Research Project Updates

Emily Palmer

Oregon State University

palmerem@oregonstate.edu

November 1, 2021

Goals for this presentation

- Introduce current project
- Get feedback on approach.
- Broad research question: Using GEEs, how can we specify a reasonable mean and variance function and working correlation matrix to model microbiome data

Ideas from Masters Project

For my masters project I used the MLTC model using Generalized Estimating Equations (GEEs) for longitudinal data assumed to be taxonomically correlated

- Assume that observations that share the same taxonomic group (at some level) have the same correlation based on the taxonomic tree. Combine with any distinct correlations from longitudinal measures.
- Global test for association between response (OTU counts/proportions) and covariates (eg disease status).
- Two part model modeling presence/absence and log-transformed abundance separately

Issues with previous model - Moving forward

- Previous method is computationally infeasible
- Challenge of picking what taxa level to aggregate to: tradeoff of how many models to run/interpretability - arbitrary
- Compositionally is not considered - interpretations challenging
- Used taxonomy instead of phylogeny instead of - taxonomy may be too coarse
- The correlation matrix may not be positive semi-definite

GEE Framework

- Consider $i = 1, \ldots, n$ samples/individuals, each with $j = 1, \ldots, p$ OTU/taxa observations. The response vector for the ith subject is $\mathrm{Y}_i = (Y_{i1}, \ldots, Y_{ip})^T$. Denote $E(Y_{ij}) = \mu_{ij}$
- The collection of observations for a sample make up a cluster. Responses within a cluster are not independent, responses between clusters are independent.
- Consider covariates $\mathrm{X}_i = (X_{i1}, \ldots, X_{iq})$ for each sample. Usually covariates have the same value for all p OTU observations in the ith subject.
- Use a link function between the covariates and mean response as $g(\mu_{ij}) = x_{ij}^T \beta$
- Variance is a function of the mean. $\mathrm{Var}(Y_{ij}) = \phi a(\mu_{ij})$

GEE Equations

- β is the solution to

$$\sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right)^T \mathrm{V}_i^{-1} (\mathrm{Y}_i - \boldsymbol{\mu}_i) = 0$$

- Where $\mathrm{V}_i = \mathrm{A}_i^{1/2} \mathrm{R} \mathrm{A}_i^{1/2}$
- A_i is a diagonal matrix whose entries are the variances
- R is the **working correlation matrix**

GEE Algorithm

Initialize β, ϕ, R

- Update μ, e_{ik}, ϕ based on current values of β
- Update R, R^{-1} using ϕ, e_{ik}
- Update β iteratively using estimating equations and current values of R, μ, ϕ, e_{ik}

Repeat until convergence

What to modify

- How to specify the working correlation matrix?
- What leads to a lack of independence?
  - Phylogenetic correlation?
  - Compositionality
  - (Longitudinal repeated measures - incorporate eventually)
- What mean and variance function to use?
  - Previously a two-part model - binomial + log transformed gaussian
  - Now - Dirichlet

Evolutionary trait model

- We borrow the idea of the evolutionary trait model (Martins and Hansen 1997) used in Microbiome data models (Xiao et al 2018)
- From a phylogenetic tree, create matrix D where d_{ij} is the distance between OTU i and j
- Use patristic distance - length of the shortest path
- Correlation between OTU j and k is

$$Cor(Y_{ij}, Y_{ik}) = C_{jk}(\rho) = e^{-2\rho d_{jk}}$$

  Where $\rho \in (0, \infty)$ and needs to be estimated. If ρ is small, C_{jk} is close to 1 indicating high correlation. If ρ is large, indicates no correlation.
- Interpretation of ρ: depth of the phylogenetic tree where groups are formed.

Potential issues

- Imposes a positive correlation requirement
- Assumes all OTUs that have the same distance are correlated the same
- Does not address compositionally

How to fix?

Add a sign term

- One potential fix is to add a sign term to allow for negative correlations
- Now let

$$\text{Corr}(Y_{ij}, Y_{ik}) = c_{ij} e^{-2\rho d_{ij}}$$

  where $c_{ij} \in \{-1, 1\}$
- How to set c_{ij}?

Parameter estimation for ρ and c_{ij}

- In the GEE algorithm, each iteration has a step to estimate R equivalent to estimating ρ and c_{ij} using the current value of $\boldsymbol{\beta}$ in the iteration.
- One approach is to use OLS to find ρ such that $\log e_{ij} e_{ik} = -2\rho d_{jk}$, where e_{ij} is the Pearson residual. However, e_{ij} can be negative. Instead take the absolute value: $\log |e_{ij} e_{ik}| = -2\rho d_{jk}$
- Another approach is to use the residuals from all subjects.

$$\left| \frac{1}{n} \sum_{i=1}^{n} e_{ij} e_{ik} \right| = e^{-2\rho d_{jk}}$$

- Let $c_{jk} = \text{sign}\left(\sum_{i=1}^{n} e_{ij} e_{ik}\right)$
- Need to use constrained optimization to ensure that $\rho > 0$.

Mean and variance function

- If the responses are proportions, we use the same mean and variance function as those from the Dirichlet$(\alpha_1, \ldots, \alpha_p)$ distribution

$$\mu_{ij} = E(Y_{ij}) = \frac{\alpha_{ij}}{\alpha_{i0}}$$

Where $\sum_{j=1}^{p} \alpha_{ij} = \alpha_{i0}$

$$V(Y_{ij}) = a(\mu_{ij}) = \frac{\alpha_{ij}(\alpha_{i0} - \alpha_{ij})}{\alpha_{i0}^2(\alpha_{i0} + 1)}$$

- Using the Dirichlet distribution we account for $\sum_{j=1}^{p} Y_{ij} = 1$
- Link using

$$\log \alpha_i = X_i \beta$$

Correlation idea from Dirichlet distribution

Dirichlet random variables have a correlation that describes dependence from the constant sum constraint of the components of the cluster.

$$Corr(Y_{ij}, Y_{ik})_{\text{Dir}} = -\frac{\alpha_{ij}\alpha_{ik}}{\alpha_{i0}^2(\alpha_{i0}+1)} \frac{1}{\sqrt{V(Y_{ij})V(Y_{ik})}}$$

This will be negative.

Compositional and Phylogenetic Correlations

- Model the correlations as a mixture of the phylogenetic correlation and compositional correlation. We will combine this covariance that arises from the Dirichlet distribution for compositional dependence.
- $$R = \omega R_{ETM} + (1 - \omega) R_{Dir}$$

  Where $\omega \in [0, 1]$
- Interpretation if compositionality or phylogenetic structure is the leading cause of correlation between responses.
- Is the sign term c_{jk} still needed?
- We now also need to estimate ω

Real data

- iHMP-IBD data
- Does metagenomic data generate a phylogenetic tree?
- Explore the data to see if this is a reasonable approach to model.

Thank you!