Suppose the d dimensional vector of complete records. Each observation $\mathbf{W}_i = (W_{i1}, \ldots, W_{id})$ can be decomposed as:

$$\begin{aligned}\mathbf{W}_i &= \mathbf{O}_i^\top \mathbf{O}_i \mathbf{W}_i + \mathbf{M}_i^\top \mathbf{M}_i \mathbf{W}_i \\ &= \mathbf{O}_i^\top \mathbf{W}_i^{(o)} + \mathbf{M}_i^\top \mathbf{W}_i^{(m)},\end{aligned}$$

where $\mathbf{W}_i^{(o)} = \mathbf{O}_i \mathbf{W}_i$ is a d_i^o-dimensional vector comprising only observed entries in the i^{th} observation, $\mathbf{W}_i^{(m)} = \mathbf{M}_i \mathbf{W}_i$ is a $d - d_i^o$-dimensional vector comprising of missing entries in the i^{th} observation, $\mathbf{O}_i$ and $\mathbf{M}_i$ are observed and missing entries extraction matrices of dimensions $d_i^o \times d$ and $(d - d_i^o) \times d$, respectively.

To obtain the marginal densities of the observed data, we first define the hierarchical structure comprising only the observed entries.

$$W_{ij}^{(o)} \mid Y_{ij}^{(o)} \sim \text{Poisson}(e^{\mathbf{o}_j^\top Y_i}) \quad \text{and} \quad \mathbf{Y}_i^{(o)} \sim \mathcal{N}_{d_i^{(o)}}(\mathbf{O}_i \boldsymbol{\mu}, \mathbf{O}_i \boldsymbol{\Sigma} \mathbf{O}_i^\top),$$

where $\mathbf{o}_j$ is the j^{th} row of the matrix $\mathbf{O}_i$.

Thus, for the mixtures of MPLN distribution, the marginal density of the observed entries can be written as

$$\begin{aligned}f(\mathbf{w}_i \mid \boldsymbol{\vartheta}) &= \sum_{g=1}^{G} \pi_g f_{\mathbf{W}_i}(\mathbf{w}_i^{(o)} \mid \mathbf{O}_i \boldsymbol{\mu}_g, \mathbf{O}_i \boldsymbol{\Sigma}_g \mathbf{O}_i^\top) \\ &= \sum_{g=1}^{G} \pi_g f_{\mathbf{W}_i}(\mathbf{O}_i \mathbf{w}_i \mid \mathbf{O}_i \boldsymbol{\mu}_g, \mathbf{O}_i \boldsymbol{\Sigma}_g \mathbf{O}_i^\top)\end{aligned}$$

In model-based clustering, an additional component membership indicator variable $\mathbf{Z}$ is introduced which is assumed to be unknown and $Z_{ig} = 1$ 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 $\mathbf{z}$ and the complete-data likelihood is defined as

$$L_c(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \mathbf{w}, \mathbf{z}) = \prod_{g=1}^{G} \prod_{i=1}^{n} \left[\pi_g f(\mathbf{w}_i^{(o)} \mid \mathbf{O}_i \boldsymbol{\mu}_g, \mathbf{O}_i \boldsymbol{\Sigma}_g \mathbf{O}_i^\top)\right]^{z_{ig}}.$$

and the complete-data log-likelihood can be written as

$$l_c(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \mathbf{w}, \mathbf{z}) = \sum_{g=1}^{G} \sum_{i=1}^{n} z_{ig} \log \pi_g + \sum_{g=1}^{G} \sum_{i=1}^{n} z_{ig} \log f(\mathbf{w}_i^{(o)} \mid \mathbf{O}_i \boldsymbol{\mu}_g, \mathbf{O}_i \boldsymbol{\Sigma}_g \mathbf{O}_i^\top).$$

The marginal probability density of observed entries in the i^{th} observation $\mathbf{W}_i^{(o)}$ can be written as:

$$f_{\mathbf{w}}(\mathbf{w}_i^{(o)}) = f_{\mathbf{w}}(\mathbf{O}_i\mathbf{w}_i) = \int_{\mathbb{R}^{d_i^o}} \left[\prod_{j=1}^{d_i^o} p(w_j^{(o)} \mid y_j^{(o)})\right] \phi_{d_i^o}(\mathbf{Y}_i^{(o)} \mid \mathbf{O}_i\boldsymbol{\mu}_g, \mathbf{O}_i\boldsymbol{\Sigma}_g\mathbf{O}_i^\top)\; d\mathbf{Y}_i^{(o)},$$

where $w_j^{(o)}$ and $y_j^{(o)}$ are the j^{th} element of $\mathbf{w}^{(o)}$ and $\mathbf{y}^{(o)}$ respectively, $p(\cdot)$ is the probability mass function of the Poisson distribution with mean $\lambda_j = e^{y_j^{(o)}}$ and $\phi_{d_i^o}(\cdot)$ is the probability density function of d_i^o-dimensional Gaussian distribution with mean $\mathbf{O}_i\boldsymbol{\mu}_g$ and covariance $\mathbf{O}_i\boldsymbol{\Sigma}_g\mathbf{O}_i^\top$. (Subedi and Browne, 2020) proposed variational approximations for approximating the marginal of $\mathbf{W}$ and developed an EM-type framework for parameter estimation for the mixtures of MPLN distributions. Here, we will develop a similar framework for parameter estimation for the mixtures of MPLN with partial records.

Suppose, we have an approximating density $q(\mathbf{y}_i^{(o)}) = q(\mathbf{O}_i\mathbf{y}_i)$, the log of the marginal density can be written as

$$\log f(\mathbf{w}_i^{(o)}) = \log f(\mathbf{O}_i\mathbf{w}_i) = F(q, \mathbf{w}_i^{(o)}) + D_{KL}(q\|f),$$

where $D_{KL}(q\|f) = \int_{\mathbb{R}^d} q(\mathbf{y}_i^{(o)}) \log \frac{q(\mathbf{y}_i^{(o)})}{f(\mathbf{y}_i^{(o)}|\mathbf{w}_i^{(o)})} d\mathbf{y}$ is the Kullback-Leibler (KL) divergence between $f(\mathbf{y}_i^{(o)} \mid \mathbf{w}_i^{(o)})$ and approximating distribution $q(\mathbf{y}_i^{(o)})$, and

$$F(q, \mathbf{w}_i^{(o)}) = \int_{\mathbb{R}^{d_i^o}} \left[\log f(\mathbf{w}_i^{(o)}, \mathbf{y}_i^{(o)}) - \log q(\mathbf{y}_i^{(o)})\right] q(\mathbf{y}_i^{(o)}) d\mathbf{y}_i^{(o)}$$

is called the evidence lower bound (ELBO). Replacing the $f(\mathbf{w}_i^{(o)})$ by the sum of ELBO and KL divergence, the complete data log-likelihood of the mixtures of MPLN distributions can be written as:

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

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

Assuming that given $Z_{ig} = 1$, we assume $q(\mathbf{y}_i^{(o)}) = \mathcal{N}_{d_i^o}(\mathbf{m}_{ig}, \mathbf{S}_{ig})$ and the ELBO for each observation $\mathbf{w}_i$ becomes

$$
\begin{aligned}
F(q_{ig}, \mathbf{w}_i^{(o)}) &= \frac{1}{2}\log|\mathbf{S}_{ig}| - \frac{1}{2}(\mathbf{m}_{ig} - \mathbf{O}_i\boldsymbol{\mu}_g)^\top(\mathbf{O}_i\boldsymbol{\Sigma}_g\mathbf{O}_i^\top)^{-1}(\mathbf{m}_{ig} - \mathbf{O}_i\boldsymbol{\mu}_g) - \frac{1}{2}\operatorname{tr}((\mathbf{O}_i\boldsymbol{\Sigma}_g\mathbf{O}_i^\top)^{-1}\mathbf{S}_{ig}) \\
&\quad + \frac{1}{2}\log|(\mathbf{O}_i\boldsymbol{\Sigma}_g\mathbf{O}_i^\top)^{-1}| + \frac{d_i^o}{2} + \mathbf{m}_{ig}^\top\mathbf{O}_i\mathbf{y}_i - \sum_{j=1}^{d_i^o}\left(e^{(m_{igj}+\frac{1}{2}S_{ig,jj})} + \log(y_{ij}^{(o)}!)\right),
\end{aligned}
$$

where m_{igj} is the j^{th} element of the $\mathbf{m}_{ig}$ and $S_{ig,jj}$ is the j^{th} diagonal element of the matrix $\mathbf{S}_{ig}$. The variational parameters that maximize the ELBO will minimize the KL divergence between the true posterior and the approximating density. Parameter estimation can be done in an iterative EM-type approach such that the following steps are iterated.

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 using only the observed data. Given $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$,

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

   Note that this involves the marginal distribution of $\mathbf{W}$ which is difficult to compute. Hence, following Subedi and Browne (2020), we use an approximation of $\mathbb{E}(Z_{ig})$ where we replace the marginal density of the exponent of ELBO such that

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

2. Given $\hat{Z}_{ig}$, variational parameters $\mathbf{m}_{ig}$ and $\mathbf{S}_{ig}$ is updated conditional on $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$ as following:

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

$$
\mathbf{S}_{ig}^{(t+1)} = \left\{(\mathbf{O}_i\boldsymbol{\Sigma}_g\mathbf{O}_i^\top)^{-1} + \mathbf{I} \odot \mathbf{exp}\left[\mathbf{m}_{ig}^{(t)} + \frac{1}{2}\operatorname{diag}\left(\mathbf{S}_{ig}^{(t)}\right)\right]\mathbf{1}_{d_i^o}^\top\right\}^{-1}
$$

   where the vector function $\mathbf{exp}\,[\mathbf{a}] = (e^{a_1}, \ldots, e^{a_{d_i^o}})^\top$ is a vector of exponential each element of the d_i^o-dimensional vector $\mathbf{a}$, $\operatorname{diag}(\mathbf{S}) = (\mathbf{S}_{11}\ldots, \mathbf{S}_{d_i^o d_i^o})$ puts the diagonal elements of the $d_i^o \times d_i^o$ matrix $\mathbf{S}$ into a d_i^o-dimensional vector, $\odot$ the Hadmard product and $\mathbf{1}_{d_i^o}$ is a d_i^o-dimensional vector of ones.

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

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

3. Given $\hat{z}_{ig}$ and the variational parameters $\mathbf{m}_{ig}$ and $\mathbf{S}_{ig}$, the updates for the parameters $\boldsymbol{\pi}$, $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$ are obtained by maximizing the variational lower bound of the complete data log-likelihood. The updates for π_g and $\boldsymbol{\mu}_g$ have closed form solutions:

$$\hat{\pi}_g = \frac{\sum_{i=1}^n \widehat{Z}_{ig}}{n}, \quad \text{and} \quad \hat{\boldsymbol{\mu}}_g = \left(\sum_{i=1}^n \widehat{Z}_{ig}\mathbf{O}_i^\top(\mathbf{O}_i\boldsymbol{\Sigma}_g\mathbf{O}_i^T)^{-1}\mathbf{O}_i\right)^{-1}\left(\sum_{i=1}^n \widehat{Z}_{ig}\mathbf{O}_i^\top(\mathbf{O}_i\boldsymbol{\Sigma}_g\mathbf{O}_i^\top)^{-1}\mathbf{m}_{ig}\right).$$

The update for $\boldsymbol{\Sigma}_g$ however does not have a closed form solution. Thus, we will utilize gradient descent to update $\boldsymbol{\Sigma}_g$. Thus, we write the approximation of the complete data log-likelihood with ELBO as:

$$\begin{aligned}
l_c =& -\sum_{g=1}^G\sum_{i=1}^n \frac{z_{ig}}{2}(\mathbf{m}_{ig} - \mathbf{O}_i\boldsymbol{\mu}_g)^\top(\mathbf{O}_i\boldsymbol{\Sigma}_g\mathbf{O}_i^\top)^{-1}(\mathbf{m}_{ig} - \mathbf{O}_i\boldsymbol{\mu}_g)\\
&- \sum_{g=1}^G\sum_{i=1}^n \frac{z_{ig}}{2}\operatorname{tr}((\mathbf{O}_i\boldsymbol{\Sigma}_g\mathbf{O}_i^\top)^{-1}\mathbf{S}_{ig}) + \sum_{g=1}^G\sum_{i=1}^n \frac{z_{ig}}{2}\log|(\mathbf{O}_i\boldsymbol{\Sigma}_g\mathbf{O}_i^\top)^{-1}| + C\\
=& -\sum_{g=1}^G\sum_{i=1}^n \frac{z_{ig}}{2}\operatorname{tr}\big[(\mathbf{m}_{ig} - \mathbf{O}_i\boldsymbol{\mu}_g)(\mathbf{m}_{ig} - \mathbf{O}_i\boldsymbol{\mu}_g)^\top(\mathbf{O}_i\boldsymbol{\Sigma}_g\mathbf{O}_i^\top)^{-1}\big]\\
&- \sum_{g=1}^G\sum_{i=1}^n \frac{z_{ig}}{2}\operatorname{tr}\big[\mathbf{S}_{ig}(\mathbf{O}_i\boldsymbol{\Sigma}_g\mathbf{O}_i^\top)^{-1}\big] - \sum_{g=1}^G\sum_{i=1}^n \frac{z_{ig}}{2}\log|(\mathbf{O}_i\boldsymbol{\Sigma}_g\mathbf{O}_i^\top)| + C,
\end{aligned}$$

where C is a constant with respect to $\boldsymbol{\Sigma}_g$. Setting

$$\boldsymbol{\Omega}_{ig} = (\mathbf{m}_{ig} - \mathbf{O}_i\boldsymbol{\mu}_g)(\mathbf{m}_{ig} - \mathbf{O}_i\boldsymbol{\mu}_g)^\top + \mathbf{S}_{ig},$$

the approximation of the complete data log-likelihood can be written as

$$l_c = -\sum_{g=1}^G\sum_{i=1}^n \frac{z_{ig}}{2}\operatorname{tr}\left[\boldsymbol{\Omega}_{ig}(\mathbf{O}_i\boldsymbol{\Sigma}_g\mathbf{O}_i^\top)^{-1}\right] - \sum_{g=1}^G\sum_{i=1}^n \frac{z_{ig}}{2}\log|(\mathbf{O}_i\boldsymbol{\Sigma}_g\mathbf{O}_i^\top)| + C.$$

Thus,

$$\nabla_{\boldsymbol{\Sigma}_g} l_c = \sum_{i=1}^n \frac{z_{ig}}{2}\mathbf{O}_i^\top(\mathbf{O}_i\boldsymbol{\Sigma}_g\mathbf{O}_i^\top)^{-1}\boldsymbol{\Omega}_{ig}(\mathbf{O}_i\boldsymbol{\Sigma}_g\mathbf{O}_i^\top)^{-1}\mathbf{O}_i - \sum_{i=1}^n \frac{z_{ig}}{2}\mathbf{O}_i^\top(\mathbf{O}_i\boldsymbol{\Sigma}_g\mathbf{O}_i^\top)^{-1}\mathbf{O}_i$$

and $\boldsymbol{\Sigma}_g$ can be updated using gradient ascent algorithm as:

$$\boldsymbol{\Sigma}_g^{(t+1)} = \boldsymbol{\Sigma}_g^{(t)} + \gamma\nabla_{\boldsymbol{\Sigma}_g} l_c,$$

where γ is the learning rate and is set to 0.001.

References

Subedi, S. and Browne, R. P. (2020), 'A family of parsimonious mixtures of multivariate poisson-lognormal distributions for clustering multivariate count data', *Stat* **9**(1), e310. e310 sta4.310.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/sta4.310*