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

Emily Palmer

Journal Club, Oregon State University

March 14, 2021

Overview

Introduction

Correlation strucutre

- Taxonomic structure of OTUs
- Modelling correlations from repeated measures
- Integrative Correlation Matrix

Microbiome Taxonomic Longitudinal Correlation model

- GEE framework

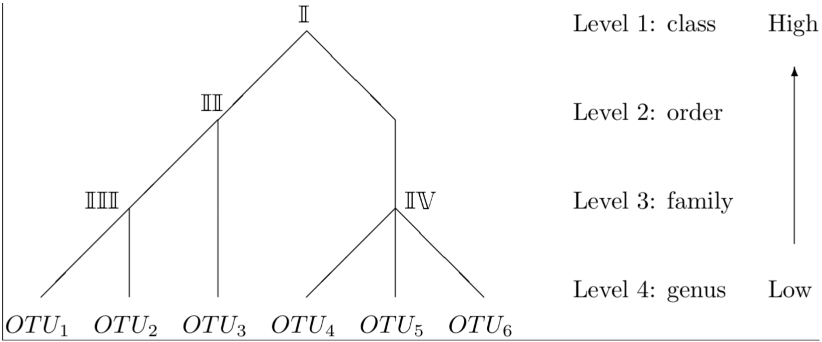
Introduction

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

Notation and Definitions

- N OTUs
- 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
- $n_{m_i,i}$: number of OTUs belonging to taxon $t_{m_i,i}$

Example



Level 1: class High

Level 2: order

Level 3: family

Level 4: genus Low

Correlation matrix of taxonomic structure - Assumptions

- Assume that OTUs that belong to the same taxa at some higher level have some correlation
- From taxonomic structure, all OTUs will belong to same taxa at hightest level, so there are $\binom{N}{2}$ possible correlations - infeasible to model
- Clusters of OTUs (otus belonging to the same taxa)
- Assume that two pairs of OTUs have the same correlation if the first common taxa of both pairs are identical
  If $\mathcal{P}^*$ and $\mathcal{P}^\dagger$ are two pairs of OTUs, with correlation ρ^* and $\rho^\dagger$. $t_{m_i^*,i^*}$ is first common taxa of $\mathcal{P}^*$ $t_{m_i^\dagger,i^\dagger}$ is first common taxa of $\mathcal{P}^\dagger$

$$\rho^* = \rho^\dagger \leftrightarrow t_{m_i^*,i^*} = t_{m_i^\dagger,i^\dagger}$$

Finding the taxonomic structure matrix

- The taxonomic structure matrix
- Go through algoritm...

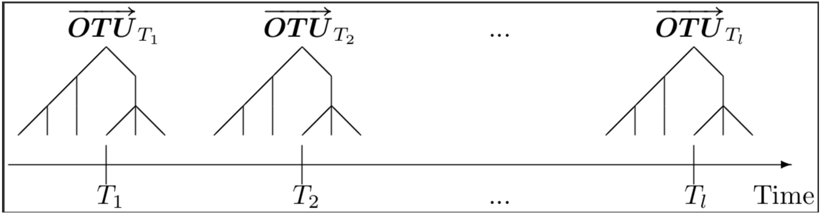
Correlations of longitudinal data

Types of correlations between pairs of time points

- Exchangable
  - 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 dentoed Ω_T

Example


$\overrightarrow{OTU}_{T_1}$
$\overrightarrow{OTU}_{T_2}$
...
$\overrightarrow{OTU}_{T_l}$
T_1
T_2
...
T_l
Time

Multiple Columns

Example for 3 timepoints

Exchangable structure

$$\begin{array}{cccc} & T_1 & T_2 & T_3 \\ T_1 & \mathbb{D} & \mathbb{i} & \mathbb{i} \\ T_2 & \mathbb{i} & \mathbb{D} & \mathbb{i} \\ T_3 & \mathbb{i} & \mathbb{i} & \mathbb{D} \end{array}$$

Toeplitz structure

$$\begin{array}{cccc} & T_1 & T_2 & T_3 \\ T_1 & \mathbb{D} & \mathbb{i} & \mathbb{ii} \\ T_2 & \mathbb{i} & \mathbb{D} & \mathbb{i} \\ T_3 & \mathbb{ii} & \mathbb{i} & \mathbb{D} \end{array}$$

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

$$\begin{array}{ccccccc}
 & (T_1,S_1) & (T_2,S_1) & (T_3,S_1) & (T_1,S_2) & (T_2,S_2) & (T_3,S_2) \\
(T_1,S_1) & \mathbb{D} & \mathbb{i} & \mathbb{i} & \mathbb{ii} & \mathbb{iii} & \mathbb{iii} \\
(T_2,S_1) & \mathbb{i} & \mathbb{D} & \mathbb{i} & \mathbb{iii} & \mathbb{ii} & \mathbb{iii} \\
(T_3,S_1) & \mathbb{i} & \mathbb{i} & \mathbb{D} & \mathbb{iii} & \mathbb{iii} & \mathbb{ii} \\
(T_1,S_2) & \mathbb{ii} & \mathbb{iii} & \mathbb{iii} & \mathbb{D} & \mathbb{i} & \mathbb{i} \\
(T_2,S_2) & \mathbb{iii} & \mathbb{ii} & \mathbb{iii} & \mathbb{i} & \mathbb{D} & \mathbb{i} \\
(T_3,S_2) & \mathbb{iii} & \mathbb{iii} & \mathbb{ii} & \mathbb{i} & \mathbb{i} & \mathbb{D}
\end{array}$$

Integrative Correlation Matrix

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

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

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

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

- y_k independent clusters $k = 1, \dots, K$
- J_k cluster length for cluster $y_k = (y_{k1}, \dots, y_{kJ_k})$
- x_{kj} the vector of covariates with length p
- $\boldsymbol{\mu}_k = (\mu_{k1}, \dots, \mu_{kJ_k})$
- Each observation y_{kj}

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

- Conditional variance of y_{kj}

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

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

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

- $\boldsymbol{D}_k = \frac{d\mu_k}{d\beta}$, $\boldsymbol{V}_k = \boldsymbol{A}_k^{1/2}\boldsymbol{R}_k(\boldsymbol{\rho})\boldsymbol{A}_k^{1/2}$,
- $\boldsymbol{A}_k = diag(\mu_{k1}\phi, \ldots, \mu_{kJ_k}\phi)$ $\boldsymbol{\rho}$ collection of all correlation coefficients in $\boldsymbol{R_k}$
- $\phi, \boldsymbol{\rho}$ also need to be estimated

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

where e_{kj} is the pearson residual

- $\hat{\boldsymbol{\rho}}$ is estimated as a funtion of ϕ and e_{kj}, depending on the correlation structure R

Hypothesis testing

In GEE theory, $\hat{\boldsymbol{\beta}}$ is asymptotically normally distributed with mean $\boldsymbol{\beta}$ and variance

Citation

An example of the `\cite` command to cite within the presentation:

This statement requires citation [Smith, 2012].

References

John Smith (2012)
Title of the publication
Journal Name 12(3), 45 – 678.