Spring research update

Emily Palmer

Oregon State University

palmerem@oregonstate.edu

April 11, 2022

Overview

- Use Generalized estimating equations to estimate regression and correlation 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 $y_i \sim \text{Dirichlet}(\alpha_{i1}, \ldots, \alpha_{ip})$
- Link function: Link covarites to α's:

$$\log(\alpha_{ij}) = x_i^T \beta_j$$

Compositional Dirichlet Correlation

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

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

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

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

- Need to estimate β, ρ, ω
- 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

GEE Algorithm - ρ, ω, ϕ step

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

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

  suject to $0 \leq \omega \leq 1, \rho > 0$
- Use "L-BFGS-B" algorithm
- Given ρ, ω, working correlation R is specified.

GEE Algorithm - β step

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

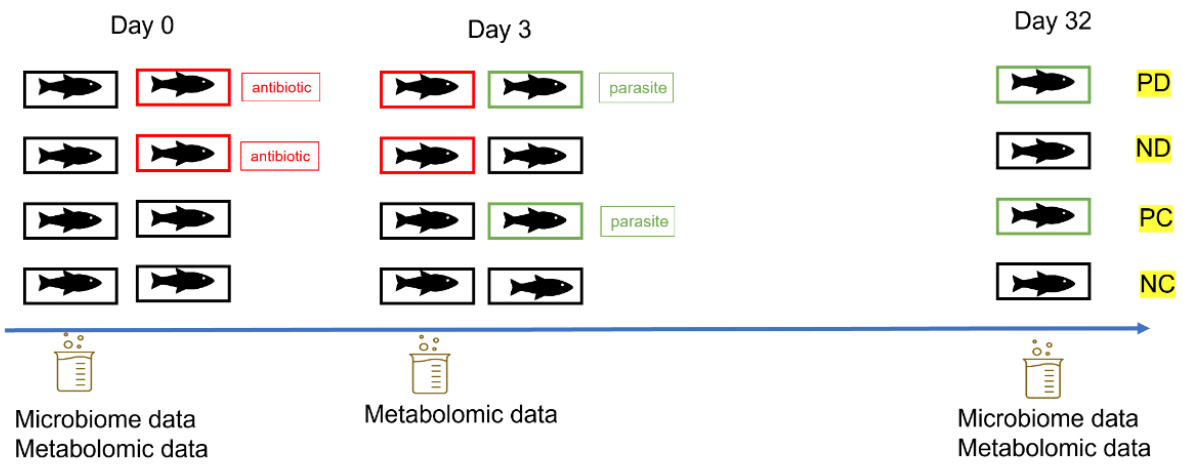
$$H = -\sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right)^t V_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} + \lambda I$$

$$\hat{\beta}^{(s+1)} = \hat{\beta}^{(s)} - \gamma H^{-1} G \qquad 0 < \gamma \leq 1$$

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.
- Check $|G^{(k+1)}| < |G^{(k)}|$ and iteratively set $\gamma^{k+1} = \frac{1}{2}\gamma^k$.

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.

Real Data - Results

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

	Phylum	Class	Order	Family	Genus		
asv_29	"Planctomycetes"	"Phycisphaerae"	"Phycisphaerales"	"Phycisphaeraceae"	"SM1A02"	"3.706"	"-1.564"
asv_2	"Fusobacteria"	"Fusobacteriia"	"Fusobacteriales"	"Fusobacteriaceae"	"Cetobacterium"	"-1.859"	"3.152"
asv_80	"Bacteroidetes"	"Flavobacteriia"	"Flavobacteriales"	"Flavobacteriaceae"	"Cloacibacterium"	"-3.267"	"1.023"
asv_5	"Bacteroidetes"	"Flavobacteriia"	"Flavobacteriales"	"Flavobacteriaceae"	"Flavobacterium"	"-1.925"	"2.378"
asv_33	"Bacteroidetes"	"Flavobacteriia"	"Flavobacteriales"	"Flavobacteriaceae"	"Flavobacterium"	"-4.326"	"0.454"
asv_19	"Bacteroidetes"	"Sphingobacteriia"	"Sphingobacteriales"	"Chitinophagaceae"	"Chitinophaga"	"-3.699"	"0.538"
asv_72	"Bacteroidetes"	"Sphingobacteriia"	"Sphingobacteriales"	"Chitinophagaceae"	"Dinghuibacter"	"-4.194"	"0.405"
asv_54	"Bacteroidetes"	"Sphingobacteriia"	"Sphingobacteriales"	"Chitinophagaceae"	"Flavihumibacter"	"-2.406"	"0.664"
asv_56	"Bacteroidetes"	"Flavobacteriia"	"Flavobacteriales"	"Cryomorphaceae"	"Fluviicola"	"-4.362"	"0.322"
asv_13	NA	NA	NA	NA	NA	"1.172"	"5.549"
asv_35	"Proteobacteria"	"Alphaproteobacteria"	"Rickettsiales"	"Mitochondria"	NA	"-1.005"	"-2.135"
asv_160	"Actinobacteria"	"Actinobacteria"	"Corynebacteriales"	"Mycobacteriaceae"	"Mycobacterium"	"-1.753"	"2.693"
asv_6	"Firmicutes"	NA	NA	NA	NA	"-4.845"	"-0.201"
asv_28	"Proteobacteria"	"Alphaproteobacteria"	"Rhodobacterales"	"Rhodobacteraceae"	"Defluviimonas"	"-3.958"	"0.426"
asv_10	"Proteobacteria"	"Alphaproteobacteria"	"Rhodobacterales"	"Rhodobacteraceae"	"Gemmobacter"	"-3.853"	"0.215"
asv_44	"Proteobacteria"	"Alphaproteobacteria"	"Rhizobiales"	NA	NA	"-3.729"	"0.663"
asv_74	"Proteobacteria"	"Alphaproteobacteria"	"Rhizobiales"	"Rhizobiaceae"	NA	"-4.174"	"0.171"
asv_23	"Proteobacteria"	"Alphaproteobacteria"	"Rhizobiales"	"Rhizobiaceae"	"Rhizobium"	"-25.261"	"-23.942"
asv_20	"Proteobacteria"	"Alphaproteobacteria"	"Rhizobiales"	"Rhizobiaceae"	"Ensifer"	"-1.751"	"-1.107"
asv_36	"Proteobacteria"	"Alphaproteobacteria"	"Rhizobiales"	"Hyphomicrobiaceae"	"Hyphomicrobium"	"-15.776"	"2.056"
asv_62	"Proteobacteria"	"Alphaproteobacteria"	"Rhizobiales"	"MNG7"	NA	"-2.738"	"1.946"
asv_8	"Proteobacteria"	"Alphaproteobacteria"	"Rhizobiales"	"Bradyrhizobiaceae"	"Bosea"	"-3.948"	"0.512"
asv_86	"Proteobacteria"	"Alphaproteobacteria"	"Rhizobiales"	"Rhizobiales_Incertae_Sedis"	"Phreatobacter"	"-4.16"	"0.415"
asv_16	"Proteobacteria"	"Alphaproteobacteria"	"Rhizobiales"	"Rhizobiales_Incertae_Sedis"	"Phreatobacter"	"-2.266"	"0.521"
asv_49	"Proteobacteria"	"Alphaproteobacteria"	"Rhizobiales"	"Bradyrhizobiaceae"	"Bradyrhizobium"	"0.196"	"-4.174"
asv_1	"Proteobacteria"	"Gammaproteobacteria"	"Aeromonadales"	"Aeromonadaceae"	"Aeromonas"	"-0.393"	"9.102"
asv_60	"Proteobacteria"	"Gammaproteobacteria"	"Pseudomonadales"	"Pseudomonadaceae"	"Pseudomonas"	"-4.083"	"0.021"
asv_34	"Proteobacteria"	"Gammaproteobacteria"	"Pseudomonadales"	"Pseudomonadaceae"	"Pseudomonas"	"-3.863"	"0.405"
asv_3	"Proteobacteria"	"Gammaproteobacteria"	"Pseudomonadales"	"Pseudomonadaceae"	"Pseudomonas"	"5.38"	"-4.593"
asv_51	"Proteobacteria"	"Gammaproteobacteria"	"Oceanospirillales"	"Oceanospirillaceae"	NA	"-3.927"	"0.394"

Current challenges - High dimensionality

- Currently working on various numerical challenges likely due to high dimension of the data:
- 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.
- Believe 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?
- How to calculate distances between different Genera?
- Ideally this will help both high-dimensionality and interpretability.

Questions & next steps

1. Adding covariates: What are covariates or groups of covariates that are more of interest? (parasite burden, antibiotics etc.)
2. Hypothesis testing: How to calculate significance? Can use sandwich estimator and asymptotic normality? How do ρ, ω play a role in this test?
3. Simulations
4. How to address high dimensionality?

Thank you!