Department of Mathematical Science, KAIST

Zero-inflated regression model for microbiome compositional data analysis

Jinhwan Suk

September 29, 2020

Statistical Microbiome data analysis

History

- Microbiome researchers are interested in testing multivariate hypotheses concerning...
  1. the effects of treatments or,
  2. the effects of experimental factors on whole assemblages of bacterial taxa
  3. estimating sample sizes for such experiments

- Multivariate methods to test for differences in bacterial taxa composition between groups of metagenomic samples.
  1. Permutation test (Mantel test)
  2. Analysis of Similarity (ANOSIM)
  3. NP-Manova

Statistical Microbiome data analysis

History

- Non-parametric methods are usually **less powerful** than parametric methods.
- The power of parametric tests heavily depends on how well the model fits data.
  1. Multinomial model
  2. Dirichlet-Multinomial model (La Rosa, 2012)
- The DM model intrinsically imposes a **negative correlation** among taxon counts.
- Actual data display *both positive and negative* correlations (Mandal, 2015)
- In addition, the DM has only one dispersion parameter.

ZIGDM regression for microbiome data

Introduction

- The GDM/ZIGDM model has not been applied to microbiome data

  The additional parameters are necessary?
- GDM and ZIGDM doesn't belong to the natural exponential family and the parameter estimation is not simple.
- They developed fast **EM algorithm** to meet that challenge.
- The ZIGDM and GDM models **fit** the gut microbiome data **significantly better** than the DM model.

Which one is desirable?

In studies of the chemical composition of rats, the proportion fat is

$$\frac{\text{fat}}{\text{fat} + \text{fat-free dry matter} + \textit{water}} \quad \textbf{vs} \quad \frac{\text{fat}}{\text{fat} + \text{fat-free dry matter}}$$

In analysis of the chemical composition of rocks, **eliminate** silica or not?

Neutrality

Sometimes, it may be desirable to eliminate a proportion, say P_1, and then to analyze the proportions

$$\frac{P_2}{1-P_1}, \frac{P_3}{1-P_1}, \ldots, \frac{P_k}{1-P_1}$$

Derivation of Generalized Dirichlet Distribution

Concepts of Independence for Proportions with a Generalization of the Dirichlet Distribution (1969, JASA)

Definition (Neutrality)

Given a vector of proportions $(P_1, P_2, \ldots, P_k)$, the proportion P_1 is said to be **neutral** if P_1 is independent of the vector

$$\left(\frac{P_2}{1-P_1}, \frac{P_3}{1-P_1}, \ldots, \frac{P_k}{1-P_1}\right)$$

Definition (Neutral vector)

Given $\boldsymbol{P}$ divided so that $P = (P_{j1}, P_{j2})$. P_{j1} is a **neutral vector** if it is independent of $W_j = \left(\frac{P_{j+1}}{1-P_1\cdots-P_j}, \ldots, \frac{P_k}{1-P_1\cdots-P_j}\right)$.

If P_{j1} is neutral for all j, then $\boldsymbol{P}$ is said to be **completely neutral**.

Derivation of Generalized Dirichlet Distribution

Concepts of Independence for Proportions with a Generalization of the Dirichlet Distribution (1969, JASA)

Theorem

We consider the random variables Z_i, $i = 1, \cdots, k$ defined by $Z_i = \frac{P_i}{1-\sum_{j=1}^{i-1} P_j}$. If $\boldsymbol{P}$ is completely neutral if and only if $Z_1, Z_2, \ldots, Z_k$ are mutually independent.

Theorem

Let $\boldsymbol{P}$ be a completely neutral model with mutually independent Z's each having a specified frequency function. Then,

$$Cov(P_i, P_j) = \left[\frac{\mathbf{E}P_j}{\mathbf{E}(1-S_i)}\right]\left(K_i Var(1-S_{i-1}) - Var(P_i)\right)$$

Note that $Cov(P_i, P_j)$ can be positive.

Generalization of the Dirichlet Distribution

Suppose $\boldsymbol{P}$ is completely neutral. Then the random variables Z_i, $i = 1, \cdots, k$ are mutually independent. Let the density of each Z_i be a univariate beta distribution.

$$\frac{1}{\mathcal{B}(a_i, b_i)} z_i^{a_i - 1} (1 - z_i)^{b_i - 1}$$

where a_i, $b_i > 0$ and $\mathcal{B}(\cdot, \cdot)$ is the beta function.
The Z's can be transformed to $P_1, \ldots, P_{k-1}$ and we obtain the density function of the P's

$$GD(\boldsymbol{a}, \boldsymbol{b}) = \frac{1}{\prod_{i=1}^{k-1} \mathcal{B}(a_i, b_i)} p_k^{b_{k-1} - 1} \prod_{i=1}^{k-1} \left[p_i^{a_i - 1} \left(\sum_{j=1}^{k} p_j \right)^{b_{i-1} - (a_i + b_i)} \right]$$

where $P_k = 1 - \sum_{i=1}^{k-1} P_i$.

GD and ZIGD

GD model for random proportions

- The GDM is given by using the GD as a prior for the multinomial distribution.
- The GD is a conjugate prior for the multinomial.

$$\boldsymbol{P} \sim \text{GD}(\boldsymbol{a}, \boldsymbol{b})$$
$$\boldsymbol{Y} \sim \text{Multinomial}(\boldsymbol{P})$$

  Then, $P|Y \sim \text{GD}(\boldsymbol{a}^*, \boldsymbol{b}^*)$ where $a_j^* = a_j + Y_j$ and $b_j^* = b_j + Y_{j+1} + \cdots + Y_{K+1}$.
- The GDM has been applied to multivariate count data with complex correlation such as RNA-seq data.

GD and ZIGD

ZIGD model for random proportions with zero components for absent taxa

- The GD model assumes all observed zeros in $\boldsymbol{Y}$ are sampling zeros.
- To model absent taxa(i.e. structural zeros), we assume Z_j follows zero-inflated Beta(ZIB) distribution with parameters (π_j, a_j, b_j)
- Transforming Z's to P's, distribution of $\boldsymbol{P}$ is referred to as $ZIGD(\boldsymbol{\pi}, \boldsymbol{a}, \boldsymbol{b})$.

$$P_1 = Z_1,\ P_j = Z_j \prod_{i=1}^{j-1} (1 - Z_i)$$

- *ZIGD* is a conjugate prior for the multinomial.

ZIGDM Regression Model

Hierarchical Model

We have n subjects measured on $K+1$ taxa.

- Y_{ij} : the observed count for taxon j in subject i.
- P_{ij} : the underlying true proportion for taxon j in subject i.

Assume that the count vector $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iK})$ follows the $ZIGDM(\boldsymbol{\pi_i}, \boldsymbol{a}_i, \boldsymbol{b}_i)$

$$\Delta_{ij} \sim \text{Bernoulli}(\pi_{ij})$$

$$Z_{ij} = 0 \text{ if } \Delta_{ij} = 1, \quad Z_{ij}|\Delta_{ij} = 0 \sim \text{Beta}(a_{ij}, b_{ij})$$

$$\left(P_{i1} = Z_{i1}, \quad P_{ij} = Z_{ij} \prod_{k=1}^{j-1} (1 - Z_{ik}) \right)$$

$$\boldsymbol{Y}_i | \boldsymbol{P}_i \sim \text{Multinomial}(\boldsymbol{P}_i, N_i), \quad N_i = \sum_{j=1}^{K+1} Y_{ij} \tag{1}$$

ZIGDM Regression Model

Linking parameters to covariates

To estimate parameters in $ZIB(\boldsymbol{\pi_i}, \boldsymbol{a}_i, \boldsymbol{b}_i)$, we model

$$\mu_{ij} = \frac{a_i}{a_i + b_i}, \quad \phi_{ij} = \frac{1}{1 + a_i + b_i}$$

μ_{ij} pertains to the mean of the Beta variable and ϕ_{ij} can be viewed as the dispersion parameter

Regression model :

$$\pi_{ij} = \frac{e^{\gamma_j^T \boldsymbol{X}_i}}{1 + e^{\gamma_j^T \boldsymbol{X}_i}}, \ \mu_{ij} = \frac{e^{\alpha_j^T \boldsymbol{X}_i}}{1 + e^{\alpha_j^T \boldsymbol{X}_i}}, \ \phi_{ij} = \frac{e^{\beta_j^T \boldsymbol{X}_i}}{1 + e^{\beta_j^T \boldsymbol{X}_i}} \tag{2}$$

We write the complete parameter as $\boldsymbol{\theta} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$.

The complete data log-likelihood expressed in terms of Z's :

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= \log\left[\prod_{i=1}^{n}\left(f(\boldsymbol{Y}_i|\boldsymbol{Z}_i)\prod_{j=1}^{K}f(\boldsymbol{Z}_{ij})\right)\right] \\ &= \sum_{i=1}^{n}\log f(\boldsymbol{Y}_i|\boldsymbol{Z}_i) + \sum_{j=1}^{K}\sum_{i=1}^{n}\big\{\Delta_{ij}\log\pi_{ij} + (1-\Delta_{ij})\log(1-\pi_{ij}) + \\ &\quad (1-\Delta_{ij})[-\log\mathcal{B}(a_{ij},b_{ij}) + (a_{ij}-1)\log Z_{ij} + (b_{ij}-1)\log(1-Z_{ij})]\big\}\end{aligned}$$

where $a_{ij} = \mu_{ij}(1/\phi_{ij}-1)$ and $b_{ij} = (1-\mu_{ij})(1/\phi_{ij}-1)$.

ZIGDM Regression Model

EM Algorithm : E-step

In the t-th E-step, we need to compute the expected complete data log-likelihood,

$$Q^*_{\boldsymbol{\theta}^{(t)}} = \sum_{j=1}^{K}\sum_{i=1}^{n} \mathbb{E}_{Z_{ij},\Delta_{ij}|Y_{ij},\boldsymbol{\theta}^{(t-1)}}\big\{\Delta_{ij}\log\pi_{ij} + (1-\Delta_{ij})\log(1-\pi_{ij}) + (1-\Delta_{ij})[-\log\mathcal{B}(a_{ij},b_{ij}) + (a_{ij}-1)\log Z_{ij} + (b_{ij}-1)\log(1-Z_{ij})]\big\}$$

We have

$$\mathbb{E}_{\Delta_{ij}|Y_{ij}}[\Delta_{ij}] = \left\{ \begin{array}{ll} 0, & \text{if } Y_{ij} > 0 \\ \frac{\pi_{ij}}{\pi_{ij}+(1-\pi_{ij})\frac{\mathcal{B}(a^*_{ij},b^*_{ij})}{\mathcal{B}(a_{ij},b_{ij})}}, & \text{for } 0 \le n \le 1 \end{array} \right\}$$

$$\mathbb{E}_{Z_{ij}|Y_{ij},\Delta_{ij}=0}[\log Z_{ij}] = \psi(a^*_{ij}) - \psi(a^*_{ij}+b^*_{ij})$$

$$\mathbb{E}_{Z_{ij}|Y_{ij},\Delta_{ij}=0}[\log(1-Z_{ij})] = \psi(b^*_{ij}) - \psi(a^*_{ij}+b^*_{ij})$$

where $a^*_{ij} = a_{ij} + Y_{ij}$ and $b^*_{ij} = b_{ij} + Y_{i(j+1)} + \cdots + Y_{i(K+1)}$.

ZIGDM Regression Model

EM Algorithm : M-step

$Q^*_{\boldsymbol{\theta}^{(t)}}$ can be rewritten as

$$Q^*_{\boldsymbol{\theta}^{(t)}} = \sum_{j=1}^{K} Q^*_{\boldsymbol{\gamma}_j^{(t)}} + \sum_{j=1}^{K} Q^*_{\boldsymbol{\alpha}_j^{(t)}, \boldsymbol{\beta}_j^{(t)}}$$

In the t-th M-step, for each taxon j, we obtain $\boldsymbol{\gamma}_j^{(t)}$ from maximizing $Q^*_{\boldsymbol{\gamma}_j^{(t)}}$ and obtain $\boldsymbol{\alpha}_j^{(t)}$ and $\boldsymbol{\beta}_j^{(t)}$ from maximizing $Q^*_{\boldsymbol{\alpha}_j^{(t)}, \boldsymbol{\beta}_j^{(t)}}$.

Association Test

Setting

Test association between composition of microbiome and covariates.

1. Association with mean

$$\text{Model} : \text{Logit}(\mu_i) = \boldsymbol{\alpha}^T x_i$$
$$H_0 : \boldsymbol{\alpha}_{*1} = \boldsymbol{\alpha}_{*2} = \cdots = \boldsymbol{\alpha}_{*K} = 0$$

2. Association with dispersion(subject-specific variation)

$$\text{Model} : \text{Logit}(\phi_i) = \boldsymbol{\beta}^T x_i$$
$$H_0 : \boldsymbol{\beta}_{*1} = \boldsymbol{\beta}_{*2} = \cdots = \boldsymbol{\beta}_{*K} = 0$$

They derived score statistic and obtain P-values using permutation technique.

Association Test

Result

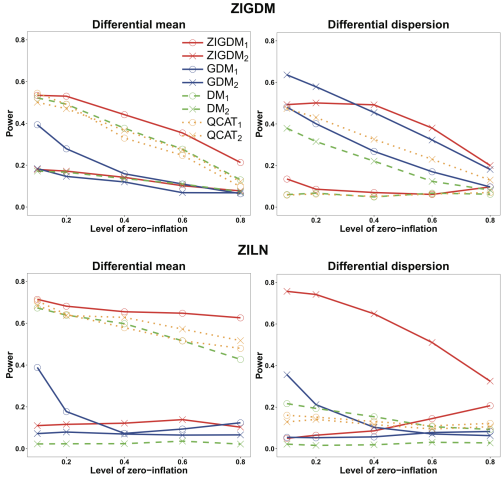

Fig. 1. Power of the permutation tests under ZIGDM and ZILN models when the sample size is 100. The pattern of variation is indicated above each graph.

Gut Microbiome and BMI

Result