Mixtures of multivariate Poisson lognormal factor analyzers for high dimensional RNAseq data

Abstract

1 Introduction

2 Methodology

2.1 Mixtures of Factor Analyzer

We assume the following hierarchical structure:

$$
\begin{aligned}
&Y_{ij} \mid X_{ij} = x_{ij} \sim \text{Poisson}\left(\exp\{x_{ij} + \log k_j\}\right)\\
&\mathbf{X}_i \mid \mathbf{U}_i = \mathbf{u}_i \sim N(\boldsymbol{\mu}_g + \boldsymbol{\Lambda}_g \mathbf{u}_i, \boldsymbol{\Psi}_g)\\
&\mathbf{U}_i \sim N(\mathbf{0}_p, \mathbf{I}_p).
\end{aligned}
$$

In model-based clustering, we also have unobserved component membership indicator variable $\mathbf{Z}$ such that $Z_{ig} = 1$ if the observation i^{th} belongs to group g and $Z_{ig} = 0$ otherwise.

Hence, the complete data now comprises of observed expression levels $\mathbf{y}$, underlying latent variable $\boldsymbol{\theta}$, and unknown group membership $\mathbf{z}$.

The complete data likelihood can therefore be written as:

$$L(\boldsymbol{\vartheta}) = \prod_{g=1}^{G}\prod_{i=1}^{n}\left[\pi_g \left\{\prod_{j=1}^{d} f(y_{ij} \mid x_{ij}, k_j)\right\} \; f(\mathbf{x}_i \mid \mathbf{u}_i, \boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g) \; f(\mathbf{u}_i)\right]^{z_{ig}}$$

where $\boldsymbol{\vartheta}$ denotes all the model parameters. The complete data log-likelihood can therefore be written as:

$$l_c(\boldsymbol{\vartheta}) = \sum_{g=1}^{G}\sum_{i=1}^{n} z_{ig}\left[\log \pi_g + \left\{\sum_{j=1}^{d} \log f(y_{ij} \mid x_{ij}, k_j)\right\} + \log f(\mathbf{x}_i \mid \mathbf{u}_i, \boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g) + \log f(\mathbf{u}_i)\right]$$

In order to utilize the EM framework for parameter estimation, we require $\mathbb{E}(Z_{ig}\mathbf{X}_i \mid \mathbf{y}_i, \mathbf{u}_i, \boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g)$, $\mathbb{E}(Z_{ig}\mathbf{X}_i\mathbf{X}_i' \mid \mathbf{y}_i, \mathbf{u}_i, \boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g)$, $\mathbb{E}(Z_{ig}\mathbf{U}_i \mid \mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g)$ and $\mathbb{E}(Z_{ig}\mathbf{U}_i\mathbf{U}_i' \mid \mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g)$.

2.2 Variational approximation of mixtures of MPLN factor analyzers

Recently, Subedi and Browne (2020) proposed an alternate framework for parameter estimation utilizing variational Gaussian approximation (VGA). Variational approximation (Wainwright et al., 2008) is an approximate inference technique that uses a computationally convenient approximating density in place of a more complex but 'true' posterior density obtained by minimizing the Kullback-Leibler (KL) divergence between the true and the approximating densities. This alleviates the computational cost that comes with MCMC-EM algorithm. Here, we propose a mixtures of Poisson log-normal distributions with factor analyzers that utilizes a variational EM framework for parameter estimation.

Parameter estimation for a mixtures of factor analyzers is typically done using alternating expectation conditional maximization framework that assumes different specification of missing data at different stages. Here, we will adapt a similar notion resulting in a two stage iterative EM-type approach.

2.3 Stage 1

In Stage 1, we treat $\mathbf{X}$ and $\mathbf{Z}$ as missing such that

$$\begin{aligned}
&Y_{ij} \mid X_{ij} = x_{ij} \sim \text{Poisson}\left(\exp\{x_{ij} + \log k_j\}\right)\\
&\mathbf{X}_i \sim N(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g),
\end{aligned}$$

where $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g^T + \boldsymbol{\Psi}_g$.Therefore, the component specific marginal density of the observed data $\mathbf{y}_i$ can be written as

$$f(\mathbf{y}_i \mid Z_{ig} = 1, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) = \int_{\mathbb{R}^d} \left[\prod_{j=1}^{d} f_p(y_{ij} \mid x_{ij}, k_j, Z_{ig} = 1)\right] f_N(\mathbf{x}_i \mid Z_{ig} = 1, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)\; d\mathbf{x}_i,$$

where x_{ij} and y_{ij} are the j^{th} element of $\mathbf{x}_i$ and $\mathbf{y}_i$ respectively, $f_p(\cdot)$ is the probability mass function of the Poisson distribution with mean $e^{x_{ij}+\log k_j}$ and $f_N(\cdot)$ is the probability density function of d-dimensional Gaussian distribution with mean $\boldsymbol{\mu}_g$ and covariance $\boldsymbol{\Sigma}_g$. Note that the marginal distribution of Y involves multiple integrals and cannot be further simplified. The the complete-data log-likelihood using the marginal density of $\mathbf{y}_i \mid Z_{ig} = 1$ can be written as:

$$l(\boldsymbol{\vartheta}) = \sum_{g=1}^{G}\sum_{i=1}^{n} z_{ig} \log \pi_g + \sum_{g=1}^{G}\sum_{i=1}^{n} z_{ig} \log f(\mathbf{y}_i \mid Z_{ig} = 1, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g).$$

As the marginal of $\mathbf{y}_i \mid Z_{ig} = 1$ cannot be simplified, we use variational Gaussian approximation to approximate the marginal log-density of the observed variable $\mathbf{y}_i$ from component g in the mixtures of MPLN distributions. Suppose, we have an approximating density $q(\mathbf{x}_{ig})$, the marginal log-density of $\mathbf{y}_i \mid Z_{ig} = 1$ can be written as

$$\begin{aligned}\log f(\mathbf{y}_i \mid Z_{ig} = 1) &= \int_{\mathbb{R}^d} \log f(\mathbf{y}_i \mid Z_{ig} = 1)\ q(\mathbf{x}_{ig})\ d\mathbf{x}_{ig} \\ &= \int_{\mathbb{R}^d} \log \frac{f(\mathbf{y}_i, \mathbf{x}_i \mid Z_{ig} = 1)/q(\mathbf{x}_{ig})}{f(\mathbf{x}_i \mid \mathbf{y}_i, Z_{ig} = 1)/q(\mathbf{x}_{ig})}\ q(\mathbf{x}_{ig})\ d\mathbf{x}_{ig} \\ &= \int_{\mathbb{R}^d} \left[\log f(\mathbf{y}_i, \mathbf{x}_i \mid Z_{ig} = 1) - \log q(\mathbf{x}_{ig})\right] q(\mathbf{x}_{ig})\ d\mathbf{x}_{ig} + D_{KL}(q_{ig} \| f_{ig}) \\ &= F(q_{ig}, \mathbf{y}_i) + D_{KL}(q_{ig} \| f_{ig}),\end{aligned}$$

where $D_{KL}(q_{ig}\|f_{ig}) = \int_{\mathbb{R}^d} q(\mathbf{x}_{ig}) \log \frac{q(\mathbf{x}_{ig})}{f(\mathbf{x}_i|\mathbf{y}_i, Z_{ig}=1)} d\mathbf{x}_{ig}$ is the Kullback-Leibler (KL) divergence between $f(\mathbf{x}_i \mid \mathbf{y}_i, Z_{ig} = 1)$ and approximating distribution $q(\mathbf{x}_{ig})$, and

$$F(q_{ig}, \mathbf{y}_i) = \int_{\mathbb{R}^d} \left[\log f(\mathbf{y}_i, \mathbf{x}_i \mid Z_{ig} = 1) - \log q(\mathbf{x}_{ig})\right] q(\mathbf{x}_{ig}) d\mathbf{x}_{ig},$$

is our evidence lower bound (ELBO) for each observation $\mathbf{y}_i$. The complete data log-likelihood of the mixtures of MPLN distributions can be written as:

$$l_c(\boldsymbol{\vartheta} \mid \mathbf{y}) = \sum_{g=1}^{G} \sum_{i=1}^{n} z_{ig} \log \pi_g + \sum_{g=1}^{G} \sum_{i=1}^{n} z_{ig} \left[F(q_{ig}, \mathbf{y}_i) + D_{KL}(q_{ig} \| f_{ig})\right].$$

To minimize the KL divergence, we maximize our ELBO. In VGA, $q(\mathbf{x}_{ig})$ is assumed to be a Gaussian distribution. Assuming $q(\mathbf{x}_{ig}) = \mathscr{N}(\mathbf{m}_{ig}, \mathbf{S}_{ig})$, the ELBO for each observation $\mathbf{y}_i$

becomes

$$F(q_{ig},\mathbf{y}_i) = -\frac{1}{2}(\mathbf{m}_{ig}-\boldsymbol{\mu}_g)'\boldsymbol{\Sigma}_g^{-1}(\mathbf{m}_{ig}-\boldsymbol{\mu}_g) - \frac{1}{2}\operatorname{tr}(\boldsymbol{\Sigma}_g^{-1}\mathbf{S}_{ig}) + \frac{1}{2}\log|\mathbf{S}_{ig}| - \frac{1}{2}\log|\boldsymbol{\Sigma}_g| + \frac{d}{2} + \mathbf{m}_{ig}'\mathbf{y}_i + \sum_{j=1}^{d}(\log k_j)\,y_{ij} - \sum_{j=1}^{d}\left\{e^{\log k_j + m_{igj} + \frac{1}{2}S_{ig,jj}} + \log(y_{ij}!)\right\}.$$

This lower bound is strictly jointly concave with respect to the mean ($\mathbf{m}_{ig}$) and variance ($\mathbf{S}_{ig}$) of the approximating distribution and hence, similar to Arridge et al. (2018) and Subedi and Browne (2020), estimation can be obtained via Newton's method and fixed-point method.

Therefore, in stage 1, we update $\mathbb{E}(Z_{ig} \mid \mathbf{y}_i)$, variational parameters $\mathbf{m}_{ig}$ and $\mathbf{S}_{ig}$, and model parameters π_g and $\boldsymbol{\mu}_g$.

1. Conditional on the variational parameters $\mathbf{m}_{ig}, \mathbf{S}_{ig}$ and on $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$, the $\mathbb{E}(Z_{ig})$ is computed. Given $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$,

   $$\mathbb{E}(Z_{ig} \mid \mathbf{y}_i) = \frac{\pi_g f(\mathbf{y} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)}{\sum_{h=1}^{G}\pi_h f(\mathbf{y} \mid \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)}.$$

   Note that this involves the marginal distribution of Y which is difficult to compute. Hence, we use an approximation of $\mathbb{E}(Z_{ig})$ where we replace the marginal density of the exponent of ELBO such that

   $$\widehat{Z}_{ig} \overset{\text{def}}{=} \frac{\pi_g \exp\left[F\left(q_{ig},\mathbf{y}_i\right)\right]}{\sum_{h=1}^{G}\pi_h \exp\left[F\left(q_{ih},\mathbf{y}_i\right)\right]}.$$

   This approximation is computationally convenient and a similar framework was utilized by Tang et al. (2015); Gollini and Murphy (2014).

2. Update the variationalparameters $\mathbf{m}_{ig}$ and $\mathbf{S}_{ig}$ as following:

(a) Fixed-point method for updating $\mathbf{S}_{ig}$ is

$$\mathbf{S}_{ig}^{(t+1)} = \left\{ \boldsymbol{\Sigma}_g^{-1} + \mathbf{I} \odot \mathbf{exp} \left[\log k_j + \mathbf{m}_{ig}^{(t)} + \frac{1}{2} \operatorname{diag} \left(\mathbf{S}_{ig}^{(t)} \right) \right] \mathbf{1}_d' \right\}^{-1}$$

where the vector function $\mathbf{exp}\,[\mathbf{a}] = (e^{a_1}, \ldots, e^{a_d})'$ is a vector of exponential each element of the d-dimensional vector $\mathbf{a}$, diag($\mathbf{S}$) $= (\mathbf{S}_{11} \ldots, \mathbf{S}_{dd})$ puts the diagonal elements of the $d \times d$ matrix $\mathbf{S}$ into a d-dimensional vector, $\odot$ the Hadmard product and $\mathbf{1}_d$ is a d-dimensional vector of ones ;

(b) Newton's method to update $\mathbf{m}_{ig}$ is

$$\mathbf{m}_{ig}^{(t+1)} = \mathbf{m}_{ig}^{(t)} - \mathbf{S}_{ig}^{(t+1)} \left\{ \mathbf{y}_i - \mathbf{exp} \left[\log k_j + \mathbf{m}_{ig}^{(t)} + \frac{1}{2} \operatorname{diag} \left(\mathbf{S}_{ig}^{(t+1)} \right) \right] - \boldsymbol{\Sigma}_g^{-1} \left(\mathbf{m}_{ig}^{(t)} - \boldsymbol{\mu}_g \right) \right\}.$$

3. Given $\widehat{Z}_{ig}$ and the variational parameters $\mathbf{m}_{ig}$ and $\mathbf{S}_{ig}$, the parameters $\boldsymbol{\pi}$ and $\boldsymbol{\mu}_g$ are updated as:

$$\widehat{\pi}_g = \frac{\sum_{i=1}^n \widehat{Z}_{ig}}{n},$$
$$\widehat{\boldsymbol{\mu}}_g = \frac{\sum_{i=1}^n \widehat{Z}_{ig} \mathbf{m}_{ig}^{(t+1)}}{\sum_{i=1}^n \widehat{Z}_{ig}}.$$

2.4 Stage 2

In Stage 2, we treat $\mathbf{X}$, $\mathbf{U}$ and $\mathbf{Z}$ as missing such that

$$Y_{ij} \mid X_{ij} = x_{ij} \sim \text{Poisson}\left(\exp\{x_{ij} + \log k_j\}\right)$$
$$\mathbf{X}_i \mid \mathbf{U}_i = \mathbf{u}_i \sim N(\boldsymbol{\mu}_g + \boldsymbol{\Lambda}_g \mathbf{u}_i, \boldsymbol{\Psi}_g),$$
$$\mathbf{U}_i \sim N(\mathbf{0}_p, \mathbf{I}_p).$$

Here, we use d for the dimensionality of the observed data and p as the dimensionality of the latent variable. Suppose, we have an approximating density $q_{\mathbf{x},\mathbf{u}}(\mathbf{x}_{ig}, \mathbf{u}_{ig})$, the marginal log-density of $\mathbf{y}_i \mid Z_{ig} = 1$ can be written as

$$
\begin{aligned}
\log f(\mathbf{y}_i \mid Z_{ig} = 1) &= \int_{\mathbb{R}^p} \int_{\mathbb{R}^d} \log f(\mathbf{y}_i \mid Z_{ig} = 1)\ q_{\mathbf{x},\mathbf{u}}(\mathbf{x}_{ig}, \mathbf{u}_{ig})\ d\mathbf{x}_{ig}\ d\mathbf{u}_{ig} \\
&= \int_{\mathbb{R}^p} \int_{\mathbb{R}^d} \log \frac{f(\mathbf{y}_i, \mathbf{x}_i, \mathbf{u}_i \mid Z_{ig} = 1)/q_{\mathbf{x},\mathbf{u}}(\mathbf{x}_{ig}, \mathbf{u}_{ig})}{f(\mathbf{x}_i, \mathbf{u}_i \mid \mathbf{y}_i, Z_{ig} = 1)/q_{\mathbf{x},\mathbf{u}}(\mathbf{x}_{ig}, \mathbf{u}_{ig})}\ q_{\mathbf{x},\mathbf{u}}(\mathbf{x}_{ig}, \mathbf{u}_{ig})\ d\mathbf{x}_{ig} \\
&= \int_{\mathbb{R}^p} \int_{\mathbb{R}^d} \Big[\log f(\mathbf{y}_i, \mathbf{x}_i, \mathbf{u}_i \mid Z_{ig} = 1) - \log q_{\mathbf{x},\mathbf{u}}(\mathbf{x}_{ig}, \mathbf{u}_{ig})\Big] q_{\mathbf{x},\mathbf{u}}(\mathbf{x}_{ig}, \mathbf{u}_{ig})\ d\mathbf{x}_{ig}\ d\mathbf{u}_{ig} \\
&+ D_{KL}(q_{ig} \| f_{ig}) \\
&= F(q_{ig}, \mathbf{y}_i) + D_{KL}(q_{ig} \| f_{ig}),
\end{aligned}
$$

where $D_{KL}(q_{ig} \| f_{ig}) = \int_{\mathbb{R}^p} \int_{\mathbb{R}^d} q_{\mathbf{x},\mathbf{u}}(\mathbf{x}_{ig}, \mathbf{u}_{ig}) \log \frac{q_{\mathbf{x},\mathbf{u}}(\mathbf{x}_{ig}, \mathbf{u}_{ig})}{f(\mathbf{x}_i, \mathbf{u}_i \mid \mathbf{y}_i, Z_{ig}=1)}\ d\mathbf{x}_{ig}\ d\mathbf{u}_{ig}$ is the Kullback-Leibler (KL) divergence between $f(\mathbf{x}_i, \mathbf{u}_i \mid \mathbf{y}_i, Z_{ig} = 1)$ and approximating distribution $q(\mathbf{x}_{ig}, \mathbf{u}_{ig})$, and

$$
F(q_{ig}, \mathbf{y}_i) = \int_{\mathbb{R}^p} \int_{\mathbb{R}^d} \Big[\log f(\mathbf{y}_i, \mathbf{x}_i, \mathbf{u}_i \mid Z_{ig} = 1) - \log q_{\mathbf{x},\mathbf{u}}(\mathbf{x}_{ig}, \mathbf{u}_{ig})\Big] q_{\mathbf{x},\mathbf{u}}(\mathbf{x}_{ig}, \mathbf{u}_{ig})\ d\mathbf{x}_{ig}\ d\mathbf{u}_{ig},
$$

is our evidence lower bound (ELBO) for each observation $\mathbf{y}_i$. If we assume $q_{\mathbf{x},\mathbf{u}}(\mathbf{x}_{ig}, \mathbf{u}_{ig})$ to be factorable such that $q_{\mathbf{x},\mathbf{u}}(\mathbf{x}_{ig}, \mathbf{u}_{ig}) = q_{\mathbf{x}}(\mathbf{x}_{ig}) q_{\mathbf{u}}(\mathbf{u}_{ig})$,

$$\begin{aligned}
F(q_{ig}, \mathbf{y}_i) = & \int_{\mathbb{R}^p} \int_{\mathbb{R}^d} \Big[\log f(\mathbf{y}_i \mid \mathbf{x}_i, \mathbf{u}_i, Z_{ig} = 1) + \log f(\mathbf{x}_i \mid \mathbf{u}_i, Z_{ig} = 1) + \log f(\mathbf{u}_i \mid Z_{ig} = 1) \\
& - \log q_{\mathbf{x}}(\mathbf{x}_{ig}) - \log q_{\mathbf{u}}(\mathbf{u}_{ig}) \Big] \; q_{\mathbf{x}}(\mathbf{x}_{ig}) \; q_{\mathbf{u}}(\mathbf{u}_{ig}) \; d\mathbf{x}_{ig} \; d\mathbf{u}_{ig} \\
& = \int_{\mathbb{R}^p} \bigg\{ \int_{\mathbb{R}^d} \Big[\log f(\mathbf{y}_i \mid \mathbf{x}_i, \mathbf{u}_i, Z_{ig} = 1) + \log f(\mathbf{x}_i \mid \mathbf{u}_i, Z_{ig} = 1) + \log f(\mathbf{u}_i \mid Z_{ig} = 1) \\
& - \log q_{\mathbf{x}}(\mathbf{x}_{ig}) - \log q_{\mathbf{u}}(\mathbf{u}_{ig}) \Big] \; q_{\mathbf{x}}(\mathbf{x}_{ig}) \; d\mathbf{x}_{ig} \bigg\} \; q_{\mathbf{u}}(\mathbf{u}_{ig}) \; d\mathbf{u}_{ig} \\
& = \int_{\mathbb{R}^p} \Bigg[\bigg\{ \int_{\mathbb{R}^d} \Big[\log f(\mathbf{y}_i \mid \mathbf{x}_i, \mathbf{u}_i, Z_{ig} = 1) + \log f(\mathbf{x}_i \mid \mathbf{u}_i, Z_{ig} = 1) - \log q_{\mathbf{x}}(\mathbf{x}_i) \Big] \; q_{\mathbf{x}}(\mathbf{x}_{ig}) \; d\mathbf{x}_{ig} \bigg\} \\
& + \log f(\mathbf{u}_i \mid Z_{ig} = 1) - \log q_{\mathbf{u}}(\mathbf{u}_{ig}) \Bigg] q_{\mathbf{u}}(\mathbf{u}_{ig}) \; d\mathbf{u}_{ig}.
\end{aligned}$$

Using the approximating density $q_{\mathbf{x}}(\mathbf{x}_{ig})$ from Stage 1 (i.e. $q_{\mathbf{x}}(\mathbf{x}_{ig}) = N(\mathbf{m}_{ig}, \mathbf{S}_{ig})$) and assuming $q_{\mathbf{u}}(\mathbf{u}_{ig}) = N(\mathbf{P}_{ig}, \mathbf{Q}_{ig})$, we get

$$\begin{aligned}
F(q_{ig},\mathbf{y}_i) &= \int_{\mathbb{R}^p}\Bigg[\mathbf{m}'_{ig}\mathbf{y}_i + \sum_{j=1}^{d}\left(\log k_j\right)y_{ij} - \sum_{j=1}^{d}\left\{e^{\log k_j + m_{igj} + \frac{1}{2}S_{ig,jj}} + \log(y_{ij}!)\right\} + \frac{1}{2}\log|\mathbf{S}_{ig}| + \frac{d}{2} \\
&\quad - \frac{1}{2}(\mathbf{m}_{ig} - \boldsymbol{\mu}_g - \boldsymbol{\Lambda}_g\mathbf{u}_i)'\boldsymbol{\Psi}_g^{-1}(\mathbf{m}_{ig} - \boldsymbol{\mu}_g - \boldsymbol{\Lambda}_g\mathbf{u}_i) - \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Psi}_g^{-1}\mathbf{S}_{ig}\right) - \frac{1}{2}\log|\boldsymbol{\Psi}_g| \\
&\quad + \log f(\mathbf{u}_i \mid Z_{ig} = 1) - \log q_{\mathbf{u}}(\mathbf{u}_{ig})\Bigg]q_{\mathbf{u}}(\mathbf{u}_{ig})\ d\mathbf{u}_{ig} \\
&= \int_{\mathbb{R}^p}\Bigg[\mathbf{m}'_{ig}\mathbf{y}_i + \sum_{j=1}^{d}\left(\log k_j\right)y_{ij} - \sum_{j=1}^{d}\left\{e^{\log k_j + m_{igj} + \frac{1}{2}S_{ig,jj}} + \log(y_{ij}!)\right\} + \frac{1}{2}\log|\mathbf{S}_{ig}| + \frac{d}{2} \\
&\quad - \frac{1}{2}(\mathbf{m}_{ig} - \boldsymbol{\mu}_{ig})^T\boldsymbol{\Psi}_g^{-1}(\mathbf{m}_{ig} - \boldsymbol{\mu}_{ig}) + (\mathbf{m}_{ig} - \boldsymbol{\mu}_g)^T\boldsymbol{\Psi}_g^{-1}\boldsymbol{\Lambda}_g\mathbf{u}_{ig} - \frac{1}{2}\mathbf{u}_{ig}^T\boldsymbol{\Lambda}_g^T\boldsymbol{\Psi}_g^{-1}\boldsymbol{\Lambda}_g\mathbf{u}_{ig} \\
&\quad - \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Psi}_g^{-1}\mathbf{S}_{ig}\right) - \frac{1}{2}\log|\boldsymbol{\Psi}_g| + \log f(\mathbf{u}_i \mid Z_{ig} = 1) - \log q_{\mathbf{u}}(\mathbf{u}_{ig})\Bigg]q_{\mathbf{u}}(\mathbf{u}_{ig})\ d\mathbf{u}_{ig} \\
&= \mathbf{m}'_{ig}\mathbf{y}_i + \sum_{j=1}^{d}\left(\log k_j\right)y_{ij} - \sum_{j=1}^{d}\left\{e^{\log k_j + m_{igj} + \frac{1}{2}S_{ig,jj}} + \log(y_{ij}!)\right\} + \frac{1}{2}\log|\mathbf{S}_{ig}| + \frac{d}{2} \\
&\quad - \frac{1}{2}(\mathbf{m}_{ig} - \boldsymbol{\mu}_{ig})^T\boldsymbol{\Psi}_g^{-1}(\mathbf{m}_{ig} - \boldsymbol{\mu}_{ig}) + (\mathbf{m}_{ig} - \boldsymbol{\mu}_g)^T\boldsymbol{\Psi}_g^{-1}\boldsymbol{\Lambda}_g\mathbf{P}_{ig} - \frac{1}{2}\mathbf{P}_{ig}^T\boldsymbol{\Lambda}_g^T\boldsymbol{\Psi}_g^{-1}\boldsymbol{\Lambda}_g\mathbf{P}_{ig} \\
&\quad - \frac{1}{2}\operatorname{tr}(\boldsymbol{\Lambda}_g^T\boldsymbol{\Psi}_g^{-1}\boldsymbol{\Lambda}_g\mathbf{Q}_{ig}) - \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Psi}_g^{-1}\mathbf{S}_{ig}\right) - \frac{1}{2}\log|\boldsymbol{\Psi}_g| - \frac{1}{2}\mathbf{P}_{ig}^T\mathbf{P}_{ig} + \frac{1}{2}\log|\mathbf{Q}_{ig}^{-1}| \\
&\quad - \frac{1}{2}\operatorname{tr}(\mathbf{Q}_{ig}) + \frac{p}{2}.
\end{aligned}$$

The variational parameters $\mathbf{P}_{ig}$ and $\mathbf{Q}_{ig}$ that maximize this ELBO will also minimize the KL divergence. Therefore, in Stage 2, we update variational parameters $\mathbf{P}_{ig}$ and $\mathbf{Q}_{ig}$, and model parameters $\boldsymbol{\Lambda}_g$ and $\boldsymbol{\Psi}_g$.

1. The update for the variational parameters $\mathbf{P}_{ig}$ and $\mathbf{Q}_{ig}$ are:

$$\begin{aligned}
\widehat{\mathbf{Q}}_{ig}^{(t+1)} &= (\mathbf{I}_p + \hat{\boldsymbol{\Lambda}}_g^T \hat{\boldsymbol{\Psi}}_g^{-1} \hat{\boldsymbol{\Lambda}}_g)^{-1}\\
&= \mathbf{I}_p - \hat{\boldsymbol{\beta}}_g\hat{\boldsymbol{\Lambda}}_g(\text{after a bit of simplification where } \hat{\boldsymbol{\beta}}_g = \hat{\boldsymbol{\Lambda}}_g^T(\hat{\boldsymbol{\Psi}}_g + \hat{\boldsymbol{\Lambda}}_g\hat{\boldsymbol{\Lambda}}_g^T)^{-1}).\\
\widehat{\mathbf{P}}_{ig}^{(t+1)} &= \widehat{\mathbf{Q}}_{ig}^{(t+1)}\hat{\boldsymbol{\Lambda}}_g^T\hat{\boldsymbol{\Psi}}_g^{-1}(\mathbf{m}_{ig}^{(t+1)} - \hat{\boldsymbol{\mu}}_g)\\
&= \hat{\boldsymbol{\beta}}_g(\mathbf{m}_{ig}^{(t+1)} - \hat{\boldsymbol{\mu}}_g)\,(\text{after a bit of simplification})\,.
\end{aligned}$$

2. The update for $\boldsymbol{\Psi}_g$ and $\boldsymbol{\Lambda}_g$ are obtained by repeating the following steps until convergence:

$$\begin{aligned}
\widehat{\boldsymbol{\Theta}}_g &= \mathbf{I}_p - \hat{\boldsymbol{\beta}}_g\hat{\boldsymbol{\Lambda}}_g + \hat{\boldsymbol{\beta}}_g\mathbf{W}_g\hat{\boldsymbol{\beta}}_g^T\\
\hat{\boldsymbol{\beta}}_g &= \hat{\boldsymbol{\Lambda}}_g^T(\hat{\boldsymbol{\Lambda}}_g\hat{\boldsymbol{\Lambda}}_g^T + \hat{\boldsymbol{\Psi}}_g)^{-1}\\
\hat{\boldsymbol{\Lambda}}_g &= \mathbf{W}_g\hat{\boldsymbol{\beta}}^T\widehat{\boldsymbol{\Theta}}_g^{-1}. \qquad (1)\\
\hat{\boldsymbol{\Psi}}_g &= \text{ diag}\left(\mathbf{W}_g - \hat{\boldsymbol{\Lambda}}_g\hat{\boldsymbol{\beta}}_g\mathbf{W}_g + \frac{\sum_{i=1}^n \widehat{Z}_{ig}\mathbf{S}_{ig}}{\sum_{i=1}^n \widehat{Z}_{ig}}\right) \qquad (2)
\end{aligned}$$

where $\mathbf{W}_g = \frac{\sum_{i=1}^n \widehat{Z}_{ig}(\mathbf{m}_{ig}^{(t+1)}-\hat{\boldsymbol{\mu}}_g)(\mathbf{m}_{ig}^{(t+1)}-\hat{\boldsymbol{\mu}}_g)^T}{\sum_{i=1}^n \widehat{Z}_{ig}}$Note: convergence is assumed when the difference in Frobenius norm of $\hat{\boldsymbol{\Lambda}}_g$ and $\hat{\boldsymbol{\Psi}}_g$ from successive iterations are both less than 10^{-6}.

3 Family of models

Constraints can be imposed on $\boldsymbol{\Lambda}_g$ and $\boldsymbol{\Psi}_g$ as described in Table 1 that results in a family of 8 models.

Table 1: Family of models based on the constraints on $\boldsymbol{\Lambda}_g$ and $\boldsymbol{\Psi}_g$.

Model ID	Loading Matrix $\boldsymbol{\Lambda}_g$	Error Variance $\boldsymbol{\Psi}_g$	Isotropic $\psi_g\mathbf{I}$	Covariance parameters
UUU	unconstrained	unconstrained	unconstrained	$G\left[dq-q\left(q-1\right)/2\right]+Gd$
UUC	unconstrained	unconstrained	constrained	$G\left[dq-q\left(q-1\right)/2\right]+G$
UCU	unconstrained	constrained	unconstrained	$G\left[dq-q\left(q-1\right)/2\right]+d$
UCC	unconstrained	constrained	constrained	$G\left[dq-q\left(q-1\right)/2\right]+1$
CUU	constrained	unconstrained	unconstrained	$\left[dq-q\left(q-1\right)/2\right]+Gd$
CUC	constrained	unconstrained	constrained	$\left[dq-q\left(q-1\right)/2\right]+G$
CCU	constrained	constrained	unconstrained	$\left[dq-q\left(q-1\right)/2\right]+d$
CCC	constrained	constrained	constrained	$\left[dq-q\left(q-1\right)/2\right]+1$

3.1 Isotropic assumption: $\boldsymbol{\Psi}_g = \psi_g\mathbf{I}_d$

$$\begin{aligned}
l(\boldsymbol{\vartheta}) &= \sum_{g=1}^{G}\sum_{i=1}^{n} z_{ig}\log\pi_g + \sum_{g=1}^{G}\sum_{i=1}^{n} z_{ig}\log f(\mathbf{y}_i \mid Z_{ig}=1, \boldsymbol{\mu}_g, \boldsymbol{\Psi}_g, \boldsymbol{\Lambda}_g).\\
&= \sum_{g=1}^{G}\sum_{i=1}^{n} z_{ig}\log\pi_g + \sum_{g=1}^{G}\sum_{i=1}^{n} z_{ig}\left\{F(q_{ig},\mathbf{y}_i) + D_{KL}(q_{ig} \mid\mid f_{ig})\right\}
\end{aligned}$$

Under the isotropic constraint (i.e. $\boldsymbol{\Psi}_g = \psi_g\mathbf{I}_d$),

$$\begin{aligned}
\sum_{g=1}^{G}\sum_{i=1}^{n} z_{ig}F(q_{ig},\mathbf{y}_i) &= \sum_{g=1}^{G}\sum_{i=1}^{n} z_{ig}\left[-\frac{\psi_g^{-1}}{2}(\mathbf{m}_{ig}-\boldsymbol{\mu}_{ig})^T(\mathbf{m}_{ig}-\boldsymbol{\mu}_{ig}) + \psi_g^{-1}(\mathbf{m}_{ig}-\boldsymbol{\mu}_g)^T\boldsymbol{\Lambda}_g\mathbf{P}_{ig}\right.\\
&\quad\left.+\frac{d}{2}\log\psi_g^{-1} - \frac{\psi_g^{-1}}{2}\mathbf{P}_{ig}^T\boldsymbol{\Lambda}_g^T\boldsymbol{\Lambda}_g\mathbf{P}_{ig} - \frac{\psi_g^{-1}}{2}\operatorname{tr}(\boldsymbol{\Lambda}_g^T\boldsymbol{\Lambda}_g\mathbf{Q}_g) - \frac{\psi_g^{-1}}{2}\operatorname{tr}\left(\mathbf{S}_{ig}\right)\right] + C,\\
&= \sum_{g=1}^{G}\left[-\frac{{\psi_g}^{-1}}{2}n_g\operatorname{tr}(\mathbf{W}_g) + {\psi_g}^{-1}n_g\operatorname{tr}(\mathbf{W}_g\boldsymbol{\Lambda}_g\boldsymbol{\beta}_g) + \frac{dn_g}{2}\log\psi_g^{-1} - \frac{{\psi_g}^{-1}n_g}{2}\operatorname{tr}(\boldsymbol{\Lambda}_g^T\boldsymbol{\Lambda}_g\mathbf{Q}_g)\right.\\
&\quad\left.-\frac{{\psi_g}^{-1}}{2}n_g\operatorname{tr}(\boldsymbol{\beta}^T\boldsymbol{\Lambda}_g^T\boldsymbol{\Lambda}_g\boldsymbol{\beta}_g\mathbf{W}_g) - \frac{\psi_g^{-1}}{2}\operatorname{tr}\left(\sum_{i=1}^{n} z_{ig}\mathbf{S}_{ig}\right)\right] + C,
\end{aligned}$$

where C is a constant with respect to ψ_g^{-1} and $n_g = \sum_{i=1}^n z_{ig}$. Taking derivative with respect to ψ_g^{-1} and setting it to 0, we get

$$d\hat{\psi}_g - \operatorname{tr}(\mathbf{W}_g) + 2\operatorname{tr}(\mathbf{W}_g\hat{\boldsymbol{\Lambda}}_g\hat{\boldsymbol{\beta}}_g) - \operatorname{tr}(\hat{\boldsymbol{\beta}}^T\hat{\boldsymbol{\Lambda}}_g^T\hat{\boldsymbol{\Lambda}}_g\hat{\boldsymbol{\beta}}_g\mathbf{W}_g) - \operatorname{tr}(\hat{\boldsymbol{\Lambda}}_g^T\hat{\boldsymbol{\Lambda}}_g\mathbf{Q}_g) - \operatorname{tr}\left(\frac{\sum_{i=1}^n z_{ig}\mathbf{S}_{ig}}{n_g}\right) = 0.$$

And,

$$\begin{aligned}
\hat{\psi}_g &= \frac{1}{d}\operatorname{tr}\left(\mathbf{W}_g - 2\mathbf{W}_g\hat{\boldsymbol{\Lambda}}_g\hat{\boldsymbol{\beta}}_g + \hat{\boldsymbol{\beta}}^T\hat{\boldsymbol{\Lambda}}_g^T\hat{\boldsymbol{\Lambda}}_g\hat{\boldsymbol{\beta}}_g\mathbf{W}_g + \hat{\boldsymbol{\Lambda}}_g^T\hat{\boldsymbol{\Lambda}}_g\mathbf{Q}_g + \frac{\sum_{i=1}^n z_{ig}\mathbf{S}_{ig}}{n_g}\right)\\
&= \frac{1}{d}\operatorname{tr}\left(\mathbf{W}_g - 2\mathbf{W}_g\hat{\boldsymbol{\Lambda}}_g\hat{\boldsymbol{\beta}}_g + \hat{\boldsymbol{\Lambda}}_g\hat{\boldsymbol{\Theta}}_g\hat{\boldsymbol{\Lambda}}_g^T + \frac{\sum_{i=1}^n z_{ig}\mathbf{S}_{ig}}{n_g}\right)\\
&= \frac{1}{d}\operatorname{tr}\left(\mathbf{W}_g - 2\mathbf{W}_g\hat{\boldsymbol{\Lambda}}_g\hat{\boldsymbol{\beta}}_g - \hat{\boldsymbol{\Lambda}}_g\hat{\boldsymbol{\Theta}}_g\hat{\boldsymbol{\Theta}}_g^{-1}\hat{\boldsymbol{\beta}}_g\mathbf{W}_g + \frac{\sum_{i=1}^n z_{ig}\mathbf{S}_{ig}}{n_g}\right)\\
&= \frac{1}{d}\operatorname{tr}\left(\mathbf{W}_g - \mathbf{W}_g\hat{\boldsymbol{\Lambda}}_g\hat{\boldsymbol{\beta}}_g + \frac{\sum_{i=1}^n z_{ig}\mathbf{S}_{ig}}{n_g}\right)\\
&= \frac{1}{d}\operatorname{tr}\left(\mathbf{W}_g - \hat{\boldsymbol{\Lambda}}_g\hat{\boldsymbol{\beta}}_g\mathbf{W}_g + \frac{\sum_{i=1}^n z_{ig}\mathbf{S}_{ig}}{n_g}\right).
\end{aligned}$$

Therefore,

$$\hat{\psi}_g = \frac{1}{d}\operatorname{tr}\left(\mathbf{W}_g - \hat{\boldsymbol{\Lambda}}_g\hat{\boldsymbol{\beta}}_g\mathbf{W}_g + \frac{\sum_{i=1}^n z_{ig}\mathbf{S}_{ig}}{n_g}\right). \qquad (3)$$

3.2 Equal Variance: $\boldsymbol{\Psi}_g = \boldsymbol{\Psi}$

Under the equal variance constraint (i.e. $\boldsymbol{\Psi}_g = \boldsymbol{\Psi}$),

$$\begin{aligned}
\sum_{g=1}^{G}\sum_{i=1}^{n} z_{ig}F(q_{ig},\mathbf{y}_i) = & \sum_{g=1}^{G}\sum_{i=1}^{n} z_{ig}\left[-\frac{1}{2}(\mathbf{m}_{ig}-\boldsymbol{\mu}_{ig})^T\boldsymbol{\Psi}^{-1}(\mathbf{m}_{ig}-\boldsymbol{\mu}_{ig}) + (\mathbf{m}_{ig}-\boldsymbol{\mu}_{g})^T\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}_g\mathbf{P}_{ig}\right.\\
& \left. +\frac{1}{2}\log|\boldsymbol{\Psi}^{-1}| - \frac{1}{2}\mathbf{P}_{ig}^T\boldsymbol{\Lambda}_g^T\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}_g\mathbf{P}_{ig} - \frac{1}{2}\operatorname{tr}(\boldsymbol{\Lambda}_g^T\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}_g\mathbf{Q}_g) - \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Psi}^{-1}\mathbf{S}_{ig}\right)\right] + C,\\
= & \sum_{g=1}^{G}\left[-\frac{n_g}{2}\operatorname{tr}(\boldsymbol{\Psi}^{-1}\mathbf{W}_g) + n_g\operatorname{tr}(\mathbf{W}_g\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}_g\boldsymbol{\beta}_g) - \frac{n_g}{2}\operatorname{tr}(\boldsymbol{\Lambda}_g^T\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}_g\mathbf{Q}_g)\right.\\
& \left. +\frac{n_g}{2}\log|\boldsymbol{\Psi}^{-1}| - \frac{{\psi_g}^{-1}}{2}n_g\operatorname{tr}(\boldsymbol{\beta}^T\boldsymbol{\Lambda}_g^T\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}_g\boldsymbol{\beta}_g\mathbf{W}_g) - \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Psi}^{-1}\sum_{i=1}^{n} z_{ig}\mathbf{S}_{ig}\right)\right] + C,
\end{aligned}$$

where C is a constant with respect to $\boldsymbol{\Psi}$ and $n_g = \sum_{i=1}^{n} z_{ig}$. Taking derivative with respect to $\boldsymbol{\Psi}^{-1}$ and setting it to 0, we get

$$n\hat{\boldsymbol{\Psi}} - \sum_{g=1}^{G} n_g\mathbf{W}_g + 2\sum_{g=1}^{G} n_g\hat{\boldsymbol{\Lambda}}_g\hat{\boldsymbol{\beta}}_g\mathbf{W}_g - \sum_{g=1}^{G} n_g\hat{\boldsymbol{\beta}}^T\hat{\boldsymbol{\Lambda}}_g^T\hat{\boldsymbol{\Lambda}}_g\hat{\boldsymbol{\beta}}_g\mathbf{W}_g - \sum_{g=1}^{G} n_g\hat{\boldsymbol{\Lambda}}_g^T\hat{\boldsymbol{\Lambda}}_g\mathbf{Q}_g - \operatorname{tr}\left(\frac{\sum_{i=1}^{n} z_{ig}\mathbf{S}_{ig}}{n_g}\right) = 0.$$

Simplifying this, we get

$$\begin{aligned}
\hat{\boldsymbol{\Psi}} = & \operatorname{diag}\left(\sum_{g=1}^{G}\frac{n_g}{n}\mathbf{W}_g - \sum_{g=1}^{G}\frac{n_g}{n}\hat{\boldsymbol{\Lambda}}_g\hat{\boldsymbol{\beta}}_g\mathbf{W}_g + \frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n} z_{ig}\mathbf{S}_{ig}\right).\\
= & \operatorname{diag}\left(\sum_{g=1}^{G}\frac{n_g}{n}\mathbf{W}_g - \sum_{g=1}^{G}\frac{n_g}{n}\hat{\boldsymbol{\Lambda}}_g\hat{\boldsymbol{\beta}}_g\mathbf{W}_g + \frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n} z_{ig}\mathbf{S}_{ig}\right)
\end{aligned}$$

3.3 Equal Variance and isotropic constraint $\boldsymbol{\Psi}_g = \psi\mathbf{I}_d$

$$\begin{aligned}
\sum_{g=1}^{G}\sum_{i=1}^{n} z_{ig}F(q_{ig},\mathbf{y}_i) = {} & \sum_{g=1}^{G}\sum_{i=1}^{n} z_{ig}\left[-\frac{\psi^{-1}}{2}(\mathbf{m}_{ig}-\boldsymbol{\mu}_{ig})^T(\mathbf{m}_{ig}-\boldsymbol{\mu}_{ig}) + \psi^{-1}(\mathbf{m}_{ig}-\boldsymbol{\mu}_g)^T\boldsymbol{\Lambda}_g\mathbf{P}_{ig}\right.\\
& \left.+\frac{d}{2}\log\psi^{-1} - \frac{\psi^{-1}}{2}\mathbf{P}_{ig}^T\boldsymbol{\Lambda}_g^T\boldsymbol{\Lambda}_g\mathbf{P}_{ig} - \frac{\psi^{-1}}{2}\operatorname{tr}(\boldsymbol{\Lambda}_g^T\boldsymbol{\Lambda}_g\mathbf{Q}_g) - \frac{\psi^{-1}}{2}\operatorname{tr}\left(\mathbf{S}_{ig}\right)\right] + C,\\
= {} & \sum_{g=1}^{G}\left[-\frac{\psi^{-1}}{2}n_g\operatorname{tr}(\mathbf{W}_g) + \psi^{-1}n_g\operatorname{tr}(\mathbf{W}_g\boldsymbol{\Lambda}_g\boldsymbol{\beta}_g) + \frac{dn_g}{2}\log\psi^{-1} - \frac{\psi^{-1}n_g}{2}\operatorname{tr}(\boldsymbol{\Lambda}_g^T\boldsymbol{\Lambda}_g\mathbf{Q}_g)\right.\\
& \left.-\frac{\psi^{-1}}{2}n_g\operatorname{tr}(\boldsymbol{\beta}^T\boldsymbol{\Lambda}_g^T\boldsymbol{\Lambda}_g\boldsymbol{\beta}_g\mathbf{W}_g) - \frac{\psi^{-1}}{2}\operatorname{tr}\left(\sum_{i=1}^{n} z_{ig}\mathbf{S}_{ig}\right)\right] + C,
\end{aligned}$$

where C is a constant with respect to ψ^{-1} and $n_g = \sum_{i=1}^{n} z_{ig}$. Taking derivative with respect to ψ_g^{-1} and setting it to 0, we get

$$\begin{aligned}
& dn\hat{\psi} - \operatorname{tr}(\sum_{g=1}^{G} n_g\mathbf{W}_g) + 2\operatorname{tr}(\sum_{g=1}^{G} n_g\mathbf{W}_g\hat{\boldsymbol{\Lambda}}_g\hat{\boldsymbol{\beta}}_g) - \operatorname{tr}(\sum_{g=1}^{G} n_g\hat{\boldsymbol{\beta}}^T\hat{\boldsymbol{\Lambda}}_g^T\hat{\boldsymbol{\Lambda}}_g\hat{\boldsymbol{\beta}}_g\mathbf{W}_g) - \operatorname{tr}(\sum_{g=1}^{G} n_g\hat{\boldsymbol{\Lambda}}_g^T\hat{\boldsymbol{\Lambda}}_g\mathbf{Q}_g)\\
& \quad - \operatorname{tr}\left(\sum_{g=1}^{G}\sum_{i=1}^{n} z_{ig}\mathbf{S}_{ig}\right) = 0.
\end{aligned}$$

And, simplifying this, we get

$$\hat{\psi} = \frac{1}{d}\operatorname{tr}\left(\sum_{g=1}^{G}\frac{n_g}{n}\mathbf{W}_g - \sum_{g=1}^{G}\frac{n_g}{n}\hat{\boldsymbol{\Lambda}}_g\hat{\boldsymbol{\beta}}_g\mathbf{W}_g + \frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n} z_{ig}\mathbf{S}_{ig}\right) \tag{4}$$

3.4 Equal Loading Matrices: $\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$

Under the equal loading matrices constraint (i.e. $\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$),

$$\begin{aligned}\sum_{g=1}^{G}\sum_{i=1}^{n} z_{ig}F(q_{ig},\mathbf{y}_i) &= \sum_{g=1}^{G}\sum_{i=1}^{n} z_{ig}\left((\mathbf{m}_{ig}-\boldsymbol{\mu}_g)^T\boldsymbol{\Psi}_g^{-1}\boldsymbol{\Lambda}\mathbf{P}_{ig} - \frac{1}{2}\mathbf{P}_{ig}^T\boldsymbol{\Lambda}^T\boldsymbol{\Psi}_g^{-1}\boldsymbol{\Lambda}\mathbf{P}_{ig} - \frac{1}{2}\operatorname{tr}(\boldsymbol{\Lambda}^T\boldsymbol{\Psi}_g^{-1}\boldsymbol{\Lambda}\mathbf{Q}_g)\right)\\ &= \sum_{g=1}^{G}\left(n_g\operatorname{tr}\left(\boldsymbol{\beta}_g\mathbf{W}_g\boldsymbol{\Psi}_g^{-1}\boldsymbol{\Lambda}\right) - \frac{n_g}{2}\operatorname{tr}\left(\boldsymbol{\beta}_g^T\boldsymbol{\Lambda}^T\boldsymbol{\Psi}_g^{-1}\boldsymbol{\Lambda}\boldsymbol{\beta}_g\mathbf{W}_g\right) - \frac{n_g}{2}tr(\boldsymbol{\Lambda}^T\boldsymbol{\Psi}_g^{-1}\boldsymbol{\Lambda}\mathbf{Q}_g)\right).\end{aligned}$$

Taking derivative with respect to $\boldsymbol{\Lambda}$ and setting it equal to 0, we get

$$\begin{aligned}0 &= \sum_{g=1}^{G} n_g\hat{\boldsymbol{\Psi}}_g^{-1}\mathbf{W}_g\hat{\boldsymbol{\beta}}_g^T - \sum_{g=1}^{G} n_g\hat{\boldsymbol{\Psi}}_g^{-1}\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\beta}}_g\mathbf{W}_g\hat{\boldsymbol{\beta}}_g^T - \sum_{g=1}^{G} n_g\hat{\boldsymbol{\Psi}}_g^{-1}\hat{\boldsymbol{\Lambda}}\mathbf{Q}_g\\ &= \sum_{g=1}^{G} n_g\hat{\boldsymbol{\Psi}}_g^{-1}\mathbf{W}_g\hat{\boldsymbol{\beta}}_g^T - \sum_{g=1}^{G} n_g\hat{\boldsymbol{\Psi}}_g^{-1}\hat{\boldsymbol{\Lambda}}\left(\hat{\boldsymbol{\beta}}_g\mathbf{W}_g\hat{\boldsymbol{\beta}}_g^T - \mathbf{I}_p - \hat{\boldsymbol{\beta}}_g\hat{\boldsymbol{\Lambda}}\right)\\ &= \sum_{g=1}^{G} n_g\hat{\boldsymbol{\Psi}}_g^{-1}\mathbf{W}_g\hat{\boldsymbol{\beta}}_g^T - \sum_{g=1}^{G} n_g\hat{\boldsymbol{\Psi}}_g^{-1}\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Theta}}_g.\end{aligned}$$

In this case, the loading matrix cannot be solved directly and must be solved in a row-by-row manner as suggested by Mcnicholas and Murphy (2008). Hence,

$$\hat{\lambda}_i = \mathbf{r}_i\left(\sum_{g=1}^{G}\frac{n_g}{\hat{\Psi}_{g,(i)}}\hat{\boldsymbol{\Theta}}_g\right)^{-1}, \tag{5}$$

where λ_i is the i^{th} row of the matrix $\boldsymbol{\Lambda}$, $\Psi_{g,(i)}$ is the i^{th} diagonal element of $\boldsymbol{\Psi}_g$ and $\mathbf{r}_i$ is the i^{th} row of the matrix $\sum_{g=1}^{G} n_g\boldsymbol{\Psi}_g^{-1}\mathbf{W}_g\hat{\boldsymbol{\beta}}_g^T$.

Table 2: Parameter Updates for the family of models based on constraints on $\boldsymbol{\Lambda}_g$ and $\boldsymbol{\Psi}_g$.

Model	Parameters	Estimates
UUU	$\boldsymbol{\beta}_g$	$\hat{\boldsymbol{\beta}}_g = \hat{\boldsymbol{\Lambda}}_g^T(\hat{\boldsymbol{\Lambda}}_g\hat{\boldsymbol{\Lambda}}_g^T + \boldsymbol{\Psi}_g)^{-1}$
	$\boldsymbol{\Theta}_g$	$\hat{\boldsymbol{\Theta}}_g = \mathbf{I}_p - \hat{\boldsymbol{\beta}}_g\hat{\boldsymbol{\Lambda}}_g + \hat{\boldsymbol{\beta}}_g\mathbf{W}_g\hat{\boldsymbol{\beta}}_g^T$
	$\boldsymbol{\Lambda}_g$	$\hat{\boldsymbol{\Lambda}}_g = \mathbf{W}_g\hat{\boldsymbol{\beta}}^T\widehat{\boldsymbol{\Theta}}_g^{-1}$
	$\boldsymbol{\Psi}_g$	$\hat{\boldsymbol{\Psi}}_g = \text{ diag}\left(\mathbf{W}_g - \hat{\boldsymbol{\Lambda}}_g\hat{\boldsymbol{\beta}}_g\mathbf{W}_g + \frac{\sum_{i=1}^n \widehat{Z}_{ig}\mathbf{S}_{ig}}{\sum_{i=1}^n \widehat{Z}_{ig}}\right)$
UUC	$\boldsymbol{\beta}_g$	$\hat{\boldsymbol{\beta}}_g = \hat{\boldsymbol{\Lambda}}_g^T(\hat{\boldsymbol{\Lambda}}_g\hat{\boldsymbol{\Lambda}}_g^T + \psi_g\mathbf{I}_d)^{-1}$
	$\boldsymbol{\Theta}_g$	$\hat{\boldsymbol{\Theta}}_g = \mathbf{I}_p - \hat{\boldsymbol{\beta}}_g\hat{\boldsymbol{\Lambda}}_g + \hat{\boldsymbol{\beta}}_g\mathbf{W}_g\hat{\boldsymbol{\beta}}_g^T$
	$\boldsymbol{\Lambda}_g$	$\hat{\boldsymbol{\Lambda}}_g = \mathbf{W}_g\hat{\boldsymbol{\beta}}^T\widehat{\boldsymbol{\Theta}}_g^{-1}$
	$\boldsymbol{\Psi}_g = \psi_g\mathbf{I}_d$	$\hat{\psi}_g = \frac{1}{d}\,\text{tr}\left(\mathbf{W}_g - \hat{\boldsymbol{\Lambda}}_g\hat{\boldsymbol{\beta}}_g\mathbf{W}_g + \frac{\sum_{i=1}^n z_{ig}\mathbf{S}_{ig}}{n_g}\right)$
UCU	$\boldsymbol{\beta}_g$	$\hat{\boldsymbol{\beta}}_g = \hat{\boldsymbol{\Lambda}}_g^T(\hat{\boldsymbol{\Lambda}}_g\hat{\boldsymbol{\Lambda}}_g^T + \boldsymbol{\Psi})^{-1}$
	$\boldsymbol{\Theta}_g$	$\hat{\boldsymbol{\Theta}}_g = \mathbf{I}_p - \hat{\boldsymbol{\beta}}_g\hat{\boldsymbol{\Lambda}}_g + \hat{\boldsymbol{\beta}}_g\mathbf{W}_g\hat{\boldsymbol{\beta}}_g^T$
	$\boldsymbol{\Lambda}_g$	$\hat{\boldsymbol{\Lambda}}_g = \mathbf{W}_g\hat{\boldsymbol{\beta}}^T\widehat{\boldsymbol{\Theta}}_g^{-1}$
	$\boldsymbol{\Psi}_g = \boldsymbol{\Psi}$	$\hat{\boldsymbol{\Psi}} = \text{ diag}\left(\sum_{g=1}^G \frac{n_g}{n}\mathbf{W}_g - \sum_{g=1}^G \frac{n_g}{n}\hat{\boldsymbol{\Lambda}}_g\hat{\boldsymbol{\beta}}_g\mathbf{W}_g + \frac{1}{n}\sum_{g=1}^G\sum_{i=1}^n z_{ig}\mathbf{S}_{ig}\right)$
UCC	$\boldsymbol{\beta}_g$	$\hat{\boldsymbol{\beta}}_g = \hat{\boldsymbol{\Lambda}}_g^T(\hat{\boldsymbol{\Lambda}}_g\hat{\boldsymbol{\Lambda}}_g^T + \psi\mathbf{I})^{-1}$
	$\boldsymbol{\Theta}_g$	$\hat{\boldsymbol{\Theta}}_g = \mathbf{I}_p - \hat{\boldsymbol{\beta}}_g\hat{\boldsymbol{\Lambda}}_g + \hat{\boldsymbol{\beta}}_g\mathbf{W}_g\hat{\boldsymbol{\beta}}_g^T$
	$\boldsymbol{\Lambda}_g$	$\hat{\boldsymbol{\Lambda}}_g = \mathbf{W}_g\hat{\boldsymbol{\beta}}^T\widehat{\boldsymbol{\Theta}}_g^{-1}$
	$\boldsymbol{\Psi}_g = \psi\mathbf{I}$	$\hat{\psi} = \frac{1}{d}\,\text{tr}\left(\sum_{g=1}^G \frac{n_g}{n}\mathbf{W}_g - \sum_{g=1}^G \frac{n_g}{n}\hat{\boldsymbol{\Lambda}}_g\hat{\boldsymbol{\beta}}_g\mathbf{W}_g + \frac{1}{n}\sum_{g=1}^G\sum_{i=1}^n z_{ig}\mathbf{S}_{ig}\right)$

Continued on next page...

Model	Parameters	Estimates
CUU	$\boldsymbol{\beta}_g$	$\hat{\boldsymbol{\beta}}_g = \hat{\boldsymbol{\Lambda}}^T(\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}^T + \boldsymbol{\Psi}_g)^{-1}$
	$\boldsymbol{\Theta}_g$	$\hat{\boldsymbol{\Theta}}_g = \mathbf{I}_p - \hat{\boldsymbol{\beta}}_g\hat{\boldsymbol{\Lambda}} + \hat{\boldsymbol{\beta}}_g\mathbf{W}_g\hat{\boldsymbol{\beta}}_g^T$
	$\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$	Updated using Equation 5.
	$\boldsymbol{\Psi}_g$	$\hat{\boldsymbol{\Psi}}_g = \operatorname{diag}\left(\mathbf{W}_g - \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\beta}}_g\mathbf{W}_g + \frac{\sum_{i=1}^n \widehat{Z}_{ig}\mathbf{S}_{ig}}{\sum_{i=1}^n \widehat{Z}_{ig}}\right)$
CUC	$\boldsymbol{\beta}_g$	$\hat{\boldsymbol{\beta}}_g = \hat{\boldsymbol{\Lambda}}^T(\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}^T + \psi_g\mathbf{I}_d)^{-1}$
	$\boldsymbol{\Theta}_g$	$\hat{\boldsymbol{\Theta}}_g = \mathbf{I}_p - \hat{\boldsymbol{\beta}}_g\hat{\boldsymbol{\Lambda}} + \hat{\boldsymbol{\beta}}_g\mathbf{W}_g\hat{\boldsymbol{\beta}}_g^T$
	$\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$	Updated using Equation 5.
	$\boldsymbol{\Psi}_g = \psi_g\mathbf{I}_d$	$\hat{\psi}_g = \frac{1}{d}\operatorname{tr}\left(\mathbf{W}_g - \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\beta}}_g\mathbf{W}_g + \frac{\sum_{i=1}^n z_{ig}\mathbf{S}_{ig}}{n_g}\right)$
CCU	$\boldsymbol{\beta}_g = \boldsymbol{\beta}$	$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\Lambda}}^T(\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}^T + \boldsymbol{\Psi})^{-1}$
	$\boldsymbol{\Theta}_g$	$\hat{\boldsymbol{\Theta}}_g = \mathbf{I}_p - \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\Lambda}} + \hat{\boldsymbol{\beta}}\mathbf{W}_g\hat{\boldsymbol{\beta}}^T$
	$\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$	Updated using Equation 5.
	$\boldsymbol{\Psi}_g = \boldsymbol{\Psi}$	$\hat{\boldsymbol{\Psi}} = \operatorname{diag}\left(\sum_{g=1}^G \frac{n_g}{n}\mathbf{W}_g - \sum_{g=1}^G \frac{n_g}{n}\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\beta}}\mathbf{W}_g + \frac{1}{n}\sum_{g=1}^G\sum_{i=1}^n z_{ig}\mathbf{S}_{ig}\right)$
CCC	$\boldsymbol{\beta}_g = \boldsymbol{\beta}$	$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\Lambda}}^T(\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}^T + \psi\mathbf{I}_d)^{-1}$
	$\boldsymbol{\Theta}_g$	$\hat{\boldsymbol{\Theta}}_g = \mathbf{I}_p - \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\Lambda}} + \hat{\boldsymbol{\beta}}\mathbf{W}_g\hat{\boldsymbol{\beta}}^T$
	$\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$	Updated using Equation 5.
	$\boldsymbol{\Psi}_g = \psi\mathbf{I}$	$\hat{\psi} = \frac{1}{d}\operatorname{tr}\left(\sum_{g=1}^G \frac{n_g}{n}\mathbf{W}_g - \sum_{g=1}^G \frac{n_g}{n}\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\beta}}\mathbf{W}_g + \frac{1}{n}\sum_{g=1}^G\sum_{i=1}^n z_{ig}\mathbf{S}_{ig}\right)$

4 Test Data

I did a small test set containing 200 datasets of $d = 5$ with $p = 2$ and $N = 1000$ simulated from a $G = 2$ model for UUU model only. Using 6 cores on my laptop, the combined run time for all 200 dataset was 225.510 seconds (∼3.7 minutes). This is a substantial improvement in the speed. Do note that I am only running $p = 2$ and $G = 2$. But again, I am running a very crude version of the code and there are ways to parallelize it and make it even faster. We can always look into that later.

```
> total_run<-200
> ptm<-proc.time()
> output_all<-list()
> output_all <- foreach(run = 1:total_run, .errorhandling = "pass") %dopar% {
+   parallel_FA(run)
+ }
> proc.time()-ptm
    user   system  elapsed
1253.871   18.508  223.510
```

Initialization was done using k-means. The mean ARI for 200 runs is 0.9977813 (sd of 0.003179812). Summary of the parameter estimates from the 200 runs:

Table 3: Summary of the average and standard errors (SE) of the estimated parameters from the 200 datasets.

	True Parameters	Average of Estimated Parameters (Standard errors)
$\boldsymbol{\mu}_1$	(6, 3, 3, 6, 3)	(6.00, 3.00, 3.00, 6.00, 3.00) SE = (0.05, 0.07, 0.06, 0.06, 0.05)
$\boldsymbol{\mu}_2$	(5, 3, 5, 3, 5)	(5.00, 3.00, 5.00, 3.00, 5.00) SE = (0.04, 0.04, 0.02, 0.02, 0.04)

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 0.86 & 0.47 & 0.35 & 0.28 & 0.23 \\ 0.47 & 1.44 & 0.44 & 0.46 & 0.26 \\ 0.35 & 0.44 & 0.98 & 0.23 & 0.32 \\ 0.28 & 0.46 & 0.23 & 1.05 & 0.12 \\ 0.23 & 0.26 & 0.32 & 0.12 & 0.67 \end{bmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 1.01 & 0.69 & 0.41 & 0.49 & 0.94 \\ 0.69 & 0.99 & 0.32 & 0.43 & 0.81 \\ 0.41 & 0.32 & 0.38 & 0.24 & 0.46 \\ 0.49 & 0.43 & 0.24 & 0.46 & 0.55 \\ 0.94 & 0.81 & 0.46 & 0.55 & 1.23 \end{bmatrix},$$

$$\text{mean } \hat{\boldsymbol{\Sigma}}_1 = \begin{bmatrix} 0.87 & 0.45 & 0.36 & 0.29 & 0.24 \\ 0.45 & 1.43 & 0.41 & 0.43 & 0.26 \\ 0.36 & 0.41 & 0.99 & 0.23 & 0.25 \\ 0.29 & 0.43 & 0.23 & 1.05 & 0.12 \\ 0.24 & 0.26 & 0.25 & 0.12 & 0.67 \end{bmatrix}, \quad \text{mean } \hat{\boldsymbol{\Sigma}}_2 = \begin{bmatrix} 1.01 & 0.69 & 0.41 & 0.48 & 0.93 \\ 0.69 & 1.00 & 0.34 & 0.41 & 0.80 \\ 0.41 & 0.34 & 0.38 & 0.23 & 0.46 \\ 0.48 & 0.41 & 0.23 & 0.46 & 0.54 \\ 0.93 & 0.80 & 0.46 & 0.54 & 1.23 \end{bmatrix},$$

$$\text{SE } \hat{\boldsymbol{\Sigma}}_1 = \begin{bmatrix} 0.07 & 0.06 & 0.06 & 0.06 & 0.05 \\ 0.06 & 0.11 & 0.07 & 0.07 & 0.05 \\ 0.06 & 0.07 & 0.09 & 0.06 & 0.05 \\ 0.06 & 0.07 & 0.06 & 0.08 & 0.05 \\ 0.05 & 0.05 & 0.05 & 0.05 & 0.06 \end{bmatrix}, \quad \text{SE } \hat{\boldsymbol{\Sigma}}_2 = \begin{bmatrix} 0.06 & 0.05 & 0.03 & 0.03 & 0.06 \\ 0.05 & 0.06 & 0.03 & 0.03 & 0.06 \\ 0.03 & 0.03 & 0.02 & 0.02 & 0.03 \\ 0.03 & 0.03 & 0.02 & 0.03 & 0.04 \\ 0.06 & 0.06 & 0.03 & 0.04 & 0.07 \end{bmatrix}.$$

References

Arridge, S. R., K. Ito, B. Jin, and C. Zhang (2018). Variational gaussian approximation for poisson data. *Inverse Problems 34*(2), 025005.

Gollini, I. and T. B. Murphy (2014). Mixture of latent trait analyzers for model-based clustering of categorical data. *Statistics and Computing 24*(4), 569–588.

Mcnicholas, P. D. and T. B. Murphy (2008). Parsimonious gaussian mixture models. *Statistics and Computing 18*(3), 285–296.

Subedi, S. and R. Browne (2020). A parsimonious family of multivariate poisson-lognormal distributions for clustering multivariate count data. *arXiv preprint arXiv:2004.06857*.

Tang, Y., R. P. Browne, and P. D. McNicholas (2015). Model based clustering of high-dimensional binary data. *Computational Statistics & Data Analysis 87*, 84–101.

Wainwright, M. J., M. I. Jordan, et al. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning 1*(1–2), 1–305.