The Full Derivation of MSFA Model

Anonymous Author(s)
Affiliation
Address
email

1 Nonparametric Factor Analysis Model

Let $\boldsymbol{X} = [\boldsymbol{x}_{1:}, \ldots, \boldsymbol{x}_{D:}]^T \in \mathbb{R}^{D \times N}$ be the data matrix in which $\boldsymbol{x}_{d:}$ denotes measured variable d and $\boldsymbol{x}_{:n}$ denotes data sample n. In standard factor analysis model, the data matrix $\boldsymbol{X}$ can be decomposed into the product of the *factor loading matrix* $\boldsymbol{A} \in \mathbb{R}^{D \times K}$ and *factor matrix* $\boldsymbol{F} = [\boldsymbol{f}_{:1}, \ldots, \boldsymbol{f}_{:N}] \in \mathbb{R}^{K \times N}$ plus noise matrix $\boldsymbol{E}$. The number of latent factors K is much smaller than D in general. The generative model of factor analysis is given by:

$$\begin{aligned} \boldsymbol{X} &= \boldsymbol{A}\boldsymbol{F} + \boldsymbol{E} \\ \boldsymbol{F} &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \\ \boldsymbol{E} &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi}) \end{aligned} \tag{1}$$

where the matrix $\boldsymbol{E} = [\boldsymbol{e}_{:1}, \ldots, \boldsymbol{e}_{:N}]$ accounts for idiosyncratic noise. $\boldsymbol{\Psi} = [\sigma_\epsilon^2 \boldsymbol{I}]$ is a diagonal matrix. We assume $\boldsymbol{F}$ has a zero mean, unit variance Gaussian prior as used in standard factor analysis [**?**]. In nonparametric Bayesian factor analysis models [**?**, **?**, **?**, **?**], the factor loading matrix $\boldsymbol{A}$ can be factorized into the Hadamard product of a real-valued matrix $\boldsymbol{G}$ and a binary-valued matrix $\boldsymbol{Z}$ in further.

$$\boldsymbol{A} = \boldsymbol{G} \odot \boldsymbol{Z} \tag{2}$$

where $\boldsymbol{G}$ and $\boldsymbol{Z}$ are of the same size as $\boldsymbol{A}$. For each data sample n, the factor analysis model is given by: $\boldsymbol{x}_{:n} = (\boldsymbol{G} \odot \boldsymbol{Z})\boldsymbol{f}_{:n} + \boldsymbol{e}_{:n}$. We assume $\boldsymbol{G}$ has a Gaussian prior $\mathcal{N}(\boldsymbol{0}, \sigma_g^2 \boldsymbol{I})$ and $\boldsymbol{Z}$ has a Dirichlet process prior $\mathcal{DP}(\beta, H_0)$ in which β is the concentration parameter. The graphical model for the nonparametric Bayesian factor analysis model and its equivalent representation are depicted in Figure (1)(2) respectively.

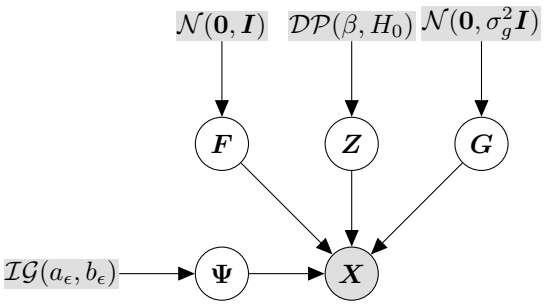

Figure 1: The Graphical Model for Nonparametric Bayesian Factor Analysis Model

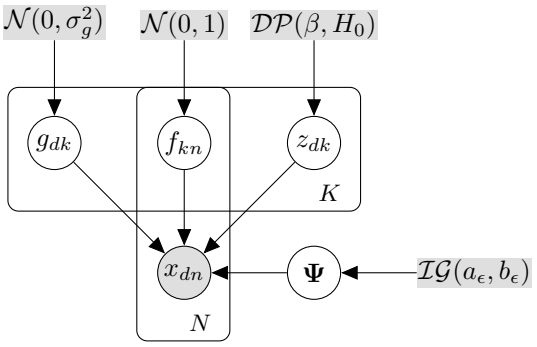

Figure 2: An Equivalent Graphical Representation for Nonparametric Bayesian Factor Analysis

We use $\mathbf{\Theta}$ to denote model parameters $\{\boldsymbol{G}, \boldsymbol{Z}, \boldsymbol{F}, \mathbf{\Psi}\}$, and $x_{dn}, g_{dk}, z_{dk}, f_{kn}$ to denote the entry of $\boldsymbol{X}, \boldsymbol{G}, \boldsymbol{Z}, \boldsymbol{F}$ respectively.

The joint distribution of the data $\boldsymbol{X}$ and model parameters $\mathbf{\Theta}$ can be written as:

$$
\begin{aligned}
p(\boldsymbol{X}, \mathbf{\Theta}) &= \prod_{n=1}^{N}\prod_{d=1}^{D}\prod_{k=1}^{K} p(x_{dn}, g_{dk}, z_{dk}, f_{kn}, \sigma_\epsilon^2, \mu_g, \sigma_g^2, \beta, a_\epsilon, b_\epsilon) \\
&= \prod_{n=1}^{N}\prod_{d=1}^{D}\prod_{k=1}^{K} p(x_{dn}|g_{dk}, z_{dk}, f_{kn}, \sigma_\epsilon^2, \mu_g, \sigma_g^2, \beta, a_\epsilon, b_\epsilon) \\
&\quad \cdot \prod_{d=1}^{D}\prod_{k=1}^{K} p(g_{dk}|\mu_g, \sigma_g^2, z_{dk}, f_{kn}, \sigma_\epsilon^2, \beta, a_\epsilon, b_\epsilon) \\
&\quad [\text{ Each entry } g_{dk} \text{ of the matrix } \mathbf{G} \text{ is given a Gaussian prior independently.}] \\
&\quad \cdot \prod_{d=1}^{D}\prod_{k=1}^{K} p(z_{dk}|\boldsymbol{z}_{-dk}, f_{kn}, \sigma_\epsilon^2, \mu_g, \sigma_g^2, \beta, a_\epsilon, b_\epsilon) \\
&\quad [\ \boldsymbol{z}_{-dk} \text{ denotes variable indicators excluding } z_{dk}.] \\
&\quad \cdot \prod_{n=1}^{N}\prod_{k=1}^{K} p(f_{kn}|\sigma_\epsilon^2, \mu_g, \sigma_g^2, \beta, a_\epsilon, b_\epsilon) \\
&\quad \cdot p(\sigma_\epsilon^2|a_\epsilon, b_\epsilon, \mu_g, \sigma_g^2, \beta) \\
&\quad \cdot p(\mu_g, \sigma_g^2, \beta, a_\epsilon, b_\epsilon) \\
&= \prod_{n=1}^{N}\prod_{d=1}^{D}\prod_{k=1}^{K} p(x_{dn}|g_{dk}, z_{dk}, f_{kn}, \sigma_\epsilon^2, \underline{\boldsymbol{z}_{-dk}, \mu_g, \sigma_g^2, \beta, a_\epsilon, b_\epsilon}) \\
&\quad \cdot \prod_{d=1}^{D}\prod_{k=1}^{K} p(g_{dk}|\mu_g, \sigma_g^2, \underline{z_{dk}, \boldsymbol{z}_{-dk}, f_{kn}, \sigma_\epsilon^2, \beta, a_\epsilon, b_\epsilon}) \\
&\quad \cdot \prod_{d=1}^{D}\prod_{k=1}^{K} p(z_{dk}|\boldsymbol{z}_{-dk}, \beta, \underline{f_{kn}, \sigma_\epsilon^2, \mu_g, \sigma_g^2, a_\epsilon, b_\epsilon}) \\
&\quad \cdot \prod_{n=1}^{N}\prod_{k=1}^{K} p(f_{kn}|\underline{\sigma_\epsilon^2, \mu_g, \sigma_g^2, \beta, a_\epsilon, b_\epsilon}) \\
&\quad \cdot p(\sigma_\epsilon^2|a_\epsilon, b_\epsilon, \underline{\mu_g, \sigma_g^2, \beta}) p(\mu_g, \sigma_g^2) p(\beta) p(a_\epsilon, b_\epsilon)
\end{aligned}
$$

The variables with underline from line 8-12 can be removed from the conditional distribution according to the independencies among variables shown in Figure (1). We use a consistent concentration parameter β for all local factor analyzers as we assume the number of local latent factors is about the same.

The likelihood is given by:

$$
\begin{aligned}
p(\boldsymbol{X}|\boldsymbol{\Theta}) &= \prod_{n=1}^{N}\prod_{d=1}^{D}\prod_{k=1}^{K} p(x_{dn}|g_{dk}, z_{dk}, f_{kn}, \sigma_\epsilon^2) \\
&= \prod_{n=1}^{N}\prod_{d=1}^{D}\prod_{k=1}^{K} \left(\frac{1}{\sqrt{2\pi\sigma_\epsilon^2}}\exp\left[-\frac{(x_{dn}-g_{dk}f_{kn})^2}{2\sigma_\epsilon^2}\right]\right)^{z_{dk}} \\
&= \prod_{n=1}^{N}\prod_{d=1}^{D} \left(\frac{1}{\sqrt{2\pi\sigma_\epsilon^2}}\exp\left[-\frac{(x_{dn}-g_{d1}f_{1n})^2}{2\sigma_\epsilon^2}\right]\right)^{z_{d1}} \cdots \left(\frac{1}{\sqrt{2\pi\sigma_\epsilon^2}}\exp\left[-\frac{(x_{dn}-g_{dK}f_{Kn})^2}{2\sigma_\epsilon^2}\right]\right)^{z_{dK}} \\
&= \prod_{n=1}^{N}\prod_{d=1}^{D} \left(\frac{1}{\sqrt{2\pi\sigma_\epsilon^2}}\exp\left[-\frac{(x_{dn}-g_{d\hat{\mathrm{k}}}f_{\hat{\mathrm{k}}n})^2}{2\sigma_\epsilon^2}\right]\right) \\
&= \frac{1}{(2\pi\sigma_\epsilon^2)^{ND/2}}\exp\left[-\frac{\sum_{n=1,d=1}^{N,D}(x_{dn}-g_{d\hat{\mathrm{k}}}f_{\hat{\mathrm{k}}n})^2}{2\sigma_\epsilon^2}\right] \\
&= \frac{1}{(2\pi\sigma_\epsilon^2)^{ND/2}}\exp\left[-\frac{tr\left([\boldsymbol{X}-(\boldsymbol{G}\odot\boldsymbol{Z})\boldsymbol{F})]^T[\boldsymbol{X}-(\boldsymbol{G}\odot\boldsymbol{Z})\boldsymbol{F}]\right)}{2\sigma_\epsilon^2}\right]
\end{aligned}
\tag{3}
$$

where we use $\hat{\mathrm{k}}$ to denote k when $z_{dk}=1$.

We derive the posterior distribution of model parameters used in our Gibbs sampling in following subsections.

1.1 Sampling g_{dk}

We place a Gaussian prior $\mathcal{N}(\mu_g, \sigma_g^2)$ on each factor loading g_{dk} independently and let $\mu_g = 0$ as follows:

$$
p(g_{dk}|\mu_g, \sigma_g^2) = \frac{1}{\sqrt{2\pi\sigma_g^2}}\exp\left[-\frac{(g_{dk}-\mu_g)^2}{2\sigma_g^2}\right] \tag{4}
$$

Given the data and the other model parameters be fixed, we can derive the posterior distribution of g_{dk} as follows:

$$
\begin{aligned}
p(g_{dk}|X,-) &= \prod_{n=1}^{N} p(g_{dk}|x_{dn}, f_{kn}, z_{dk}, \boldsymbol{z}_{-dk}, \mu_g, \sigma_g^2, \sigma_\epsilon^2) \\
&= \prod_{n=1}^{N} \frac{p(x_{dn}, g_{dk}|z_{dk}, \boldsymbol{z}_{-dk}, f_{kn}, \sigma_\epsilon^2, \mu_g, \sigma_g^2)}{p(x_{dn}|z_{dk}, \boldsymbol{z}_{-dk}, f_{kn}, \sigma_\epsilon^2, \mu_g, \sigma_g^2)} \\
&= \prod_{n=1}^{N} \frac{p(x_{dn}|g_{dk}, \boldsymbol{z}_{dk}, \boldsymbol{z}_{-dk}, f_{kn}, \sigma_\epsilon^2)p(g_{dk}|\mu_g, \sigma_g^2)}{p(x_{dn}|z_{dk}, \boldsymbol{z}_{-dk}, f_{kn}, \sigma_\epsilon^2, \mu_g, \sigma_g^2)}
\end{aligned}
$$

$$\propto \prod_{n=1}^{N} p(x_{dn}|g_{dk}, z_{dk}, \boldsymbol{z}_{-dk}, f_{kn}, \sigma_\epsilon^2) p(g_{dk}|\mu_g, \sigma_g^2)$$

$[\because p(x_{dn}|z_{dk}, \boldsymbol{z}_{-dk}, f_{kn}, \sigma_\epsilon^2, \mu_g, \sigma_g^2)$ is normalization constant given that the other parameters fixed.$]$

$$\propto \prod_{n=1}^{N} \left(\frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} \exp\left[-\frac{(x_{dn} - g_{dk} f_{kn})^2}{2\sigma_\epsilon^2} \right] \right)^{z_{dk}} \frac{1}{\sqrt{2\pi\sigma_g^2}} \exp\left[-\frac{(g_{dk} - \mu_g)^2}{2\sigma_g^2} \right]$$

where z_{dk} plays the role of feature (i.e.latent factor) selection as variable d can be associated to only one factor $\boldsymbol{f}_{k:}$ indicated by $z_{dk} = 1$ with the associating strength measured by g_{dk}. We use $-$ to denote the "rest" of the model parameters.

When $z_{dk} = 1$,

$$\begin{aligned} p(g_{dk}|X, -) &\propto \exp\left[-\frac{\sum_n (x_{dn} - g_{dk} f_{kn})^2}{2\sigma_\epsilon^2} \right] \exp\left[-\frac{(g_{dk} - \mu_g)^2}{2\sigma_g^2} \right] \\ &\quad [\because \text{ only the exponential terms involve } g_{dk}] \\ &\propto \exp\left[-\frac{\sum_n (x_{dn}^2 + g_{dk}^2 f_{kn}^2 - 2 x_{dn} g_{dk} f_{kn})}{2\sigma_\epsilon^2} - \frac{{g_{dk}}^2 + \mu_g^2 - 2\mu_g g_{dk}}{2\sigma_g^2} \right] \\ &\propto \exp\left[-\frac{g_{dk}^2}{2} \left(\frac{\sum_n f_{kn}^2}{\sigma_\epsilon^2} + \frac{1}{\sigma_g^2} \right) + g_{dk} \left(\frac{\sum_n x_{dn} f_{kn}}{\sigma_\epsilon^2} + \frac{\mu_g}{\sigma_g^2} \right) - \left(\frac{\sum_n x_{dn}^2}{2\sigma_\epsilon^2} + \frac{\mu_g}{2\sigma_g^2} \right) \right] \end{aligned} \tag{5}$$

Since the product of two Gaussians is a Gaussian, we rewrite this in the form

$$\begin{aligned} p(g_{dk}|X, -) &\propto \exp\left[-\frac{g_{dk}^2}{2} \left(\frac{\sum_n f_{kn}^2}{\sigma_\epsilon^2} + \frac{1}{\sigma_g^2} \right) + g_{dk} \left(\frac{\sum_n x_{dn} f_{kn}}{\sigma_\epsilon^2} + \frac{\mu_g}{\sigma_g^2} \right) - \left(\frac{\sum_n x_{dn}^2}{2\sigma_\epsilon^2} + \frac{\mu_g}{2\sigma_g^2} \right) \right] \\ &= \exp\left[-\frac{\lambda_{dk}}{2} (g_{dk}^2 - 2 g_{dk} \mu_{dk} + {\mu_{dk}}^2) \right] \\ &= \exp\left[-\frac{\lambda_{dk}}{2} (g_{dk} - \mu_{dk})^2 \right] \end{aligned} \tag{6}$$

Matching coefficient of g_{dk}^2, we find λ_{dk} is given by

$$\lambda_{dk} = \frac{\sum_n f_{kn}^2}{\sigma_\epsilon^2} + \frac{1}{\sigma_g^2} \tag{7}$$

Matching coefficient of g_{dk}, we find μ_{dk} is given by

$$\mu_{dk} = \left(\frac{\sum_n f_{kn}^2}{\sigma_\epsilon^2} + \frac{1}{\sigma_g^2} \right)^{-1} \left(\frac{\sum_n x_{dn} f_{kn}}{\sigma_\epsilon^2} + \frac{\mu_g}{\sigma_g^2} \right) \tag{8}$$

Hence, given the data and the other parameters be fixed, we can sample $g_{dk} \sim \mathcal{N}(\mu_{dk}, \lambda_{dk})$ in which μ_{dk}, λ_{dk}, that are the mean and precision parameters of the posterior distribution of g_{dk}, can be calculated according to:

$$\begin{aligned} \lambda_{dk} &= \begin{cases} \frac{\sum_n f_{kn}^2}{\sigma_\epsilon^2} + \frac{1}{\sigma_g^2} & \text{if } z_{dk} = 1 \\ \frac{1}{\sigma_g^2} & \text{if } z_{dk} = 0 \end{cases} \\ \mu_{dk} &= \begin{cases} \left(\frac{\sum_n f_{kn}^2}{\sigma_\epsilon^2} + \frac{1}{\sigma_g^2} \right)^{-1} \left(\frac{\sum_n x_{dn} f_{kn}}{\sigma_\epsilon^2} + \frac{\mu_g}{\sigma_g^2} \right) & \text{if } z_{dk} = 1 \\ \mu_g & \text{if } z_{dk} = 0 \end{cases} \end{aligned} \tag{9}$$

1.2 Sampling z_{dk}

Our goal is to partition the measure variables into K clusters but K is unbounded a priori. The Dirichlet process (DP) defines a distribution over clustering where the number of clusters does not need to be specified a priori. The mixing proportions, $\omega_k = \frac{\sum_{d=1}^{D} z_{dk}}{D}$, are given a symmetric Dirichlet prior with the concentration parameters β:

$$p(\omega_1, \ldots, \omega_K|\beta) \sim \text{Dirichlet}(\beta/K, \ldots, \beta/K) = \frac{\Gamma(\beta)}{\Gamma(\beta/K)^K} \prod_{k=1}^{K} \omega_k^{\beta/K-1} \tag{10}$$

where the mixing proportions must be positive and sum to one. We use $z_{dk} = 1$ to indicate that the measured variable d is partitioned into cluster k. As we enforce the disjoint partition over variables, one variable can be grouped into only one cluster and thus $\sum_{k=1}^{K} z_{dk} = 1$. Given the mixing proportions, the prior for the occupation numbers, $m_k = \sum_{d=1}^{D} z_{dk}$, is multinomial and the joint distribution of the indicators becomes:

$$p(\{z_{dk}\}_{d=1,k=1}^{D,K}|\omega_1, \ldots, \omega_K) = \prod_{k=1}^{K} \omega_k^{m_k} \tag{11}$$

Using the standard Dirichlet integral, we can integrate out the mixing proportions and write the prior directly in terms of the indicators as follows:

$$\begin{aligned}
p(\{z_{dk}\}_{d=1,k=1}^{D,K}|\beta) &= \int p(\{z_{dk}\}_{d=1,k=1}^{D,K}|\omega_1, \ldots, \omega_K) p(\omega_1, \ldots, \omega_K|\beta) \mathrm{d}\omega_1 \ldots \mathrm{d}\omega_K \\
&= \frac{\Gamma(\beta)}{\Gamma(\beta/K)^K} \int \prod_{k=1}^{K} \omega_k^{m_k+\beta/K-1} \mathrm{d}\omega_k \\
&= \frac{\Gamma(\beta)}{\Gamma(\beta/K)^K} \int \frac{\prod_k \Gamma(m_k+\beta/K)}{\Gamma(\sum_k m_k+\beta)} \frac{\Gamma(\sum_k m_k+\beta)}{\prod_k \Gamma(m_k+\beta/K)} \prod_{k=1}^{K} \omega_k^{m_k+\beta/K-1} \mathrm{d}\omega_k \\
&= \frac{\Gamma(\beta)}{\Gamma(\beta/K)^K} \frac{\prod_k \Gamma(m_k+\beta/K)}{\Gamma(\sum_k m_k+\beta)} \\
&= \frac{\Gamma(\beta)}{\Gamma(D+\beta)} \prod_{k=1}^{K} \frac{\Gamma(m_k+\beta/K)}{\Gamma(\beta/K)} \\
&[\because \int \frac{\Gamma(\sum_k m_k+\beta)}{\prod_k \Gamma(m_k+\beta/K)} \prod_{k=1}^{K} \omega_k^{m_k+\beta/K-1} \mathrm{d}\omega_k = 1]
\end{aligned} \tag{12}$$

We use $\boldsymbol{z}_{-dk}$ denotes all the indicators except $\boldsymbol{z}_{d:} = [z_{d1}, \ldots, z_{dK}]$. By keeping $\boldsymbol{z}_{-dk}$ to be fixed, we obtain the conditional prior for the single indicator z_{dk} :

$$p(z_{dk}|\boldsymbol{z}_{-dk}, \beta) = \begin{cases} \frac{m_{-dk}+\beta/K}{D-1+\beta} & \text{if } z_{dk} = 1 \\ \frac{D-1+\beta-m_{-dk}-\beta/K}{D-1+\beta} & \text{if } z_{dk} = 0 \end{cases} \tag{13}$$

where the subscript $-d$ denotes all indexes except d and m_{-dk} is the number of variables, excluding $\boldsymbol{x}_{d:}$, that are grouped into cluster k.

By taking the limit $K \to \infty$, we get the conditional prior for a single indicator as:

$$p(z_{dk}|\boldsymbol{z}_{-dk}, \beta) = \begin{cases} \frac{m_{-dk}}{D-1+\beta} & \text{if } z_{dk} = 1 \\ \frac{D-1+\beta-m_{-dk}}{D-1+\beta} & \text{if } z_{dk} = 0 \end{cases} \tag{14}$$

Given the data $\boldsymbol{X}$ and model parameters be fixed excluding z_{dk}, we can calculate the posterior distribution of z_{dk} as:

$$
\begin{aligned}
p(z_{dk}=1|\boldsymbol{X},-) &= \prod_{n=1}^{N} p(z_{dk}=1|x_{dn},g_{dk},f_{kn},\boldsymbol{z}_{-dk},\beta) \\
&= \prod_{n=1}^{N} \frac{p(z_{dk}=1,x_{dn}|g_{dk},f_{kn},\boldsymbol{z}_{-dk},\beta)}{p(x_{dn}|g_{dk},f_{kn},\boldsymbol{z}_{-dk},\beta)} \\
&\propto \prod_{n=1}^{N} p(z_{dk}=1,x_{dn}|g_{dk},f_{kn},\boldsymbol{z}_{-dk},\beta) \\
&= \prod_{n=1}^{N} p(x_{dn}|g_{dk},z_{dk}=1,f_{kn},\boldsymbol{z}_{-dk},\beta)p(z_{dk}=1|g_{dk},f_{kn},\boldsymbol{z}_{-dk},\beta) \\
&= \prod_{n=1}^{N} p(x_{dn}|g_{dk},z_{dk}=1,f_{kn},\boldsymbol{z}_{-dk},\beta)p(z_{dk}=1|g_{dk},f_{kn},\boldsymbol{z}_{-dk},\beta) \\
&= \prod_{n=1}^{N} p(x_{dn}|g_{dk},z_{dk}=1,f_{kn},\beta)p(z_{dk}=1|\boldsymbol{z}_{-dk},\beta) \\
&= \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} \exp\left[-\frac{(x_{dn}-g_{dk}f_{kn})^2}{2\sigma_\epsilon^2}\right] \frac{m_{-dk}}{D-1+\beta} \\
&= \frac{1}{(2\pi\sigma_\epsilon^2)^{N/2}} \exp\left[-\frac{\sum_{n=1}^{N}(x_{dn}-g_{dk}f_{kn})^2}{2\sigma_\epsilon^2}\right] \frac{m_{-dk}}{D-1+\beta} \\
&\propto \exp\left[-\frac{tr\left([\boldsymbol{x}_{d:}-g_{dk}\boldsymbol{f}_{k:}]^T[\boldsymbol{x}_{d:}-g_{dk}\boldsymbol{f}_{k:}]\right)}{2\sigma_\epsilon^2}\right] \frac{m_{-dk}}{D-1+\beta}
\end{aligned}
$$

Alternatively, we can integrate out g_{dk} to sample z_{dk} as:

$$
\begin{aligned}
p(z_{dk}=1|\boldsymbol{X},-) &= \int p(z_{dk}=1,g_{dk}|\boldsymbol{X},\boldsymbol{\Theta})\mathrm{d}g_{dk} \\
&= \int \prod_{n=1}^{N} p(z_{dk}=1,g_{dk}|x_{dn},f_{kn},\boldsymbol{z}_{-dk},\beta,\mu_g,\sigma_g^2)\mathrm{d}g_{dk} \\
&= \int \prod_{n=1}^{N} \frac{p(z_{dk}=1,g_{dk},x_{dn}|f_{kn},\boldsymbol{z}_{-dk},\beta,\mu_g,\sigma_g^2)}{p(x_{dn}|f_{kn},\boldsymbol{z}_{-dk},\beta,\mu_g,\sigma_g^2)}\mathrm{d}g_{dk} \\
&\propto \int \prod_{n=1}^{N} p(z_{dk}=1,g_{dk},x_{dn}|f_{kn},\boldsymbol{z}_{-dk},\beta,\mu_g,\sigma_g^2)\mathrm{d}g_{dk} \\
&= \int \prod_{n=1}^{N} p(x_{dn}|g_{dk},z_{dk}=1,f_{kn},\boldsymbol{z}_{-dk},\beta,\mu_g,\sigma_g^2) \\
&\times p(z_{dk}=1|g_{dk},f_{kn},\boldsymbol{z}_{-dk},\beta,\mu_g,\sigma_g^2)p(g_{dk}|\mu_g,\sigma_g^2)\mathrm{d}g_{dk} \\
&= \int \prod_{n=1}^{N} p(x_{dn}|g_{dk},z_{dk}=1,f_{kn})p(z_{dk}=1|\boldsymbol{z}_{-dk},\beta)p(g_{dk}|\mu_g,\sigma_g^2)\mathrm{d}g_{dk} \\
&= p(z_{dk}=1|\boldsymbol{z}_{-dk},\beta)\int \prod_{n=1}^{N} p(x_{dn}|g_{dk},z_{dk}=1,f_{kn})p(g_{dk}|\mu_g,\sigma_g^2)\mathrm{d}g_{dk}
\end{aligned}
$$

$$
\begin{aligned}
&= p(z_{dk}=1|\boldsymbol{z}_{-dk},\beta)\int \frac{1}{(2\pi\sigma_\epsilon^2)^{\frac{N}{2}}}\exp\left[-\frac{\sum_n(x_{dn}-g_{dk}f_{kn})^2}{2\sigma_\epsilon^2}\right]\frac{1}{(2\pi\sigma_g^2)^{\frac{1}{2}}}\exp\left[-\frac{(g_{dk}-\mu_g)^2}{2\sigma_g^2}\right]\mathrm{d}g_{dk}\\
&= p(z_{dk}=1|\boldsymbol{z}_{-dk},\beta)\int \frac{1}{(2\pi\sigma_\epsilon^2)^{\frac{N}{2}}}\exp\left[-\frac{\sum_n(x_{dn}^2+g_{dk}^2f_{kn}^2-2x_{dn}g_{dk}f_{kn})}{2\sigma_\epsilon^2}\right]\\
&\times\frac{1}{(2\pi\sigma_g^2)^{\frac{1}{2}}}\exp\left[-\frac{g_{dk}^2+\mu_g^2-2\mu_g g_{dk}}{2\sigma_g^2}\right]\mathrm{d}g_{dk}\\
&= p(z_{dk}=1|\boldsymbol{z}_{-dk},\beta)\int \frac{1}{(2\pi\sigma_\epsilon^2)^{\frac{N}{2}}}\exp\left[-\frac{g_{dk}^2}{2}(\frac{\sum_n f_{kn}^2}{\sigma_\epsilon^2}+\frac{1}{\sigma_g^2})+g_{dk}(\frac{\sum_n x_{dn}f_{kn}}{\sigma_\epsilon^2}+\frac{\mu_g}{\sigma_g^2})\right]\\
&\times\frac{1}{(2\pi\sigma_g^2)^{\frac{1}{2}}}\exp\left[-\frac{\sum_n x_{dn}^2}{2\sigma_\epsilon^2}-\frac{\mu_g}{2\sigma_g^2}\right]\mathrm{d}g_{dk}\\
&[\text{ Recall equation 6}]\\
&= p(z_{dk}=1|\boldsymbol{z}_{-dk},\beta)\int \frac{1}{(2\pi\sigma_\epsilon^2)^{\frac{N}{2}}}\exp\left[-\frac{{g_{dk}}^2}{2}\lambda_{dk}+g_{dk}\lambda_{dk}\mu_{dk}+\frac{\lambda_{dk}{\mu_{dk}}^2}{2}-\frac{\lambda_{dk}{\mu_{dk}}^2}{2}\right]\\
&\times\frac{1}{(2\pi\sigma_g^2)^{\frac{1}{2}}}\exp\left[-\frac{\sum_n x_{dn}^2}{2\sigma_\epsilon^2}-\frac{\mu_g}{2\sigma_g^2}\right]\mathrm{d}g_{dk}\\
&= p(z_{dk}=1|\boldsymbol{z}_{-dk},\beta)\int \frac{1}{(2\pi\sigma_\epsilon^2)^{\frac{N}{2}}}\exp\left[-\frac{\lambda_{dk}}{2}(g_{dk}-\mu_{dk})^2\right]\exp\left[\frac{\lambda_{dk}{\mu_{dk}}^2}{2}\right]\\
&\times\frac{1}{(2\pi\sigma_g^2)^{\frac{1}{2}}}\exp\left[-\frac{\sum_n x_{dn}^2}{2\sigma_\epsilon^2}-\frac{\mu_g}{2\sigma_g^2}\right]\mathrm{d}g_{dk}\\
&= p(z_{dk}=1|\boldsymbol{z}_{-dk},\beta)\int \exp\left[-\frac{\lambda_{dk}}{2}(g_{dk}-\mu_{dk})^2\right]\mathrm{d}g_{dk}\ \frac{1}{(2\pi\sigma_\epsilon^2)^{\frac{N}{2}}}\exp\left[\frac{\lambda_{dk}{\mu_{dk}}^2}{2}\right]\\
&\times\frac{1}{(2\pi\sigma_g^2)^{\frac{1}{2}}}\exp\left[-\frac{\sum_n x_{dn}^2}{2\sigma_\epsilon^2}-\frac{\mu_g}{2\sigma_g^2}\right]\\
&= p(z_{dk}=1|\boldsymbol{z}_{-dk},\beta)\frac{(2\pi)^{\frac{1}{2}}}{{\lambda_{dk}}^{\frac{1}{2}}}\frac{1}{(2\pi\sigma_\epsilon^2)^{\frac{N}{2}}}\exp\left[\frac{\lambda_{dk}{\mu_{dk}}^2}{2}\right]\frac{1}{(2\pi\sigma_g^2)^{\frac{1}{2}}}\exp\left[-\frac{\sum_n x_{dn}^2}{2\sigma_\epsilon^2}-\frac{\mu_g}{2\sigma_g^2}\right]\\
&= p(z_{dk}=1|\boldsymbol{z}_{-dk},\beta)\frac{1}{{\lambda_{dk}}^{\frac{1}{2}}}\frac{1}{(2\pi\sigma_\epsilon^2)^{\frac{N}{2}}}\exp\left[\frac{\lambda_{dk}{\mu_{dk}}^2}{2}\right]\frac{1}{\sigma_g}\exp\left[-\frac{\sum_n x_{dn}^2}{2\sigma_\epsilon^2}-\frac{\mu_g}{\sigma_g^2}\right]\\
&\propto p(z_{dk}=1|\boldsymbol{z}_{-dk},\beta)\frac{1}{{\lambda_{dk}}^{\frac{1}{2}}}\exp\left[\frac{\lambda_{dk}{\mu_{dk}}^2}{2}\right]\\
&\propto \frac{m_{-dk}}{D-1+\beta}\frac{1}{{\lambda_{dk}}^{\frac{1}{2}}}\exp\left[\frac{\lambda_{dk}{\mu_{dk}}^2}{2}\right]
\end{aligned}
$$

1.3 Sampling β

According to reference [**?**, **?**], the prior distribution of the number of variable clusters K can be written as

$$p(K|\beta,D)\propto c_D(K)D!\beta^K\frac{\Gamma(\beta)}{\Gamma(\beta+D)} \tag{15}$$

and $c_D(K)=p(K|\beta=1,D)$, can be computed by using recurrence formula for Stirling numbers. For $\beta>0$, the Gamma function in (15) can be computed as

$$\frac{\Gamma(\beta)}{\Gamma(\beta+D)}=\frac{(\beta+D)\mathcal{B}(\beta+1,D)}{\beta\Gamma(D)} \tag{16}$$

where $\mathcal{B}(\beta+1,D)$ is a Beta distribution with parameter $\beta+1,D$.

We put a Gamma prior on β and compute the posterior distribution of β as follows:

$$\begin{aligned} p(\beta|\boldsymbol{X}, K) &\propto p(\beta|K) \\ &\propto p(K|\beta)p(\beta) \\ &\propto \beta^K \frac{(\beta+D)\mathcal{B}(\beta+1, D)}{\beta\Gamma(D)} p(\beta) \end{aligned} \tag{17}$$

By introducing an auxiliary variable u, we can rewrite this in the form:

$$\begin{aligned} p(\beta|\boldsymbol{X}, K) &\propto \beta^K \frac{(\beta+D)\mathcal{B}(\beta+1, D)}{\beta\Gamma(D)} p(\beta) \\ &\propto p(\beta)\beta^{K-1}(\beta+D)\int_0^1 u^\beta (1-u)^{D-1} \mathrm{d}u \end{aligned} \tag{18}$$

Hence we can write the joint distribution of β and u as

$$p(\beta, u|x, K) \propto p(\beta)\beta^{K-1}(\beta+D)u^\beta(1-u)^{D-1} \qquad (\beta \in (0, +\infty], u \in (0,1)) \tag{19}$$

Given the Gamma prior $\mathcal{G}(a_\beta, b_\beta)$ for β, we calculate the conditional distribution of β and u as follows:

$$\begin{aligned} p(u|\beta, K) &\propto u^\beta(1-u)^{D-1} \qquad (u \in (0,1)) \\ p(\beta|K, u) &\propto \beta^{K-1}(\beta+D)u^\beta(1-u)^{D-1}\mathcal{G}(a_\beta, b_\beta) \\ &\propto \beta^{K-1}(\beta+D)u^\beta(1-u)^{D-1}\frac{b_\beta^{a_\beta}}{\Gamma(a_\beta)}\beta^{a_\beta-1}\exp(-\beta b_\beta) \\ &\propto \beta^{a_\beta+K-2}(\beta+D)e^{-\beta(b_\beta-\log(u))} \\ &\propto \beta^{a_\beta+K-1}e^{-\beta(b_\beta-\log(u))} + D\beta^{a_\beta+K-2}e^{-\beta(b_\beta-\log(u))} \\ &\propto \frac{\Gamma(a_\beta+K)}{[b_\beta-\log(u)]^{a_\beta+K}}\frac{[b_\beta-\log(u)]^{a_\beta+K}}{\Gamma(a_\beta+K)}\beta^{a_\beta+K-1}e^{-\beta(b_\beta-\log(u))} \\ &+ D\frac{\Gamma(a_\beta+K-1)}{[b_\beta-\log(u)]^{a_\beta+K-1}}\frac{[b_\beta-\log(u)]^{a_\beta+K-1}}{\Gamma(a_\beta+K-1)}\beta^{a_\beta+K-2}e^{-\beta(b_\beta-\log(u))} \\ &\propto \frac{\Gamma(a_\beta+K)}{[b_\beta-\log(u)]^{a_\beta+K}}\mathcal{G}(a_\beta+K, b_\beta-log(u)) \\ &+ D\frac{\Gamma(a_\beta+K-1)}{[b_\beta-\log(u)]^{a_\beta+K-1}}\mathcal{G}(a_\beta+K-1, b_\beta-log(u)) \end{aligned} \tag{20}$$

According to the conditional posteriors derived above, we can sample β and u as follows:

$$\begin{aligned} &(u|\beta, K) \sim \mathcal{B}(\beta+1, D) \\ &(\beta|u, K) \sim \pi_K\mathcal{G}(a_\beta+K, b_\beta-log(u)) + (1-\pi_K)\mathcal{G}(a_\beta+K-1, b_\beta-log(u)) \end{aligned} \tag{21}$$

where the posterior distribution of β is a mixture of two gamma densities with the weights π_K defined by

$$\frac{\pi_K}{(1-\pi_K)} = \frac{(a_\beta+K-1)}{D(b_\beta-log(u))} \tag{22}$$

1.4 Sampling the factor matrix F

Given our model is represented as follows:

$$\begin{aligned} p(\boldsymbol{f}_n) &= \mathcal{N}(\boldsymbol{f}_n|\boldsymbol{0}, \boldsymbol{I}) \\ p(\boldsymbol{x}_n) &= \mathcal{N}\left(\boldsymbol{x}_n|\boldsymbol{\mu} + (\boldsymbol{G} \odot \boldsymbol{Z})\boldsymbol{f}_n, \boldsymbol{\Psi}\right) \end{aligned} \tag{23}$$

We can calculate the joint distribution of $\boldsymbol{f}_n$ and $\boldsymbol{x}_n$ as follows:

$$p\left(\begin{bmatrix}\boldsymbol{f}_n \\ \boldsymbol{x}_n\end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{f}_n \\ \boldsymbol{x}_n\end{bmatrix} | \begin{bmatrix}\boldsymbol{0} \\ \boldsymbol{\mu}\end{bmatrix}, \begin{bmatrix}\boldsymbol{I} & (\boldsymbol{G} \odot \boldsymbol{Z})^T \\ (\boldsymbol{G} \odot \boldsymbol{Z}) & (\boldsymbol{G} \odot \boldsymbol{Z})(\boldsymbol{G} \odot \boldsymbol{Z})^T + \boldsymbol{\Psi}\end{bmatrix}\right) \tag{24}$$

Apply the Gaussian conditioning formulas to the joint distribution derived above, we can get the conditional distribution of the factor matrix $\boldsymbol{F}$ as follows:

$$\begin{aligned} p(\boldsymbol{f}_n|\boldsymbol{x}_n) &= \mathcal{N}(\boldsymbol{f}_n|\boldsymbol{\mu}_f, \boldsymbol{\Sigma}_f) \\ \boldsymbol{\Sigma}_f &= \boldsymbol{I} - (\boldsymbol{G} \odot \boldsymbol{Z})^T \left((\boldsymbol{G} \odot \boldsymbol{Z})(\boldsymbol{G} \odot \boldsymbol{Z})^T + \boldsymbol{\Psi}\right)^{-1} (\boldsymbol{G} \odot \boldsymbol{Z}) \\ \boldsymbol{\mu}_f &= (\boldsymbol{G} \odot \boldsymbol{Z})^T \left((\boldsymbol{G} \odot \boldsymbol{Z})(\boldsymbol{G} \odot \boldsymbol{Z})^T + \boldsymbol{\Psi}\right)^{-1} \boldsymbol{f}_n \end{aligned} \tag{25}$$

1.5 Sampling σ_ϵ^2

$\boldsymbol{\Psi} = [\sigma_\epsilon^2 \boldsymbol{I}]$ is a diagonal matrix. We place a inverse Gamma prior on σ_ϵ^2 as:

$$\begin{aligned} p(\sigma_\epsilon^2|a_\epsilon, b_\epsilon) &= \mathcal{IG}(a_\epsilon, b_\epsilon) \\ &= \frac{b_\epsilon^{a_\epsilon}}{\Gamma(a_\epsilon)} (\sigma_\epsilon^2)^{-a_\epsilon - 1} \exp\left[-\frac{b_\epsilon}{\sigma_\epsilon^2}\right] \end{aligned} \tag{26}$$

We can get that the posterior is also inverse Gamma distributed:

$$\begin{aligned} p(\sigma_\epsilon^2|x_{dn}, g_{dk}, z_{dk}, f_{kn}, a_\epsilon, b_\epsilon) &= \prod_{n=1}^{N}\prod_{d=1}^{D}\prod_{k=1}^{K} \frac{p(x_{dn}|g_{dk}, z_{dk}, f_{kn}, \sigma_\epsilon^2) p(\sigma_\epsilon^2|a_\epsilon, b_\epsilon)}{p(x_{dn}|g_{dk}, z_{dk}, f_{kn}, a_\epsilon, b_\epsilon)} \\ &\propto \prod_{n=1}^{N}\prod_{d=1}^{D}\prod_{k=1}^{K} p(x_{dn}|g_{dk}, z_{dk}, f_{kn}, \sigma_\epsilon^2) p(\sigma_\epsilon^2|a_\epsilon, b_\epsilon) \\ &= \prod_{n=1}^{N}\prod_{d=1}^{D}\prod_{k=1}^{K} \left(\frac{1}{(2\pi\sigma_\epsilon^2)} \exp\left[-\frac{(x_{dn} - g_{dk}f_{kn})^2}{2\sigma_\epsilon^2}\right]\right)^{z_{dk}} \frac{b_\epsilon^{a_\epsilon}}{\Gamma(a_\epsilon)} (\sigma_\epsilon^2)^{-a_\epsilon - 1} \exp\left[-\frac{b_\epsilon}{\sigma_\epsilon^2}\right] \\ &= (\sigma_\epsilon^2)^{-a_\epsilon - \frac{ND}{2} - 1} \exp\left[-\frac{b_\epsilon + \frac{\sum_k^K \left(\sum_{d=1,n=1}^{D,N} (x_{dn} - g_{dk}f_{kn})^2\right)^{z_{dk}}}{2}}{\sigma_\epsilon^2}\right] \\ &= \mathcal{IG}\left(a_\epsilon + \frac{ND}{2}, b_\epsilon + \frac{\sum_k^K \left(\sum_{d=1,n=1}^{D,N} (x_{dn} - g_{dk}f_{kn})^2\right)^{z_{dk}}}{2}\right) \\ &= \mathcal{IG}\left(a_\epsilon + \frac{ND}{2}, b_\epsilon + \frac{tr\left([\boldsymbol{X} - (\boldsymbol{G} \odot \boldsymbol{Z})\boldsymbol{F})]^T [\boldsymbol{X} - (\boldsymbol{G} \odot \boldsymbol{Z})\boldsymbol{F}]\right)}{2}\right) \end{aligned} \tag{27}$$

1.6 Sampling σ_g^2

σ_g^2 is given a inverse-Gamma prior $\mathcal{IG}(a_g, b_g)$. The local graphical model of the matrix $\boldsymbol{G}$ is depicted in Figure(3).

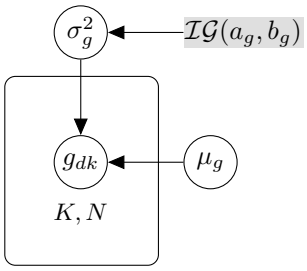

Figure 3: The Local Graphical Model for The Factor Loading Matrix

The posterior distribution of σ_g^2 can be calculated as:

$$
\begin{aligned}
p(\sigma_g^2|\boldsymbol{G}, \mu_g, a_g, b_g) &= \prod_{k=1}^{K}\prod_{d=1}^{D} p(\sigma_g^2|g_{dk}, \mu_g, a_g, b_g) \\
&= \prod_{k=1}^{K}\prod_{d=1}^{D} \frac{p(g_{dk}|\mu_g, \sigma_g^2)p(\sigma_g^2|a_g, b_g)}{p(g_{dk}|\mu_g, a_g, b_g)} \\
&\propto \prod_{k=1}^{K}\prod_{d=1}^{D} p(g_{dk}|\mu_g, \sigma_g^2)p(\sigma_g^2|a_g, b_g) \\
&\propto \prod_{k=1}^{K}\prod_{d=1}^{D} \frac{1}{(2\pi\sigma_g^2)} \exp\left[-\frac{(g_{dk}-\mu_g)^2}{2\sigma_g^2}\right] \frac{b_g^{a_g}}{\Gamma(a_g)}(\sigma_g^2)^{-a_g-1} \exp\left[-\frac{b_g}{\sigma_g^2}\right] \\
&= (\sigma_g^2)^{-a_g-\frac{DK}{2}-1} \exp\left[-\frac{b_g + \frac{\sum_{d=1,k=1}^{D,K}(g_{dk}-\mu_g)^2}{2}}{\sigma_g^2}\right] \\
&= \mathcal{IG}\left(a_g + \frac{DK}{2}, b_g + \frac{\sum_{d=1,k=1}^{D,K}(g_{dk}-\mu_g)^2}{2}\right)
\end{aligned}
$$

2 The Nonparametric Bayesian Mixture of Sparse Factor Analysers

Consider observed data $\boldsymbol{X} \in \mathbb{R}^{D \times N}$ where we have D measured variables for each data sample $\boldsymbol{x}_n$. Our goal is to cluster the samples into the different groups parameterised by group-specific covariance structure and mean, and for each group of samples, to partition the variables into the clusters where the clustered variables can be interpreted by common latent factors. Our generative model can be represented as:

$$
\begin{aligned}
x_{dn} &\sim \sum_k \pi_k \mathcal{N}\left(\mu_d^{(k)} + \sum_j (g_{dj}^{(k)} \circ z_{dj}^{(k)}) f_{jn}, {\sigma_d^{(k)}}^2\right) \\
c_n &\sim \text{Multi}(\pi_1, \ldots, \pi_K) \\
\pi_k &\sim \text{GEM}(\alpha) \\
\boldsymbol{\mu}^{(k)} &\sim \mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0) \\
g_{dj}^{(k)} &\sim \mathcal{N}(0, {\sigma_g^{(k)}}^2) \\
c_d^{(k)} &\sim \text{Multi}(\lambda_1^{(k)}, \ldots, \lambda_{\mathcal{J}}^{(k)}) \\
\lambda_j^{(k)} &\sim \text{GEM}(\beta) \\
z_{dj}^{(k)} &= \mathbb{1}(c_d = j) \\
f_{jn} &\sim \mathcal{N}(0, I)
\end{aligned}
\tag{28}
$$

where the samples are drawn from one of the mixture components indicated by variable c_n with the mixing proportions π_k as defined in classical Dirichlet process mixture model. Through factor modelling of each mixture component independently, we can reconstruct the covariances by the sampled factor loading matrices. In our model for the grouped samples linked to a particular mixture component, $c_d^{(k)}$, used to indicate the clustering of variables for samples associated with group k, is drawn from a Dirichlet Process and $g_{dj}^{(k)}$ is used to model the strength of association between variable d and factor j. By introducing indicator variable $z_{dj}^{(k)} = \mathbb{1}(c_d = j)$, the factor loading matrix can be written as $\boldsymbol{W}^{(k)} = \boldsymbol{G}^{(k)} \circ \boldsymbol{Z}^{(k)}$. We use a group-specific parameter $\sigma_d^{(k)}$ to model the idiosyncratic noise. The covariance associated with this component can be modelled as $\boldsymbol{\Sigma}^{(k)} = \boldsymbol{W}^{(k)} {\boldsymbol{W}^{(k)}}^T + {\sigma_d^{(k)}}^2 I_D$.

2.1 Hyperparameter setting

We put Inverse Gamma priors on ${\sigma_g^{(k)}}^2$ and ${\sigma_d^{(k)}}^2$ independently for each component k and put Gamma priors on α and β.

$$
\begin{aligned}
{\sigma_g^{(k)}}^2 &\sim \mathcal{IG}(a_g, b_g) \\
{\sigma_d^{(k)}}^2 &\sim \mathcal{IG}(a_d, b_d) \\
\alpha &\sim \mathcal{G}(a_\alpha, b_\alpha) \\
\beta &\sim \mathcal{G}(a_\beta, b_\beta)
\end{aligned}
\tag{29}
$$

3 Inference

We use a Gibbs sampler to exploit the posterior distribution over the model parameters $\boldsymbol{\mu}^{(k)}, g_{dj}^{(k)}, c_d^{(k)}, \boldsymbol{f}_n, c_n$ and the hyperparameters $\alpha, \beta, \sigma_d^{(k)}, \sigma_g^{(k)}$. The posterior distributions used in Gibbs sampling are summarized here.

Sampling the real-valued matrix $\boldsymbol{G}^{(j)}$:

As we place a normal distributed prior $\mathcal{N}(0, {\sigma_g^{(k)}}^2)$ on the entries $g_{dj}^{(k)}$ of $\boldsymbol{G}^{(k)}$ independently, we can sample $g_{dj}^{(k)}$ from its posterior as follows:

$$p(g_{dj}^{(k)}|\{\boldsymbol{x}_n, \boldsymbol{f}_n\}_{c_n=k}, {\sigma_g^{(k)}}^2, \boldsymbol{\mu}^{(k)}) \sim \mathcal{N}(\mu_{d,j}, \lambda_{d,j}^{-1}) \tag{30}$$

where

$$\begin{aligned}
\lambda_{d,j} &= \begin{cases} \frac{\sum_{c_n=k} f_{jn}^2}{{\sigma_d^{(k)}}^2} + \frac{1}{{\sigma_g^{(k)}}^2} & \text{if } c_d = j \\ \frac{1}{{\sigma_g^{(k)}}^2} & \text{otherwise} \end{cases} \\
\mu_{d,j} &= \begin{cases} \left(\frac{\sum_{c_n=k} f_{jn}^2}{{\sigma_d^{(k)}}^2} + \frac{1}{{\sigma_g^{(k)}}^2} \right)^{-1} \left(\frac{\sum_{c_n=k} (x_{dn} - \mu_d^{(k)}) f_{jn}}{{\sigma_d^{(k)}}^2} \right) & \text{if } c_d = j \\ 0 & \text{otherwise} \end{cases}
\end{aligned} \tag{31}$$

where $\mu_d^{(k)}$ denote the element d of the mean of mixture component k.

We place an inverse-Gamma prior $\mathcal{IG}(a_g, b_g)$ on the hyperparameter ${\sigma_g^{(k)}}^2$ and its posterior can be calculated as:

$$p({\sigma_g^{(k)}}^2|\boldsymbol{G}^{(k)}, a_g, b_g) = \mathcal{IG}\left(a_g + \frac{DJ^{(k)}}{2}, b_g + \frac{tr(\boldsymbol{G}^{(k)T}\boldsymbol{G}^{(k)})}{2}\right)$$

Sampling $\boldsymbol{Z}^{(j)}$:

We sample the entries $\{z_{dk}^{(j)}\}$ of matrix $\boldsymbol{Z}^{(j)}$ from its posterior for each mixture component independently:

$$\begin{aligned}
&p(z_{dk}^{(j)} = 1|\boldsymbol{x}_{(\mathcal{C}_j)}, g_{dk}^{(j)}, \boldsymbol{z}_{-dk}^{(j)}, f_{kn}, \beta, {\sigma_d^{(j)}}^2) \\
&\propto \prod_{n \in \mathcal{C}_j} p(x_{dn}|g_{dk}^{(j)}, z_{dk}^{(j)} = 1, f_{kn}, \beta, {\sigma_d^{(j)}}^2) p(z_{dk}^{(j)} = 1|\boldsymbol{z}_{-dk}^{(j)}, \beta) \\
&= \frac{1}{(2\pi {\sigma_d^{(j)}}^2)^{|\mathcal{C}_j|/2}} \exp\left[-\frac{\sum_{n \in \mathcal{C}_j} (x_{dn} - g_{dk}^{(j)} f_{kn})^2}{2{\sigma_d^{(j)}}^2} \right] \frac{m_{-dk}^{(j)}}{D - 1 + \beta}
\end{aligned} \tag{32}$$

where $\boldsymbol{z}_{-dk}$ denotes all the indicators except $\boldsymbol{z}_d = [z_{d1}, \ldots, z_{dK}]$, and $m_{-dk}^{(j)}$ denotes the number of variables associated with the latent factor k in component j. **Sampling $\boldsymbol{F}$:**

For sampling of the factor matrix $\boldsymbol{F}$, we sample the columns of $\boldsymbol{F}$ independently from its posterior $p(\boldsymbol{f}_n|\boldsymbol{x}_n, \boldsymbol{G}^{(c_n)}, \boldsymbol{Z}^{(c_n)}, \boldsymbol{\mu}^{(c_n)})$

$$\begin{aligned}
p(\boldsymbol{f}_n|\boldsymbol{x}_n) &= \mathcal{N}(\boldsymbol{f}_n|\mu_f^{(c_n)}, \Sigma_f^{(c_n)}) \\
\Sigma_f^{(c_n)} &= I - (G^{(c_n)} \odot Z^{(c_n)})^T \left((G^{(c_n)} \odot Z^{(c_n)})(G^{(c_n)} \odot Z^{(c_n)})^T + {\sigma_d^{(c_n)}}^2 I \right)^{-1} (G^{(c_n)} \odot Z^{(c_n)}) \\
\mu_f^{(c_n)} &= (G^{(c_n)} \odot Z^{(c_n)})^T \left((G^{(c_n)} \odot Z^{(c_n)})(G^{(c_n)} \odot Z^{(c_n)})^T + {\sigma_d^{(c_n)}}^2 I \right)^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}^{(c_n)})
\end{aligned} \tag{33}$$

Sampling the idiosyncratic noise term: We place an inverse-Gamma prior $\mathcal{IG}(a_d, b_d)$ over ${\sigma_d^{(k)}}^2$ and the posterior is also inverse-Gamma distributed:

$$p({\sigma_d^{(k)}}^2|.) \propto \mathcal{IG}\left(a_d + \frac{N^{(k)}D}{2}, b_d + \frac{tr\left(\boldsymbol{E}^T\boldsymbol{E}\right)}{2}\right)$$

where $\boldsymbol{E} = \sum_{c_n=k}(\boldsymbol{x}_n - \boldsymbol{\mu}^{(k)} - (G^{(k)} \odot Z^{(k)})\boldsymbol{f}_n)$.

Sampling $\{c_n\}$: We cluster each data point according to the posterior of the indicator variable c_n that can be calculated as

$$\begin{aligned} & p(c_n = j|\boldsymbol{x}_n, \boldsymbol{c}_{-n}, \alpha, \boldsymbol{\mu}^{(j)}, \boldsymbol{A}^{(j)}, \boldsymbol{\Psi}^{(j)}) \\ \propto\ & p(c_n = j|\boldsymbol{c}_{-n}, \alpha) p(\boldsymbol{x}_n|c_n = j, \boldsymbol{\mu}^{(j)}, \boldsymbol{A}^{(j)}, \boldsymbol{\Psi}^{(j)}, \boldsymbol{c}_{-n}) \\ \propto\ & \frac{N_{-n}^{(j)}|\boldsymbol{A}^{(j)}\boldsymbol{A}^{(j)T} + \boldsymbol{\Psi}^{(j)}|^{-\frac{1}{2}}}{N - 1 + \alpha} \cdot \\ & \cdot \exp\left(-\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}^{(j)})^T(\boldsymbol{A}^{(j)}\boldsymbol{A}^{(j)T} + \boldsymbol{\Psi}^{(j)})^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}^{(j)})\right) \end{aligned} \tag{34}$$

where $|\cdot|$ denotes the determinant of its matrix argument, $N_{-n}^{(j)}$ denotes the number of data samples excluding $\boldsymbol{x}_n$ that are associated with component j, and $\boldsymbol{c}_{-n}$ denotes all the indicators for data samples except c_n.

Sampling the idiosyncratic noise term: Since we placed an inverse Gamma prior $\mathcal{IG}(a_\epsilon, b_\epsilon)$ over ${\sigma_d^{(j)}}^2$, the posterior is also inverse Gamma distributed:

$$p({\sigma_d^{(j)}}^2|-) \propto \mathcal{IG}\left(a_\epsilon + \frac{N^{(j)}D}{2}, b_\epsilon + \frac{\mathrm{tr}\left(\Gamma^T\Gamma\right)}{2}\right) \tag{35}$$

where we use $-$ to denote the rest of the model parameters, and $\Gamma = \hat{\boldsymbol{x}}_{(\mathcal{C}_j)} - (G^{(j)} \odot Z^{(j)})\boldsymbol{f}_{(\mathcal{C}_j)}$.

Sampling hyperparameters: We sample the concentration parameter α from a joint distribution by introducing an auxiliary variable u as

$$\begin{aligned} p(\alpha|u, J) \sim\ & \gamma_J \mathcal{G}(a_\alpha + J, b_\alpha - \log(u)) \\ & + (1 - \gamma_J)\mathcal{G}(a_\alpha + J - 1, b_\alpha - \log(u)) \\ p(u|\alpha, N) \sim\ & \mathrm{Beta}(\alpha + 1, N) \end{aligned} \tag{36}$$

with the weights γ_J defined by

$$\frac{\gamma_J}{(1 - \gamma_J)} = \frac{a_\alpha + J - 1}{N(b_\alpha - \log(u))} \tag{37}$$

where N denotes the number of data samples and $\log$ denotes the natural logarithm of its argument.

We use the same sampling procedure to sample the second-layer DP concentration parameter β conditioned upon the maximum number of latent factors in the mixture distribution K and data dimension D.