Resampling-based Inferences for Compositional Regression with Application to Beef Cattle Microbiomes

Sujin Lee[1], Sungkyu Jung[1], Jeferson Lourenco[2], Dean Pringle[2], and Jeongyoun Ahn[3, *]

[1]Department of Statistics, Seoul National University
[2]Department of Animal and Dairy Science, University of Georgia
[3]Department of Industrial and Systems Engineering, KAIST
[*]Corresponding author: jyahn@kaist.ac.kr

September 19, 2022

Abstract

Gut microbiomes are increasingly found to be associated with many health related characteristics of human as well as animal. A regression with compositional microbiomes covariates are commonly used to identify important bacterial taxa that are related to various phenotype responses. Often the dimension of microbiome taxa easily exceeds the number of available samples, which creates a serious challenge in estimation and inference of the model. The sparse log-contrast regression method is useful for such case as it can yield a model estimate that depends on only a small number of taxa. However, a formal statistical inference procedure for individual regression coefficients has not been properly established yet. We propose a new estimation and inference procedure for linear regression models with extremely low-sample sized compositional predictors. Under the compositional log-contrast regression framework, the proposed approach consists of two steps. The first step is to screen relevant predictors by fitting a log-contrast model with a sparse penalty. The screened-in variables are used as predictors in the non-sparse log-contrast model in the second step, where each of the regression coefficients is tested using nonparametric, resampling-based methods such as permutation and bootstrap. The performances of the proposed methods are evaluated by a simulation study, which shows they outperform traditional approaches based on normal assumptions or large sample asymptotics. Application to steer microbiomes data successfully identifies key bacterial taxa that are related to important cattle quality measures.

Availability and implementation: The sequencing data used in this article are publicly available and can be found at: https://www.mg-rast.org using the accession number: mgm4909317.3. The source code can be accessed at https://github.com/Sujin-L/Resampling-based-Inference.

1 Introduction

In recent years the human gut microbiomes study has become a growing area of research, since it has been increasingly clear that the microbiomes play a major role in health

and disease in humans (NIH Human Microbiome Portfolio Analysis Team, 2019). Extensive research has also been conducted on microbiomes of livestock such as pigs and cows (O'Hara et al., 2020; Bergamaschi et al., 2020). Similarly to humans, animal microbiomes are believed to be related to various traits of animals. For example, in the case of beef cattle, feed efficiency, fat thickness, yield grade, and marbling scores are a few examples of important economical features that are believed to be related to gut microbiomes in the animal (Lourenco et al., 2020).

In studies of microbiomes, the 16S ribosomal RNA (rRNA) gene targeted sequencing is commonly used for analysis (Li, 2015). From the sequences, operational taxonomic units (OTUs) are assigned to different taxonomic ranks, such as species, genus, family, class and phylum. For a given sample and taxonomic level, the resulting data are tables of read counts for each taxon. Since there is no information in the actual read numbers per sample, relative abundance values are subject for analysis. Such data sets are referred to as compositional data (Aitchison, 1982).

Our motivating data problem is a regression analysis with microbiome data from beef cattle steers. The data used in this study came from 20 animals. Three different parts of their gastrointestinal tract were sampled: rumen, cecum and rectum (i.e., feces). Microbial DNA was purified from the samples, and amplification of the 16S rRNA gene was performed as previously described by Lourenco et al. (2020). Paired-end DNA sequences were merged, quality-filtered, and clustered into OTUs at 97% similarity using the QIIME v1.9.1 pipeline (Caporaso et al., 2010). Taxonomic analyses were performed at several different levels, and the results presented in the current study are the ones obtained at the family-level, which resulted in a matrix of 36 taxa and 20 samples, whose rows are percentage with sum to 100. The data can be viewed as a 20×36 matrix ($\mathbf{X}$), whose rows represent the animal, and the columns correspond to the family level taxon.

Four different traits regarded as important in beef cattle production systems were considered in this study: residual feed intake, yield grade, back fat thickness, and percentage of lipid in the carcass. A main interest of this study was to identify microbial taxa that are related to those key traits of a cattle. In order to do so, we considered a regression model with each of the traits as the response variable and the microbiome information as predictors, which presents two main challenges. First, with the given sample size (20) and the number of taxa (36), we have a large p small n, i.e., High Dimension, Low Sample Size (HDLSS) problem, thus a regularization is inevitable since the regression model is not estimable due to over-parametrization. Second, even when a model is estimated, determining whether or not individual coefficients are significant is not straightforward since the classical test framework such as t-test will not be properly justified.

Lee et al. (2016) and Javanmard and Montanari (2014) proposed an inference method that can be applied for lasso estimates of regression coefficients in the case when $n < p$. However, since we consider a lasso estimation with an additional constraint, these methods are not readily applicable. Therefore, in this work, we propose a two-stage procedure for *HDLSS compositional regression* modeling, with an emphasis on identifying significant predictors. Our proposed procedure consists of two parts: variable screening followed by individual significance tests. First, we screen out non-significant predictor variables using compositional lasso (Lin et al., 2014; Bates and Tibshirani, 2019; Shi et al., 2016). The screening step is introduced in order to reduce the variability in the subsequent inferences in the next step. The initial compositional model is regularized using a lasso so as to

encourage more variables to be screened for the subsequent inferential step. Second we fit a compositional regression model with the screened-in variables only, for which we test the significance of individual variables. For this non-conventional testing problem, we consider several alternatives the classical Student's t-test, such as permutation and bootstrap. Note that a main purpose of this study is to explore taxa that are significantly related to the response and not necessarily to build a predictive model. Nonetheless we discuss how a predictive regression model can be constructed based on the chosen variables in Section 3.2.3.

The rest of the article is organized as follows. In Section 2, we introduce a constrained linear log-contrast model for compositional covariates and discuss the least squares estimators and their properties. Section 3 presents our two-stage procedure for estimating a sparse log-contrast model with only statistically significant predictors. An extensive simulation study to evaluate the proposed approach is presented in Section 4 and the motivating beef cattle microbiome data are analyzed and discussed in Section 5. Finally Section 6 concludes this work. Proofs of theoretical results and additional results tables are in the Supplementary Material.

2 Log-contrast Regression for Compositional Data

Let $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^T \in \mathbb{R}^{n\times p}$ be the matrix of observed compositional variables, where $\mathbf{x}_i$ is a composition vector that lies in the $(p-1)$-dimensional positive simplex $\mathbb{S}^{p-1} = \{(x_1, \cdots, x_p) : x_j > 0\ j = 1, \cdots, p,\ \sum_{j=1}^{p} x_j = 1\}$. Now suppose that we also observe a continuous variable that is assumed to be related to the compositional variables in $\mathbf{X}$. Let $\mathbf{y} \in \mathbb{R}^n$ denote the vector of the observed response. Relating $\mathbf{X}$ and $\mathbf{y}$ via a linear regression model encounters many problems due to the geometrical constraints of the compositional variables. For instance, one cannot interpret the partial regression coefficients in a usual way since it is impossible to increase one variable while holding other predictors constant due to the fixed sum.

The log-contrast model (Aitchison and Bacon-Shone, 1984) is a popular alternative regression model for compositional predictors. First we assume that all compositions are strictly positive. If this is not the case, replacing zeros by a small positive number is a common remedy. See Lubbe et al. (2021) for different strategies for the zero replacement and their comparisons. Then we define the log-ratio values, i.e., $z_{ij} = \log(x_{ij}/x_{ip})$, $j = 1, \ldots, p-1$, in which the pth variable is used as a reference. The linear log-contrast model is

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{Z}^p\boldsymbol{\beta}_{\setminus p} + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{Z}^p \in \mathbb{R}^{n\times(p-1)}$ is the matrix with the log-ratio values z_{ij}; $\boldsymbol{\beta}_{\setminus p} = (\beta_1, \cdots, \beta_{p-1})^T$ is the regression coefficient vector; $\boldsymbol{\epsilon}$ is the n-vector of independent noise with zero mean and variance σ^2. To avoid the ambiguity of choosing the reference component, the model (1) is often re-formulated as the following model with a linear constraint by letting $\beta_p = -\sum_{j=1}^{p-1}\beta_j$ (Lin et al., 2014; Srinivasan et al., 2021):

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \sum_{j=1}^{p}\beta_j = 0, \tag{2}$$

where $\mathbf{1} = (1, \cdots, 1)^T$, $\mathbf{Z} = \log\mathbf{X}$ is the $n \times p$ matrix and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$.

Estimating the model (2) is done via a constrained least squares method when the sample size n is larger than p. Note that we can re-write the model (2) as a centered model by estimating the intercept term μ by $\hat{\mu} = \bar{\mathbf{y}} - \bar{\mathbf{Z}}\hat{\boldsymbol{\beta}}$:

$$\mathbf{y} - \bar{\mathbf{y}} = (\mathbf{Z} - \bar{\mathbf{Z}})\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \mathbf{1}^T\boldsymbol{\beta} = 0,$$

where $\bar{\mathbf{y}} = \bar{y}\mathbf{1}$ with $\bar{y} = \sum_{i=1}^n y_i/n$ and $\bar{\mathbf{Z}} = (\bar{\mathbf{z}}_1, \cdots, \bar{\mathbf{z}}_p)$ and $\bar{\mathbf{z}}_j = \bar{z}_j\mathbf{1}$ with $\bar{z}_j = \sum_{i=1}^n \log x_{ij}/n$. For notational convenience, suppose $\mathbf{y}$ and $\mathbf{Z}$ are centered response and predictor variables respectively. Then we rewrite the model as

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \mathbf{1}^T\boldsymbol{\beta} = 0, \tag{3}$$

and we obtain the following objective function

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}\|_2^2, \quad \text{subject to } \mathbf{1}^T\boldsymbol{\beta} = 0, \tag{4}$$

where $\|\cdot\|_2$ denotes the ℓ_2 norm. The following lemma gives a closed-form solution to the above minimization problem.

Lemma 1. *The closed-form solution of* (4) *is obtained as follows:*

$$\hat{\boldsymbol{\beta}} = [\mathbf{I} - (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{1}(\mathbf{1}^T(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{1})^{-1}\mathbf{1}^T]\hat{\boldsymbol{\beta}}_{ols} \tag{5}$$

where $\hat{\boldsymbol{\beta}}_{ols} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{y}$.

It is straightforward to see that $\hat{\boldsymbol{\beta}}$ is unbiased and its variance is given by

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2\mathbf{M}(\mathbf{Z}^T\mathbf{Z})^{-1}, \tag{6}$$

where $\mathbf{M} = \mathbf{I} - (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{1}(\mathbf{1}^T(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{1})^{-1}\mathbf{1}^T$.

Note that the variance of the constrained estimator (5) has a smaller variance than the ordinary least squares estimator $\hat{\boldsymbol{\beta}}_{ols}$ as the variance in (6) can be re-rewritten as

$$\begin{aligned}\mathrm{Var}(\hat{\boldsymbol{\beta}}) =& \sigma^2(\mathbf{Z}^T\mathbf{Z})^{-1} \\ &- \sigma^2(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{1}(\mathbf{1}^T(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{1})^{-1}\mathbf{1}^T(\mathbf{Z}^T\mathbf{Z})^{-1}.\end{aligned}$$

To test the significance of each regression coefficient, we consider testing the following hypotheses: For $j = 1, \ldots, p$,

$$H_0 : \beta_j = 0 \quad vs. \quad H_1 : \beta_j \neq 0. \tag{7}$$

The following proposition states that we can carry out a t-test with the constrained estimator $\hat{\beta}_j$ and the associated standard error.

Proposition 2. *Suppose that* $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I)$. *Under the null hypothesis* $H_0 : \beta_j = 0$, *the following test statistic has the Student's t distribution with* $n - p + 1$ *degrees of freedom, i.e.,*

$$t_j = \frac{\hat{\beta}_j}{\hat{se}(\hat{\beta}_j)} \sim t_{n-p+1}, \tag{8}$$

where $\hat{se}(\hat{\beta}_j) = \sqrt{\widehat{Var}(\hat{\boldsymbol{\beta}})_{jj}}$, $\widehat{Var}(\hat{\boldsymbol{\beta}})_{jj}$ *is the* j*th diagonal entry of* $\hat{\sigma}^2\mathbf{M}(\mathbf{Z}^T\mathbf{Z})^{-1}$, $\hat{\sigma}^2 = \mathrm{SSE}(\hat{\boldsymbol{\beta}})/(n - p + 1)$, *and* $\mathrm{SSE}(\hat{\boldsymbol{\beta}}) = (\mathbf{y} - \mathbf{Z}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{Z}\hat{\boldsymbol{\beta}})$.

The level α test can be carried out by rejecting the null hypothesis if $|t_j| > t_{\alpha/2,n-p+1}$, where $t_{\alpha,n-p+1}$ is $(1-\alpha)$th quantile of t_{n-p+1} distribution. Moreover, a $100(1-\alpha)\%$ confidence interval for β_j is given by

$$\left[\hat{\beta}_j - t_{\alpha/2,n-p+1}\hat{se}(\hat{\beta}_j),\ \hat{\beta}_j + t_{\alpha/2,n-p+1}\hat{se}(\hat{\beta}_j)\right].$$

Note that the above parametric inference cannot work for our motivating cattle data with $p > n$. Estimation of the model (2) will suffer from overfitting and high variances. Even if the model can be estimated, the t-test cannot be used unless the errors are normal, since the sample size is too small for normal approximation. In the next section, we present a sequential approach with which we overcome these problems and are able to make inference on the significance of predictor variables.

3 Inference for High-dimensional Log-contrast Regression

3.1 Screening via Sparse Log-contrast Regression

A regularized regression is an effective way to overcome high-dimensionality of the data. In this section we consider a popular sparse regularization method, lasso regression (Tibshirani, 1996). Applying the ℓ_1 regularization approach to the log-contrast model (2), Lin et al. (2014) proposed a constrained convex optimization problem for the following model

$$\hat{\boldsymbol{\beta}}_\lambda = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\frac{1}{2n}\|\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1\right), \quad \text{subject to } \mathbf{1}^T\boldsymbol{\beta} = 0, \tag{9}$$

where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$, $\lambda > 0$ is a regularization parameter, and $\|\cdot\|_1$ denotes ℓ_1 norm.

Since the constrained lasso optimization problem (9) does not have a closed form solution, one needs to rely on a numerical algorithm such as quadratic programming (Brodie et al., 2009; Bondell and Reich, 2009) or the alternating direction method of multipliers (ADMM) (Lin et al., 2014; Fang et al., 2015). One can also obtain a solution path by using the method proposed by Jeon et al. (2020). In this work, we use R package `glmnet` with a weighting trick suggested by Bates and Tibshirani (2019). That is, we augment the data with an observation with all covariates equal to one, and the response value zero. By assigning this value a dominatingly large weight, one can force the resulting solution $\hat{\boldsymbol{\beta}}_\lambda$ to have an arbitrary small $\sum_{j=1}^p \hat{\beta}_j$.

The number of screened variables in this step is determined by the sparsity tuning parameter λ. Lin et al. (2014) suggest to use λ that minimizes the generalized information criterion (GIC) (Fan and Tang, 2013):

$$\text{GIC}(\lambda) = \log(\hat{\sigma}_\lambda^2) + (s_\lambda - 1)\frac{\log\log(n)}{n}\log(\max(p, n))$$

where $\hat{\sigma}_\lambda^2 = \|\mathbf{y} - \mathbf{Z}\hat{\boldsymbol{\beta}}_\lambda\|_2^2/n$ and s_λ is the number of nonzero coefficients in $\hat{\boldsymbol{\beta}}_\lambda$. Note that because of the zero-sum constraint, the effective number of free parameters is $s_\lambda - 1$ for $s_\lambda \geq 2$. However, we have found that tuning based on GIC tends to choose too many

predictors, as will be seen in Sections 4 and 5. Thus we propose to select λ by a cross-validation based on the mean squared error (MSE) defined as below:

$$\text{MSE}(\lambda) = \sum_{i=1}^{r} (y_i - \hat{y}_{i,\lambda})^2,$$

where y_i represents the ith sample in test data, $\hat{y}_{i,\lambda}$ is its fitted value based on the coefficients obtained with the tuning parameter λ, and r is the test sample size of each fold. In all our empirical studies, we used a 10-fold cross-validation. Since MSE, as a function of λ, does not usually yield a smooth trajectory, we employ `loess` smoothing for better stability. Let the smoothed version of the MSE be denoted by $m(\lambda)$. Then we choose the smallest λ that yields the minimum $m(\lambda)$ plus its standard error:

$$\hat{\lambda} = \min\{\lambda : m(\lambda) \leq m(\lambda_{min}) + s.e.(m(\lambda_{min}))\}. \tag{10}$$

Note that this rule is similar in spirit to the one-standard error rule in Hastie et al. (2009), except that we favor a *larger* model. We then apply the chosen $\hat{\lambda}$ in (9) using the whole training data. Variables with nonzero coefficient estimates are screened in, and we record the indices set $J = \{j : \hat{\beta}_j(\hat{\lambda}) \neq 0, j = 1, \ldots, p\}$. Note that in rare, but theoretically possible, cases where no variable is screened, we adjust $\hat{\lambda}$ so that at least two variables must be screened. Specifically, we set $\hat{\lambda} = \max\{\lambda : \#\{i : \hat{\beta}_i(\lambda) \neq 0\} \geq 2\}$. There was a small fraction of runs in the simulation study with an extremely small sample size ($n = 20$) when such adjustment was necessary, while there was no such case in the real data analysis, as seen in Table 3. A reviewer raised a question about comparison of the proposed screening idea with ones based on GIC, MSE, and also the typical one-standard error rule that encourages higher parsimony. We investigate this question by implementing the screening rules to our motivating cattle microbiomes data in Section 5.

3.2 Testing Individual Predictors

Recall that J is the set of indices of screened variables. Consider the reduced log-contrast model with screened variables:

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \text{subject to } \beta_j = 0, j \notin J, \quad \sum_{j \in J} \beta_j = 0. \tag{11}$$

In below we discuss a few resampling-based approaches for formally testing whether each of the variables in J is significantly related to the response, controlling for other variables in J.

3.2.1 Permutation Tests

There exist many permutation methods for individual significance testing in multiple regression. Manly (2018) discuss permutation of the response values to nullify the relationship between the response and the predictors. Freedman and Lane (1983) and Kennedy (1995) propose to permute the residuals under the reduced model, i.e., the null model under $H_0 : \beta_j = 0$, while Ter Braak (1992) use residuals from the full model. All these methods give empirical p-values by comparing the observed test statistic with the reference

distribution generated by random permutations. Anderson and Robinson (2001) show that all three methods become asymptotically equivalent as the sample size n increases.

In this work, we take an approach similar to Manly (2018). After the screening step in Section 3.1, we estimate the model (11) and obtain the estimates $\hat{\beta}_j$, $j \in J$, using (5) and calculate the associated t-statistics t_j using (8). In order to obtain reference distributions for these statistics, we randomly permute the entries of the response vector $\mathbf{y}$ and denote the permuted response by $\mathbf{y}^*$. Then the model (11) is fit with $\mathbf{y}^*$ and the unpermuted $\mathbf{Z}$, yielding the estimates $\hat{\beta}_j^*$ and the associated t-statistics t_j^*, $j \in J$. This process is repeated M times, producing the permutation distribution of $\hat{\beta}_j^*$ and t_j^*. We use $M = 1000$ for all our empirical studies. The empirical p-value for testing $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$ in the model (11) is then defined as the tail probability in the permutation distribution (Hope, 1968), i.e.,

$$P_p(t_j) = \frac{1}{M} \sum_{m=1}^{M} I\big(|t_j^{*m}| \geq |t_j|\big), \tag{12}$$

$$P_p(\hat{\beta}_j) = \frac{1}{M} \sum_{m=1}^{M} I\left(|\hat{\beta}_j^{*m}| \geq |\hat{\beta}_j|\right), \tag{13}$$

where $I(\cdot)$ is the indicator function. The null hypothesis (7) is rejected if the empirical p-value is less than the level of significance α. We denote these tests by $\phi_{t_j,\alpha}$ and $\phi_{\hat{\beta}_j,\alpha}$, respectively. Note that while Manly (2018) only used t_j, here we also consider $\hat{\beta}_j$, whose empirical performance in terms of variable selection is sometimes superior; see Section 4.

In the following theorem, we show that the permutation tests above asymptotically maintain the nominal size α for a large n, under the exchangeable errors condition. Note that the exchangeable errors condition is significantly weaker than requiring i.i.d. Gaussian errors. We also assume that the random permutation is equi-probable, i.e. each permutation $\mathbf{y}^*$ occurs with probability $1/n!$, which is a standard assumption that holds naturally.

Theorem 3. *Assume that in (11), the errors $\{\epsilon_i : i = 1, \ldots, n\}$ are exchangeable. Then, the permutation tests $\phi_{\hat{\beta}_j,\alpha}$ and $\phi_{t_j,\alpha}$ are (asymptotically) unbiased in the following sense. (The following statements remain true if $\phi_{\hat{\beta}_j,\alpha}$ is replaced by $\phi_{t_j,\alpha}$.)*

(i) Suppose that $|J| = 2$. Let $\alpha \in (0, 1)$ be such that $\alpha = k/n!$ for some $k = 1, 2, \ldots, n!$. Then for each $j \in J$, the test $\phi_{\hat{\beta}_j,\alpha}$ is exact in the sense that the type I error rate of $\phi_{\hat{\beta}_j,\alpha}$ is exactly α.

(ii) Suppose that $|J| > 2$. Let $j \in J$ be fixed. If, under the null hypothesis $H_0 : \beta_j = 0$, $E(y_i^2)$ is bounded below, and $E(|y_i|^3)$ is bounded above, then the type I error rate of $\phi_{\hat{\beta}_j,\alpha}$ approaches to α as the sample size increases. More precisely, we have $\lim_{n\to\infty} Pr(P_p(\hat{\beta}_j) < \alpha) = \alpha$ for any $\alpha \in (0, 1)$.

3.2.2 Bootstrap Confidence Interval

Bootstrap confidence interval (CI) (Efron, 1982; Davison and Hinkley, 1997) is an effective tool for quantifying the variability of a statistic in a non-parametric way. The hypotheses in (7) can be tested based on whether or not a two-sided bootstrap CI contains zero.

We consider both percentile bootstrap and studentized bootstrap in this work. The percentile bootstrap interval with $100(1-\alpha)\%$ confidence is simply the interval between the $100(\alpha/2)$th and $100(1-\alpha/2)$th percentiles of a bootstrap distribution of $\hat{\beta}_j$. The percentile bootstrap CI assumes that there exists a monotone transformation of $\hat{\beta}_j$, say $\xi(\hat{\beta}_j)$, such that $\xi(\hat{\beta}_j) - \xi(\beta_j)$ is a pivot, whereas the studentized bootstrap CI assumes that $(\hat{\beta}_j - \beta_j)/\hat{se}(\hat{\beta}_j)$ is a pivot, where $\hat{se}(\hat{\beta}_j)$ is the standard error estimate of $\hat{\beta}_j$.

Similarly to the permutation procedure, we first obtain the estimate $\hat{\beta}_j$ for the screened model (11), as given by (5). We generate B bootstrap data sets with which we compute the bootstrap replications $\hat{\beta}_j^{*b}$ and $t_j^{*b} = (\hat{\beta}_j^{*b}(b) - \hat{\beta}_j)/\hat{se}(\hat{\beta}_j^{*b})$, where $\hat{se}(\hat{\beta}_j^{*b})$ is the standard error estimate of $\hat{\beta}_j$, given by Proposition 2, with $\hat{\beta}_j$ replaced by $\hat{\beta}_j^{*b}$, for $b = 1, \cdots, B$. We used $B = 1000$ in this work. Then from the bootstrap distribution of $\hat{\beta}_j^{*1}, \ldots, \hat{\beta}_j^{*B}$, the percentile bootstrap CI is constructed as follows:

$$\hat{\mathcal{I}}_{p,\alpha} = \left[\hat{\beta}_{j,\text{lo}},\ \hat{\beta}_{j,\text{up}}\right] = \left[\hat{\beta}_j^{*(\alpha/2)},\ \hat{\beta}_j^{*(1-\alpha/2)}\right], \tag{14}$$

where $\hat{\beta}_j^{*(\alpha/2)}$ and $\hat{\beta}_j^{*(1-\alpha/2)}$ respectively denote the $100(\alpha/2)$th and $100(1-\alpha/2)$th percentiles of the bootstrap distribution. Alternatively, the studentized bootstrap CI is constructed as follows:

$$\hat{\mathcal{I}}_{s,\alpha} = \left[\hat{\beta}_j - \hat{t}_j^{*(1-\alpha/2)}\hat{se}(\hat{\beta}_j),\ \hat{\beta}_j - \hat{t}_j^{*(\alpha/2)}\hat{se}(\hat{\beta}_j)\right], \tag{15}$$

where $t_j^{*(\alpha/2)}$ and $t_j^{*(1-\alpha/2)}$ respectively denote the $100(\alpha/2)$th and $100(1-\alpha/2)$th percentiles of the bootstrap distribution of $t_j^{*1}, \ldots, t_j^{*B}$. The bootstrap CIs can be used for testing (7) at level α by rejecting $H_0 : \beta_j = 0$ if the CI contains zero.

The following theorem shows that the coverage probability for either bootstrap CI converges to the nominal level as the sample size increases (Carpenter and Bithell, 2000), which implies that the hypothesis tests based on the bootstrap CIs are asymptotically exact.

Theorem 4. *Assume that the model* (11) *is true. Then the coverage probabilities of the bootstrap CIs in* (14) *and* (15) *converge to* $1-\alpha$ *as* n *increases. Specifically,*

(i) $Pr(\beta_j \in \hat{\mathcal{I}}_{p,\alpha}) = 1 - \alpha + O_p\left(\frac{1}{\sqrt{n}}\right)$ *for* $j \in J$.

(ii) $Pr(\beta_j \in \hat{\mathcal{I}}_{s,\alpha}) = 1 - \alpha + O_p\left(\frac{1}{n}\right)$ *for* $j \in J$.

3.2.3 Estimation of the Final Model

As mentioned in Section 1, the purpose of our proposed procedure is to explore the associations of individual compositional taxa with a response variable. Nevertheless a final predictive model can be constructed based on the found associations, as follows. Suppose that s number of coefficients are found to be significant based on the test procedures as discussed in the previous subsection. If $s \geq 2$, then the final model estimate is obtained by fitting the log-contrast model (2) with those s variables. However the case when $s = 1$ needs a special treatment. Due to the compositional nature of the variables, if x is found

to be significant, then it unavoidably implies that $1 - x$ is also significant. Thus we use the following model:

$$y_i = \mu + \beta_1 \log(x_i) + \beta_2 \log(1 - x_i) + \epsilon_i \quad \text{subject to } \beta_1 + \beta_2 = 0,$$

which leads to following unrestricted, simple regression model

$$y_i = \mu + \beta_1 \log\left(\frac{x_i}{1 - x_i}\right) + \epsilon_i.$$

4 Simulation Study

In this section we compare performances of the various inference methods that were discussed in the previous section: parametric inference described in Proposition 2, the permutation methods (12) with t-statistic and (13) with $\hat{\beta}$, the percentile bootstrap CI (14), the studentized bootstrap CI (15), and also the union of significant variables from the four resampling-based methods. They are respectively denoted as "t-test","perm(t)", "perm(β)", "boot(p)", "boot(s)", and "union" (see Section 5.) As described in Section 3, these inferential methods are applied to the screened model (11). We also include the compositional lasso with GIC (Lin et al., 2014) and clr-lasso (Susin et al., 2020) for comparison. The clr-lasso is the usual lasso regression with the predictors given by the clr-transformed compositions. The clr-transformation of a composition $\mathbf{x}$ is

$$\text{clr}(\mathbf{x}) = [\log(x_1/g(\mathbf{x})), \cdots, \log(x_p/g(\mathbf{x}))],$$

where $g(\mathbf{x}) = (\prod_{j=1}^{p} x_j)^{1/p}$ is a geometric mean of the composition. We used the cross-validated tuning parameter for the clr-lasso as done in Susin et al. (2020).

We generate the data from the following two models, which are replicated 1000 times respectively. The first model is a logistic-normal model similar to the examples in Lin et al. (2014). The second model is supposed to mimic our motivating cattle data, in the sense that we directly use the observed microbiome covariates and the estimated values of the coefficients as the true values. All computations are conducted in `R` (version 4.1.2). The total computation time for performing all four proposed resampling-based methods on a single data set of size $(n, p) = (50, 36)$ is 5.533 seconds (on average); these times are 35.80 and 6.386 seconds for the cases $(n, p) = (100, 200)$ and $(20, 36)$, respectively. Computation times are measured on a Macbook Pro (Intel Core i7, 2.6 GHz, 16 GB RAM).

- **Model 1**

  We generate an $n \times p$ data matrix $\mathbf{W} = (w_{ij})$ from a multivariate normal distribution $N_p(\theta, \mathbf{\Sigma})$, and then obtain the covariate matrix $\mathbf{X} = (x_{ij})$ by transforming $x_{ij} = \exp(w_{ij}) / \sum_{k=1}^{p} \exp(w_{ik})$. To reflect the fact that the components of a composition in metagenomic data often differ by orders of magnitude, we let $\theta = (\theta_j)$ with $\theta_j = \log(0.5p)$ for $j = 1, \cdots, 5$ and $\theta_j = 0$ otherwise. As for the covariance matrix $\mathbf{\Sigma}$, we let $\mathbf{\Sigma} = 3(\rho^{|i-j|})$ with $\rho = 0.2$ or 0.5. Finally we generate the response values using the model in (3) with $\boldsymbol{\beta}_{true} = (1, -0.8, 0.6, 0, 0, -1.5, -0.5, 1.2, 0, \cdots, 0)^T$ and $\sigma = 3$. We set $(n, p) = (50, 36), (100, 200)$, and $(20, 36)$. Note that the last sample size-dimension setting is from the cattle data example. Extensive simulation studies with $n = (20, 50, 100)$, $p = (10, 36, 100, 200)$ are provided in the Supplementary Material Section S1.

- **Model 2**

  We consider the following model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_{true} + \boldsymbol{\epsilon},$$

  where the $p = 36$ covariates in $\mathbf{X}$ are from the logarithms of microbiome taxa compositions measured in the rumen sample of the cattle microbiome data set, $n = 20$, and $\boldsymbol{\beta}_{true} = (-1.06, 1.06, 0, \cdots, 0)^T$, which is the final coefficient estimate for the data with RFI as the response, as will be seen in Section 5. For this model four different error distributions are considered: $\epsilon \sim N(0, \sigma^2)$ with $\sigma = 1,\ 3$, and $\epsilon \sim \text{Laplace}(0, \sigma)$, with $\sigma = \sqrt{1/2}, \sqrt{9/2}$.

As for the performance criteria we use the following measures for variable selectivity: false negatives (FN), false positives (FP), positive predictive value (PPV) defined as

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}},$$

and Matthew's correlation coefficient (MCC) by Matthews (1975) defined as

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP}) \times (\text{TP} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TN} + \text{FN})}},$$

where TP and TN denote true positives and true negatives respectively. Note that MCC ranges in the interval $[-1, +1]$, with extreme vales -1 (for only when TP = 0 and TN = 0) and $+1$ (for only when FP = 0 and FN = 0) reached in the case of complete misclassification and perfect classification, respectively, while MCC = 0 represents no better than random prediction (Chicco and Jurman, 2020).

Table 1 displays the results under Model 1 with different correlation levels among the covariates and Table 2 shows the results from Model 2 with normal errors and Laplace errors. While the resampling-based methods such as permutation and bootstrap seem generally better than GIC, clr-lasso, and t-test, there are some findings worth mentioning. We have found that GIC performs decently when n is not small, however, its performance deteriorates when $n = 20$. It is also noticeable that when the sample size is small, the four resampling-based methods tend to select variables conservatively and often yield model estimates with less variables than desired. This is in contrast to that the other three methods tends to choose more variables and inevitably yield higher FP.

The high PPV values of perm(β) are due to that the method yields extremely sparse model estimates with only few signal variables. It is clear that when the sample size is too small such as $(n, p) = (20, 36)$, most methods suffer from inflated FN as it becomes challenging to detect signals. In particular, perm(β) tends to be the most conservative in detecting signals, which implies a possible lack of power. On the other hand, in large n settings, perm(t) tends to yield high FPs by being too liberal in signal detection. In this regard, the boot(p) seems more effective for small sized data considering the overall balance of the performance measures in both settings. It is also interesting to see that t-test and perm(t) show very similar results throughout, even though the latter is consistently a bit better.

Table 1: Average performance measures (with standard errors) of the eight methods based on 1000 repetitions under Model 1. Note that there are six true signal variables in this setting, i.e., $\|\boldsymbol{\beta}\|_0 = 6$.

ρ	(n, p)	Method	FN	FP	$\|\hat{\boldsymbol{\beta}}\|_0$	MCC	PPV
0.2	(50, 36)	GIC	2.94 (0.06)	1.42 (0.06)	4.49 (0.1)	0.51 (0.01)	0.76 (0.01)
		clr-lasso	1.98 (0.05)	3.24 (0.12)	7.26 (0.15)	0.56 (0.01)	0.67 (0.01)
		t-test	1.65 (0.03)	3.18 (0.08)	7.53 (0.09)	0.59 (0.01)	0.63 (0.01)
		perm(t)	1.69 (0.03)	3.06 (0.08)	7.37 (0.09)	0.59 (0.01)	0.64 (0.01)
		perm(β)	4.03 (0.04)	0.06 (0.01)	2.03 (0.04)	0.47 (0.01)	0.97 (0.00)
		boot(p)	2.72 (0.05)	0.94 (0.04)	4.23 (0.06)	0.58 (0.01)	0.81 (0.01)
		boot(s)	3.00 (0.05)	0.74 (0.03)	3.75 (0.07)	0.55 (0.01)	0.84 (0.01)
		union	1.68 (0.03)	3.11 (0.08)	7.43 (0.08)	0.59 (0.01)	0.63 (0.01)
	(100, 200)	GIC	2.72 (0.05)	0.85 (0.04)	4.13 (0.07)	0.64 (0.01)	0.85 (0.01)
		clr-lasso	1.47 (0.04)	5.74 (0.24)	10.27 (0.26)	0.63 (0.00)	0.60 (0.01)
		t-test	0.94 (0.03)	11.61 (0.22)	16.67 (0.23)	0.51 (0.00)	0.35 (0.00)
		perm(t)	0.97 (0.03)	11.43 (0.22)	16.46 (0.22)	0.51 (0.00)	0.35 (0.00)
		perm(β)	3.27 (0.04)	0.07 (0.01)	2.81 (0.04)	0.63 (0.01)	0.98 (0.00)
		boot(p)	1.45 (0.04)	5.26 (0.11)	9.81 (0.12)	0.59 (0.00)	0.52 (0.01)
		boot(s)	1.66 (0.05)	4.93 (0.14)	9.27 (0.16)	0.57 (0.01)	0.52 (0.01)
		union	0.94 (0.03)	11.81 (0.24)	16.87 (0.24)	0.51 (0.00)	0.34 (0.00)
	(20, 36)	GIC	1.33 (0.03)	17.28 (0.07)	21.95 (0.07)	0.15 (0.00)	0.21 (0.00)
		clr-lasso	3.56 (0.06)	4.93 (0.15)	7.37 (0.20)	0.23 (0.01)	0.44 (0.01)
		t-test	4.89 (0.05)	1.73 (0.10)	2.84 (0.14)	0.15 (0.01)	0.52 (0.01)
		perm(t)	4.73 (0.05)	2.08 (0.11)	3.35 (0.15)	0.17 (0.01)	0.50 (0.01)
		perm(β)	5.83 (0.02)	0.09 (0.01)	0.26 (0.02)	0.04 (0.00)	0.65 (0.03)
		boot(p)	4.35 (0.04)	1.48 (0.05)	3.13 (0.07)	0.28 (0.01)	0.57 (0.01)
		boot(s)	4.66 (0.05)	2.9 (0.13)	4.24 (0.17)	0.13 (0.01)	0.41 (0.01)
		union	3.69 (0.05)	4.18 (0.13)	6.5 (0.16)	0.26 (0.01)	0.46 (0.01)
0.5	(50, 36)	GIC	3.80 (0.05)	1.23 (0.06)	3.43 (0.10)	0.39 (0.01)	0.73 (0.01)
		clr-lasso	3.17 (0.05)	2.58 (0.11)	5.41 (0.15)	0.43 (0.01)	0.67 (0.01)
		t-test	2.29 (0.04)	2.9 (0.08)	6.62 (0.09)	0.52 (0.01)	0.62 (0.01)
		perm(t)	2.32 (0.04)	2.79 (0.07)	6.47 (0.09)	0.52 (0.01)	0.62 (0.01)
		perm(β)	4.41 (0.04)	0.15 (0.01)	1.74 (0.04)	0.38 (0.01)	0.92 (0.01)
		boot(p)	3.17 (0.05)	1.07 (0.04)	3.90 (0.07)	0.51 (0.01)	0.77 (0.01)
		boot(s)	3.38 (0.05)	0.92 (0.04)	3.54 (0.07)	0.48 (0.01)	0.78 (0.01)
		union	2.28 (0.04)	2.86 (0.07)	6.59 (0.09)	0.52 (0.01)	0.62 (0.01)
	(100, 200)	GIC	3.77 (0.04)	0.91 (0.04)	3.14 (0.07)	0.47 (0.01)	0.79 (0.01)
		clr-lasso	2.95 (0.04)	4.3 (0.22)	7.35 (0.25)	0.51 (0.00)	0.63 (0.01)
		t-test	1.80 (0.03)	9.39 (0.20)	13.59 (0.21)	0.46 (0.00)	0.35 (0.00)
		perm(t)	1.81 (0.04)	9.3 (0.20)	13.48 (0.21)	0.47 (0.00)	0.36 (0.00)
		perm(β)	3.74 (0.04)	0.19 (0.01)	2.45 (0.04)	0.54 (0.01)	0.93 (0.01)
		boot(p)	2.25 (0.04)	4.92 (0.10)	8.66 (0.11)	0.51 (0.01)	0.48 (0.01)
		boot(s)	2.39 (0.04)	4.55 (0.10)	8.17 (0.12)	0.50 (0.01)	0.48 (0.01)
		union	1.79 (0.03)	9.58 (0.20)	13.79 (0.21)	0.46 (0.00)	0.35 (0.00)
	(20, 36)	GIC	1.54 (0.03)	16.95 (0.08)	21.4 (0.08)	0.14 (0.00)	0.21 (0.00)
		clr-lasso	4.10 (0.05)	4.36 (0.15)	6.25 (0.20)	0.18 (0.01)	0.42 (0.01)
		t-test	5.04 (0.04)	1.54 (0.09)	2.49 (0.12)	0.14 (0.01)	0.49 (0.01)
		perm(t)	4.90 (0.04)	1.87 (0.10)	2.96 (0.13)	0.15 (0.01)	0.48 (0.01)
		perm(β)	5.81 (0.02)	0.14 (0.01)	0.33 (0.02)	0.04 (0.00)	0.57 (0.03)
		boot(p)	4.75 (0.03)	1.35 (0.04)	2.60 (0.06)	0.22 (0.01)	0.52 (0.01)
		boot(s)	4.91 (0.04)	2.56 (0.12)	3.65 (0.15)	0.11 (0.01)	0.40 (0.01)
		union	4.14 (0.04)	3.66 (0.13)	5.53 (0.16)	0.22 (0.01)	0.43 (0.01)

The findings from the extensive simulations in the Supplementary Material generally agree with what we have in the paper. All methods perform better in large n setting as expected. GIC performs well when n is relatively large, however, it tends to choose too many variables when p is larger than n. The resampling-based methods are better in small sized data. The perm(β) method tends to yield much sparse model estimates

Table 2: Average performance measures (with standard errors) of eight methods based on 1000 repetitions under Model 2. Note that there are three true signal variables in this setting, i.e., $\|\boldsymbol{\beta}\|_0 = 2$.

Error	σ	Method	FN	FP	$\|\hat{\boldsymbol{\beta}}\|_0$	MCC	PPV
normal	1	GIC	0.06 (0.01)	15.97 (0.24)	17.92 (0.24)	0.28 (0.01)	0.17 (0.01)
		clr-lasso	0.42 (0.02)	5.15 (0.16)	6.73 (0.17)	0.43 (0.01)	0.37 (0.01)
		t test	0.60 (0.03)	2.48 (0.12)	3.88 (0.13)	0.51 (0.01)	0.57 (0.01)
		perm(t)	0.51 (0.02)	2.52 (0.12)	4.01 (0.13)	0.56 (0.01)	0.58 (0.01)
		perm(β)	1.33 (0.02)	0.04 (0.01)	0.71 (0.02)	0.40 (0.01)	0.96 (0.01)
		boot(p)	0.58 (0.02)	0.75 (0.03)	2.17 (0.04)	0.66 (0.01)	0.73 (0.01)
		boot(s)	0.87 (0.03)	2.68 (0.13)	3.81 (0.15)	0.42 (0.01)	0.56 (0.01)
		union	0.30 (0.02)	3.83 (0.14)	5.54 (0.15)	0.59 (0.01)	0.52 (0.01)
	3	GIC	0.39 (0.02)	18.99 (0.25)	20.6 (0.26)	0.13 (0.00)	0.10 (0.00)
		clr-lasso	1.57 (0.02)	3.37 (0.18)	3.81 (0.20)	0.07 (0.01)	0.17 (0.01)
		t test	1.57 (0.02)	1.59 (0.08)	2.02 (0.09)	0.16 (0.01)	0.27 (0.01)
		perm(t)	1.54 (0.02)	1.76 (0.09)	2.21 (0.10)	0.16 (0.01)	0.27 (0.01)
		perm(β)	1.74 (0.02)	0.63 (0.02)	0.89 (0.03)	0.13 (0.01)	0.32 (0.02)
		boot(p)	1.58 (0.02)	1.21 (0.03)	1.63 (0.04)	0.17 (0.01)	0.27 (0.01)
		boot(s)	1.58 (0.02)	2.10 (0.11)	2.52 (0.12)	0.13 (0.01)	0.24 (0.01)
		union	1.41 (0.02)	2.77 (0.12)	3.36 (0.13)	0.18 (0.01)	0.24 (0.01)
laplace	$\frac{1}{\sqrt{2}}$	GIC	0.07 (0.01)	15.58 (0.24)	17.5 (0.24)	0.28 (0.01)	0.17 (0.01)
		clr-lasso	0.43 (0.02)	4.74 (0.15)	6.31 (0.17)	0.43 (0.01)	0.38 (0.01)
		t test	0.58 (0.03)	2.28 (0.12)	3.70 (0.13)	0.54 (0.01)	0.59 (0.01)
		perm(t)	0.53 (0.02)	2.29 (0.12)	3.76 (0.12)	0.57 (0.01)	0.61 (0.01)
		perm(β)	1.31 (0.02)	0.06 (0.01)	0.74 (0.02)	0.41 (0.01)	0.94 (0.01)
		boot(p)	0.54 (0.02)	0.71 (0.03)	2.17 (0.04)	0.69 (0.01)	0.75 (0.01)
		boot(s)	0.90 (0.03)	2.45 (0.13)	3.54 (0.14)	0.42 (0.01)	0.58 (0.01)
		union	0.29 (0.02)	3.57 (0.14)	5.28 (0.14)	0.60 (0.01)	0.54 (0.01)
	$\frac{3}{\sqrt{2}}$	GIC	0.41 (0.02)	18.91 (0.25)	20.50 (0.27)	0.13 (0.00)	0.09 (0.00)
		clr-lasso	1.57 (0.02)	3.10 (0.16)	3.54 (0.18)	0.07 (0.01)	0.16 (0.01)
		t test	1.55 (0.02)	1.67 (0.08)	2.12 (0.09)	0.16 (0.01)	0.26 (0.01)
		perm(t)	1.53 (0.02)	1.80 (0.09)	2.27 (0.10)	0.16 (0.01)	0.26 (0.01)
		perm(β)	1.72 (0.02)	0.70 (0.03)	0.98 (0.03)	0.13 (0.01)	0.31 (0.02)
		boot(p)	1.58 (0.02)	1.18 (0.03)	1.60 (0.04)	0.17 (0.01)	0.27 (0.01)
		boot(s)	1.59 (0.02)	1.96 (0.10)	2.37 (0.11)	0.13 (0.01)	0.23 (0.01)
		union	1.43 (0.02)	2.66 (0.11)	3.23 (0.12)	0.18 (0.01)	0.23 (0.01)

which indicates that it has a relatively smaller power. On the other hand, perm(t) tends to have large FP. The boot(p) seems to be most effective for small sized data considering the overall balance of the performance measures. The boot(s) is overall similar to boot(p), though slightly worse.

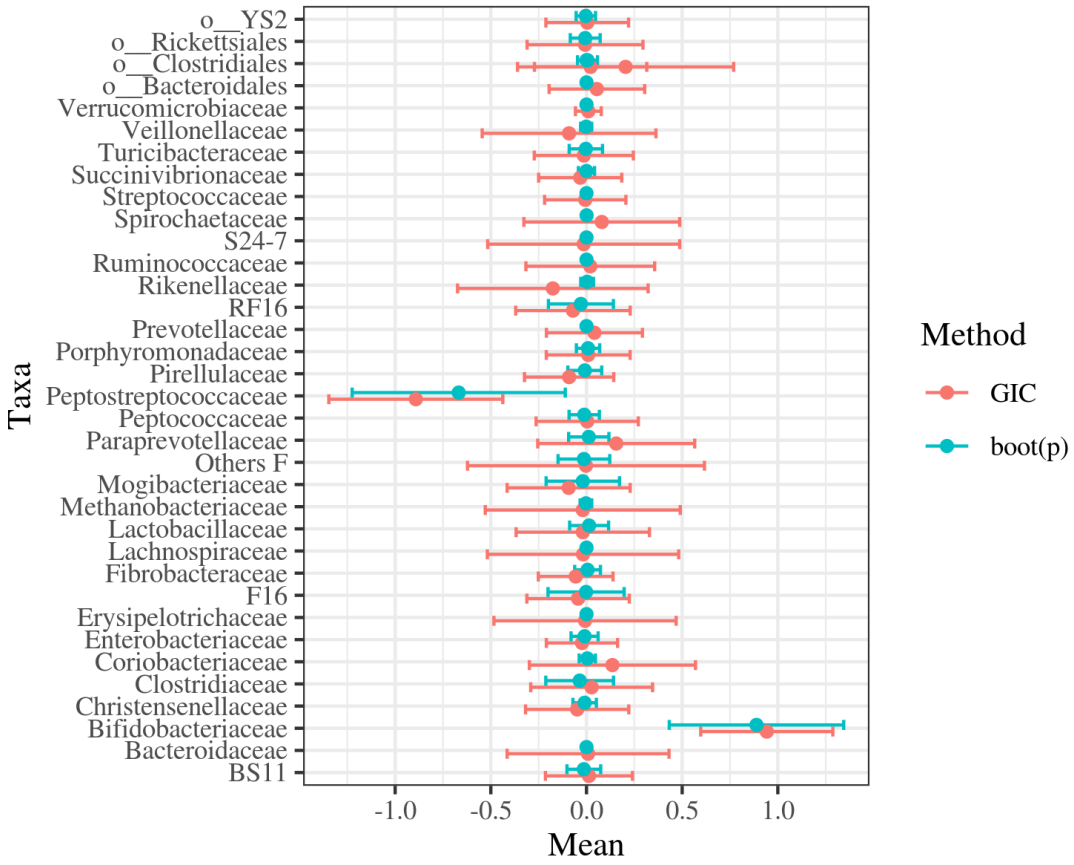
The effect of higher correlation can be seen by comparing the $\rho = 0.2$ part and the $\rho = 0.5$ part in Table 1. Overall, the compared methods have worse MCC and FN but better FP under higher correlation. Even though all methods suffer from the collinearity regardless of the sample size and dimension, it seems that the bootstrap methods are most robust. From Table 2 we can get an idea on which methods can deal with larger noise levels in the data. Regardless of the error distribution, most methods suffer from larger noise levels with respect to MCC and PPV, while the boot(p) is the best in comparison.

In addition to the variable selectivity results appearing in Tables 1 and 2, we have also investigated the accuracy of the coefficient estimation. Figure 1 shows the error plot, i.e., average of the coefficient estimates $\pm$ standard errors, under Model 2 with normal errors with $\sigma = 1$, for GIC and boot(p) methods. Recall from Table 1 that GIC chooses too many variables (17.92 on average), whereas boot(p) chooses only 2.17 variables. (There are two

true signal variables.) We also observe in Figure 1 that the coefficients corresponding to most of the noise variables are correctly estimated by the boot(p) method to be zero with very narrow CIs, while the coefficient estimates from GIC are highly variable, and are oftentimes biased. As for the signal coefficients (Peptostreptococcaceae, Bifidobacteriaceae), both GIC and boot(p) are biased towards zero. The difference between the bias from boot(p) and GIC does not stand out, when compared to the standard error. A reviewer has pointed out that the bias of boot(p) for the signal coefficients is intrinsic to the conservative nature of the method, which helps in correctly estimating the zero coefficients.

In summary, the simulation study suggests that resampling-based methods are able to select signal variables with better accuracy than existing approaches under various scenarios. However, even though bootstrap methods performed a bit better than permutation methods in this study, we suggest that rather than relying on a particular method, one should implement multiple methods and consider inspecting the union of the selected variables. We also apply this idea to the real data in the next section.

Figure 1: Error plot of the estimated coefficients under Model 2 with normal error with $\sigma = 1$. Note that there are two true signal variables which are $\beta_{\text{Peptostreptococcaceae}} = -1.06$, $\beta_{\text{Bifidobacteriaceae}} = 1.06$.



5 Steer Quality Prediction with Microbiomes

In this section, we attempt to identify important bacterial taxa in steers' rumen, cecum and feces that are related to key quality traits of beef cattle: residual food intake (RFI), back fat, lipid, and yield grade (YG). Since there are three sets of microbiomes with four response variables, there will be 12 regression models to fit, each with 36 covariates and 20 sample size. In Section 5.1 we give a detailed description of the analysis with rumen microbiomes to predict RFI. The results of all modelings will be presented later in Section 5.2 with detailed discussions on scientific implications on the findings. For the zero values in the data, we replace them by the smallest nonzero entry of each data matrix $\mathbf{X}$, after which we re-closure each row to make it a composition. (Aitchison, 1982).

Table 3 summarizes the results from the variable screening step with the compositional lasso. We report the number of screened taxa based on GIC, MSE, MSE + 1-se, and MSE − 1-se. Note the MSE + 1-se is the proposed screening rule, discussed in Section 3.1, while MSE − 1-se is the *usual* one standard error rule, e.g., implemented in the R package `glmnet`. Considering $p = 36$, the GIC rule screens in too many taxa except for RFI models, while the MSE and MSE − 1-se rules screened in too few.

Table 3: Number of screened taxa in the cattle microbiome data. The subscripts r, c, f denote microbiome predictors from rumen, cecum, and feces, respectively.

	RFI_r	RFI_c	RFI_f	Back_r	Back_c	Back_f	Lipid_r	Lipid_c	Lipid_f	YG_r	YG_c	YG_f
GIC	1	1	20	15	26	22	24	24	24	24	26	27
MSE	2	6	4	7	3	1	1	0	4	5	2	1
MSE + 1-se	4	6	4	11	3	3	6	2	4	9	3	2
MSE − 1-se	1	4	4	5	2	1	0	0	4	5	1	1

5.1 Regressing RFI on Rumen Microbiomes

In modeling with RFI as the response and rumen microbiome compositions as predictors, four bacterial taxa, namely Peptostreptococcaceae, Bifidobacteriaceae, Clostridiales, and Fibrobacteraceae, are screened at the screening step. The individual significance test results for the four taxa are shown in Table 4. We also assess how stable the selection is by using the "stability selection" proposed by Meinshausen and Bühlmann (2010). We take 100 sub-samples of size $n/2$ compute the frequency of the selection, which is shown in the last column of the table. All four screened taxa have 75 or higher selection frequencies out of 100, while the rest of $36 - 4 = 32$ taxa have been selected less than 75 times, which indicates that the screening results are quite stable. See the Supplementary Material for the detailed procedure and the results.

As seen with simulated data in the previous section, perm(β) method is most conservative in that it only selects Peptostreptococcaceae, where the other three methods choose Bifidobacteriaceae as well as Peptostreptococcaceae. Then the final model can be based on only one bacterium Peptostreptococcaceae:

$$\widehat{\text{RFI}} = -9.63 - 1.04 \times \log \frac{\text{Peptostreptococcaceae}}{1 - \text{Peptostreptococcaceae}},$$

Table 4: Four bacterial taxa in rumen found to be significantly related to RFI. * indicates p-value < 0.05 or confidence interval not containing zero.

Taxa	$\hat{\beta}$	perm(t)	perm(β)	boot(p)	boot(s)	selection frequency
Peptostreptococcaceae	−1.09	0.01*	0.02*	[−1.74, −0.47]*	[−1.75, −0.57]*	0.76
Bifidobacteriaceae	0.90	0.05*	0.09	[0.04, 1.72]*	[0.04, 2.29]*	0.77
o__Clostridiales	0.52	0.18	0.28	[−0.40, 1.35]	[−1.09, 1.56]	0.75
Fibrobacteraceae	−0.33	0.19	0.30	[−0.71, 0.10]	[−0.72, 0.21]	0.81

or based on the two bacterial taxa:

$$\widehat{\text{RFI}} = -3.56 - 1.06 \times \log(\text{Peptostreptococcaceae}) + 1.06 \times \log(\text{Bifidobacteriaceae}).$$

According to Welch et al. (2021), RFI is calculated as the difference between observed and expected feed intake based on metabolic body weight and level of body weight gain. Consequently, if an animal eats less than expected within that level of gain, that animal is considered more efficient than its counterparts. Thus being, because lower values of RFI are desirable, the bacterial taxa correlated negatively with this trait should be the ones that are more desirable. This was the case of Fibrobacteraceae, which had a negative association with RFI values. Fibrobacteraceae have been recognized as main cellulose degraders in ruminant gut systems (Ozbayram et al., 2018). Therefore, a greater abundance of this group likely contribute to a greater degradation of fibrolytic material in the rumen of the most feed-efficient cattle, resulting in more efficient animals (with lower RFI values), and consequently the observed negative correlation between RFI and Fibrobacteraceae. This relation, however, turns out to be non-significant in our analysis. Instead, our result indicates stronger associations of Peptostreptococcaceae and Bifidobacteriaceae to RFI, which invites a further investigation on the biological mechanism.

Results from the current study are in contrast with the ones reported by Welch et al. (2020) who found no correlations between RFI and the ruminal abundances of Bifidobacteriaceae and Peptostreptococcaceae in beef steers. However, the negative association that we detected between RFI and Peptostreptococcaceae might be partially explained by the high rate of production of acetate observed by members of this family (Slobodkin, 2014); given that acetate is an important source of energy in ruminants. On the other hand, bacteria from the family Bifidobacteriaceae are often associated with a healthy gastrointestinal tract, and are often included in probiotics (Gomes and Malcata, 1999; Maldonado-Gómez et al., 2016). Thus, our findings suggest that better feed efficiency, recognizable as lower RFI values, is not necessarily an indicator of better ruminal health.

5.2 Results of All Regressions

We have applied our proposed two-stage procedure to the rest of the combinations of a response and microbiome samples, all of which are with the sample size $n = 20$ and the numbers of taxa $p = 36$. Table 5 summarizes the numbers of significant taxa found by compared approaches. We can see that the GIC-based compositional lasso (Lin et al., 2014) and the cross-validated clr-lasso (Susin et al., 2020) choose either too many (more than 20 taxa out of 36) or none, while the proposed four methods consistently select zero to four taxa. Since the methods do not necessarily agree, we compute the intersection and

union of the selected taxa by the four methods. The intersection result indicates that there are at least some level of consistency in the chosen variables. However in cases with small samples, we suggest to use the union of the results and investigate the collective selection instead of relying on a particular method. Table 6 lists the taxa found to be significant in each case, along with how many methods out of the four declared them so and the signs of association with the response in the final model. See the Supplementary Material Section S2 for the detailed regression results.

Regarding cattle carcass traits, literature linking the microbiome to those traits is scarce. However, Krause et al. (2020) reported a positive association between back fat thickness and the ruminal abundance of Peptococcaceae, which is in line with our study, however, different than our findings, those authors did not find a correlation between back fat and Christensenelaceae. Concerning lipid content in the carcass, Krause et al. (2020) found a negative association between this trait and ruminal abundance of Succinivibrionaceae and F16, and found a positive association with Coriobacteriaceae.

Microorganisms from the family Succinivibrionaceae ferment carbohydrates to succinate and acetate (Santos and Thompson, 2014). It has been shown that acetate induces lower intramuscular adipose tissue lipid and adipocyte volume in beef cattle steers (Smith et al., 2018). Those findings are in line with our results, given that Succinivibrionaceae abundance was negatively correlated with the percentage of lipid measured in the carcass (Table 6). Moreover, this negative correlation was detected by the three and four statistical methods in two different sample types respectively: ruminal and fecal samples; indicating a strong biological evidence that Succinivibrionaceae decreases the amount of lipid in beef carcasses.

Table 5: Number of taxa found to be significant by various approaches. The subscripts r, c, f denote microbiome predictors from rumen, cecum, and feces, respectively.

	RFI_r	RFI_c	RFI_f	$Back_r$	$Back_c$	$Back_f$	$Lipid_r$	$Lipid_c$	$Lipid_f$	YG_r	YG_c	YG_f
GIC	1	1	20	15	26	22	24	24	24	24	26	27
clr-lasso	20	19	19	16	0	19	18	0	18	20	20	22
perm(t)	2	1	1	3	1	1	1	2	3	4	1	2
perm(β)	1	0	1	0	1	1	1	2	2	1	1	2
boot(p)	2	0	2	0	1	0	4	2	3	3	1	0
boot(s)	2	0	1	0	1	1	0	2	3	1	2	2
intersection	1	0	1	0	1	0	0	2	2	0	1	0
union	2	1	2	3	1	1	4	2	4	4	2	2

6 Conclusion

In regression problems with microbiomes as covariates and a key health-related trait as the response, identifying significantly related predictor bacterial taxa is crucial for uncovering the intricate biological mechanism. Even though many regression approaches have been suggested for compositional data, a formal hypothesis testing has not been fully developed yet. We consider the popular sparse log-contrast regression model in this work. The classical hypothesis testing methods for this model are not applicable under the HDLSS situation and non-normal errors. We explored a few non-parametric, resampling-based alternatives for individual regression coefficients testing. Specifically, two permutation tests and two types of bootstrap confidence intervals are considered. From empirical studies, it

Table 6: List of bacterial taxa that are found be significantly related to the four responses in three microbiome data sets. The numbers in the parentheses indicate the number of methods according to which the associated taxa is found significant. Also shown in the parentheses are the signs of the associations. For example, Coriobacteriaceae in cecum is found to be positively associated with Back fat in steers by the all four methods.

	rumen	cecum	feces
RFI	Peptostreptococcaceae (4, −) Bifidobacteriaceae (3, +)	BS11 (1, −)	Lactobacillaceae (4, +) BS11 (1, −)
Back fat	Christensenellaceae (1, +) Peptococcaceae (1, +) Others F (1, −)	Coriobacteriaceae (4, +)	Peptococcaceae (3, −)
Lipid	Succinivibrionaceae (3, −) Coriobacteriaceae (1, +) Fibrobacteraceae (1, −) F16 (1, −)	o__Bacteroidales (4, −) Bifidobacteriaceae (4, +)	Streptococcaceae (4, +) Succinivibrionaceae (4, −) o__Bacteroidales (3, −)
Yield grade	Peptococcaceae (3, +) Others F (3, −) Bifidobacteriaceae (2, +) Coriobacteriaceae (1, +)	Coriobacteriaceae (4, +) Peptococcaceae (1, −)	Peptococcaceae (3, −) o__Rickettsiales (3, +)

is found that generally the resampling-based methods are as effective at least as existing approaches or better at detecting signal predictor variables. Since they do not necessarily agree, we propose to employ an ensemble of methods to make a decision. Application of this approach to the real microbiome data from steers revealed key bacterial taxa that are relevant to beef quality traits.

As for the significance of the chosen variables, since we test the significance of individual variables that are included in the screened model, there is no mathematical guarantee that the variables that are ultimately selected will still be significant in the final model. However, we believe there are high chances that they remain significant in the final model. A heuristic reasoning is that if a variable's contribution is meaningful in the presence of the others, it will probably still be significant in the presence of a reduced set of the other variables. We empirically check out this conjecture in the motivating cattle microbiome data. Once the final models are determined, we test the variables individually again within the final model. We have found that for an overwhelming majority of cases, except for only three cases out of 36, all variables in the final models are statistically significant.

Acknowledgements

This work was partially supported by National Research Foundation of Korea grants 2019R1A2C2002256 and 2021R1A2C1093526.

References

Aitchison, J. (1982), "The statistical analysis of compositional data," *Journal of the Royal Statistical Society: Series B (Methodological)*, 44, 139–160.

Aitchison, J. and Bacon-Shone, J. (1984), "Log contrast models for experiments with mixtures," *Biometrika*, 71, 323–330.

Anderson, M. J. and Robinson, J. (2001), "Permutation tests for linear models," *Australian & New Zealand Journal of Statistics*, 43, 75–88.

Bates, S. and Tibshirani, R. (2019), "Log-ratio lasso: scalable, sparse estimation for log-ratio models," *Biometrics*, 75, 613–624.

Bergamaschi, M., Tiezzi, F., Howard, J., Huang, Y. J., Gray, K. A., Schillebeeckx, C., McNulty, N. P., and Maltecca, C. (2020), "Gut microbiome composition differences among breeds impact feed efficiency in swine," *Microbiome*, 8, 1–15.

Bondell, H. D. and Reich, B. J. (2009), "Simultaneous factor selection and collapsing levels in ANOVA," *Biometrics*, 65, 169–177.

Brodie, J., Daubechies, I., De Mol, C., Giannone, D., and Loris, I. (2009), "Sparse and stable Markowitz portfolios," *Proceedings of the National Academy of Sciences*, 106, 12267–12272.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., et al. (2010), "QIIME allows analysis of high-throughput community sequencing data," *Nature Methods*, 7, 335–336.

Carpenter, J. and Bithell, J. (2000), "Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians," *Statistics in Medicine*, 19, 1141–1164.

Chicco, D. and Jurman, G. (2020), "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, 21, 6.

Davison, A. C. and Hinkley, D. V. (1997), *Bootstrap Methods and Their Application*, no. 1, Cambridge University Press.

Efron, B. (1982), *The Jackknife, the Bootstrap and Other Resampling Plans*, SIAM.

Fan, Y. and Tang, C. Y. (2013), "Tuning parameter selection in high dimensional penalized likelihood," *Journal of the Royal Statistical Society: Series B (Methodological)*, 531–552.

Fang, H., Huang, C., Zhao, H., and Deng, M. (2015), "CCLasso: correlation inference for compositional data through Lasso," *Bioinformatics*, 31, 3172–3180.

Freedman, D. and Lane, D. (1983), "A nonstochastic interpretation of reported significance levels," *Journal of Business & Economic Statistics*, 1, 292–298.

Gomes, A. M. and Malcata, F. X. (1999), "Bifidobacterium spp. and Lactobacillus acidophilus: biological, biochemical, technological and therapeutical properties relevant for use as probiotics," *Trends in Food Science & Technology*, 10, 139–157.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009), *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer Series in Statistics.

Hope, A. C. (1968), "A simplified Monte Carlo significance test procedure," *Journal of the Royal Statistical Society: Series B (Methodological)*, 30, 582–598.

Javanmard, A. and Montanari, A. (2014), "Confidence intervals and hypothesis testing for high-dimensional regression," *The Journal of Machine Learning Research*, 15, 2869–2909.

Jeon, J.-J., Kim, Y., Won, S., and Choi, H. (2020), "Primal path algorithm for compositional data analysis," *Computational Statistics & Data Analysis*, 106958.

Kennedy, F. E. (1995), "Randomization tests in econometrics," *Journal of Business & Economic Statistics*, 13, 85–94.

Krause, T. R., Lourenco, J. M., Welch, C. B., Rothrock, M. J., Callaway, T. R., and Pringle, T. D. (2020), "The relationship between the rumen microbiome and carcass merit in Angus steers," *Journal of Animal Science*, 98, skaa287.

Lee, J. D., Sun, D. L., Sun, Y., Taylor, J. E., et al. (2016), "Exact post-selection inference, with application to the lasso," *Annals of Statistics*, 44, 907–927.

Li, H. (2015), "Microbiome, metagenomics, and high-dimensional compositional data analysis," *Annual Review of Statistics and Its Application*, 2, 73–94.

Lin, W., Shi, P., Feng, R., and Li, H. (2014), "Variable selection in regression with compositional covariates," *Biometrika*, 101, 785–797.

Lourenco, J. M., Kieran, T. J., Seidel, D. S., Glenn, T. C., Silveira, M. F. d., Callaway, T. R., and Stewart Jr, R. L. (2020), "Comparison of the ruminal and fecal microbiotas in beef calves supplemented or not with concentrate," *PloS One*, 15, e0231533.

Lubbe, S., Filzmoser, P., and Templ, M. (2021), "Comparison of zero replacement strategies for compositional data with large numbers of zeros," *Chemometrics and Intelligent Laboratory Systems*, 210, 104248.

Maldonado-Gómez, M. X., Martínez, I., Bottacini, F., O'Callaghan, A., Ventura, M., van Sinderen, D., Hillmann, B., Vangay, P., Knights, D., Hutkins, R. W., et al. (2016), "Stable engraftment of Bifidobacterium longum AH1206 in the human gut depends on individualized features of the resident microbiome," *Cell Host & Microbe*, 20, 515–526.

Manly, B. F. (2018), *Randomization, Bootstrap and Monte Carlo Methods in Biology*, Chapman and Hall/CRC.

Matthews, B. W. (1975), "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405, 442–451.

Meinshausen, N. and Bühlmann, P. (2010), "Stability selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72, 417–473.

NIH Human Microbiome Portfolio Analysis Team (2019), "A review of 10 years of human microbiome research activities at the US National Institutes of Health, Fiscal Years 2007-2016," *Microbiome*, 7, 1–19.

O'Hara, E., Neves, A. L., Song, Y., and Guan, L. L. (2020), "The role of the gut microbiome in cattle production and health: driver or passenger?" *Annual Review of Animal Biosciences*, 8, 199–220.

Ozbayram, E. G., Ince, O., Ince, B., Harms, H., and Kleinsteuber, S. (2018), "Comparison of rumen and manure microbiomes and implications for the inoculation of anaerobic digesters," *Microorganisms*, 6, 15.

Santos, E. d. O. and Thompson, F. (2014), *The Family Succinivibrionaceae*, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 639–648.

Shi, P., Zhang, A., and Li, H. (2016), "Regression analysis for microbiome compositional data," *The Annals of Applied Statistics*, 10, 1019–1040.

Slobodkin, A. (2014), "The Family Peptostreptococcaceae," in *The Prokaryotes: Firmicutes and Tenericutes*, pp. 291–302.

Smith, S., Blackmon, T., Sawyer, J., Miller, R., Baber, J., Morrill, J., Cabral, A., and Wickersham, T. (2018), "Glucose and acetate metabolism in bovine intramuscular and subcutaneous adipose tissues from steers infused with glucose, propionate, or acetate," *Journal of Animal Science*, 96, 921–929.

Srinivasan, A., Xue, L., and Zhan, X. (2021), "Compositional knockoff filter for high-dimensional regression analysis of microbiome data," *Biometrics*, 77, 984–995.

Susin, A., Wang, Y., Lê Cao, K.-A., and Calle, M. L. (2020), "Variable selection in microbiome compositional data analysis," *NAR Genomics and Bioinformatics*, 2, lqaa029.

Ter Braak, C. J. (1992), "Permutation versus bootstrap significance tests in multiple regression and ANOVA," in *Bootstrapping and Related Techniques*, Springer, pp. 79–85.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, 58, 267–288.

Welch, C. B., Lourenco, J. M., Davis, D. B., Krause, T. R., Carmichael, M. N., Rothrock, M. J., Pringle, T. D., and Callaway, T. R. (2020), "The impact of feed efficiency selection on the ruminal, cecal, and fecal microbiomes of Angus steers from a commercial feedlot," *Journal of Animal Science*, 98, skaa230.

Welch, C. B., Lourenco, J. M., Krause, T. R., Seidel, D. S., Fluharty, F. L., Pringle, T. D., and Callaway, T. R. (2021), "Evaluation of the fecal bacterial communities of Angus steers with divergent feed efficiencies across the lifespan from weaning to slaughter," *Frontiers in Veterinary Science*, 8.