June 16

To-Do

□ Writeup current research.

□ Create bibliography

□ Make current document of all definitions

□ Writeup dirichlet properties

□ Citations

□ Biological definitions

□ GEE notation

□ Writeup of identifiability issue

June 17 - Meeting

This week I focused on taking a step back and writing up current research, make a place for all definitions used in one place.

Todo: Literature search on dirichlet, look at link function, see if used logit,

Implement penalty first.

Look up logit idea, remember

Continue

Start career search

Wednesday Jun 22

- □ Work on writeup
  - Create bibliography
  - Make current document of all definitions
  - Writeup dirichlet properties
  - Citations
  - Biological definitions
  - GEE notation
  - Writeup of identifiability issue
- □ Literature search on dirichlet logit link function
- □ Implemement penalty

Literature search on dirichlet link

douma2019a Use the log link like we were under normal parametrization: $\log(\alpha_c) = \eta_c = X_c\beta_c$.

Alternatively,

$$E(p_c) = \mu_c$$

and $var[p_c] = \frac{\mu_c(1-\mu_c)}{\phi+1}$, where $\phi = \alpha_0$.

Then the link function for μ is the multinomial logit function.

Somewhat reccomends alternate parametrization. Then only $C-1$ values fitted. treat C as baseline category

ϕ as the precision parameter, can model with log function.

So model both μ and ϕ? seperately?

Zero inflation: offers different transformation

$$p^* = \frac{p(n-1) + \frac{1}{C}}{n}$$

Also reccomended by **maier2014** This compresses the data symmetrically around .5 so extreme values are more affected.

Reccomended R packages: DIRMULT, BRMS, DirichletReg

Notes zero-augmented Dirichlet regression **tsagris2017**

Uses the logit link? or the one with a 1 on the bottom.

Could be interesting to look further into this method

June 23

□ Work on writeup for first pomofocus

□ Add penalty.

Change in beta order wouldnt change the GEE equations, but wondering how current formulation outputs old formulation? Explore code more.

Start on the penalty

I dont quite understand the derivatives parts...

We want to heavily penalize

- Add penalty $\lambda \sum_{j=1}^{p} \beta_j^2$ to GEE equation
- Add penalty

Meeting

:

Messed up on penalty, shoudl be matrix of 1s! Try this again.

Lambda larger,

Why does it have noise in only one clump?

Try larger sample size? Is it variation? Seems like.

Add diagonal back to hessian.

Try continuous X

Something with intercept. TODO: Constraint has to be for each beta. Ie the intercept betas have to sum to 0, same for all other covariate betas.

Interpretability on alphas. Cant interpret on betas since

Start trying the logit link function. Will need to re-derive the matrix of partial derivatives.

June 27th

☑ First & Second pomofocus on the writeup

- Some spell check and GEE description, working on bib files.

□ Third and pomofocus on literature review

- Reading Luna, Mansbach, and Shaw 2020
- Also zotero custimization

Literature review: Luna, Mansbach, and Shaw 2020

Could be potential for Journal club. Abstract: Longitudinal negative binomial mixed effects model Use hazard function Joint model: Longitudinal and time to event data.

Method for determining how much changes in microbiome affect disease onset. Introduction One previous method to use smoothing splines to determine time intervals in which compositions are different between groups. Prev methods only identify associations, not determine how much.

Methods: Joint model: finds associations between time-dependent covariates and event times. Uses longitudinal submodel to model time dependent covariates, and then using those in a time to event model.

Longitudinal submodel uses read counts (non Gaussian, overdispersed) Uses neg binom distribution to model overdispersed count data.

Model subject specific taxon abundances over time using negative binom linear mixed effects model. (log link) Include offset of log of total sequence reads

Event submodel: use predicted relative abundances from mixed model into a hazard function.

Tuesday June 28

□ 1st Pomofocus: Read paper

- Read some of Holmes, Harris, and Quince 2012

□ 2nd Pomofocus: Writeup

□ 3rd Pomofocus: State of research, organize, plan.

Read paper Dirichlet Multinomial Mixtures Holmes, Harris, and Quince 2012

Cluster into different groups, each group has its own Dirichlet mixture.

Writeup progress

Working more on GEE section: Things still needed to be done include:

□ modifying the partials derivation,

□ Introduction!

□ GEE algorithm

□ More on identifiability

Understanding notes from meeting

Consider a discrete valued covariate x. This mimics what I expect commonly used data to be like, for example disease status or antibiotic usage.

Consider the simulation values I have (although slightly simpler). Let $n = 500$ for simulations, which seems to give a decently valued estimate.

Let $p = 10$, the number of ASVs, let $q = 2$, for one covariate and an intercept. $x_i \in (0, 1)$, with a 50/50 split. so,

$$\mathbf{x} = (\overbrace{0 \cdots 0}^{250}, \overbrace{1 \cdots 1}^{250})$$

Then, assume the true β is

$$\beta = \begin{pmatrix} \beta_{11} = 0 \\ \vdots \\ \beta_{p1} = 0 \\ \beta_{12} = 1 \\ \vdots \\ \beta_{p/2,2} = 1 \\ \beta_{p/2+1,2} = -1 \\ \vdots \\ \beta_{p2} = -1 \end{pmatrix}$$

This follows our **constraint**:

$$\sum_{j=1}^{p} \beta_{jk} = 0 \quad \text{for all } k$$

Then we have:

$$\begin{aligned}\log(\alpha_{ij}) &= x_i\beta\\ \alpha_{ij} &= e^{x_i}\beta\end{aligned}$$

$$\begin{pmatrix}\alpha_{i=1,j=1}\\ \vdots\\ \alpha_{i=1,j=p/2+1}\\ \vdots\\ \alpha_{i=251,j=1}\\ \vdots\\ \alpha_{i=251,j=p/2+1}\\ \vdots\end{pmatrix} = \begin{pmatrix}e^{\beta_{11}+x\beta_{12}} = e^{0+0\cdot 1}\\ \vdots\\ e^{0+0\cdot -1}\\ \vdots\\ e^{0+1}\\ \vdots\\ e^{0-1}\\ \vdots\end{pmatrix}$$

Thus $\alpha_{ij} = e^0 = 1$ for $i = 1, \dots, 250, j = 1, \dots, p$, and for $i > 250$, $\alpha_{ij} = e^1$ for $j \leq p/2$ and $\alpha_{ij} = e^{-1}$ for $j > p/2$

$$\alpha_{ij} = \begin{cases}1 & i \leq 250\\ e^1 & i > 250, j \leq p/2\\ e^{-1} & i > 250, j > p/2\end{cases}$$

Idea: since there is this weird interpretation problem, maybe we should focus on correlation output and build networks??

References

Holmes, Ian, Keith Harris, and Christopher Quince (Feb. 2012). "Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics". In: *PLOS ONE* 7.2, e30126. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0030126.

Luna, Pamela N., Jonathan M. Mansbach, and Chad A. Shaw (Dec. 2020). "A Joint Modeling Approach for Longitudinal Microbiome Data Improves Ability to Detect Microbiome Associations with Disease". In: *PLOS Computational Biology* 16.12, e1008473. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1008473.