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

Martin, I.

Author: Martin, I.
Title: Mixed models for correlated compositional data: applied to microbiome studies in Indonesia
Issue Date: 2019-10-08

Summary

Rapid urbanization is almost always accompanied by a transition from infectious diseases to noncommunicable inflammatory disorders as a dominant cause of morbidity. This is likely due to changing lifestyle and environmental factors. To understand a precise mechanism of this transition, a population study was done in Nangapanda, Ende district in Indonesia. This area was chosen as chronic parasitic worm infections were endemic and lifestyle changes occurred at a rapid pace. Despite its detrimental effect on human health, parasitic helminth infections are associated with a strong modulation of immune responses which explains a low prevalence of inflammatory disorders in areas endemic for parasitic worms. To investigate the complex biological mechanism underlying this association, clinical and biomedical data were collected from subjects using a household-based cluster-randomized, double-blind, placebo-controlled trial. Specifically, studies in this thesis focus on analyzing the relationships between helminth infections, gut microbiome composition, and immune responses. Considering the complexities of the data gathered in this study, the available methods appeared to be limited, hence the development of statistical methodology is another focus of this thesis.

Chapter 1 provides a general introduction to the thesis with regard to the collected data, the research questions and the available statistical methods for the analysis purpose. The pyrosequencing procedure to obtain microbiome profiles for each sample is briefly described. Such a process imposes a compositional structure on the microbiome data which needs to be accounted for in the modeling. In addition, multiple observations from the same subject were collected, which yields a correlation structure between measurements. Statistical tools used to analyze microbiome data are reviewed and challenges for proper modeling of this type of data in a repeated measurement design are discussed.

Chapter 2 describes the application of a recently developed statistical method to model the microbiome data collected for this study. This model assumes that microbiome data are realizations of a multinomial distribution. In order to account for the presence of extra variation, the parameters of this multinomial distribution are assumed to be random effects following a conjugate distribution. However, the method is only valid for independent multinomial observations

and thus the correlated structure in our data due to repeated measurements is left unaccounted for in the modeling. Therefore, the analyses were carried out at each time point separately. The effect of helminth infection on the gut microbiome composition is analyzed using the data at pre-treatment while the effect of anthelminthic treatment on microbiome composition is assessed at post-treatment. To investigate whether the treatment has a different effect on subjects who were helminth-infected compared to helminth-uninfected, an interaction term between infection status and treatment is included in the model. It appears that only in subjects who received anthelminthic treatment and remained infected at both pre- and post-treatment, the ratio of Bacteroidetes to Firmicutes and the ratio of Actinobacteria to Firmicutes significantly differed compared to other groups. The method here is limited to the analysis of data from one time point, hence the alteration of microbiome composition over time cannot be analyzed. Chapter 3 attempts to develop a model to address this.

In Chapter 3, a statistical method for modeling repeatedly measured microbiome data is developed which addresses the correlation structure between a subject's multiple observations. This is done by introducing a normally distributed random effect. Three different covariance structures for the normally distributed random effects are considered. Firstly, we assume a univariate subject-specific random effect where the random effect for each bacterial category at different time points is the same. Secondly, each category has a different random effect with category-specific variance. Finally, it is assumed that the multivariate random effects have a common variance for all categories. A simulation study was conducted to investigate the performance of the proposed method in estimating the fixed effects and standard deviations of the random effects. It appeared that the estimates of the fixed effects are not affected by the choice of the covariance structure of the normally distributed random effect. For our application, the conclusion based on the analysis in Chapter 2 with regard to the fixed effect is confirmed, i.e. subjects who were infected at baseline and remained infected at post-treatment showed also an alteration in their Bacteroidetes to Firmicutes ratio in our extended model. To assess model fit, we computed the marginal correlations between and within categories over time. It appears that the marginal correlation in the data is well captured using the model with a multivariate random effect having a common variance for all categories.

In the next two chapters, the interplay between helminth infection, the gut microbiome composition, and immune responses which were characterized by the whole blood cytokine responses to antigens is studied. It is known that the removal of helminth infection by anthelminthic treatment restores immune responsiveness. It is also hypothesized that certain gut bacteria influences immune responses. Our aim in Chapter 4 is to gain insights into the mechanism underlying this interplay using the observations at both time points. A linear mixed model is applied to the data with a cytokine response to specific antigen as an outcome variable. For this model, the predictors are bacterial proportion or diversity and their interaction term with helminth infection status. We restricted our analysis to three bacterial categories, namely Actinobacteria, Bacteroidetes and Firmicutes as these bacterial phyla were associated with helminth infection in the analysis of Chapter 2 and 3.

In this study, we observed that a gain in the proportion of Bacteroidetes is significantly associated with a decrease in concentration of IL-10 to LPS in helminth-uninfected subjects. This association is dampened in helminth-infected subjects. This finding confirms the hypothesized relationship that the removal of helminth infections restores immune responsiveness and that gut bacteria influences immune responses. Several limitations of this analysis can be noted: each bacterial proportion is assumed to be independent and its association with cytokine responses is analyzed separately. Thus, the compositional feature of microbiome data is ignored. In addition, the measurement error of the microbiome data is left unaccounted for in this model which potentially leads to biased estimators of the regression parameters. Furthermore, the association between helminth infection and bacterial proportion is not quantified in this model. Therefore, a statistical method developed in Chapter 5 which attempts to address these limitations.

In Chapter 5, our aim is to build a statistical model for the association between helminth infections and both microbiome and cytokine responses simultaneously by considering all sources of variability in the data. First of all, cytokine responses as continuous outcomes and gut microbiome as multivariate count outcomes observed from the same individuals are correlated. Secondly, the correlation between the same type of observations at different time points is expected. Finally, specific to microbiome data, there is an additional variability due to overdispersion and measurement error. The cytokine response and the microbiome composition are assumed to follow a normal and a multinomial distribution, respectively. A set of latent variables which is assumed to follow a multivariate normal distribution is incorporated to account for the additional variabilities in the data. The measurement error is modeled with multidimensional normally distributed random effect, i.e., each category has a different random effect which is assumed to be correlated. This is done to allow for more flexibility in modeling the extra variation in each category. The joint probability distribution is formulated and parameters are estimated by maximizing the joint likelihood with numerical quadratures. A simulation study is carried out to investigate the performance of the estimator for the fixed effects as well as random effect parameters in comparison with the method introduced in Chapter 4 (the naive method). The joint model outperforms the naive method. In the data application, it is shown that the correlation between microbiome composition and cytokine responses are small. When analyzing the marginal correlation using the proposed method, it appears that the marginal correlation does not fit to the observed one.

Chapter 6 summarizes and describes the results of the analyses performed and the statistical methods used in Chapter 2 to 5. The aim of this chapter is to evaluate the evidence for causality of the identified associations among helminth infection, the gut microbiome composition, and cytokine responses using data from the randomized controlled trial. We found that treatment has an effect on both the gut microbiome composition and cytokine responses via removing helminth infection and that the gut microbiome has a direct effect on cytokine responses. The directed acyclic graphs (DAGs) are used to visualize the direction of the causal effect and several potential sources of biases are included in this DAGs, namely unobserved confounders and measurement errors. The statistical methods developed in this thesis account for the additional variation due to unobserved confounders and measurement errors via the inclusion of the random effect. Specific to microbiome data, there are two possibilities of distributional assumption for a random effect, namely using the conjugate and normally distributed. In this thesis, we have explored these assumptions. It appears that models with random effects having a conjugate distribution fit the microbiome data well when considering how well its marginal correlation capture the observed correlation. Using the findings from the literature as well as from our analyses, we conclude that treatment has a causal effect on helminth infection and that helminth infection has direct effects on both the gut microbiome and the cytokine responses. It appears that the correlation between the gut microbiome and cytokine responses is small, hence the evaluation of their effect is not carried out.

Finally, data from randomized controlled trials, as is the case in this thesis are beneficial to examine causal relationships between variables involved. Furthermore, observations which are repeatedly measured provide information on how a specific outcome evolves over time. Unfortunately, the studies in this thesis use data from small subsamples from the larger trial which possibly decreases the statistical power to detect effects. One solution to address this limitation is by integrating different sources of observation, as we did in Chapter 5.