Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis

Emily Palmer

Oregon State University

palmerem@oregonstate.edu

February 10, 2022

Biostatistics (2019) **20**, 4, *pp.* 698–713
doi:10.1093/biostatistics/kxy025
Advance Access publication on June 24, 2018

Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis

ZHENG-ZHENG TANG*

Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI 53715, USA and Wisconsin Institute for Discovery, Madison, WI 53715, USA

tang@biostat.wisc.edu

GUANHUA CHEN

Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI 53715, USA

Background

- Microbial data contains an abundance of zeros
- To properly model these data, zeros must be taken into account
- Using a distribution that more appropriately approximates the true zero abundances gives statistical tests more power.
- Types of zeros:
  - Structural zeros: True zeros - taxa are absent
  - Sampling zeros: Observed zeros, undersampling

Dirichlet Multinomial

- Frequently used to model microbial counts
- Imposes negative correlation
- Only one dispersion parameter, so modeling different dispersion parameters and zero-inflation levels among multiple taxa is not considered

Generalized Dirichlet Distribution

- $K+1$ taxa
- N total sequencing reads
- $\boldsymbol{Y} = (Y_1, \ldots, Y_K)$, $Y_{K+1} = N - \sum_{j=1}^{K} Y_j$ read counts
- Unobserved underlying proportions $\boldsymbol{P} = (P_1, \ldots, P_K), P_{K+1} = 1 - \sum_{j=1}^{K} P_j$
- The Generalized Dirichlet constructs the proportions P out of mutually independent Beta variables: $\boldsymbol{Z} = (Z_1, \ldots, Z_K)$, $Z_j \sim Beta(a_j, b_j)$,

$$P_1 = Z_1, \quad P_j = Z_j \prod_{i=1}^{j-1}(1 - Z_i), j = 2, \ldots, K$$

Generalized Dirichlet Density

- Density function for unobserved proportions $\boldsymbol{P}$

$$f(\boldsymbol{P}) = \prod_{j=1}^{K} \frac{1}{\mathcal{B}(a_j, b_j)} P_j^{a_j - 1}(1 - P_1 - \cdots - P_j)^{c_j}$$

  where $c_j = b_j - a_{j+1} - b_{j+1}$ for $j = 1, \ldots, K-1$ and $c_K = b_K - 1$
- $\boldsymbol{P} \sim GD(\boldsymbol{a}, \boldsymbol{b})$
- Can also obtain mutually independent Beta RVs from a GD distribution:

$$Z_1 = P_1, Z_j = \frac{P_j}{1 - \sum_{i=1}^{j-1} P_i}, j = 2, \ldots, K$$

Generalized Dirichlet Multinomial (GDM)

- The GDM distribution is created by using the GD as a prior for the multinomial distribution
- $\boldsymbol{Y}$ is multinomially distributed with $GD(\boldsymbol{a}, \boldsymbol{b})$ prior on proportion parameters $\boldsymbol{P}$
- Posterior probability of $(\boldsymbol{P}|\boldsymbol{Y})$ is $GD(\boldsymbol{a}^*, \boldsymbol{b}^*)$ with
  $\boldsymbol{a}^* = (a_1^*, \ldots, a_K^*), \boldsymbol{b}^* = (b_1^*, \ldots, b_K^*)$,
  $a_j^* = a_j + Y_j, b_j^* = b_j + Y_{J+1} + \cdots + Y_{K+1}, j = 1, \ldots, K$
- GD distribution assumes all taxa have positive proportions (every taxa present in the sample)
- Observed zeros are sampling zeros using GD distribution

Zero-inflated Generalized Dirichlet distribution

- Introduce zero-inflation to model structural zeros (absent taxa)
- Assume Z_j follows a zero-inflated Beta (ZIB) distribution $Z_j \sim ZIB(\pi_j, a_j, b_j)$
  - π_j is the probability $Z_j = 0$
- Create the ZIGD distribution by replacing Beta distributed Z_j with ZIB Z_j in GD

$$P_1 = Z_1, \quad P_j = Z_j \prod_{i=1}^{j-1}(1 - Z_i), j = 2, \ldots, K$$

- $\boldsymbol{P} \sim ZIGD(\boldsymbol{\pi}, \boldsymbol{a}, \boldsymbol{b})$

ZIGD properties

- $Z_j = 0$ implies $P_j = 0$
- $\Delta_j = I(Z_j = 0) = I(P_j = 0) \sim Bern(\pi_j)$
- If we assume $\Delta_1 = \ldots = \Delta_K = 0$, the ZIGD is the GD
- Assume there are L taxa present in a sample, $U = (u_1, \ldots, u_L)$ indeces for present taxa. $\bar{U}$ indeces of absent taxa
- M observed zero count taxa with indeces $V = (v_1, \ldots, v_M)$
- $U \cap V$ taxa present in the sample but have zero counts due to undersampling

ZIGD conjugate prior for multinomial

- $\boldsymbol{Y}$ follows a multinomial with ZIGD$(\pi, \boldsymbol{a}, \boldsymbol{b})$ prior on proportion parameters $\boldsymbol{P}$
- $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_K)$
- $\Delta_j = I(P_j = 0)$, so $f(\boldsymbol{P}, \boldsymbol{\Delta}) = f(\boldsymbol{P})$
- Posterior probability of proportions given observed counts:

$$f(\boldsymbol{P}|\boldsymbol{Y}) = f(\boldsymbol{P}, \boldsymbol{\Delta}|\boldsymbol{Y}) = f(\boldsymbol{P}|\boldsymbol{\Delta}, \boldsymbol{Y}) f(\boldsymbol{\Delta}|\boldsymbol{Y})$$

ZIGD conjugate prior for multinomial

- Since $P_j = 0$ when $\Delta_j = 1$ (taxon has proportion of 0 if it is absent from a sample)

$$f(\boldsymbol{P}|\boldsymbol{\Delta}) = I(\boldsymbol{P}_{\bar{U}} = 0) f(\boldsymbol{P}_U|\boldsymbol{\Delta}_U = 0, \boldsymbol{\Delta}_{\bar{U}} = 1)$$

- Since $f(\boldsymbol{P}_U|\boldsymbol{\Delta}_U = 0, \boldsymbol{\Delta}_{\bar{U}} = 1) \sim GD(\boldsymbol{a}_U, \boldsymbol{b}_U)$, and the GD is conjugate prior for the multinomial, the posterior probability is

$$f(\boldsymbol{P}_U|\boldsymbol{\Delta}_U = 0, \boldsymbol{\Delta}_{\bar{U}} = 1, \boldsymbol{Y}) \sim GD(\boldsymbol{a}^*_U, \boldsymbol{b}^*_U)$$

- $\Delta_j = 0$ when $Y_j > 0$ (a taxon is present in the sample if the count is positive)

$$f(\boldsymbol{\Delta}, |\boldsymbol{Y}) = I(\boldsymbol{\Delta}_{\bar{V}} = 0) f(\boldsymbol{\Delta}_V|\boldsymbol{Y}_V = 0, Y_{\bar{V}} > 0)$$

- The mass function for $\boldsymbol{\Delta}_{\mathcal{V}}$ given $\boldsymbol{Y}_{\mathcal{V}} = 0, \boldsymbol{Y}_{\bar{\mathcal{V}}} > 0$ is then

$$
\begin{aligned}
& f(\boldsymbol{\Delta}_{\mathcal{V}} \mid \mathbf{Y}_{\mathcal{V}} = \mathbf{0}, \mathbf{Y}_{\overline{\mathcal{V}}} > \mathbf{0}) \\
& \propto f(\boldsymbol{\Delta}_{\mathcal{V}}) f(\mathbf{Y}_{\mathcal{V}} = \mathbf{0}, \mathbf{Y}_{\overline{\mathcal{V}}} > \mathbf{0} \mid \boldsymbol{\Delta}_{\mathcal{V}}, \boldsymbol{\Delta}_{\overline{\mathcal{V}}} = \mathbf{0}) \\
& = f(\boldsymbol{\Delta}_{\mathcal{V}}) \int_{\mathbf{P}} f(\mathbf{Y}_{\mathcal{V}} = \mathbf{0}, \mathbf{Y}_{\overline{\mathcal{V}}} > \mathbf{0} \mid \mathbf{P}, \boldsymbol{\Delta}_{\mathcal{V}}, \boldsymbol{\Delta}_{\overline{\mathcal{V}}} = \mathbf{0}) f(\mathbf{P} \mid \boldsymbol{\Delta}_{\mathcal{V}}, \boldsymbol{\Delta}_{\overline{\mathcal{V}}} = \mathbf{0}) \mathrm{d}\mathbf{P} \\
& \propto \prod_{i \in \mathcal{V}} \left\{ \pi_j^{\Delta_j} (1 - \pi_j)^{(1 - \Delta_j)} \right\} \prod_{i \in \mathcal{U} \cap \mathcal{V}} \left\{ \frac{\mathcal{B}(a_j^*, b_j^*)}{\mathcal{B}(a_j, b_j)} \right\} \\
& = \prod_{j \in \mathcal{V}} \left\{ \pi_j^{\Delta_j} \left[(1 - \pi_j) \frac{\mathcal{B}(a_j^*, b_j^*)}{\mathcal{B}(a_j, b_j)} \right]^{(1 - \Delta_j)} \right\}.
\end{aligned}
$$

- Thus the posterior probability $f(\boldsymbol{P}|\boldsymbol{Y})$ follows a ZIGD with the zero-inflation on the taxa having observed zero counts
- Probability of an observed zero being structural is

$$
\frac{\pi_j}{\pi_j + (1 - \pi_j) \frac{\mathcal{B}(a_j^*, b_j^*)}{\mathcal{B}(a_j, b_j)}} \quad (j \in \mathcal{V})
$$

ZIGDM Regression Model

- n subjects measured on $K+1$ taxa
- Y_{ij}, P_{ij} observed count and true proportion for taxon j in subject i.
- $\boldsymbol{X}_i$ d-dimensional vector of intercept, covariates, and confounding variables
- Assume $\boldsymbol{Y}_i$ follows a ZIGDM$(\boldsymbol{\pi}_i, \boldsymbol{a}_i, \boldsymbol{b}_i)$
- The model is then

$$\Delta_{ij} \sim \text{Bernoulli}(\pi_{ij}), \;\; j = 1, \ldots, K,$$

$$Z_{ij} = 0 \text{ if } \Delta_{ij} = 1, \;\; Z_{ij} \mid \Delta_{ij} = 0 \sim \text{Beta}(a_{ij}, b_{ij}), \;\; j = 1, \ldots, K,$$

$$P_{i1} = Z_{i1}, \;\; P_{ij} = Z_{ij} \prod_{k=1}^{j-1} (1 - Z_{ik}), \;\; j = 2, \ldots, K,$$

$$\mathbf{Y}_i \mid \mathbf{P}_i \sim \text{Multinomial}(\mathbf{P}_i, N_i), \text{ where } \mathbf{P}_i = (P_{i1}, \ldots, P_{iK}) \text{ and } N_i = \sum_{j=1}^{K+1} Y_{ij}.$$

Link functions

- With this model, we can link π_i, $\boldsymbol{a}_i$, and $\boldsymbol{b}_i$ to $\boldsymbol{X}_i$
- π_{ij} is the probability of absence
- a_{ij}, b_{ij} control abundance distribution at the presence
- Reparametrize $\mu_{ij} = \frac{a_{ij}}{a_{ij}+b_{ij}}$ and $\sigma_{ij} = \frac{1}{1+a_{ij}+b_{ij}}$ as the mean of the Beta variable and the dispersion parameter (since the variance of the Beta variables are of the form $\mu_{ij}(1-\mu_{ij})\sigma_{ij}$)
- Use logit link functions ($\mu_{ij}, \mu_{ij}, \sigma_{ij}$ all between 0 and 1)

$$\pi_{ij} = \frac{e^{\boldsymbol{\gamma}_j^{\mathrm{T}}\mathbf{x}_i}}{1+e^{\boldsymbol{\gamma}_j^{\mathrm{T}}\mathbf{x}_i}}, \quad \mu_{ij} = \frac{e^{\boldsymbol{\alpha}_j^{\mathrm{T}}\mathbf{x}_i}}{1+e^{\boldsymbol{\alpha}_j^{\mathrm{T}}\mathbf{x}_i}}, \quad \text{and} \quad \sigma_{ij} = \frac{e^{\boldsymbol{\beta}_j^{\mathrm{T}}\mathbf{x}_i}}{1+e^{\boldsymbol{\beta}_j^{\mathrm{T}}\mathbf{x}_i}}, \quad j = 1, \ldots, K,$$

- $\boldsymbol{\gamma}_j = (\gamma_{1j}, \ldots, \gamma_{dj})$, $\boldsymbol{\alpha}_j = (\alpha_{1j}, \ldots, \alpha_{dj})$, $\boldsymbol{\beta}_j = (\beta_{1j}, \ldots, \beta_{dj})$ regression coefficients for taxon j.

Estimating parameters

- Denote $\boldsymbol{\theta} = (\gamma_1, \ldots, \gamma_K, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$ the complete set of parameters
- Likelihood-based inference on $\boldsymbol{\theta}$ is difficult since the observed log-likelihood function is complicated
- Complete data log-likelihood in terms of Z (instead of $\boldsymbol{P}$)

$$l(\boldsymbol{\theta}) = \log\left[\prod_{i=1}^{n}\left\{f(\mathbf{Y}_i \mid \mathbf{Z}_i)\prod_{j=1}^{K} f(Z_{ij})\right\}\right]$$

$$= \sum_{i=1}^{n} \log\{f(\mathbf{Y}_i \mid \mathbf{Z}_i)\}$$

$$+ \sum_{j=1}^{K}\sum_{i=1}^{n}\Big\{\Delta_{ij}\log\pi_{ij} + (1-\Delta_{ij})\log(1-\pi_{ij}) +$$

$$(1-\Delta_{ij})\left[-\log(\mathcal{B}(a_{ij}, b_{ij})) + (a_{ij}-1)\log(Z_{ij}) + (b_{ij}-1)\log(1-Z_{ij})\right]\Big\}$$

EM algorithm for estimating parameters

- In the t-th E-step, compute the expected complete data log-likelihood:

$$Q_\theta^* = \sum_{j=1}^{K}\sum_{i=1}^{n} \mathrm{E}\Big\{\Delta_{ij}\log\pi_{ij} + (1-\Delta_{ij})\log(1-\pi_{ij}) + (1-\Delta_{ij})\left[-\log(\mathcal{B}(a_{ij},b_{ij})) + (a_{ij}-1)\log Z_{ij} + (b_{ij}-1)\log(1-Z_{ij})\right]\Big\}$$

- Expectation is with respect to the posterior distributions of $(\boldsymbol{\Delta}_i|\boldsymbol{Y}_i;\boldsymbol{\theta}^{(t-1)})$ and $(\boldsymbol{Z}_i|\boldsymbol{\Delta}_i,\boldsymbol{Y}_i;\boldsymbol{\theta}^{t-1})$
- $\boldsymbol{\theta}^{(t-1)}$ the parameter estimates in the $(t-1)$-th M-step

E - step

$$\Delta_{ij}^* = \mathrm{E}\left(\Delta_{ij} \mid \mathbf{Y}_i\right) = \begin{cases} 0 & \text{if } Y_{ij} > 0 \\ \dfrac{\pi_{ij}}{\pi_{ij} + (1-\pi_{ij})\dfrac{\mathcal{B}\left(a_{ij}^*, b_{ij}^*\right)}{\mathcal{B}\left(a_{ij}, b_{ij}\right)}} & \text{if } Y_{ij} = 0 \end{cases},$$

$$A_{ij}^* = \mathrm{E}\left(\log Z_{ij} \mid \mathbf{Y}_i, \Delta_{ij} = 0\right) = \psi\left(a_{ij}^*\right) - \psi\left(a_{ij}^* + b_{ij}^*\right),$$

$$B_{ij}^* = \mathrm{E}\left(\log(1 - Z_{ij}) \mid \mathbf{Y}_i, \Delta_{ij} = 0\right) = \psi\left(b_{ij}^*\right) - \psi\left(a_{ij}^* + b_{ij}^*\right)$$

- $a_{ij}^* = a_{ij} + Y_{ij}$, $b_{ij}^* = b_{ij} + Y_{i(j+1)} + \cdots + Y_{i(K+1)}$, $\psi()$ is the digamma function
- Rewrite Q_θ^* as:

$$Q_\theta^* = \sum_{j=1}^{K} Q_{\gamma_j}^* + \sum_{j=1}^{K} Q_{\alpha_j, \beta_j}^*$$

- $Q_{\gamma j}^* = \sum_{i=1}^{n} \{\Delta_{ij}^* \log \pi_{ij} + (1 - \Delta_{ij}^*) \log(1 - \pi_{ij})\}$ and $Q_{\alpha_j, \beta_j}^* = \sum_{i=1}^{n} (1 - \Delta_{ij}^*)\{-\log(\mathcal{B}(a_{ij}, b_{ij})) + (a_{ij} - 1)A_{ij}^* + (b_{ij} - 1)B_{ij}^*\}$

M-Step

- In the t-th M-step, for each taxon j, obtain $\boldsymbol{\gamma}_j^{(t)}$ from maximizing $Q^*_{\gamma_j}$, and obtain $\boldsymbol{\alpha}_j^{(t)}, \boldsymbol{\beta}_j^{(t)}$ by maximizing $Q^*_{\alpha_j,\beta_j}$
- Computation for optimization is the same as a logisitic and weighted Beta regression
- Parameters for individual taxa updated independently, estimation can be done in parallel for different parameters
- Results in a computationally efficient EM algorithm relying on simple posterior estimation calculations and parallel parameter updates for taxa

Association tests

- Test null hypothesis that covariates are not associated with the mean $H_0 : \boldsymbol{\alpha}_{*1} = \cdots = \boldsymbol{\alpha}_{*K} = 0$
- Or covariates are not associated with dispersion $H_0 : \boldsymbol{\beta}_{*1} = \cdots = \boldsymbol{\beta}_{*K} = 0$
- $\boldsymbol{\alpha}_{*j}, \boldsymbol{\beta}_{*j}$ subsets of $\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j$ corresponding to covariates of intersest
- Can test using score, Wald, or likelihood-ratio statistics
- This paper uses score statistics, which where computationally faster and more stable
- Use permutation techniques for p values since asymptotic approximations may not be accurate when most observations are zero
- Permute covariate of interest and calculate score test statistic in each permutation

Thank you!