Finite Mixtures of Multivariate Poisson-Log Normal Factor Analyzers for Clustering Count Data

Anjali Silva* Steven J. Rothstein†
Paul D. McNicholas‡ Sanjeena Subedi§

Abstract

A mixture of multivariate Poisson-log normal factor analyzers is introduced by imposing constraints on the covariance matrix, which resulted flexible models for clustering purposes. In particular, a class of four parsimonious mixture models based on the mixtures of factor analyzers model are introduced. A Markov chain Monte Carlo alternating expectation-conditional maximization algorithm (MCMC-AECM) is used for model parameter estimation, and information criteria are used for model selection. The proposed models are explored in the context of clustering discrete data arising from RNA sequencing studies. Using real and simulated data, the models are shown to give favorable clustering performance.

1 Introduction

Model-based clustering is a technique that utilizes finite mixture models to cluster data (Wolfe, 1965; McLachlan and Basford, 1988; McLachlan and Peel, 2000; McNicholas, 2016). The general distribution function for mixture models can be given as $f(\mathbf{y}|\pi_1, \ldots, \pi_G, \boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_G) = \sum_{g=1}^{G} \pi_g f_g(\mathbf{y}|\boldsymbol{\vartheta}_g)$, where G is the total number of clusters, $f_g(\cdot)$ is the distribution function with parameters $\boldsymbol{\vartheta}_g$, and $\pi_g > 0$ is the mixing weight of the g^{th} component such that $\sum_{g=1}^{G} \pi_g = 1$. Mixture model-based clustering methods can be over-parameterized in high-dimensional spaces, especially as the number of clusters increases. Subspace clustering allows to cluster data in low-dimensional subspaces, while keeping all the dimensions and by introducing restrictions to mixture parameters (Bouveyron and Brunet, 2014). Restrictions are introduced to the model parameters with the aim of obtaining parsimonious models, which are sufficiently flexible for clustering purposes.

*Department of Mathematics and Statistics, University of Guelph, Guelph, ON, Canada.
†Department of Molecular and Cellular Biology, University of Guelph, Guelph, ON, Canada.
‡Department of Mathematics and Statistics, McMaster University, Hamilton, ON, Canada.
§Department of Mathematical Sciences, Binghamton University, Binghamton, New York, USA.

The factor analysis model by Spearman (1904) assumes that a p-dimensional vector of observed variables, $\mathbf{X}$, can be modeled by a q-dimensional vector of latent factors, where $q<p$. As a result, the factor analysis model is useful in modeling the covariance structure of high-dimensional data using a small number of latent variables. The mixture of factor analyzers model was later introduced (Ghahramani and Hinton, 1997) and this model is able to concurrently perform clustering and, within each cluster, local dimensionality reduction. Consider n independent p-dimensional continuous variables $\mathbf{X}_1, \ldots, \mathbf{X}_n$, which come from a heterogeneous population with G subgroups. In mixture of factor analyzers framework, $\mathbf{X}_i$ is modelled as

$$\mathbf{X}_i = \boldsymbol{\mu}_g + \boldsymbol{\Lambda}_g \mathbf{U}_{ig} + \boldsymbol{\epsilon}_{ig},$$

for $i = 1, \ldots, n$ and $g = 1, \ldots, G$. Here, $\boldsymbol{\mu}_g$ is a $p \times 1$ vector of gth component mean, $\boldsymbol{\Lambda}_g$ is a $p \times q$ matrix of gth component factor loadings, $\mathbf{U}_{ig} \sim \mathcal{N}_q(\mathbf{0}, \mathbf{I}_q)$ is a $q \times 1$ vector of gth component latent factors, and $\boldsymbol{\epsilon}_{ig} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Psi}_g)$ is a $p \times 1$ vector of gth component errors with $\boldsymbol{\Psi}_g = \text{diag}(\psi_{g1}, \ldots, \psi_{gp})$. Note that the $\mathbf{U}_{ig}$ are independently distributed and are independent of the $\boldsymbol{\epsilon}_{ig}$, which are also independently distributed. Under this model, the density of $\mathbf{X}_i$ from a mixture of factor analyzers model is

$$f(\mathbf{x}_i \mid \boldsymbol{\Theta}) = \sum_{g=1}^{G} \pi_g f_g(\mathbf{x}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g \boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g),$$

Here, $\boldsymbol{\Theta} = (\pi_1, \ldots, \pi_G, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_G, \boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_G, \boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_G)$. Further, $\mathbf{X}_i | \mathbf{u}_{ig} \sim \mathcal{N}_p(\boldsymbol{\mu}_g + \boldsymbol{\Lambda}_g \mathbf{u}_{ig}, \boldsymbol{\Psi}_g)$. It should be noted that the $\boldsymbol{\Lambda}_g$ is not uniquely defined for $q > 1$. Therefore the q-dimensional space in which the factors lie can be determined, but the directions of these factors cannot be determined. However, this does not affect the clustering algorithm, because $\boldsymbol{\Lambda}_g \boldsymbol{\Lambda}_g'$ is unique. The number of free gth component covariance parameters that are reduced using factor analysis model is

$$\frac{1}{2}p(p+1) - \Big[pq + p - \frac{1}{2}q(q-1)\Big] = \frac{1}{2}\Big[(p-q)^2 - (p+q)\Big],$$

given that $(p-q)^2 > (p+q)$ (Lawley and Maxwell, 1962; McNicholas, 2016).

In 2008, this work was extended and a family of eight parsimonious Gaussian mixture models (PGMMs; McNicholas and Murphy, 2008) were introduced with parsimonious covariance structures. The PGMM family arises by considering the general mixture of factor analyzers model ($\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g \boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g$) and by allowing the constraints $\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$, $\boldsymbol{\Psi}_g = \boldsymbol{\Psi}$, and the isotropic constraint $\boldsymbol{\Psi}_g = \psi_g \mathbf{I}_p$. These covariance structures can have as few as $pq - q(q-1)/2 + 1$ free parameters or as many as $G[pq - q(q-1)/2 + p]$ free parameters, where $q = 1, 2, \ldots$. The number of covariance parameters are linear in data dimensionality making this family well suited for analysis of high-dimensional data (McNicholas and Murphy, 2010). The constraints allow for assuming a common structure in the component covariance matrix $\boldsymbol{\Sigma}_g$, if appropriate, and this enables a parsimonious model.

Previously, a model-based clustering methodology using mixtures of multivariate Poisson-log normal distribution (MPLN; Aitchison and Ho, 1989) was developed to analyze multivariate count measurements from RNA-seq studies (Silva et al., 2017). A p-dimensional random variable following a G-component mixtures of MPLN distribution is said to have a total of $G-1+Gp+Gp(p+1)/2$ free parameters. Here, $G-1$ parameters are contributed by the mixing proportions, Gp from the means and $Gp(p+1)/2$ from the covariance matrices. Since the largest contribution is through the covariance matrices, it is a natural focus for the introduction of parsimony.

In this work, a family of mixtures of MPLN factor analyzers that is analogous to the PGMM family is developed, by considering the constraints $\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}_g = \boldsymbol{\Psi}$. This family is referred to as the parsimonious mixtures of MPLN factor analyzers family (PMPLNFA). The proposed model simultaneously performs factor analysis and cluster analysis, by assuming that the discrete observed data have been generated by a factor analyzer model with continuous latent variables. Details of parameter estimation are provided, and both real and simulated data illustrations are used to demonstrate the clustering ability.

2 Methodology

2.1 Mixtures of MPLN Factor Analyzers

For genes $i \in \{1,\ldots,n\}$ and samples $j \in \{1,\ldots,p\}$, the MPLN distribution is modified to give

$$\begin{aligned} Y_{ij}|\theta_{ijg} &\sim \mathscr{P}(\exp\{\theta_{ijg}+\log s_j\}) \\ (\theta_{i1g},\ldots,\theta_{ipg})' &\sim \mathscr{N}_p(\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g), \end{aligned}$$

where $\mathcal{N}_p(\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)$ is the p-dimensional normal distribution with mean vector $\boldsymbol{\mu}_g$ and covariance matrix $\boldsymbol{\Sigma}_g$. The $\mathscr{P}$ denotes the Poisson distribution with parameters $\exp\{\theta_{ijg}\}$ and the $\mathbf{s}$ is a known vector of constants that represents the differences in library sizes for each sample j. Here, the mean of Y is $\mathbb{E}(Y_j)=\exp\{\boldsymbol{\mu}_{jg}+\frac{1}{2}\sigma_{jjg}\}\boldsymbol{m}_{jg}$ and the variance is $\mathbb{V}\mathrm{ar}(Y_j)=\boldsymbol{m}_{jg}+\boldsymbol{m}_{jg}^2(\exp\{\sigma_{jjg}\}-1)$.

Let $(\theta_{i1g},\ldots,\theta_{ipg})' \sim \mathcal{N}_p(\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)$ denote the transpose of p-dimensional vector $\boldsymbol{\theta}$ for a given observation i that follows the normal distribution. In the mixture of factor analyzers framework, $\boldsymbol{\theta}'_{ig}$ is modelled as

$$\boldsymbol{\theta}'_{ig} = \boldsymbol{\mu}_g + \boldsymbol{\Lambda}_g\mathbf{U}_{ig} + \boldsymbol{\epsilon}_{ig}, \tag{1}$$

for $i=1,\ldots,n$ and $g=1,\ldots,G$. Here, $\boldsymbol{\mu}_g$ is a $p\times 1$ vector of gth component mean, $\boldsymbol{\Lambda}_g$ is a $p\times q$ matrix of gth component factor loadings, $\mathbf{U}_{ig}\sim\mathcal{N}_q(\mathbf{0},\mathbf{I}_q)$ is a $q\times 1$ vector of gth component latent factors, and $\boldsymbol{\epsilon}_{ig}\sim\mathcal{N}_p(\mathbf{0},\boldsymbol{\Psi}_g)$ is a $p\times 1$ vector of gth component errors with $\boldsymbol{\Psi}_g=\mathrm{diag}(\psi_{g1},\ldots,\psi_{gp})$. The marginal distribution of $\boldsymbol{\theta}'_{ig}$ arising from the model in (1) is

$\mathcal{N}(\boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g)$. Conditional on $\mathbf{u}_{ig}$, this results $\boldsymbol{\theta}_{ig}'|\mathbf{u}_{ig} \sim \mathcal{N}_p(\boldsymbol{\mu}_g + \boldsymbol{\Lambda}_g\mathbf{u}_{ig}, \boldsymbol{\Psi}_g)$. Under this model, a G-component mixture of MPLN factor analyzers has the distribution

$$\begin{aligned} f(\mathbf{y};\boldsymbol{\Theta}) &= \sum_{g=1}^{G} \pi_g f_{\mathbf{Y}}(\mathbf{y}|\boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g) \\ &= \sum_{g=1}^{G} \pi_g \int_{\mathbb{R}^p} \left(\prod_{j=1}^{p} f(y_{ij}|\boldsymbol{\theta}_{ijg}, s_j) \right) f(\boldsymbol{\theta}_{ig}|\boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g)\ d\boldsymbol{\theta}_{ig}. \end{aligned}$$

By considering the constraints $\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}_g = \boldsymbol{\Psi}$, four different PMPLNFA models are introduced (see Table 1). Parameter estimation of these models are carried out using the MCMC-AECM algorithm.

Table 1: Nomenclature and covariance structure for the members of the PMPLNFA family.

Model ID	Loading matrix $\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$	Error variance $\boldsymbol{\Psi}_g = \boldsymbol{\Psi}$	Variance $\boldsymbol{\Sigma}_\mathbf{g}$	Free covariance parameters
CC	Constrained	Constrained	$\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$	$[pq - q(q-1)/2] + p$
CU	Constrained	Unconstrained	$\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}_g$	$[pq - q(q-1)/2] + Gp$
UC	Unconstrained	Constrained	$\boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}$	$G[pq - q(q-1)/2] + p$
UU	Unconstrained	Unconstrained	$\boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g$	$G[pq - q(q-1)/2] + Gp$

2.2 Parameter Estimation

At the first stage of the MCMC-AECM algorithm, when estimating $\boldsymbol{\vartheta}_1 = (\boldsymbol{\theta}_g, \pi_g, \boldsymbol{\mu}_g; g = 1, \ldots, G)$, the group labels $\mathbf{z}_{ig}$ and $\boldsymbol{\theta}_{ig}, i = 1, \ldots, n; g = 1, \ldots, G$, are the missing data. Hence, the complete-data log-likelihood for the mixture model is

$$\begin{aligned} l_{c1}(\boldsymbol{\vartheta}_1) = &\sum_{g=1}^{G} n_g \log \pi_g - \sum_{i=1}^{n}\sum_{g=1}^{G}\sum_{j=1}^{p} z_{ig} \exp\{\theta_{ijg} + \log s_j\} + \sum_{i=1}^{n}\sum_{i=g}^{G} z_{ig}(\boldsymbol{\theta}_{ig} + \log \mathbf{s})\mathbf{y}_i' \\ &- \sum_{i=1}^{n}\sum_{g=1}^{G}\sum_{j=1}^{p} z_{ig} \log y_{ijg}! - \frac{np}{2}\log 2\pi - \frac{1}{2}\sum_{g=1}^{G} n_g \log|\boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g| \\ &- \frac{1}{2}\sum_{g=1}^{G} n_g \mathrm{tr}\big\{\mathbf{S}_g\big(\boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g\big)^{-1}\big\}. \end{aligned}$$

Here $n_g = \sum_{i=1}^{n} z_{ig}$ and $\boldsymbol{S}_g$ represents the sample covariance matrix for component g, which has the form

$$\boldsymbol{S}_g = \frac{1}{n_g}\sum_{i=1}^{n} z_{ig}\, \mathbb{E}\left((\boldsymbol{\theta}_{ig} - \boldsymbol{\mu}_g^{(t)})(\boldsymbol{\theta}_{ig} - \boldsymbol{\mu}_g^{(t)})'\right). \tag{2}$$

At each E-step, the conditional expected value of $\boldsymbol{\theta}_{ig}$ and conditional expected value of group membership variable, z_{ig}, are respectively updated as

$$\begin{aligned}
&\mathbb{E}(\boldsymbol{\theta}_{ig}|\mathbf{y}_i) \simeq \frac{1}{N}\sum_{k=1}^{N}\boldsymbol{\theta}_{ig}^{(k)} \simeq \boldsymbol{\theta}_{ig}^{(t)},\\
&\mathbb{E}(Z_{ig}|\mathbf{y}_i,\boldsymbol{\theta}_{ig},\mathbf{s}) = \frac{\pi_g f(\mathbf{y}_i|\boldsymbol{\theta}_{ig}^{(t)},\mathbf{s}) f(\boldsymbol{\theta}_{ig}|\boldsymbol{\mu}_g^{(t)},\boldsymbol{\Lambda}_g^{(t)},\boldsymbol{\Psi}_g^{(t)})}{\sum_{h=1}^{G}\pi_h^{(t)} f(\mathbf{y}_i|\boldsymbol{\theta}_{ih}^{(t)},\mathbf{s}) f(\boldsymbol{\theta}_{ih}|\boldsymbol{\mu}_h^{(t)},\boldsymbol{\Lambda}_h^{(t)},\boldsymbol{\Psi}_h^{(t)})} =: z_{ig}^{(t)}.
\end{aligned} \tag{3}$$

Here, $\boldsymbol{\theta}_{ig}^{(k)}$ is the random sample simulated via `RStan` package for iterations $k = 1, \ldots, B$. As the values from initial iterations are discarded from further analysis to minimize bias, the number of iterations used for parameter estimation is N, where $N < B$. The expected value of the complete-data log-likelihood at first stage is

$$\begin{aligned}
\mathcal{Q}_1 \simeq &\sum_{g=1}^{G} n_g^{(t)} \log \pi_g^{(t)} - \sum_{i=1}^{n}\sum_{g=1}^{G}\sum_{j=1}^{p} z_{ig}^{(t)} \exp\{\mathbb{E}(\theta_{ijg}) + \log s_j\}\\
&+ \sum_{i=1}^{n}\sum_{i=g}^{G} z_{ig}^{(t)}(\mathbb{E}(\boldsymbol{\theta}_{ig}) + \log \mathbf{s})\mathbf{y}_i' - \sum_{i=1}^{n}\sum_{g=1}^{G}\sum_{j=1}^{p} z_{ig}\log y_{ijg}! - \frac{np}{2}\log 2\pi\\
&- \frac{1}{2}\sum_{g=1}^{G} n_g^{(t)} \log|\boldsymbol{\Lambda}_g^{(t)}\boldsymbol{\Lambda}_g^{(t)\prime} + \boldsymbol{\Psi}_g^{(t)}| - \frac{1}{2}\sum_{g=1}^{G} n_g^{(t)}\mathrm{tr}\big\{\mathbf{S}_g\big(\boldsymbol{\Lambda}_g^{(t)}\boldsymbol{\Lambda}_g^{(t)\prime} + \boldsymbol{\Psi}_g^{(t)}\big)^{-1}\big\}.
\end{aligned}$$

Here $n_g^{(t)} = \sum_{i=1}^{n} z_{ig}^{(t)}$ and $n = \sum_{g=1}^{G} n_g$. Note, the $z_{ig}^{(t)}$ replaces z_{ig} in $\boldsymbol{S}_g$, cf. (2). Maximizing $\mathcal{Q}_1$ with respect to π_g and $\boldsymbol{\mu}_g$, respectively, leads to the parameter updates,

$$\pi_g^{(t+1)} = \frac{n_g^{(t)}}{n}, \qquad \boldsymbol{\mu}_g^{(t+1)} = \frac{\sum_{i=1}^{n} z_{ig}^{(t)}\,\mathbb{E}(\boldsymbol{\theta}_{ig})}{n_g^{(t)}}.$$

At the second stage of the MCMC-AECM algorithm, when estimating $\boldsymbol{\vartheta}_2 = (\boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g; g = 1, \ldots, G)$, the group labels z_{ig}, the latent factors $\mathbf{u}_{ig}$, and $\boldsymbol{\theta}_{ig}, i = 1, \ldots, n; g = 1, \ldots, G$, are taken to be the missing data. Thus, the complete-data log-likelihood is

$$\begin{aligned}
l_{c2}(\boldsymbol{\vartheta}_2) = C + \sum_{g=1}^{G}\Bigg[&- \frac{n_g^{(t)}}{2}\log|\boldsymbol{\Psi}_g^{(t)}| - \frac{n_g^{(t)}}{2}\mathrm{tr}\{\boldsymbol{\Psi}_g^{(t)^{-1}}\boldsymbol{S}_g\}\\
&+ \sum_{i=1}^{n} z_{ig}^{(t)}(\boldsymbol{\theta}_i^{(t)} - \boldsymbol{\mu}_g^{(t+1)})'\boldsymbol{\Psi}_g^{(t)^{-1}}\boldsymbol{\Lambda}_g^{(t)}\boldsymbol{u}_i - \frac{1}{2}\mathrm{tr}\big\{\boldsymbol{\Lambda}_g^{(t)\prime}\boldsymbol{\Psi}_g^{(t)^{-1}}\boldsymbol{\Lambda}_g^{(t)}\sum_{i=1}^{n} z_{ig}^{(t)}\boldsymbol{u}_i\boldsymbol{u}_i'\big\}\Bigg].
\end{aligned}$$

Here C is a constant with respect to $\boldsymbol{\Lambda}_g$ and $\boldsymbol{\Psi}_g$. The expected value of the complete-data log-likelihood for the second stage is

$$\begin{aligned}\mathcal{Q}_2 \simeq C + \frac{1}{2}\sum_{g=1}^{G} n_g^{(t)} \Bigg[&\log|\boldsymbol{\Psi}_g^{(t)-1}| - \mathrm{tr}\big\{\boldsymbol{\Psi}_g^{(t)-1}\boldsymbol{S}_g\big\} + 2\mathrm{tr}\big\{\boldsymbol{\Psi}_g^{(t)-1}\boldsymbol{\Lambda}_g^{(t)}\hat{\boldsymbol{\beta}}_g^{(t)}\boldsymbol{S}_g\big\} \\ &- \mathrm{tr}\big\{\boldsymbol{\Lambda}_g^{(t)\prime}\boldsymbol{\Psi}_g^{(t)-1}\boldsymbol{\Lambda}_g^{(t)}\boldsymbol{\Phi}_g^{(t)}\big\}\Bigg].\end{aligned} \tag{4}$$

Here, $\hat{\boldsymbol{\beta}}_g^{(t)}$ is a $q \times p$ matrix that is given by $\hat{\boldsymbol{\beta}}_g^{(t)} = \boldsymbol{\Lambda}_g^{(t)\prime}\big(\boldsymbol{\Lambda}_g^{(t)}\boldsymbol{\Lambda}_g^{(t)\prime} + \boldsymbol{\Psi}_g^{(t)}\big)^{-1}$ and $\boldsymbol{\Phi}_g^{(t)}$ is a symmetric $q \times q$ matrix that is given by $\boldsymbol{\Phi}_g^{(t)} = \mathbf{I}_q - \hat{\boldsymbol{\beta}}_g^{(t)}\boldsymbol{\Lambda}_g^{(t)} + \hat{\boldsymbol{\beta}}_g^{(t)}\boldsymbol{S}_g\hat{\boldsymbol{\beta}}_g^{(t)\prime}$. The $z_{ig}^{(t+1)}$ and $\boldsymbol{\theta}_{ig}^{(t+1)}$ are computed as in (3) with the estimates of $\pi_g^{(t+1)}$ and $\boldsymbol{\mu}_g^{(t+1)}$ as calculated in the first stage of the MCMC-AECM algorithm. In turn, $\boldsymbol{\mu}_g^{(t+1)}$ replaces $\boldsymbol{\mu}_g^{(t)}$ and $z_{ig}^{(t+1)}$ replaces z_{ig} in $\boldsymbol{S}_g$, cf. (2). Differentiating $\mathcal{Q}_2$ with respect to $\boldsymbol{\Lambda}_g$ and $\boldsymbol{\Psi}_g^{-1}$, respectively, leads to

$$\boldsymbol{\Lambda}_g^{(t+1)} = \boldsymbol{S}_g\hat{\boldsymbol{\beta}}_g^{(t)\prime}\boldsymbol{\Phi}_g^{(t)^{-1}}, \qquad \boldsymbol{\Psi}_g^{(t+1)} = \mathrm{diag}\{\boldsymbol{S}_g - \boldsymbol{\Lambda}_g^{(t+1)}\hat{\boldsymbol{\beta}}_g^{(t)}\boldsymbol{S}_g\}.$$

The form of complete-data log-likelihood and the parameter estimates will vary depending on which of the four models in the PMPLNFA family is under consideration. Note, the parameter updates are analogous to those given by McNicholas and Murphy (2008) for the Gaussian case. The MCMC-AECM algorithm iteratively updates the parameters until convergence. The resulting $z_{ig}^{(t+1)}$ values at the convergence of MCMC-AECM algorithm are estimates of the posterior probability of the group membership for each observation and can be used to cluster observations into G clusters. The difference in the number of free parameters estimated using mixtures of MPLN model (Silva et al., 2017) and the PMPLNFA model are illustrated in Figure 1 and Figure 2.

2.3 Convergence

Convergence of MCMC-AECM algorithm is determined following the criteria outlined by Silva et al. (2017). To determine whether the MCMC chains have converged to the posterior distribution, the *potential scale reduction factor* (Gelman and Rubin, 1992) and the *effective number of samples* (Gelman et al., 2013) is used. To check if the likelihood has reached its maximum, the Heidelberger and Welch's convergence diagnostic (Heidelberger and Welch, 1983) is used.

2.4 Initialization

Initialization of $\boldsymbol{\mu}_g$ and $\boldsymbol{S}_g$ is done following the criteria outlined by Silva et al. (2017). The matrices $\boldsymbol{\Lambda}_g$ and $\boldsymbol{\Psi}_g$ are initialized following McNicholas and Murphy (2008) using

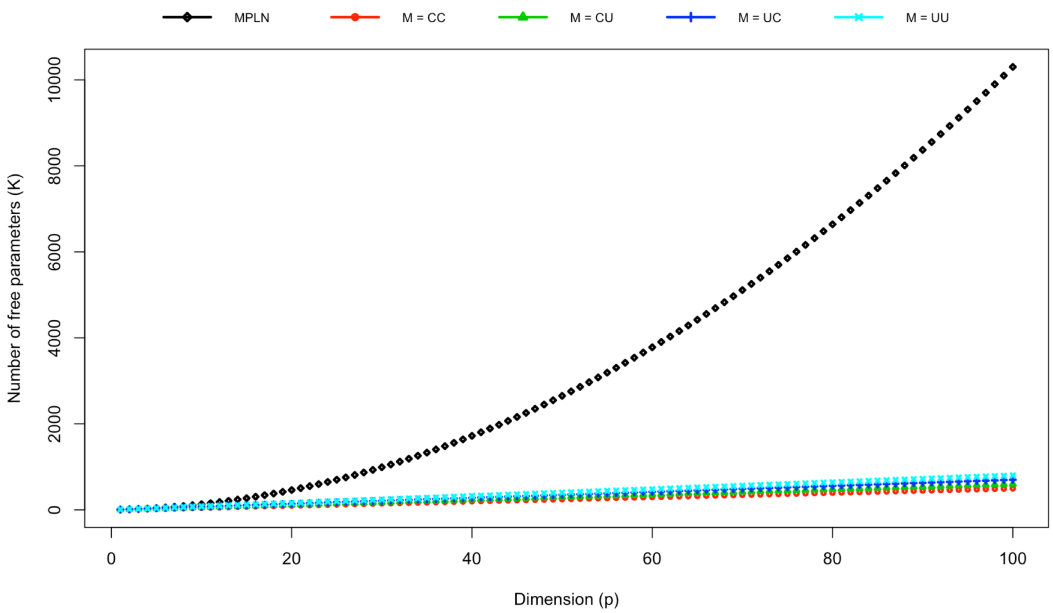

Figure 1: Scatter plot illustrating how the number of free parameters K grows with data dimensionality p for the mixtures of MPLN model and the four members of the PMPLNFA family (M = CC, CU, UC, UU). Here $G = 2, q = 2$, and $p = 1 : 100$.

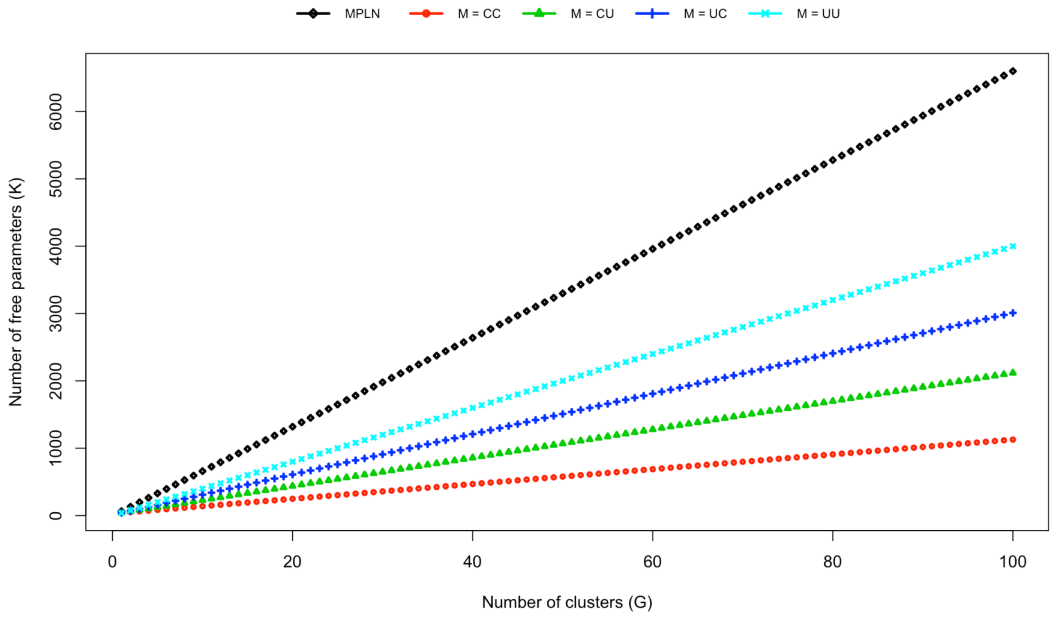

Figure 2: Scatter plot illustrating how the number of free parameters K grows with the number of clusters G for the mixtures of MPLN model and the four members of the PMPLNFA family. Here $G = 1 : 100, q = 2$, and $p = 10$.

the eigen-decomposition of $\hat{\boldsymbol{S}}_g^{(0)}$ as follows. The initial values of the jth column of $\boldsymbol{\Lambda}_g$ are set as $\gamma_j^{(0)} = \sqrt{d_j}\rho_{ij}$, where d_j is the jth largest eigenvalue of $\hat{\boldsymbol{S}}_g^{(0)}$ and ρ_{ij} is the ith eigenvector corresponding to the jth largest eigenvalue of $\hat{\boldsymbol{S}}_g^{(0)}$ for $i \in \{1, \ldots, p\}$ and $j \in \{1, \ldots, q\}$. The $\boldsymbol{\Psi}_g$ is then initialized as $\hat{\boldsymbol{\Psi}}_g^{(0)} = \text{diag}(\hat{\boldsymbol{S}}_g^{(0)} - \hat{\boldsymbol{\Lambda}}_g^{(0)}\hat{\boldsymbol{\Lambda}}_g^{(0)'})$. For initialization of $\hat{z}_{ig}$, two algorithms are provided: k-means and random. For k-means initialization, k-means clustering is performed on the dataset and the resulting cluster memberships are used for the initialization of $\hat{z}_{ig}$. For random initialization, random values are chosen for $\hat{z}_{ig} \in [0, 1]$ such that $\sum_{i=1}^{n} z_{ig} = 1$ for all i. If multiple initialization runs are considered, the $\hat{z}_{ig}$ values corresponding to the run with the highest log-likelihood value are used for downstream analysis.

2.5 Parallel Implementation

The algorithm is parallelized using the `parallel` package (R Core Team, 2017) and `foreach` package (Revolution Analytics and Weston, 2015), to run each combination of cluster, G, latent factor, q and component scale matrix, M in parallel, each one on a different processor. All data analyses were done using the parallelized code.

2.6 Model Selection and Performance Assessment

For this analysis, four model selection criteria are used. These include the Akaike information criterion (AIC; Akaike, 1973), the Bayesian information criterion (BIC; Schwarz, 1978), a variation of the AIC used by Bozdogan (1994) called AIC3, and the integrated completed likelihood (ICL; Biernacki et al., 2000). AIC, BIC, AIC3 and ICL are used to select the best PMPLNFA model in terms of the number of clusters, latent factors, and the structure of the component scale matrices. Performance assessment is done using adjusted Rand index (ARI; Hubert and Arabie, 1985).

2.7 Transcriptome Data Analysis

To illustrate the use of PMPLNFA, it was applied to a co-expression analysis of differentially expressed genes identified in the RNA-seq study of cranberry beans (*P. vulgaris*) by Freixas-Coutin et al. (2017). The study was conducted to evaluate if the changes in the seed coat transcriptome were associated with proanthocyanidin levels as a function of seed development. For this purpose, RNA-seq was used to monitor the transcriptional dynamics in the seed coats of darkening and non-darkening cranberry beans at three developmental stages: early, intermediate and mature. The data are available on the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) under the BioProject PRJNA380220.

The study identified 1336 differentially expressed genes, which were used for clustering. The raw read counts for genes were obtained from Binary Alignment/Map files using samtools (Li et al., 2009) and HTSeq (Anders et al., 2015). The median value from the 3 replicates per each developmental stage was used. The genes were clustered for a range of $G = 1, \ldots, 10$ using k-means initialization with 3 runs. To identify if co-expressed genes are implicated in similar biological processes or pathways, a gene ontology (GO) enrichment analysis was performed on the gene clusters using the Singular Enrichment Analysis tool available on AgriGO (Du et al., 2010). A significance level of 5% was used with Fisher statistical testing and Yekutieli multi-test adjustment.

2.8 Simulation Study

Simulation studies were conducted to illustrate the ability to recover the true underlying parameters by the PMPLNFA. The count range in all simulated datasets represented count range and library sizes observed in RNA-seq data. Two dimensionality sizes were considered, $p = 6$ and $p = 10$, each with $q = 2$ latent factors for $G = 2$ clusters with $n = 500$ and $\pi_1 = 0.6$. For $p = 6$, data was generated from all the models with covariance structure $M =$ CC, CU, UC and UU. For $p = 10$, the covariance structure was set to be $M =$ CC and CU. The last simulation setting had $p = 6$, $q = 3$, $G = 1$ and $M =$ CC. For each of the settings (see Table 2), two datasets were generated. Each dataset was run for a clustering range of $G = 1, \ldots, 3$, a latent factor range of $q = 1, \ldots, 3$ and for all four component scale matrices listed in Table 1. All data analyses were performed using k-means initialization with 3 runs.

Table 2: Various settings used in the simulation study for PMPLNFA.

Setting	Cluster, G	Latent factor, q	Component scale matrix, M	Dimensions
1	2	2	CC	500 x 6
2	2	2	CU	500 x 6
3	2	2	UC	500 x 6
4	2	2	UU	500 x 6
5	2	2	CC	500 x 10
6	2	2	CU	500 x 10
7	1	3	CC	500 x 6

For both real and simulation data analyses, the normalization factors representing library size estimate for samples were obtained using the trimmed mean of M values from `calcNormFactors` function of `edgeR` package (Robinson and Oshlack, 2010; McCarthy et al., 2012). All data analyses were performed on a MacBook Pro with 3.1 GHz quad-core Intel Core i7 processor and 16 GB RAM, and the Joyce high performance computing cluster at the McMaster University, Hamilton, ON, Canada. The Joyce system is an Intel(R) Xeon(R) CPU E5-4627 v2 with 32 cores, 3.3 GHz and 256 GB RAM.

2.9 Software availability

The source code is made available at `https://github.com/anjalisilva/mixMPLNFA` and is released under the open source MIT license.

3 Results

3.1 Transcriptome Data Analysis

All information criteria selected a model with $G = 5, q = 3$ and $M = \text{UU}$ for the cranberry bean RNA-seq study. The clustering results via BIC for transcriptome data are summarized on Figure 3. The expression patterns among the different clusters for this model are provided in Figure 4. The compositions of genes in Clusters 1 through 5 were as follows: 722, 59, 180, 111 and 264. Cluster 1 genes were expressed roughly constantly across all samples. The GO enrichment analysis identified genes belonging to oxidation reduction, dehydrogenase activity, binding and electron carrier activity. Cluster 2 genes showed variable expression and belonged to pathogenesis and multi-organism process.

Cluster 3 genes showed higher expression in early and intermediate developmental stages relative to mature developmental stage, across both the darkening and non-darkening varieties. These genes belonged to lipid/ fatty acid biosynthesis, lipid/ fatty acid metabolic process, synthase activity and transferase activity. Cluster 4 expression patterns revealed higher expression in early developmental stage and gradual decrease during intermediate followed by mature developmental stage. GO enrichment analysis identified binding, hydrolase activity, oxidoreductase activity and catalytic activity. Finally, Cluster 5 expression patterns revealed lower expression in intermediate developmental stage relative to other developmental stages, regardless of the variety. These genes belonged to hydrolase activity, ATPase activity, pyrophosphatase activity, nucleoside-triphosphatase activity and aromatic amino acid family metabolic process.

Due to the similarities in Clusters 1 and 4, with respect to GO enrichment terms, these clusters were further analyzed. Although both clusters were enriched for binding, Cluster 1 was enriched for iron ion binding (GO:0005506), heme binding (GO:0020037) and tetrapyrrole binding (GO:0046906). Conversely, Cluster 4 was enriched for coenzyme binding (GO:0050662) and flavin adenine dinucleotide (FAD)-binding (GO:0050660). In terms of oxidation reduction, Cluster 1 was enriched for oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen (GO:0016705). However, Cluster 4 was enriched for oxidoreductase activity, acting on the CH-OH group of donors, nicotinamide adenine dinucleotide (NAD) or nicotinamide adenine dinucleotide phosphate (NADP) as acceptor (GO:0016616) and oxidoreductase activity, acting on CH-OH group of donors (GO:0016614). Overall, the results obtained from GO annotations suggest distinct themes among the five clusters.

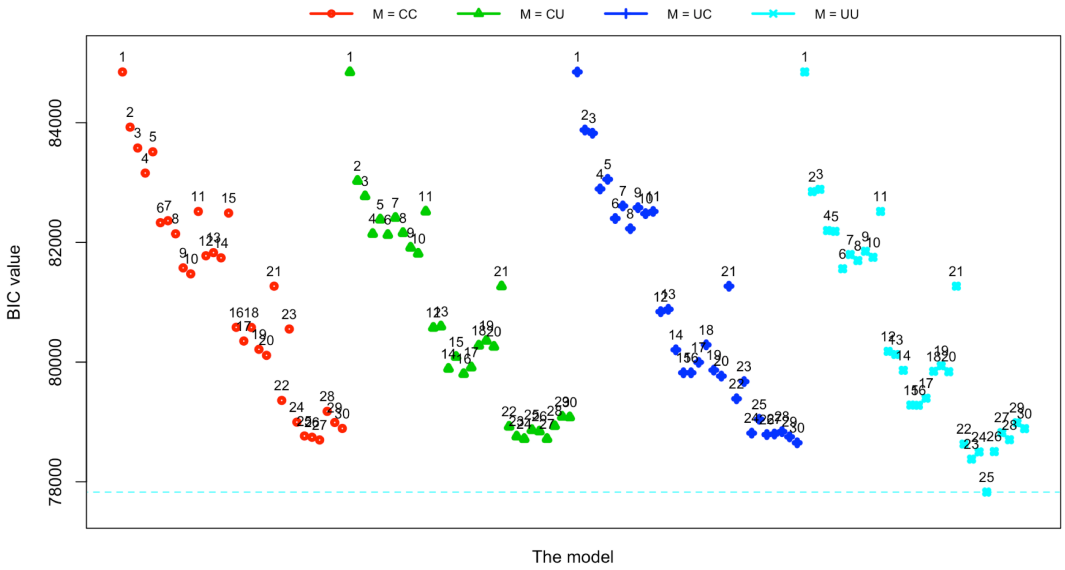

Figure 3: Plot of BIC value versus the model for the cranberry bean RNA-seq data. Numbers denote the corresponding model and the color denotes the component scale matrix M. Model 1 is $(G = 1, q = 1)$, model 2 is $(G = 2, q = 1)$, model 3 is $(G = 3, q = 1)$, ... The dotted line indicates the model selected by BIC, which is $G = 5, q = 3, M = \text{UU}$.

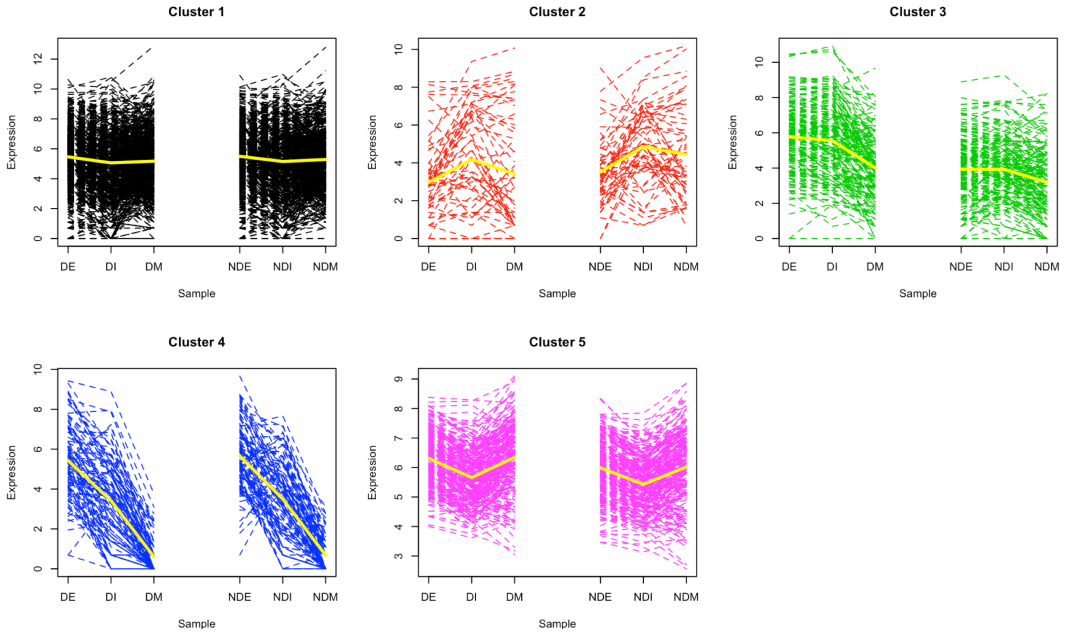

Figure 4: The expression patterns for the five clusters in the $G = 5, q = 3$ and $M = \text{UU}$ model selected by all information criteria for the cranberry bean RNA-seq data. The expression is in log-transformed counts. The yellow line represents the mean expression level for each cluster. The samples are DE: darkening early, DI: darkening intermediate, DM: darkening mature, NDE: non-darkening early, NDI: non-darkening intermediate and NDM: non-darkening mature.

3.2 Simulation Study

The clustering results obtained using different model selection criteria and corresponding average ARI values are summarized in Table 3. The model selection criteria gave comparable results in terms of the number of clusters and the number of latent factors, however, occasionally AIC and AIC3 failed to select the correct number of clusters. In such situations, AIC and AIC3 selected a higher number of clusters. Across all settings, the ARI values were equal to or very close to one, indicating that the algorithm is able to assign observations to the proper clusters, i.e., the clusters that were originally used to generate the simulation datasets. In terms of selecting the component scale matrix, both AIC and AIC3 selected the incorrect model for some settings. In these situations, the unconstrained loading matrix or the unconstrained error variance was selected. Overall, the simulation experiments illustrate that the proposed algorithm is able to recover proper clusters, latent factors and the component scale matrix for PMPLNFA.

Table 3: Model selection results of the clusters (average ARI, standard deviation), latent factors and component scale matrices for each simulation setting using different model selection criteria.

	Cluster, G				Latent factor, q				Component scale matrix, M			
Setting	BIC	ICL	AIC	AIC3	BIC	ICL	AIC	AIC3	BIC	ICL	AIC	AIC3
1	2 (0.99, 0.01)	2 (0.99, 0.01)	2 (0.99, 0.01)	2 (0.99, 0.01)	2	2	2	2	CC	CC	UC	CC
2	2 (1.00, 0.00)	2 (1.00, 0.00)	2 (1.00, 0.00)	2 (1.00, 0.00)	2	2	2	2	CU	CU	CU	CU
3	2 (1.00, 0.00)	2 (1.00, 0.00)	2 (1.00, 0.00)	2 (1.00, 0.00)	2	2	2	2	UC	UC	UU	UU
4	2 (1.00, 0.00)	2 (1.00, 0.00)	2 (1.00, 0.00)	2 (1.00, 0.00)	2	2	2	2	UU	UU	UU	UU
5	2 (1.00, 0.00)	2 (1.00, 0.00)	2–3 (0.99, 0.01)	2–3 (0.99, 0.01)	2	2	2	2	CC	CC	CC/CU	CC/CU
6	2 (1.00, 0.00)	2 (1.00, 0.00)	2 (1.00, 0.00)	2 (1.00, 0.00)	2	2	2	2	CU	CU	CU	CU
7	1 (1.00, 0.00)	1 (1.00, 0.00)	1 (1.00, 0.00)	1 (1.00, 0.00)	3	3	3	3	CC	CC	CC	CC

4 Discussion

A mixture of factor analyzers model for MPLN distribution as well as a family of mixture models based thereon is introduced. This is the first use of a mixture of MPLN factor analyzer distributions within the literature. To our knowledge, this is also the first use of a mixture of discrete factor analyzers within the literature. The proposed models are well-suited to high-dimensional applications as the number of scale parameters is linear in data

dimensionality for all four models, as opposed to in traditional MPLN, where the parameters grow quadratically. The PMPLNFA family of models can be easily extended to consider an isotropic noise. Further, a MPLN factor analysis model can be obtained as a special case of the mixture of PMPLNFA, i.e., with $G = 1$. Extensions applicable to factor analyzers model, such as the mixture of common factor analyzers (Baek et al., 2010) can be applied to the PMPLNFA family in future. A mixture of common factor analyzers is a restrictive form of the mixture of factor analyzers model and can be useful when the number of clusters and dimensionality are very large.

Acknowledgments

This research was supported by the Ontario Graduate Fellowship (Silva), Arthur Richmond Memorial Scholarship (Silva), and the Canada Natural Sciences and Engineering Research Council grant 400920-2013 (Subedi).

References

Aitchison, J. and C. H. Ho (1989). The multivariate Poisson-log normal distribution. *Biometrika 76*(4), 643–653.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, pp. 267–281. Springer Verlag.

Anders, S., P. T. Pyl, and W. Huber (2015). HTSeq-a Python framework to work with high-throughput sequencing data. *Bioinformatics 31*(2), 166–169.

Annis, J., B. J. Miller, and T. J. Palmeri (2016). Bayesian inference with Stan: A tutorial on adding custom distributions. *Behavior Research Methods 49*, 1–24.

Baek, J., G. J. McLachlan, and L. K. Flack (2010). Mixtures of factor analyzers with common factor loadings: applications to the clustering and visualization of high-dimensional data. *IEEE Trans Pattern Anal Mach Intelligence 32*, 1298–1309.

Biernacki, C., G. Celeux, and G. Govaert (2000). Assessing a mixture model for clustering with the integrated classification likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*(7), 719–725.

Bouveyron, C. J. and C. Brunet (2014). Model-based clustering of high-dimensional data: A review. *Computational Statistics and Data Analysis 71*, 52–78.

Bozdogan, H. (1994). Mixture-model cluster analysis using model selection criteria and a new informational measure of complexity. In *Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach: Volume 2 Multivariate Statistical Modeling*, pp. 69–113. Dordrecht: Springer Netherlands.

Du, Z., X. Zhou, Y. Ling, Z. Zhang, and Z. Su (2010). agriGO: a GO analysis toolkit for the agricultural community. *Nucleic Acids Research 38*, W64–W70.

Freixas-Coutin, J. A., S. Munholland, A. Silva, S. Subedi, L. Lukens, W. L. Crosby, K. P. Pauls, and G. G. Bozzo (2017). Proanthocyanidin accumulation and transcriptional responses in the seed coat of cranberry beans (*Phaseolus vulgaris* L) with different susceptibility to postharvest darkening. *BMC Plant Biology 17*(89).

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013). *Bayesian Data Analysis*. Boca Raton, FL: Chapman & Hall/CRC Press.

Gelman, A. and D. B. Rubin (1992). Inference from iterative simulation using multiple sequences. *Stat Sci 7*(4), 457–472.

Ghahramani, Z. and G. E. Hinton (1997). The EM algorithm for mixtures of factor analyzers. Technical Report CRG-TR-96-1 University of Toronto.

Heidelberger, P. and P. D. Welch (1983). Simulation run length control in the presence of an initial transient. *Operations Research 31*(6), 1109–1144.

Hubert, L. and P. Arabie (1985). Comparing partitions. *Journal of Classification 2*, 193–218.

Junk-Knievel, D. C., A. Vandenberg, and K. E. Bett (2008). Slow darkening in pinto bean (*Phaseolus vulgaris* L) seed coats is controlled by a single major gene. *Crop Science 48*(1), 189–193.

Lawley, D. N. and A. E. Maxwell (1962). Factor analysis as a statistical method. *Journal of the Royal Statistical Society: Series D 12*, 209–229.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and . G. P. D. P. Subgroup (2009). The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics 25*(16), 2078–2079.

McCarthy, J. D., Y. Chen, and K. G. Smyth (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research 40*(10), 4288–4297.

McLachlan, G. J. and K. E. Basford (1988). *Mixture Models Inference and Applications to Clustering*. New York: Marcel Dekker.

McLachlan, G. J. and D. Peel (2000). *Finite Mixture Models*. New York: Wiley.

McNicholas, P. D. (2016). *Mixture Model-based Classification*. Boca Raton: Chapman and Hall/CRC Press.

McNicholas, P. D. and T. B. Murphy (2010). Model-based clustering of microarray expression data via latent Gaussian mixture models. *Bioinformatics 26*, 2705–2712.

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

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Revolution Analytics and S. Weston (2015). *foreach: Provides Foreach Looping Construct for R*. R package version 1.4.3.

Robinson, M. D. and A. Oshlack (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology 11*, R25.

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

Silva, A., S. J. Rothstein, P. D. McNicholas, and S. Subedi (2017). A Multivariate Poisson-Log Normal Mixture Model for Clustering Transcriptome Sequencing Data. arXiv preprint arXiv:1711.11190.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology 15*, 72–101.

Stan Development Team (2016). RStan: the R interface to Stan. R package version 2.15.1.

Wolfe, J. H. (1965). A computer program for the maximum likelihood analysis of types. Technical Bulletin 65–15, US Naval Personnel Research Activity.