Spring research update

Emily Palmer

Oregon State University

palmerem@oregonstate.edu

April 10, 2022

Approach

- Use Generalized estimating equations to estimate regression parameters and covariance parameters for microbiome relative abundances
- Assume the mean and variance follow those of the Dirichlet distribution
- Assume the correlation structure between ASVs in a sample depends both on compositionality and phylogenetic similarity

Notation and setup

- Let y_{ij} be the relative abundance of the jth ASV in the ith sample. $i = 1, \ldots, n, j = 1, \ldots, p$
- Assume that $E(y_{ij}) = \mu_{ij} = \frac{\alpha_{ij}}{\alpha_{i0}}$ where $\alpha_{i0} = \sum_{j=1}^{p} \alpha_{i0}$. and α_i are the parameters of if $\mathrm{y}_i \sim \mathrm{Dirichlet}(\alpha_1, \ldots, \alpha_p)$
- Link function: Link covarites to α's:

$$\log(\alpha_{ij}) = \mathrm{x_i}^T \boldsymbol{\beta}_j$$

Compositional Dirichlet Correlation

- Since ASVs are in relative abundances, we believe there is a negative correlation arising from compositionality.
- Dirichlet correlation: for $j \neq k$

$$R_{i,D} = Cor(y_{ij}, y_{ik}) = -\frac{\alpha_{ij}\alpha_{ik}}{\alpha_{i0}^2(\alpha_{i0}+1)\sqrt{V(y_{ij})V(y_{ik})}}$$

$$V_{ij} = \frac{\alpha_{ij}(\alpha_{i0}-\alpha_{ij})}{\alpha_{i0}^2(\alpha_{i0}+1)}$$

GEE Algorithm

- Need to estimate β, ρ, ω
- Fisher scoring algorithm
- Alternate by using current values of ρ and ω (And thus R_i) to find β and using current value of β (and thus α, ϵ) to calculate ρ, ω

Evolutionary Trait Correlation

- 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 ASV i and j.
- Use patristic distance - length of the shortest path.
- Correlation between ASV j and k is

  $$R_{i,ETM} = Cor(y_{ij}, y_{ik}) = e^{-2\rho d_{jk}}$$

  Where $\rho > 0$ and needs to be estimated.
- If ρ is small, R_{ijk} 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.

Working Correlation matrix

- Use weighted sum of Dirichlet compositional correlation and evolutionary trait model correlation

$$R = \omega R_{Dir} + (1 - \omega) R_{ETM}$$

GEE Equations

The GEE equations are

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

- Where $V_i = \frac{1}{\phi} A_i^{\frac{1}{2}} R_i A_i^{\frac{1}{2}}$
- $A_i = diag(V(Y_{ij}))$
- $\left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right)^t = \frac{1}{\alpha_{i0}^2} (\alpha_{i0} \mathrm{diag}(\alpha_i) - \alpha_i \alpha_i^t) \otimes X_i$

GEE Algorithm - ρ, ω, ϕ step

GEE algorithm goes between steps for estimating ρ, ω, and ϕ and step for estimating β.

- $e_{ij} = y_{ij} - \mu_{ij}$
- $\phi = \left(\frac{1}{n*p-(p*q-1)} \sum_{i=1}^{n} \sum_{j=1}^{p} e_{ij}^2 \right)^{-1}$
- ω, ρ minimize

$$\sum (\phi e_{ij} e_{ik} - [\omega R_{ijk,D} + (1-\omega) e^{-2\rho D_{jk}}])^2$$

  suject to $0 \leq \omega \leq 1, \rho > 0$

- Given ρ, ω, working correlation R is specified.

GEE Calculate ω, ρ

- Use "L-BFGS-B" algorithm (Byrd et. al. (1995))
- Option in `optim()` that allows bounds to be specified $0 \le \omega \le 1, \rho > 0$
- This uses a limited-memory modification of the BFGS quasi-Newton method.

GEE Algorithm - β step

$$G = \sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right)^t \mathrm{V}_i^{-1} (\mathrm{Y_i} - \boldsymbol{\mu}_i)$$

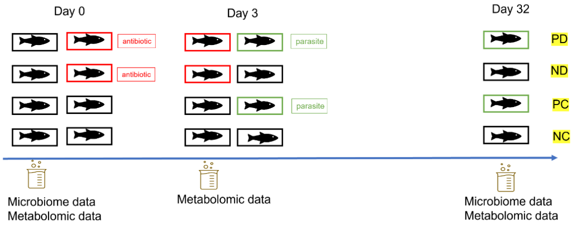
$$H = -\sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right)^t V_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} + \lambda I$$

$$\hat{\beta}^{(s+1)} = \hat{\beta}^{(s)} - \gamma H^{-1} G$$

β-step considerations

- H is or is close to singular. Add a small diagonal constant. λ (depending on eigenvalues of H)
- Appears that steps are too large, which leads to quick divergence and infinite estimates.
- Currently implementing a line search algorithm. Check sum of squared GEE values and iteratively set $\gamma^{k+1} = \frac{1}{2}\gamma^{k}$ until there is a reduction.

Real data - Zebrafish data



- Use only day 32
- Filter taxa to only include if present in 30% of samples
- Use covariate of parasite/no parasite introduced
- Using 68 samples and 39 taxa.
- Initially using data to get algorithm up and running, now transitioning to looking for results.

Real Data - Results

- $\omega = 0.79$
- $\rho = 4.4$
- 80% of estimated correlation is due to the compositional structure, 20% is from phylogenetic correlation,
- $\beta =$
- Sandwhich etimator
- Include heatmap of correlations?

$$(X^tX)^{-1}[\sum_i (y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)^t X_i](X^tX)^{-1}$$

Current challenges

- Currently working on various numerical challenges:
- When data are not filtered so strictly (eg 10% or 20%), algorithm would not converge or was unstable. A less strict filter results in both a larger percent zero data but also an increase in parameters.
- Belieive to be due to close to singular hessian matrix.
- Stability of estimates of ρ, ω sensitive to initial starting values of each step.

Questions - Phylogenetic analysis

- Currently analysis is done on the ASV level.
- Could we perform analysis on the Genus level?
- How would we consolidate the phylogenetic tree (and calculate distances) at the Genus level?
- Is this reasonable/appropriate?

Questions & next steps

1. What are covariates or groups of covariates that are more of interest?
2. How to calculating significance? Can use sandwhich estimator and asymptotic normality. Permutations?

Thank you!