October 27

October 21

Today I worked mostly on coding the alpha step (correlation) for the evolutionary trait model.

Unfortunately, this assumption forces all correlations to be positive (since the function is $e^{-2\rho d_{ij}}$). This might be too restrictive of an assumption and is not (to my knowledge) biologically supported (and in fact has been biologically rejected theres a paper cited in my masters project)

I focused on coding for the nls part of the updating alpha step. We have the upper triangular (or lower triangular - its arbitrary) part of the distance matrix which has dimensions $p \times p$ where p is the number of OTUs.

Currently no longitudinal component yet.

Had to figure out how to manipulate data.

For the i sample, we find the corresponding d_{jk} where both j and k are indeces (here) for the OTUs. the response then is $e_{ij}e_{ik}$.

Trying to remember if we decided to make it be nonlinear least squares regression or take the log.

Lets just try the linear version?

But what if $e_{ij}e_{ik}$ is negative? So we cant take the log... They also had this in the AR1 case in the Liang paper.

Additionally, ρ has to be in $(0, \infty)$, ie positive.

Reminder: large ρ means no correlation, small ρ means high correlation.

After trying to take the log of the normalized residuals, it turns out that almost half of the multiplied residuals are negative, making the log infinity. this does not seem like a good basis for running regression as we would have to exclude them all. Is it the absolute value that is calculated??

** Look up how AR1 GEE correlation is usually calculated. - from Liang paper.

Do we use an intercept?

Currently failing with error

Error in nls($y\ exp(-2 * x * a)$, data = alpha df, start = c(a = 1)) : singular gradient

Also current variance is really small. but comparitively is it/ idk. Probably the problem is we have negative residuals. and the exponential function is positive. What to do!

Current approach: Take absolute values. This makes the code actually run.

Now we need to invert R.

GeeM code uses R as the block matrix of the entire sample structure. so instead of $p \times p$ it is $np \times np$

Is inverting equivalent??

How do i do this inverting?? Or should I do the old version of systems of equations??

First idea: convert R into Block R and invert using baser functions.

Can I use the fixed format inversion function?

Can I invert the singular R and then make it into a block matrix?

For a block diagonal matrix the inverse is the inverse of the blocks (apparently)

We only need to be fancy when cluster sizes are not the same.

Using solve - R is not invertable! Because of such high values? Why are values so high!

rho has to be positive- currently negative...

Use the linear model form because that can ensure that rho is positive? jk thats not true

Force intercept at 0? Will this ensure positive rho? Would significance mean anything here?

Still need to do sandwich estimates. Look at convergence criteria.

Oct 15 Meeting

The working correlation assumption using the evolutionary trait model $C(\rho) = e^{-2\rho d_{ij}}$ will assume all positive correlations. This might be too restrictive of an assumption.

What distribution to use for the link and variance functions? Perhaps log-normal, two parts.

New idea: (from Chenyang's GEE clustering paper) Use her ideas of the mean and variance links using Dirichlet multinomial. This will help address compositionally.

How to incorporate her working correlation matrix with mine? Weighted average? New parameter: $\omega\Sigma_1 + (1 - \omega)\Sigma_2$

How to calculate rho? Use linear or non-linear regression to find estimates in each iteration - REgress (log?) on eijeik e2rhod But residuals might be negative?

Look up more how Ziegler paper made sure the estimation of α it was not negative?. Because α could be negative but they are using logs.

October 12

First, consider samples taken on m individuals, indexed $i = 1, \ldots, m$.

Each individual has p_i response OTU measurements taken on it. For simplicity, assume $p_i = p$ for all i. OTUs measurements indexed $j = 1, \ldots, p$

$$\mathbf{y}_i = (y_{i1}, \ldots, y_{ij}, \ldots, y_{ip})$$

On each sample there are measured q covariates, such as age, disease status, etc, indexed $k = 1, \ldots, q$

$$\mathbf{X}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{ik}, \ldots, \mathbf{x}_{iq})^T$$

This matrix $\mathbf{X}_i$ has dimension $p \times q$.

In most cases $\mathbf{x_{ik}} = x_{ik}\mathbf{1}_{p\times 1}$, ie values for the kth covariate of the ith sample is the same for all p OTU observations.

We are trying to code a GEE model for microbiome data. Using [1], based on the evolutionary trait model for correlations between OTUs.

Set up notation for GEEs:

Consider the mean vector $\boldsymbol{\mu}_i = (E(y_{i1}), \ldots, E(y_{ip}))^t$ and the mean is linked to the covariates through the link

$$g(\boldsymbol{\mu}_i) = \mathbf{X}_i \boldsymbol{\beta}$$

Where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)^t$

The variance of the response is a function of the mean

$$\mathbf{Var}(\mathbf{y}_{ik}) = \phi a_{ik} = \phi a(\mu_{ik})$$

The generalized estimating equations find the solution β to satisfy

$$\sum_{i=1}^{m} \left(\frac{d\boldsymbol{\mu}_i}{d\boldsymbol{\beta}}\right)^t \left(\mathbf{A}_i^{\frac{1}{2}} \mathbf{R}_i(\alpha) \mathbf{A}_i^{\frac{1}{2}}\right)^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) = 0$$

$\mathbf{R}_i(\alpha)$ is known as the **working correlation matrix**, which is parametrized by the vector α.

$\mathbf{A}_i = \phi \text{diag}(a(\mu_{i1}))$

Currently with this formulation, the interpretation for β_k is for the expected change in the population average response from changes in the covariate.

The GEE algorithm

The GEE algorithm is an iterative algorithm based off of alternating between solving the estimating equation for β and solving the estimating equation for ϕ and α.

Trying to get GEE code running. Using ohio dataset from geeM package and data1 from glmmTree package for sanity checks and correctness. Will need to work on filtering the glmmTree dataset more as in current forumulation will not run.

Also going through the geeM code to make modifications to get the basic setup!

References

[1] J. Xiao, L. Chen, S. Johnson, Y. Yu, X. Zhang, and J. Chen, “Predictive Modeling of Microbiome Data Using a Phylogeny-Regularized Generalized Linear Mixed Model,” *Frontiers in Microbiology*, vol. 9, p. 1391, June 2018.