Supplemental material for "Conditional regression based on a multivariate zero-inflated logistic normal model for microbiome relative abundance data"

Zhigang Li[1,2,3,4*], Katherine Lee[5], Margaret R. Karagas[3,4], Juliette C. Madan[3,4,6], Anne G. Hoen[1,3,4], James O'Malley[1,7], Hongzhe Li[8]

[1]Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth, 1 Medical Center Drive, Lebanon, NH 03756, USA, [2]Department of Biostatistics, University of Florida, Gainesville, FL, 32611, USA, [3]Children's Environmental Health and Disease Prevention Research Center at Dartmouth, Hanover, New Hampshire, [4]Department of Epidemiology, Geisel School of Medicine at Dartmouth, 1 Medical Center Drive, Lebanon, NH 03756, USA, [5]Phillips Exeter Academy, Exeter, NH 03833, USA, [6]Division of Neonatology, Department of Pediatrics, Children's Hospital at Dartmouth, Lebanon, New Hampshire, [7]The Dartmouth Institute for Health Policy and Clinical Practice, Geisel School of Medicine at Dartmouth, 1 Medical Center Drive, Lebanon, NH 03756, USA and [8]Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA

S1). Regression model for the discrete part.

In addition to the regression model (2) which regresses the mean of the non-zero RA on covariates, we can also model the discrete part of the distribution in relation to the covariates by allowing the parameters $\{p_1, p_2, \dots, p_{1,2\dots,K+1}\}$ depend on the covariates thru the following regression equations:

$$logit(p_1) = (1, x^T)\alpha_1$$

$$logit(p_2) = (1, x^T)\alpha_2$$

$$\dots.$$

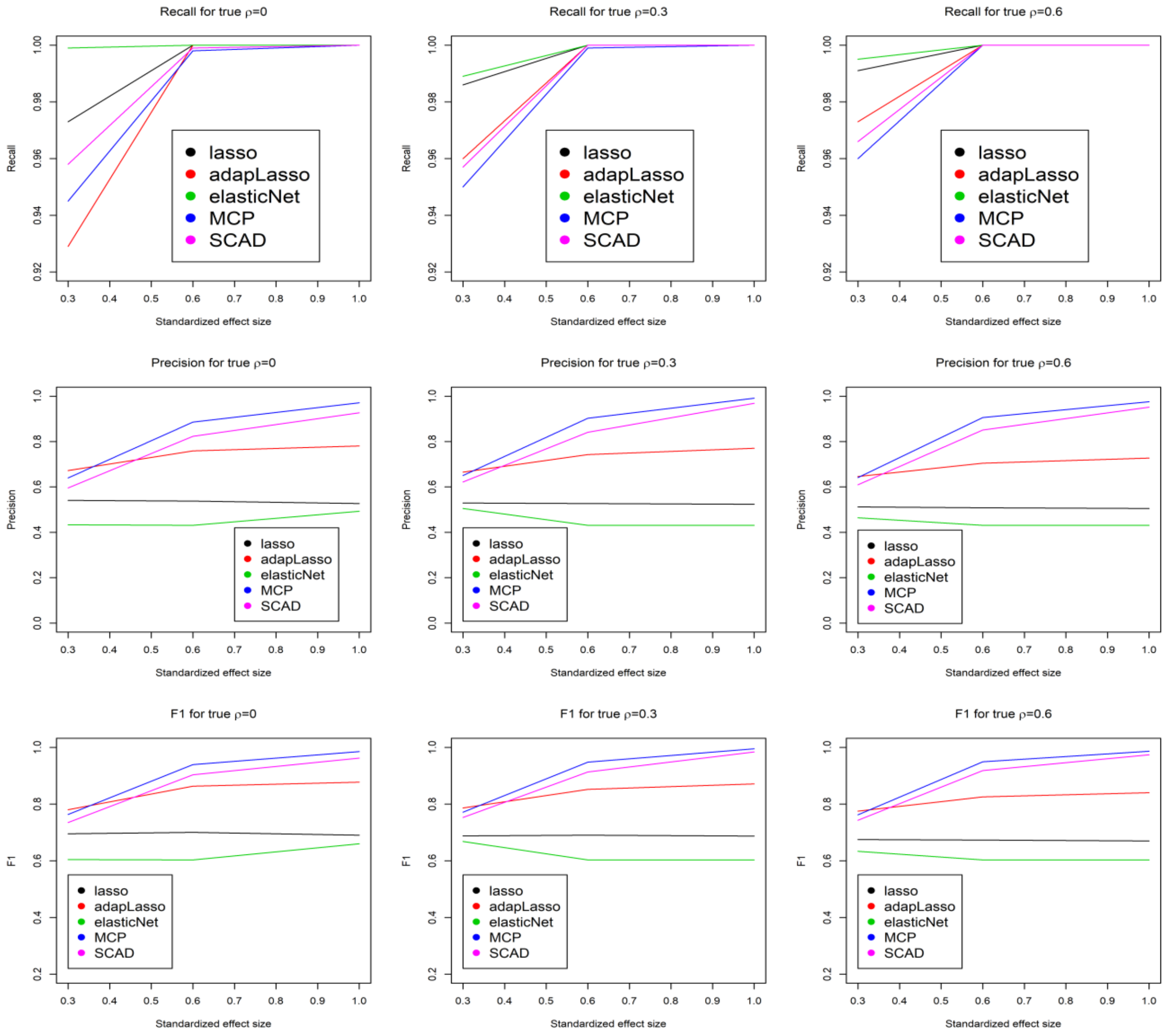
$$logit(p_{1,2\dots,K+1}) = (1, x^T)\alpha_{1,2\dots,K+1},$$

where x denotes the covariate vector (subject index is suppressed). The parameters α_1, α_2,…, $\alpha_{1,2\dots,K+1}$ are all $(Q+1)$-dimensional parameter vectors including intercepts. The complete log-likelihood function denoted by ℓ_c, can be written as following:

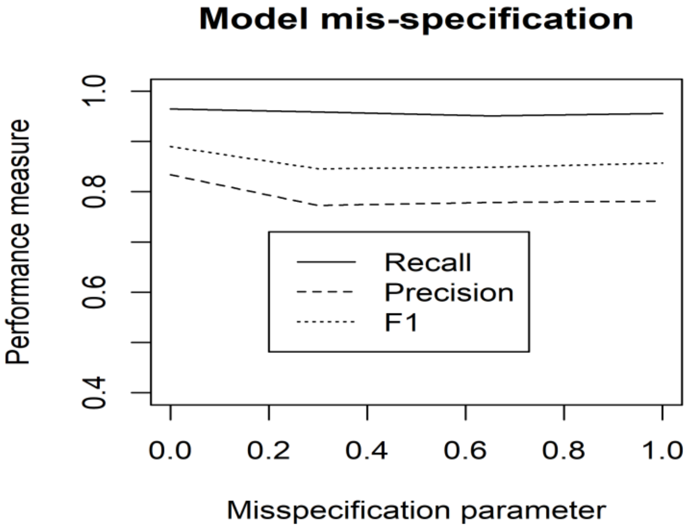
$$\begin{aligned}\ell_c &= \sum_i \log\left(p^i_{k_1,\dots,k_L}\right) + \sum_i 0.5\log\left|\mathrm{A}_i\Sigma\mathrm{A}_i^T\right|^{-1} \\ &\quad - 0.5\sum_i \left(U^i_{k_1,\dots,k_L} - A_iX_i\beta\right)^T (\mathrm{A}_i\Sigma\mathrm{A}_i^T)^{-1}\left(U^i_{k_1,\dots,k_L} - A_iX_i\beta\right) + \text{constant} \\ &= \sum_i \log\left(\text{expit}\left((1, x_i^T)\alpha_{k_1,\dots,k_L}\right)\right) + \sum_i 0.5\log\left|\mathrm{A}_i\Sigma\mathrm{A}_i^T\right|^{-1} \\ &\quad - 0.5\sum_i \left(U^i_{k_1,\dots,k_L} - A_iX_i\beta\right)^T (\mathrm{A}_i\Sigma\mathrm{A}_i^T)^{-1}\left(U^i_{k_1,\dots,k_L} - A_iX_i\beta\right) + \text{constant},\end{aligned}$$

where $\text{expit}(x) = \frac{\exp(x)}{1+\exp(x)}$ and i is subject index. It is straightforward to see that the parameters α_1, α_2,…, $\alpha_{1,2\dots,K+1}$ can be treated as nuisance parameters if the target of inference is β.

S2). Comparison across different regularization approaches. In this comparison, 300 subjects were generated with 400 taxa and 6 covariates were generated. Under the ZILN model, there are 2394 regression coefficients in model. We set 1588 coefficients to 0, and thus the model size (ie, the number of non-zero coefficients) is 1204. All the non-zero coefficients were set to have the same value in the range from 0.3 to 1. The standard deviation σ is set to be 1 and thus the regression coefficients are standardized effect sizes. We also set $\rho_X = 0.5$, $\rho = 0.5$ and data sparsity level was set to be 0.54. The regression coefficient vector β has a smaller length and is less sparse than the main setting presented in the paper. We did this set of simulation prior to the main simulation. The results (see figures below) showed that all the regularization approaches have good recall rates (>0.9) which means that they are all very powerful for picking up true non-zero coefficients. However, the LASSO approaches including elastic net and adaptive LASSO had worse precision rates because they tend to pick up more false positives (ie, zero coefficients in this case). MCP and SCAD had better precision rates than LASSO approaches. MCP was a little bit better than SCAD and had the highest F1 scores.



S3). Robustness to distribution misspecification. We assessed the performance of our approach when the distribution is mis-specified. We add a perturbation to the residual so that the distribution is mis-specified: $U^i = X_i\beta + (1-\gamma)\varepsilon^i + \gamma\sigma(\delta^i - 1)$ where $0 \le \gamma \le 1$, i is subject index and δ^i is a random vector with each element following the chi-square distribution with degrees of freedom of 1. The two random vectors ε^i and δ^i are independent. Notice that $\gamma = 0$ corresponds to the correctly specified distribution, $\gamma = 1$ corresponds to a completely mis-specified distribution, and $0 < \gamma < 1$ corresponds to a partially mis-specified distribution. We studied four cases: $\gamma = 0, 0.3, 0.65, 1$ and plotted the performance measures in the following figure. In these simulations, we set $\rho_X = 0.85$, SNR=4.5 and the non-zero elements of the β vector were generated from a uniform distribution over the interval $[-7,-4) \cup (4,7]$. Other settings are the same as described at the beginning of Section 3.2. It showed that the recall rate is very robust to the mis-specification. Precision and F1 dropped a little bit when γ departs from 0, but they remained stable thereafter.



S4). Sensitivity analysis for choosing different reference taxon. To study the performance of our approach with respect to different reference taxon, we randomly selected two reference taxa and conducted two set of simulation. In the simulations, we set $\rho_X = 0.85$, $\gamma = 0$, SNR=4.5 and the non-zero elements of the β vector were generated from a uniform distribution over the interval $[-7,-4) \cup (4,7]$. Other settings are the same as described at the beginning of Section 3.2. The results (see table below) showed that the impact of the reference taxon is minimally different compared with the case where the true reference taxon (ie, the reference taxon used in the data generation) is used. Recall rate is fairly stable across all cases. Precision and F1 had less than 8% drop for the randomly selected reference taxon, but they stayed stable across the two cases with randomly selected reference taxon.

	True reference taxon	Randomly selected taxon 1	Randomly selected taxon 2
Recall	0.965	0.948	0.938
Precision	0.834	0.781	0.772
F1	0.89	0.853	0.819

S5). Proof of $\mathbf{1}_{(Y_k>0)} = \boldsymbol{Z_k}$ for $\boldsymbol{k} = \mathbf{1}, \dots, \boldsymbol{K} + \mathbf{1}$.

Proof: by the definition in the first paragraph of Section 2.2, we have the observed RA vector

$$Y = \left(\frac{Y_1^* Z_1}{\sum_{j=1}^{K+1} Y_j^* Z_j}, \dots, \frac{Y_{K+1}^* Z_{K+1}}{\sum_{j=1}^{K+1} Y_j^* Z_j}\right)^T, \qquad (S5-1)$$

where Y_k^* is the true RA of the kth taxon and $Z_k = 1/0$ to indicate the observed kth taxon being positive/zero. By the definition of Z_k, we know that $Z_k = 1$ implies $Y_k > 0$. Next, we prove that $Y_k > 0$ implies $Z_k = 1$. From equation $(S5-1)$, the observed RA for the kth taxon is $Y_k = \frac{Y_k^* Z_k}{\sum_{j=1}^{K+1} Y_j^* Z_j}$, $k = 1, \dots, K+1$. We can further write this equality as:

$$Y_k = \begin{cases} 0, & Z_k = 0 \\ \dfrac{Y_k^*}{\sum_{j=1}^{k-1} Y_j^* Z_j + Y_k^* + \sum_{j=k+1}^{K+1} Y_j^* Z_j}, & Z_k = 1 \end{cases}, \quad k = 1, \dots, K+1$$

From the above equality, it is straightforward to see that $Y_k > 0$ implies $Z_k = 1$. Taken together, we have $1_{(Y_k>0)} = Z_k, k = 1, \dots, K+1$, where $1_{(.)}$ is an indicator function.