Contents

Motivation

Start writing what might be a good introduction for any paper or chapter.

Include citations!

Definitions

- ASV
- OTU
- Zero-Inflation
- Phylogenetic tree
- Compositional
- Covariates/variables/ coefficients

Notation

- Number of samples: n
- Sample index: $i : i = 1, \dots, n$
- Number of ASVs: k
- ASV index: $j : j = 1, \dots, k$
- Number of covariates: q. Note that q does not include the intercept.
- Covariate index: $k : k = 1, \dots, q$
- Responses: Transformed counts into relative abundances.
  Note that an alternative transformation to relative abundance might be good.

$$y_{ij}$$

$$\mathbf{y}_{np\times 1} = \begin{pmatrix} y_{i=1,j=1} \\ \vdots \\ y_{i=1,j=p} \\ \vdots \\ y_{i=n,j=1} \\ \vdots \\ y_{i=n,j=p} \end{pmatrix}$$

- Design matrix $\mathbf{x}$ ('little x').

$$\mathbf{x}_{n\times q} = \begin{pmatrix} x_{i=1,k=1} & \cdots & x_{i=1,k=q} \\ x_{i=2,k=1} & \cdots & x_{i=2,k=q} \\ \vdots & \ddots & \vdots \\ x_{i=n,k=1} & \cdots & x_{i=n,k=q} \end{pmatrix}$$

We don't have coefficient values measured on each ASV (i.e. no j index), they are only measured on each sample. Thus $x_{ij} = x_i$ for all $j = 1, \dots, p$. Then we convert to design matrix big X:

- Design matrix X:

$X_{np\times pq} = x \otimes I_p$

$$= \left(\begin{array}{c|c|c|c} p\left\{\begin{matrix} x_{i=1,k=1} & 0 & \cdots & 0 \\ 0 & x_{i=1,k=1} & \cdots & \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & x_{i=1,k=1} \end{matrix}\right. & \begin{matrix} x_{i=1,k=2} & 0 & \cdots & 0 \\ 0 & x_{i=1,k=2} & \cdots & \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & x_{i=1,k=2} \end{matrix} & \cdots & \\ \hline 0 & x_{i=1,k=1} & 0\cdots0 \quad 0 & x_{i=1,k=2} \\ \vdots & & & \\ & 0\cdots0 & x_{i=1,k=1} & \\ x_{i=2,k=1} & 0 & \cdots \quad x_{i=2,k=2} & \cdots \\ \vdots & & & \end{array} \right.$$

- Parameter vector β with entries β_{jk}

$$\boldsymbol{\beta}_{pq\times 1} = \begin{pmatrix} \beta_{j=1,k=1} \\ \beta_{j=2,k=1} \\ \vdots \\ \beta_{j=p,k=1} \\ \beta_{j=1,k=2} \\ \vdots \\ \beta_{j=1,k=q} \\ \vdots \\ \beta_{j=p,k=q} \end{pmatrix}$$

- Link:

  Link covariates to response.

  First assume that $\mathbf{y}$ has the same mean and covariance structure as if $\mathbf{y}_i \sim Dir(\alpha_{1p}, \dots, \alpha_{ip})$ for all $i = 1, \dots, n$.

  Then,

$$\begin{aligned} g(\alpha) = \log(\alpha) &= X\beta \\ \log(\alpha_{ij}) &= x_i\beta_j \end{aligned}$$

  Where $\beta_j = (\beta_{j,k=1}, \dots \beta_{j,k=q})^t$, and $x_i = (x_{i1}, \dots, x_{iq})$

  This link function makes sense since $\alpha > 0$.

Dirichlet ideas

Assume $\eta_i = (\eta_{i1}, \dots, \eta_{ip}) \sim Dir(\alpha_{i1}, \dots, \alpha_{ip})$, and $\sum_{j=1}^{p} \alpha_{ij} = \alpha_{i0}$

Dirichlet Expected value

$$E(\eta_{ij}) = \frac{\alpha_{ij}}{\alpha_{i0}}$$

Dirichlet variance

$$Var(\eta_{ij}) = \frac{\alpha_{ij}(\alpha_{i0} - \alpha_{ij})}{\alpha_{i0}^2(\alpha_{i0} + 1)}$$

Note: $0 \le Var(\eta_{ij}) \le 1$

Dirichlet covariance

For $i \neq j$

$$Cov(\eta_{is}, \eta_{it}) = -\frac{\alpha_{is}\alpha_{it}}{\alpha_{i0}^2(\alpha_{i0} + 1)}$$

Note this matrix is singular. Note this quantity is always negative

Dirichlet correlation

For $i \neq j$

$$\begin{aligned} Cor(\eta_{is}, \eta_{it}) &= \frac{Cov(\eta_{is}, \eta_{it})}{\sqrt{Var(\eta_{is})Var(\eta_{it})}} \\ &= -\frac{\alpha_{is}\alpha_{it}}{\alpha_{i0}^2(\alpha_{i0} + 1)} \cdot \sqrt{\frac{\alpha_{i0}^2(\alpha_{i0} + 1)}{\alpha_{is}(\alpha_{i0} - \alpha_{is})}}\sqrt{\frac{\alpha_{i0}^2(\alpha_{i0} + 1)}{\alpha_{it}(\alpha_{i0} - \alpha_{it})}} \\ &= -\sqrt{\frac{\alpha_{is}\alpha_{it}}{(\alpha_{i0} - \alpha_{is})(\alpha_{i0} - \alpha_{it})}} \end{aligned}$$

Note this is always negative

GEEs

The collection of measurements taken on a single sample are assumed to be correlated. The entries in $\mathbf{y}_i$ for a given i are correlated.

One such analysis method useful for correlated data are Generalized Estimating Equations. Originally proposed by LIANG and ZEGER 1986 for longitudinal data, this method uses a working correlation structure, which allows for any specified correlation structure. Even if the correlation structure is misspecified, the resulting estimates are consistent.

The GEE method is a marginal model, only needing the first two moments and a link function to specify the method. In other words, this method does not perform any calculations on the full distribution likelihood. Therefore this method is useful in cases when the likelihood is intractable or very difficult to work with.

We assume that

$$\begin{aligned} E(y_{ij}) &= \mu_{ij} = a'(\theta_{ij}) \\ Var(y_{ij}) &= a''(\theta_{ij}) \end{aligned}$$

Where $\theta_{ij} =$

ϕ is then the dispersion parameter.

We link the expected mean to a set of linear predictors by

$$g(\mu_i) = \mathbf{x}\boldsymbol{\beta}$$

Where g is a known invertible link function.

Instead of directly to the mean μ_i, for this parametrization of the Dirichlet distribution, we link the α Dirichlet hyperparameters to a linear predictor. To link αs, the log link is appropriate, since $\alpha > 0$

Thus:

$$\log \alpha_{ij} = x_i \beta_j$$

See above for more details.

The GEE equations are

$$\sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right)^t_{pq \times p} \mathbf{V}^{-1}_{i_{p\times p}} (\mathbf{Y_i} - \boldsymbol{\mu}_i)_{p\times 1} = 0$$

The estimator $\hat{\beta}$ is the solution to this set of equations.

- $\boldsymbol{V}_i = A_i^{\frac{1}{2}} R_i A_i^{\frac{1}{2}}$: Working covariance matrix
- $diag(A_i) = \sqrt{Var(\eta_i)}$: Square root of Dirichlet variance.
- R_i: working correlation matrix, see below, mixture of Dirichlet and phylogenetic correlation.
- $\left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)$: Matrix of partial derivatives, defined below
- Y_i: response ASV proportions $0 < Y_i < 1$
- μ_i: Expected mean: $E(Y_i) = \frac{\alpha_i}{\alpha_{i0}}$

Derivation of $\left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)$

$$
\begin{aligned}
\left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)^t &= \left(\frac{\partial}{\partial \boldsymbol{\beta}} \frac{\boldsymbol{\alpha}_i}{\alpha_{i0}}\right)^t \\
&= \left(\frac{\partial}{\partial \boldsymbol{\beta}} \frac{e^{\mathbf{x}_i \boldsymbol{\beta}}}{\sum_{j=1}^{p} e^{\mathbf{x_{ij}} \boldsymbol{\beta}}}\right)^t \\
&= \begin{pmatrix}
\frac{\partial}{\partial \boldsymbol{\beta}_1} \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}_1}}{\sum_{j=1}^{p} e^{\mathbf{x_i}^T \boldsymbol{\beta}_j}} & \cdots & \frac{\partial}{\partial \boldsymbol{\beta}_1} \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}_p}}{\sum_{j=1}^{p} e^{\mathbf{x}_i^T \boldsymbol{\beta}_j}} \\
\vdots & \ddots & \vdots \\
\frac{\partial}{\partial \boldsymbol{\beta}_p} \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}_1}}{\sum_{j=1}^{p} e^{\mathbf{x}_i^T \boldsymbol{\beta}_j}} & \cdots & \frac{\partial}{\partial \boldsymbol{\beta}_p} \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}_p}}{\sum_{j=1}^{p} e^{\mathbf{x}_i^T \boldsymbol{\beta}_j}}
\end{pmatrix} \\
&= \frac{1}{\alpha_{i0}^2} \begin{pmatrix}
x_1\alpha_1\alpha_0 - \alpha_1 x_1 \alpha_1 & -\alpha_2 x_1 \alpha_1 & \cdots & -\alpha_p x_1 \alpha_1 \\
x_2\alpha_1\alpha_0 - \alpha_1 x_2 \alpha_1 & -\alpha_2 x_2 \alpha_1 & \cdots & -\alpha_p x_2 \alpha_1 \\
\vdots & \vdots & \ddots & \vdots \\
x_q\alpha_1\alpha_0 - \alpha_1 x_q \alpha_1 & -\alpha_2 x_q \alpha_1 & \cdots & \vdots \\
-\alpha_1 x_1 \alpha_2 & x_1\alpha_2\alpha_0 - \alpha_2 x_2 \alpha_2 & & \\
\vdots & \vdots & \ddots & \vdots \\
-\alpha_1 x_q \alpha_p & \cdots & & x_q\alpha_p\alpha_0 - \alpha_p x_p \alpha_p
\end{pmatrix}_{pq \times p} \\
&= \frac{1}{a_{i0}} \begin{pmatrix}
x_1\alpha_1 & 0 & \cdots & 0 \\
x_2\alpha_1 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
x_q\alpha_1 & 0 & \cdots & 0 \\
0 & x_1\alpha_2 & \cdots & 0 \\
0 & x_2\alpha_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & x_1\alpha_p
\end{pmatrix} - \frac{1}{\alpha_{i0}^2} \begin{pmatrix}
\alpha_1^2 x_1 & \alpha_1\alpha_2 x_1 & \cdots & \alpha_1\alpha_p x_1 \\
\alpha_1^2 x_2 & \alpha_1\alpha_2 x_2 & \cdots & \alpha_1\alpha_p x_2 \\
\vdots & \vdots & \ddots & \vdots \\
\alpha_1^2 x_q & \alpha_1\alpha_2 x_q & \cdots & \alpha_1\alpha_p x_q \\
\alpha_1\alpha_2 x_1 & \alpha_2^2 x_1 & \cdots & \alpha_2\alpha_p x_1 \\
\alpha_1\alpha_2 x_2 & \alpha_2^2 x_2 & \cdots & \alpha_2\alpha_p x_2 \\
\vdots & \vdots & \ddots & \vdots \\
\alpha_1\alpha_p x_q & \alpha_2\alpha_p x_q & \cdots & \alpha_p^2 x_q
\end{pmatrix} \\
&= \frac{1}{\alpha_{i0}}[I_p \otimes \mathbf{x}_i] diag(\boldsymbol{\alpha}_i) - \frac{1}{\alpha_{i0}^2} \boldsymbol{\alpha}_i \boldsymbol{\alpha}_i^t \otimes \mathbf{x}_i
\end{aligned}
$$

GEE algorithm

Let G be the GEE equation, and H be the hessian.

Identifiability issues

Note that we can write the mean:

$$\begin{aligned}
\mu_{ij} &= \frac{\alpha_{ij}}{\alpha_{i0}} \\
&= \frac{e^{x_i \beta_j}}{\sum_{s=1}^{p} e^{x_i \beta_s}} \\
&= \frac{e^{x_i \beta_j}}{e^{x_i \beta_1} + \cdots + e^{x_i \beta_p}} \\
&= \frac{e^{\gamma}}{e^{\gamma}} \frac{e^{x_i \beta_j}}{e^{x_i \beta_1} + \cdots + e^{x_i \beta_p}} \\
&=
\end{aligned}$$

Thus μ_{ij} is the same regardless of if $\beta = (\beta_1, \dots, \beta_p)$ or $\beta = (\beta_1 + \gamma, \cdots, \beta_p + \gamma)$.

Thus the mean is non-identifiable.

Check if this is the same for the variance.

Penalty

To solve this problem, we first attempt to add a penalty.

Penalty:

$$\lambda \left(\sum_{j=1}^{p} \beta_p \right)^2 = \lambda \beta^t 1 1^t \beta$$

New GEE eqn

$$\begin{aligned} G^* &= G + \frac{\partial \text{Penalty}}{\partial \beta} \\ &= G + 2\lambda 1 1^t \beta \end{aligned}$$

New Hessian:

$$\begin{aligned} H^* &= H + \frac{\partial^2 \text{Penalty}}{\partial^2} \\ &= H + 2\lambda 1 1^t \end{aligned}$$

Approaches from here

References

LIANG, KUNG-YEE and SCOTT L. ZEGER (Apr. 1986). “Longitudinal Data Analysis Using Generalized Linear Models”. In: *Biometrika* 73.1, pp. 13–22. ISSN: 0006-3444. DOI: `10.1093/biomet/73.1.13`.