Masters Project

Emily Palmer

Group Meeting, Oregon State University

May 12, 2021

Motivation

PLOS COMPUTATIONAL BIOLOGY

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

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

Bo Chen, Wei Xu

Version 2 Published: September 8, 2020 • https://doi.org/10.1371/journal.pcbi.1008108

Overview of Model

- Correlation structure from taxonomic information
- Correlation structure from longitudinal data or repeated measures
- Two part model, using generalized estimating equations for parameter estimation
  - Presence/Absence
  - Relative Abundance of positive counts
- Results from paper show this method to have accurate Type I error, unbiased estimation of model parameters, and more powerful than some existing methods.

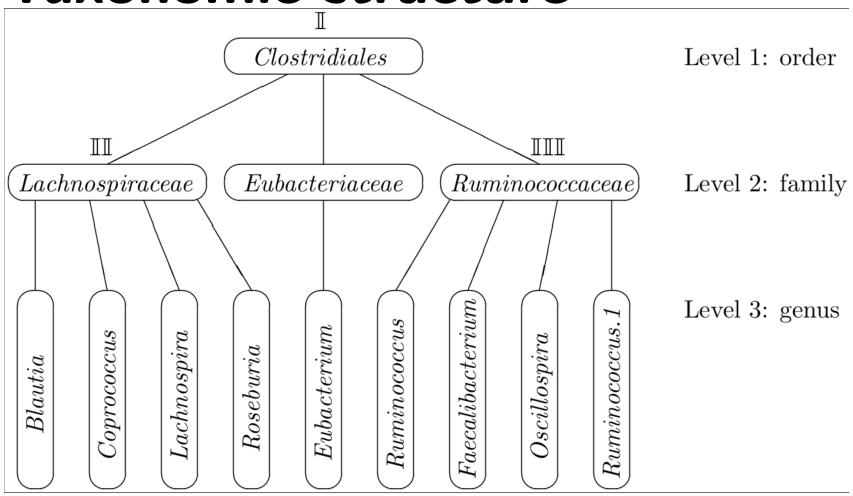
Application in paper

- Twin obesity data Turnbaugh et. al. (2009)
- 54 families (clusters), 2 twins each, 2 time points
- Observations for 9 OTUs (only order Clostridiales)
- Obesity indication for each twin
- Some clusters are incomplete

Correlation matrix of taxonomic structure

- Assume that OTUs that belong to the same taxa at a higher level have some correlation

Taxonomic structure



Gamma matrix

$$\Gamma = \begin{pmatrix} \mathbb{D} & \mathbb{I} & \mathbb{I} & \mathbb{I} & \mathbb{IIII} & \mathbb{IIII} & \mathbb{IIII} & \mathbb{IIII} & \mathbb{IIII} \\ \mathbb{I} & \mathbb{D} & \mathbb{I} & \mathbb{I} & \mathbb{IIII} & \mathbb{IIII} & \mathbb{IIII} & \mathbb{IIII} & \mathbb{IIII} \\ \mathbb{I} & \mathbb{I} & \mathbb{D} & \mathbb{I} & \mathbb{IIII} & \mathbb{IIII} & \mathbb{IIII} & \mathbb{IIII} & \mathbb{IIII} \\ \mathbb{I} & \mathbb{I} & \mathbb{I} & \mathbb{D} & \mathbb{IIII} & \mathbb{IIII} & \mathbb{IIII} & \mathbb{IIII} & \mathbb{IIII} \\ \mathbb{IIII} & \mathbb{IIII} & \mathbb{IIII} & \mathbb{IIII} & \mathbb{D} & \mathbb{IIII} & \mathbb{IIII} & \mathbb{IIII} & \mathbb{IIII} \\ \mathbb{IIII} & \mathbb{IIII} & \mathbb{IIII} & \mathbb{IIII} & \mathbb{IIII} & \mathbb{D} & \mathbb{II} & \mathbb{II} & \mathbb{II} \\ \mathbb{IIII} & \mathbb{IIII} & \mathbb{IIII} & \mathbb{IIII} & \mathbb{IIII} & \mathbb{II} & \mathbb{D} & \mathbb{II} & \mathbb{II} \\ \mathbb{IIII} & \mathbb{IIII} & \mathbb{IIII} & \mathbb{IIII} & \mathbb{IIII} & \mathbb{II} & \mathbb{II} & \mathbb{D} & \mathbb{II} \\ \mathbb{IIII} & \mathbb{IIII} & \mathbb{IIII} & \mathbb{IIII} & \mathbb{IIII} & \mathbb{II} & \mathbb{II} & \mathbb{II} & \mathbb{D} \end{pmatrix}.$$

Combining with Longitudinal/Repeated Measure data

- Correlation matrix for repeated measures - structure flexible

$$\mathbf{\Omega} = \begin{pmatrix} \mathbb{D} & \mathbb{1} & \mathbb{11} & \mathbb{111} \\ \mathbb{1} & \mathbb{D} & \mathbb{111} & \mathbb{11} \\ \mathbb{11} & \mathbb{111} & \mathbb{D} & \mathbb{1} \\ \mathbb{111} & \mathbb{11} & \mathbb{1} & \mathbb{D} \end{pmatrix}.$$

- Integrative correlation matrix - This will be a block matrix indicating the distinct correlations of all combinations of time points and OTUs
- If N is the number of OTUs, and M the number of repeated measures, this integrative correlation matrix will have dimension $(N \times M) \times (N \times M)$
- This grows very quickly

Working correlation matrix in R

- The package `geepack` estimates the regression and covariance parameters
- Requires a specified working correlation matrix that is $\binom{N \times M}{2} \times$ number correlation parameters for one cluster
- For n clusters, this will have dimension $\left(n \times \binom{N \times M}{2} \right) \times$ number correlation parameters
- Correlations are linear combinations of the columns of the covariates based on the upper triangular part of the integrative correlation matrix.

Adjusting for Incomplete Clusters

- If we have the same number of observations per cluster, the working correlation matrix will be the same for each cluster
- Often, there will be some missing data for a cluster, missing time points, etc.
- The corresponding row and column of the integrative correlation matrix needs to be removed. This adjustment needs to be made for each cluster
- Corresponding rows of the working correlation matrix will need to be removed to run the code.

Scaling the model

Scaling the model

- Paper focused on data with only 9 OTUs
- How does this scale to a dataset with more common numbers of OTUs (> 1000?)
- Focus currently only on taxonomic correlation aspect

Scaling the model

- As-is, this method does not scale well to larger datasets

American Gut data

- Focus at genus level at one body sample site
- 14300 taxa and
- 260 correlation parameters to estimate
- Integrative correlation matrix will have dimension 14300×14300
- Working correlation matrix will have dimension $\binom{14300}{2} \times 240$ for one cluster
- Matrix is too large for R

Filter more?

- Filter taxa based on threshold of genus sparsity
- Reduces to 1200 taxa and
- 72 correlation parameters to estimate.
- Integrative correlation matrix will have dimension 1200×1200
- Working correlation matrix will have dimension $\binom{1200}{2} \times 240$ for one cluster
- $\left(3000 \times \binom{1200}{2}\right) \times 240$ for all clusters
- Matrix is again too large for R

Discussion

- Better ways to scale this model?
- Another implementation of fitting GEEs in R?
- Focus on groups of OTUs individually?
- Would aggregating to a higher taxa level help?
- American Gut covariates to use?