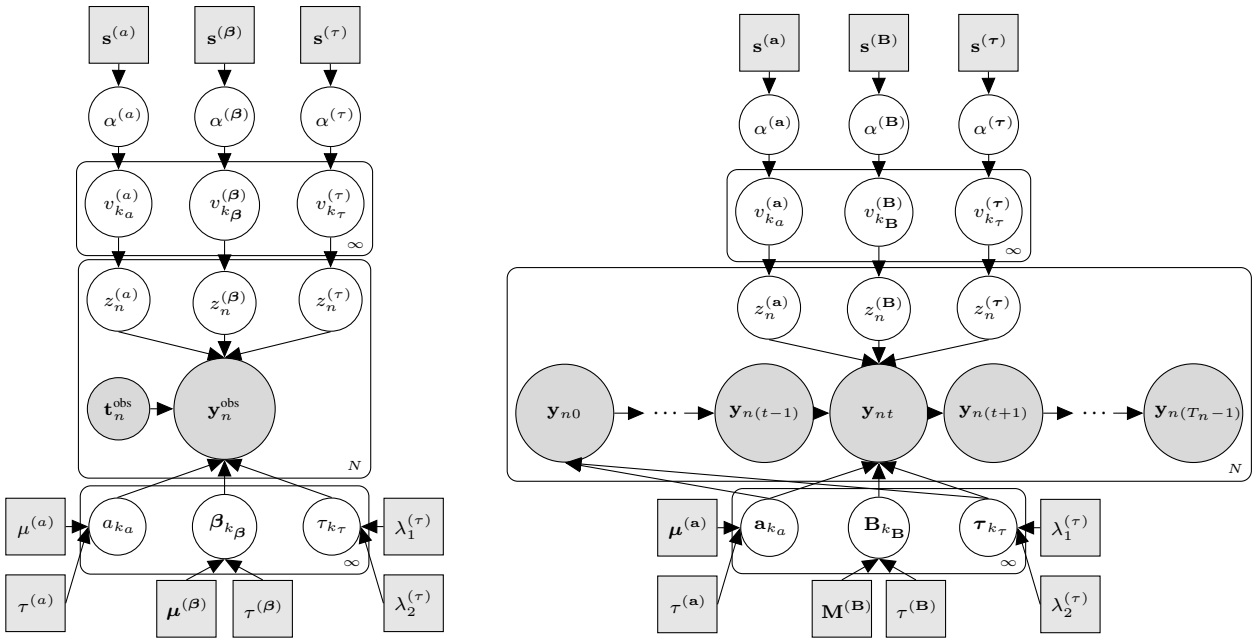


Figure 2: Plate diagram of MMM with **left)** nonlinear growth model and **right)** multivariate autoregressive model.

maintaining competitive performance.

**VAR** We test the model on both small and large datasets with varying time lengths, containing up to $T = 10$ time points and $D = 3$ variables. The ground truth number of facet clusters is $K_{\mathbf{a}} = 3,\ K_{\mathbf{B}} = 3$ and $K_{\boldsymbol{\sigma}} = 3$. The results in Table A.2 and visual representations in Supplement A.1 demonstrate the model's performance.

For the smaller dataset ($N = 1,000$), the relative $L_2$ errors for the intercept, coefficient, and noise facets are $0.002$, $0.017$, and $0.028$, respectively, with corresponding ARIs of $0.915$, $1.0$, and $0.997$, indicating near-perfect clustering accuracy. In the larger dataset ($N = 6,000$), estimation accuracy for the intercept and noise facets improves further (errors decrease to 0.001 and 0.012), while the error for the coefficient matrix slightly increases to 0.09. Nevertheless, ARIs remain high (0.843 for $\mathbf{a}$, 0.982 for $\mathbf{B}$, and 0.989 for $\boldsymbol{\sigma}$), demonstrating accurate clustering and estimation performance across varying-length multivariate time series.

## 4 APPLICATIONS

In this section, we apply the NLG and VAR models to two distinct time series datasets derived from the English Longitudinal Study of Ageing (ELSA) [Banks et al., 2023]. ELSA is a nationally representative dataset of individuals aged $50$ and older, residing in private households and originally derived from the Health Survey for England in 2002. Comprehensive methodological details on ELSA can be found in Pacchiotti et al. [2021].

**ELSA Income Data** The ELSA income data used in this study is consistent with Marshall et al. [2024]. A final sample of $13,002$ respondents is selected by including only individuals who participated in at least two waves of ELSA and reported incomes ranging between $£0$ and $£1000$ per week to reduce the influence of outliers. The income variable used in this analysis is the equivalised total income at the "benefit unit" level, which includes either a single individual or a couple with any dependent children [Marshall et al., 2024]. This equivalisation process adjusts the reported income values so that they represent the income of a single-person household, making it possible to compare households fairly regardless of their size. Additionally, all income data is adjusted for inflation from 2002 to 2019, using 2018/19 as the base year. The analysis considers ages between $50$ to $90$, observing up to nine time points across nine waves for each individual spanning at most 18 years of their life. Due to this longitudinal framework, the missing data rate is significantly high at $86.8\%$.

The single-facet clustering analysis reported by Marshall et al. [2024] identified ten distinct income trajectory clusters. These clusters were later consolidated into four broader categories of later-life income trajectories based on stable and similar income levels following the statutory retirement age of $65$. The resulting categories were labelled as "Luxury" (retirement income at or above $£500$ per week), "Comfortable" ($£300$ to $£500$ per week in retirement), "Always Poor" (generally below $£300$ per week in retirement) and "Boom to Bust" (income rising to $£600$ per week by age 70, then declining to around $£200$ after age 80). A critical differentiating factor among the income trajectory clusters within these broader groups was the degree of income volatility

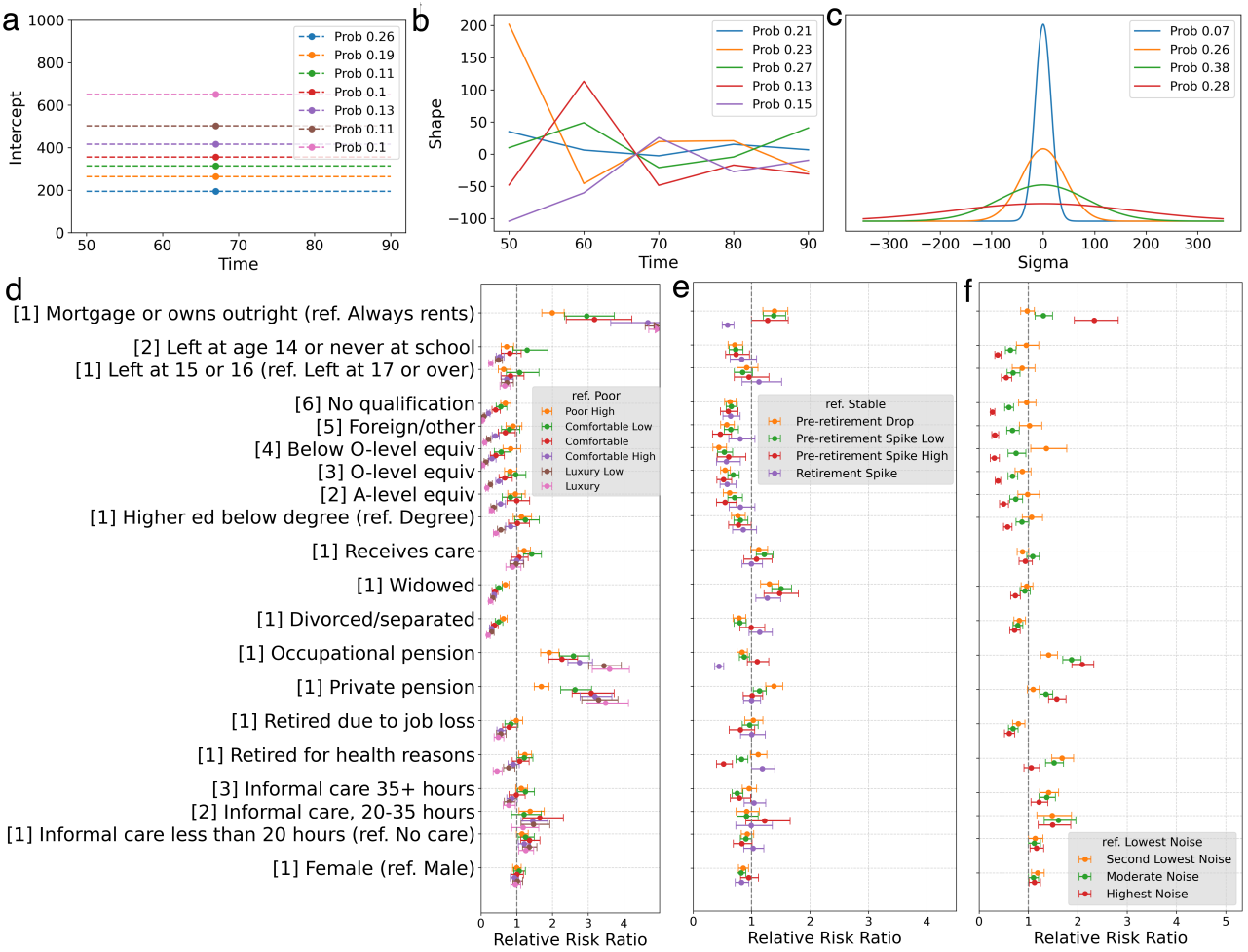



Figure 3: **a)** Seven intercept clusters named “Poor", “Poor High", “Comfortable Low", “Comfortable", “Comfortable High", “Luxury Low" and “Luxury" that broadly fall into the categories of “Poor", “Comfortable" and “Luxury", at the retirement age of 67. **b)** Five shape clusters, “Stable" income, “Pre-retirement Drop" in income, “Pre-retirement Spike Low" in income, “Pre-retirement Spike High" in income and “Retirement Spike" in income. **c)** Four noise clusters. Relative risk ratios (RRRs) for **d)** intercept **e)** shape and **f)** noise clusters.

experienced between the ages of 50 and 65. Three distinct volatility patterns were identified: a pre-retirement income drop, a spike in pre-retirement income, and stable income trajectories.

We applied our NLG model with an intercept shift aligned to the retirement age of 67 (proof in Supplement E) to the same dataset to learn multi-facet clustering results. The shift was applied to gain better interpretability in the context of retirement age as done by Marshall et al. [2024], and also since the missing rate was lower around this age Supplement A.2). The dataset spans $T = 41$ time steps corresponding to ages 50 through 90. We used linear B-splines ($p = 2$) with 3 equidistant internal knots positioned at ages 60, 70 and 80, following the approach in Marshall et al. [2024]. Given the income range of £0 to £1000, the prior mean of the intercept is appropriately set at 500 with a standard deviation of 300, to reflect the central tendency and variability within this range. Moreover, we set truncation level $\ell = 20$ with a pruning probability threshold of 0.05 (see Supplement A.2), and specify the prior for $\alpha^{(f)}$ as $\text{Gamma}(300, 5000)$. To ensure a robust exploration of the optimization landscape, we conducted 50 parallel runs of our Variational Bayesian framework.

The estimated clusters in each facet are visually presented in Figure 3. The model identified seven intercept clusters at 193.9, 263.1, 313.69, 356.01, 415.45, 502.32 and 650.52. We refer to these clusters as “Poor", “Poor High", “Comfortable Low", “Comfortable", “Comfortable High", “Luxury Low" and “Luxury", aligning with the findings of Marshall et al. [2024]. Moreover, five distinct income trajectory shapes were identified: “Stable income" (Cluster 1), “Pre-retirement Drop in income" (Cluster 2), “Pre-retirement

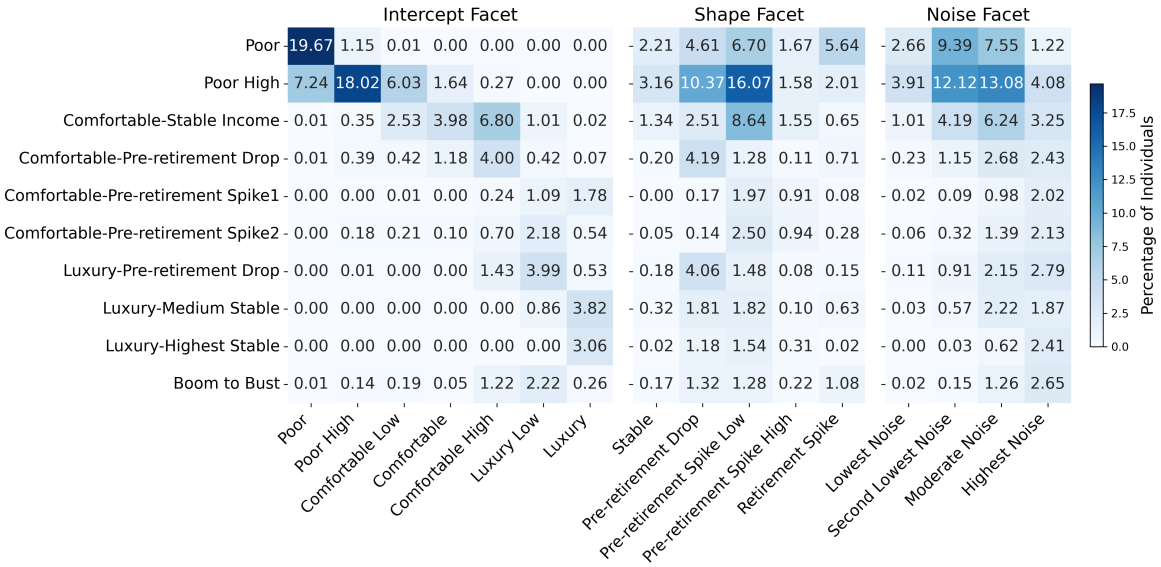



Figure 4: Contingency table of cluster assignments between multi-facet (column) and single-facet (row) clustering presented as percentage over whole population.

Spike Low in income" (Cluster 3), "Pre-retirement Spike High in income" (Cluster 4), and "Retirement Spike in income" (Cluster 5). These findings are largely consistent with those of Marshall et al. [2024]. The four noise clusters reveal significant variability in income trajectories. The noise variable captures the volatility of income around an average trend, and it was not captured previously, thus, adding an additional dimension to the analysis of income trajectory.

Following Marshall et al. [2024], we explore the drivers of these distinct patterns using multinomial regression models with social class, gender, and precarity in housing, pensions, relationships, care, and retirement as predictors. We observe that lower education decreases the likelihood of belonging to "Luxury" cluster compared to "Poor" cluster (RRR $< 0.5$, $p < 0.001$) while having occupational pension increases the likelihood of belonging to "Luxury" cluster compared to "Poor" cluster (RRR $= 3.597$, $p < 0.001$). Compared to "Stable" cluster, lower education decreases the likelihood of belonging to "Pre-retirement Spike" clusters (RRR $< 0.7$, $p < 0.02$). Compared to "Stable" cluster, being widowed increases the likelihood of belonging to "Pre-retirement Drop" cluster (RRR $= 1.505$, $p < 0.001$) while having occupational pension decreases the likelihood of belonging to "Pre-retirement Drop" cluster (RRR $= 0.872$, $p < 0.01$). Compared to "Lowest Noise" cluster, lower education decreases the likelihood of belonging to other higher noise clusters, e.g., "Highest Noise" (RRR $< 0.5$, $p < 0.001$), while an opposite effect is observed for having an occupational pension.

We compare the cluster assignments from our multi-facet model with those from the single-facet model explored by Marshall et al. [2024], which identified ten clusters, and summarize this in Figure 4. We observe general consistency between the two clustering approaches when viewed through a multi-faceted lens. In the intercept facet, a significant proportion of individuals assigned to "Poor" and "Poor High" clusters in the single-facet model have also been assigned to "Poor" and "Poor High" clusters in the multi-facet model, although some "Poor High" individuals in the former have been assigned to "Comfortable Low" in the latter. A similar situation is also observed in the "Luxury" clusters in the single-facet model that have been aligned with "Luxury" cluster in the multi-facet model except "Luxury-Pre-retirement Drop" cluster that has been aligned with "Comfortable High" and "Luxury Low" clusters. Similarly, a significant proportion of individuals assigned to "Luxury-Pre-retirement Drop" and "Comfortable-Pre-retirement Drop" clusters in single-facet model have also been assigned to "Pre-retirement Drop" cluster in the multi-facet clustering model. A similar situation is observed for "Pre-retirement Spike" clusters in the single-facet model that have been aligned to "Pre-retirement Spike Low" and "Pre-retirement Spike High" clusters in the multi-facet model. In the context of the noise facet, a significant proportion of individuals from most clusters in the single-facet clustering except "Poor", "Poor High", "Comfortable-Stable Income", and "Luxury-Medium Stable Income" have been assigned to the "Highest Noise" cluster, while the rest of the clusters align with "Moderate Noise" and "Second Lowest Noise" clusters with "Poor" and "Poor High" also aligning with the "Lowest Noise" cluster in the multi-facet model.

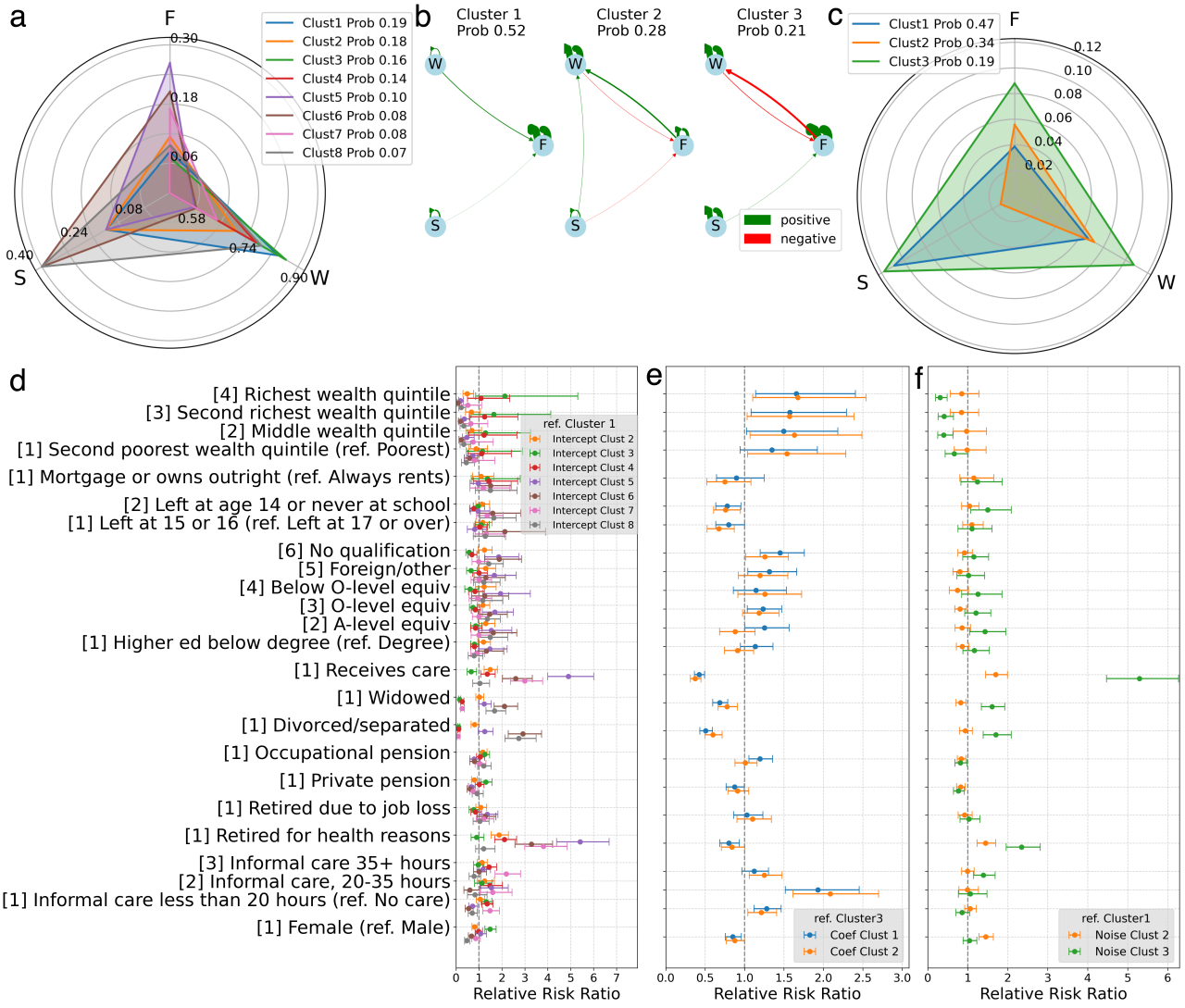



Figure 5: **a)** Eight intercept clusters. **b)** Three coefficient matrix clusters. **c)** Three noise clusters. Relative risk ratios (RRRs) for **d)** intercept **e)** coefficients and **f)** noise clusters.

**ELSA Multivariate Data** The ELSA multivariate dataset analysed includes 6640 individuals with varying time lengths. For each individual, a minimum of 4 time points is observed across nine survey waves. We examine the trajectories of three variables for each individual: "frailty", "wellbeing" and "social isolation". Frailty is quantified using a frailty index [Marshall et al., 2015], ranging from 0 to 56. Wellbeing is measured using CASP-19 [Howel, 2012], with a scale ranging from 0 to 57, and social isolation is assessed using an existing scale of 0 to 6 [Davies et al., 2021]. To ensure consistency, all scales are standardized to range between 0 and 1. Additionally, missing entries in each time series are imputed using mean imputation for each individual.

We applied our first-order VAR model to this dataset, which consists of up to $T_n \in [4, 9]$ time steps and $D = 3$ variables. Since the values for each variable were standardized to fall between 0 and 1, the prior mean of the intercepts was set to 0.5, with a prior standard deviation of 0.1. The truncation level was set to $\ell = 20$ with a pruning threshold of 0.05, and the prior for $\alpha^{(f)}$ was specified as $\mathrm{Gamma}(300, 5000)$. To ensure robust optimization, we performed 50 parallel runs using our VB framework.

The estimated clusters in each facet are visually displayed in Figure 5. We found eight intercept clusters representing varying levels of frailty, wellbeing and social isolation. The coefficients cluster 2 and 3 largely align with the intuitive notion that wellbeing has a negative impact on frailty. However, cluster 1 and 2 demonstrate a more counterintuitive relationship where the opposite is observed, i.e., wellbeing positively impacting frailty and vice versa. Cluster 3 shows a strong negative influence of frailty on wellbeing.

The noise facet reveals significant variability in social isolation trajectories in cluster 1 and 3 compared to cluster 2, while cluster 2 and 3 are clusters least and most driven by noise respectively.

Similar to our analysis in the previous section, we explore the drivers of these patterns using covariates. Due to the smaller size of the data and the large number of clusters, we mostly observe broad confidence intervals from the multinomial regression. However, some interesting patterns appear nonetheless. Compared to intercept cluster 1, being divorced or widowed increases (RRR $> 1.5$, $p < 0.001$) the likelihood of belonging to cluster 6 (higher frailty and lower wellbeing than cluster 1) and cluster 8 (higher social isolation than cluster 1). An opposite effect is observed for intercept clusters 3, 4 (lower social isolation than cluster 1) and cluster 7 (lower wellbeing and social isolation than cluster 1) (RRR $< 0.3$, $p < 0.001$). Compared to coefficient cluster 3 (intuitive direction of wellbeing negatively affecting frailty), receiving care decreases the likelihood of belonging to clusters 1 and 2 (counterintuitive direction of wellbeing positively affecting frailty, RRR $< 0.5$, $p < 0.001$). Compared to noise cluster 1, receiving care increases the likelihood of belonging to cluster 3 (higher noise variance in wellbeing and frailty compared to cluster 1, RRR $= 5.294$, $p < 0.001$). Compared to noise cluster 1, being widowed decreases the likelihood of belonging to cluster 2 (lower noise variance in social isolation compared to cluster 1, RRR $= 0.823$, $p < 0.02$).

## 5 DISCUSSION

Traditional time series clustering methods, like those used by Marshall et al. [2024] and Bulteel et al. [2016], typically produce a single clustering solution using all facets simultaneously and require extensive post-analysis to interpret the clusters. In contrast, nonparametric Bayesian multi-facet clustering model disentangles multiple facets within a dataset, each described by distinct characteristics. This enhances interpretability by providing clearer insight into why clusters form and what defines them. Additionally, tying facets directly to model parameters offers an intuitive way to explain clustering outcomes.

In this paper, we present an extension to existing multi-facet mixture models. First, we incorporate a variational Bayesian framework, which offers enhanced computational efficiency and is particularly well-suited for large-scale datasets, in contrast to traditional MCMC sampling methods. Second, we incorporate Dirichlet process priors to simultaneously learn the number of facet clusters, removing the need to predefine this value. Third, we apply multi-facet clustering in the context of nonlinear growth models. Fourth, we capture an additional dimension, i.e., the noise characteristics as a facet for both nonlinear growth model and multivariate regression model. Fifth, we demonstrate the versatility of the proposed method through two detailed time series model applications tested on real datasets.

This multi-facet clustering framework can be further generalised to a broader class of time series models by adopting alternative likelihood functions. For example, a Hidden Markov Model (HMM) can be used for categorical response data, where the facets may correspond to the columns of the transition matrix, capturing distinct state dynamics. This flexibility allows the multi-facet approach to be adapted to diverse temporal modelling contexts, enabling interpretable clustering based on model-specific structural elements.

While the method and analysis have notable strengths, some limitations remain. From a conceptual perspective, the independence of facets apriori is a strong assumption that considers any combinations of parameters from the facets to be feasible. However, in practice, this might not happen, thus, creating a model mismatch as shown in the introduction Figure 1. This limitation can be addressed by allowing facets to be dependent, but this can increase the number of cluster probabilities being estimated, potentially impacting computational efficiency and model identifiability. Another aspect of the multi-facet clustering is the choice of facets. For instance, in the VAR model, the coefficient matrix facet can be further decomposed into row-wise facets to capture variable-specific interaction patterns (as discussed in [Kundu and Lukemire, 2024]). This alternative facet specification may lead to different clustering outcomes. Therefore, the definition and granularity of facets are inherently subjective and should be guided by the research question and the interpretability of the underlying parameter components. From an implementation perspective, a tuning of the pruning threshold may be necessary to address the splitting of similar clusters. This process can be improved and validated more extensively on simulated data.

Multi-facet clustering offers a unique and exciting direction to clustering complex temporal data in an interpretable manner. This study takes a step in this direction by implementing this method on several temporal models, and applying this approach to complex real-world applications. The framework offers valuable insights into complex real-world phenomena and provides a flexible, interpretable, and computationally efficient approach for analysing multi-faceted real datasets.

**Acknowledgements**

Research reported in this publication was supported by the National Institute On Aging of the National Institutes of Health under Award Number R01AG017644. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. ELSA is funded by the NIHR Policy Research Programme (HEI) 198_1074_03. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. Wang is supported by the United Kingdom Research and Innovation (grant EP/S02431X/1), UKRI Centre for Doctoral Training in Biomedical AI at the University of Edinburgh, School of Informatics. Richards is supported by the National Institute for Health and Care Research (NIHR) under its Artificial Intelligence for Multiple and Long-Term Conditions Programme (project references NIHR203982), and Seth is partly supported by the National Institute for Health and Care Research (NIHR) under its Artificial Intelligence for Multiple and Long-Term Conditions Programme (reference number NIHR202639 and NIHR203982). The views expressed are those of the author and not necessarily those of the NIHR or the Department of Health and Social Care. Seth is partly supported by the Legal & General Group (research grant to establish the independent Advanced Care Research Centre at the University of Edinburgh). The funder had no role in the conduct of the study, interpretation or the decision to submit for publication. The views expressed are those of the authors and not necessarily those of Legal & General.

**References**

J. H. Ahlberg, E. N. Nilson, and J. L. Walsh. *The Theory of Splines and Their Applications*. Issn. Elsevier Science, 1967. ISBN 9780080955452.

Charles E. Antoniak. Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *The Annals of Statistics*, 2(6):1152 – 1174, 1974. doi: https://doi.org/10.1214/aos/1176342871.

J. Banks, G. David Batty, J. Breedvelt, K. Coughlin, R. Crawford, M. Marmot, J. Nazroo, Z. Oldfield, N. Steel, A. Steptoe, M. Wood, and P. Zaninotto. English Longitudinal Study of Ageing: Waves 0-9, 1998-2019. *(SN5050; Version 39) UK Data Service*, 2023. doi: https://doi.org/10.5255/ukda-sn-5050-26.

David M. Blei and Michael I. Jordan. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1): 121 – 143, 2006. doi: https://doi.org/10.1214/06-ba104.

Kirsten Bulteel, Francis Tuerlinckx, Annette Brose, and Eva Ceulemans. Clustering Vector Autoregressive Models: Capturing Qualitative Differences in Within-Person Dynamics. *Frontiers in Psychology*, 7, October 2016. ISSN 1664-1078. doi: https://doi.org/10.3389/fpsyg.2016.01540.

Miguel Á. Carreira-Perpiñán and Christopher K. I. Williams. On the Number of Modes of a Gaussian Mixture. In *Scale Space Methods in Computer Vision*, pages 625–640. Springer Berlin Heidelberg, 2003. ISBN 978-3-540-44935-5.

Guoqing Chao, Shiliang Sun, and Jinbo Bi. A Survey on Multiview Clustering. *IEEE Transactions on Artificial Intelligence*, 2(2):146–168, 2021. doi: https://doi.org/10.1109/tai.2021.3065894.

Katie Davies, Asri Maharani, Tarani Chandola, Chris Todd, and Neil Pendleton. The longitudinal relationship between loneliness, social isolation, and frailty in older adults in England: a prospective analysis. *The Lancet Healthy Longevity*, 2(2):e70–e77, January 2021. doi: https://doi.org/10.1016/s2666-7568(20)30038-6.

A. P. Dawid. Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika*, 68(1):265–274, 04 1981. ISSN 0006-3444. doi: https://doi.org/10.1093/biomet/68.1.265.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, page 226–231. AAAI Press, 1996.

Fabian Falck, Haoting Zhang, Matthew Willetts, George Nicholson, Christopher Yau, and Chris Holmes. Multi-Facet Clustering Variational Autoencoders. In *Advances in Neural Information Processing Systems*, volume 34, pages 8676–8690. Curran Associates, Inc., June 2021. ISBN 9781713845393.

Thomas S. Ferguson. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209 – 230, 1973. doi: https://doi.org/10.1214/aos/1176342360.

Chris Fraley and Adrian E Raftery. Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002. doi: https://doi.org/10.1198/016214502760047131.

Jerome H. Friedman and Jacqueline J. Meulman. Clustering Objects on Subsets of Attributes (with Discussion). *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 66(4):815–849, 10 2004. ISSN 1369-7412. doi: https://doi.org/10.1111/j.1467-9868.2004.02059.x.

Giuliano Galimberti and Gabriele Soffritti. Model-Based Methods to Identify Multiple Cluster Structures in a Data Set. *Computational Statistics & Data Analysis*, 52(1):

520–536, September 2007. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2007.02.019.

Satyajit Ghosh, Kshitij Khare, and George Michailidis. High-Dimensional Posterior Consistency in Bayesian Vector Autoregressive Models. *Journal of the American Statistical Association*, 114(526), 2019. ISSN 0162-1459. doi: https://doi.org/10.1080/01621459.2018.1437043.

A Gordon. *Classification*. Chapman and Hall/CRC, June 1999. doi: https://doi.org/10.1201/9780367805302.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*, chapter 14.3.12 Hierarchical clustering. Springer, 2009.

Denise Howel. Interpreting and evaluating the CASP-19 quality of life measure in older people. *Age and ageing*, 41:612–7, 03 2012. doi: https://doi.org/10.1093/ageing/afs023.

S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79 – 86, 1951. doi: https://doi.org/10.1214/aoms/1177729694.

Suprateek Kundu and Joshua Lukemire. Flexible Bayesian Product Mixture Models for Vector Autoregressions. *J. Mach. Learn. Res.*, 25, 2024.

Kenichi Kurihara, Max Welling, and Yee Whye Teh. Collapsed variational Dirichlet process mixture models. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, page 2796–2801. Morgan Kaufmann Publishers Inc., 2007.

Kart-Leong Lim and Han Wang. Fast approximation of variational Bayes Dirichlet process mixture using the maximization–maximization algorithm. *International Journal of Approximate Reasoning*, 93:153–177, 2018. ISSN 0888-613x. doi: https://doi.org/10.1016/j.ijar.2017.11.001.

Helmut Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer Publishing Company, Incorporated, 2007. ISBN 3540262393.

J. B. MacQueen. *Some methods for classification and analysis of multivariate observations*, pages 281–297. University of California Press, 1967.

Alan Marshall, James Nazroo, Gindo Tampubolon, and Bram Vanhoutte. Cohort differences in the levels and trajectories of frailty among older people in England. *Journal of Epidemiology & Community Health*, 69(4):316–321, 2015. ISSN 0143-005x. doi: https://doi.org/10.1136/jech-2014-204655.

Alan Marshall, Chima Eke, Bruce Guthrie, Carys Pugh, and Sohan Seth. Income Trajectories and Precarity in Later Life. *Journal of Population Ageing*, January 2024. doi: https://doi.org/10.1007/s12062-023-09437-2.

Ramsés H. Mena and Stephen G. Walker. On the Bayesian Mixture Model and Identifiability. *Journal of Computational and Graphical Statistics*, 24(4):1155–1169, 2015. ISSN 10618600.

Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012. ISBN 0262018020.

Richi Nayak and Khanh Luong. *Multi-Aspect Learning: Methods and Applications*, volume 242 of *Intelligent Systems Reference Library*. Springer International Publishing, 2023. ISBN 978-3-031-33559-4 978-3-031-33560-0. doi: https://doi.org/10.1007/978-3-031-33560-0.

Donglin Niu, Jennifer Dy, and Zoubin Ghahramani. A Nonparametric Bayesian Model for Multiple Clustering with Overlapping Feature Views. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22, pages 814–822. Pmlr, March 2012.

Bruno Pacchiotti, David Hussey, and Gary Bennett. The dynamics of ageing: The 2018-2019 english longitudinal study of ageing (Wave 9) technical report, 2021.

Jayaram Sethuraman. A Constructive Definition of the Dirichlet Prior. *Statistica Sinica*, 4:639–650, 01 1994.

Matthew Stephens. Dealing with the Multimodal Distributions of Mixture Model Parameters. *Preprint, Department of Statistics, University of Oxford*, 1996.

Hye Suk, Stephen West, Kimberly Fine, and Kevin Grimm. Nonlinear Growth Curve Modeling Using Penalized Spline Models: A Gentle Introduction. *Psychological Methods*, 24, 08 2018. doi: https://doi.org/10.1037/met0000193.

T. Warren Liao. Clustering of time series data—a survey. *Pattern Recognition*, 38(11):1857–1874, 2005. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2005.01.025.

J. Wilpon and L. Rabiner. A modified K-means clustering algorithm for use in isolated work recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(3):587–594, 1985. doi: https://doi.org/10.1109/tassp.1985.1164581.

Rui Xu and Donald Wunsch. Survey of Clustering Algorithms. *Neural Networks, IEEE Transactions on*, 16:645–678, 06 2005. doi: https://doi.org/10.1109/tnn.2005.845141.

Wei Zong, Danyang Li, Marianne L Seney, Colleen A Mcclung, and George C Tseng. Model-based multifacet clustering with high-dimensional omics applications. *Biostatistics*, page kxae020, 07 2024. ISSN 1465-4644. doi: https://doi.org/10.1093/biostatistics/kxae020.

# Supplementary Material

**Luwei Wang**[1] **Kieran Richards**[1] **Sohan Seth**[1]

[1]School of Informatics, University of Edinburgh, UK

## A ADDITIONAL RESULTS

### A.1 SIMULATION

In this section, we show additional results for simulations. Specifically, we first examine learning outcomes under various truncation levels for both the NLG and VAR models without cluster pruning, as shown in Figure A.1. The findings suggest that learning outcomes are generally stable when the truncation level exceeds the true number of clusters. However, a notable issue arises where some large clusters may split into multiple smaller, similar clusters. This phenomenon is observed in the visual representations of estimations for both models. Figure A.3 and Figure A.4 provide visual representations of NLG estimations on incomplete datasets across different truncation levels, while Figure A.5 and Figure A.6 depict estimations for VAR under varying truncation levels. The ARIs fluctuate due to cluster splitting issues, which can slightly degrade the ARI values. We then investigate the relationship between ELBO and truncation levels in Figure A.2. The figure shows that the highest ELBO is achieved at the correct truncation level, which aligns with the notion of cluster pruning during the learning process.

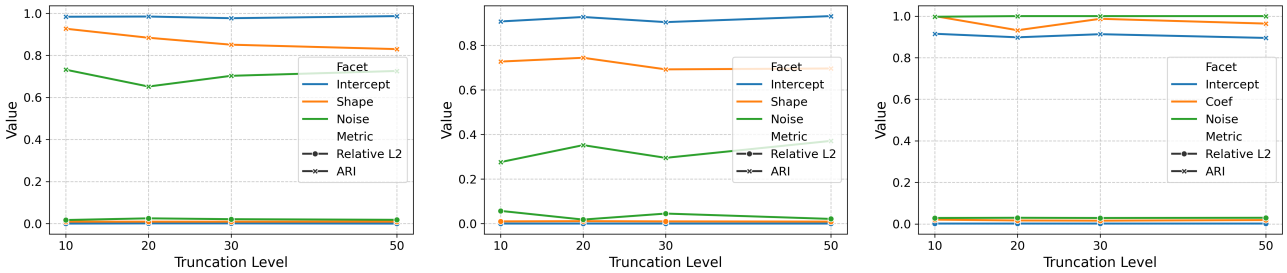
We demonstrate the effects of cluster pruning and tuning the hyperparameters of $\alpha^{(f)}$'s prior in simulation for the VAR model, as shown in Table A.3. To ensure a fair comparison, we use the same random seed for all runs to eliminate the influence of initialization, along with a fixed maximum of $500$ iterations and a truncation level of $20$. Our results show that combining cluster pruning with hyperparameter tuning of the $\alpha^{(f)}$ prior not only facilitates convergence to the optimal number of clusters but also accelerates convergence, requiring fewer iterations. Moreover, we explore the impact of tuning the pruning threshold based on ELBO, as shown in Table A.4, using the same prior settings, maximum iteration and truncation level.

Table A.1: The Average Relative $L_2$ Errors and ARIs of Facets for NLG Simulation.

| Dataset | rel$L_2$ $a$ | rel$L_2$ $\boldsymbol{\beta}$ | rel$L_2$ $\sigma$ | ARI $a$ | ARI $\boldsymbol{\beta}$ | ARI $\sigma$ |
|---|---|---|---|---|---|---|
| $N = 2,400; \text{NA}\% = 0$ | 0.000 | 0.006 | 0.007 | 0.999 | 1.0 | 0.999 |
| $N = 2,400; \text{NA}\% = 20$ | 0.000 | 0.010 | 0.006 | 1.0 | 0.996 | 0.995 |
| $N = 15,000; \text{NA}\% = 0$ | 0.000 | 0.009 | 0.017 | 0.984 | 0.930 | 0.734 |
| $N = 15,000; \text{NA}\% = 50$ | 0.000 | 0.011 | 0.046 | 0.886 | 0.73 | 0.426 |

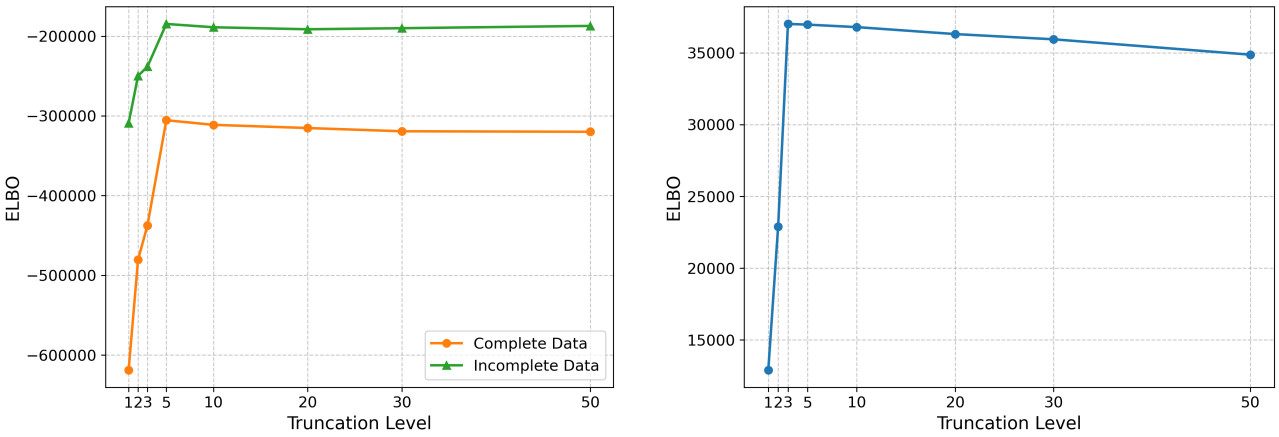
Table A.2: The Average Relative $L_2$ Errors and ARIs of Facets for VAR Simulation.

| Dataset | rel$L_2$ $\mathbf{a}$ | rel$L_2$ $\mathbf{B}$ | rel$L_2$ $\boldsymbol{\sigma}$ | ARI $\mathbf{a}$ | ARI $\mathbf{B}$ | ARI $\boldsymbol{\sigma}$ |
|---|---|---|---|---|---|---|
| $N = 1,000; \text{NA}\% = 0$ | 0.002 | 0.017 | 0.028 | 0.915 | 1.0 | 0.997 |
| $N = 6,000; \text{NA}\% = 0$ | 0.001 | 0.09 | 0.012 | 0.843 | 0.982 | 0.989 |




(a) NLG complete simulated data (b) NLG incomplete simulated data (c) VAR simulated data

Figure A.1: The relative $L_2$ errors and ARIs vs. different truncation levels for simulations.




(a) NLG model (ground truth at 5) (b) VAR model (ground truth at 3)

Figure A.2: The ELBO vs. different truncation levels for simulations. Both models achieve their highest ELBO when the truncation level matches the ground truth number of clusters.

Table A.3: Demonstration of Effects of Cluster Pruning and Hyperparameters Tuning of $\alpha^{(f)}$ Prior by VAR Simulation.

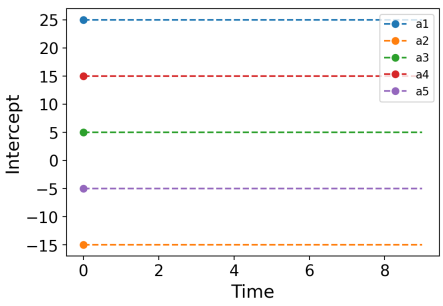
| Cluster pruning | $\alpha^{(f)}$ prior tuning | Iterations used | #Cluster $\mathbf{a}$ | #Cluster $\mathbf{B}$ | #Cluster $\boldsymbol{\sigma}$ |
|---|---|---|---|---|---|
| No | No | 500 | 4 | 12 | 3 |
| No | Yes | 500 | 3 | 8 | 3 |
| Yes | No | 500 | 3 | 4 | 3 |
| Yes | Yes | 470 | 3 | 3 | 3 |
| True #Clusters | | | 3 | 3 | 3 |

Table A.4: Demonstration of Tuning Cluster Pruning Threshold by NLG Simulation.

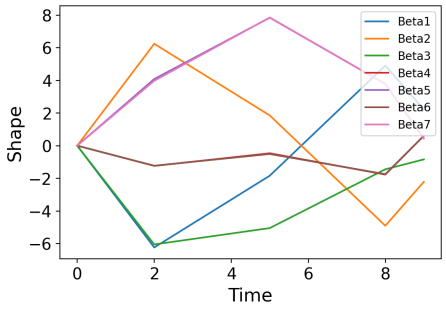
| Pruning threshold | ELBO | #Cluster $a$ | #Cluster $\boldsymbol{\beta}$ | #Cluster $\sigma$ |
|---|---|---|---|---|
| No pruning | -189642.87 | 6 | 10 | 6 |
| 0.01 | -186681.43 | 5 | 7 | 7 |
| 0.02 | -185241.61 | 5 | 6 | 6 |
| 0.04 | -184177.22 | 5 | 5 | 5 |
| 0.06 | -183491.80 | 5 | 5 | 5 |
| 0.07 | **-183135.41** | 5 | 5 | 5 |
| 0.08 | -190081.36 | 5 | 5 | 5 |
| True #Clusters | | 5 | 5 | 5 |

Table A.5: Average runtime of the BMF-NLG model using different inference methods implemented in RSTAN.

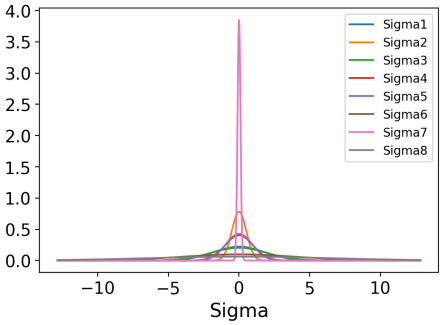
| Size $N$ | HMC | ADVI | MLE |
|---|---|---|---|
| 250 | 1h | 3min | 30s |
| 300 | 4h | 10min | 80s |
| 1200 | 15h | 1h | 7min |
| 2400 | 60h | 2h | 20min |
| 15000 | >10days | >2days | 20h |



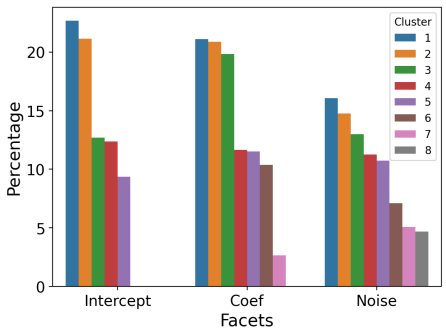

(a) $a$




(b) $\boldsymbol{\beta}$

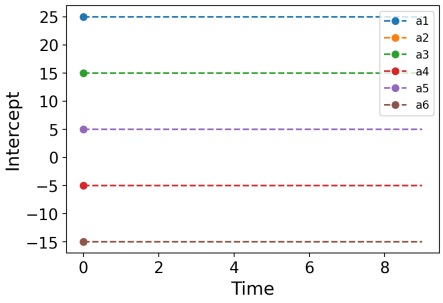



(c) $\sigma$

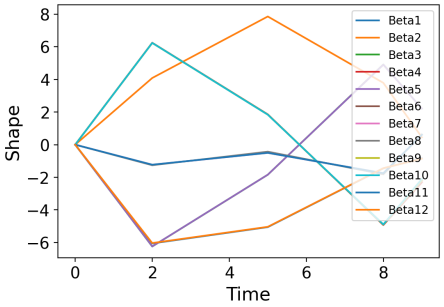



(d) $\boldsymbol{\pi}$ for facets

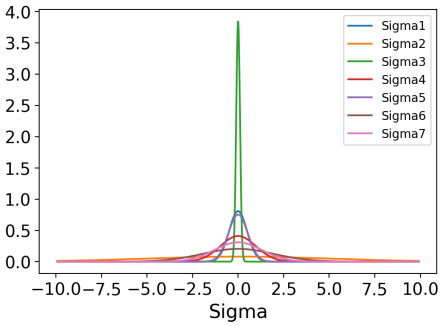
Figure A.3: Parameter estimations of the NLG model on the large incomplete dataset at truncation level 10. The true number of clusters for all three facets is 5.



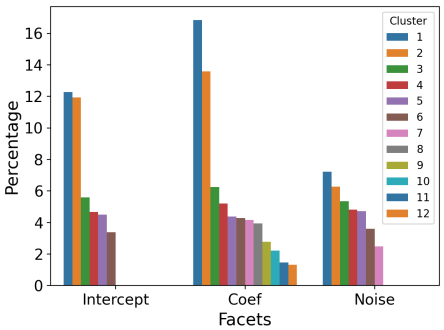

(a) $a$




(b) $\boldsymbol{\beta}$

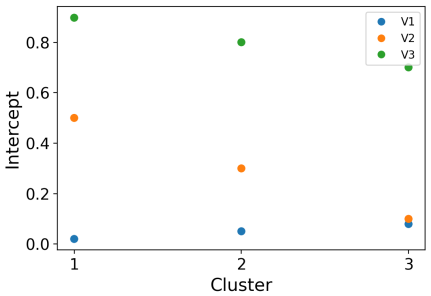



(c) $\sigma$

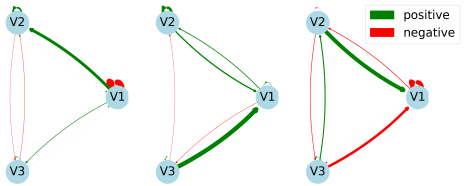



(d) $\boldsymbol{\pi}$ for facets

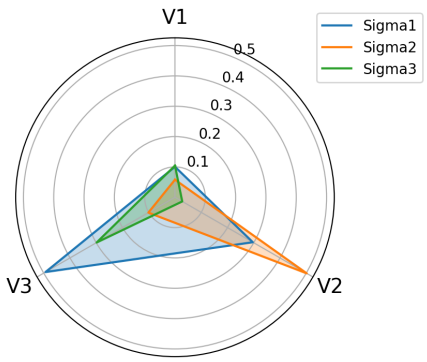
Figure A.4: Parameter estimations of the NLG model on the large incomplete dataset at truncation level 50.



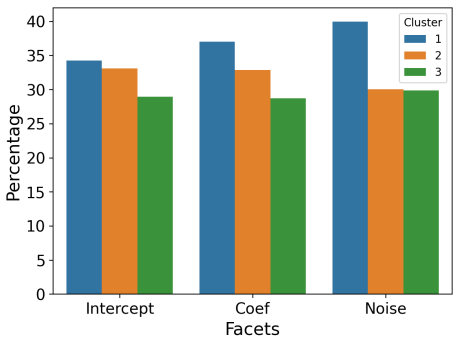

(a) $\mathbf{a}$




(b) $\mathbf{B}$

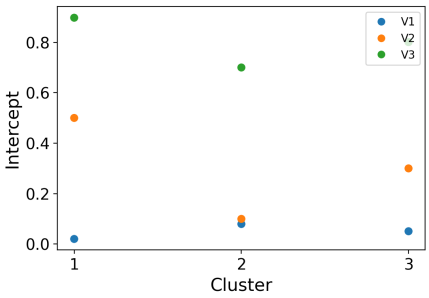



(c) $\boldsymbol{\sigma}$

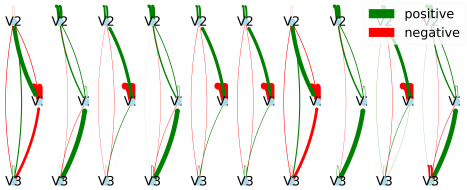



(d) $\boldsymbol{\pi}$ for facets

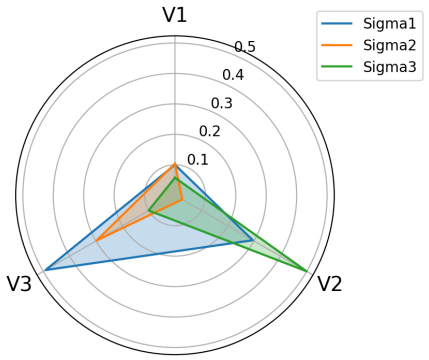
Figure A.5: Parameter estimations of the VAR model on the simulated dataset when truncation level is 10. The true number of clusters for all three facets is 3.



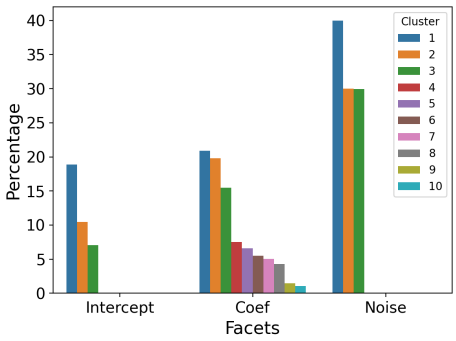

(a) $\mathbf{a}$




(b) $\mathbf{B}$




(c) $\boldsymbol{\sigma}$




(d) $\boldsymbol{\pi}$ for facets

Figure A.6: Parameter estimations of the VAR model on the simulated dataset when truncation level is 50. For the coefficient facet in (b), Cluster 1 and Cluster 7 are the same, Cluster 2, Cluster 4, Cluster 8 and Cluster 10 are the same while Cluster 3, Cluster 5, Cluster 6 and Cluster 9 are the same.

## A.2 APPLICATIONS

In this section, we present additional results for the application. Figure A.7 shows the missing rate in the ELBO income dataset at different ages. We see the missing rate at the retirement age 67 is relatively lower. As a result, we changed the intercept facet from age 50 to retirement age 67. Figure A.8 displays the impact of different cluster pruning thresholds on real datasets. Due to the high missing rate and high noise in the real datasets, we choose the pruning threshold based on the number of clusters and the smallest cluster probability we expect (0.05 for both datasets).

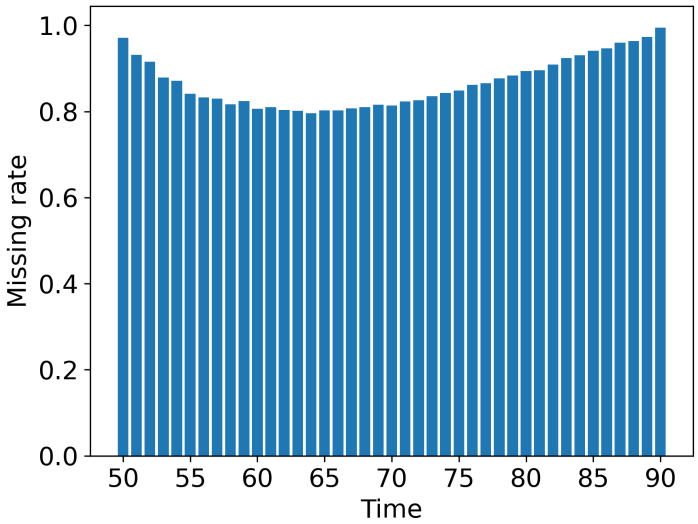
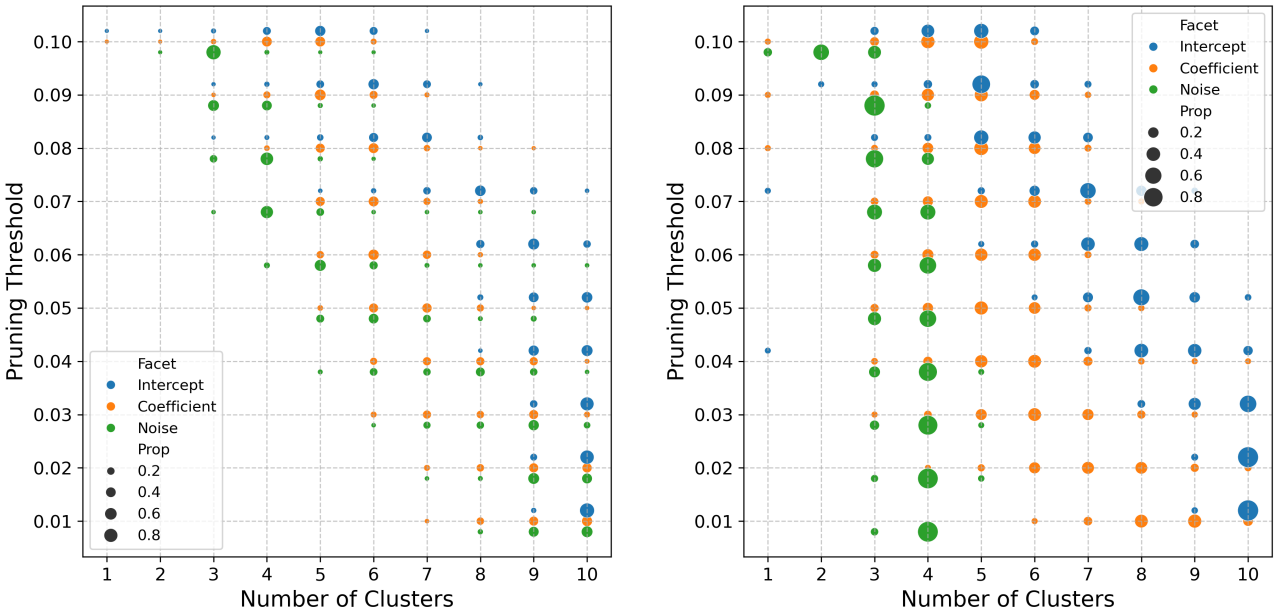



Figure A.7: The missing rate of ELSA data for the NLG model at each age time point.




(a) ELSA income data

(b) ELSA multivariate data

Figure A.8: Different pruning thresholds vs. number of clusters for real datasets. We compute the proportion of the number of clusters learned from multiple runs.

# B ADDITIONAL EXAMPLE MODEL DETAILS

## B.1 NLG

The mathematical expression of the nonlinear growth model is

$$f_n^{(p)}(t) = a_n + \sum_{i=1}^{L} \beta_{ni} B_{i,p}(t)$$

where the spline function can be built from a linear combination of a collection of B-splines $\{B_{i,p}(t)\}_{i=1}^{L}$ of degree $p-1$ with coefficient $\beta_i$. Note that if we set $M = 0$ and $p = 2$, meaning no internal knots and use linear B-splines, it simply forms the linear growth model.

Assume an additive Gaussian noise for each trajectory at various time points $\epsilon_{nt} \sim \mathcal{N}(0, \tau_n)$ and we let $\tau = \frac{1}{\sigma^2}$ known as precision. Given the individual cluster assignments $z_n^{(a)}$, $z_n^{(\boldsymbol{\beta})}$ and $z_n^{(\tau)}$ for each facet and corresponding cluster parameters, we have the observable data distributed as

$$\mathbf{y}_n \,|\, z_n^{(a)} = k_a, z_n^{(\boldsymbol{\beta})} = k_{\boldsymbol{\beta}}, z_n^{(\tau)} = k_\tau \quad \sim \quad \mathcal{N}_T(a_{k_a} + \boldsymbol{\beta}_{k_{\boldsymbol{\beta}}} \mathcal{B}(\mathbf{t}), \tau_{k_\tau} \mathbf{I}) \tag{B.3}$$

where $\mathcal{B}(\mathbf{t}) \in \mathbb{R}^{L \times T}$ is the basis spline matrix of order $p$ and $\tau_{k_\tau}\mathbf{I}$ is the precision matrix with scalar precision parameter. The cluster numbers $k_a, k_{\boldsymbol{\beta}}, k_\tau$ can go to infinity.

We specify the base distributions for facet parameters as conjugate priors. That is,

$$\begin{aligned} G_0^{(a)} &\sim \mathcal{N}(\mu^{(a)}, \tau^{(a)}) \\ G_0^{(\boldsymbol{\beta})} &\sim \mathcal{N}_L(\boldsymbol{\mu}^{(\boldsymbol{\beta})}, \tau^{(\boldsymbol{\beta})}\mathbf{I}) \\ G_0^{(\tau)} &\sim \text{Gamma}(\lambda_1^{(\tau)}, \lambda_2^{(\tau)}) \end{aligned} \tag{B.4}$$

where $\tau^{(a)}$ and $\tau^{(\boldsymbol{\beta})}\mathbf{I}$ are precisions. Therefore, the corresponding variational distributions should be in the same distribution family as priors.

For incomplete data, we have $p(\mathbf{y}_n \,|\, \mathbf{z}_n, \mathbf{a}, \boldsymbol{\beta}, \boldsymbol{\tau}) = p(\mathbf{y}_n^{\text{obs}} \,|\, \mathbf{z}_n, \mathbf{a}, \boldsymbol{\beta}, \boldsymbol{\tau}) p(\mathbf{y}_n^{\text{miss}} \,|\, \mathbf{z}_n, \mathbf{a}, \boldsymbol{\beta}, \boldsymbol{\tau})$. Denote $\mathbf{t}_n^{\text{obs}}$ observed time points for each $n$ where $\left|\mathbf{t}_n^{\text{obs}}\right| = T_n^{\text{obs}}$ then $\mathbf{y}_n^{\text{obs}}$ will depend on $\mathbf{t}_n^{\text{obs}}$. So, the likelihood B.3 becomes distribution 1. It is obvious that with the marginal distribution 1 of observed data, we can simply replace all $\mathbf{y}_n$, $\mathcal{B}(\mathbf{t})$ and $T$ with $\mathbf{y}_n^{\text{obs}}$, $\mathcal{B}(\mathbf{t}_n^{\text{obs}})$ and $T_n^{\text{obs}}$ in the update rules.

## B.2 VAR

We define the first-order vector autoregressive model for an individual time series $n$ as:

$$\mathbf{y}_{nt} = \mathbf{a}_n + \mathbf{B}_n(\mathbf{y}_{n(t-1)} - \mathbf{a}_n) + \boldsymbol{\epsilon}_{nt}$$

where $\mathbf{y}_{nt}$ and $\mathbf{y}_{n(t-1)} \in \mathbb{R}^D$ are $D$ dimensional vector of time series values at time points $t$ and $t-1$ and $\mathbf{a}_n$ denotes the intercept term. $\mathbf{B}_n$ is a $D \times D$ matrix containing the regression coefficients where $\mathbf{B}_{n,ij}$ refers to the coefficient of $y_{n(t-1),j}$ in the linear function for $y_{nt,i}$. We assume $\boldsymbol{\epsilon}_{nt} \sim \mathcal{N}_D(0, \text{diag}(\boldsymbol{\tau}_n))$ is the time-invariant noise term with diagonal precision matrix parametrised by $\boldsymbol{\tau}_n \in \mathbb{R}_+^D$.

The distribution of individual time series is already stated in the main text and the joint likelihood of the entire time series is

$$p(\mathbf{y}_n = [\mathbf{y}_{n0}, \ldots, \mathbf{y}_{n(T-1)}] \,|\, \mathbf{z}_n, \mathbf{a}, \mathbf{B}, \boldsymbol{\tau}) = p(\mathbf{y}_{n0} \,|\, \mathbf{z}_n, \mathbf{a}, \boldsymbol{\tau}) \prod_{t=1}^{T-1} p(\mathbf{y}_{nt} \,|\, \mathbf{y}_{n(t-1)}, \mathbf{z}_n, \mathbf{a}, \mathbf{B}, \boldsymbol{\tau})$$

where the individual time series matrix $\mathbf{y}_n \in \mathbb{R}^{D \times T}$. This forms a Matrix normal distribution [Dawid, 1981]:

$$\mathbf{y}_n \,|\, z_n^{(a)} = k_a, z_n^{(\mathbf{B})} = k_{\mathbf{B}}, z_n^{(\tau)} = k_\tau \quad \sim \quad \mathcal{MN}_{D,T}\left(\mathbf{a}_{k_a}\mathbf{1}^\intercal + \left[\mathbf{0}, \mathbf{B}_{k_{\mathbf{B}}}([\mathbf{y}_{n0}, \ldots, \mathbf{y}_{n(T-2)}] - \mathbf{a}_{k_a}\mathbf{1}^\intercal)\right], \text{diag}(\boldsymbol{\tau}_{k_\tau}), \mathbf{I}_T\right).$$

The extension to varying length time series is straightforward by considering different time series lengths $T_n$ for each individual.

We specify the base distributions for facet parameters as conjugate priors. That is,

$$\begin{aligned} G_0^{(\mathbf{a})} &\sim \mathcal{N}_D(\boldsymbol{\mu}^{(a)}, \tau^{(a)}\mathbf{I}) \\ G_0^{(\mathbf{B})} &\sim \mathcal{MN}_{D,D}(\mathbf{M}^{(\mathbf{B})}, \mathrm{diag}(\tau^{(\mathbf{B})}), \mathbf{I}) \\ G_0^{(\tau)} &\sim \mathrm{Gamma}(\lambda_1^{(\tau)}, \lambda_2^{(\tau)}) \text{ for } \tau_d,\ d = 1, \ldots, D \end{aligned} \tag{B.5}$$

# C ADDITIONAL INFORMATION ON VARIATIONAL INFERENCE

Variational inference focuses on minimizing the Kullback-Leibler (KL) divergence [Kullback and Leibler, 1951] between a variational distribution, denoted as $q(\boldsymbol{\Theta})$, and the true posterior distribution $p(\boldsymbol{\Theta} \mid \mathbf{Y})$. Specifically, let $q_{\boldsymbol{\nu}}(\boldsymbol{\Theta})$ be a family of distributions parameterized by variational parameters $\boldsymbol{\nu}$. The objective is to minimize the KL divergence between $q_{\boldsymbol{\nu}}(\boldsymbol{\Theta})$ and $p(\boldsymbol{\Theta} \mid \mathbf{Y})$, given by:

$$D_{\mathrm{KL}}(q_{\boldsymbol{\nu}}(\boldsymbol{\Theta}) \,\|\, p(\boldsymbol{\Theta} \mid \mathbf{Y})) = \mathbb{E}_q[\log q_{\boldsymbol{\nu}}(\boldsymbol{\Theta})] - \mathbb{E}_q[\log p(\boldsymbol{\Theta}, \mathbf{Y})] + \log p(\mathbf{Y}).$$

Since this term is constant with respect to the variational parameters, it is equivalent to maximizing a lower bound on the log model evidence $\log p(\mathbf{Y})$, referred to as the evidence lower bound (ELBO). The generic ELBO for MMM is:

$$\begin{aligned} \text{ELBO} &= \mathbb{E}_q[\log p(\boldsymbol{\Theta}, \mathbf{Y})] - \mathbb{E}_q[\log q(\boldsymbol{\Theta})] \\ &= \sum_{f=1}^{F} \sum_{k_f=1}^{\infty} \mathbb{E}_q[\log p(\boldsymbol{\theta}_{k_f}^{(f)})] + \sum_{f=1}^{F} \mathbb{E}_q[\log p(\mathbf{v}^{(f)} \mid \alpha^{(f)})] + \sum_{n=1}^{N} \sum_{f=1}^{F} \mathbb{E}_q[\log p(z_n^{(f)} \mid \mathbf{v}^{(f)})] \\ &\quad + \sum_{n=1}^{N} \mathbb{E}_q[\log p(\mathbf{y}_n \mid \mathbf{z}_n, \{\boldsymbol{\theta}_{k_f}^{(f)}\}_{f=1}^{F})] - \mathbb{E}_q[\log q(\boldsymbol{\Theta})]. \end{aligned} \tag{C.6}$$

The cluster assignments for a sample across all facets are collectively encoded as $\mathbf{z}_n = (z_n^{(1)}, \ldots, z_n^{(F)})$. Under the stick-breaking construction, facet parameters are sampled as $\boldsymbol{\theta}_{k_f}^{(f)} \sim G_0^{(f)}$, with stick lengths $v_{k_f}^{(f)} \sim \mathrm{Beta}(1, \alpha^{(f)})$ determining probabilities $\pi_{k_f}^{(f)}(\mathbf{v}^{(f)}) = v_{k_f}^{(f)} \prod_{i=1}^{k_f-1}(1 - v_i^{(f)})$. We also place Gamma prior to $\alpha^{(f)}$: $\alpha^{(f)} \sim \mathrm{Gamma}(s_1^{(f)}, s_2^{(f)})$. Thus, the generic variational distributions for MMM are:

$$\begin{aligned} \boldsymbol{\theta}_k^{(f)} &\sim p(\boldsymbol{\theta}_k^{(f)} \mid \boldsymbol{\lambda}_k^{(f)^*}) \\ v_k^{(f)} &\sim \mathrm{Beta}(\alpha_{k1}^{(f)^*}, \alpha_{k2}^{(f)^*}) \\ z_n^{(f)} &\sim \mathrm{Cat}(\boldsymbol{\pi}_n^{(f)^*}) \\ \alpha^{(f)} &\sim \mathrm{Gamma}(s_1^{(f)^*}, s_2^{(f)^*}). \end{aligned} \tag{C.7}$$

It can be shown that the generic update rules for variational parameters can be accomplished by computing the following equations ( Supplement C.1):

$$\begin{aligned} &\boldsymbol{\lambda}_k^{(f)^*} = \mathbb{E}_q[g(\boldsymbol{\Theta}_{-\boldsymbol{\theta}_k^{(f)}}, \mathbf{Y})] \\ &\alpha_{k1}^{(f)^*} = 1 + \sum_{n=1}^{N} \pi_{nk}^{(f)^*};\ \alpha_{k2}^{(f)^*} = \frac{s_1^{(f)^*}}{s_2^{(f)^*}} + \sum_{n=1}^{N} \sum_{i=k+1}^{\ell} \pi_{ni}^{(f)^*} \\ &\pi_{nk}^{(f)^*} \propto \exp\left( \mathbb{E}_q[\log v_k^{(f)}] + \sum_{i=1}^{k-1} \mathbb{E}_q[\log(1 - v_i^{(f)})] + S_{nk}^{(f)} \right) \\ &s_1^{(f)^*} = s_1^{(f)} + \ell - 1;\ s_2^{(f)^*} = s_2^{(f)} - \sum_{k=1}^{\ell-1} \mathbb{E}_q[\log(1 - v_k^{(f)})] \end{aligned} \tag{C.8}$$

where $g(\mathbf{\Theta}_{-\boldsymbol{\theta}_k^{(f)}}, \mathbf{Y})$ are the parameters of the distribution for $\boldsymbol{\theta}_k^{(f)}$ when conditioning on the remaining latent variables and the observations i.e. $p(\boldsymbol{\theta}_k^{(f)} \mid \mathbf{\Theta}_{-\boldsymbol{\theta}_k^{(f)}}, \mathbf{Y})$. The update for $v_k^{(f)}$ is independent of model specification while $S_{nk}^{(f)}$ depends on the likelihood and different facets. Iterating these update rules optimizes ELBO in Equation C.6 with respect to the variational parameters defined in Equation C.7. The algorithm converges when the change in ELBO falls below a predefined threshold.

### C.1 DERIVATION OF UPDATE RULES FOR VARIATIONAL PARAMETERS IN NLG

We take the NLG model for example to give the full derivation steps. Likelihood function of $\mathbf{y}_n \mid \mathbf{z}_n, \mathbf{a}, \boldsymbol{\beta}, \boldsymbol{\tau}$:

$$\mathbf{y}_n \mid z_n^{(a)} = k_a, z_n^{(\boldsymbol{\beta})} = k_{\boldsymbol{\beta}}, z_n^{(\tau)} = k_\tau \quad \sim \quad \mathcal{N}_T(a_{k_a} + \boldsymbol{\beta}_{k_{\boldsymbol{\beta}}} \mathcal{B}(\mathbf{t}), \tau_{k_\tau} \mathbf{I}) \in \mathbb{R}^{1 \times T}$$

where $\mathcal{B}(\mathbf{t}) \in \mathbb{R}^{L \times T}$ is the basis spline matrix of order $p$ and $\tau_{k_\tau} \mathbf{I}$ is the precision matrix with scalar precision parameter.

Conjugate priors and base distributions on latent variables $\mathbf{\Theta}$ where we consider assigning priors to the concentration parameter in the beta distribution:

$$\begin{aligned}
a_{k_a} &\sim \mathcal{N}(\mu^{(a)}, \tau^{(a)}) \text{ for } k_a = 1, \ldots, \infty \text{ and scalar precision parameter } \tau^{(a)} \\
\boldsymbol{\beta}_{k_{\boldsymbol{\beta}}} &\sim \mathcal{N}_L(\boldsymbol{\mu}^{(\boldsymbol{\beta})}, \tau^{(\boldsymbol{\beta})} \mathbf{I}) \text{ for } k_{\boldsymbol{\beta}} = 1, \ldots, \infty \text{ and scalar precision parameter } \tau^{(\boldsymbol{\beta})} \\
\tau_{k_\tau} &\sim \text{Gamma}(\lambda_1^{(\tau)}, \lambda_2^{(\tau)}) \text{ for } k_\tau = 1, \ldots, \infty \\
z_n^{(a)} &\sim \text{Cat}(\boldsymbol{\pi}^{(a)}(\mathbf{v}^{(a)})) \\
z_n^{(\boldsymbol{\beta})} &\sim \text{Cat}(\boldsymbol{\pi}^{(\boldsymbol{\beta})}(\mathbf{v}^{(\boldsymbol{\beta})})) \\
z_n^{(\tau)} &\sim \text{Cat}(\boldsymbol{\pi}^{(\tau)}(\mathbf{v}^{(\tau)})) \\
v_{k_a}^{(a)} &\sim \text{Beta}(1, \alpha^{(a)}) \text{ for } k_a = 1, \ldots, \infty \\
v_{k_{\boldsymbol{\beta}}}^{(\boldsymbol{\beta})} &\sim \text{Beta}(1, \alpha^{(\boldsymbol{\beta})}) \text{ for } k_{\boldsymbol{\beta}} = 1, \ldots, \infty \\
v_{k_\tau}^{(\tau)} &\sim \text{Beta}(1, \alpha^{(\tau)}) \text{ for } k_\tau = 1, \ldots, \infty \\
\alpha^{(a)} &\sim \text{Gamma}(s_1^{(a)}, s_2^{(a)}) \\
\alpha^{(\boldsymbol{\beta})} &\sim \text{Gamma}(s_1^{(\boldsymbol{\beta})}, s_2^{(\boldsymbol{\beta})}) \\
\alpha^{(\tau)} &\sim \text{Gamma}(s_1^{(\tau)}, s_2^{(\tau)})
\end{aligned}$$

Assume the variational distribution (i.e. approximate posterior distribution) with truncation level $\ell$ by considering truncated stick-breaking representations:

$$\begin{aligned}
q(a_k) &\sim \mathcal{N}({\mu_k^{(a)}}^*, {\tau_k^{(a)}}^*) \\
q(\boldsymbol{\beta}_k) &\sim \mathcal{N}_L({\boldsymbol{\mu}_k^{(\boldsymbol{\beta})}}^*, {\mathbf{\Lambda}_k}^*) \\
q(\tau_k) &\sim \text{Gamma}({\lambda_{k1}^{(\tau)}}^*, {\lambda_{k2}^{(\tau)}}^*) \\
q(z_n^{(f)}) &\sim \text{Cat}({\boldsymbol{\pi}_n^{(f)}}^*) \text{ for facets } a, \boldsymbol{\beta}, \tau \\
q(v_k^{(f)}) &\sim \text{Beta}({\boldsymbol{\alpha}_k^{(f)}}^*) \text{ for facets } a, \boldsymbol{\beta}, \tau \\
q(\alpha^{(f)}) &\sim \text{Gamma}({\mathbf{s}^{(f)}}^*) \text{ for facets } a, \boldsymbol{\beta}, \tau
\end{aligned}$$

Therefore, the equation for the joint factorized variational distribution is as follows:

$$\begin{aligned}
q(\mathbf{\Theta}) = &\prod_{k=1}^{\ell} \{q(a_k \mid {\mu_k^{(a)}}^*, {\tau_k^{(a)}}^*) q(\boldsymbol{\beta}_k \mid {\boldsymbol{\mu}_k^{(\boldsymbol{\beta})}}^*, {\mathbf{\Lambda}_k}^*) q(\tau_k \mid {\boldsymbol{\lambda}_k^{(\tau)}}^*)\} \\
&\times \prod_{k=1}^{\ell-1} \{q(v_k^{(a)} \mid {\boldsymbol{\alpha}_k^{(a)}}^*) q(v_k^{(\boldsymbol{\beta})} \mid {\boldsymbol{\alpha}_k^{(\boldsymbol{\beta})}}^*) q(v_k^{(\tau)} \mid {\boldsymbol{\alpha}_k^{(\tau)}}^*)\}
\end{aligned}$$

$$\times \prod_{n=1}^{N} \{q(z_n^{(a)} \mid {\boldsymbol{\pi}_n^{(a)}}^*) q(z_n^{(\boldsymbol{\beta})} \mid {\boldsymbol{\pi}_n^{(\boldsymbol{\beta})}}^*) q(z_n^{(\tau)} \mid {\boldsymbol{\pi}_n^{(\tau)}}^*)\}$$
$$\times q(\alpha^{(a)} \mid {\mathbf{s}^{(a)}}^*) q(\alpha^{(\boldsymbol{\beta})} \mid {\mathbf{s}^{(\boldsymbol{\beta})}}^*) q(\alpha^{(\tau)} \mid {\mathbf{s}^{(\tau)}}^*)$$

We then derive the true conditional posterior distributions for each parameter and the corresponding update rules for variational parameters. First, the joint probability of $\boldsymbol{\Theta}$ and $\mathbf{Y}$ is as follows:

$$\begin{aligned}
p(\boldsymbol{\Theta}, \mathbf{Y}) = &\prod_{n=1}^{N} \left\{ p(\mathbf{y}_n \mid \mathbf{z}_n, \mathbf{a}, \boldsymbol{\beta}, \boldsymbol{\tau}) p(z_n^{(a)} \mid \boldsymbol{\pi}^{(a)}(\mathbf{v}^{(a)})) p(z_n^{(\boldsymbol{\beta})} \mid \boldsymbol{\pi}^{(\boldsymbol{\beta})}(\mathbf{v}^{(\boldsymbol{\beta})})) p(z_n^{(\tau)} \mid \boldsymbol{\pi}^{(\tau)}(\mathbf{v}^{(\tau)})) \right\} \\
&\times \prod_{k=1}^{\infty} \{ p(v_k^{(a)} \mid 1, \alpha^{(a)}) p(v_k^{(\boldsymbol{\beta})} \mid 1, \alpha^{(\boldsymbol{\beta})}) p(v_k^{(\tau)} \mid 1, \alpha^{(\tau)}) \} \\
&\times \prod_{k=1}^{\infty} \{ p(a_k \mid \mu^{(a)}, \tau^{(a)}) p(\boldsymbol{\beta}_k \mid \boldsymbol{\mu}^{(\boldsymbol{\beta})}, \tau^{(\boldsymbol{\beta})} \mathbf{I}) p(\tau_k \mid \lambda_1^{(\tau)}, \lambda_2^{(\tau)}) \} \\
&\times p(\alpha^{(a)} \mid s_1^{(a)}, s_2^{(a)}) p(\alpha^{(\boldsymbol{\beta})} \mid s_1^{(\boldsymbol{\beta})}, s_2^{(\boldsymbol{\beta})}) p(\alpha^{(\tau)} \mid s_1^{(\tau)}, s_2^{(\tau)})
\end{aligned}$$

The ELBO is expressed as:

$$\begin{aligned}
\text{ELBO}_{NLG} = &\mathbb{E}_q[\log p(\mathbf{a})] + \mathbb{E}_q[\log p(\boldsymbol{\beta})] + \mathbb{E}_q[\log p(\boldsymbol{\tau})] \\
&+ \mathbb{E}_q[\log p(\mathbf{v}^{(a)} \mid \alpha^{(a)})] + \mathbb{E}_q[\log p(\mathbf{v}^{(\boldsymbol{\beta})} \mid \alpha^{(\boldsymbol{\beta})})] + \mathbb{E}_q[\log p(\mathbf{v}^{(\tau)} \mid \alpha^{(\tau)})] \\
&+ \sum_{n=1}^{N} (\mathbb{E}_q[\log p(z_n^{(a)} \mid \mathbf{v}^{(a)})] + \mathbb{E}_q[\log p(z_n^{(\boldsymbol{\beta})} \mid \mathbf{v}^{(\boldsymbol{\beta})})] + \mathbb{E}_q[\log p(z_n^{(\tau)} \mid \mathbf{v}^{(\tau)})]) \\
&+ \mathbb{E}_q[\log p(\alpha^{(a)} \mid \mathbf{s}^{(a)})] + \mathbb{E}_q[\log p(\alpha^{(\boldsymbol{\beta})} \mid \mathbf{s}^{(\boldsymbol{\beta})})] + \mathbb{E}_q[\log p(\alpha^{(\tau)} \mid \mathbf{s}^{(\tau)})] \\
&+ \sum_{n=1}^{N} \mathbb{E}_q[\log p(\mathbf{y}_n \mid \mathbf{z}_n, \mathbf{a}, \boldsymbol{\beta}, \boldsymbol{\tau})] - \mathbb{E}_q[\log q(\boldsymbol{\Theta})].
\end{aligned} \tag{C.9}$$

The terms in the third row using indicator random variables $z_n$ in Equation C.9 can be rewritten as:

$$\begin{aligned}
\mathbb{E}_q[\log p(z_n \mid \mathbf{v})] &= \mathbb{E}_q \left[ \log \left( \prod_{k=1}^{\infty} (1 - v_k)^{\mathbf{1}[z_n > k]} {v_k}^{\mathbf{1}[z_n = k]} \right) \right] \\
&= \sum_{k=1}^{\infty} q(z_n > k) \mathbb{E}_q[\log(1 - v_k)] + q(z_n = k) \mathbb{E}_q[\log v_k] \\
&= \sum_{k=1}^{\ell} q(z_n > k) \mathbb{E}_q[\log(1 - v_k)] + q(z_n = k) \mathbb{E}_q[\log v_k].
\end{aligned}$$

Recall that $\mathbb{E}_q[\log(1 - v_k)] = 0$ and $q(z_n > \ell) = 0$ and we know:

$$\begin{aligned}
q(z_n = k) &= \pi_{nk}^* \\
q(z_n > k) &= \sum_{i=k+1}^{\ell} \pi_{ni}^* \\
\mathbb{E}_q[\log v_k] &= \boldsymbol{\Psi}(\alpha_{k1}^*) - \boldsymbol{\Psi}(\alpha_{k1}^* + \alpha_{k2}^*) \\
\mathbb{E}_q[\log(1 - v_k)] &= \boldsymbol{\Psi}(\alpha_{k2}^*) - \boldsymbol{\Psi}(\alpha_{k1}^* + \alpha_{k2}^*)
\end{aligned}$$

where the digamma function, denoted by $\boldsymbol{\Psi}$, arises from the derivative of the log normalization factor in the beta distribution. Note that this generic derivation does not rely on a particular model.

### C.1.1 Parameters that Do Not Depend on Particular Model

**For $\alpha^{(f)}$ of the Beta distribution**:

$$
\begin{aligned}
p(\alpha^{(f)} \mid \boldsymbol{\Theta}_{-\alpha^{(f)}}, \mathbf{Y}) &\propto p(\alpha^{(f)} \mid s_1^{(a)}, s_2^{(a)}) \prod_{k=1}^{\infty} p(v_k^{(f)} \mid 1, \alpha^{(f)}) \\
&\propto {\alpha^{(f)}}^{s_1^{(a)}-1} \exp\{-s_2^{(a)}\alpha^{(f)}\} \prod_{k=1}^{\infty} \alpha^{(f)}(1-v_k^{(f)})^{\alpha^{(f)}-1} \\
&\propto {\alpha^{(f)}}^{s_1^{(f)}-1} {\alpha^{(f)}}^{\max(k)} \exp\{-s_2^{(f)}\alpha^{(f)}\} \prod_{k=1}^{\infty} \exp\{(\alpha^{(f)}-1)\log(1-v_k^{(f)})\} \\
&\propto {\alpha^{(f)}}^{s_1^{(f)}+\max(k)-1} \exp\{-s_2^{(f)}\alpha^{(f)}\} \exp\left\{(\alpha^{(f)}-1)\sum_{k=1}^{\infty}\log(1-v_k^{(f)})\right\} \\
&\propto {\alpha^{(f)}}^{s_1^{(f)}+\max(k)-1} \exp\left\{-\left(s_2^{(f)} - \sum_{k=1}^{\infty}\log(1-v_k^{(f)})\right)\alpha^{(f)}\right\}
\end{aligned}
$$

Thus, ${s_1^{(f)}}^* = s_1^{(f)} + \ell - 1$ and ${s_2^{(f)}}^* = s_2^{(f)} - \sum_{k=1}^{\ell-1} \mathbb{E}_q[\log(1-v_k^{(f)})]$.

**The true conditional distribution for $v_k^{(f)}$ is**:

$$
\begin{aligned}
p(v_k^{(f)} \mid \boldsymbol{\Theta}_{-v_k^{(f)}}, \mathbf{Y}) &\propto p(v_k^{(f)} \mid 1, \alpha^{(f)}) \prod_{n=1}^{N} p(z_n^{(f)} \mid \boldsymbol{\pi}^{(f)}(\mathbf{v}^{(f)})) \\
&\propto \exp\left\{(\alpha^{(f)}-1)\log(1-v_k^{(f)}) + \sum_{n=1}^{N}\log\left(\prod_{k=1}^{\infty}(1-v_k^{(f)})^{\mathbf{1}[z_n^{(f)}>k]} {v_k^{(f)}}^{\mathbf{1}[z_n^{(f)}=k]}\right)\right\} \\
&\propto \exp\left\{(\alpha^{(f)}-1)\log(1-v_k^{(f)}) + \sum_{n=1}^{N}\{\mathbf{1}[z_n^{(f)}>k]\log(1-v_k^{(f)}) + \mathbf{1}[z_n^{(f)}=k]\log v_k^{(f)}\}\right\} \\
&\propto \exp\left\{\sum_{n=1}^{N}\mathbf{1}[z_n^{(f)}=k]\log v_k^{(f)} + (\alpha^{(f)} + \sum_{n=1}^{N}\mathbf{1}[z_n^{(f)}>k] - 1)\log(1-v_k^{(f)})\right\}
\end{aligned}
$$

Thus, ${\alpha_{k1}^{(f)}}^* = 1 + \sum_{n=1}^{N} {\pi_{nk}^{(f)}}^*$ and ${\alpha_{k2}^{(f)}}^* = \frac{{s_1^{(f)}}^*}{{s_2^{(f)}}^*} + \sum_{n=1}^{N}\sum_{i=k+1}^{\ell} {\pi_{ni}^{(f)}}^*$.

### C.1.2 Facets Parameters Specific for NLG

**For intercept $a_k$**:

$$
\begin{aligned}
p(a_k \mid \boldsymbol{\Theta}_{-a_k}, \mathbf{Y}) &\propto p(a_k \mid \mu^{(a)}, \tau^{(a)}) \prod_{n=1}^{N} p(\mathbf{y}_n \mid \mathbf{z}_n, \mathbf{a}, \boldsymbol{\beta}, \boldsymbol{\tau})^{\mathbf{1}[z_n^{(a)}=k]} \\
&\propto \exp\left(\frac{\tau^{(a)}(a_k-\mu^{(a)})^2}{-2}\right) \\
&\times \prod_{n=1}^{N}\left\{\exp\left(\frac{\tau_{z_n^{(\tau)}}(a_{z_n^{(a)}} - \mathbf{m}_{z_n^{(\boldsymbol{\beta})}})(a_{z_n^{(a)}} - \mathbf{m}_{z_n^{(\boldsymbol{\beta})}})^{\intercal}}{-2}\right)\right\}^{\mathbf{1}[z_n^{(a)}=k]} \\
&\quad \text{where } \mathbf{m}_{z_n^{(\boldsymbol{\beta})}} = \mathbf{y}_n - \boldsymbol{\beta}_{z_n^{(\boldsymbol{\beta})}}\mathcal{B} \in \mathbb{R}^{1\times T} \\
&\propto \exp\left(\frac{\tau^{(a)}(a_k-\mu^{(a)})^2 + \sum_{n=1}^{N}\{\mathbf{1}[z_n^{(a)}=k]\tau_{z_n^{(\tau)}}(T a_{z_n^{(a)}}^2 - 2a_{z_n^{(a)}}\sum \mathbf{m}_{z_n^{(\boldsymbol{\beta})}} + \left\|\mathbf{m}_{z_n^{(\boldsymbol{\beta})}}\right\|^2)\}}{-2}\right)
\end{aligned}
$$

$$\propto \exp\left(\frac{(\tau^{(a)} + T\sum_{n=1}^{N}\{\mathbf{1}[z_n^{(a)}=k]\tau_{z_n^{(\tau)}}\})a_k^2 - 2(\tau^{(a)}\mu^{(a)} + \sum_{n=1}^{N}\{\mathbf{1}[z_n^{(a)}=k]\tau_{z_n^{(\tau)}}\sum \mathbf{m}_{z_n^{(\boldsymbol{\beta})}}\})a_k}{-2}\right)$$

$$\begin{aligned} q(a_k) &\propto \exp\{\mathbb{E}_{q(\boldsymbol{\Theta}_{-a_k})}[\log p(a_k \mid \boldsymbol{\Theta}_{-a_k}, \mathbf{Y})]\} \\ &\propto \exp\left\{\frac{\tau^{(a)} + T\mathbb{E}_q\left[\sum_{n=1}^{N}\{\mathbf{1}[z_n^{(a)}=k]\tau_{z_n^{(\tau)}}\}\right])a_k^2 - 2(\tau^{(a)}\mu^{(a)} + \mathbb{E}_q\left[\sum_{n=1}^{N}\{\mathbf{1}[z_n^{(a)}=k]\tau_{z_n^{(\tau)}}\sum \mathbf{m}_{z_n^{(\boldsymbol{\beta})}}\}\right])a_k}{-2}\right\} \end{aligned}$$

where we have $\mathbb{E}_q[\tau_{z_n^{(\tau)}}] = \sum_{j=1}^{\ell} q(z_n^{(\tau)}=j)\mathbb{E}_q[\tau_j] = \sum_{j=1}^{\ell} {\pi_{nj}^{(\tau)}}^* \frac{{\lambda_{j1}^{(\tau)}}^*}{{\lambda_{j2}^{(\tau)}}^*}$ and $\mathbb{E}_q[\boldsymbol{\beta}_{z_n^{(\boldsymbol{\beta})}}] = \sum_{j=1}^{\ell} {\pi_{nj}^{(\boldsymbol{\beta})}}^* {\boldsymbol{\mu}_j^{(\boldsymbol{\beta})}}^*$. Thus,

$${\tau_k^{(a)}}^* = \tau^{(a)} + T\sum_{n=1}^{N}\left\{{\pi_{nk}^{(a)}}^*\sum_{j=1}^{\ell}\left({\pi_{nj}^{(\tau)}}^*\frac{{\lambda_{j1}^{(\tau)}}^*}{{\lambda_{j2}^{(\tau)}}^*}\right)\right\}$$

and

$${\mu_k^{(a)}}^* = \frac{\tau^{(a)}\mu^{(a)} + \sum_{n=1}^{N}\left\{{\pi_{nk}^{(a)}}^*\left(\sum_{j=1}^{\ell}{\pi_{nj}^{(\tau)}}^*\frac{{\lambda_{j1}^{(\tau)}}^*}{{\lambda_{j2}^{(\tau)}}^*}\right)\left(\sum_T \mathbf{y}_n - \sum_T\left[\left(\sum_{j=1}^{\ell}{\pi_{nj}^{(\boldsymbol{\beta})}}^*{\boldsymbol{\mu}_j^{(\boldsymbol{\beta})}}^*\right)\mathcal{B}\right]\right)\right\}}{{\tau_k^{(a)}}^*}$$

**For coefficient row vector $\boldsymbol{\beta}_k$:**

$$\begin{aligned} p(\boldsymbol{\beta}_k \mid \boldsymbol{\Theta}_{-\boldsymbol{\beta}_k}, \mathbf{Y}) &\propto p(\boldsymbol{\beta}_k \mid \boldsymbol{\mu}^{(\boldsymbol{\beta})}, \tau^{(\boldsymbol{\beta})}\mathbf{I})\prod_{n=1}^{N} p(\mathbf{y}_n \mid \mathbf{z}_n, \mathbf{a}, \boldsymbol{\beta}, \boldsymbol{\tau})^{\mathbf{1}[z_n^{(\boldsymbol{\beta})}=k]} \\ &\propto \exp\left(\frac{\tau^{(\boldsymbol{\beta})}(\boldsymbol{\beta}_k - \boldsymbol{\mu}^{(\boldsymbol{\beta})})(\boldsymbol{\beta}_k - \boldsymbol{\mu}^{(\boldsymbol{\beta})})^\intercal}{-2}\right) \\ &\times \prod_{n=1}^{N}\left\{\exp\left(\frac{\tau_{z_n^{(\tau)}}(\boldsymbol{\beta}_k\mathcal{B} + a_{z_n^{(a)}} - \mathbf{y}_n)(\boldsymbol{\beta}_k\mathcal{B} + a_{z_n^{(a)}} - \mathbf{y}_n)^\intercal}{-2}\right)\right\}^{\mathbf{1}[z_n^{(\boldsymbol{\beta})}=k]} \\ &\propto \exp\left(\frac{\tau^{(\boldsymbol{\beta})}(\boldsymbol{\beta}_k\boldsymbol{\beta}_k^\intercal - 2\boldsymbol{\beta}_k\boldsymbol{\mu}^{(\boldsymbol{\beta})\intercal}) + \sum_{n=1}^{N}\{\mathbf{1}[z_n^{(\boldsymbol{\beta})}=k]\tau_{z_n^{(\tau)}}(\boldsymbol{\beta}_k\mathcal{B} - \mathbf{m}_{z_n^{(a)}})(\boldsymbol{\beta}_k\mathcal{B} - \mathbf{m}_{z_n^{(a)}})^\intercal\}}{-2}\right) \end{aligned}$$

where $\mathbf{m}_{z_n^{(a)}} = \mathbf{y}_n - a_{z_n^{(a)}} \in \mathbb{R}^{1\times T}$

$$\begin{aligned} &\propto \exp\left(\frac{1}{-2}\left[\boldsymbol{\beta}_k(\tau^{(\boldsymbol{\beta})}\mathbf{I} + \sum_{n=1}^{N}\{\mathbf{1}[z_n^{(\boldsymbol{\beta})}=k]\tau_{z_n^{(\tau)}}\mathcal{B}\mathcal{B}^\intercal\})\boldsymbol{\beta}_k^\intercal\right.\right. \\ &\quad \left.\left. -\ 2\boldsymbol{\beta}_k(\tau^{(\boldsymbol{\beta})}\boldsymbol{\mu}^{(\boldsymbol{\beta})\intercal} + \sum_{n=1}^{N}\{\mathbf{1}[z_n^{(\boldsymbol{\beta})}=k]\tau_{z_n^{(\tau)}})\mathcal{B}\mathbf{m}_{z_n^{(a)}}^\intercal\right]\right) \end{aligned}$$

Thus,

$$\boldsymbol{\Lambda}_k^* = \tau^{(\boldsymbol{\beta})}\mathbf{I} + \sum_{n=1}^{N}\left\{{\pi_{nk}^{(\boldsymbol{\beta})}}^*\left(\sum_{j=1}^{\ell}{\pi_{nj}^{(\tau)}}^*\frac{{\lambda_{j1}^{(\tau)}}^*}{{\lambda_{j2}^{(\tau)}}^*}\right)\mathcal{B}\mathcal{B}^\intercal\right\} \in \mathbb{R}^{L\times L}$$

and

$${\boldsymbol{\mu}_k^{(\boldsymbol{\beta})}}^* = \left\{\tau^{(\boldsymbol{\beta})}\boldsymbol{\mu}^{(\boldsymbol{\beta})} + \sum_{n=1}^{N}\left\{{\pi_{nk}^{(\boldsymbol{\beta})}}^*\left(\sum_{j=1}^{\ell}{\pi_{nj}^{(\tau)}}^*\frac{{\lambda_{j1}^{(\tau)}}^*}{{\lambda_{j2}^{(\tau)}}^*}\right)\left(\mathbf{y}_n - \left(\sum_{j=1}^{\ell}{\pi_{nj}^{(a)}}^*{\mu_j^{(a)}}^*\right)\right)\mathcal{B}^\intercal\right\}\right\}({\boldsymbol{\Lambda}_k}^*)^{-1} \in \mathbb{R}^{1\times L}$$

where $\mathbb{E}_q[a_{z_n^{(a)}}] = \sum_{j=1}^{\ell}{\pi_{nj}^{(a)}}^*{\mu_j^{(a)}}^*$.

**For precision scalar $\tau_k$**:

$$
\begin{aligned}
p(\tau_k \mid \boldsymbol{\Theta}_{-\tau_k}, \mathbf{Y}) &\propto p(\tau_k \mid \lambda_1^{(\tau)}, \lambda_2^{(\tau)}) \prod_{n=1}^{N} p(\mathbf{y}_n \mid \mathbf{z}_n, \mathbf{a}, \boldsymbol{\beta}, \boldsymbol{\tau})^{\mathbf{1}[z_n^{(\tau)}=k]} \\
&\propto \tau_k^{\lambda_1^{(\tau)}-1} \exp\{-\lambda_2^{(\tau)}\tau_k\} \\
&\times \tau_k^{\frac{T}{2}\sum_{n=1}^{N}\mathbf{1}[z_n^{(\tau)}=k]} \prod_{n=1}^{N} \left\{ \exp\left( \frac{\tau_{z_n^{(\tau)}}(\mathbf{y}_n - \boldsymbol{\mu}_{z_n^{(a)}, z_n^{(\boldsymbol{\beta})}})(\mathbf{y}_n - \boldsymbol{\mu}_{z_n^{(a)}, z_n^{(\boldsymbol{\beta})}})^\intercal}{-2} \right) \right\}^{\mathbf{1}[z_n^{(\tau)}=k]} \\
&\quad \text{where } \boldsymbol{\mu}_{z_n^{(a)}, z_n^{(\boldsymbol{\beta})}} = a_{z_n^{(a)}} + \boldsymbol{\beta}_{z_n^{(\boldsymbol{\beta})}}\mathcal{B} \in \mathbb{R}^{1\times T} \\
&\propto \tau_k^{\lambda_1^{(\tau)} + \frac{T}{2}\sum_{n=1}^{N}\mathbf{1}[z_n^{(\tau)}=k]-1} \\
&\times \exp\left( -\tau_k \left\{ \lambda_2^{(\tau)} + \frac{\sum_{n=1}^{N}\{\mathbf{1}[z_n^{(\tau)}=k](\mathbf{y}_n - \boldsymbol{\mu}_{z_n^{(a)}, z_n^{(\boldsymbol{\beta})}})(\mathbf{y}_n - \boldsymbol{\mu}_{z_n^{(a)}, z_n^{(\boldsymbol{\beta})}})^\intercal\}}{2} \right\} \right)
\end{aligned}
$$

Thus,

$$
\lambda_{k1}^{(\tau)^*} = \lambda_1^{(\tau)} + \frac{T}{2}\sum_{n=1}^{N} \pi_{nk}^{(\tau)^*}
$$

and

$$
\lambda_{k2}^{(\tau)^*} = \lambda_2^{(\tau)} + \frac{1}{2}\sum_{n=1}^{N} \left\{ \pi_{nk}^{(\tau)^*} \mathbb{E}_q \left[ \left\| \mathbf{y}_n - \boldsymbol{\beta}_{z_n^{(\boldsymbol{\beta})}}\mathcal{B} - a_{z_n^{(a)}} \right\|^2 \right] \right\}
$$

### C.1.3 Parameters for Cluster Assignments

**For $z_n^{(f)}$ of any facet**:

$$
p(z_n^{(f)} \mid \boldsymbol{\Theta}_{-z_n^{(f)}}, \mathbf{Y}) \propto p(z_n^{(f)} = k \mid \boldsymbol{\pi}^{(f)}(\mathbf{v}^{(f)})) p(\mathbf{y}_n \mid \mathbf{z}_n, \{\boldsymbol{\theta}_{k_f}^{(f)}\}_{f=1}^{F})
$$

Hence,

$$
\pi_{nk}^{(f)^*} \propto \exp\left( \mathbb{E}_q[\log v_k^{(f)}] + \sum_{i=1}^{k-1} \mathbb{E}_q[\log(1 - v_i^{(f)})] + S_{nk}^{(f)} \right)
$$

where $S_{nk}^{(f)}$ depends on the likelihood and different facets and

$$
\begin{aligned}
\mathbb{E}_q[\log v_k^{(f)}] &= \boldsymbol{\Psi}(\alpha_{k1}^{(f)^*}) - \boldsymbol{\Psi}(\alpha_{k1}^{(f)^*} + \alpha_{k2}^{(f)^*}) \\
\mathbb{E}_q[\log(1 - v_k^{(f)})] &= \boldsymbol{\Psi}(\alpha_{k2}^{(f)^*}) - \boldsymbol{\Psi}(\alpha_{k1}^{(f)^*} + \alpha_{k2}^{(f)^*})
\end{aligned}
$$

with the digamma function denoted by $\boldsymbol{\Psi}$.

**So $z_n^{(a)}$ of intercept**:

$$
\begin{aligned}
p(z_n^{(a)} = k \mid \boldsymbol{\Theta}_{-z_n^{(a)}}, \mathbf{Y}) &\propto p(z_n^{(a)} = k \mid \boldsymbol{\pi}^{(a)}(\mathbf{v}^{(a)})) p(\mathbf{y}_n \mid \mathbf{z}_n, \mathbf{a}, \boldsymbol{\beta}, \boldsymbol{\tau}) \\
&\propto v_k^{(a)} \prod_{i=1}^{k-1}(1 - v_i^{(a)}) \exp\left( \frac{\tau_{z_n^{(\tau)}}(\mathbf{y}_n - a_k - \boldsymbol{\beta}_{z_n^{(\boldsymbol{\beta})}}\mathcal{B})(\mathbf{y}_n - a_k - \boldsymbol{\beta}_{z_n^{(\boldsymbol{\beta})}}\mathcal{B})^\intercal}{-2} \right) \\
&\propto \exp\left( \log v_k^{(a)} + \sum_{i=1}^{k-1} \log(1 - v_i^{(a)}) + \frac{-1}{2}\tau_{z_n^{(\tau)}}(\mathbf{y}_n - a_k - \boldsymbol{\beta}_{z_n^{(\boldsymbol{\beta})}}\mathcal{B})(\mathbf{y}_n - a_k - \boldsymbol{\beta}_{z_n^{(\boldsymbol{\beta})}}\mathcal{B})^\intercal \right)
\end{aligned}
$$

Thus,

$$
\pi_{nk}^{(a)^*} \propto \exp\left( \mathbb{E}_q[\log v_k^{(a)}] + \sum_{i=1}^{k-1} \mathbb{E}_q[\log(1 - v_i^{(a)})] + S_{nk}^{(a)} \right)
$$

where

$$S_{nk}^{(a)} = \frac{-1}{2}\left(\sum_{j=1}^{\ell} {\pi_{nj}^{(\tau)}}^* \frac{\lambda_{j1}^{(\tau)*}}{\lambda_{j2}^{(\tau)*}}\right)\left\|\mathbf{y}_n - {\mu_k^{(a)}}^* - \left(\sum_{j=1}^{\ell} {\pi_{nj}^{(\boldsymbol{\beta})}}^* {\boldsymbol{\mu}_j^{(\boldsymbol{\beta})}}^*\right)\mathcal{B}\right\|^2$$

**We can obtain a similar result for** ${\pi_{nk}^{(\boldsymbol{\beta})}}^* \propto \exp\left(\mathbb{E}_q[\log v_k^{(\boldsymbol{\beta})}] + \sum_{i=1}^{k-1}\mathbb{E}_q[\log(1-v_i^{(\boldsymbol{\beta})})] + S_{nk}^{(\boldsymbol{\beta})}\right)$ where

$$S_{nk}^{(\boldsymbol{\beta})} = \frac{-1}{2}\left(\sum_{j=1}^{\ell} {\pi_{nj}^{(\tau)}}^* \frac{\lambda_{j1}^{(\tau)*}}{\lambda_{j2}^{(\tau)*}}\right)\left\|\mathbf{y}_n - \left(\sum_{j=1}^{\ell} {\pi_{nj}^{(a)}}^* {\mu_j^{(a)}}^*\right) - {\boldsymbol{\mu}_k^{(\boldsymbol{\beta})}}^*\mathcal{B}\right\|^2$$

**For** $z_n^{(\tau)}$ **of noise**:

$$\begin{aligned}
p(z_n^{(\tau)} = k \mid \boldsymbol{\Theta}_{-z_n^{(\tau)}}, \mathbf{Y}) &\propto p(z_n^{(\tau)} = k \mid \boldsymbol{\pi}^{(\tau)}(\mathbf{v}^{(\tau)}))p(\mathbf{y}_n \mid \mathbf{z}_n, \mathbf{a}, \boldsymbol{\beta}, \boldsymbol{\tau}) \\
&\propto v_k^{(\tau)}\prod_{i=1}^{k-1}(1-v_i^{(\tau)})\tau_k^{\frac{T}{2}}\exp\left(\frac{\tau_k(\mathbf{y}_n - a_{z_n^{(a)}} - \boldsymbol{\beta}_{z_n^{(\boldsymbol{\beta})}}\mathcal{B})(\mathbf{y}_n - a_{z_n^{(a)}} - \boldsymbol{\beta}_{z_n^{(\boldsymbol{\beta})}}\mathcal{B})^\intercal}{-2}\right) \\
&\propto \exp\left(\log v_k^{(\tau)} + \sum_{i=1}^{k-1}\log(1-v_i^{(\tau)}) + \frac{T}{2}\log\tau_k \right. \\
&\qquad \left. + \frac{-1}{2}\tau_k(\mathbf{y}_n - a_{z_n^{(a)}} - \boldsymbol{\beta}_{z_n^{(\boldsymbol{\beta})}}\mathcal{B})(\mathbf{y}_n - a_{z_n^{(a)}} - \boldsymbol{\beta}_{z_n^{(\boldsymbol{\beta})}}\mathcal{B})^\intercal\right)
\end{aligned}$$

Thus,

$${\pi^{(\tau)}}^*_{nk} \propto \exp\left(\mathbb{E}_q[\log v_k^{(\tau)}] + \sum_{i=1}^{k-1}\mathbb{E}_q[\log(1-v_i^{(\tau)})] + S_{nk}^{(\tau)}\right)$$

where

$$S_{nk}^{(\tau)} = \frac{T}{2}\log\frac{\lambda_{k1}^{(\tau)*}}{\lambda_{k2}^{(\tau)*}} + \frac{-1}{2}\frac{\lambda_{k1}^{(\tau)*}}{\lambda_{k2}^{(\tau)*}}\left\|\mathbf{y}_n - \left(\sum_{j=1}^{\ell} {\pi_{nj}^{(a)}}^* {\mu_j^{(a)}}^*\right) - \left(\sum_{j=1}^{\ell} {\pi_{nj}^{(\boldsymbol{\beta})}}^* {\boldsymbol{\mu}_j^{(\boldsymbol{\beta})}}^*\right)\mathcal{B}\right\|^2$$

### C.1.4 ELBO Computation

For ELBO in Equation C.9:

$$\begin{aligned}
&\mathbb{E}_q[\log p(\mathbf{y}_n \mid \mathbf{z}_n, \mathbf{a}, \boldsymbol{\beta}, \boldsymbol{\tau})] = -\frac{T}{2}\log 2\pi + \frac{T}{2}\mathbb{E}_q[\log\tau_{z_n^{(\tau)}}] - \frac{1}{2}\mathbb{E}_q[\tau_{z_n^{(\tau)}}]\left\{\mathbf{y}_n\mathbf{y}_n^\intercal - 2\sum_T(\mathbb{E}_q[a_{z_n^{(a)}}]\mathbf{y}_n) - 2\mathbb{E}_q[\boldsymbol{\beta}_{z_n^{(\boldsymbol{\beta})}}]\mathcal{B}\mathbf{y}_n^\intercal\right. \\
&\left. + T\mathbb{E}_q[a^2_{z_n^{(a)}}] + 2\sum_T(\mathbb{E}_q[a_{z_n^{(a)}}]\mathbb{E}_q[\boldsymbol{\beta}_{z_n^{(\boldsymbol{\beta})}}]\mathcal{B}) + \mathbb{E}_q[\boldsymbol{\beta}_{z_n^{(\boldsymbol{\beta})}}]\mathcal{B}\mathcal{B}^\intercal\mathbb{E}_q[\boldsymbol{\beta}_{z_n^{(\boldsymbol{\beta})}}^\intercal]\right\} \\
&\mathbb{E}_q[\log p(a_k \mid \mu^{(a)}, \tau^{(a)})] = -\log\sqrt{2\pi} + \frac{1}{2}\log\tau^{(a)} - \frac{1}{2}\tau^{(a)}\left(\frac{1}{{\tau_k^{(a)}}^*} + {\mu_k^{(a)}}^{*2} - 2\mu^{(a)}{\mu_k^{(a)}}^* + {\mu^{(a)}}^2\right) \\
&\mathbb{E}_q[\log q(a_k \mid {\mu_k^{(a)}}^*, {\tau_k^{(a)}}^*)] = -\log\sqrt{2\pi} + \frac{1}{2}\log{\tau_k^{(a)}}^* - \frac{1}{2}{\tau_k^{(a)}}^*\left(\frac{1}{{\tau_k^{(a)}}^*} + {\mu_k^{(a)}}^{*2} - 2{\mu_k^{(a)}}^*{\mu_k^{(a)}}^* + {\mu_k^{(a)}}^{*2}\right) \\
&\qquad\qquad = -\log\sqrt{2\pi} + \frac{1}{2}\log{\tau_k^{(a)}}^* - \frac{1}{2} \\
&\mathbb{E}_q[\log p(\boldsymbol{\beta}_k \mid \boldsymbol{\mu}^{(\boldsymbol{\beta})}, \tau^{(\boldsymbol{\beta})}] = -L\log\sqrt{2\pi} + \frac{L}{2}\log\tau^{(\boldsymbol{\beta})} - \frac{1}{2}\tau^{(\boldsymbol{\beta})}\left(\mathrm{tr}((\boldsymbol{\Lambda}_k{}^*)^{-1}) + \left\|{\boldsymbol{\mu}_k^{(\boldsymbol{\beta})}}^*\right\|^2 - 2\boldsymbol{\mu}^{(\boldsymbol{\beta})\intercal}{\boldsymbol{\mu}_k^{(\boldsymbol{\beta})}}^* + \left\|\boldsymbol{\mu}^{(\boldsymbol{\beta})}\right\|^2\right) \\
&\mathbb{E}_q[\log q(\boldsymbol{\beta}_k \mid {\boldsymbol{\mu}_k^{(\boldsymbol{\beta})}}^*, \boldsymbol{\Lambda}_k{}^*)] = -L\log\sqrt{2\pi} + \frac{1}{2}\log|\boldsymbol{\Lambda}_k{}^*| - \frac{L}{2} \\
&\mathbb{E}_q[\log p(\tau_k \mid \boldsymbol{\lambda}^{(\tau)})] = \lambda_1^{(\tau)}\log\lambda_2^{(\tau)} - \log\Gamma(\lambda_1^{(\tau)}) + (\lambda_1^{(\tau)} - 1)\left\{\boldsymbol{\Psi}({\lambda_{k1}^{(\tau)}}^*) - \log{\lambda_{k2}^{(\tau)}}^*\right\} - \lambda_2^{(\tau)}\frac{\lambda_{k1}^{(\tau)*}}{\lambda_{k2}^{(\tau)*}}
\end{aligned}$$

$$\mathbb{E}_q[\log q(\tau_k \mid \boldsymbol{\lambda}_k^{(\tau)^*})] = \lambda_{k1}^{(\tau)^*} \log \lambda_{k2}^{(\tau)^*} - \log \Gamma(\lambda_{k1}^{(\tau)^*}) + (\lambda_{k1}^{(\tau)^*} - 1)\left\{\boldsymbol{\Psi}(\lambda_{k1}^{(\tau)^*}) - \log \lambda_{k2}^{(\tau)^*}\right\} - \lambda_{k1}^{(\tau)^*}$$

$$\mathbb{E}_q[\log p(v_k^{(f)} \mid \alpha^{(f)})] = (\frac{s_1^{(f)^*}}{s_2^{(f)^*}} - 1)\{\boldsymbol{\Psi}(\alpha_{k2}^{(f)^*}) - \boldsymbol{\Psi}(\alpha_{k1}^{(f)^*} + \boldsymbol{\alpha}_{k2}^{(f)^*})\} - \{\boldsymbol{\Psi}(s_1^{(f)^*}) - \log s_1^{(f)^*}\}$$

$$\begin{aligned}\mathbb{E}_q[\log q(v_k^{(f)} \mid \boldsymbol{\alpha}_k^{(f)^*})] = {} & (\alpha_{k1}^{(f)^*} - 1)\{\boldsymbol{\Psi}(\alpha_{k1}^{(f)^*}) - \boldsymbol{\Psi}(\alpha_{k1}^{(f)^*} + \alpha_{k2}^{(f)^*})\} + (\alpha_{k2}^{(f)^*} - 1)\{\boldsymbol{\Psi}(\alpha_{k2}^{(f)^*}) - \boldsymbol{\Psi}(\alpha_{k1}^{(f)^*} + \alpha_{k2}^{(f)^*})\} \\ & - \log \mathcal{B}(\alpha_{k1}^{(f)^*}, \alpha_{k2}^{(f)^*})\end{aligned}$$

$$\mathbb{E}_q[\log p(z_n^{(f)} \mid \mathbf{v}^{(f)})] = \sum_{k=1}^{\ell}\left\{\left(\sum_{i=k+1}^{\ell} \pi_{ni}^{(f)^*}\right)\{\boldsymbol{\Psi}(\alpha_{k2}^{(f)^*}) - \boldsymbol{\Psi}(\alpha_{k1}^{(f)^*} + \boldsymbol{\alpha}_{k2}^{(f)^*})\} + \pi_{nk}^{(f)^*}\{\boldsymbol{\Psi}(\alpha_{k1}^{(f)^*}) - \boldsymbol{\Psi}(\alpha_{k1}^{(f)^*} + \alpha_{k2}^{(f)^*})\}\right\}$$

$$\mathbb{E}_q[\log q(z_n^{(f)} \mid \boldsymbol{\pi}_n^{(f)^*})] = \log \max_i \pi_{ni}^{(f)^*}$$

$$\mathbb{E}_q[\log p(\alpha^{(f)} \mid \mathbf{s}^{(a)})] = s_1^{(f)} \log s_2^{(f)} - \log \Gamma(s_1^{(f)}) + (s_1^{(f)} - 1)\{\boldsymbol{\Psi}(s_1^{(f)^*}) - \log s_1^{(f)^*}\} - s_2^{(f)} \frac{s_1^{(f)^*}}{s_2^{(f)^*}}$$

$$\mathbb{E}_q[\log q(\alpha^{(f)} \mid \mathbf{s}^{(a)^*})] = s_1^{(f)^*} \log s_1^{(f)^*} - \log \Gamma(s_1^{(f)^*}) + (s_1^{(f)^*} - 1)\{\boldsymbol{\Psi}(s_1^{(f)^*}) - \log s_1^{(f)^*}\} - s_1^{(f)^*}$$

## C.2 DERIVATION OF UPDATE RULES FOR VARIATIONAL PARAMETERS IN VAR

### C.2.1 Facets Parameters Specific for VAR

**For intercept vector $\mathbf{a}_k$:**

$$
\begin{aligned}
p(\mathbf{a}_k \mid \boldsymbol{\Theta}_{-\mathbf{a}_k}, \mathbf{Y}) &\propto p(\mathbf{a}_k \mid \boldsymbol{\mu}^{(a)}, \tau^{(a)}\mathbf{I}) \prod_{n=1}^{N} p(\mathbf{y}_n \mid \mathbf{z}_n, \mathbf{a}, \mathbf{B}, \boldsymbol{\tau})^{\mathbf{1}[z_n^{(a)}=k]} \\
&\propto \exp\left(\frac{\tau^{(a)}(\mathbf{a}_k - \boldsymbol{\mu}^{(a)})^\intercal(\mathbf{a}_k - \boldsymbol{\mu}^{(a)})}{-2}\right) \prod_{n=1}^{N}\left\{\exp\left(\frac{(\mathbf{y}_{n0} - \mathbf{a}_k)^\intercal \mathrm{diag}(\boldsymbol{\tau}_{z_n^{(\tau)}})(\mathbf{y}_{n0} - \mathbf{a}_k)}{-2}\right)\right\}^{\mathbf{1}[z_n^{(a)}=k]} \\
&\times \prod_{n=1}^{N}\prod_{t=1}^{T-1}\left\{\exp\left(\frac{(\mathbf{y}_{nt} - \mathbf{a}_k - \mathbf{B}_{z_n^{(\mathbf{B})}}\mathbf{y}_{n(t-1)} + \mathbf{B}_{z_n^{(\mathbf{B})}}\mathbf{a}_k)^\intercal \mathrm{diag}(\boldsymbol{\tau}_{z_n^{(\tau)}})(\mathbf{y}_{nt} - \mathbf{a}_k - \mathbf{B}_{z_n^{(\mathbf{B})}}\mathbf{y}_{n(t-1)} + \mathbf{B}_{z_n^{(\mathbf{B})}}\mathbf{a}_k)}{-2}\right)\right\}^{\mathbf{1}[z_n^{(a)}=k]} \\
&\propto \exp\left(\frac{\tau^{(a)}(\mathbf{a}_k - \boldsymbol{\mu}^{(a)})^\intercal(\mathbf{a}_k - \boldsymbol{\mu}^{(a)}) + \sum_{n=1}^{N}\{\mathbf{1}[z_n^{(a)}=k](\mathbf{y}_{n0} - \mathbf{a}_k)^\intercal \mathrm{diag}(\boldsymbol{\tau}_{z_n^{(\tau)}})(\mathbf{y}_{n0} - \mathbf{a}_k)\}}{-2}\right) \\
&\times \exp\left(\frac{\sum_{n=1}^{N}\mathbf{1}[z_n^{(a)}=k]\sum_{t=1}^{T-1}\{(\mathbf{y}_{nt} - \mathbf{a}_k - \mathbf{B}_{z_n^{(\mathbf{B})}}\mathbf{y}_{n(t-1)} + \mathbf{B}_{z_n^{(\mathbf{B})}}\mathbf{a}_k)^\intercal \mathrm{diag}(\boldsymbol{\tau}_{z_n^{(\tau)}})(\mathbf{y}_{nt} - \mathbf{a}_k - \mathbf{B}_{z_n^{(\mathbf{B})}}\mathbf{y}_{n(t-1)} + \mathbf{B}_{z_n^{(\mathbf{B})}}\mathbf{a}_k)\}}{-2}\right) \\
&\propto \exp\left(\frac{1}{-2}\left[\mathbf{a}_k^\intercal\left\{\tau^{(a)}\mathbf{I} + T\sum_{n=1}^{N}\mathbf{1}[z_n^{(a)}=k]\mathrm{diag}(\boldsymbol{\tau}_{z_n^{(\tau)}}) + (T-1)\sum_{n=1}^{N}\mathbf{1}[z_n^{(a)}=k]\mathbf{B}_{z_n^{(\mathbf{B})}}^\intercal \mathrm{diag}(\boldsymbol{\tau}_{z_n^{(\tau)}})\mathbf{B}_{z_n^{(\mathbf{B})}}\right.\right.\right. \\
&-(T-1)\sum_{n=1}^{N}\mathbf{1}[z_n^{(a)}=k]\mathbf{B}_{z_n^{(\mathbf{B})}}^\intercal \mathrm{diag}(\boldsymbol{\tau}_{z_n^{(\tau)}}) - (T-1)\sum_{n=1}^{N}\mathbf{1}[z_n^{(a)}=k]\mathrm{diag}(\boldsymbol{\tau}_{z_n^{(\tau)}})\mathbf{B}_{z_n^{(\mathbf{B})}}\Big\}\mathbf{a}_k \\
&- 2\left\{\tau^{(a)}\boldsymbol{\mu}^{(a)\intercal} + \sum_{n=1}^{N}\mathbf{1}[z_n^{(a)}=k]\mathbf{y}_{n0}^\intercal \mathrm{diag}(\boldsymbol{\tau}_{z_n^{(\tau)}}) + \sum_{n=1}^{N}\mathbf{1}[z_n^{(a)}=k]\{\sum_{t=1}^{T-1}\mathbf{y}_{nt}^\intercal\}\mathrm{diag}(\boldsymbol{\tau}_{z_n^{(\tau)}})\right. \\
&-\sum_{n=1}^{N}\mathbf{1}[z_n^{(a)}=k]\{\sum_{t=1}^{T-1}\mathbf{y}_{n(t-1)}^\intercal\}\mathbf{B}_{z_n^{(\mathbf{B})}}^\intercal \mathrm{diag}(\boldsymbol{\tau}_{z_n^{(\tau)}}) + \sum_{n=1}^{N}\mathbf{1}[z_n^{(a)}=k]\{\sum_{t=1}^{T-1}\mathbf{y}_{n(t-1)}^\intercal\}\mathbf{B}_{z_n^{(\mathbf{B})}}^\intercal \mathrm{diag}(\boldsymbol{\tau}_{z_n^{(\tau)}})\mathbf{B}_{z_n^{(\mathbf{B})}} \\
&\left.\left.\left.-\sum_{n=1}^{N}\mathbf{1}[z_n^{(a)}=k]\{\sum_{t=1}^{T-1}\mathbf{y}_{nt}^\intercal\}\mathrm{diag}(\boldsymbol{\tau}_{z_n^{(\tau)}})\mathbf{B}_{z_n^{(\mathbf{B})}}\right\}\mathbf{a}_k\right]\right)
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\Lambda_k^{(a)^*} &= \tau^{(a)}\mathbf{I} + T\sum_{n=1}^{N}\pi_{nk}^{(a)^*}\mathrm{diag}(\mathbb{E}_q[\boldsymbol{\tau}_{z_n^{(\tau)}}]) + (T-1)\sum_{n=1}^{N}\pi_{nk}^{(a)^*}\mathbb{E}_q[\mathbf{B}_{z_n^{(\mathbf{B})}}^\intercal]\mathrm{diag}(\mathbb{E}_q[\boldsymbol{\tau}_{z_n^{(\tau)}}])\mathbb{E}_q[\mathbf{B}_{z_n^{(\mathbf{B})}}] \\
&\quad - (T-1)\sum_{n=1}^{N}\pi_{nk}^{(a)^*}\mathbb{E}_q[\mathbf{B}_{z_n^{(\mathbf{B})}}^\intercal]\mathrm{diag}(\mathbb{E}_q[\boldsymbol{\tau}_{z_n^{(\tau)}}]) - (T-1)\sum_{n=1}^{N}\pi_{nk}^{(a)^*}\mathrm{diag}(\mathbb{E}_q[\boldsymbol{\tau}_{z_n^{(\tau)}}])\mathbb{E}_q[\mathbf{B}_{z_n^{(\mathbf{B})}}]
\end{aligned}
$$

and

$$
\begin{aligned}
\boldsymbol{\mu}_k^{(a)^*} &= (\Lambda_k^{(a)^*})^{-1}\left\{\tau^{(a)}\boldsymbol{\mu}^{(a)} + \sum_{n=1}^{N}\pi_{nk}^{(a)^*}\mathrm{diag}(\mathbb{E}_q[\boldsymbol{\tau}_{z_n^{(\tau)}}])\mathbf{y}_{n0} + \sum_{n=1}^{N}\pi_{nk}^{(a)^*}\mathrm{diag}(\mathbb{E}_q[\boldsymbol{\tau}_{z_n^{(\tau)}}])\{\sum_{t=1}^{T-1}\mathbf{y}_{nt}\}\right. \\
&-\sum_{n=1}^{N}\pi_{nk}^{(a)^*}\mathrm{diag}(\mathbb{E}_q[\boldsymbol{\tau}_{z_n^{(\tau)}}])\mathbb{E}_q[\mathbf{B}_{z_n^{(\mathbf{B})}}]\{\sum_{t=1}^{T-1}\mathbf{y}_{n(t-1)}\} + \sum_{n=1}^{N}\pi_{nk}^{(a)^*}\mathbb{E}_q[\mathbf{B}_{z_n^{(\mathbf{B})}}^\intercal]\mathrm{diag}(\mathbb{E}_q[\boldsymbol{\tau}_{z_n^{(\tau)}}])\mathbb{E}_q[\mathbf{B}_{z_n^{(\mathbf{B})}}]\{\sum_{t=1}^{T-1}\mathbf{y}_{n(t-1)}\} \\
&\left.-\sum_{n=1}^{N}\pi_{nk}^{(a)^*}\mathbb{E}_q[\mathbf{B}_{z_n^{(\mathbf{B})}}^\intercal]\mathrm{diag}(\mathbb{E}_q[\boldsymbol{\tau}_{z_n^{(\tau)}}])\{\sum_{t=1}^{T-1}\mathbf{y}_{nt}\}\right\} \in \mathbb{R}^{D\times 1}
\end{aligned}
$$

**For coefficient matrix $\mathbf{B}_k$:**

$$p(\mathrm{vec}(\mathbf{B}_k) \mid \mathbf{\Theta}_{-\mathbf{B}_k}, \mathbf{Y}) \propto p(\mathrm{vec}(\mathbf{B}_k) \mid \mathrm{vec}(\mathbf{M}^{(\mathbf{B})}), \tau^{(\mathbf{B})}\mathbf{I}) \prod_{n=1}^{N} p(\mathbf{y}_n \mid \mathbf{z}_n, \mathbf{a}, \mathbf{B}, \boldsymbol{\tau})^{\mathbf{1}[z_n^{(\mathbf{B})}=k]}$$

$$\propto \exp\left(\frac{\tau^{(\mathbf{B})}\mathrm{vec}(\mathbf{B}_k - \mathbf{M}^{(\mathbf{B})})^\intercal \mathrm{vec}(\mathbf{B}_k - \mathbf{M}^{(\mathbf{B})})}{-2}\right)$$

$$\times \prod_{n=1}^{N} \left\{ \exp\left(\frac{\mathrm{tr}(\mathrm{diag}(\boldsymbol{\tau}_{z_n^{(\tau)}})(\mathbf{Y}_{n,-0} - \mathbf{M}_{z_n^{(a)}})(\mathbf{Y}_{n,-0} - \mathbf{M}_{z_n^{(a)}})^\intercal)}{-2}\right)\right\}^{\mathbf{1}[z_n^{(\mathbf{B})}=k]}$$

where we let $\mathbf{Y}_{n,-0} = [\mathbf{y}_{n1}, \ldots, \mathbf{y}_{n(T-1)}] \in \mathbb{R}^{D\times(T-1)}$ and $\mathbf{Y}_{n,-(T-1)} = [\mathbf{y}_{n0}, \ldots, \mathbf{y}_{n(T-2)}] \in \mathbb{R}^{D\times(T-1)}$

so $\mathbf{M}_{z_n^{(a)}} = \mathbf{a}_{z_n^{(a)}}\mathbf{1}^\intercal + \mathbf{B}_k(\mathbf{Y}_{n,-(T-1)} - \mathbf{a}_{z_n^{(a)}}\mathbf{1}^\intercal) \in \mathbb{R}^{D\times(T-1)}$

Thus the precision matrix and mean matrix are

$$\Lambda_k^{(\mathbf{B})^*} = \tau^{(\mathbf{B})}\mathbf{I}_{D^2} + \sum_{n=1}^{N} \pi_{nk}^{(\mathbf{B})^*} \left\{(\mathbf{Y}_{n,-(T-1)} - \mathbb{E}_q[\mathbf{a}_{z_n^{(a)}}]\mathbf{1}^\intercal)(\mathbf{Y}_{n,-(T-1)} - \mathbb{E}_q[\mathbf{a}_{z_n^{(a)}}]\mathbf{1}^\intercal)^\intercal \otimes \mathrm{diag}(\mathbb{E}_q[\boldsymbol{\tau}_{z_n^{(\tau)}}])\right\}$$

$$\mathbf{M}_k^{(\mathbf{B})^*} = \mathrm{mat}\left[\Lambda_k^{(\mathbf{B})^{*-1}}\mathrm{vec}\left(\tau^{(\mathbf{B})}\mathbf{M}^{(\mathbf{B})} + \sum_{n=1}^{N} \pi_{nk}^{(\mathbf{B})^*}\mathrm{diag}(\mathbb{E}_q[\boldsymbol{\tau}_{z_n^{(\tau)}}])(\mathbf{Y}_{n,-0} - \mathbb{E}_q[\mathbf{a}_{z_n^{(a)}}]\mathbf{1}^\intercal)(\mathbf{Y}_{n,-(T-1)} - \mathbb{E}_q[\mathbf{a}_{z_n^{(a)}}]\mathbf{1}^\intercal)^\intercal\right)\right]$$

**For precision vector $\boldsymbol{\tau}_k$:**

$$p(\boldsymbol{\tau}_k \mid \mathbf{\Theta}_{-\boldsymbol{\tau}_k}, \mathbf{Y}) \propto \prod_{d=1}^{D} p(\tau_{kd} \mid \lambda_1^{(\tau)}, \lambda_2^{(\tau)}) \prod_{n=1}^{N} p(\mathbf{y}_n \mid \mathbf{z}_n, \mathbf{a}, \mathbf{B}, \boldsymbol{\tau})^{\mathbf{1}[z_n^{(\tau)}=k]}$$

$$\propto \prod_{d=1}^{D}\left\{\tau_{kd}^{\lambda_1^{(\tau)}-1}\exp(-\lambda_2^{(\tau)}\tau_{kd})\right\}\prod_{n=1}^{N}\left\{\left(\prod_{d=1}^{D}\tau_{kd}^{\frac{1}{2}}\right)\exp\left(\frac{(\mathbf{y}_{n0} - \mathbf{a}_{z_n^{(a)}})^\intercal\mathrm{diag}(\boldsymbol{\tau}_k)(\mathbf{y}_{n0} - \mathbf{a}_{z_n^{(a)}})}{-2}\right)\right\}^{\mathbf{1}[z_n^{(\tau)}=k]}$$

$$\prod_{n=1}^{N}\prod_{t=1}^{T-1}\left\{\left(\prod_{d=1}^{D}\tau_{kd}^{\frac{1}{2}}\right)\exp\left(\frac{(\mathbf{y}_{nt} - \mathbf{a}_{z_n^{(a)}} - \mathbf{B}_{z_n^{(\mathbf{B})}}\mathbf{y}_{n(t-1)} + \mathbf{B}_{z_n^{(\mathbf{B})}}\mathbf{a}_{z_n^{(a)}})^\intercal\mathrm{diag}(\boldsymbol{\tau}_k)(\mathbf{y}_{nt} - \mathbf{a}_{z_n^{(a)}} - \mathbf{B}_{z_n^{(\mathbf{B})}}\mathbf{y}_{n(t-1)} + \mathbf{B}_{z_n^{(\mathbf{B})}}\mathbf{a}_{z_n^{(a)}})}{-2}\right)\right\}^{\mathbf{1}[z_n^{(\tau)}=k]}$$

$$\propto \prod_{d=1}^{D}\left\{\tau_{kd}^{(\lambda_1^{(\tau)} + \frac{T}{2}\sum_{n=1}^{N}\mathbf{1}[z_n^{(\tau)}=k]-1)}\right\}\exp\left(-\tau_{kd}\left\{\lambda_2^{(\tau)} + \frac{1}{2}\sum_{n=1}^{N}\mathbf{1}[z_n^{(\tau)}=k](y_{n0,d} - a_{z_n^{(a)},d})^2\right.\right.$$

$$\left.\left. + \frac{1}{2}\sum_{n=1}^{N}\mathbf{1}[z_n^{(\tau)}=k]\sum_{t=1}^{T-1}(y_{nt,d} - a_{z_n^{(a)},d} - \mathbf{B}_{z_n^{(\mathbf{B})},d\cdot}^\intercal\mathbf{y}_{n(t-1)} + \mathbf{B}_{z_n^{(\mathbf{B})},d\cdot}^\intercal\mathbf{a}_{z_n^{(a)}})^2\right\}\right)$$

Thus, $\boldsymbol{\tau}_k$ follows independent Gamma distribution with parameters for each $\tau_{kd}$ to be:

$$\lambda_{kd,1}^{(\tau)^*} = \lambda_1^{(\tau)} + \frac{T}{2}\sum_{n=1}^{N}\pi_{nk}^{(\tau)^*}$$

and

$$\lambda_{kd,2}^{(\tau)^*} = \lambda_2^{(\tau)} + \frac{1}{2}\sum_{n=1}^{N}\pi_{nk}^{(\tau)^*}\left\{(y_{n0,d} - \mathbb{E}_q[a_{z_n^{(a)},d}])^2 + \sum_{t=1}^{T-1}(y_{nt,d} - \mathbb{E}_q[a_{z_n^{(a)},d}] - \mathbb{E}_q[\mathbf{B}_{z_n^{(\mathbf{B})},d\cdot}^\intercal]\mathbf{y}_{n(t-1)} + \mathbb{E}_q[\mathbf{B}_{z_n^{(\mathbf{B})},d\cdot}^\intercal]\mathbb{E}_q[\mathbf{a}_{z_n^{(a)}}])^2\right\}$$

We have the following results for the expectations in terms of variational parameters:

$$\mathbb{E}_q[\mathbf{a}_{z_n^{(a)}}] = \sum_{j=1}^{\ell}\pi_{nj}^{(a)^*}\boldsymbol{\mu}_j^{(a)^*}$$

$$\mathbb{E}_q[\mathbf{B}_{z_n^{(\mathbf{B})}}] = \sum_{j=1}^{\ell} {\pi_{nj}^{(\mathbf{B})}}^* \mathbf{M}_j^{(\mathbf{B})^*}$$

$$\mathbb{E}_q[\boldsymbol{\tau}_{z_n^{(\tau)}}] = \sum_{j=1}^{\ell} {\pi_{nj}^{(\tau)}}^* [\frac{{\lambda_{j1,1}^{(\tau)}}^*}{{\lambda_{j1,2}^{(\tau)}}^*}, \cdots, \frac{{\lambda_{jD,1}^{(\tau)}}^*}{{\lambda_{jD,2}^{(\tau)}}^*}]^\intercal$$

### C.2.2 Parameters for Cluster Assignments

**For $z_n^{(a)}$ of intercept**:

$$p(z_n^{(a)} = k \,|\, \boldsymbol{\Theta}_{-z_n^{(a)}}, \mathbf{Y}) \propto p(z_n^{(a)} = k \,|\, \boldsymbol{\pi}^{(a)}(\mathbf{v}^{(a)})) p(\mathbf{y}_n \,|\, \mathbf{z}_n, \mathbf{a}, \mathbf{B}, \boldsymbol{\tau})$$
$$\propto v_k^{(a)} \prod_{i=1}^{k-1} (1 - v_i^{(a)}) \exp\left( \frac{(\mathbf{y}_{n0} - \mathbf{a}_k)^\intercal \text{diag}(\boldsymbol{\tau}_{z_n^{(\tau)}})(\mathbf{y}_{n0} - \mathbf{a}_k)}{-2} \right)$$
$$\times \exp\left( \frac{\text{tr}(\text{diag}(\boldsymbol{\tau}_{z_n^{(\tau)}})(\mathbf{Y}_{n,-0} - \mathbf{M}_{z_n^{(\mathbf{B})}})(\mathbf{Y}_{n,-0} - \mathbf{M}_{z_n^{(\mathbf{B})}})^\intercal)}{-2} \right)$$
where $\mathbf{Y}_{n,-0} = [\mathbf{y}_{n1}, \ldots, \mathbf{y}_{n(T-1)}] \in \mathbb{R}^{D \times (T-1)}$ and $\mathbf{Y}_{n,-(T-1)} = [\mathbf{y}_{n0}, \ldots, \mathbf{y}_{n(T-2)}] \in \mathbb{R}^{D \times (T-1)}$
and $\mathbf{M}_{z_n^{(\mathbf{B})}} = \mathbf{a}_k \mathbf{1}^\intercal + \mathbf{B}_{z_n^{(\mathbf{B})}} (\mathbf{Y}_{n,-(T-1)} - \mathbf{a}_k \mathbf{1}^\intercal) \in \mathbb{R}^{D \times (T-1)}$

Thus,

$${\pi_{nk}^{(a)}}^* \propto \exp\left( \mathbb{E}_q[\log v_k^{(a)}] + \sum_{i=1}^{k-1} \mathbb{E}_q[\log(1 - v_i^{(a)})] + S_{nk}^{(a)} \right)$$

where

$$S_{nk}^{(a)} = \frac{-1}{2} \Big\{ (\mathbf{y}_{n0} - \mathbb{E}_q[\mathbf{a}_k])^\intercal \,\text{diag}(\mathbb{E}_q[\boldsymbol{\tau}_{z_n^{(\tau)}}])(\mathbf{y}_{n0} - \mathbb{E}_q[\mathbf{a}_k]) + \text{tr}\Big( \text{diag}(\mathbb{E}_q[\boldsymbol{\tau}_{z_n^{(\tau)}}])$$
$$\times (\mathbf{Y}_{n,-0} - \mathbb{E}_q[\mathbf{a}_k]\mathbf{1}^\intercal - \mathbb{E}_q[\mathbf{B}_{z_n^{(\mathbf{B})}}](\mathbf{Y}_{n,-(T-1)} - \mathbb{E}_q[\mathbf{a}_k]\mathbf{1}^\intercal))(\mathbf{Y}_{n,-0} - \mathbb{E}_q[\mathbf{a}_k]\mathbf{1}^\intercal - \mathbb{E}_q[\mathbf{B}_{z_n^{(\mathbf{B})}}](\mathbf{Y}_{n,-(T-1)} - \mathbb{E}_q[\mathbf{a}_k]\mathbf{1}^\intercal))^\intercal \Big) \Big\}$$

**Similarly,** ${\pi_{nk}^{(\mathbf{B})}}^* \propto \exp\left( \mathbb{E}_q[\log v_k^{(\mathbf{B})}] + \sum_{i=1}^{k-1} \mathbb{E}_q[\log(1 - v_i^{(\mathbf{B})})] + S_{nk}^{(\mathbf{B})} \right)$ where

$$S_{nk}^{(\mathbf{B})} = \frac{-1}{2} \Big\{ \Big( \mathbf{y}_{n0} - \mathbb{E}_q[\mathbf{a}_{z_n^{(a)}}] \Big)^\intercal \,\text{diag}(\mathbb{E}_q[\boldsymbol{\tau}_{z_n^{(\tau)}}])(\mathbf{y}_{n0} - \mathbb{E}_q[\mathbf{a}_{z_n^{(a)}}]) + \text{tr}\Big( \text{diag}(\mathbb{E}_q[\boldsymbol{\tau}_{z_n^{(\tau)}}])$$
$$(\mathbf{Y}_{n,-0} - \mathbb{E}_q[\mathbf{a}_{z_n^{(a)}}]\mathbf{1}^\intercal - \mathbb{E}_q[\mathbf{B}_k](\mathbf{Y}_{n,-(T-1)} - \mathbb{E}_q[\mathbf{a}_{z_n^{(a)}}]\mathbf{1}^\intercal))(\mathbf{Y}_{n,-0} - \mathbb{E}_q[\mathbf{a}_{z_n^{(a)}}]\mathbf{1}^\intercal - \mathbb{E}_q[\mathbf{B}_k](\mathbf{Y}_{n,-(T-1)} - \mathbb{E}_q[\mathbf{a}_{z_n^{(a)}}]\mathbf{1}^\intercal))^\intercal \Big) \Big\}$$

**For $z_n^{(\tau)}$ of noise**:

$$p(z_n^{(\tau)} = k \,|\, \boldsymbol{\Theta}_{-z_n^{(\tau)}}, \mathbf{Y}) \propto p(z_n^{(\tau)} = k \,|\, \boldsymbol{\pi}^{(\tau)}(\mathbf{v}^{(\tau)})) p(\mathbf{y}_n \,|\, \mathbf{z}_n, \mathbf{a}, \mathbf{B}, \boldsymbol{\tau})$$
$$\propto v_k^{(a)} \prod_{i=1}^{k-1} (1 - v_i^{(a)}) \left( \prod_{d=1}^{D} \tau_{kd}^{\frac{1}{2}} \right) \exp\left( \frac{(\mathbf{y}_{n0} - \mathbf{a}_{z_n^{(a)}})^\intercal \text{diag}(\boldsymbol{\tau}_k)(\mathbf{y}_{n0} - \mathbf{a}_{z_n^{(a)}})}{-2} \right)$$
$$\times \left( \prod_{d=1}^{D} \tau_{kd}^{\frac{T-1}{2}} \right) \exp\left( \frac{\text{tr}(\text{diag}(\boldsymbol{\tau}_k)(\mathbf{Y}_{n,-0} - \mathbf{M}_{z_n^{(a)}, z_n^{(\mathbf{B})}})(\mathbf{Y}_{n,-0} - \mathbf{M}_{z_n^{(a)}, z_n^{(\mathbf{B})}})^\intercal)}{-2} \right)$$

Thus, ${\pi^{(\tau)}}^*_{nk} \propto \exp\left( \mathbb{E}_q[\log v_k^{(\tau)}] + \sum_{i=1}^{k-1} \mathbb{E}_q[\log(1 - v_i^{(\tau)})] + S_{nk}^{(\tau)} \right)$ where

$$S_{nk}^{(\tau)} = \frac{T}{2} \sum_{d=1}^{D} \log \mathbb{E}_q[\tau_{kd}] + \frac{-1}{2} \Big\{ \Big( \mathbf{y}_{n0} - \mathbb{E}_q[\mathbf{a}_{z_n^{(a)}}] \Big)^\intercal \,\text{diag}(\mathbb{E}_q[\boldsymbol{\tau}_k])(\mathbf{y}_{n0} - \mathbb{E}_q[\mathbf{a}_{z_n^{(a)}}]) + \text{tr}\left( \text{diag}(\mathbb{E}_q[\boldsymbol{\tau}_k]) \times \right.$$
$$\left. (\mathbf{Y}_{n,-0} - \mathbb{E}_q[\mathbf{a}_{z_n^{(a)}}]\mathbf{1}^\intercal - \mathbb{E}_q[\mathbf{B}_{z_n^{(\mathbf{B})}}](\mathbf{Y}_{n,-(T-1)} - \mathbb{E}_q[\mathbf{a}_{z_n^{(a)}}]\mathbf{1}^\intercal))(\mathbf{Y}_{n,-0} - \mathbb{E}_q[\mathbf{a}_{z_n^{(a)}}]\mathbf{1}^\intercal - \mathbb{E}_q[\mathbf{B}_{z_n^{(\mathbf{B})}}](\mathbf{Y}_{n,-(T-1)} - \mathbb{E}_q[\mathbf{a}_{z_n^{(a)}}]\mathbf{1}^\intercal))^\intercal \right) \Big\}$$

### C.2.3 ELBO Computation

$$\mathbb{E}_q[\log p(\mathbf{y}_n \mid \mathbf{z}_n, \mathbf{a}, \mathbf{B}, \boldsymbol{\tau})] = \left(\frac{1}{2}\sum_{d=1}^{D}\mathbb{E}_q[\log \tau_{z_n^{(\tau)},d}]\right) - \frac{1}{2}(\mathbf{y}_{n0} - \mathbb{E}_q[\mathbf{a}_{z_n^{(a)}}])^\intercal \mathrm{diag}(\mathbb{E}_q[\boldsymbol{\tau}_{z_n^{(\tau)}}])(\mathbf{y}_{n0} - \mathbb{E}_q[\mathbf{a}_{z_n^{(a)}}])$$
$$+ \sum_{t=1}^{T-1}\left\{\left(\frac{1}{2}\sum_{d=1}^{D}\mathbb{E}_q[\log \tau_{z_n^{(\tau)},d}]\right) - \frac{1}{2}(\mathbf{y}_{nt} - \mathbb{E}_q[\mathbf{a}_{z_n^{(a)}}] - \mathbb{E}_q[\mathbf{B}_{z_n^{(\mathbf{B})}}]\mathbf{y}_{n(t-1)} + \mathbb{E}_q[\mathbf{B}_{z_n^{(\mathbf{B})}}]\mathbb{E}_q[\mathbf{a}_{z_n^{(a)}}])^\intercal \mathrm{diag}(\mathbb{E}_q[\boldsymbol{\tau}_{z_n^{(\tau)}}])\right.$$
$$\left.\times\ (\mathbf{y}_{nt} - \mathbb{E}_q[\mathbf{a}_{z_n^{(a)}}] - \mathbb{E}_q[\mathbf{B}_{z_n^{(\mathbf{B})}}]\mathbf{y}_{n(t-1)} + \mathbb{E}_q[\mathbf{B}_{z_n^{(\mathbf{B})}}]\mathbb{E}_q[\mathbf{a}_{z_n^{(a)}}])\right\}$$
$$\mathbb{E}_q[\log p(\mathbf{a}_k \mid \boldsymbol{\mu}^{(a)}, \tau^{(a)}\mathbf{I})] = -\frac{D}{2}\log(2\pi) + \frac{D}{2}\log\tau^{(a)} - \frac{\tau^{(a)}}{2}(\left\|\boldsymbol{\mu}_k^{(a)^*} - \boldsymbol{\mu}^{(a)}\right\|^2 + \mathrm{tr}((\Lambda_k^{(a)^*})^{-1}))$$
$$\mathbb{E}_q[\log q(\mathbf{a}_k \mid \boldsymbol{\mu}_k^{(a)^*}, \Lambda_k^{(a)^*})] = -\frac{D}{2}\log(2\pi) + \frac{1}{2}\log\left|\Lambda_k^{(a)^*}\right|$$
$$\mathbb{E}_q[\log p(\mathrm{vec}(\mathbf{B}_k) \mid \mathrm{vec}(\mathbf{M}^{(\mathbf{B})}), \tau^{(\mathbf{B})}\mathbf{I}] = -\frac{D^2}{2}\log(2\pi) + \frac{D^2}{2}\log\tau^{(\mathbf{B})} - \frac{\tau^{(\mathbf{B})}}{2}(\left\|\mathrm{vec}(\mathbf{M}_k^{(\mathbf{B})^*} - \mathbf{M}^{(\mathbf{B})})\right\|^2 + \mathrm{tr}((\Lambda_k^{(\mathbf{B})^*})^{-1}))$$
$$\mathbb{E}_q[\log q(\mathrm{vec}(\mathbf{B}_k) \mid \mathrm{vec}(\mathbf{M}_k^{(\mathbf{B})^*}), \Lambda_k^{(\mathbf{B})^*})] = -\frac{D^2}{2}\log(2\pi) + \frac{1}{2}\log\left|\Lambda_k^{(\mathbf{B})^*}\right|$$
$$\mathbb{E}_q[\log p(\boldsymbol{\tau}_k \mid \boldsymbol{\lambda}^{(\tau)})] = \sum_{d=1}^{D}\left(\lambda_1^{(\tau)}\log\lambda_2^{(\tau)} - \log\Gamma(\lambda_1^{(\tau)}) + (\lambda_1^{(\tau)} - 1)\left\{\boldsymbol{\Psi}(\lambda_{kd,1}^{(\tau)^*}) - \log\lambda_{kd,2}^{(\tau)^*}\right\} - \lambda_2^{(\tau)}\frac{\lambda_{kd,1}^{(\tau)^*}}{\lambda_{kd,2}^{(\tau)^*}}\right)$$
$$\mathbb{E}_q[\log q(\boldsymbol{\tau}_k \mid \boldsymbol{\lambda}_k^{(\tau)^*})] = \sum_{d=1}^{D}\left(\lambda_{kd,1}^{(\tau)^*}\log\lambda_{kd,2}^{(\tau)^*} - \log\Gamma(\lambda_{kd,1}^{(\tau)^*}) + (\lambda_{kd,1}^{(\tau)^*} - 1)\left\{\boldsymbol{\Psi}(\lambda_{kd,1}^{(\tau)^*}) - \log\lambda_{kd,2}^{(\tau)^*}\right\} - \lambda_{kd,1}^{(\tau)^*}\right)$$
$$\mathbb{E}_q[\log p(v_k^{(f)} \mid \alpha^{(f)})] = (\frac{s_1^{(f)^*}}{s_2^{(f)^*}} - 1)\{\boldsymbol{\Psi}(\alpha_{k2}^{(f)^*}) - \boldsymbol{\Psi}(\alpha_{k1}^{(f)^*} + \boldsymbol{\alpha}_{k2}^{(f)^*})\} - \{\boldsymbol{\Psi}(s_1^{(f)^*}) - \log s_1^{(f)^*}\}$$
$$\mathbb{E}_q[\log q(v_k^{(f)} \mid \boldsymbol{\alpha}_k^{(f)^*})] = (\alpha_{k1}^{(f)^*} - 1)\{\boldsymbol{\Psi}(\alpha_{k1}^{(f)^*}) - \boldsymbol{\Psi}(\alpha_{k1}^{(f)^*} + \alpha_{k2}^{(f)^*})\} + (\alpha_{k2}^{(f)^*} - 1)\{\boldsymbol{\Psi}(\alpha_{k2}^{(f)^*}) - \boldsymbol{\Psi}(\alpha_{k1}^{(f)^*} + \alpha_{k2}^{(f)^*})\}$$
$$- \log\mathbf{B}(\alpha_{k1}^{(f)^*}, \alpha_{k2}^{(f)^*})$$
$$\mathbb{E}_q[\log p(z_n^{(f)} \mid \mathbf{v}^{(f)})] = \sum_{k=1}^{\ell}\left\{\left(\sum_{i=k+1}^{\ell}\pi_{ni}^{(f)^*}\right)\{\boldsymbol{\Psi}(\alpha_{k2}^{(f)^*}) - \boldsymbol{\Psi}(\alpha_{k1}^{(f)^*} + \boldsymbol{\alpha}_{k2}^{(f)^*})\} + \pi_{nk}^{(f)^*}\{\boldsymbol{\Psi}(\alpha_{k1}^{(f)^*}) - \boldsymbol{\Psi}(\alpha_{k1}^{(f)^*} + \alpha_{k2}^{(f)^*})\}\right\}$$
$$\mathbb{E}_q[\log q(z_n^{(f)} \mid \boldsymbol{\pi}_n^{(f)^*})] = \log\max_i \pi_{ni}^{(f)^*}$$
$$\mathbb{E}_q[\log p(\alpha^{(f)} \mid \mathbf{s}^{(a)})] = s_1^{(f)}\log s_2^{(f)} - \log\Gamma(s_1^{(f)}) + (s_1^{(f)} - 1)\{\boldsymbol{\Psi}(s_1^{(f)^*}) - \log s_1^{(f)^*}\} - s_2^{(f)}\frac{s_1^{(f)^*}}{s_2^{(f)^*}}$$
$$\mathbb{E}_q[\log q(\alpha^{(f)} \mid \mathbf{s}^{(a)^*})] = s_1^{(f)^*}\log s_1^{(f)^*} - \log\Gamma(s_1^{(f)^*}) + (s_1^{(f)^*} - 1)\{\boldsymbol{\Psi}(s_1^{(f)^*}) - \log s_1^{(f)^*}\} - s_1^{(f)^*}$$

# D PROOF FOR B-SPLINES

Given a set of $N$ B-splines $\{B_{i,p}(t)\}_{i=0}^{N-1}$ of degree $p-1$ with coefficients $\{\beta_i\}_{i=0}^{N-1}$, denote $B_{0,p}(t)$ as a B-spline controlling the intercept such that $B_{0,p}(t=0) \neq 0$ while $B_{i,p}(t=0) = 0$ for all $i \neq 0$. Then the linear combination of the collection including $B_{0,p}(t)$ is equivalent to the linear combination of B-splines without $B_{0,p}(t)$ but plus additional explicit intercept.

*Proof.* We know $\sum_{i=0}^{N-1} B_{i,p}(t) = 1$ by definition. With $B_{0,p}(t)$, the linear combination is

$$\begin{aligned}\sum_{i=0}^{N-1}\beta_i B_{i,p}(t) &= \beta_0 B_{0,p}(t) + \sum_{i=1}^{N-1}\beta_i B_{i,p}(t) - \sum_{i=0}^{N-1}\beta_0 B_{i,p}(t) + \beta_0 \\ &= \beta_0 B_{0,p}(t) + \sum_{i=1}^{N-1}\beta_i B_{i,p}(t) - \beta_0 B_{0,p}(t) - \sum_{i=1}^{N-1}\beta_0 B_{i,p}(t) + \beta_0\end{aligned}$$

$$= \sum_{i=1}^{N-1} (\beta_i - \beta_0) B_{i,p}(t) + \beta_0$$
$$= \sum_{i=1}^{N-1} \beta_i^{(new)} B_{i,p}(t) + \beta_0 \quad (\text{ let } \beta_i^{(new)} = \beta_i - \beta_0)$$

From the last equation, we see that the first term is the linear combination of B-splines without $B_{0,p}(t)$ and the second term can be seen as an additional intercept term. □

# E PROOF FOR B-SPLINES INTERCEPT SHIFT

Given a set of $N$ B-splines $\{B_{i,p}(t)\}_{i=1}^{N}$ of degree $p-1$ with coefficients $\{\beta_i\}_{i=1}^{N}$, excluding the intercept $B_{0,p}(t)$, denote $t_{\text{tar}}$ as a targeted time point where we want a new intercept $\beta_0^{(new)}$ to represent its value. Then the B-splines function with this new intercept is equivalent to shifting downward all B-splines by $B_{i,p}(t_{\text{tar}})$.

*Proof.* By Supplement D we know the function of B-splines can be expressed as $f(t) = \beta_0 + \sum_{i=1}^{N} \beta_i B_{i,p}(t)$. When at time $t_{\text{tar}}$, let $\beta_0^{(new)} = f(t_{\text{tar}}) = \beta_0 + \sum_{i=1}^{N} \beta_i B_{i,p}(t_{\text{tar}})$, then

$$\begin{aligned} f(t) &= \beta_0 + \sum_{i=1}^{N} \beta_i B_{i,p}(t) + \beta_0^{(new)} - \beta_0^{(new)} \\ &= \beta_0^{(new)} + \left( \beta_0 + \sum_{i=1}^{N} \beta_i B_{i,p}(t) - \beta_0 - \sum_{i=1}^{N} \beta_i B_{i,p}(t_{\text{tar}}) \right) \\ &= \beta_0^{(new)} + \sum_{i=1}^{N} \beta_i \left( B_{i,p}(t) - B_{i,p}(t_{\text{tar}}) \right) \end{aligned}$$

□