Cluster Analysis Of Microbiome Data Via Mixtures Of Dirichlet-Multinomial Regression Models

by

Drew Neish

A Thesis
presented to
The University Of Guelph

In partial fulfilment of requirements
for the degree of
Master of Science
in
Mathematics and Statistics

Guelph, Ontario, Canada

© Drew Neish, December, 2015

ABSTRACT

CLUSTER ANALYSIS OF MICROBIOME DATA VIA MIXTURES OF DIRICHLET-MULTINOMIAL REGRESSION MODELS

Drew Neish
University of Guelph, 2015

Advisors:
Dr. S. Dang (Subedi)
Dr. Z. Feng

The human gut microbiome is a source of genetic and metabolic diversity, and exploring the relationship between biological/environmental covariates and the resulting taxonomic composition of the gut microbial community is an active area of research. Previously, a Dirichlet-multinomial regression framework has been suggested to model this relationship, but it did not account for any latent group structure which has been observed across microbiome samples which share similar biota compositions (known as enterotypes). Here, a finite mixture of Dirichlet-multinomial regression models is proposed and illustrated in order to account for the enterotype structure and allow for a probabilistic investigation of the relationship between bacterial abundance and biological/environmental covariates within each inferred enterotype. Furthermore, finite mixtures of regression models which incorporate the concomitant effect of the covariates on the resulting mixing proportions are also proposed and examined within the Dirichlet-multinomial framework.

Acknowledgements

I would first and foremost like to thank my advisors Dr. Sanjeena Dang and Dr. Zeny Feng for their continuous guidance, expertise, and patience throughout the course of my research. Their mentorship has broadened my academic horizons and it has been a great pleasure to work with them. My masters experience has been booth productive and enriching, and without their support this work could not have been accomplished.

I would also like to thank all the faculty in the Department of Mathematics and Statistics for their instruction, assistance, and contagious enthusiasm for their research which I hope to carry with me in my future.

Finally, I would like to express my sincere gratitude to my friends and family for keeping me sane with their love and encouragement every step of the way.

Table of Contents

List of Tables

List of Figures

Chapter 1

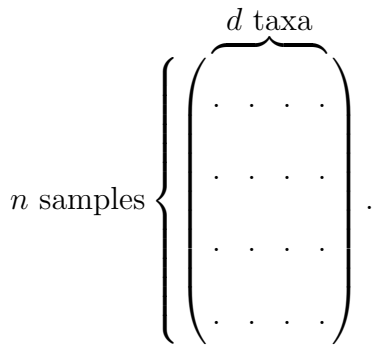
Introduction

1.1 The microbiome

It has been estimated that the amount of microorganisms residing on or within the human body outnumber human cells ten-fold (Ley et al., 2006), and we coexist with this microbiota in a mutualistic relationship which provides us with beneficial metabolic and genetic capabilities we alone do not possess (Backhed et al., 2005). Specifically, the human microbiota have been shown to play roles in digestive enzyme activity (Cantarel et al., 2012), pathogen protection (Round and Mazmanian, 2009), and vitamin synthesis (LeBlanc et al., 2013). This relationship can however become detrimental, and changes in the microbiota can lead to diabetes (Qin et al., 2012), obesity (Turnbaugh et al., 2009), inflammatory bowel disease (Greenblum et al., 2012), colorectal cancer (Tjalsma et al., 2012), and many other diseases/factors which impact the health of an individual. The microbiome refers to the collection of genomic information from the microbiota, and compositional information regarding which organisms populate the human microbiome as well as their effects on human health is currently an active area of research.

Next generation DNA sequencing technologies have allowed for the explo-

ration of microbiome composition without the need for isolation and culturing (Streit and Schmitz, 2004). Typically, microbial metagenomic data results from either targeted amplicon sequencing or shotgun metagenomics. Targeted amplicon sequencing involves the PCR amplification of a specific gene region and then assigning the targeted amplicon sequences to samples with the use of barcodes; a variable region of the 16S rRNA gene is most commonly chosen due to its omnipresence in bacterial organisms and its fast evolving regions which are useful for identification (Kuczynski et al., 2012). For shotgun metagenomics, DNA is extracted from all cells in a community and is subsequently randomly sheared in many short sequences, and these fragments are then sequenced to obtain reads of the different community members from throughout the genome. In either case, it is then possible to classify sequence reads generated from each technique against known taxa to identify the organisms present in each sample. In many situations, microorganisms will not have been taxonomically classified at the species level, so lower resolution phylogenetic levels such as phylum or genus must be used. An alternate strategy is to classify sequences into operational taxonomic units (OTUs) at a specified similarity level which is independent of taxonomic rank (Eckburg et al., 2005). The taxonomic counts for each organism in each sample are then represented in a matrix showing the frequency of each taxon in each sample; thereby displaying the microbiome composition across n individuals over d taxa at a pre-specified level:



1.2 Enterotypes

The gut microbiome is of particular interest and hosts the largest number of microbes in the human body. With more than 3×10^6 genes, the set of human gut microbial genes is approximately 150 times larger than the human gene complement (Qin et al., 2010). Within individual samples, the observed species diversity is high, but unevenly distributed with Firmucites, Bacteroidetes, and Actinobacteria being the dominating phyla (Zoetendal et al., 2008). Arumugam et al. (2011) explored the constitution of the gut microbiome across individuals which differed on a variety of factors and published their discovery of three discrete enterotypes; where an enterotype defined to be a classification of a living organism based on the composition of its gut microbiome. The authors made use of the partitioning around medoids (PAM) algorithm (Kaufman and Rousseeuw, 1987) to identify the three underlying enterotypes amongst microbiome samples, and concluded that enterotype classification is mainly dependent on the amounts of three bacterial genera: Bacteroides, Prevetella, and Ruminoccus present in the sample. Their dataset was small (39 individuals) but they suggested that enterotypes are most strongly influenced by nutrient

intake, and are not dictated by age, weight, gender, or national divisions. The cause of enterotype differentiation is still unkown, but another possible explanation comes from the longitudinal study of Bergström et al. (2014) which shows the intestines of infants are randomly colonized by different pioneering species of microbes only allowing certain species to follow, and enterotype establishment occurs between 9 and 36 months. Arumugam et al. (2011) also suggested that practical importance of these enterotypes include differential response to diet or drug intake. Whether enterotypes actually represent discrete clusters, or simply concentrated areas within a gradient of microbiome samples is still an active area of debate requiring more data to be resolved (Yong, 2012).

Wu et al. (2011) examined the fecal samples of 98 healthy individuals and gut bacterial counts were obtained based on 16S rRNA sequencing. A food frequency questionnaire was also used to collect dietary information on 214 micro-nutrients. They again used the PAM method for clustering, but found there were only two underlying enterotypes distinguished primarily by levels of Bacteroides and Prevotella. Their work suggested that enterotypes are strongly influenced by long-term dietary patterns.

1.3 Statistical approaches for microbiome data

Chen and Li (2013) expanded on the work of Wu et al. (2011) by developing statistical methods to examine the relationship between the covariates associated with gut microbiome composition. They utilized a sparse Dirichlet-multinomial regression

technique to examine the relationship between nutrients and bacterial count in each sample. Their work also utilized a penalized likelihood method for variable selection and estimation in order to identify most influential nutrients with respect to bacterial counts. This framework did not however acknowledge the underlying enterotype structure and possible differing covariate effects on taxa counts between enterotypes.

Holmes et al. (2012) suggested the use of model-based clustering on microbial metagenomic data for a probabilistic investigation of the underlying group structure. They made use of Dirichlet-multinomial mixture models to cluster samples and handle the heterogeneity and over dispersion present in microbiome datasets. This framework proved useful and, as a classifier, demonstrated better performance than the use of random forests on real microbiome samples. Their work suggested the benefit of using Dirichlet-multinomial mixtures to model gut microbiome data, however, the effect of covariates on resulting taxa counts was not explored.

The focus of this work is to improve on current statistical approaches which relate microbiome composition data to different environmental or biological covariates. The Dirichlet-multinomial regression approach of Chen and Li (2013) shows good performance but does not take into account the enterotype clustering observed by Arumugam et al. (2011), Holmes et al. (2012), Wu et al. (2011), and others. To account for this underlying group structure, the use of finite mixtures of Dirichlet-multinomial regression models are proposed. Both the traditional finite mixture model framework as well as an extension to account for concomitant effects of the covariates are explored to better model the relationship between covariates and resulting taxa counts.

Parameter estimation for a mixture of Dirichlet-multinomial regression models is accomplished using a generalized expectation-maximization algorithm (Dempster et al., 1977). To increase the log-likelihood in each iteration, we utilize a minorization-maximization algorithm and a Newton-Raphson algorithm as optimization strategies. Clustering performance and parameter estimation using both approaches are also compared to those from a standard optimization method: the simulated annealing method from the "optim" function implemented in the base stats package in R (R Core Team, 2015).

This thesis is arranged as follows. Chapter 2 details the development of relevant methodologies as well as criteria for model selection and performance assessment. Chapter 3 presents description of the simulation models and summary of the results from the simulation studies. In Chapter 4, we will apply our proposed methods to analyze real data. The thesis is concluded with discussion in Chapter 5.

Chapter 2

Methodology

2.1 Finite mixture models

Finite mixture models provide probabilistic framework to represent heterogeneity from a finite number of latent classes within an overall population. They have a wide range of applications in many disciplines, and can be applied to data where true group membership is not known or to provide parameter estimates to model multi-modal distributions (see Titterington et al. (1985), McLachlan and Peel (2000), and Frühwirth-Schnatter (2006) for a survey).

Similar to the finite mixture model notation as described in McLachlan and Peel (2000), we let each subpopulation be indexed by $g = 1, ..., G$. Let π_g denote the mixing proportion which can represent the prevalence of each underlying group g, where $\pi_g > 0$ and $\sum_{g=1}^{G} \pi_g = 1$. Let $f_g(\mathbf{y}_i|\theta_g)$ represent the distribution function of $\mathbf{y}_i$ with group g specific parameters θ_g. The finite mixture model for response $\mathbf{y}_i$ given is then given by:

$$f(\mathbf{y}_i; \theta) = \sum_{g=1}^{G} \pi_g f_g(\mathbf{y}_i|\theta_g). \tag{2.1}$$

The observed data is regarded as being incomplete, and we must introduce a latent indicator variable z_{ig} for the group membership of subject i defined as: $z_{ig} = 1$

if subject i belongs to group g, and 0 otherwise. The complete data likelihood for n samples is given by:

$$L_c(\boldsymbol{\pi}, \boldsymbol{\theta}) = \prod_{i=1}^{n} \sum_{g=1}^{G} \left[\pi_g f_g(\mathbf{y}_i|\theta_g)\right]^{z_{ig}}, \tag{2.2}$$

and the complete data log-likelihood is then:

$$\begin{aligned} \ell_c(\boldsymbol{\pi}, \boldsymbol{\theta}) &= \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \left[\log(\pi_g) + \log f_g(\mathbf{y}_i|\theta_g)\right] \\ &= \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \log(\pi_g) + \sum_{i=1}^{n} \sum_{g=1}^{G} \log f_g(\mathbf{y}_i|\theta_g). \end{aligned} \tag{2.3}$$

Estimation of the parameters in finite mixture models is achieved in a maximum likelihood framework with a generalized expectation-maximization algorithm (discussed further in Section 2.3.1).

2.1.1 Finite mixtures of regression models

Finite mixtures of regression (FMR) models (DeSarbo and Cron, 1988) provide a way to model the relationship between the response $\mathbf{y}_i$ and covariate $\mathbf{x}_i$. The finite mixture model (2.1) can be extended to an FMR model to incorporate the group-specific covariate effects on the response variable. FMR models for response $\mathbf{y}_i$ given the covariate $\mathbf{x}_i$ and their associated coefficients $\boldsymbol{\beta}$ are then given by:

$$f(\mathbf{y}_i; \boldsymbol{\beta}, \mathbf{x}_i) = \sum_{g=1}^{G} \pi_g f(\mathbf{y}_i; \boldsymbol{\beta}_g, \mathbf{x}_i). \tag{2.4}$$

Here π_g is the mixing proportion and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_G)$, where $\boldsymbol{\beta}_g$ is the matrix of regression coefficients of the g^{th} component, indicating the different covariate effects on the group-specific response.

2.1.2 Finite mixtures of regression models with concomitant variables

Covariate data may hold information relevant to the underlying mixing proportions, so the FMR formulation (2.4) can be extended to exploit information from the covariates to construct a more informed model. In this framework, the probability of the latent group membership will be functionally related to the concomitant variables (the covariates) following some distribution (Dayton and Macready, 1988). Finite mixtures of regression models with concomitant variables (FMRC) for response $\mathbf{y}_i$ given $\mathbf{x}_i$ and $\boldsymbol{\beta}$ are then given by:

$$f(\mathbf{y}_i; \boldsymbol{\beta}, \mathbf{x}_i) = \sum_{g=1}^{G} \pi_{g|\mathbf{x}_i} f(\mathbf{y}_i; \boldsymbol{\beta}_g, \mathbf{x}_i), \tag{2.5}$$

where $\pi_{g|\mathbf{x}_i} = P(z_{ig} = 1|\mathbf{x}_i)$, with $\pi_{g|\mathbf{x}_i} > 0$ and $\sum_{g=1}^{G} \pi_{g|\mathbf{x}_i} = 1$ for any observation i. These constraints are satisfied by the multinomial logit model, where the component weights (mixing proportions) will depend on the concomitant coefficients $\mathbf{v}$:

$$\pi_{g|\mathbf{x}_i} = \frac{e^{v_g{}'\mathbf{x}_i}}{\sum_{g'=1}^{G} e^{v_g'{}'\mathbf{x}_i}}. \tag{2.6}$$

For identifiability, the first group is used as a reference group with coefficient set to zero ($v_1 = 0$). FMR models are only concerned with the distribution of $\mathbf{Y}|\mathbf{X}$, whereas the FMRC framework will be able to model both $\mathbf{Y}|\mathbf{X}$ as well as a logistic model of the concomitant variables. Computationally, the coefficients $\mathbf{v}$ are estimated through the *multinom* function from the package *nnet* (Venables and Ripley, 2002), which uses a single layer neural network to fit the multinomial log-linear model.

Wedel (2002) demonstrated the benefit of incorporating covariate effects when attempting to accurately cluster data in a FMRC framework. Many variables

may influence which enterotype a subject will belong to in the gut microbiome setting (e.g. diet (Wu et al., 2011; Hildebrandt et al., 2009), drug uptake (Antunes and Finlay, 2011), host genetics (Turnbaugh et al., 2009), and unknown environmental factors (Adlerberth, 2008)), suggesting important information could be missed if group estimation relies solely on resulting taxa counts and ignores covariate data.

2.2 Dirichlet-multinomial model for count data

2.2.1 Multinomial sampling

Begin with a matrix $\mathbf{Y}$ with entries y_{ij} which represent the abundance of bacterial taxa j observed in sample i, where $i \in \{1, ..., n\}$ and $j \in \{1, ..., d\}$. Each row of $\mathbf{Y}$ then corresponds to the counts of each of the d taxa for sample i, and is denoted $\mathbf{y}_i$. One way to model this count data for each sample will be with the use of a multinomial model. Let p_{ij} denote the probability that an individual microbe from the i^{th} sample belongs to the j^{th} taxon, where the constraint $\sum_{j=1}^{d} p_{ij} = 1$ is imposed for each observation i. Then let $y_{i+} = \sum_{j=1}^{d} y_{ij}$ be the total number of observations across categories for each sample. The multinomial distribution of the counts for a sample $\mathbf{y}_i = \{y_{i1}, ..., y_{id}\}$ parameterized by $\mathbf{p}_i = \{p_{i1}, ..., p_{id}\}$ is given by the probability density function for the multinomial distribution:

$$f_M(\mathbf{y}_i, \mathbf{p}_i) = \frac{y_{i+}}{y_{i1}!...y_{id}!} \prod_{j=1}^{d} p_{ij}^{y_{ij}}. \tag{2.7}$$

Mean and variance for each sample $\mathbf{y}_i$ are then given by:

$$E(\mathbf{y}_i) = (y_{i+})\mathbf{p}_i, \qquad Var(\mathbf{y}_i) = (y_{i+})\mathbf{p}_i(1_d - \mathbf{p}_i),$$

where 1_d represents a vector of length d with the value 1 in each entry.

2.2.2 Dirichlet distribution

Actual variability in microbiome composition data is generally greater than a traditional multinomial model would predict due to the heterogeneity amongst samples. To account for this overdispersion, we assume the probability vector $\mathbf{p}_i$ used in the multinomial model (2.7) is itself a random variable following a Dirichlet distribution. The Dirichlet distribution is the multivariate generalization of the Beta distribution and is parameterized by a vector of positive real numbers $\boldsymbol{\alpha}_i = \{\alpha_{i1}, ..., \alpha_{id}\}$. Dirichlet distributions are commonly used as prior distributions, and is of particular benefit for this application as it is the conjugate prior of the multinomial distribution. Let $\alpha_{i+} = \sum_{j=1}^{d} \alpha_{ij}$, then its probability density function for a vector $\mathbf{p}_i = \{p_{i1}, ..., p_{id}\}$ is given by:

$$f_D(\mathbf{p}_i, \boldsymbol{\alpha}_i) = \frac{\Gamma(\alpha_{i+})}{\prod_{j=1}^{d} \Gamma(\alpha_{ij})} \prod_{j=1}^{d} p_{ij}^{\alpha_{ij}-1}, \tag{2.8}$$

where $\Gamma(\cdot)$ is the gamma function. The mean and variance of variable $\mathbf{p}_i$ are:

$$E(\mathbf{p}_i) = \frac{\boldsymbol{\alpha}_i}{\alpha_{i+}}, \qquad Var(\mathbf{p}_i) = \frac{\boldsymbol{\alpha}_i(\alpha_{i+} - \boldsymbol{\alpha}_i)}{(1 + \alpha_{i+})\alpha_{i+}^2}.$$

The term α_{i+} effectively acts as a precision parameter; as α_{i+} increases, the values will have smaller variance and become more concentrated about the mean, and as α_{i+} decreases, values will exhibit wider variation.

2.2.3 Dirichlet-multinomial distribution

This is a compound probability distribution of the multinomial and Dirichlet distributions defined in (2.7) and (2.8). The Dirichlet-multinomial (DM) distribution (Mosimann, 1962) can be viewed as drawing a vector $\mathbf{p}_i$ from a Dirichlet distribution parameterized by $\boldsymbol{\alpha}_i$, and a vector of samples $\mathbf{y}_i$ is then drawn from a multinomial distribution with probability vector $\mathbf{p}_i$. Using the compound of the Dirichlet prior with the observed multinomial samples and marginalizing over the parameter $\mathbf{p}$ gives the probability density function of a DM distribution:

$$\begin{aligned} f_{DM}(\mathbf{y}_i, \boldsymbol{\alpha}_i) &= \int_{\mathrm{p}} f_M(\mathbf{y}_i, \mathbf{p}_i) \cdot f_D(\mathbf{p}_i, \boldsymbol{\alpha}_i)\, \mathrm{dp} \\ &= \int_{\mathrm{p}} \frac{y_{i+}}{y_{i1}!...y_{id}!} \prod_{j=1}^{d} p_{ij}^{y_{ij}} \cdot \frac{\Gamma(\alpha_{i+})}{\prod_{j=1}^{d} \Gamma(\alpha_{ij})} \prod_{j=1}^{d} p_{ij}^{\alpha_{ij}-1}\, \mathrm{dp} \\ &= \frac{y_{i+}}{y_{i1}!...y_{id}!} \frac{\Gamma(\alpha_{i+})}{\prod_{j=1}^{d} \Gamma(\alpha_{ij})} \cdot \int_{\mathrm{p}} \prod_{j=1}^{d} p_{ij}^{y_{ij}+\alpha_{ij}-1}\, \mathrm{dp} \\ &= \frac{y_{i+}}{y_{i1}!...y_{id}!} \frac{\Gamma(\alpha_{i+})}{\prod_{j=1}^{d} \Gamma(\alpha_{ij})} \cdot \frac{\prod_{j=1}^{d} \Gamma(y_{ij} + \alpha_{ij})}{\Gamma(\sum_{j=1}^{d} y_{ij} + \alpha_{ij})}. \end{aligned} \tag{2.9}$$

The DM distribution is also known as the Dirichlet compound multinomial distribution or multivariate Pólya distribution. The mean and variance for each sample $\mathbf{y}_i$ are given by:

$$E(\mathbf{y}_i) = (y_{i+})E(\mathbf{p}_i), \qquad Var(\mathbf{y}_i) = (y_{i+})E(\mathbf{p}_i)(1 - E(\mathbf{p}_i))\frac{y_{i+} + \alpha_{i+}}{1 + \alpha_{i+}}.$$

Haldane (1941) demonstrated that the DM distribution can be represented without the use of gamma functions:

$$f_{DM}(\mathbf{y}_i, \boldsymbol{\alpha}_i) = \frac{y_{i+}}{y_{i1}!...y_{id}!} \cdot \frac{\prod_{j=1}^{d} \alpha_{ij}(\alpha_{ij}+1)\cdots(\alpha_{ij}+y_{ij}-1)}{\alpha_{i+}(\alpha_{i+}+1)\cdots(\alpha_{i+}+y_{i+}-1)}, \tag{2.10}$$

where rising polynomial terms in place of gamma functions leads to increased tractability.

2.2.4 DM regression

The DM distribution is advantageous for modelling taxa proportions of a microbiome sample compared to a classic multinomial model as it is able to model over dispersion. But if one is interested in modelling the relationship of microbiome composition with respect to certain environmental/biological covariates, then a DM regression model can be employed. Let $\mathbf{X}$ be the design matrix for $(p+1)$ covariates $(k \in \{0, ..., p\})$ for n samples, where $\mathbf{X}$ is augmented with a n-vector consisting entirely of 1's as the first column, x_{ik} then represents the value of the k^{th} covariate for sample i. Chen and Li (2013) extended the DM distribution to incorporate covariates through a link function between the distribution parameter α_{ij} and covariates via the following log-linear model:

$$\alpha_{ij}(\mathbf{x}_i) = e^{\mathbf{x}_i^T \boldsymbol{\beta}_j}, \tag{2.11}$$

where $\mathbf{x}_i$ is the i^{th} row vector of $\mathbf{X}$ and is $(p+1)$-dimensional, and $\boldsymbol{\beta}_j$ is the $(p+1)$-dimensional vector of coefficients for taxa j. Let $\boldsymbol{\beta} = (\beta_{kj})_{(p+1)\times d}$ represent the regression coefficient matrix, and β_{kj} is then the coefficient for the j^{th} taxa with respect to the k^{th} covariate; effectively measuring the sign and magnitude of the effect

of the the k^{th} covariate on the j^{th} taxon. The intercept entries β_{0j} can then be viewed as baseline abundancies for taxon j. Note that the log-linear link was chosen for computational simplicity, but it also reflects biological plausibility as microorganisms display exponential growth in favourable environments. With the use of this link function, the DM distribution for a sample $\mathbf{y}_i$ given coefficients $\boldsymbol{\beta}$ and covariates $\mathbf{x}_i$ is then (using the gamma formulation from (2.9)):

$$f_{DM}(\mathbf{y}_i;\boldsymbol{\beta},\mathbf{x}_i) = \frac{y_{i+}}{y_{i1}!...y_{id}!}\frac{\Gamma(\sum_{j=1}^{d} e^{\mathbf{x}_i^T\boldsymbol{\beta}_j})}{\prod_{j=1}^{d}\Gamma(e^{\mathbf{x}_i^T\boldsymbol{\beta}_j})}\frac{\prod_{j=1}^{d}\Gamma(y_{ij}+e^{\mathbf{x}_i^T\boldsymbol{\beta}_j})}{\Gamma(\sum_{j=1}^{d}(y_{ij}+e^{\mathbf{x}_i^T\boldsymbol{\beta}_j}))}. \tag{2.12}$$

The log-likelihood function for coefficient matrix $\boldsymbol{\beta}$ will take the form:

$$\begin{aligned}\ell(\boldsymbol{\beta};\mathbf{Y},\mathbf{X}) = \sum_{i=1}^{n}\Bigg[&\log\Gamma\left(\sum_{j=1}^{d} e^{\mathbf{x}_i^T\boldsymbol{\beta}_j}\right) - \log\left(\prod_{j=1}^{d}\Gamma(e^{\mathbf{x}_i^T\boldsymbol{\beta}_j})\right)\\ &+\log\left(\prod_{j=1}^{d}\Gamma(y_{ij}+e^{\mathbf{x}_i^T\boldsymbol{\beta}_j})\right) - \log\left(\Gamma\left(\sum_{j=1}^{d} y_{ij} + e^{\mathbf{x}_i^T\boldsymbol{\beta}_j}\right)\right)\Bigg].\end{aligned} \tag{2.13}$$

Replacing gamma functions with the rising polynomials of Haldane (1941), and using the simplification noted in Zhang (2014) representing the increasing polynomial terms through a summation over indicator variable l will simplify (2.13) to:

$$\begin{aligned}\ell(\boldsymbol{\beta};\mathbf{Y},\mathbf{X}) = \sum_{i=1}^{n}\sum_{j=1}^{d}\sum_{l=0}^{y_{ij}-1}\log(e^{\mathbf{x}_i^T\boldsymbol{\beta}_j}+l) - \sum_{i=1}^{n}\sum_{l=0}^{y_{i+}-1}\log\left(\sum_{j=1}^{d} e^{\mathbf{x}_i^T\boldsymbol{\beta}_j}+l\right)\\ + \sum_{i=1}^{n}\log\left(\frac{y_{i+}}{y_{i1}!...y_{id}!}\right).\end{aligned} \tag{2.14}$$

As the difference of two concave terms plus a constant, (2.14) is not necessarily concave and is challenging to optimize.

2.3 Finite mixtures of DM regression models

We extend equation (2.12) to finite mixtures of DM regression models. Finite mixtures of DM regression models for response $\mathbf{y}_i$ given $\mathbf{x}_i$ and $\boldsymbol{\beta}$ is given by:

$$f(\mathbf{y}_i; \boldsymbol{\beta}, \mathbf{x}_i) = \sum_{g=1}^{G} \pi_g f_{DM}(\mathbf{y}_i; \boldsymbol{\beta}_g, \mathbf{x}_i), \tag{2.15}$$

where $\boldsymbol{\beta_g}$ denotes the coefficient matrix for each group g:

$$\boldsymbol{\beta}_g = \begin{bmatrix} \beta_{01} & \beta_{02} & \dots & \beta_{0d} \\ \beta_{11} & \beta_{12} & \dots & \beta_{1d} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{p1} & \beta_{p2} & \dots & \beta_{pd} \end{bmatrix}.$$

.

The likelihood for finite mixtures of DM regression models can be written as:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \sum_{g=1}^{G} \pi_g f_{DM}(\mathbf{y}_i; \boldsymbol{\beta}_g, \mathbf{x}_i), \tag{2.16}$$

and given group assignments z_{ig} for the membership of subject i, the complete data likelihood for n samples is given by:

$$L_c(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{g=1}^{G} \left[\pi_g f_{DM}(\mathbf{y}_i; \boldsymbol{\beta}_g, \mathbf{x}_i)\right]^{z_{ig}}, \tag{2.17}$$

and the complete data log-likelihood is:

$$\ell_c(\boldsymbol{\beta}) = \sum_{g=1}^{G} \sum_{i=1}^{n} z_{ig} \log f_{DM}(\mathbf{y}_i; \boldsymbol{\beta}_g, \mathbf{x}_i) + \sum_{g=1}^{G} \sum_{i=1}^{n} z_{ig} \log \pi_g. \tag{2.18}$$

The extension to FMRC in the DM regression framework replaces π_g in (2.15) with $\pi_{g|\mathbf{x}_i}$ as defined in (2.6).

2.3.1 Generalized expectation-maximization algorithm

A generalized expectation-maximization (GEM) algorithm is used to find maximum likelihood estimates the coefficient matrices $\boldsymbol{\beta}_g$. This technique alternates between an expectation step (E-step) to calculate the expectation of the complete data log-likelihood (2.18) given the current parameter estimates, and a maximization step (M-step) to solve for parameter values which optimize the expected complete data likelihood function found in the E-step. Parameter estimates found in the previous iteration are used to determine the distribution for the latent variables z_{ig} used in the next E-step.

Applying laws of conditional probability, we can solve for the expected value of z_{ig} given values of $\mathbf{x}_i$ and $\mathbf{y}_i$ with:

$$\begin{aligned}
\mathrm{E}(z_{ig}|\mathbf{y}_i, \mathbf{x}_i) &= 1 \cdot P(z_{ig} = 1|\mathbf{y}_i, \mathbf{x}_i) + 0 \cdot P(z_{ig} = 0|\mathbf{y}_i, \mathbf{x}_i) \\
&= P(z_{ig} = 1|\mathbf{y}_i, \mathbf{x}_i) \\
&= \frac{P(z_{ig} = 1)P(\mathbf{y}_i|\mathbf{x}_i, z_{ig} = 1)}{P(\mathbf{y}_i|\mathbf{x}_i)} \\
&= \frac{P(z_{ig} = 1)P(\mathbf{y}_i|\mathbf{x}_i, z_{ig} = 1)}{\sum_{g=1}^{G} P(z_{ig} = 1)P(\mathbf{y}_i|\mathbf{x}_i, z_{ig} = 1)} \\
&= \frac{\pi_g P(\mathbf{y}_i|\mathbf{x}_i, z_{ig} = 1)}{\sum_{g=1}^{G} \pi_g P(\mathbf{y}_i|\mathbf{x}_i, z_{ig} = 1)}. \qquad (2.19)
\end{aligned}$$

The values of $\mathrm{E}(\pi_g)$ are updated by averaging group membership over n samples:

$$\mathrm{E}(\pi_g) = \frac{\sum_{i=1}^{n} \mathrm{E}(z_{ig}|\mathbf{x}_i, \mathbf{y}_i)}{n}. \qquad (2.20)$$

The matrix $\mathbf{Z}$ consisting of elements z_{ig} can be initialized through a variety of techniques. In the data analysis section of this thesis, initialization with the use of

a PAM algorithm via the *cluster* package (Maechler et al., 2015) in R on the response matrix $\mathbf{Y}$ is used to obtain preliminary estimates of the group assignments prior to running the GEM algorithm. The coefficient matrices $\boldsymbol{\beta}_g$ are initialized such that each entry β_{gkj} is set to 0; this was chosen as an appropriate starting point as it leads to a weak uniform assumption on the α_{ij} terms (i.e. $e^0 = 1 = \alpha_{ij}$).

In the M-step, values of $\boldsymbol{\beta}_g$ which increase the complete data log-likelihood (2.18) are found with three different algorithms whose performances are examined and compared: simulated annealing, Newton-Raphson, and the minorization-maximization technique. Note that this is a GEM algorithm and not a true EM algorithm because these optimization strategies only guarantee an increase in the log-likelihood at each iteration but not necessarily maximization; this discrepancy arises due to the non-concavity of the log-likelihood (2.18), leading to difficulties in true maximization at each parameter update

The GEM algorithm is halted when it reaches some convergence criteria; in this thesis, the convergence criteria was chosen to be when the difference in log likelihood between consecutive iterations is less than 10^{-4}.

2.3.2 Deterministic annealing

As previously mentioned, in order to place a weak uniform prior assumption on the α_{ij} terms, the coefficient matrices $\boldsymbol{\beta}_g$ are initialized such that each entry β_{gkj} is set to 0, but in some situations this initialization may be far from the truth and the deterministic annealing approach of Ueda and Nakano (1998) is used to recover from a poor starting value. Deterministic annealing modifies the GEM framework such that

the E-step includes averaging the complete data log-likelihood ℓ_c over the distribution proportional to the current estimate of the conditional density of the complete data raised to an exponent τ. This means the current estimate of the posterior probability of the i^{th} observation belonging to the g^{th} group will become:

$$\mathrm{E}(z_{ig}|\mathbf{x}_i, \mathbf{y}_i) = \left[\frac{\pi_g P(\mathbf{y}_i|\mathbf{x}_i, z_{ig} = 1)}{\sum_{g=1}^{G} \pi_g P(\mathbf{y}_i|\mathbf{x}_i, z_{ig} = 1)} \right]^{\tau}. \tag{2.21}$$

Ueda and Nakano (1998) recommend beginning with a value of τ close to 0 ($0 \leq \tau \leq 1$), then increasing τ in each iteration of the GEM algorithm with $\tau^{(t+1)} = c\tau^{(t)}$, where c is on the interval $1 \leq c \leq 1.5$ until $\tau^{(t+1)} = 1$ (for our purposes, setting $c = 1.10$ and $\tau^{(0)} = 0.1$ proved to be effective). The deterministic annealing technique effectively protects against the negative effects of a poor starting value by letting the estimates of component densities to overlap considerably in the first few iterations (see McLachlan and Peel (2000) for further discussion).

2.4 Optimization strategies

2.4.1 Simulated annealing

Simulated annealing (Kirkpatrick et al., 1983) is a probabilistic metaheuristic approach to find parameters $\boldsymbol{\beta}_g$ which optimize the complete data log-likelihood function (2.18). It is an adapted version of the Metropolis-Hastings algorithm which explores neighbouring parameter values from the current parameter estimates in order to move the system to a state of lower energy. The key advantage of simulated annealing is that it allows for the system to escape local optima through hill climbing

moves which worsen the objective function in the search for a global optimum. For this application, it is implemented via the *optim* function in the base *stats* package in R. Note that this implementation uses a variant given by Bélisle (1992) which performs well on rough surfaces. Simulated annealing tends to perform slowly in practice, but was chosen as it displays stability and is a very robust algorithm (Goffe et al., 1994).

2.4.2 Newton-Raphson

Analytic solutions for the gradient and hessian matrix of the complete data log-likelihood are derived and a Newton-Raphson (NR) algorithm is also employed as an optimization strategy (see Appendix A for full derivation).

Given the group assignment matrix $\mathbf{Z}$ and using the link $\alpha_{gij}(\mathbf{x}_i) = e^{\mathbf{x}_i^T \boldsymbol{\beta}_{gj}}$, the score function for each group g with respect to taxa j and covariate k is given by:

$$S(\beta_{gkj}) = \sum_{i=1}^{n} z_{ig}\alpha_{gij}\left(\sum_{l=0}^{y_{ij}-1} c_{ij}\frac{l}{l+\alpha_{gij}} - \sum_{l=1}^{y_{i+}-1}\frac{1}{l+\alpha_{gi+}}\right)x_{ik}, \tag{2.22}$$

where $c_{ij} = 1$ if $y_{ij} \geq 1$, otherwise $c_{ij} = 0$ if $y_{ij} = 0$. $S(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_G)$ is populated with entries $S(\beta_{gkj})$ at each iteration of the GEM algorithm.

The entries for the information matrix for each group g, $I_g(\boldsymbol{\beta})$, composed of entries $I_g(\boldsymbol{\beta})_{kj,k'j'} = \frac{-\partial^2 \log L_c}{\partial\beta_{gkj}\partial\beta_{gk'j'}}$ is given by:

$$I_g(\boldsymbol{\beta})_{kj,k'j'} = \begin{cases} -\sum_{i=1}^{n} z_{ig}\ \alpha_{gij}\ x_{ik}\ x_{ik'}[c_{ij}\sum_{l=0}^{y_{ij}-1}\frac{l}{(l+\alpha_{gij})^2} - \sum_{l=1}^{y_{i+}-1}(\frac{\alpha_{gi+}-\alpha_{gid}+l}{(\alpha_{gi+}+l)^2})] & \text{for } j = j' \\ -\sum_{i=1}^{n} z_{ig}\ \alpha_{gid}\ \alpha_{gid'}\ x_{ik}\ x_{ik'}(\sum_{l=1}^{y_{i+}-1}\frac{l}{(\alpha_{gi+}+l)^2}) & \text{for } j \neq j'. \end{cases}$$

(2.21)

Updates at each iteration are of the form:

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} + S\left(\hat{\boldsymbol{\beta}}^{(t)}\right)\left[I\left(\hat{\boldsymbol{\beta}}^{(t)}\right)\right]^{-1}, \tag{2.24}$$

and note that only one NR update is computed to increase the log-likelihood at each iteration in the GEM algorithm. Lange (1995) demonstrated that the use of one NR update at each GEM iteration demonstrates similar convergence properties to a traditional EM algorithm but with increased speed.

NR can be unstable due to the non-concavity of the complete data log-likelihood, which can become especially problematic in a regression framework when there is no good starting point available. There is also high computational cost associated with calculating the information matrix at each iteration, especially when dealing with high dimensionality. To deal with this issue, we follow the approach of Chen and Li (2013) and only the diagonal entries of the Hessian are calculated for ease of computation, as we are assuming the β_{gjk} terms are independent. However, unlike the NR derivation of Chen and Li (2013), this derivation is based on rising polynomials and not gamma functions, meaning the calculations are in finite terms and will not be approximations.

2.4.3 Minorization-maximization algorithm

The minorization-maximization (MM) algorithm of Hunter and Lange (2000) is a two-step procedure which works by first constructing a surrogate function $g(\theta)$

that "minorizes" the original objective function $f(\theta)$; i.e. $g(\theta)$ must satisfy:

$$f(\theta) \geq g(\theta|\theta^{(t)}) \quad \forall\theta,$$

$$f(\theta^{(t)}) = g(\theta^{(t)}|\theta^{(t)}).$$

In other words, $g(\theta|\theta^{(t)})$ lies below $f(\theta)$ and is tangent at the point $\theta = \theta^{(t)}$. Note that if $-g(\theta|\theta^{(t)})$ minorizes $-f(\theta)$ at $\theta = \theta^{(t)}$ then $g(\theta)$ is said to majorize $f(\theta)$ (and one will then be working with a majorization-minimization algorithm).

Examples of commonly used minorizing functions include the use of supporting hyperplanes and Taylor series approximations as seen in Figures 2.1 and 2.2 (see Lange (1999) for further discussion).

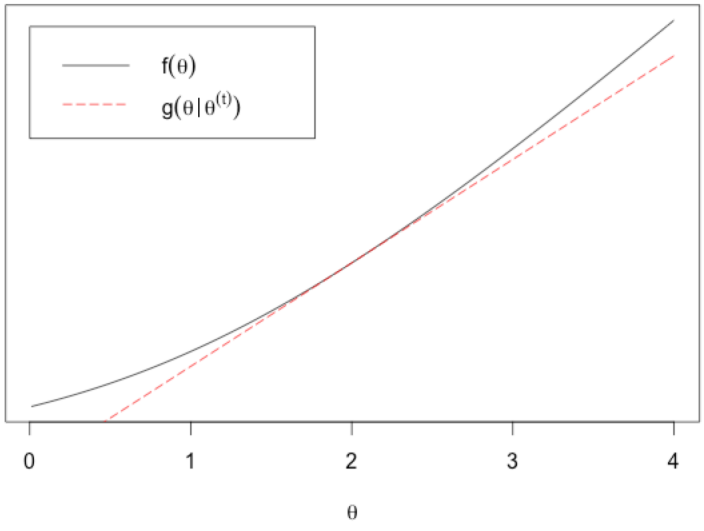


Figure 2.1: Supporting hyperplane $g(.)$ minorizes an objective function $f(.)$, $\theta^{(t)} = 2$.

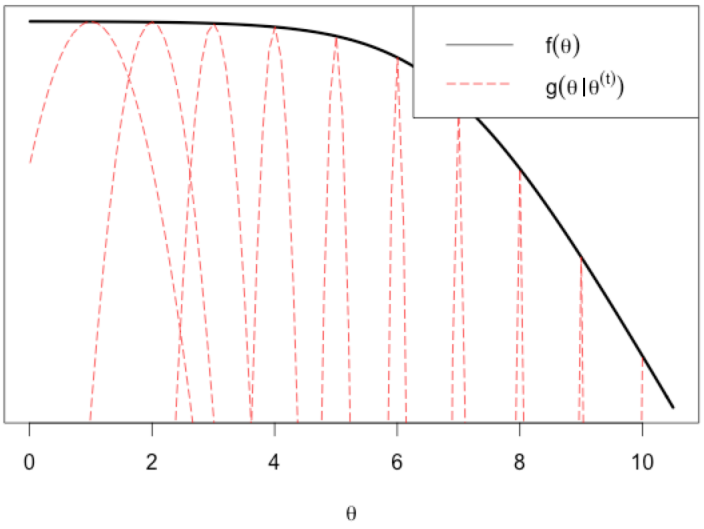


Figure 2.2: Second degree taylor series approximations $g(.)$ minorizes an objective function $f(.)$, $\theta^{(t)} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$.

The next step in the MM procedure is to find the value of θ which maximizes $g(\theta)$:

$$\theta^{(t+1)} = \text{argmax } g(\theta|\theta^{(t)}).$$

This algorithm will guarantee ascension as:

$$f(\theta^{(t+1)}) \geq g(\theta^{(t+1)}|\theta^{(t)}) \geq g(\theta^{(t)}) = f(\theta^{(t)}).$$

This ascent property provides numerical stability to the MM algorithm, and has previously been successfully applied to DM distribution fitting (Zhou and Lange, 2010). Advantages of this procedure come from careful choices of the surrogate function(s) $g(\theta)$; by using surrogate functions that are simpler to work with than the objective function $f(\theta)$, one can greatly reduce the amount of computational power and time necessary. Furthermore, it is also possible to reduce a high dimensional

problem into several low dimensional ones and compute iterations in parallel if possible.

In order to minorize the complete data log-likelihood function (2.18), the surrogate functions used by Zhang (2014) for DM regression are extended to incorporate the group assignment variables z_{ig}. First, note that (2.18) can be re-written as:

$$\begin{aligned}\ell_c = &\sum_{g=1}^{G}\sum_{i=1}^{n} z_{ig}\log\pi_g + \sum_{g=1}^{G}\sum_{i=1}^{n} z_{ig}\sum_{j=1}^{d}\sum_{l=0}^{y_{ij}-1}\log(e^{\mathbf{x}_i^T\boldsymbol{\beta}_j}+l)\\ &-\sum_{g=1}^{G}\sum_{i=1}^{n} z_{ig}\sum_{l=0}^{y_{i+}-1}\log\left(\sum_{j=1}^{d}e^{\mathbf{x}_i^T\boldsymbol{\beta}_j}+l\right)+\sum_{g=1}^{G}\sum_{i=1}^{n} z_{ig}\log\left(\frac{y_{i+}}{y_{i1}!\ldots y_{id}!}\right). \qquad (2.25)\end{aligned}$$

With respect to $\boldsymbol{\beta}$, the first and last terms are constant, so the goal is to maximize the second and third term. Jensen's inequality is applied to the second term of ℓ_c:

$$\begin{aligned}\log\sum_i c_i x_i &\geq \sum_i \frac{c_i x_i^{(t)}}{\sum_i c_i x_i^{(t)}}\log\left(\frac{\sum_i c_i x_i^{(t)}}{c_i x_i^{(t)}}c_i x_i\right)\\ &= \sum_i \frac{c_i x_i^{(t)}}{\sum_i c_i x_i^{(t)}}\log(x_i) + C^{(t)}, \quad c_i, x_i > 0,\end{aligned}$$

where $C^{(t)}$ collects the constants irrelevant to optimization, and the supporting hyperplane property is applied to the third term of ℓ_c:

$$-\log(x) \geq -\log(x)^{(t)} - \frac{x - x^{(t)}}{x^{(t)}}.$$

This gives us the surrogate function:

$$\begin{aligned}g(\boldsymbol{\beta}|\boldsymbol{\beta}^{(t)}) = &\sum_{g=1}^{G}\sum_{i=1}^{n} z_{ig}\sum_{j=1}^{d}\sum_{l=0}^{y_{ij}-1}\frac{e^{\mathbf{x}_i^T\boldsymbol{\beta}_{gj}^{(t)}}}{e^{\mathbf{x}_i^T\boldsymbol{\beta}_{gj}^{(t)}}+l}(\mathbf{x}_i^T\boldsymbol{\beta}_{gj})\\ &-\sum_{g=1}^{G}\sum_{i=1}^{n} z_{ig}\sum_{j=1}^{d}\sum_{l=0}^{y_{i+}-1}\frac{e^{\mathbf{x}_i^T\boldsymbol{\beta}_{gj}^{(t)}}}{\sum_{j=1}^{d}(e^{\mathbf{x}_i^T\boldsymbol{\beta}_{gj}^{(t)}}+l)} + C^{(t)}. \qquad (2.26)\end{aligned}$$

Construction of $g(\boldsymbol{\beta}|\boldsymbol{\beta}^{(t)})$ constitutes the first M of the MM algorithm. In order to maximize this surrogate function, the iteratively reweighted Poisson regression (IRPR) procedure of Zhang (2014) is extended to a mixture model framework.

2.4.3.1 Iteratively reweighted Poisson regression

The log-likelihood for a weighted Poisson model is:

$$\ell = \sum_{i=1}^{n} w_i(-\mu_i + y_i \log(\mu_i) - \log(y_i!)), \tag{2.27}$$

where $\mu = e^{(\mathbf{x}^T\boldsymbol{\beta})}$. And (2.27) can be re-written as:

$$\begin{aligned}\ell &= \sum_{i=1}^{n} w_i y_i \mathbf{x}^T\boldsymbol{\beta} - \sum_{i=1}^{n} w_i e^{(\mathbf{x}^T\boldsymbol{\beta})} - \sum_{i=1}^{n} w_i \log(y_i!) \\ &= \sum_{i=1}^{n} w_i y_i \mathbf{x}^T\boldsymbol{\beta} - \sum_{i=1}^{n} w_i e^{(\mathbf{x}^T\boldsymbol{\beta})} + C. \end{aligned} \tag{2.28}$$

The surrogate function (2.26) can be expressed in the form of the weighted Poisson regression likelihood from (2.28):

$$\begin{aligned} g(\boldsymbol{\beta}|\boldsymbol{\beta}^{(t)}) = &\sum_{g=1}^{G}\sum_{i=1}^{n}\sum_{j=1}^{d}\sum_{l=0}^{y_{ij}-1} z_{ig}\frac{e^{\mathbf{x}_i^T\boldsymbol{\beta}_{gj}^{(t)}}}{e^{\mathbf{x}_i^T\boldsymbol{\beta}_{gj}^{(t)}}+l}(\mathbf{x}_i^T\boldsymbol{\beta}_{gj}) \\ &- \sum_{g=1}^{G}\sum_{i=1}^{n}\sum_{j=1}^{d}\sum_{l=0}^{y_{i+}-1} \frac{z_{ig}}{\sum_{j=1}^{d}(e^{\mathbf{x}_i^T\boldsymbol{\beta}_{gj}^{(t)}}+l)} + C^{(t)}. \end{aligned}$$

At each iteration, d x G independent weighted poisson regressions are performed, with working weight:

$$w_{gij}^{(t)} = \sum_{l=0}^{y_{i+}-1} \frac{z_{ig}}{\sum_{j=1}^{d} e^{\mathbf{x}_i^T\boldsymbol{\beta}_{gj}^{(t)}}+l}, \tag{2.29}$$

and response:

$$y_{gij}^{(t)} = \left(\sum_{l=0}^{y_{ij}-1} z_{ig}\frac{e^{\mathbf{x}_i^T\boldsymbol{\beta}_{gj}^{(t)}}}{e^{\mathbf{x}_i^T\boldsymbol{\beta}_{gj}^{(t)}}+l}\right)(w_{gij}^{(t)})^{-1}. \tag{2.30}$$

Computationally, solutions for the updates to the working weights and responses are found with the $glm.fit$ command in the base *stats* package in R. Note that the log-likelihood is increased with the use of only one IRPR iteration within each M-step of the GEM algorithm.

2.5 Model selection criteria and performance assessment

2.5.1 Selecting the number of latent classes

When fitting a mixture model to observed data, the underlying number of groups is often unknown. In order to estimate this latent group structure, a number of models are fit with different values for g, and the values of the Integrated Classification Likelihood (ICL) (Biernacki et al., 1998) are compared. ICL has shown to be more robust than BIC to violations of mixture model assumptions (Biernacki et al., 2000) and is effective in choosing the number of mixture components resulting in a clustering structure with the greatest evidence (McLachlan and Peel, 2000). In particular, an approximation of ICL which relies on the Bayesian Information Criteria (Schwarz et al., 1978) known as ICL-BIC (Biernacki et al., 2000) is used in this thesis. The definitions of BIC and ICL-BIC given by McLachlan and Peel (2000) are used here; specifically,

$$\text{BIC} = -2\ell(\theta) + \psi \log(n), \tag{2.31}$$

where $\ell(\theta)$ is the log-likelihood defined in 2.18, n is the total number of observations, and ψ is the number of free parameters in the model. ICL-BIC uses the BIC to approximate ICL by penalizing BIC with an extra term $\text{E}(\mathbf{z})$, which is the estimated

mean entropy of the posterior probabilities of the clusters given the observed data:

$$\mathrm{E}(z_{ig}|\mathbf{y}_i, \mathbf{x}_i) = \frac{\pi_g P(\mathbf{y}_i|\mathbf{x}_i, z_{ig}=1)}{\sum_{g=1}^{G} \pi_g P(\mathbf{y}_i|\mathbf{x}_i, z_{ig}=1)},$$

$$\mathrm{E}(\mathbf{z}) = -\sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig} \log z_{ig}. \tag{2.32}$$

The term $\mathrm{E}(\mathbf{z})$ will effectively penalize clusters with poor separation. ICL-BIC is computed as follows:

$$\text{ICL-BIC} = -2\ell(\theta) + \psi \log(n) + 2\mathrm{E}(\mathbf{z}). \tag{2.33}$$

Note for the rest of this thesis we will denote ICL-BIC as ICL, as ICL-BIC is an approximation for the true ICL.

2.5.2 Adjusted Rand index

The adjusted Rand index (ARI) of Hubert and Arabie (1985) is used to measure the similarity between estimated cluster assignments and the true cluster assignments from simulations in Section 3. An ARI value of 1 indicates perfect agreement, 0 indicates random labeling, and negative values indicate less agreement than expected by random labeling. ARI is defined as:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - (\sum_i \binom{n_{i+}}{2}\sum_j \binom{n_{+j}}{2})/\binom{n}{2}}{\frac{1}{2}(\sum_i \binom{n_{i+}}{2}\sum_j \binom{n_{+j}}{2}) - (\sum_i \binom{n_{i+}}{2}\sum_j \binom{n_{+j}}{2})/\binom{n}{2}}, \tag{2.34}$$

where A and B are two different clusterings of the same set of n observations, and elements n_{ij}, n_{i+}, and n_{+j}, are defined in in terms of a contingency table illustrated with Table 2.1. ARI is calculated with the use of the *mclust* (Fraley et al., 2012) package.

$A \setminus B$	B_1	B_2	$\cdots$	B_G	Sums
A_1	n_{11}	n_{12}	$\cdots$	n_{1G}	n_{1+}
A_2	n_{21}	n_{22}	$\cdots$	n_{2G}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_G	n_{G1}	n_{G2}	$\cdots$	n_{GG}	n_{G+}
Sums	n_{+1}	n_{+2}	$\cdots$	n_{+G}	n

Table 2.1: Notation for contingency table for comparing two clustering assignments.

Chapter 3

Simulation studies

3.1 Simulation design

Data sets are simulated to mimic n microbiome samples, where each sample contains information for p covariates, counts of d bacterial taxa, and can be assigned to one of G groups. Three varieties of data sets are simulated to represent three distinct scenarios.

- Scenario 1: only the distribution of $\mathbf{Y}|\mathbf{X}$ influences the group structure.
- Scenario 2: the covariates exhibit a concomitant influence on the group structure through a logistic model.
- Scenario 3: both the distribution of $\mathbf{Y}|\mathbf{X}$ and $\mathbf{X}$ influences the group structure.

3.1.1 Scenario 1

The response matrix $\mathbf{Y}$ consists of a $n = 1000$ microbiome samples simulated with counts from $d = 3$ bacterial taxa. Only 3 simulated taxa were used as it has been shown that enterotypes can be adequately characterized by the levels of the genera Bacteroides, Prevotella, and Ruminococcus (Arumugam et al., 2011). The covariate

matrix $\mathbf{X}$ is generated from a multivariate Gaussian distribution, $\mathbf{X} \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma}_{i,j} = 0.1^{|i-j|}$. $\mathbf{X}$ is augmented with an intercept vector consisting entirely of 1's, so $\mathbf{X}$ is of dimension $n \times (p+1)$. The underlying number of groups is chosen to be $G = 2$, with mixing proportions set to be $\boldsymbol{\pi} = (0.41, 0.59)$.

The entries for the coefficient matrices for each group $\boldsymbol{\beta}_g$ are drawn from a uniform distribution on the interval $[-2g, 2g]$. The DM distribution parameter for each sample $\boldsymbol{\alpha}_{gi}$ is found with $\boldsymbol{\alpha}_{gi}(\mathbf{x}_i) = \sum_{g=1}^{G} e^{\mathbf{x}_i^T \boldsymbol{\beta}_g} z_{ig}$, and the entries $\mathbf{y}_i$ from response matrix $\mathbf{Y}$ are drawn from a DM distribution parameterized with $\boldsymbol{\alpha}_{gi}$ and $\sum_{j=1}^{d} Y_{ij}$ follows a Gaussian distribution with mean 100 and variance 80. Random draws from a DM distribution are accomplished through the *rdirm* function from the R package MGLM (Zhang and Zhou, 2013).

100 response matrices $\mathbf{Y}$ are drawn randomly with fixed values for $\boldsymbol{\beta}$ and $\mathbf{X}$, and group assignments drawn from a multinomial distribution parameterized by $\boldsymbol{\pi} = (0.41, 0.59)$. This design represents a traditional FMR simulation, where the covariates have no influence on the underlying group structure. The simulated groups display strong separation and with only minor overlap, and as group separation in the FMR framework depends only on the coefficient matrices $\boldsymbol{\beta}_g$, this separation can be visualized through 3-dimensional scatterplots of the linear predictor $\mathbf{X}\boldsymbol{\beta}$ as seen in Figure 3.1 (generated from the R package *rgl* (Adler et al., 2015)).

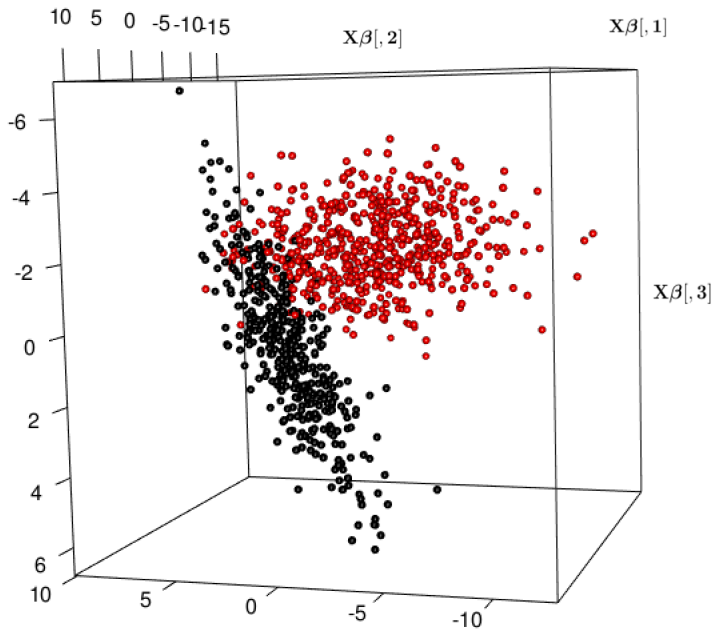


Figure 3.1: Scatterplot of the linear predictor $\mathbf{X}\boldsymbol{\beta}$ for Scenario 1. Data points from group 1 shown in black, and data points from group 2 are shown in red.

3.1.2 Scenario 2

The response matrix $\mathbf{Y}$ again consists of a $n = 1000$ microbiome samples simulated with counts from $d = 3$ bacterial taxa. The covariate matrix $\mathbf{X}$ is generated following the same design as Scenario 1. The underlying number of groups is chosen to be $G = 3$, with mixing proportions set to be $\boldsymbol{\pi} = (0.24, 0.28, 0.48)$.

The entries for the coefficient matrices for each group $\boldsymbol{\beta}_g$ are drawn from a uniform distribution on the interval $[-2g, 2g]$, and the entries for the coefficient matrix $\mathbf{v}$ which dictates the concomitant effect of the covariates on the mixing proportions $\pi_{g|\mathbf{x}_i}$ from (2.6) is generated from a uniform distribution on the interval

$[0, 4]$. The DM distribution parameter for each sample $\boldsymbol{\alpha}_{gi}$ is again found with $\boldsymbol{\alpha}_{gi}(\mathbf{x}_i) = \sum_{g=1}^{G} e^{\mathbf{x}_i^T \boldsymbol{\beta}_g} z_{ig}$, and the entries $\mathbf{y}_i$ from response matrix $\mathbf{Y}$ are drawn from a DM distribution parameterized with $\boldsymbol{\alpha}_{gi}$ and $\sum_{j=1}^{d} Y_{ij}$ follows a Gaussian distribution with mean 100 and variance 80.

100 response matrices $\mathbf{Y}$ are drawn randomly with fixed values for $\boldsymbol{\beta}$ and $\mathbf{X}$, and group assignments are generated about $\boldsymbol{\pi} = (0.24, 0.28, 0.48)$ through small random uniform additions to the fixed coefficent matrix $\mathbf{v}$ in each simulation in order to mimic draws from a multinomial distribution. This design represents an FMRC simulation, where the covariates influence the mixing proportions through a multinomial logit model.

3.1.3 Scenario 3

The covariate matrix $\mathbf{X}$ in this design is composed of G different submatrices $\mathbf{X}_g$ which are generated according to group specific means and covariances $MVN(\mu_g, \Sigma_g)$. The response matrix $\mathbf{Y}$ again consists of a $n = 1000$ microbiome samples simulated with counts from $d = 3$ bacterial taxa. The underlying number of groups is chosen to be $G = 4$, with mixing proportions $\boldsymbol{\pi} = (0.15, 0.21, 0.28, 0.36)$.

The entries for the coefficient matrices for each group $\boldsymbol{\beta}_g$ are again drawn from a uniform distribution on the interval $[-2g, 2g]$, and the DM distribution parameter for each sample $\boldsymbol{\alpha}_{gi}$ is found with $\boldsymbol{\alpha}_{gi}(\mathbf{x}_i) = \sum_{g=1}^{G} e^{\mathbf{x}_i^T \boldsymbol{\beta}_g} z_{ig}$, and the entries $\mathbf{y}_i$ from response matrix $\mathbf{Y}$ are drawn from a Dirichlet-multinomial distribution parameterized with $\boldsymbol{\alpha}_{gi}$ and $\sum_{j=1}^{d} Y_{ij}$ follows a Gaussian distribution with mean 100 and variance 80.

100 response matrices $\mathbf{Y}$ are drawn randomly with fixed values for $\boldsymbol{\beta}$ and $\mathbf{X}$, and group assignments are drawn from a multinomial distribution parameterized by $\boldsymbol{\pi} = (0.15, 0.21, 0.28, 0.36)$.

3.2 Simulation results

In each scenario, the performance of GEM algorithms using the FMR and FMRC approaches are compared. Furthermore, the performances of the different optimization strategies: Newton-Rhapson (NR), minorization-maximization (MM), and simulated annealing (SANN) are compared within the FMR and FMRC frameworks. Each dataset/algorithm combination is evaluated on its ability to identify the true number of underlying groups, the true mixing proportions $\boldsymbol{\pi}$, and the proximity to the true coefficient matrices $\boldsymbol{\beta}_g$ for each group.

3.2.1 Scenario 1 results

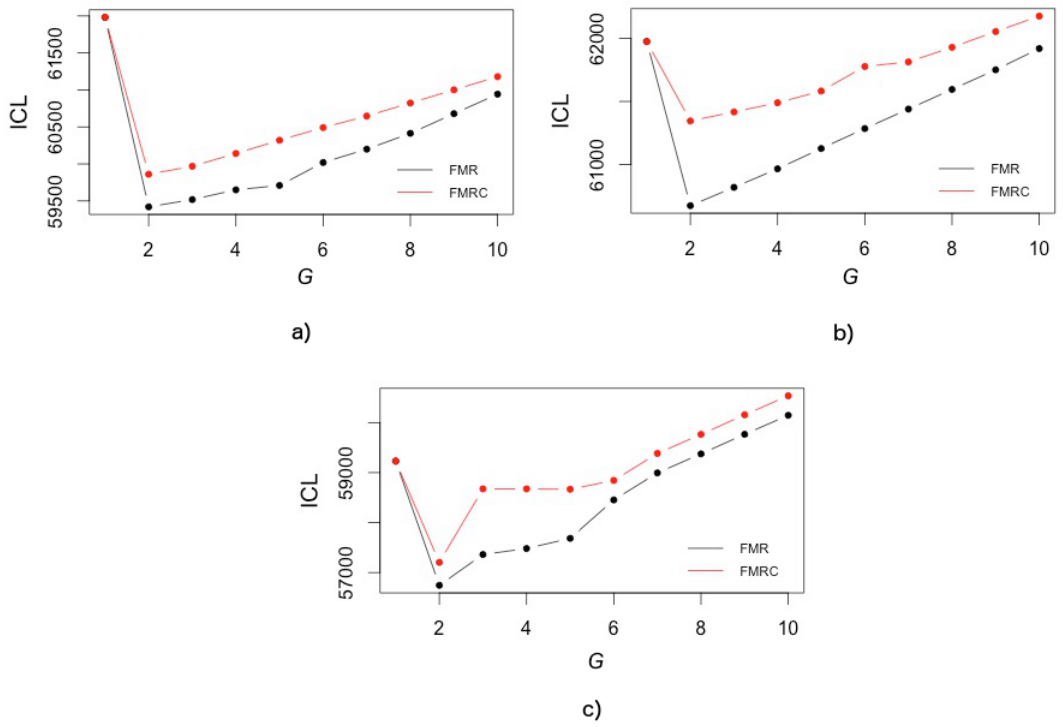


Figure 3.2: ICL values for models fit with number of components G ranging from 1 to 10 on one of the hundred simulated data sets under Scenario 1. Output from NR, MM, and SANN optimization is represented in a), b), and c), respectively.

ICL values shown in Figure 3.2 with each maximization technique within FMR and FMRC frameworks are lowest at $G = 2$, meaning the best fitting model uses two underlying groups. The selection of $G = 2$ by each algorithm accurately matches the design of Scenario 1. Although Figure 3.2 only shows the output from one simulation, all algorithms consistently displayed similar patterns on each of the 100 simulated data sets. Also note that FMR always had lower ICL values than

FMRC, implying the extra coefficients involved to model concomitant variables in FMRC are unnecessary as they worsen the ICL. This is expected as data sets were generated with no concomitant influence on the mixing proportions in Scenario 1.

Group 1													
	True $\boldsymbol{\beta}_1$	β_{101}	β_{102}	β_{103}	β_{111}	β_{112}	β_{113}	β_{121}	β_{122}	β_{123}	β_{131}	β_{132}	β_{133}
		1.96	1.79	0.84	1.64	-1.53	2.00	1.99	0.30	-0.87	0.12	-1.47	0.52
	$\hat{\boldsymbol{\beta}}_1$	$\hat{\beta}_{101}$	$\hat{\beta}_{102}$	$\hat{\beta}_{103}$	$\hat{\beta}_{111}$	$\hat{\beta}_{112}$	$\hat{\beta}_{113}$	$\hat{\beta}_{121}$	$\hat{\beta}_{122}$	$\hat{\beta}_{123}$	$\hat{\beta}_{131}$	$\hat{\beta}_{132}$	$\hat{\beta}_{133}$
FMR	**NR**	1.91(0.11)	1.73(0.10)	0.79(0.10)	1.56(0.17)	-1.60(0.16)	1.94(0.16)	1.88(0.12)	0.20(0.10)	-0.97(0.11)	0.11(0.11)	-1.49(0.13)	0.51(0.13)
	MSE	0.0146	0.0136	0.0125	0.0353	0.0305	0.0292	0.0265	0.0200	**0.0221**	0.0122	0.0173	0.0170
	MM	1.90(0.05)	1.82(0.08)	0.84(0.08)	1.74(0.06)	-1.36(0.04)	2.01(0.04)	1.77(0.06)	0.08(0.04)	-1.13(0.05)	0.23(0.08)	-1.31(0.05)	0.65(0.05)
	MSE	**0.0061**	**0.0073**	**0.0064**	**0.0136**	0.0305	**0.0017**	0.0520	0.0500	0.0701	0.0185	0.0281	0.0194
	SANN	1.91(0.13)	1.73(0.13)	0.75(0.09)	1.61(0.34)	-1.55(0.36)	2.00(0.32)	1.96(0.10)	0.27(0.12)	-0.93(0.16)	0.05(0.26)	-1.52(0.20)	0.47(0.27)
	MSE	0.0194	0.0205	0.0162	0.1165	0.1300	0.1024	**0.0109**	**0.0153**	0.0292	0.0725	0.0425	0.0754
FMRC	**NR**	1.80(0.28)	1.63(0.26)	0.67(0.27)	1.51(0.17)	-1.63(0.14)	1.87(0.16)	1.82(0.14)	0.17(0.14)	-1.02(0.16)	0.15(0.09)	-1.42(0.10)	0.56(0.09)
	MSE	0.1040	0.0932	0.1018	0.0458	**0.0296**	0.0425	0.0485	0.0365	0.0481	**0.0090**	**0.0125**	**0.0097**
	MM	1.89(0.10)	1.85(0.09)	0.81(0.08)	1.73(0.09)	-1.37(0.11)	2.00(0.09)	1.77(0.09)	0.08(0.08)	-1.13(0.07)	0.23(0.08)	-1.32(0.06)	0.63(0.06)
	MSE	0.0149	0.0117	0.0073	0.0162	0.0377	0.0081	0.0565	0.0548	0.0725	0.0185	0.0261	0.0157
	SANN	1.84(0.19)	1.66(0.19)	0.69(0.17)	1.53(0.24)	-1.66(0.23)	1.87(0.20)	1.78(0.18)	0.08(0.19)	-1.12(0.22)	0.01(0.26)	-1.61(0.26)	0.41(0.27)
	MSE	0.0505	0.0530	0.0514	0.0697	0.0698	0.0569	0.0765	0.0845	0.1109	0.0797	0.0872	0.0850
Group 2													
	True $\boldsymbol{\beta}_2$	β_{201}	β_{202}	β_{203}	β_{211}	β_{212}	β_{213}	β_{221}	β_{222}	β_{223}	β_{231}	β_{232}	β_{233}
		-3.68	-1.04	-1.89	-3.05	3.56	-0.06	-3.92	-1.58	0.02	2.53	-0.87	-1.28
	$\hat{\boldsymbol{\beta}}_2$	$\hat{\beta}_{201}$	$\hat{\beta}_{202}$	$\hat{\beta}_{203}$	$\hat{\beta}_{211}$	$\hat{\beta}_{212}$	$\hat{\beta}_{213}$	$\hat{\beta}_{221}$	$\hat{\beta}_{222}$	$\hat{\beta}_{223}$	$\hat{\beta}_{231}$	$\hat{\beta}_{232}$	$\hat{\beta}_{233}$
FMR	**NR**	-2.82(0.10)	-1.28(0.08)	-2.00(0.06)	-2.17(0.14)	2.73(0.24)	-0.24(0.08)	-2.37(0.21)	-0.77(0.08)	0.05(0.07)	1.45(0.15)	-0.82(0.09)	-1.11(0.08)
	MSE	0.7496	0.0640	0.0157	0.794	0.7465	0.0388	2.4466	0.6625	**0.0058**	1.1889	**0.0106**	0.0353
	MM	-3.78(0.13)	-1.19(0.07)	-1.99(0.08)	-2.85(0.09)	3.56(0.07)	0.08(0.05)	-3.72(0.10)	-1.46(0.08)	0.22(0.09)	2.36(0.09)	-1.01(0.08)	-1.47(0.07)
	MSE	0.0269	0.0274	0.0164	**0.0481**	**0.0049**	**0.0221**	**0.0500**	**0.0208**	0.0481	**0.0370**	0.0260	0.0410
	SANN	-3.46(0.22)	-0.95(0.19)	-1.92(0.12)	-2.80(0.26)	3.65(0.29)	-0.03(0.16)	-3.61(0.21)	-1.40(0.13)	0.17(0.21)	2.36(0.19)	-0.82(0.10)	-1.35(0.10)
	MSE	0.0968	0.0442	**0.0153**	0.1301	0.0922	0.0265	0.1402	0.0493	0.0666	0.065	0.0125	**0.0149**
FMRC	**NR**	-2.75(0.11)	-1.35(0.10)	-2.17(0.10)	-1.75(0.21)	2.45(0.27)	-0.25(0.10)	-1.77(0.17)	-0.70(0.09)	0.39(0.07)	1.20(0.20)	-0.77(0.09)	-1.09(0.12)
	MSE	0.8770	0.1061	0.0884	1.7341	1.305	0.0461	4.6514	0.7825	0.1418	1.8089	0.0181	0.0505
	MM	-3.77(0.12)	-1.19(0.08)	-1.98(0.09)	-2.84(0.12)	3.52(0.11)	0.08(0.08)	-3.72(0.10)	-1.44(0.08)	0.18(0.09)	2.35(0.09)	-1.00(0.09)	-1.44(0.10)
	MSE	**0.0225**	0.0289	0.0162	0.0585	0.0137	0.0260	**0.0500**	0.0260	0.0337	0.0405	0.0250	0.0356
	SANN	-3.35(0.36)	-1.10(0.07)	-2.06(0.18)	-2.50(0.25)	3.70(0.24)	0.08(0.17)	-3.60(0.35)	-1.57(0.20)	0.09(0.24)	2.19(0.29)	-1.04(0.18)	-1.54(0.16)
	MSE	0.2385	**0.0085**	0.0613	0.3650	0.0772	0.0485	0.2249	0.0401	0.0625	0.1997	0.0613	0.0932

Table 3.1: Mean estimates of coefficient matrices $\boldsymbol{\beta}_g$ (empirical standard deviations in parentheses) from the hundred data sets under Scenario 1 using FMR and FMRC frameworks with NR, MM, and SANN for optimization.

Coefficient estimates shown in Table 3.1 for each optimization strategy within FMR and FMRC frameworks show decent performance in respect to low mean squared error (MSE) values. The use of the MM algorithm results in the smallest standard errors and MSE values. FMR displays lower standard errors than FMRC

which is expected as the Scenario 1 is generated in an FMR framework.

		mean ARI	**ARI range**	$\hat{\boldsymbol{\pi}}$
FMR	**NR**	0.96	[0.94, 0.99]	(0.41, 0.59)
	MM	0.96	[0.92, 0.99]	(0.42, 0.58)
	SANN	0.96	[0.94, 0.98]	(0.41, 0.59)
FMRC	**NR**	0.95	[0.92, 0.97]	(0.41, 0.59)
	MM	0.95	[0.91, 0.98]	(0.42, 0.58)
	SANN	0.93	[0.93, 0.97]	(0.42, 0.58)

Table 3.2: Summary statistics of clustering results from the hundred data sets under Scenario 1 using FMR and FMRC frameworks with NR, MM, and SANN for optimization. The value $\hat{\boldsymbol{\pi}}$ represents the average mixing proportions over 100 simulations.

All maximization strategies displayed high ARI values as seen in Table 3.2, meaning the estimated group assignments were close to the true clusters in each case. Each algorithm also showed good performance in estimating the true mixing proportions $\boldsymbol{\pi} = (0.41, 0.59)$.

3.2.2 Scenario 2 results

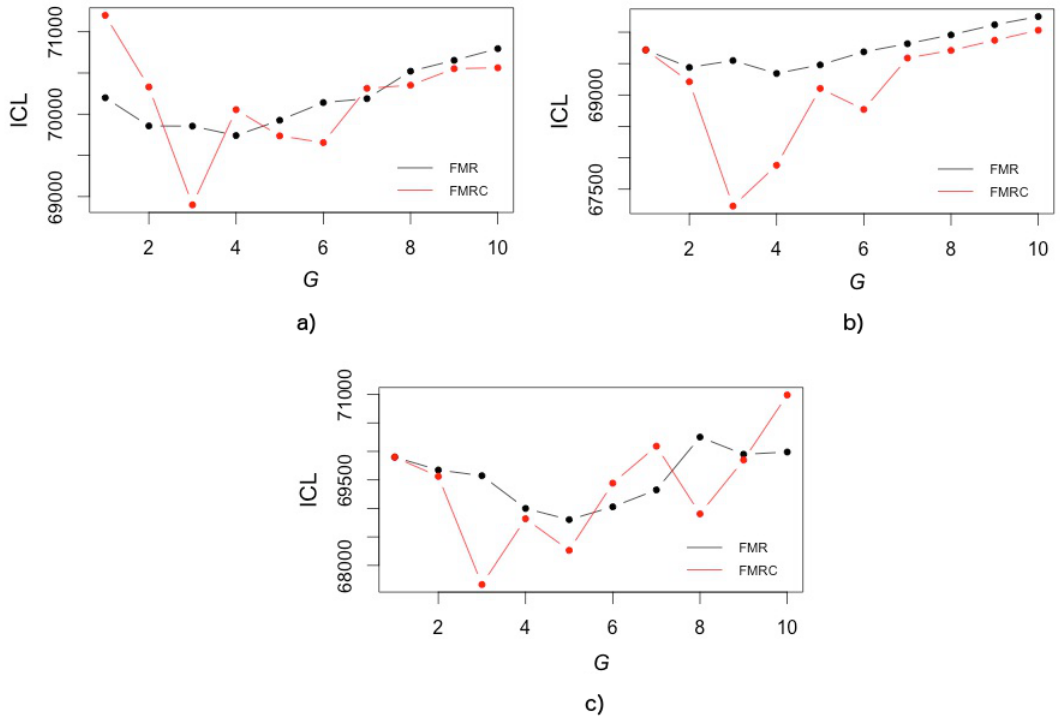


Figure 3.3: ICL values for models fit with number of components G ranging from 1 to 10 on one of the hundred simulated data sets under Scenario 2. Output from NR, MM, and SANN optimization is represented in a), b), and c), respectively.

ICL values shown in Figure 3.3 with each optimization technique achieve the lowest value at $G = 3$ within the FMRC framework. The selection of $G = 3$ within the FMRC framework accurately matches the design of Scenario 2. Also note that FMRC generally shows lower ICL values than FMR, implying the extra coefficients involved to model concomitant variables in FMRC are necessary to improve the fit of the model for this dataset. Although Figure 3.3 only shows the output from one

simulation, all FMRC estimation consistently selected $G = 3$ each of the 100 simulated data sets, whereas FMR estimation was unable to consistently select $G = 3$ and chose $G = 4$ or $G = 5$ in the majority of simulations. Since FMR estimation was unable to accurately select correct number of groups, coefficient estimates from this method are not displayed in Table 3.3 or Table 3.4.

Group 1													
	True $\boldsymbol{\beta}_1$	β_{101}	β_{102}	β_{103}	β_{111}	β_{112}	β_{113}	β_{121}	β_{122}	β_{123}	β_{131}	β_{132}	β_{133}
		-1.80	-1.69	1.06	-1.35	-1.97	-1.74	-1.42	-1.28	1.84	-0.61	0.21	-1.51
	$\hat{\boldsymbol{\beta}}_1$	$\hat{\beta}_{101}$	$\hat{\beta}_{102}$	$\hat{\beta}_{103}$	$\hat{\beta}_{111}$	$\hat{\beta}_{112}$	$\hat{\beta}_{113}$	$\hat{\beta}_{121}$	$\hat{\beta}_{122}$	$\hat{\beta}_{123}$	$\hat{\beta}_{131}$	$\hat{\beta}_{132}$	$\hat{\beta}_{133}$
FMR	**NR**	-	-	-	-	-	-	-	-	-	-	-	-
	MSE	-	-	-	-	-	-	-	-	-	-	-	-
	MM	-	-	-	-	-	-	-	-	-	-	-	-
	MSE	-	-	-	-	-	-	-	-	-	-	-	-
	SANN	-	-	-	-	-	-	-	-	-	-	-	-
	MSE	-	-	-	-	-	-	-	-	-	-	-	-
FMRC	**NR**	-1.56(0.71)	-1.09(0.41)	0.49(0.77)	-0.43(0.72)	-1.68(0.78)	-0.91(0.85)	-1.60(0.80)	-1.18(0.66)	0.97(1.00)	-0.39(0.74)	0.42(0.98)	-1.11(0.82)
	MSE	0.5617	0.5281	0.9178	1.3648	0.6925	1.4114	0.6724	0.4456	1.7569	0.5960	1.0045	0.8324
	MM	-1.82(0.16)	-1.90(0.13)	1.03(0.13)	-0.72(0.09)	-1.28(0.08)	-1.27(0.08)	-1.46(0.08)	-1.34(0.07)	1.44(0.07)	-0.48(0.07)	0.51(0.08)	-1.25(0.09)
	MSE	**0.0260**	**0.0610**	**0.0178**	0.4050	0.4825	**0.2273**	**0.0080**	**0.0085**	**0.1649**	**0.0218**	0.0964	**0.0757**
	SANN	-2.02(0.22)	-1.92(0.22)	0.78(0.29)	-0.79(0.25)	-1.35(0.27)	-1.14(0.25)	-1.46(0.16)	-1.35(0.15)	1.27(0.21)	-0.31(0.16)	0.48(0.14)	-1.00(0.20)
	MSE	0.0968	0.1013	0.1625	**0.3761**	**0.4573**	0.4225	0.0272	0.0274	0.3690	0.1156	**0.0925**	0.3001
Group 2													
	True $\boldsymbol{\beta}_2$	β_{201}	β_{202}	β_{203}	β_{211}	β_{212}	β_{213}	β_{221}	β_{222}	β_{223}	β_{231}	β_{232}	β_{233}
		2.71	-1.27	0.33	2.00	3.37	3.42	0.85	-2.61	1.44	3.63	-2.26	3.92
	$\hat{\boldsymbol{\beta}}_2$	$\hat{\beta}_{201}$	$\hat{\beta}_{202}$	$\hat{\beta}_{203}$	$\hat{\beta}_{211}$	$\hat{\beta}_{212}$	$\hat{\beta}_{213}$	$\hat{\beta}_{221}$	$\hat{\beta}_{222}$	$\hat{\beta}_{223}$	$\hat{\beta}_{231}$	$\hat{\beta}_{232}$	$\hat{\beta}_{233}$
FMR	**NR**	-	-	-	-	-	-	-	-	-	-	-	-
	MSE	-	-	-	-	-	-	-	-	-	-	-	-
	MM	-	-	-	-	-	-	-	-	-	-	-	-
	MSE	-	-	-	-	-	-	-	-	-	-	-	-
	SANN	-	-	-	-	-	-	-	-	-	-	-	-
	MSE	-	-	-	-	-	-	-	-	-	-	-	-
FMRC	**NR**	1.39(1.07)	-0.95(0.43)	-0.63(0.81)	1.54(1.06)	2.17(0.71)	1.20(0.80)	0.29(0.42)	-1.31(1.24)	0.07(0.71)	1.92(0.82)	-1.65(1.22)	2.13(1.04)
	MSE	2.8873	0.2873	1.5777	1.3352	1.9441	5.5684	0.49	3.2276	2.381	3.5965	1.8605	4.2857
	MM	2.58(0.16)	-1.55(0.14)	0.16(0.20)	1.35(0.07)	2.58(0.09)	2.61(0.10)	0.66(0.16)	-2.43(0.11)	1.30(0.35)	3.00(0.10)	-2.24(0.12)	3.20(0.21)
	MSE	**0.0425**	**0.0980**	**0.0689**	**0.4274**	**0.6322**	**0.6661**	**0.0617**	**0.0445**	**0.1421**	**0.4069**	**0.0148**	**0.5625**
	SANN	1.01(0.43)	-1.60(0.35)	0.50(0.43)	-0.23(0.35)	1.74(0.48)	1.86(0.49)	0.56(0.41)	-1.84(0.67)	0.72(0.49)	2.68(0.44)	-2.46(0.76)	1.65(0.50)
	MSE	3.0749	0.2314	0.2138	5.0954	2.8873	2.6737	0.2522	1.0418	0.7585	1.0961	0.6176	5.4029
Group 3													
	True $\boldsymbol{\beta}_3$	β_{301}	β_{302}	β_{303}	β_{311}	β_{312}	β_{313}	β_{321}	β_{322}	β_{323}	β_{331}	β_{332}	β_{333}
		-1.12	-0.05	-3.21	4.76	-5.51	-1.60	0.90	1.69	-1.89	4.73	-4.95	3.47
	$\hat{\boldsymbol{\beta}}_3$	$\hat{\beta}_{301}$	$\hat{\beta}_{302}$	$\hat{\beta}_{303}$	$\hat{\beta}_{311}$	$\hat{\beta}_{312}$	$\hat{\beta}_{313}$	$\hat{\beta}_{321}$	$\hat{\beta}_{322}$	$\hat{\beta}_{323}$	$\hat{\beta}_{331}$	$\hat{\beta}_{332}$	$\hat{\beta}_{333}$
FMR	**NR**	-	-	-	-	-	-	-	-	-	-	-	-
	MSE	-	-	-	-	-	-	-	-	-	-	-	-
	MM	-	-	-	-	-	-	-	-	-	-	-	-
	MSE	-	-	-	-	-	-	-	-	-	-	-	-
	SANN	-	-	-	-	-	-	-	-	-	-	-	-
	MSE	-	-	-	-	-	-	-	-	-	-	-	-
FMRC	**NR**	-1.00(0.61)	0.02(0.40)	-2.28(0.34)	2.63(0.93)	-2.98(1.55)	-0.81(1.10)	0.29(0.53)	0.54(0.69)	-1.23(0.34)	3.30(0.99)	-3.78(1.19)	1.93(1.00)
	MSE	**0.3865**	0.1649	0.9805	5.4018	8.8034	1.8341	0.6530	1.7986	0.5512	3.025	2.785	3.3716
	MM	-1.79(0.14)	0.05(0.11)	-3.44(0.16)	3.27(0.12)	-3.36(0.06)	-1.58(0.06)	0.58(0.07)	0.74(0.06)	-1.21(0.10)	3.27(0.10)	-3.78(0.04)	2.14(0.09)
	MSE	0.4685	**0.0221**	**0.0785**	2.2345	**4.6261**	**0.0040**	**0.1073**	**0.9061**	0.4724	2.1416	**1.3705**	1.777
	SANN	-1.69(0.38)	-0.55(0.64)	-3.75(0.34)	3.36(0.37)	-2.65(0.94)	-0.66(0.33)	0.31(0.12)	0.58(0.19)	-1.28(0.29)	3.40(0.47)	-2.30(1.02)	3.20(0.37)
	MSE	0.4693	0.6596	0.4072	**2.0969**	9.0632	0.9925	0.3625	1.2682	**0.4562**	**1.9898**	8.0629	**0.2098**

Table 3.3: Mean estimates of coefficient matrices $\boldsymbol{\beta}_g$ (empirical standard deviations in parentheses) from the hundred data sets under Scenario 2 using FMR and FMRC frameworks with NR, MM, and SANN for optimization.

In Scenario 2, the MM algorithm shows greater performance in terms of low

standard errors and MSE values for the coefficient estimates in Table 3.3. The NR and SANN algorithms exhibit the highest MSE values and therefore poorest performance in parameter estimation.

		mean ARI	**ARI range**	$\hat{\boldsymbol{\pi}}$
FMR	**NR**	-	-	-
	MM	-	-	-
	SANN	-	-	-
FMRC	**NR**	0.94	[0.80, 1.00]	(0.24, 0.29, 0.47)
	MM	0.98	[0.89, 1.00]	(0.24, 0.28, 0.48)
	SANN	0.86	[0.82, 0.89]	(0.21, 0.31, 0.48)

Table 3.4: Summary statistics of clustering results from the hundred data sets under Scenario 2 using FMR and FMRC frameworks with NR, MM, and SANN for optimization. The value $\hat{\boldsymbol{\pi}}$ represents the average mixing proportions over 100 simulations.

Within the FMRC framework, the MM maximization strategy displayed the highest ARI values as seen in Table 3.4, with SANN displaying the lowest ARI values. The NR algorithm displayed the widest variation in ARI, which may be related to the NR techniques susceptibility to local optima. Each did however show good performance in estimating the true mixing proportions $\boldsymbol{\pi} = (0.24, 0.28, 0.48)$.

3.2.3 Scenario 3 results

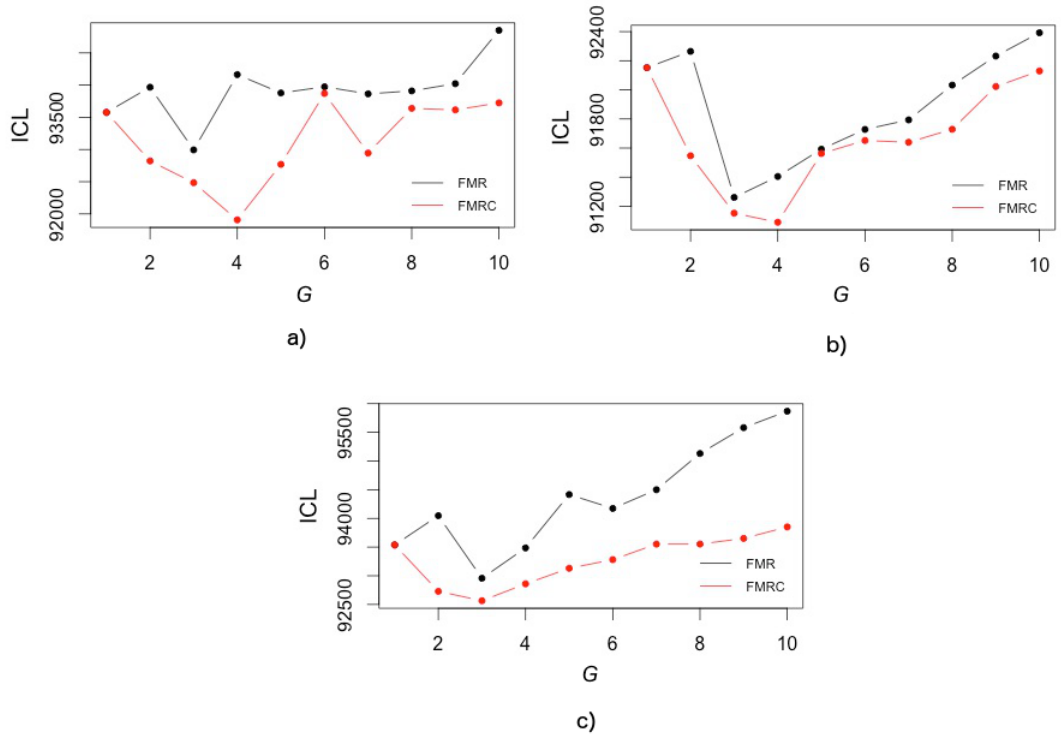


Figure 3.4: ICL values for models fit with number of components G ranging from 1 to 10 on one of the hundred simulated data sets under Scenario 3. Output from NR, MM, and SANN optimization is represented in a), b), and c), respectively.

ICL values shown in Figure 3.4 are lowest at $G = 4$ for the proposed NR and MM maximization techniques within the FMRC framework, but the SANN approach reaches its lowest ICL value at $G = 3$. This means the NR and MM techniques are selecting the model with the correct number of groups used in this simulation, but the SANN approach is not. Although Figure 3.4 only shows the output from one simulation, FMRC estimation consistently selected $G = 4$ with NR and MM optimization

in each of the 100 simulated data sets, but SANN optimization selected $G = 3$ in the majority of simulations. FMR estimation also displayed similar behaviour to Figure 3.4 in each of the 100 simulated data sets, consistently selecting $G = 3$ groups. Since FMR estimation and SANN optimization within the FMRC framework were unable to accurately select correct number of groups, coefficient estimates from these method are not displayed in Table 3.5. The coefficient estimates for $\boldsymbol{\beta}$ are also not displayed for Scenario 3 as the estimation techniques used (FMR and FMRC) are not constructed with group-specific means and covariances in the covariates and thus exhibited very poor estimates of $\boldsymbol{\beta}$.

		mean ARI	**ARI range**	$\hat{\boldsymbol{\pi}}$
FMR	**NR**	-	-	-
	MM	-	-	-
	SANN	-	-	-
FMRC	**NR**	0.87	[0.78, 0.98]	(0.13, 0.23, 0.26, 0.38)
	MM	0.93	[0.75, 0.99]	(0.13, 0.22, 0.28, 0.37)
	SANN	-	-	-

Table 3.5: Summary statistics of clustering results from the hundred data sets under Scenario 3 using FMR and FMRC frameworks with NR, MM, and SANN for optimization. The value $\hat{\boldsymbol{\pi}}$ represents the average mixing proportions over 100 simulations.

The NR and MM maximization methods both display fairly high mean ARI values with wide ranges. Each algorithm also showed good performance in estimating the true mixing proportions $\boldsymbol{\pi} = (0.15, 0.21, 0.28, 0.36)$.

Chapter 4

Real data analysis

The data set analyzed in this thesis uses gut microbiome information from a cross-sectional study of 98 healthy volunteers, as well as accompanying covariate information from 214 micronutrients obtained through a food questionnaire (further details on the data set and the study can be found in Wu et al. (2011)). The data set, which is made available through the $PenLNM$ package in R, contains taxa counts of the genera Bacteroides, Prevotella, and Ruminococcus for the 98 individuals. Furthermore, the sparse DM regression of Chen and Li (2013) selected 11 nutrients through a group ℓ_1 penalty to be most influential on resulting genera, and only these 11 nutrients are used in our analysis; these nutrients are: total polyunsaturated fat in grams (poly), methionine in grams (meth), sucrose in grams (sucr), animal protein in grams (aprot), vitamin E food fortification only in milligrams (es2mg), maltose in grams (malt), added germ from wheat in grams (germa), choline from phosphatidylcholine in milligrams (ptdcho), taurine in milligrams (tau), naringenin in milligrams (unarg), and eriodictylol in milligrams (uerid).

4.1 Results

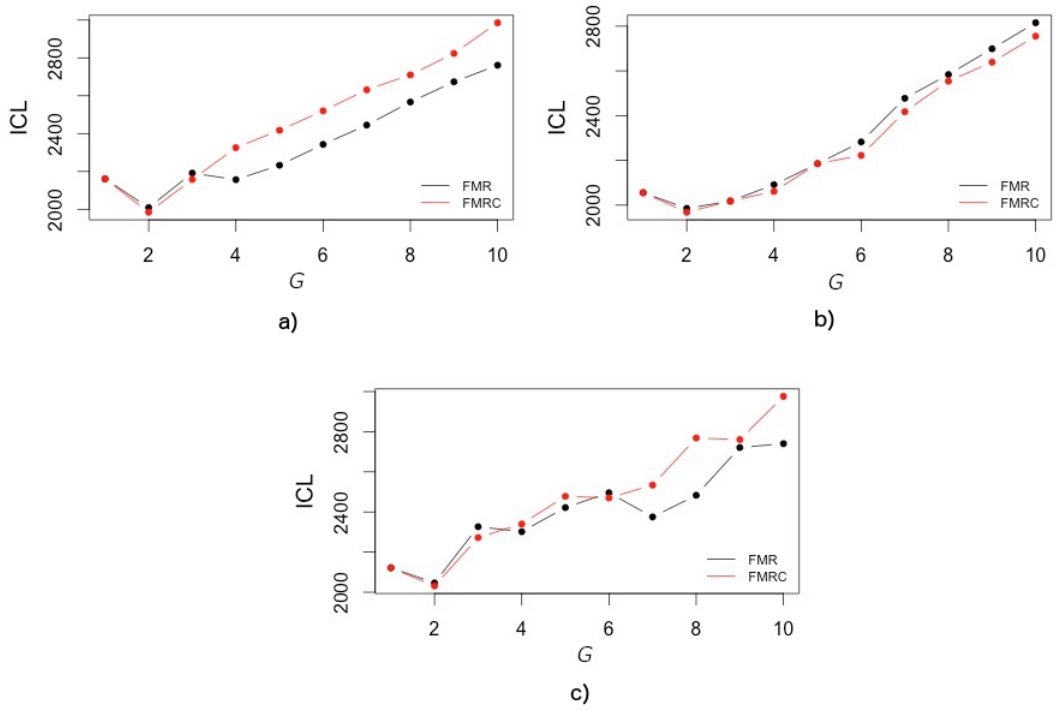


Figure 4.1: ICL values for FMR and FMRC models with number of components G ranging from 1 to 10 on on the Wu data set. Output from NR, MM, and SANN optimization is represented in a), b), and c), respectively.

ICL values shown in Figure 4.1 are lowest for FMR and FMRC at $G = 2$ in all optimization strategies, meaning each technique is selecting the best fitting model to have two underlying groups on this data set. FMRC produces a slightly lower value at $G = 2$ in each method, indicating a small concomitant effect of the covariates is present on the mixing proportions.

Coefficient estimates with NR, MM, and SANN optimization are displayed

in Tables 4.1, 4.2, and 4.3, respectively. In each case, group 1 is relatively high in Bacteroides and Ruminococcus, and group 2 is relatively high in Prevotella. The data set exhibits large differences in direction and magnitude of the coefficients between enterotypes within the FMR and FMRC framework in each optimization algorithm.

	Estimated $\hat{\boldsymbol{\beta}}$ values with FMR			Estimated $\hat{\boldsymbol{\beta}}$ values with FMRC		
	Bact.	Prev.	Rumi.	Bact.	Prev.	Rumi.
$\boldsymbol{\beta}_1$: int.	2.09	-4.97	-1.16	0.66	-3.30	-0.62
poly	-0.97	7.46	-0.62	-1.44	10.01	-0.05
meth	3.82	-22.30	8.02	16.50	-15.39	15.59
sucr	-3.55	-1.12	2.43	-10.80	2.79	1.76
aprot	-2.71	45.63	-2.02	-5.08	13.37	-4.32
es2mg	2.51	16.85	-0.60	0.48	10.52	1.17
malt	-2.19	15.21	-0.78	-0.72	11.88	2.00
germa	-5.67	0.03	-2.06	0.56	2.03	-1.40
ptdcho	0.37	-9.93	-1.96	-0.28	-34.35	-3.63
tau	-2.88	-41.87	-8.36	7.10	0.9	-9.29
unarg	-0.70	13.63	2.42	-4.71	8.31	-1.35
uerid	-0.74	-7.33	-0.54	-4.73	-12.23	-0.75
	Bact.	Prev.	Rumi.	Bact.	Prev.	Rumi.
$\boldsymbol{\beta}_2$: int.	0.88	-1.47	-1.69	3.55	-0.20	-0.99
poly	2.43	14.01	1.64	3.66	21.61	2.95
meth	-21.41	3.66	-20.75	-15.50	-20.39	-1.35
sucr	-2.53	8.77	-3.30	-2.51	-2.36	2.39
aprot	22.02	-14.05	16.88	12.31	8.14	4.37
es2mg	4.94	2.51	6.43	-8.81	11.67	0.88
malt	-6.89	-3.63	-4.94	1.51	-9.23	-1.70
germa	-0.64	6.69	-0.39	1.82	3.07	0.72
ptdcho	2.31	-5.03	2.69	8.73	1.24	1.10
tau	10.04	14.68	6.38	-7.58	-2.73	-0.67
unarg	3.57	7.85	-0.93	-8.65	19.55	3.36
uerid	1.04	9.78	5.64	-2.24	15.22	1.47

Table 4.1: Estimates of coefficient matrices $\boldsymbol{\beta}_g$ from the Wu data set with FMR and FMRC through NR optimization.

	Estimated $\hat{\boldsymbol{\beta}}$ values with FMR	**Estimated $\hat{\boldsymbol{\beta}}$ values with FMRC**

$\boldsymbol{\beta}_1$:

	Bact.	Prev.	Rumi.
int.	3.76	-5.99	-0.13
poly	-2.51	12.86	-1.98
meth	39.90	-13.98	41.69
sucr	-1.22	0.98	7.99
aprot	-21.11	69.77	-19.65
es2mg	1.05	31.06	1.45
malt	-3.98	23.20	0.32
germa	-1.84	-2.99	-3.60
ptdcho	-10.40	-17.92	-10.01
tau	-10.19	-85.43	-19.61
unarg	1.03	20.41	4.63
uerid	1.83	-12.14	-1.03

	Bact.	Prev.	Rumi.
int.	1.48	-6.53	-1.33
poly	1.45	20.48	-0.33
meth	14.20	-21.15	13.15
sucr	-3.35	8.73	2.04
aprot	-9.81	20.93	-5.59
es2mg	-2.02	25.56	-0.65
malt	-0.87	24.39	0.68
germa	0.86	1.47	0.02
ptdcho	-4.89	-47.18	-2.91
tau	-0.65	-19.55	-8.68
unarg	0.05	16.35	-0.18
uerid	6.11	-19.27	3.06

$\boldsymbol{\beta}_2$:

	Bact.	Prev.	Rumi.
int.	1.85	-0.86	-1.34
poly	7.71	21.66	7.00
meth	-57.04	-29.15	-55.02
sucr	2.78	11.91	-6.53
aprot	56.93	-5.11	46.37
es2mg	1.13	-3.85	5.61
malt	-11.40	-9.52	-8.77
germa	5.70	16.97	3.02
ptdcho	13.37	1.00	9.27
tau	-0.21	18.33	3.26
unarg	4.84	8.70	0.07
uerid	1.72	10.31	8.13

	Bact.	Prev.	Rumi.
int.	0.84	-0.83	-1.60
poly	1.02	26.24	2.00
meth	-8.99	-26.28	-6.54
sucr	-0.33	2.74	3.15
aprot	10.85	-14.23	8.91
es2mg	-1.63	17.12	1.28
malt	-2.82	-7.43	-1.02
germa	0.96	-5.56	0.06
ptdcho	6.93	3.52	2.39
tau	-0.29	2.60	1.14
unarg	-10.99	18.62	1.88
uerid	2.32	19.68	1.85

Table 4.2: Estimates of coefficient matrices $\boldsymbol{\beta}_g$ on the Wu data set with FMR and FMRC through MM optimization.

	Estimated $\hat{\boldsymbol{\beta}}$ values with FMR			Estimated $\hat{\boldsymbol{\beta}}$ values with FMRC		
	Bact.	Prev.	Rumi.	Bact.	Prev.	Rumi.
$\boldsymbol{\beta}_1$: int.	0.94	-1.74	-1.40	8.67	-1.17	2.34
poly	-8.90	4.00	-2.56	2.38	0.38	4.36
meth	4.00	-5.76	8.12	12.59	-2.47	7.50
sucr	-0.40	6.00	0.61	-6.83	-9.34	1.94
aprot	15.01	5.02	-6.79	1.28	-2.02	12.94
es2mg	9.22	4.41	-5.30	9.24	12.48	3.31
malt	-0.95	3.19	0.93	0.59	2.06	-0.85
germa	0.72	0.19	0.96	1.05	-1.74	-1.98
ptdcho	-4.90	-5.51	-1.71	5.75	7.22	3.70
tau	8.00	3.25	0.20	-1.35	-4.54	-1.42
unarg	-3.18	3.68	-1.29	-4.04	0.20	4.68
uerid	-3.89	-0.57	-1.85	-3.02	-11.74	1.99
	Bact.	Prev.	Rumi.	Bact.	Prev.	Rumi.
$\boldsymbol{\beta}_2$: int.	7.41	-5.17	1.12	1.25	-2.06	-1.51
poly	-3.76	-1.20	-18.90	-1.50	6.28	-1.01
meth	-2.30	1.10	-4.46	3.99	-1.85	4.74
sucr	-9.56	8.32	2.67	0.66	4.85	1.43
aprot	-3.12	14.09	3.35	9.93	-2.86	2.72
es2mg	-1.35	2.41	0.78	1.40	3.45	4.49
malt	1.61	0.11	-10.88	-3.27	4.26	4.16
germa	-9.25	14.02	9.76	1.45	10.06	-0.41
ptdcho	2.97	-2.81	1.85	5.76	-6.62	4.20
tau	6.96	-9.32	4.10	-9.81	-4.89	-7.50
unarg	-4.33	1.02	-2.31	-2.82	0.57	-1.23
uerid	-6.40	-1.12	-4.06	-4.85	4.61	-6.00

Table 4.3: Estimates of coefficient matrices $\boldsymbol{\beta}_g$ from the Wu data set with FMR and FMRC through SANN optimization.

		$\hat{\boldsymbol{\pi}}$	$\boldsymbol{\ell}$
FMR	**NR**	(0.58, 0.42)	-8578.21
	MM	(0.56, 0.44)	-8520.18
	SANN	(0.63, 0.37)	-8598.53
FMRC	**NR**	(0.67, 0.33)	-8528.61
	MM	(0.53, 0.47)	**-8497.50**
	SANN	(0.67, 0.33)	-8558.25

Table 4.4: Estimates of the mixing proportions for the Wu data set using NR, MM, and SANN optimization strategies within FMR and FMRC frameworks. Values for the log-likelihood $\boldsymbol{\ell}$ are displayed for each estimation technique are displayed as well.

Table 4.4 presents estimates for $\boldsymbol{\pi}$ on the Wu data set. With each approach, the two groups are fairly evenly split, with the group high in Bacteroides (group 1) consistently being larger. Based on the log-likelihood criterion, the MM algorithm within the FMRC framework is the best fitting model.

Chapter 5

Discussion

This work illustrates a mixture model framework to better relate biological/environmental covariates to the composition of the microbiome by accounting for the underlying enterotype structure observed across samples. Simulations were performed under different assumptions for the distribution of covariates, as well as an application to a real data set which linked nutrient intake to resulting taxa counts. Furthermore, the performances of traditional FMR and FMRC to incorporate the covariate effect on the mixing proportions were compared. Although this thesis was defined in respect to microbiome data, note however that this methodology can be applied other varieties of overdispersed count datasets with covariate information and where there is believed to be an underlying latent class.

From the simulation studies, the MM algorithm showed the best performance with respect to identification of the underlying number of groups, proximity of parameter estimates to the true values, and low standard deviations. The use of the heuristic approach of SANN optimization was useful in Scenario 1 and Scenario 2, but showed poor performance in Scenario 3 and was unable to identify the true number of groups. The NR algorithm showed moderate performance in most areas, but still had high standard deviations relative to MM, suggesting the NR methods

susceptibility to falling in local optima.

For Scenario 1, FMR estimation proved the most effective as expected since the data was simulated from an FMR framework, but FMRC estimated still displayed good performance, suggesting little negative impacts to applying FMRC estimation on FMR simulated data in this framework. For Scenario 2, FMRC estimation showed significant improvements in performance over FMR estimation, showing the necessity of incorporating the concomitant effect in this scenario. For Scenario 3, FMRC estimation shows improvement in ARI and identifying the underlying number of groups when compared with FMR estimation. Scenario 3 also displays a strong advantage of FMRC estimation as no assumptions were made on the distribution of the covariates $\mathbf{X}$, but FMRC was still able to accurately model the latent groups.

In the real data analysis section, the best fitting model from all optimization approaches through ICL comparisons uses two underlying groups. This result agrees with the findings of Wu et al. (2011) where they found these 98 microbiome samples can be best clustered into two enterotypes. In each maximization strategy, enterotype 1 represents the samples which are relatively high in Bacteroides and Ruminococcus, and enterotype 2 represents samples which had higher levels of Prevotella. This finding also agrees with Wu et al. (2011), where they found levels of Bacteroides and Prevotella to be the distinguishing genera between enterotypes, and samples high in Ruminococcus would join the Bacteroides enterotype. The MM and NR methods agree on the signs of coefficients for the $\boldsymbol{\beta}_g$ matrices, but magnitudes differ. Results from SANN maximization do not show many similarities to the MM and NR results. From the MM and NR algorithms on the Wu dataset, the effects of animal protein

and methionine on resulting taxa counts in particular greatly differ based on an individual's enterotype, suggesting the importance of accounting for enterotypes in exploring covariate effects on microbiome composition.

Further work can be done to handle the zero-inflation often present in taxa counts from microbiome data, where more zero counts appear than expected from the DM model; possible solutions could come from the multilevel negative binomial regression modeling of Moghimbeigi et al. (2008), or the multilevel Poisson regression modeling of Lee et al. (2006). The DM model also does not have a very flexible covariance structure to model relationships between the counts, which suggests the possible benefit of the more flexible Gaussian-multinomial model in which the probability vectors $\mathbf{p}$ from the multinomial distribution will follow a logistically transformed multivariate Gaussian distribution (Aitchison, 1985). Additionally, the methods outlined in this thesis are not intended for high dimensional data due to high computational cost of computing the complete data log-likelihood with increasing numbers of taxa and covariates, and there is no scheme being employed to remove covariates with little impact on resulting taxa counts. The former issue can be fixed by utilizing the online computation schemes of Zhang (2014) on DM regression where only a small batch of data points are stored in memory at each iteration, and the latter issue can be resolved with the regularization techniques of Chen and Li (2013) and Lin et al. (2014).

Bibliography

Adler, D., Murdoch, D., and others (2015). *rgl: 3D Visualization Using OpenGL*. R package version 0.95.1367.

Adlerberth, I. (2008). Factors influencing the establishment of the intestinal microbiota in infancy. *Nestle Nutrition Workshop Series*, 136:13–29.

Aitchison, J. (1985). A general class of distributions on the simplex. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 136–146.

Antunes, L. and Finlay, B. (2011). A comparative analysis of the effect of antibiotic treatment and enteric infection on intestinal homeostasis. *Gut Microbes*, 2(2):105–108.

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D., Fernandes, G., Tap, J., Bruls, T., Batto, J., et al. (2011). Enterotypes of the human gut microbiome. *Nature*, 473(7346):174–180.

Backhed, F., Ley, R., Sonnenburg, J., Peterson, D., and Gordon, J. (2005). Host-bacterial mutualism in the human intestine. *Science*, 307(5717):1915–1920.

Bélisle, C. (1992). Convergence theorems for a class of simulated annealing algorithms on r^d. *Journal of Applied Probability*, pages 885–895.

Bergström, A., Skov, T., Bahl, M., Roager, H., Christensen, L., Ejlerskov, K., Mølgaard, C., Michaelsen, K., and Licht, T. (2014). Establishment of intestinal microbiota during early life: a longitudinal, explorative study of a large cohort of danish infants. *Applied and Environmental Microbiology*, 80(9):2889–2900.

Biernacki, C., Celeux, G., and Govaert, G. (1998). Assessing a mixture model for clustering eith integrated classification likelihood. Technical Report 3521, Rhône-Alpes: INRIA.

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

Cantarel, B., Lombard, V., and Henrissat, B. (2012). Complex carbohydrate utilization by the healthy human microbiome. *PloS One*, 7(6):e28742.

Chen, J. and Li, H. (2013). Variable selection for sparse dirichlet-multinomial regression with an application to microbiome data analysis. *The Annals of Applied Statistics*, 7(1).

Dayton, C. and Macready, G. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association*, 83(401):173–178.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38.

DeSarbo, W. and Cron, W. (1988). A maximum likelihood methodology for cluster-wise linear regression. *Journal of Classification*, 5(2):249–282.

Eckburg, P., Bik, E., Bernstein, C., Purdom, E., Dethlefsen, L., Sargent, M., Gill, S., Nelson, K., and Relman, D. (2005). Diversity of the human intestinal microbial flora. *Science*, 308(5728):1635–1638.

Fraley, C., Raftery, A., Murphy, T., and Scrucca, L. (2012). *mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation*.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer Science & Business Media.

Goffe, W., Ferrier, G., and Rogers, J. (1994). Global optimization of statistical functions with simulated annealing. *Journal of Econometrics*, 60(1):65–99.

Greenblum, S., Turnbaugh, P., and Borenstein, E. (2012). Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proceedings of the National Academy of Sciences*, 109(2):594–599.

Haldane, J. (1941). The fitting of binomial distributions. *Annals of Eugenics*, 11(1):179–181.

Hildebrandt, M., Hoffmann, C., Sherrill-Mix, S., Keilbaugh, S., Hamady, M., Chen, Y., Knight, R., Ahima, R., Bushman, F., and Wu, G. (2009). High-fat diet de-

termines the composition of the murine gut microbiome independently of obesity. *Gastroenterology*, 137(5):1716–1724.

Holmes, I., Harris, K., and Quince, C. (2012). Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PLoS One*, 7(2):e30126.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1):193–218.

Hunter, D. and Lange, K. (2000). Quantile regression via an mm algorithm. *Journal of Computational and Graphical Statistics*, 9(1):60–77.

Kaufman, L. and Rousseeuw, P. (1987). *Clustering by means of medoids*. North-Holland.

Kaufman, L. and Rousseeuw, P. (1990). Partitioning around medoids (program pam). *Finding groups in data: an introduction to cluster analysis*, pages 68–125.

Kirkpatrick, S., Gelatt, C., Vecchi, M., et al. (1983). Optimization by simulated annealing. *Science*, 220(4598):671–680.

Kuczynski, J., Lauber, C., Walters, W., Parfrey, L., Clemente, J., Gevers, D., and Knight, R. (2012). Experimental and analytical tools for studying the human microbiome. *Nature Reviews Genetics*, 13(1):47–58.

Lange, K. (1995). A gradient algorithm locally equivalent to the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 425–437.

Lange, K. (1999). *Numerical analysis for statisticians*. New York: Springer-Verlag.

LeBlanc, J., Milani, C., de Giori, G., Sesma, F., van Sinderen, D., and Ventura, M. (2013). Bacteria as vitamin suppliers to their host: a gut microbiota perspective. *Current Opinion in Biotechnology*, 24(2):160–168.

Lee, A., Wang, K., Scott, J., Yau, K., and McLachlan, G. (2006). Multi-level zero-inflated poisson regression modelling of correlated count data with excess zeros. *Statistical Methods in Medical Research*, 15(1):47–61.

Ley, R., Peterson, D., and Gordon, J. (2006). Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell*, 124(4):837–848.

Lin, W., Shi, P., Feng, R., and Li, H. (2014). Variable selection in regression with compositional covariates. *Biometrika*, page asu031.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2015). *cluster: Cluster Analysis Basics and Extensions*.

McLachlan, G. and Peel, D. (2000). Finite mixture models. *Wiley Series in Probability and Statistics*.

Moghimbeigi, A., Eshraghian, M., Mohammad, K., and Mcardle, B. (2008). Multilevel zero-inflated negative binomial regression modeling for over-dispersed count data with extra zeros. *Journal of Applied Statistics*, 35(10):1193–1202.

Mosimann, J. (1962). On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika*, pages 65–82.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen,

T., Pons, N., Levenez, F., Yamada, T., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65.

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418):55–60.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Round, J. and Mazmanian, S. (2009). The gut microbiota shapes intestinal immune responses during health and disease. *Nature Reviews Immunology*, 9(5):313–323.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Streit, W. and Schmitz, R. (2004). Metagenomics–the key to the uncultured microbes. *Current Opinion in Microbiology*, 7(5):492–498.

Titterington, D., Smith, A., and Makov, U. (1985). Statistical analysis of finite mixture distributions. *John Wiley&Sons Ltd, Chichester*.

Tjalsma, H., Boleij, A., Marchesi, J., and Dutilh, B. (2012). A bacterial driver–passenger model for colorectal cancer: beyond the usual suspects. *Nature Reviews Microbiology*, 10(8):575–582.

Turnbaugh, P., Hamady, M., Yatsunenko, T., Cantarel, B., Duncan, A., Ley, R.,

Sogin, M., Jones, W., Roe, B., Affourtit, J., et al. (2009). A core gut microbiome in obese and lean twins. *Nature*, 457(7228):480–484.

Ueda, N. and Nakano, R. (1998). Deterministic annealing em algorithm. *Neural Networks*, 11(2):271–282.

Venables, W. and Ripley, B. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Wedel, M. (2002). Concomitant variables in finite mixture models. *Statistica Neerlandica*, 56(3):362–375.

Wu, G., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y., Keilbaugh, S., Bewtra, M., Knights, D., Walters, W., Knight, R., et al. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science*, 334(6052):105–108.

Yong, E. (2012). Gut microbial enterotypes become less clear-cut. *Nature News*.

Zhang, Y. (2014). *Selected Topics in Statistical Computing*. PhD thesis, North Carolina State University.

Zhang, Y. and Zhou, H. (2013). *cluster: MGLM: R package and Matlab toolbox for multivariate categorical data analysis*.

Zhou, H. and Lange, K. (2010). Mm algorithms for some discrete multivariate distributions. *Journal of Computational and Graphical Statistics*, 19(3):645–665.

Zoetendal, E., Rajilić-Stojanović, M., and De Vos, W. (2008). High-throughput di-

versity and functionality analysis of the gastrointestinal tract microbiota. *Gut*, 57(11):1605–1615.

Appendix A

Derivation of Newton-Raphson method for mixtures of Dirichlet-multinomial regression models

Begin with the response matrix $\mathbf{Y}$:

$$\mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1d} \\ y_{21} & y_{22} & \cdots & y_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nd} \end{bmatrix}$$

indexed by $i = 1, \ldots, n$ observations, and $j = 1, \ldots, d$ taxa. Suppose there are G components in a given population, the rows of $\mathbf{Y}$, $\mathbf{y}_i$, will follow the DM distribution defined in (2.9) indexed by parameter $\boldsymbol{\alpha}_{gi}$, where $\boldsymbol{\alpha}_{gi} = (\alpha_{gi1}, \alpha_{gi2}, \ldots, \alpha_{gid})$. The

likelihood for a mixture of DM distributions is given as:

$$\begin{aligned} L(\boldsymbol{\alpha}, \boldsymbol{\pi}; \mathbf{Y}) &= \prod_{i=1}^{n} \sum_{g=1}^{G} \pi_g f_{DM}(\mathbf{y}_i | \boldsymbol{\alpha}_{gi}) \\ &= \prod_{i=1}^{n} \sum_{g=1}^{G} \pi_g \frac{y_{i+}}{y_{i1}! \ldots y_{id}!} \frac{\Gamma(\alpha_{gi+})}{\prod_{j=1}^{d} \Gamma(\alpha_{gij})} \frac{\prod_{j=1}^{d} \Gamma(y_{ij} + \alpha_{gij})}{\Gamma(\sum_{j=1}^{d} y_{ij} + \alpha_{gij})}. \end{aligned}$$

Let $c_{ij} = 1$ if $y_{ij} \geq 1$, otherwise $c_{ij} = 0$ if $y_{ij} = 0$, then we have:

$$\begin{aligned} \frac{\Gamma(y_{ij} + \alpha_{gij})}{\Gamma(\alpha_{gij})} &= \left[\frac{(y_{ij} + \alpha_{gij} - 1) \ldots \alpha_{gij} \Gamma(\alpha_{gij})}{\Gamma(\alpha_{gij})} \right]^{c_{ij}} \left[\frac{\Gamma(\alpha_{gij})}{\Gamma(\alpha_{gij})} \right]^{1 - c_{ij}} \\ &= [(y_{ij} + \alpha_{gij} - 1) \ldots \alpha_{gij}]^{c_{id}}. \end{aligned}$$

As this is in a regression framework, the link function from (2.11) is used, so $\alpha_{gij} = e^{\mathbf{x_i} \boldsymbol{\beta}_{\mathbf{gj}}}$, where $\boldsymbol{\beta}_{gj} = (\beta_{g1j}, \beta_{g2j}, \ldots, \beta_{gpj})$. So $L(\boldsymbol{\alpha}, \boldsymbol{\pi}; \mathbf{Y})$ can be re-written as:

$$\begin{aligned} L(\boldsymbol{\alpha}, \boldsymbol{\pi}; \mathbf{Y}) &= \prod_{i=1}^{n} \sum_{g=1}^{G} \pi_g \frac{y_{i+}}{y_{i1}! \ldots y_{id}!} \frac{\Gamma(\alpha_{gi+})}{\prod_{j=1}^{d} \Gamma(\alpha_{gij})} \frac{\prod_{j=1}^{d} \Gamma(y_{ij} + \alpha_{gij})}{\Gamma(\sum_{j=1}^{d} y_{ij} + \alpha_{gij})} \\ &= \prod_{i=1}^{n} \sum_{g=1}^{G} \pi_g \frac{y_{i+}}{y_{i1}! \ldots y_{id}!} \frac{\Gamma(\alpha_{gi+})}{(y_{i+} - 1 + \alpha_{gi+}) \ldots \alpha_{gi+} \Gamma(\alpha_{gi+})} \prod_{j=1}^{d} [(y_{ij} - 1 + \alpha_{gij}) \ldots (\alpha_{gij})]^{c_{ij}} \\ &= \prod_{i=1}^{n} \sum_{g=1}^{G} \pi_g \frac{y_{i+}}{y_{i1}! \ldots y_{id}!} \frac{\prod_{j=1}^{d} [(y_{ij} - 1 + \alpha_{gij}) \ldots (\alpha_{gij})]^{c_{ij}}}{(y_{i+} - 1 + \alpha_{gi+}) \ldots \alpha_{gi+}}. \end{aligned}$$

Let z_{ig} be the unobserved variable where $z_{ig} = 1$ if observation i belongs to the g^{th} component, and $z_{ig} = 0$ otherwise. The complete data likelihood for $\boldsymbol{\pi}, \boldsymbol{\alpha}$ is:

$$L_c(\boldsymbol{\alpha}, \boldsymbol{\pi}) = \prod_{i=1}^{n} \prod_{g=1}^{G} [\pi_g f_{DM}(\mathbf{y}_i \boldsymbol{\alpha}_{gi})]^{z_{ig}}.$$

Complete data log-likelihood will be given by:

$$\begin{aligned}\ell_c(\boldsymbol{\alpha},\boldsymbol{\pi}) &= \prod_{i=1}^{n}\prod_{g=1}^{G} z_{ig}[\log(\pi_g)+\log(f_{DM}(\mathbf{y}_i,\boldsymbol{\alpha}_{gi}))] \\ &= \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}[\log(\pi_g)+\log\frac{y_{i+}}{y_{i1}!\ldots y_{id}!}+\sum_{j=1}^{d} c_{ij}\log((y_{ij}-1+\alpha_{gij})\ldots(\alpha_{gij})) \\ &\qquad\qquad -\log((y_{i+}-1+\alpha_{gi+})\ldots(\alpha_{gi+}))].\end{aligned}$$

Given intial value of $z_{ig}^{(0)}$ and ignoring constant terms, we maximize $\ell_c(\boldsymbol{\alpha},\boldsymbol{\pi})$ with respect to β_{gkj} in the M-step of the GEM algorithm.

$$\begin{aligned}\frac{\partial \ell_c}{\partial \beta_{gkj}} &= \sum_{i=1}^{n} z_{ig}^{(0)}\left[\frac{\partial\log(\pi_g)}{\partial\beta_{gkj}}+\frac{\partial\log(f_{DM}(\mathbf{y}_i\boldsymbol{\alpha}_{gi}))}{\partial\beta_{gjk}}\right] \\ &= \sum_{i=1}^{n} z_{ig}^{(0)}\left[\sum_{j=1}^{d} c_{ij}\left[\frac{\partial\log(y_{ij}-1+\alpha_{gij})}{\partial\beta_{gkj}}+\cdots+\frac{\partial\log(\alpha_{gij})}{\partial\beta_{gkj}}\right]\right. \\ &\qquad\left. -\left(\frac{\partial\log(y_{i+}-1+\alpha_{gi+})}{\partial\beta_{gk+}}+\cdots+\frac{\partial\log(\alpha_{gi+})}{\partial\beta_{gkj}}\right)\right] \\ &= \sum_{i=1}^{n} z_{ig}^{(0)}\left\{\alpha_{gij}x_{ik}\left[c_{ij}\sum_{l=0}^{y_{ij}-1}\frac{1}{\alpha_{gij}+l}-\sum_{l=1}^{y_{i+}-1}\frac{1}{\alpha_{gi+}+l}\right]\right\}\end{aligned}$$

$$\frac{\partial^2 \ell_c}{\partial\beta_{gkj}\partial\beta_{gk'j}} = \sum_{i=1}^{n} z_{ig}^{(0)}\left\{\alpha_{gij}x_{ik}x_{ik'}\left[c_{ij}\sum_{l=0}^{y_{ij}-1}\frac{1}{(\alpha_{gij}+l)^2}-\sum_{l=1}^{y_{i+}-1}\frac{\alpha_{gi+}-\alpha_{gij}+l}{(\alpha_{gi+}+l)^2}\right]\right\}$$

$$\frac{\partial^2 \ell_c}{\partial\beta_{gkj}\partial\beta_{gk'j'}} = \sum_{i=1}^{n} z_{ig}^{(0)}\left\{\alpha_{gij}\alpha_{gij'}x_{ik}x_{ik'}\left[\sum_{l=1}^{y_{i+}-1}\frac{1}{(\alpha_{gi+}+l)^2}\right]\right\}$$

In summary, in the M-step we find the MLE of $\boldsymbol{\beta}=(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_G)$, where $\boldsymbol{\beta}_g$

is matrix of coefficients:

$$\boldsymbol{\beta}_g = \begin{bmatrix} \beta_{01} & \beta_{02} & \dots & \beta_{0d} \\ \beta_{11} & \beta_{12} & \dots & \beta_{1d} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{p1} & \beta_{p2} & \dots & \beta_{pd} \end{bmatrix}.$$

And the score function vector is of the form:

$$S(\boldsymbol{\beta}) = \left(\frac{\partial \ell_c}{\partial \beta_{101}}, \frac{\partial \ell_c}{\partial \beta_{111}}, \dots, \frac{\partial \ell_c}{\partial \beta_{1p1}}, \frac{\partial \ell_c}{\partial \beta_{102}}, \frac{\partial \ell_c}{\partial \beta_{112}}, \dots, \frac{\partial \ell_c}{\partial \beta_{1p2}}, \dots, \frac{\partial \ell_c}{\partial \beta_{10d}}, \frac{\partial \ell_c}{\partial \beta_{11d}}, \dots, \right.$$
$$\left. \frac{\partial \ell_c}{\partial \beta_{1pd}}, \frac{\partial \ell_c}{\partial \beta_{201}}, \dots, \frac{\partial \ell_c}{\partial \beta_{2pd}}, \dots, \frac{\partial \ell_c}{\partial \beta_{G01}}, \dots, \frac{\partial \ell_c}{\partial \beta_{Gpd}} \right),$$

$$S(\boldsymbol{\beta}) = \begin{bmatrix} \sum_{i=1}^{n} z_{i1}^{(0)} \alpha_{1i1} \left[c_{i1} \sum_{l=0}^{y_{i1}-1} \frac{1}{\alpha_{1i1}+l} - \sum_{l=1}^{y_{i+}-1} \frac{1}{\alpha_{1i+}+l} \right] x_{i0} \\ \vdots \\ \sum_{i=1}^{n} z_{i1}^{(0)} \alpha_{1i1} \left[c_{i1} \sum_{l=0}^{y_{i1}-1} \frac{1}{\alpha_{1i1}+l} - \sum_{l=1}^{y_{i+}-1} \frac{1}{\alpha_{1i+}+l} \right] x_{ip} \\ \vdots \\ \sum_{i=1}^{n} z_{ig}^{(0)} \alpha_{gij} \left[c_{ij} \sum_{l=0}^{y_{ij}-1} \frac{1}{\alpha_{gij}+l} - \sum_{l=1}^{y_{i+}-1} \frac{1}{\alpha_{gi+}+l} \right] x_{ik} \\ \vdots \\ \sum_{i=1}^{n} z_{iG}^{(0)} \alpha_{Gid} \left[c_{id} \sum_{l=0}^{y_{id}-1} \frac{1}{\alpha_{Gid}+l} - \sum_{l=1}^{y_{i+}-1} \frac{1}{\alpha_{Gi+}+l} \right] x_{ip} \end{bmatrix}.$$

The information matrix $I(\boldsymbol{\beta})$ can be written as:

$$I(\boldsymbol{\beta}) = \begin{bmatrix} I_1(\boldsymbol{\beta}) & & & \\ & \ddots & & \\ & & I_g(\boldsymbol{\beta}) & \\ & & & \ddots \\ & & & & I_G(\boldsymbol{\beta}) \end{bmatrix},$$

where $I_g(\boldsymbol{\beta})$ is composed of entries:

$$I_g(\boldsymbol{\beta})_{kj,k'j'} = -\frac{\partial^2 \ell_c}{\partial \beta_{gkj} \partial \beta_{gk'j'}}$$
$$= \begin{cases} -\sum_{i=1}^{n} z_{ig}^{(0)} \alpha_{gij} x_{ik} x_{ik'} \left[c_{ij} \sum_{l=0}^{y_{ij}-1} \frac{1}{(\alpha_{gij}+l)^2} - \sum_{l=1}^{y_{i+}-1} \frac{\alpha_{gi+}-\alpha_{gij}+l}{(\alpha_{gi+}+l)^2} \right] & \text{for } j = j' \\ -\sum_{i=1}^{n} z_{ig}^{(0)} \alpha_{gij} \alpha_{gij'} x_{ik} x_{ik'} \left[\sum_{l=1}^{y_{i+}-1} \frac{1}{(\alpha_{gi+}+l)^2} \right] & \text{for } j \neq j' \end{cases}$$

In order to find $\hat{\boldsymbol{\beta}}^{(1)}$, the next iteration in the NR method using an initial estimate $\hat{\boldsymbol{\beta}}^{(0)}$ will have the form:

$$\hat{\boldsymbol{\beta}}^{(1)} = \hat{\boldsymbol{\beta}}^{(0)} + S\left(\hat{\boldsymbol{\beta}}^{(0)}\right) \left[I\left(\hat{\boldsymbol{\beta}}^{(0)}\right) \right]^{-1},$$

and general update will be:

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} + S\left(\hat{\boldsymbol{\beta}}^{(t)}\right) \left[I\left(\hat{\boldsymbol{\beta}}^{(t)}\right) \right]^{-1}.$$