Mixed models for correlated compositional data: applied to microbiome studies in Indonesia

Martin, I.

1 Introduction

1.1 The human gut microbiome

For more than a decade, studying human microbes and their collective genome (termed as microbiome) has became an additional method to study human function beyond protein coding genes [Grice and Segre (2012); Yadav et al. (2018)]. The whole human body is inhabited by microbes with the gastro-intestinal tract harboring the most abundant and diverse species [Ley et al. (2006); Gupta et al. (2017)]. The bacterial community interacts with the host and contributes to processes varying from metabolic homeostasis [Yadav et al. (2018)] to the development of the immune system [Gensollen et al. (2016); Thursby and Juge (2017)]. Hence disruption of the microbial community is linked to development of a variety of diseases, for instance obesity and several metabolic disorders [Sonnenburg and Bäckhed (2016)]. With the advent of high-throughput sequencing technologies, the goal of microbiome studies has shifted from mapping and cataloguing genes related to bacteria to characterizing the microbial community in relation to health and diseases [Rodrigues Hoffmann et al. (2016)]. However, these technologies have not been accompanied by the development of statistical tools necessary to analyze the data generated. This thesis presents epidemiological work related to human gut microbiome, accompanied by the development of statistical methods to analyze them. The remaining of this introduction section provides information on available features of microbiome data, microbiome related epidemiological studies and statistical methodologies.

The microbial data analyzed in this thesis were obtained using high-throughput sequencing technology. This technology worked by targeting the specific region of 16S rRNA gene that is unique to bacteria. Robinson et al. (2016) have reviewed the whole process of sequencing the 16S rRNA gene to extract the microbial data. The process begins by Polymerase Chain Reaction (PCR) amplification of these rRNA genes, followed by sequencing of the PCR products and alignment to a reference database, for instance the Ribosomal Database Project (RDP) [Cole et al. (2014)]. The sequence reads are then clustered into operational taxonomical units (OTU) which is usually based on 97% similarity (a proxy of species) [Mysara et al. (2017)]. The procedure then continues with the taxonomy annotation in which the dataset can be viewed in the format shown in Table 1.1 where rows represent samples and columns represent taxa. The total reads for each sample are usually different due to technical difficulties in loading the same molar amount to an instrument [Gloor et al. (2017)]. Usually a normalization or rarefaction is done to obtain the same number of total reads [Weiss et al. (2017)].

Taxon **Samples**	**1**	**2**	$\dots$	$\boldsymbol{J}$	**Total reads**
1	c_{11}	c_{12}	$\dots$	c_{1J}	c_{1+}
2	c_{21}	c_{22}	$\dots$	c_{2J}	c_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$\boldsymbol{N}$	c_{N1}	c_{N2}	$\dots$	c_{NJ}	c_{N+}

Table 1.1: The format of the taxonomical count of microbiome data.

As a result of high-throughput sequencing, several important features of the microbiome data can be highlighted. First of all, it is a compositional data [Weiss et al. (2017)]; it consists of multiple proportions of various organisms that sums up to a constant [Gloor et al. (2017)]. The total reads per sample are up to the number of molar concentration loaded into the instrument. Thus, it is not possible to assume independence between different taxa as decreasing number of one bacterial taxa increases the other, or vice versa. Secondly, microbiome data vary highly due to unknown reasons [Turnbaugh et al. (2007)]. This could be due to sampling or individual heterogeneity, which needs to be taken into account in the statistical model. Finally, a variation in microbial taxa could occur as a result of measurement error [Li (2015)].

1.2 Soil-transmitted helminthiasis and immunity

According to the World Health Organization, approximately 1.5 billion of people are infected with soil-transmitted helminths worldwide [Collender et al. (2015)], majority species of *Ascaris lumbricoides*, *Necator americanus*, *Ancylostoma duodenale* and *Trichuris trichiura*. Infected individuals live mostly in low and middle income countries. Due to poor level of hygiene and sanitation, these parasites enter through the skin or through orofecal route and reach the gastro-intestinal tract, which then cause malnutrition, growth stunting and physical impairment [Crompton and Nesheim (2002); Hall et al. (2008); Albonico et al. (2008)]. Aside from this negative impact, there seems also a positive impact; helminths are associated with lower incidence of metabolic disorders [Wiria et al. (2012)]. This may be related to the ability of helminths to modulate the host's immune system [Wiria et al. (2012); McSorley and Maizels (2012); Gazzinelli-Guimaraes and Nutman (2018)] or due to their parasitism, consuming energy from their host [Wiria et al. (2012)].

Given the fact that helminths reside in the same niche, it has been hypothesized that gut microbiota and the helminth parasites may interact and thus modulate the immune system (reviewed in Leung et al. (2018)). This three-way relationship has been largely analyzed in laboratory animal models, and less in humans [Wegener Parfrey et al. (2017); Reynolds et al. (2015)].

1.3 The randomized controlled trial in a repeated measurement setting

The randomized controlled trial design is ideal to analyze the causal effect of treatment on outcomes of interest [Hernán and Robins (2006a); Hernan and Robins (2018)]. The dataset used in this thesis was obtained from a household-based cluster-randomized, double-blind, placebo-controlled trial conducted in an area endemic for helminth infections in Indonesia with the main purpose of analyzing the association between immunological responses and helminth infections [Wiria et al. (2010)]. Briefly, irrespective of their infection status, subjects were randomized into treatment or placebo arm. An anthelminthic treatment was administered every three months for a total period of 21 months. The stool samples were collected at two different time-points, namely before and 21 months after the first treatment to identify helminth infection status and collect microbial data. The outcomes modelled in this thesis are the microbiome composition and immune responses. Using this randomized treatment design, the causal effect of anthelminthic treatment on helminth infections, microbiome composition and immune responses is assessed.

In addition, the longitudinal design enables one to study the dynamic association between helminth infections and outcomes of interests over time. At 21 months after the first treatment, irrespective of the treatment allocation, subjects may either get a new infection, remain infected, or may be helminth-uninfected. Specifically, this design enables one to take into account gut microbiome changes over time due to changing lifestyles. The challenge in the statistical analyses of data from a repeated measurements design is the fact that observations from the same individuals are correlated. For valid inference these correlations have to be addressed. The following sections introduce statistical methodologies for correlated data.

1.4 Methodology of compositional data

The sequencing process of 16S rRNA to obtain the microbiome data has been described earlier and the format of the dataset used in the analysis has been tabulated in Table 1.1. Each sample has counts of sequence reads that are clustered in multiple categories with arbitrary total reads imposed by the instrument. As it has been described above, the important feature of this data is the compositional structure, in which counts for each category cannot be considered as an independent realization. Thus, it is important to analyze this data using a multivariate approach.

Typically, analyses in microbiome studies aim to characterize the relationship between the microbiome composition and biological, clinical or environmental features [Xia and Sun (2017)]. For this purpose, either a nonparametric or parametric approach can be used. For instance, nonparametric ecological distances or dissimilarity measures, such as alpha and beta diversity measures are commonly used for comparisons between groups. By doing so, no parameters are estimated. Note that while no specific distribution needs to be assumed for the outcome variable, the hypothesis testing may still have assumptions which might be violated. For instance, the assumption that observations are obtained from the distribution that has the same shape and are independent [Lumley et al. (2002)]. In addition, for small sample sizes, the statistical power might be limited to detect differences. The parametric approach may provide a solution for these issues. In this thesis, the parametric approach is deployed because: the interest is in modelling complex relationships in a relatively small study.

When considering the microbiome data in Table 1.1, each cell's entry c_{ij} represents the count for subject i belonging to taxa j. Count data are typically modelled using the Poisson distribution. When considering multivariate count data with a fixed total count, the multinomial distribution is used. These parametric distributions have a restricted assumption, namely the variances of the responses are specified by the means. This appears not to hold for microbiome data. As it has been pointed out in Section 1.1, microbiome data vary highly due to presence of many zeros, outliers or heterogeneity caused by sampling mechanisms and differences between individuals. An extra variation is thus observed, which is known as overdispersion.

To account for overdispersion in modelling multivariate count data, a compound distribution is used, i.e., when observations are drawn from multinomial or Poisson distribution, its mean parameters are assumed to be random variable following a specific distribution. Two often used distributions are the conjugate distribution (Gamma for Poisson and Dirichlet for multinomial) and the normal distribution. The advantage of using the conjugate distribution is that the marginal distribution has a closed form formula which consequently reduces computational time and directly take the measurement error into account [Li (2015)]. However, the drawback of this model is the restriction of the number of parameters to model the covariance structure [Li (2015)]. As an alternative, multivariate normally distributed random effects can be used [Hartzel et al. (2016); Hedeker (2003)]. The trade off between these options is in the computational burden.

Finally, as studies in this thesis were done in a repeated measurement setting, the compositional data are observed at two time-points. Hence, a correlation structure is imposed on the data by multiple sources, namely the correlation between different bacterial categories at the same time-point and the correlation within the same categories at different time-points.

1.5 General mixed models

In a repeated measurement setting, one is interested in analyzing the progression of outcomes over time in relation to predictors. In the simplest case, this outcome-predictor relationship is assumed to be linear, and linear regression can be used. However, it should be noted that under a repeated measurement setting, observations are not independent since observations from the same subjects are likely to be more similar than observations from others. Thus the variability in the observations could be due to variability of the observations within subjects and between subjects. The formulation of this linear regression needs to be extended to account for this extra variation.

Starting from linear model context, it is assumed that for subject i, a n_i numbers of observations are collected $\boldsymbol{Y}_i = \{Y_{i1}, \ldots, Y_{in_i}\}$. The observations from subject i, $\boldsymbol{Y}_i$ are modelled as individual response trajectory $\boldsymbol{\mu}_i$ and an independent residual term $\boldsymbol{e}_i$.

$$\boldsymbol{Y}_i = \boldsymbol{\mu}_i + \boldsymbol{e}_i \tag{1.1}$$

with $\boldsymbol{e}_i$ is a random variable assumed to follow the multivariate normal distribution with zero mean and variance of $\Sigma_e = \sigma_e^2 \mathbf{I}_{n_i}$ (denoted as $\boldsymbol{e}_i \sim \text{MVN}(\mathbf{0}, \Sigma_e)$). The profile response $\boldsymbol{\mu}_i$ is then linked to the design matrix $\boldsymbol{X}_i$ via the following equation

$$\boldsymbol{\mu}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{A}_i\boldsymbol{b}_i, \tag{1.2}$$

in which $\boldsymbol{X}_i$ is the design matrix for the fixed effect part (population) and $\boldsymbol{A}_i$ is the design matrix for the random effect part (subject-specific trajectory). The parameter $\boldsymbol{\beta}$ represents the population part and $b_i \sim \text{N}\left\{0, \sigma_b^2\right\}$ are the subject-specific effect. This is known as a linear mixed model which consists of a fixed population parameter β and a subject-specific effect b_i [Laird and Ware (1982)]. The estimate of the regression parameters $\boldsymbol{\beta}$ and subject-specific variability σ_b^2 are obtained by maximizing the full likelihood which has a closed form formula.

The above formulation of the model is for the continuous outcome, where the response conditional on the covariates follows the normal distribution. For other outcomes, the generalized linear model needs to be extended to account for subject-specific effects [Molenberghs and Verbeke (2005)]. The expected value of the responses $\boldsymbol{\mu}_i$ in equation (1.1) is linked with the predictor via the so-called link function g. For instance, in the case of a count response, the function of its expected value μ_i becomes

$$g(\mu_i) = \log(\mu_i) = \boldsymbol{X}_i\boldsymbol{\beta}_i$$

in which $\boldsymbol{\beta}_i$ is defined as in equation (1.1) and the logarithm function is used as a link function. For the multinomial distribution often used for microbiome data, the logit link function is used.

1.6 Outline of the thesis

This thesis is a collection of five articles which are published or submitted for publication. It is organized in two major parts. Each of this part consists of an epidemiological and a methodological paper to aid in understanding the underlying biological mechanisms which motivate development of the statistical method in the topic related to microbiome and immunity in relation to helminth infections. The last chapter discusses all results of these two parts.

The focus in **Part I** is modelling the association between helminth infections and the gut microbiome composition in a randomized-controlled trial setting. In **Chapter 2**, an advanced statistical method was utilized to analyze this association while addressing the multivariate structure of the compositional data. The correlation between bacterial taxa is accounted for by introducing a conjugate distributed random effect. This method does not account for the correlation between observations. As the data were repeatedly measured, the correlation of the observation between different time points was left unaccounted for in the statistical model.

In **Chapter 3**, a statistical framework to analyze microbiome data was developed to account for the additional correlation due to repeated measurements. We extended the Dirichlet - multinomial regression model used in **Chapter 2** to include an extra normally distributed random effect. Several covariance structures for the normally distributed random effects were considered. The conjugate distributed random effect serves as an estimate for the correlation between categories. In addition, a loglinear model approach is introduced to aid interpretation of the regression estimates.

It has been described earlier in this chapter that there is a hypothesized three-way relationship between helminth infection, gut microbiome and human immune response. Thus, the aim of **Part II** of this thesis is to analyze these complex relationships. In **Chapter 4**, the linear mixed model with subject-specific random effect is utilized to study the role of gut microbiome on human immune response in the presence or absence of helminth infections. Stimulated cytokine responses were considered as a marker of immune response while bacterial proportion and helminth infection were included as covariates. Since helminth infection might be a confounder for both microbiome and cytokine responses, it was included as covariate in the model, however the relationship between helminth infection and microbiome composition was not modelled. In addition, the presence of measurement error in the microbiome data was not taken into account in the model.

In **Chapter 5** of this thesis, the challenges in modelling the relationships between helminth infection, gut microbiome and immune response were addressed by using a joint model. The correlation between observation of the same subject as well as the measurement error present in the microbiome sequencing data were taken into account. A flexible covariance structure was used for this purpose, namely by including multivariate normally distributed random effects in the model. The proposed mixed regression model is able to model the association between helminth infection on both outcomes and the association between the different outcomes, while accounting for the correlation over time-points and between bacterial categories. Finally, **Chapter 6** discusses the relationship between helminth infections, gut microbiome and cytokine responses by using the directed acyclic graph. Using results from our and other published studies, the potential biases of estimated associations are identified and discussed. Possible solutions are presented.