

Generalized estimating equation modeling on correlated microbiome sequencing data with longitudinal measures

Emily Palmer

Journal Club, Oregon State University

May 11, 2021

Overview

Introduction

Correlation structure

- Taxonomic correlation structure of OTUs

- Correlation structure from longitudinal repeated measures

- Integrative Correlation Matrix

Microbiome Taxonomic Longitudinal Correlation model

- General GEE framework

- Two-part model

Introduction

Challenges of applying regression models on association studies of microbiome composition and environmental factors

- ▶ Many OTUs, potentially correlated
- ▶ Repeated Measures (longitudinal, other repeated measures)
- ▶ OTU data has excessive zeros

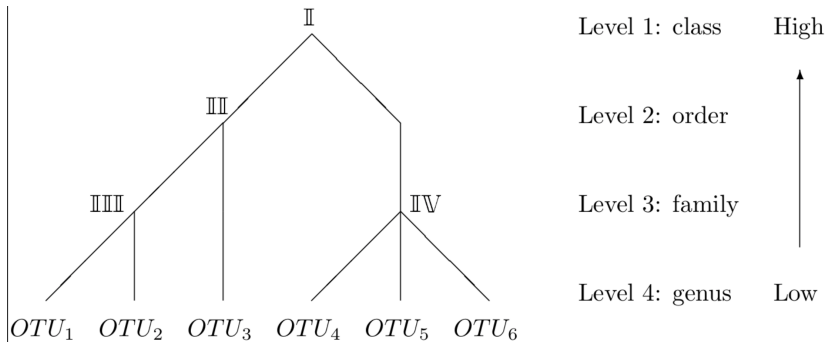
Goals

- ▶ Estimate correlations between multiple OTUs
- ▶ Incorporate correlations into models with longitudinal OTU measures
- ▶ Estimate predictors effects using GEEs
- ▶ Two-part Microbiome Taxonomic Longitudinal Correlation (MLTC) model

Correlation matrix of taxonomic structure - Assumptions

- ▶ Assume that OTUs that belong to the same taxa at some higher level have some correlation
- ▶ All OTUs will belong to same taxa at highest level, so there are $\binom{N}{2}$ possible correlations - infeasible to model
- ▶ Assume that two pairs of OTUs have the same correlation if the first common taxa of both pairs are identical

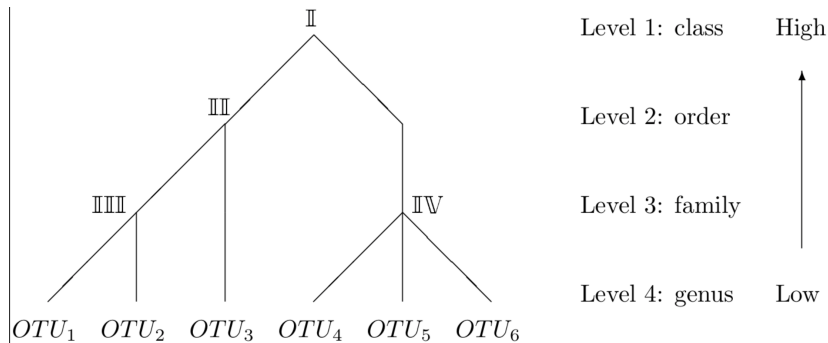
Example



Notation and Definitions

- ▶ I levels:
 - ▶ 1st taxonomic level is the level at which all observed N OTUs belong to the same taxon but not at one level lower
- ▶ M_i : number of taxa at level i ($M_1 = 1, M_I = N$)
- ▶ $t_{m_i,i}$: taxon at level i for $m_i = 1, \dots, M_i$
- ▶ $n_{m_i,i}$: number of OTUs belonging to taxon $t_{m_i,i}$.
 $\mathbf{n}_i = (n_{1i}, \dots, n_{M_i,i})$

Example



$$M_1 = 1, M_2 = 2, M_3 = 3, M_4 = 6,$$

$$\mathbf{n}_1 = 6, \mathbf{n}_2 = (3, 3), \mathbf{n}_3 = (2, 1, 3), \mathbf{n}_4 = (1, 1, 1, 1, 1, 1)$$

II represents correlation of same class different orders,

III correlation of same order different families,

IIII, IV same family

The taxonomic structure matrix Γ

| | OTU_1 | OTU_2 | OTU_3 | OTU_4 | OTU_5 | OTU_6 |
|---------|---------|---------|---------|---------|---------|---------|
| OTU_1 | D | IIII | III | I | I | I |
| OTU_2 | IIII | D | III | I | I | I |
| OTU_3 | III | III | D | I | I | I |
| OTU_4 | I | I | I | D | IV | IV |
| OTU_5 | I | I | I | IV | D | IV |
| OTU_6 | I | I | I | IV | IV | D |

Finding the taxonomic structure matrix

- ▶ Create $l - 1$ $N \times N$ block matrices

$$\mathbf{\Gamma}_i = \begin{pmatrix} \mathbf{B}_{1,i} & & \\ & \ddots & \\ & & \mathbf{B}_{M_i,i} \end{pmatrix}$$

For $m_i = 1, \dots, M_i$, each block $\mathbf{B}_{m_i,i}$ is an $n_{m_i,i} \times n_{m_i,i}$ matrix, with diagonal entries \mathbb{D} and off diagonal entries $\sum_{h=0}^{i-1} M_h + m_i$

- ▶ Create interim correlation after replacement at level i ($\mathbf{\Gamma}^{(i)}$)
 - ▶ For $i = 1$, $\mathbf{\Gamma}^{(1)} = \mathbf{\Gamma}_1$
 - ▶ For $i = 2, \dots, l - 1$, Replace the block diagonal entries of $\mathbf{\Gamma}^{(i-1)}$ with $\mathbf{B}_{m_i,i}$, but keep all other entries the same.
- ▶ Sort all elements from largest to smallest. Different ranks are the distinct correlations to estimate

Example

$$\Gamma_1 = \begin{pmatrix} \mathbb{D} & 1 & 1 & 1 & 1 & 1 \\ 1 & \mathbb{D} & 1 & 1 & 1 & 1 \\ 1 & 1 & \mathbb{D} & 1 & 1 & 1 \\ 1 & 1 & 1 & \mathbb{D} & 1 & 1 \\ 1 & 1 & 1 & 1 & \mathbb{D} & 1 \\ 1 & 1 & 1 & 1 & 1 & \mathbb{D} \end{pmatrix}, \Gamma_2 = \begin{pmatrix} \mathbb{D} & 2 & 2 & & & \\ 2 & \mathbb{D} & 2 & & & \\ 2 & 2 & \mathbb{D} & & & \\ & & & \mathbb{D} & 3 & 3 \\ & & & 3 & \mathbb{D} & 3 \\ & & & 3 & 3 & \mathbb{D} \end{pmatrix}$$

$$\Gamma_3 = \begin{pmatrix} \mathbb{D} & 4 & & & & \\ 4 & \mathbb{D} & & & & \\ & & \mathbb{D} & & & \\ & & & \mathbb{D} & 6 & 6 \\ & & & 6 & \mathbb{D} & 6 \\ & & & 6 & 6 & \mathbb{D} \end{pmatrix}.$$

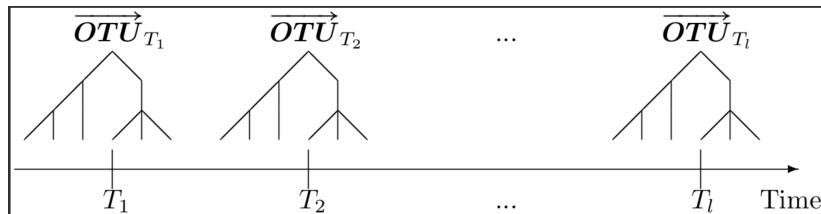
$$\Gamma^{(3)} = \begin{pmatrix} \mathbb{D} & 4 & 2 & 1 & 1 & 1 \\ 4 & \mathbb{D} & 2 & 1 & 1 & 1 \\ 2 & 2 & \mathbb{D} & 1 & 1 & 1 \\ 1 & 1 & 1 & \mathbb{D} & 6 & 6 \\ 1 & 1 & 1 & 6 & \mathbb{D} & 6 \\ 1 & 1 & 1 & 6 & 6 & \mathbb{D} \end{pmatrix}$$

Γ can be represented by $(\mathbf{n}_1, \dots, \mathbf{n}_I)$

The taxonomic structure matrix Γ

| | OTU_1 | OTU_2 | OTU_3 | OTU_4 | OTU_5 | OTU_6 |
|---------|---------|---------|---------|---------|---------|---------|
| OTU_1 | D | IIII | III | I | I | I |
| OTU_2 | IIII | D | III | I | I | I |
| OTU_3 | III | III | D | I | I | I |
| OTU_4 | I | I | I | D | IV | IV |
| OTU_5 | I | I | I | IV | D | IV |
| OTU_6 | I | I | I | IV | IV | D |

Correlation structure from longitudinal repeated measures



Types of correlations between pairs of time points

- ▶ Exchangeable
 - ▶ Assumes all correlations are equal to each other
- ▶ Toeplitz
 - ▶ Assumes time points with equal temporal distance have equal correlation
- ▶ Unstructured
 - ▶ Assumes each pair has a different correlations
 - ▶ Most complicated structure for correlation parameter estimation

Correlation structure matrix for the the same individual is denoted Ω_T

Example Correlation Matrices for 3 timepoints

Exchangable structure

| | T_1 | T_2 | T_3 |
|-------|----------------|----------------|----------------|
| T_1 | \mathbb{D} | \mathfrak{i} | \mathfrak{i} |
| T_2 | \mathfrak{i} | \mathbb{D} | \mathfrak{i} |
| T_3 | \mathfrak{i} | \mathfrak{i} | \mathbb{D} |

Toeplitz structure

| | T_1 | T_2 | T_3 |
|-------|-----------------|----------------|-----------------|
| T_1 | \mathbb{D} | \mathfrak{i} | \mathfrak{ii} |
| T_2 | \mathfrak{i} | \mathbb{D} | \mathfrak{i} |
| T_3 | \mathfrak{ii} | \mathfrak{i} | \mathbb{D} |

Combining longitudinal and sample correlation

When both longitudinal and sample correlations exist, the repeated measure correlation matrix is all combinations of time points and repeated samples

| | (T_1, S_1) | (T_2, S_1) | (T_3, S_1) | (T_1, S_2) | (T_2, S_2) | (T_3, S_2) |
|--------------|------------------|------------------|------------------|------------------|------------------|------------------|
| (T_1, S_1) | \mathbb{D} | \mathfrak{i} | \mathfrak{i} | \mathfrak{ii} | \mathfrak{iii} | \mathfrak{iii} |
| (T_2, S_1) | \mathfrak{i} | \mathbb{D} | \mathfrak{i} | \mathfrak{iii} | \mathfrak{ii} | \mathfrak{iii} |
| (T_3, S_1) | \mathfrak{i} | \mathfrak{i} | \mathbb{D} | \mathfrak{iii} | \mathfrak{iii} | \mathfrak{ii} |
| (T_1, S_2) | \mathfrak{ii} | \mathfrak{iii} | \mathfrak{iii} | \mathbb{D} | \mathfrak{i} | \mathfrak{i} |
| (T_2, S_2) | \mathfrak{iii} | \mathfrak{ii} | \mathfrak{iii} | \mathfrak{i} | \mathbb{D} | \mathfrak{i} |
| (T_3, S_2) | \mathfrak{iii} | \mathfrak{iii} | \mathfrak{ii} | \mathfrak{i} | \mathfrak{i} | \mathbb{D} |

Incorporating taxonomic structure with repeated measures

$\mathbf{\Omega}$ with dimension L , for $a, b = 1, \dots, N$,

$$\mathbf{\Omega}(\Gamma_{ab}) = \begin{pmatrix} \rho_{(\Gamma_{ab}, \mathbf{\Omega}_{11})} & \cdots & \rho_{(\Gamma_{ab}, \mathbf{\Omega}_{1L})} \\ \vdots & \ddots & \vdots \\ \rho_{(\Gamma_{ab}, \mathbf{\Omega}_{L1})} & \cdots & \rho_{(\Gamma_{ab}, \mathbf{\Omega}_{LL})} \end{pmatrix}.$$

$$\mathbf{R} = \begin{pmatrix} \mathbf{\Omega}^{11} & \cdots & \mathbf{\Omega}^{1N} \\ \vdots & \ddots & \vdots \\ \mathbf{\Omega}^{N1} & \cdots & \mathbf{\Omega}^{NN} \end{pmatrix}$$

Dimension of $\mathbf{R} = (N \times L) \times (N \times L)$

Diagonals of $\mathbf{R} = \rho(\mathbb{D}, \mathbb{D})$ are 1, off-diagonals need to be estimated

Example R

For two correlated OTUs and two repeated measures at different time points

$$\mathbf{R} = \begin{pmatrix} \rho_{(\mathbb{D},\mathbb{D})} & \rho_{(\mathbb{D},\mathbb{I})} & \rho_{(\mathbb{I},\mathbb{D})} & \rho_{(\mathbb{I},\mathbb{I})} \\ \rho_{(\mathbb{D},\mathbb{I})} & \rho_{(\mathbb{D},\mathbb{D})} & \rho_{(\mathbb{I},\mathbb{I})} & \rho_{(\mathbb{I},\mathbb{D})} \\ \rho_{(\mathbb{I},\mathbb{D})} & \rho_{(\mathbb{I},\mathbb{I})} & \rho_{(\mathbb{D},\mathbb{D})} & \rho_{(\mathbb{D},\mathbb{I})} \\ \rho_{(\mathbb{I},\mathbb{I})} & \rho_{(\mathbb{I},\mathbb{D})} & \rho_{(\mathbb{D},\mathbb{I})} & \rho_{(\mathbb{D},\mathbb{D})} \end{pmatrix}$$

- ▶ $\rho_{(\mathbb{D},\mathbb{D})} = 1$
- ▶ $\rho_{(\mathbb{D},\mathbb{I})}, \rho_{(\mathbb{I},\mathbb{D})}$ correlation between two time points and two OTUs
- ▶ $\rho_{(\mathbb{I},\mathbb{I})}$ correlation from different OTU and different time points

Introduction to MTLC:

MTLC:

- ▶ Estimate predictor effects
- ▶ Estimate correlation coefficients between OTUs, longitudinal measures and other repeated measures
- ▶ Perform hypothesis testing of predictor effects

Generalized estimating equation framework

- ▶ $\mathbf{y}_k = (y_{k1}, \dots, y_{kJ_k})$ clusters, length J_k for $k = 1, \dots, K$
- ▶ \mathbf{x}_{kj} the vector of covariates with length p , $j = 1, \dots, J_k$
- ▶ $\boldsymbol{\mu}_k = (\mu_{k1}, \dots, \mu_{kJ_k})$ mean of \mathbf{y}_k
- ▶ Each observation y_{kj}

$$g(\mu_{kj}) = \mathbf{x}_{kj}'\boldsymbol{\beta}$$

- ▶ Conditional variance of y_{kj}

$$\text{Var}(y_{kj}|\mathbf{x}_{kj}) = v(\boldsymbol{\mu}_{kj})\phi$$

v is the variance function depending on the distribution of y_{kj} ,
 ϕ is dispersion parameter

cont.

- ▶ Estimate β by solving the generalized estimating equation

$$U(\beta) = \sum_{k=1}^K \mathbf{D}'_k \mathbf{V}_k^{-1} (\mathbf{y}_k - \boldsymbol{\mu}_k) = 0$$

- ▶ $\mathbf{D}_k = \frac{d\boldsymbol{\mu}_k}{d\beta}$, $\mathbf{V}_k = \mathbf{A}_k^{1/2} \mathbf{R}_k(\rho) \mathbf{A}_k^{1/2}$,
- ▶ $\mathbf{A}_k = \text{diag}(\mu_{k1}\phi, \dots, \mu_{kJ_k}\phi)$ ρ collection of all correlation coefficients in \mathbf{R}_k
- ▶ $\mathbf{R}_k(\rho)$ is the working correlation matrix following correlation structure R

cont.

- ▶ ϕ, ρ also need to be estimated

$$\hat{\phi} = \frac{1}{\sum_{k=1}^K J_k - \rho} \sum_{k=1}^K \sum_{j=1}^{J_k} e_{kj}^2$$

where e_{kj} is the Pearson residual

- ▶ $\hat{\rho}$ is estimated as a function of ϕ and e_{kj} , depending on the correlation structure R
- ▶ Iterative - switch between estimating β from fixed value of $\hat{\phi}$ and $\hat{\beta}$ and estimating ϕ and ρ for a fixed value of $\hat{\beta}$

Hypothesis testing

- ▶ From GEE theory $\hat{\beta}$ is asymptotically normally distributed with mean β and variance

$$V_{\beta} = (\sum_{k=1}^K \mathbf{D}'_k \mathbf{V}_k^{-1} \mathbf{D}_k)^{-1} \{ \sum_{k=1}^K \mathbf{D}'_k \mathbf{V}_k^{-1} \text{Cov}(\mathbf{y}_k) \mathbf{V}_k^{-1} \mathbf{D}_k \} (\sum_{k=1}^K \mathbf{D}'_k \mathbf{V}_k^{-1} \mathbf{D}_k)^{-1}$$

- ▶ Wald test statistic for testing $H_0 : \mathbf{C}\beta = \mathbf{c}$

$$\mathbf{W} = (\mathbf{C}\hat{\beta} - \mathbf{c})' (\mathbf{C}\hat{\mathbf{V}}_{\beta} \mathbf{C}')^{-1} (\mathbf{C}\hat{\beta} - \mathbf{c})$$

- ▶ $\mathbf{W} \xrightarrow{d} \chi^2_{(q)}$ where q is the rank of \mathbf{C}

Estimating predictors effects on OTUs

- ▶ Two part model - two separate GEE models
 - ▶ Convert quantitative OTU observations to binary outcomes indicating prevalence of OTU in each observation - assessing predictor effects on OTU prevalence
 - ▶ Relative abundance of non-zero observation - assume RA follows normal distribution after log transformation - predictor effects on positive RA
 - ▶ Combine test statistics from two models for overall predictor effects

OTU GEE Model

- ▶ Assume each OTU observation y_{kj} follows a mixture of Bernoulli and log-normal distribution.
- ▶ OTU prevalence: $y_{kj}^{(0)}$ follows a Bernoulli distribution with $P(Y_{kj}^{(0)} = 1) = \mu_{kj}^{(0)}$
- ▶ Log-transform positive RAs: $y_{kj}^{(+)}$ follows a normal distribution

$$F(y) = \begin{cases} 1 - \mu_{kj}^{(0)} & y = 0 \\ 1 - \mu_{kj}^{(0)} + \mu_{kj}^{(0)} \Phi(\log y) & y > 0 \end{cases}$$

GEE Model for OTUs

- ▶ For OTU prevalence, use logit link function

$$\log \frac{\mu_{jk}^{(0)}}{1 - \mu_{jk}^{(0)}} = \mathbf{x}'_{kj} \boldsymbol{\beta}^{(0)}$$

- ▶ For Log-transform RA, use identity link function

$$\mu_{jk}^{(+)} = \mathbf{x}'_{kj} \boldsymbol{\beta}^{(0)}$$

- ▶ Use GEE framework to find parameter estimates $\hat{\boldsymbol{\beta}}^{(0)}$ and $\hat{\boldsymbol{\beta}}^{(+)}$

Hypothesis testing

- ▶ Test if the predictors have effects on either the prevalence of OTUs or the quantitative amount of RA,

$$H_0 : \mathbf{C}^{(0)}\boldsymbol{\beta}^{(0)} = \mathbf{c}^{(0)} \text{ and } H_0 : \mathbf{C}^{(+)}\boldsymbol{\beta}^{(+)} = \mathbf{c}^{(+)}$$

- ▶ Calculate Wald test statistics $W^{(0)}$ and $W^{(+)}$
- ▶ Cauchy combination test

$$W_{MTLC} = 0.5\tan[(0.5 - p^{(0)})\pi] + 0.5\tan[(0.5 - p^{(+)})\pi] \xrightarrow{d} \text{Cauchy}(0, 1)$$

Estimating correlation coefficients

- ▶ Estimated values of correlation coefficients $\hat{\rho}^{(0)}$ and $\hat{\rho}^{(+)}$ may be different.
- ▶ When Pearson correlations are available to compute, the GEE estimates are similar.

Discussion

- ▶ MTLC accounts for taxonomic correlation structure and longitudinal correlation structure
- ▶ MTLC has accurate Type I error, unbiased estimation of model parameters and robust power performance
- ▶ Correlation estimation is consistent
- ▶ Does not put a constraint on range of correlation coefficient
- ▶ Recommend using a subset of OTUs as model is time consuming when $N > 1000$.

Thank you!

References



Chen B, Xu W (2020)

Generalized estimating equation modeling on correlated microbiome sequencing data with longitudinal measures

PLoS Comput Biol 16(9): e1008108.

<https://doi.org/10.1371/journal.pcbi.1008108>.