Spring research update

Emily Palmer

Oregon State University

palmerem@oregonstate.edu

April 6, 2022

Overview

Goals

- Assume the mean and variance follow Dirichlet equations
- Assume the correlation structure between ASVs in a sample depends on compositionality and phylogenetic similarity
- Use GEEs to estimate regression parameters and covariance parameters

Notation and setup

- Let y_{ij} be the relative abundance of the jth ASV in the ith sample. $i = 1, \dots, n, j = 1, \dots, p$
- 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 $\mathrm{y}_i \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_p)$
- Link function: Link covarites to α's:

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

and

$$\alpha_{ij} = e^{\mathrm{x_i}^T \boldsymbol{\beta}_j}$$

Compositional Dirichlet Correlation

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

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

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

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

  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.

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

Algorithm - ρ, ω, ϕ step

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

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

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

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

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

Algorithm - β step

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

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

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

Where $0 < \gamma < 1$ is calculated by line search

Real data (intro)

a

Real Data - Results

a

Next steps and questions

a