The Annals of Applied Statistics
2020, Vol. 14, No. 3, 1471–1492
https://doi.org/10.1214/20-AOAS1354

A BAYESIAN MODEL OF MICROBIOME DATA FOR SIMULTANEOUS IDENTIFICATION OF COVARIATE ASSOCIATIONS AND PREDICTION OF PHENOTYPIC OUTCOMES

BY MATTHEW D. KOSLOVSKY[1,*], KRISTI L. HOFFMAN[2], CARRIE R. DANIEL[3] AND MARINA VANNUCCI[1,†]

[1]*Department of Statistics, Rice University,* [*]*mkoslovsky12@gmail.com;* [†]*marina@rice.edu*
[2]*Alkek Center for Metagenomics & Microbiome Research, Baylor College of Medicine, Kristi.Hoffman@bcm.edu*
[3]*Department of Epidemiology, The University of Texas MD Anderson Cancer Center, CDaniel@mdanderson.org*

One of the major research questions regarding human microbiome studies is the feasibility of designing interventions that modulate the composition of the microbiome to promote health and to cure disease. This requires extensive understanding of the modulating factors of the microbiome, such as dietary intake, as well as the relation between microbial composition and phenotypic outcomes, such as body mass index (BMI). Previous efforts have modeled these data separately, employing two-step approaches that can produce biased interpretations of the results. Here, we propose a Bayesian joint model that simultaneously identifies clinical covariates associated with microbial composition data and predicts a phenotypic response using information contained in the compositional data. Using spike-and-slab priors, our approach can handle high-dimensional compositional as well as clinical data. Additionally, we accommodate the compositional structure of the data via balances and overdispersion typically found in microbial samples. We apply our model to understand the relations between dietary intake, microbial samples and BMI. In this analysis we find numerous associations between microbial taxa and dietary factors that may lead to a microbiome that is generally more hospitable to the development of chronic diseases, such as obesity. Additionally, we demonstrate on simulated data how our method outperforms two-step approaches and also present a sensitivity analysis.

1. Introduction. Human microbiome research seeks to better understand the role of our microbial communities and how they interact with their host, respond to their environment and influence disease (Xia and Sun (2017)). For example, current findings suggest that the microbiome is responsive to diet, as well as other factors, and may influence various metabolic conditions, such as obesity (Maruvada et al. (2017), Sonnenburg and Bäckhed (2016), Dao et al. (2016)). Insights into the relations between microbial composition and both endogenous and exogenous factors may help researchers design personalized intervention strategies to modulate and maintain a healthy microbiome community (Knights et al. (2011), Xu and Knight (2015)). However, complex environmental interactions with the microbiome challenge our understanding of community function and its impact on health (Shetty et al. (2017)).

Human microbiome studies typically have two main objectives: (1) identifying factors that characterize the composition of the microbiome and (2) predicting biological, genetic, clinical, or experimental conditions using microbial abundance data (Xia and Sun (2017)). For both objectives, analysis is challenged for various reasons, including vast amounts of intra- and intersubject heterogeneity in taxonomic abundance as well as the compositional structure

Received April 2020.
Key words and phrases. Bayesian statistics, joint modeling, multivariate count data, prediction, variable selection.

and high dimensionality of the data. While each of these objectives have been extensively researched separately, we are unaware of any attempts to jointly model all of the data to achieve both objectives simultaneously.

For objective (1), there are various methods available to infer relations between covariates and multivariate count data (Zhang et al. (2017)). For microbial count data, researchers have previously used Dirichlet-multinomial models, since these models can handle overdispersed data that arise from within- and between-subject variability in microbial data (La Rosa et al. (2012), Wadsworth et al. (2017), Chen and Li (2013)). In exploratory research studies researchers have used penalized likelihood approaches to simultaneously shrink unassociated covariates' regression coefficients to zero and estimate the effects of associated covariates (Chen and Li (2013), Wang and Zhao (2017a)). The proven efficiency of these methods comes at a price, as optimization routines are challenged by complex data structures (Wang and Zhao (2017a)), and they do not fully capture the uncertainty of model selection. Alternatively, Bayesian methods are available which capture uncertainty in the model by exploring the model space using Markov chain Monte Carlo (MCMC) algorithms. Wadsworth et al. (2017) recently developed a Bayesian approach for identifying Kegg orthology pathways that were associated with microbial abundance data using spike-and-slab priors for the regression coefficients. In confirmatory settings Mao, Chen and Ma (2017) demonstrate how including covariates in a Bayesian graphical compositional regression model can improve accuracy in testing results and reduce false discoveries.

For objective (2), researchers may be interested in using microbial abundances to predict outcomes of interest, such as body mass index (BMI) (Lin et al. (2014), Wang and Zhao (2017b)). Microbial abundance data are an example of multivariate compositional data where the magnitude of a single component depends on the sum of all the components' counts. This dependency causes inferential biases and computational challenges if the compositional data are modeled in their raw form. To properly model compositional data, log-ratio transformations are used. Various log-ratio transformations have been proposed, including additive, centered and isometric (Aitchison (1986), Egozcue et al. (2003)). Isometric log-ratio transformations, in particular, allow researchers to properly model compositional data using balances to make inference on subsets of the taxa, as opposed to individual taxon (Morton et al. (2017), Pinto et al. (2017)). Balances are defined proportionally to the difference in the mean of the log-transformed abundances between two groups and are scale invariant. Thus, they can equivalently be constructed with raw counts or the relative proportion of counts. Additionally, researchers can use prior knowledge of structure in the data to construct balances (Fišerová and Hron (2011), Morton et al. (2017)). Once the raw compositional data are appropriately transformed, they can be used in standard analysis methods, such as linear regression and principle components analysis (Chen, Zhang and Li (2017), Gloor et al. (2017), Garcia et al. (2013), Hron, Filzmoser and Thompson (2012), Lin et al. (2014), Mert et al. (2018), Pinto et al. (2017), Shi, Zhang and Li (2016), Silverman et al. (2017), Bruno, Greco and Ventrucci (2016)).

In this work we propose a Bayesian joint modeling approach that simultaneously identifies clinical covariates associated with microbial composition data and predicts a phenotypic response using information contained in the compositional data. We conjecture that separate, two-step approaches may underestimate model uncertainty since the microbial composition data are typically treated as fixed when used to predict phenotypic responses. This may produce biased interpretation of the model (Chatfield (1995)). On the contrary, our joint modeling of all the data allows researchers to make inference on the relation between clinical measures and health outcomes, via their relation to the composition of the microbiome. Additionally, if there is a true relation between microbial composition and the phenotypic outcome, properly accommodating microbial heterogeneity based on clinical measures may result in a

more accurate prediction. Our method is designed to accommodate high-dimensional microbial and clinical measures data, overdispersion in the count data as well as the structure of the compositional data.

We apply our method to understand the relation between dietary intake and taxonomic composition of the microbiome and BMI. We have available dietary assessments and oral and fecal microbiome data from an ancillary study conducted among healthy obese and lean individuals from the Houston, TX, area (Versace et al. (2015)). The study was designed to assess eating behavior and the microbiome in selfreported healthy individuals. In our analysis, we find numerous associations between microbial taxa and dietary factors that may lead to a microbiome that is generally more hospitable to the development of chronic diseases, such as obesity. Additionally, we use simulated data to compare selection performance and predictive ability of our proposed method with respect to various two-step approaches that first select covariates associated with multivariate count data and then perform variable selection on balances, constructed using estimated count probabilities for the prediction of a phenotypic outcome.

In Section 2 we introduce our proposed joint model and describe the posterior inference. In Section 3 we apply our method to data collected to investigate the relation between diet, microbial samples and BMI. In Section 4 we perform a simulation study aimed at comparing performance with alternative approaches and present a sensitivity analysis. In Section 5 we provide concluding remarks.

2. Methods. Let y_i be the observed phenotypic outcome for the ith subject, $i = 1, \ldots, N$. Also, let $z_i' = (z_{i,1}, \ldots, z_{i,J})$ represent a J-dimensional vector of microbial taxa abundance counts and $x_i' = (x_{i,1}, \ldots, x_{i,P})$ be a vector of P dietary covariates collected on the ith subject. In the Bayesian paradigm, inference is drawn from the posterior distribution which is proportional to the likelihood of the observed data times the prior distribution of the parameters in the model. Here, we jointly model the compositional count and response data by parameterizing their likelihoods with a shared parameter (i.e., the probability of the compositional taxa).

In our joint modeling we first assume that taxa counts z_i follow a Multinomial distribution

$$z_i \sim \text{Multinomial}(\dot{z}_i | \psi_i), \tag{1}$$

with $\dot{z}_i = \sum_{j=1}^{J} z_{i,j}$ and ψ_i defined on the J-dimensional simplex

$$S^{J-1} = \left\{ (\psi_{i,1}, \ldots, \psi_{i,J}) : \psi_{i,j} \geq 0, \forall j, \sum_{j=1}^{J} \psi_{i,j} = 1 \right\}.$$

To account for overdispersion in the multivariate count data, we specify a conjugate prior on the taxa probabilities,

$$\psi_i \sim \text{Dirichlet}(\gamma_i), \tag{2}$$

with the J-dimensional vector $\gamma_i = (\gamma_{i,j} > 0, \forall j \in J)$, similarly to Wadsworth et al. (2017) and La Rosa et al. (2012). Note that, if we were only interested in identifying dietary covariates associated with the taxa count data, we could integrate out the ψ_i and model z_i with a Dirichlet-multinomial(γ_i), similar to Wadsworth et al. (2017). However, for our joint model we estimate ψ since it serves as the shared parameter between the likelihood of the phenotypic response $\boldsymbol{Y}$ and compositional data $\boldsymbol{Z}$, as described below. Next, we incorporate dietary covariate effects into the model by using a log-linear regression framework. Specifically, we set $\lambda_{i,j} = \log(\gamma_{i,j})$ and assume

$$\lambda_{i,j} = \alpha_j + \sum_{p=1}^{P} \varphi_{jp} x_{i,p}, \tag{3}$$

where $\varphi_j = (\varphi_{j1}, \ldots, \varphi_{jP})$ represents the covariates' potential relation with the jth compositional taxon and α_j is a taxon-specific intercept term. By exponentiating (3) we ensure positive hyperparameters for the Dirichlet distribution. Note that, while this analysis focuses on dietary factors, other covariates, for example, age, sex, medication use, could be included in $\boldsymbol{x}$ as well.

Under this parameterization the number of potential models to choose from when performing model selection grows quickly, even for small covariate spaces. For example, $P = 10$ covariates and just $J = 2$ compositional taxa results in over a million potential models. To reduce the dimension of the model, we employ multivariate variable selection spike-and-slab priors (Richardson, Bottolo and Rosenthal (2011), Stingo et al. (2010)) that identify dietary covariates that are associated with each compositional taxon, as opposed to spike-and-slab constructions that select variables as relevant to either all or none of the responses (Brown, Vannucci and Fearn (1998)). Here, we assume the covariates' inclusion in the model is characterized by a latent, $J \times P$-dimensional inclusion vector $\boldsymbol{\zeta}$. With this formulation, $\zeta_{jp} = 1$ indicates that covariate p is associated with compositional taxon j and zero otherwise. The prior for φ_{jp}, given ζ_{jp}, follows a mixture of a normal distribution and a Dirac-delta function at zero, δ_0, and is commonly referred to as the spike-and-slab prior. Specifically,

$$\varphi_{jp}|\zeta_{jp}, r_j^2 \sim \zeta_{jp} \cdot N\big(0, r_j^2\big) + (1 - \zeta_{jp}) \cdot \delta_0(\varphi_{jp}), \tag{4}$$

where r_j^2 is set large to impose a vague prior for the regression coefficients in the case of covariate inclusion. We assume each ζ_{jp} follows a Bernoulli prior, $p(\zeta_{jp}) \sim \text{Bernoulli}(\omega_{jp})$, where $\omega_{jp} \sim \text{Beta}(a, b)$. Integrating out ω_{jp} leads to

$$p(\zeta_{jp}) = \frac{\text{Beta}(\zeta_{jp} + a, 1 - \zeta_{jp} + b)}{\text{Beta}(a, b)}. \tag{5}$$

Hyperparameters a and b can be set to impose various levels of sparsity in the model. Lastly, we assume the intercept terms α_j follow a $N(0, \sigma_j^2)$, where σ_j^2 are set large to impose vague priors.

Next, we model the relation between the phenotypic response $\boldsymbol{Y}$ and the compositional data $\boldsymbol{Z}$ via a multivariable linear regression model. Typically, raw (or relative) compositional data used to construct balances for regression modeling are treated as fixed. In our joint model we assume they are random and calculate balances using the compositional taxa probabilities $\boldsymbol{\psi}$. As such, our model is related to the broad class of methods that make distributional assumptions for covariates to reduce inferential biases (Carroll et al. (2006), Shi, Zhang and Li (2016), Tadesse et al. (2005)).

Let the observed outcome y_i be related to an M-dimensional set of balances following

$$y_i = \alpha_0 + \sum_{m=1}^{M} \beta_m B(\psi)_{i,m} + \epsilon_i, \tag{6}$$

where α_0 is an intercept term, β_m is a regression coefficient for its respective balance as a function of ψ, $B(\psi)_{i,m}$ and $\epsilon_i \sim N(0, \sigma^2)$. Note that this formulation can easily be extended to include other covariates, in addition to the balances, that may be associated with the phenotypic response. To demonstrate how to construct a balance, consider two nonoverlapping partitions of ψ, ψ_+ and ψ_-. The balance calculated for this partition is defined as

$$B(\psi_+, \psi_-) = \sqrt{\frac{|\psi_-||\psi_+|}{|\psi_-| + |\psi_+|}} \log\left[\frac{g(\psi_+)}{g(\psi_-)}\right], \tag{7}$$

where $|\cdot|$ is the dimension of a given subset and $g(\cdot)$ is the geometric mean defined as $(\prod_{r=1}^{|\psi|} \psi_r)^{1/|\psi|}$. In our approach balances are constructed using sequential binary separation (Egozcue and Pawlowsky-Glahn (2005)), producing $M = J - 1$ potential balances in the

model. It is important to note that prediction performance of the model does not depend on the order in which the partitions are defined (Egozcue and Pawlowsky-Glahn (2005)). Additionally, log-ratio transformations cannot handle observed zero counts and require adjustments based on assumptions of their occurrence (Martín-Fernández et al. (2015)). To handle zero values for the $\boldsymbol{\psi}$, we use a multiplicative replacement strategy in which zero values are replaced with relatively small pseudovalues, and the corresponding probability vector is scaled to sum to one (Martin-Fernandez, Barceló-Vidal and Pawlowsky-Glahn (2000)). Note that this strategy does not affect the DM portion of the model. There, zero counts are admissible.

In practice, the dimension of the balance space can be large relative to N. To induce sparsity on the dimension space of the balances, we take a similar strategy as above and assume that the prior for β_m, conditioned upon a latent indicator ξ_m and σ^2, follows

$$\beta_m|\xi_m, \sigma^2 \sim \xi_m \cdot N\big(0, h_\beta \sigma^2\big) + (1 - \xi_m) \cdot \delta_0(\beta_m), \tag{8}$$

and, similarly,

$$p(\xi_m) = \frac{\text{Beta}(\xi_m + a_m, 1 - \xi_m + b_m)}{\text{Beta}(a_m, b_m)}. \tag{9}$$

The prior for the intercept term is $\alpha_0|\sigma^2 \sim N(0, h_{\alpha_0}\sigma^2)$. Large values for the hyperparameters h_{α_0} and h_β impose vague priors on the intercept term and regression coefficients, respectively. To complete the prior specification of the model, we set $\sigma^2 \sim \text{Inverse-gamma}(a_0, b_0)$, with $a_0 > 0$ and $b_0 > 0$. A graphical representation of our model is provided in Figure 1.

To summarize, our joint model assumes that the distribution of the phenotypic response $\boldsymbol{Y}$ and taxa abundance counts $\boldsymbol{Z}$ are conditionally independent given the compositional taxa probabilities $\boldsymbol{\psi}$. Specifically, we assume

$$f(\boldsymbol{Y}|\boldsymbol{\psi}) f(\boldsymbol{Z}|\boldsymbol{\psi}) p(\boldsymbol{\psi}|\boldsymbol{x}), \tag{10}$$

where $f(\boldsymbol{Y}|\boldsymbol{\psi})$ models the prediction of the phenotypic response, based on balances calculated using the compositional taxa probabilities $\boldsymbol{\psi}$ and $f(\boldsymbol{Z}|\boldsymbol{\psi})p(\boldsymbol{\psi}|\boldsymbol{x})$ characterizes the associations between taxa abundance counts and clinical covariates. In the Supplementary Material we provide a simulation study demonstrating the model's invariance to balance specification and how balance sparsity can improve prediction performance (Koslovsky et al. (2020)).

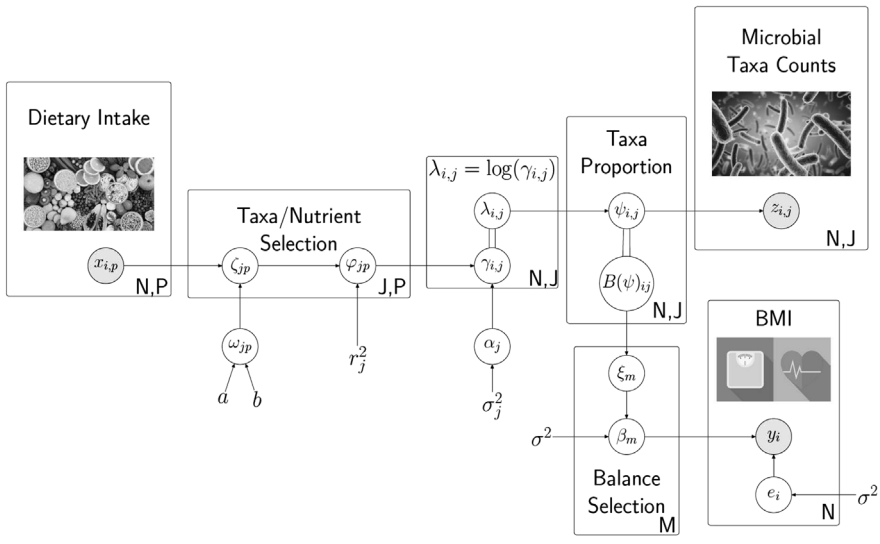

FIG. 1. *Graphical representation of the proposed Bayesian joint model for identifying dietary intake covariates associated with microbial taxa and predicting BMI.*

2.1. *Posterior inference*. We implement a Metropolis–Hastings algorithm within a Gibbs sampler. Inspired by techniques used in Bayesian nonparametrics (James, Lijoi and Prünster (2009), Argiento, Bianchini and Guglielmi (2015)), we adopt a data augmentation approach for the Dirichlet-multinomial portion of the model, which avoids Metropolis–Hastings updates for the taxa proportion parameters $\boldsymbol{\psi}$ and greatly aids scalability. First, we integrate out α_0, $\boldsymbol{\beta}$, and σ^2 in the conditional likelihood for $\boldsymbol{Y}$ to obtain a multivariate t-distribution,

$$\boldsymbol{Y} \sim \boldsymbol{t}_{2a_0}\left(\mathbf{0}_N, \frac{b_0}{a_0}(I_N + h_\alpha \mathbf{1}_N \mathbf{1}_N' + h_\beta \boldsymbol{B}(\boldsymbol{\psi})_\xi \boldsymbol{B}(\boldsymbol{\psi})_\xi')\right),$$

with $\mathbf{0}_N$ an N-dimensional vector of zeros, I_N an $N \times N$ identity matrix, $\mathbf{1}_N$ an N-dimensional vector of ones and $\boldsymbol{B}(\boldsymbol{\psi})_\xi$ the matrix of balances included in the model. Next, we introduce latent variables $c_{i,j}$ such that $\psi_{i,j} = c_{i,j}/T_i$ with $T_i = \sum_{j=1}^J c_{i,j}$ and reparameterize equation (1) as

$$z_i \sim \text{Multinomial}(\dot{z}_i | c_i / T_i),$$

where $c_i' = (c_{i,1}, \ldots, c_{i,J})$, and $c_{i,j} \sim \text{Gamma}(\gamma_{i,j}, 1)$. Then, we write the joint distribution of z_i and ψ_i, in terms of c_i, as

$$p(z_i, c_i | \gamma_i) \propto \frac{c_{i,1}^{z_{i,1}} \times \cdots \times c_{i,J}^{z_{i,J}}}{T_i^{\dot{z}_i}} \prod_{j=1}^J \frac{1}{\Gamma(\gamma_{i,j})} c_{i,j}^{\gamma_{i,j}-1} \exp(-c_{i,j}). \tag{11}$$

To avoid the calculation of the $T_i^{\dot{z}_i}$ terms, we introduce auxiliary parameters $\boldsymbol{u}' = (u_1, \ldots, u_N)$, such that $u_i | T_i \sim \text{Gamma}(\dot{z}_i, T_i)$. Using the gamma identity

$$\frac{1}{T_i^{\dot{z}_i}} = \int_0^\infty \frac{1}{\Gamma(\dot{z}_i)} u_i^{\dot{z}_i - 1} \exp(-T_i u_i) \partial u_i,$$

we can express (11) as

$$p(z_i, c_i | \gamma_i) \propto \int_0^\infty \frac{1}{\Gamma(\dot{z}_i)} u_i^{\dot{z}_i - 1} \exp(-T_i u_i) c_{i,1}^{z_{i,1}} \times \cdots \times c_{i,J}^{z_{i,J}} \prod_{j=1}^J \frac{1}{\Gamma(\gamma_{i,j})} c_{i,j}^{\gamma_{i,j}-1} \exp(-c_{i,j}) \partial u_i.$$

Using (10) and transforming ψ_i with c_i, the joint posterior distribution simplifies as proportional to

$$f(\boldsymbol{Y}|\boldsymbol{\xi}, \boldsymbol{c}) f(\boldsymbol{Z}|\boldsymbol{c}) p(\boldsymbol{c}|\boldsymbol{\alpha}, \boldsymbol{\varphi}, \boldsymbol{\zeta}, \boldsymbol{x}) p(\boldsymbol{\xi}) p(\boldsymbol{\alpha}) p(\boldsymbol{\varphi}|\boldsymbol{\zeta}) p(\boldsymbol{\zeta}) p(\boldsymbol{u}|\boldsymbol{c}),$$

where the integral obtained from the data augmentation technique is naturally estimated as a part of the full MCMC routine.

The generic iteration of the MCMC comprises of the following updates:

- Update each α_j: Propose $\alpha_j' \sim N(\alpha_j, 0.5)$. Accept α_j' with probability

$$\min\left\{\frac{p(\boldsymbol{c}|\boldsymbol{\alpha}', \boldsymbol{\varphi}, \boldsymbol{\zeta}, \boldsymbol{x}) p(\alpha_j')}{p(\boldsymbol{c}|\boldsymbol{\alpha}, \boldsymbol{\varphi}, \boldsymbol{\zeta}, \boldsymbol{x}) p(\alpha_j)}, 1\right\}.$$

- Jointly update a ζ_{jp} and φ_{jp} following the two-step approach proposed by Savitsky, Vannucci and Sha (2011).

Between-model step: Randomly select a ζ_{jp}. If $\zeta_{jp} = 1$, perform a Delete step, otherwise perform an Add step.

– Delete—Propose $\zeta_{jp}' = 0$ and $\varphi_{jp}' = 0$. Accept proposal with probability

$$\min\left\{\frac{p(\boldsymbol{c}|\boldsymbol{\alpha}, \boldsymbol{\varphi}', \boldsymbol{\zeta}', \boldsymbol{x}) p(\zeta_{jp}')}{p(\boldsymbol{c}|\boldsymbol{\alpha}, \boldsymbol{\varphi}, \boldsymbol{\zeta}, \boldsymbol{x}) p(\varphi_{jp}|\zeta_{jp}) p(\zeta_{jp})}, 1\right\}.$$

– Add—Propose $\zeta'_{jp} = 1$. Then, sample a $\varphi'_{jp} \sim N(\varphi_{jp}, 0.5)$.
Accept proposal with probability

$$\min\left\{\frac{p(\boldsymbol{c}|\boldsymbol{\alpha}, \boldsymbol{\varphi}', \boldsymbol{\zeta}', \boldsymbol{x})p(\varphi'_{jp}|\zeta'_{jp})p(\zeta'_{jp})}{p(\boldsymbol{c}|\boldsymbol{\alpha}, \boldsymbol{\varphi}, \boldsymbol{\zeta}, \boldsymbol{x})p(\zeta_{jp})}, 1\right\}.$$

Within-model step:

– Propose a $\varphi'_{jp} \sim N(\varphi_{jp}, 0.5)$ for each covariate currently selected in the model ($\zeta_{jp} = 1$). Accept each proposal with probability

$$\min\left\{\frac{p(\boldsymbol{c}|\boldsymbol{\alpha}, \boldsymbol{\varphi}', \boldsymbol{\zeta}, \boldsymbol{x})p(\varphi'_{jp}|\zeta_{jp})}{p(\boldsymbol{c}|\boldsymbol{\alpha}, \boldsymbol{\varphi}, \boldsymbol{\zeta}, \boldsymbol{x})p(\varphi_{jp}|\zeta_{jp})}, 1\right\}.$$

- Update each $c_{i,j}$ via a Gibbs step:

– Gamma$(c_{i,j}|z_{i,j} + \gamma_{i,j}, u_i + 1)$.

- Update each u_i via a Gibbs step:

– Gamma$(u_i|\dot{z}_i, T_i)$.

- Update ξ_m via an Add/Delete step: Select a random ξ_m. If $\xi_m = 1$, perform a Delete step ($\xi'_m = 0$), otherwise perform an Add Step ($\xi'_m = 1$). For both Add and Delete steps, accept proposal with probability

$$\min\left\{\frac{f(\boldsymbol{Y}|\boldsymbol{\xi}', \boldsymbol{c})p(\xi'_m)}{f(\boldsymbol{Y}|\boldsymbol{\xi}, \boldsymbol{c})p(\xi_m)}, 1\right\}.$$

For implementation the algorithm is initiated at a set of arbitrary parameter values and then used to generate samples of the posterior distribution. After burn-in, a procedure which involves removing a subset of samples that may be influenced by initialization, the remaining samples are used for inference. To determine inclusion in the model, the marginal posterior probability of inclusion (MPPI) for each of the covariates and balances is determined by taking the average of their respective inclusion indicator's MCMC samples. Note that a covariate has a unique inclusion indicator for each of the compositional taxon. Commonly, variables are included in the model if their MPPI ≥ 0.50 (Barbieri and Berger (2004)). Alternatively, Newton et al. (2004) propose using a threshold based on a Bayesian false discovery rate to control for multiplicity.

To evaluate the prediction accuracy of the model, cross-validation can be performed by fitting the model on a subset of the data (training set) and evaluating prediction performance on the remaining data (testing set) by calculating the prediction mean squared error. To obtain predictions of the testing outcomes, $\ddot{\boldsymbol{\mathcal{Y}}}$, set

$$\hat{\ddot{\boldsymbol{\mathcal{Y}}}} = \hat{\alpha}_0 + \frac{1}{S}\sum_{s=1}^{S} \boldsymbol{B}(\ddot{\boldsymbol{\psi}})\hat{\boldsymbol{\beta}}_{\boldsymbol{\xi}^s}, \tag{12}$$

where $\hat{\alpha}_0 = (n + h_{\alpha_0}^{-1})^{-1}\mathbf{1}'_n\boldsymbol{Y}$ and

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{\xi}^s} = \left(\boldsymbol{B}(\boldsymbol{\psi}^s)'_{\boldsymbol{\xi}^s}\boldsymbol{B}(\boldsymbol{\psi}^s)_{\boldsymbol{\xi}^s} + h_\beta^{-1}\boldsymbol{I}_{|\boldsymbol{\xi}^s|}\right)^{-1}\boldsymbol{B}(\boldsymbol{\psi}^s)'_{\boldsymbol{\xi}^s}\boldsymbol{Y}, \tag{13}$$

with $\boldsymbol{B}(\ddot{\boldsymbol{\psi}})$ the matrix of balances from the testing set, $\boldsymbol{B}(\boldsymbol{\psi}^s)_{\boldsymbol{\xi}^s}$ the matrix of balances selected in the s^{th} MCMC iteration of the training model and $|\boldsymbol{\xi}^s|$ the number of balances selected in $\boldsymbol{B}(\boldsymbol{\psi}^s)_{\boldsymbol{\xi}^s}$, following Brown, Vannucci and Fearn (1998). Since the $\ddot{\boldsymbol{\psi}}$ used to calculate the balances are not observed for the testing set, we estimate them as

$$\ddot{\psi}_{i,j} = \frac{\ddot{z}_{i,j} + \hat{\lambda}_{i,j}}{\sum_{j=1}^{J}\ddot{z}_{i,j} + \hat{\lambda}_{i,j}}, \tag{14}$$

where

$$\hat{\lambda}_{i,j} = \exp\left(\frac{1}{S}\sum_{s=1}^{S}\left(\alpha_j^s + \sum_{p=1}^{P}\varphi_{jp}^s \ddot{x}_{i,p}\right)\right), \tag{15}$$

$\ddot{z}_i$ and $\ddot{x}_i$ represent the multivariate counts and covariates observed for the ith testing subject and α_j^s and φ_{jp}^s are MCMC samples obtained from the training model. When splitting the data is impractical due to small sample sizes, leave-one-out cross-validation approximation procedures can be used, for example, following the approach proposed by Vehtari, Gelman and Gabry (2017). This approach approximates leave-one-out (LOO) cross-validation with the expected log pointwise predictive density (e.p.l.d.). By using Pareto smoothed importance sampling (PSIS) for estimation, it provides a more stable estimate compared to the method of Gelfand (1996).

3. Case study on diet and the microbiome. We applied our joint model to dietary assessment, oral and fecal microbiome data from an ancillary study conducted among healthy obese and lean individuals from the Houston, TX, area (Versace et al. (2015)). In addition to dietary intake, physical activity and eating behavior questionnaires, participants provided stool and oral swab samples for microbiome analysis. Participant height and weight were also measured. Adults, 21 to 55 years of age were recruited to maximize variability in usual diet/eating habits and BMI, while minimizing extraneous factors known to influence the oral and/or fecal microbiome. Individuals who used antibiotics within the past 30 days, were current smokers, had any chronic or acute condition that required exclusionary medications or dietary restrictions, reported substantial weight changes ($\pm$5 kg) in the past three months and women who were recently pregnant/lactating were excluded from the study. Approximately two-thirds of the sample were female, and 40% were obese. Participants provided fresh stool samples using an in-home collection kit with sterile swab and no storage media between their first and second in-person visit. Study staff also collected an oral (buccal) swab sample from the participant at the in-person visit.

Habitual dietary intake data were collected via the 134-food item National Cancer Institute Dietary History Questionnaire (DHQ) II, enabling evaluation of food groups, macronutrients, vitamins, minerals and eight dietary supplements (Millen et al. (2005), Subar et al. (2001)). DHQ II responses were processed via the National Cancer Institute's Diet*Calc software and initially produced 214 variables of estimated daily nutrient and food group intake. Of these, 140 variables were aggregated or excluded due to redundancy and/or low variation. The remaining 74 nutrient and food group variables were adjusted for caloric intake prior to analysis (Willett (1998)). Only participants whose total energy intake was considered plausible ($800 <$ kcal < 4200 and $600 <$ kcal < 3500, for men and women, respectively) were included in this analysis. Two 24-hour dietary recalls were compared to each individual's DHQ data to assess accuracy and consistency but not included in the current analysis.

For microbiome assessment, stool and oral swab specimens underwent total genomic DNA extraction and 16S rDNA sequencing, as described previously (Gopalakrishnan et al. (2018), Hoffman et al. (2018)). While highly conserved, the 16S rRNA gene is commonly used for bacterial identification due to regions of high variability (Li (2015)). Sequencing was performed via the Illumina MiSeq platform and targeted the V4 region. Resulting reads were processed and clustered into operational taxonomic units (OTUs) using UPARSE (Edgar (2013)) at an identity threshold of 95%. OTUs were mapped using a V4-optimized version of the SILVA database (v.123). To reduce the number of spurious relationships detected, we further limited analysis to only those OTUs identified in at least 10% of participants. This resulted in 245 and 185 taxon for the fecal and oral samples, respectively. For consistency,

only participants who provided both stool and oral swab specimens were used in this analysis, resulting in a sample size of $N = 56$.

The objective of our study was to identify relations between OTUs in microbial samples and dietary covariates, while simultaneously predicting body mass index (BMI) using our proposed joint model. In two separate analyses we modeled fecal and oral microbial samples and compared their predictive performance for BMI, controlling for age and sex by having them as fixed covariates in the model. Prior to analysis, the dietary data were standardized to mean zero and variance one. Additionally, the BMI measures were centered at the sample mean. For inference we set hyperparameters $h_{\alpha_0} = h_\beta = 1$, $a_0 = b_0 = 2$ and $\sigma^2 = r^2 = 10$. Additionally, we set the hyperparameters for the beta-binomial priors to $a = a_m = 1$, $b = 9$ and $b_m = 4$ for both models. This corresponded to a 10% and 20% prior probability of inclusion for dietary factors and balances, respectively. Note, these priors were chosen since they obtained the best prediction performance in our sensitivity analysis (see end of Section 3.1). The MCMC algorithm was run for 50,000 iterations, with the first 25,000 treated as burn-in and thinned every 10th sample. In this analysis runtimes were 16.6 and 15.8 minutes for the fecal and oral models, respectively, on a 2.5 GHz dual-core Intel Core i5 processor with eight GB RAM. Trace plots of the log-posterior distribution indicate good convergence and mixing. Covariate and balance inclusion was determined using the median model approach (i.e., MPPI ≥ 0.50).

3.1. *Results*. Figures 2 and 3 show the marginal posterior probabilities of inclusion (MPPI) for dietary covariates, indexed across compositional taxa, for the model fit to the oral and fecal microbial data. Figures 4 and 5 present heatmaps of the associations between dietary covariates and microbial abundances identified in the oral and fecal models, respectively. For interpretability, taxa are assigned to their likely representative bacterial genera using Basic Local Alignment Search Tool (BLAST) (Zhang et al. (2000)). Further details of the relations between selected pairs of taxa and dietary covariates are found in the Supplementary Material (Koslovsky et al. (2020)). Six balances calculated from the oral microbial sample were identified as associated with BMI, compared to seven balances from the fecal sample. As for prediction, due to the study's relatively small sample size, we chose to compare accuracy

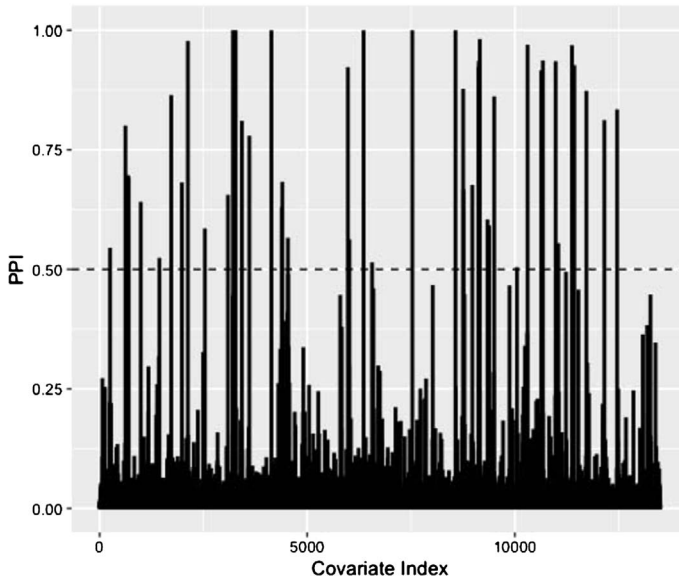

FIG. 2. *Marginal posterior probabilities of inclusion for dietary covariates indexed across compositional taxa using the oral microbial data. Dashed line represents the median model threshold* (0.50).

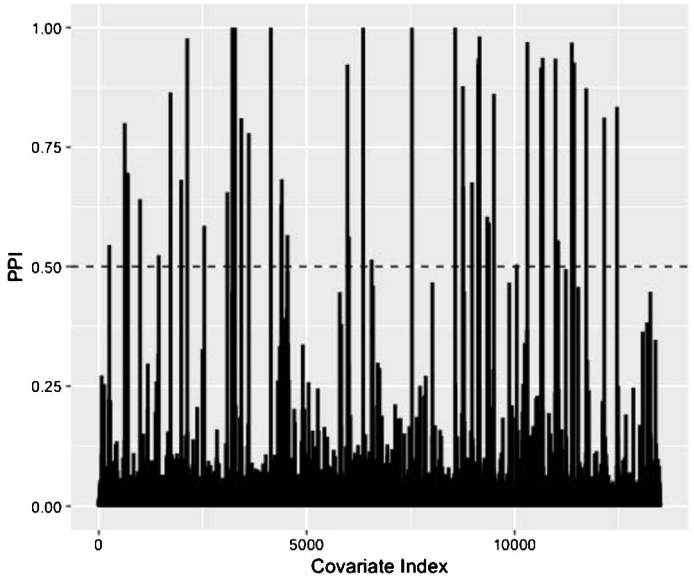

FIG. 3. *Marginal posterior probabilities of inclusion for dietary covariates indexed across compositional taxa using the fecal microbial data. Dashed line represents the median model threshold* (0.50).

of the results using the approach proposed by Vehtari, Gelman and Gabry (2017). We used the R package loo (Vehtari, Gelman and Gabry (2016)), which requires the pointwise log-likelihood, $f(y_i|\xi^s, \psi_i^s)$, for each subject $i = 1, \ldots, N$ calculated at each MCMC iteration $s = 1, \ldots, S$ and produces an estimated $\widehat{\text{e.p.l.d.}}$ value, with larger values implying a superior model. In our analysis the models fit with the oral and fecal data provided similar results ($\widehat{\text{e.p.l.d.}}_{\text{ORAL}} = -200.2$ versus $\widehat{\text{e.p.l.d.}}_{\text{FECAL}} = -201.6$, respectively).

Causal links between diet, the gut microbiome and BMI/obesity are becoming clearer (Maruvada et al. (2017), Ridaura et al. (2013), Turnbaugh (2017)). Microbiota play a key role in the extraction, absorption and storage of energy from dietary intake. Some of the most compelling findings are for "Western-style" dietary patterns which are typically characterized by low intake of fiber-rich plant foods and high intake of meat and added sugars, leading to a microbiome that is generally more hospitable and supportive of the development of obesity and other chronic diseases (Turnbaugh (2017), Valdes et al. (2018)). Interestingly, across both the oral and fecal microbiome, we observed several associations with "Western-style" dietary factors and their counterparts, for example, different nutrient rich and prebiotic vegetable food groups as well as various B vitamins (those largely found in animal sources) and antioxidant nutrients derived from both dietary intake and supplement use.

Looking at the results on the fecal microbiome, we observe several dietary relationships with Bacteroides, including lactose and consumption of dark green vegetables. Several Bacteroides species (a common and abundant genus within Bacteroidetes phylum) and their interactions with diet have been implicated in obesity (David et al. (2014), Kovatcheva-Datchary et al. (2015), Wu et al. (2011)). Bacteroides have a broad capacity to use diverse types of carbohydrates or dietary polysaccharides which include glucose, sucrose and starch for energy and can "step up" when dietary fiber intake is low, tapping into other sources of energy for the gut (Gurry et al. (2018), Marcobal et al. (2011), Sonnenburg et al. (2010)). In a recent and similarly conducted epidemiologic study of healthy adults, low fiber intake was associated with higher Bacteroides uniformis (Lin et al. (2018)). As with Bacteroides, Escherichia also metabolizes carbohydrates for energy and has been associated with intestinal inflammation in animal models fed a Western (high-fat/high-sugar) diet (Agus et al. (2016),

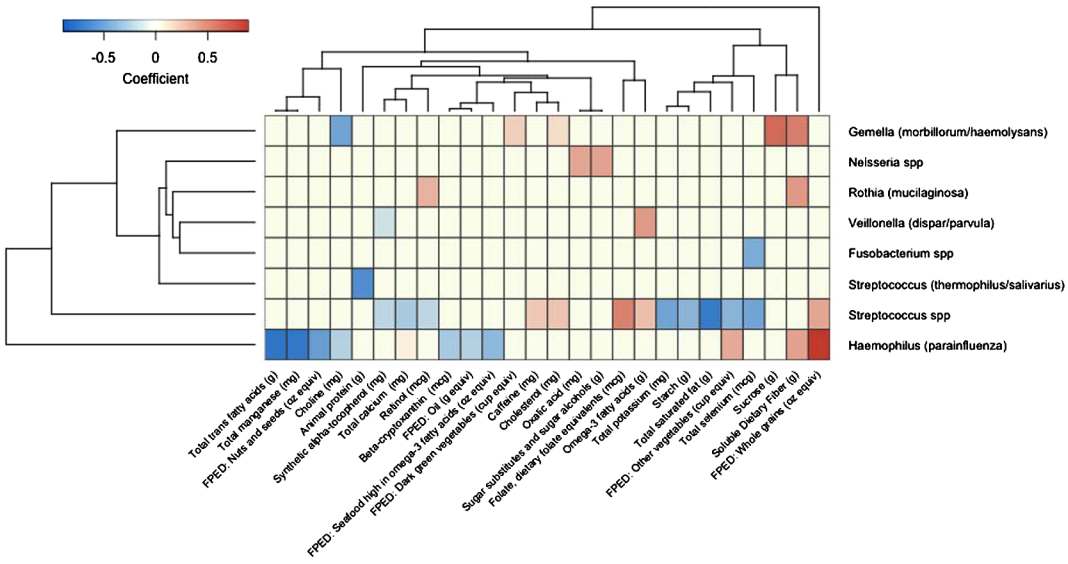

FIG. 4. *Estimated regression coefficients for corresponding dietary factors and oral microbial abundance associations, grouped by BLAST assignment and identified using the proposed joint model.*

Martinez-Medina et al. (2014)), consistent with the Escherichia-sucrose association observed in this analysis. We additionally found a number of associations linked to Prevotella, including maltose, galactose and sugar substitutes and alcohols—commonly found in snack foods. Greater levels of Prevotellaceae have been observed in obese individuals (Zhang et al. (2009)), and Prevotella copri specifically has been found in higher abundance among overweight and obese type 2 diabetics (Leite et al. (2017)).

Similar to the gut microbiome, the oral microbiome may also be shaped by dietary habits (Fan et al. (2018), Hansen et al. (2018), Ercolini et al. (2015), Kato et al. (2017), Peters et al. (2018)). Differences in the diversity and abundance of oral bacteria between overweight/obese and healthy weight individuals have now been documented in several studies

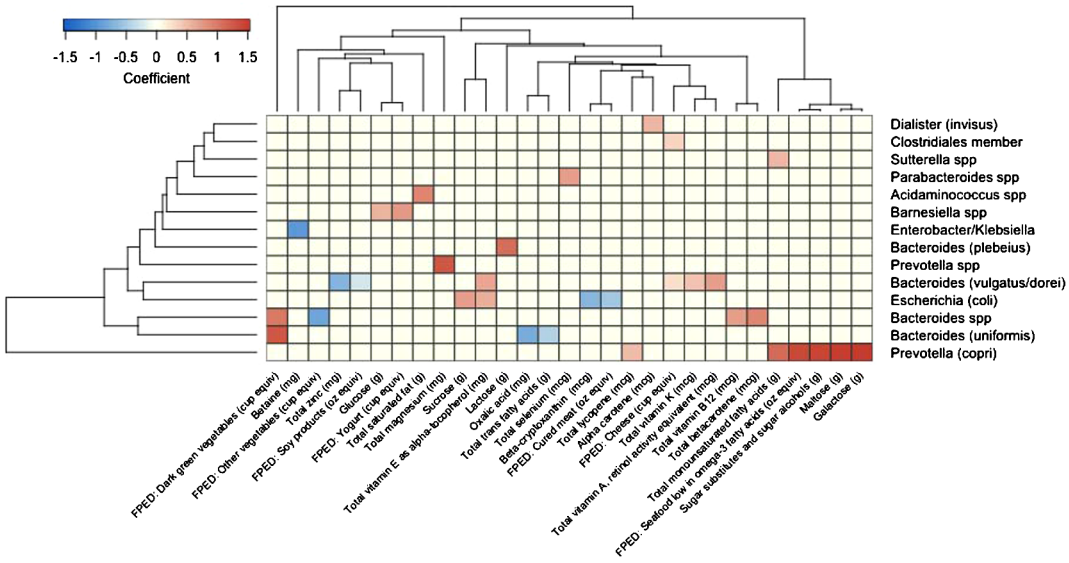

FIG. 5. *Estimated regression coefficients for corresponding dietary factors and fecal microbial abundance associations, grouped by BLAST assignment and identified using the proposed joint model.*

(Goodson et al. (2009), Haffajee and Socransky (2009), Zeigler et al. (2012)). In particular, Yang et al. (2019) found increased oral Gemella and Streptococcus oligofermentans among obese persons in a large ($n > 1{,}500$) cohort study. Gemellaceae and Streptococcaceae were also more abundant in obese subjects whose saliva suppressed aromatic compounds from wine, and the authors note that altered sensory responses may result in greater food intake (Piombino et al. (2014)). While Gemella was linked to both sucrose and cholesterol intake in our study, Streptococcus was negatively associated with key Western-diet components, namely, starch, animal protein and total saturated fat. This is likely explained by species-level variation which cannot be definitely determined by 16S sequencing, but it is important to note that Streptococcus members are the most abundant bacteria of the mouth (Huttenhower et al. (2012)). Taken overall, the current evidence suggests that the microbiome may be a reflection of obesity (or leanness) as well as a cause of it, largely, via diet-microbiome interactions (Komaroff (2017), Ridaura et al. (2013)).

While there are no methods available for direct comparison to our joint model, we compared the results of our analysis to two, two-step approaches that first select dietary covariates associated with fecal and oral multivariate count data and then perform variable selection on balances, constructed using estimated count probabilities for prediction of BMI. In the first step we used a recently proposed Bayesian variable selection method for Dirichlet-multinomial regression models (DM-BVS) (Wadsworth et al. (2017)) and a penalized approach introduced by Chen and Li (2013) (CL). For the CL approach the group penalty was set to 20%, and the model with the lowest Bayesian information criterion was used for inference (Schwarz (1978)). In the second step we fit a multiple linear regression model and performed variable selection on the balances calculated using the estimated $\boldsymbol{\psi}$ obtained in step one. The method for obtaining estimates of $\boldsymbol{\psi}$ differed across models, as explained in Section 4. We applied Bayesian variable selection for the DM-BVS approach and the lasso for the CL approach (George and McCulloch (1997), Tibshirani (1996)). We refer to the Bayesian and penalized two-step approaches as DMLM-Bayes and DMLM-Pen, respectively. Both methods were compared in regards to their selection of covariates as well as their model fit.

For the oral microbial data the DMLM-Pen and DMLM-Bayes approaches selected 61 and 45 covariate-taxon relations, respectively (see Supplementary Material Tables S3 and S4 (Koslovsky et al. (2020))). Using the DMLM-Pen approach, only two relations were also found using our joint model. However, using the DMLM-Bayes approach, seven relations were also found using our method. For the fecal microbial data, the DMLM-Pen and DMLM-Bayes approaches selected 15 and 23 covariate-taxon relations, respectively (see Supplementary Tables S5 and S6 (Koslovsky et al. (2020))). Similarly, using the DMLM-Pen approach, only three relations were also found using our joint model. However, using the DMLM-Bayes approach, 11 relations were also found using our method. Additionally, to assess model fit the mean squared error (MSE) for the joint model applied to the fecal data was 11.45, compared to 2874.34 and 27.98 for the DMLM-Pen and DMLM-Bayes approaches, respectively. Similarly, the MSE for the oral data was 90.94 with the joint model and 2993.28 and 126.16 with the DMLM-Pen and DMLM-Bayes approaches, respectively. While all models fit the fecal data better, our joint model demonstrated superior model fit for both the fecal and oral data.

We performed a sensitivity analysis to assess the sensitivity of the results produced by the joint model to prior specification. Specifically, we investigated differences in the selection and prediction results with $a = a_m = 1$ and $b = b_m = \{1, 4, 9\}$ as well as $a_0 = 2$ and $b_0 = \{2, 4, 16, 256\}$ for both the fecal and oral models separately. As expected, the number of covariates and balances selected in the model increased as the prior probability of inclusion increased. Similar to the sensitivity analysis on simulated data (Section 4), prediction performance diminished as b_0 increased.

4. Simulation study. In this section we evaluate the selection performance and predictive ability of our proposed joint model using simulated data. Performance is compared to the two two-step approaches presented in the case study. The method for obtaining estimates of $\boldsymbol{\psi}$ differs across models, as explained below.

We simulated $N = 50$ subjects with $P = 50$ covariates and $J = 150$ compositional taxa. Covariates $\boldsymbol{x}$ were simulated from a $N_P(\mathbf{0}, \Sigma)$, where $\Sigma_{i,j} = \omega^{|i-j|}$ and $\omega = 0.4$. In each of the replicate datasets, we randomly selected 10 of the 7500 covariate-taxon combinations to be associated with the compositional data. Corresponding regression coefficients φ were randomly sampled from $\pm[0.75, 1.25]$. Intercept terms $\boldsymbol{\alpha}$ were simulated from a uniform$[-2.3, 2.3]$. The multivariate count data $\boldsymbol{Z}$ were sampled from a Multinomial$(\dot{z}_i, \psi_i^*)$, where $\dot{z}_i \sim$ uniform$[2500, 7500]$ and $\psi_i^* \sim$ Dirichlet$(\boldsymbol{\gamma}_i^*)$, where $\boldsymbol{\gamma}_i^* = (\gamma_{i,1}^*, \gamma_{i,2}^*, \ldots, \gamma_{i,J}^*)$. Each $\gamma_{i,j}^* = \frac{\gamma_{i,j}}{\sum_{j=1}^J \gamma_{i,j}} \frac{1-d}{d}$, $j = 1, \ldots, J$, where $\gamma_{i,j}$ was determined using equation (3) and d serves as an overdispersion parameter which was set at 0.01, similar to Chen and Li (2013), Wadsworth et al. (2017). Thus, the data generating model differs from our model assumptions. We used a pseudovalue of 6.67×10^{-5} to replace zero values of $\psi_{i,j}$, which corresponds to the maximum roundoff error, 0.5, divided by the maximum possible value of $\dot{z}_i$, 7500. This is done to prevent taking the log of zero when calculating balances. We then generated the response data as $y_i = \alpha_0 + \boldsymbol{B}^*(\psi_i^*)'\boldsymbol{\beta} + \epsilon_i$, where $\alpha_0 = 0$, $\boldsymbol{\beta}$ is a $J-1$-dimensional vector of regression coefficients, $\boldsymbol{B}^*(\psi_i^*)$ are the balances calculated using sequential binary separation and $\epsilon_i \sim N(0, 1)$. Of the $J-1$ regression coefficients, five were randomly sampled from $\pm[1.25, 1.75]$ and the rest were set equal to zero.

When running the MCMC, we set hyperparameters $h_{\alpha_0} = h_\beta = 1$ as well as $a = 1$, and $b = \{9, 99, 999\}$, representing a prior expectation of 10%, 1% and 0.1% of the total number of covariates included in the model. For balance selection, a_m and b_m were set similarly. Before analysis, $\boldsymbol{y}$ was mean-centered, and covariates and balances were standardized to mean zero and variance one. Note that, in our joint model, balances are standardized at each MCMC iteration since they are recalculated using the current iteration's $\boldsymbol{\psi}_i$. Simulations were run for 20,000 iterations and thinned to every 10th iteration. This resulted in 2000 iterations, of which the first 1000 iterations were treated as burn-in and the remaining 1000 used for inference. Each run was initiated with a random 1% of the 7500 covariate-taxon combinations' and 5% of the 149 balances' inclusion indicators active. Covariates and balances were determined to be associated with the compositional and response data, respectively, if their MPPI ≥ 0.50 (Barbieri and Berger (2004)). Results we report below were obtained by averaging over 30 replicated datasets.

For variable selection, all methods were assessed on the basis of sensitivity (1-false negative rate), specificity (1-false positive rate) and Matthew's correlation coefficient (MCC) (a measure of overall selection accuracy). These are defined as:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{FN} + \text{TP}},$$

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}},$$

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

where TN, TP, FN and FP represent the true negatives, true positives, false negatives and false positives, respectively. To assess prediction performance, we trained the models on the 50 samples used for variable selection and tested the models on an additional 50 samples generated similarly. Prediction accuracy was assessed with the predicted mean squared error (PMSE), defined as $\sum_{i=1}^{50}(\mathcal{Y}_i - \hat{\mathcal{Y}}_i)^2$, where $\mathcal{Y}_i$ is from the testing set and $\hat{\mathcal{Y}}_i$ is its predicted

value. To obtain predictions of the outcomes with our joint model, we followed equation (12). Model fit was assessed with mean squared error (MSE), defined as $\sum_{i=1}^{50}(Y_i - \hat{Y}_i)^2$, where Y_i is from the training set. To obtain an estimate of the outcomes, $\hat{Y}_i$, we followed the approach used to calculate the PMSE, replacing $\boldsymbol{\mathcal{Y}}$ and $\boldsymbol{B}(\ddot{\boldsymbol{\psi}})$, with $\boldsymbol{Y}$ and $\boldsymbol{B}(\boldsymbol{\psi}^s)_{\boldsymbol{\xi}^s}$, respectively. Estimates for the DMLM-Bayes approach were obtained similarly, with the exception that the average of the S MCMC samples of $\boldsymbol{\psi}$ from the first step were used to construct $\boldsymbol{B}(\boldsymbol{\psi})$ in the second. For the DMLM-Pen approach, the testing balances are estimated using a similar approach as above, replacing the average of the MCMC samples in equations (14) and (15) with the CL model estimates.

4.1. *Results*. Tables 1 and 2 report results for the proposed joint model (JM), the two-step Bayesian approach (DMLM-Bayes) and the two-step penalized approach (DMLM-Pen) in terms of sensitivity, specificity, MCC, MSE and PMSE averaged over 30 simulations with standard errors in parentheses. For the Bayesian models, results are assessed over various beta-binomial priors for a covariate's probability of inclusion. Note that JM and DMLM-Bayes have similar performance for covariate selection since the underlying models are the same. Thus, we only compare to the DMLM-Pen approach in Table 1. For the selection of covariates associated with the multivariate count data, both of the models showed high specificity (Table 1). Note that models may have selected unassociated terms and still obtained a specificity of 1.00 due to rounding. However, the JM outperformed the DMLM-Pen approach in terms of sensitivity and MCC. The JM with hyperparameters $a = 1$ and $b = 99$ performed the best overall. As expected, the number of covariates selected was reduced as the mean of the inclusion prior decreased for the Bayesian approach. The DMLM-Pen approach selected the most covariates on average, leading it to have the lowest MCC overall. For the selection of balances associated with the continuous response, we found similar results for all of the models in terms of specificity (Table 2). For the Bayesian methods the number of selected balances, as well as the sensitivity and MCC, went down as the prior probability of inclusion decreased. Overall, the DMLM-Bayes model with hyperparameters $a = 1$ and $b = 9$ performed the best, with a similar performance achieved by the JM. We observed the worst performance in terms of sensitivity, specificity and MCC for the DMLM-Pen approach, as a result of the poor estimation of $\boldsymbol{\psi}$ from the DM portion of the model. Trace plots of the log posterior showed good mixing, and no observed trends in the plots after burn-in suggested model convergence across the simulations. In simulation results not shown, all of the methods maintained extremely high specificity for the null model which contained no true relations between covariates and the compositional data. Also, we found that selection performance was not sensitive to replacement pseudovalues for $\psi_{i,j} = 0$ during sampling.

TABLE 1

Covariate selection simulation results for the proposed joint model (JM) and two-step penalized DMLM-Pen approach in terms of sensitivity (Sens.), specificity (Spec.) and Matthew's correlation coefficient (MCC) averaged over 30 simulations with standard deviations in parentheses. For Bayesian models, results are assessed over various beta-binomial priors for covariates' probability of inclusion, $\boldsymbol{\zeta}$

		Covariates			
	Selection prior	Selected	Sens.	Spec.	MCC
JM	$a = 1, b = 9$	12.93 (4.15)	0.83 (0.16)	1.00 (0.00)	0.74 (0.13)
	$a = 1, b = 99$	7.47 (2.01)	0.71 (0.18)	1.00 (0.00)	0.82 (0.12)
	$a = 1, b = 999$	6.20 (2.34)	0.61 (0.24)	1.00 (0.00)	0.78 (0.16)
DMLM-Pen	–	25.00 (21.67)	0.30 (0.19)	1.00 (0.00)	0.20 (0.06)

TABLE 2

Balance selection simulation results for the proposed joint model (JM), the two-step Bayesian approach (DMLM-Bayes) and the two-step penalized approach (DMLM-Pen) in terms of sensitivity (Sens.), specificity (Spec.) and Matthew's correlation coefficient (MCC) averaged over 30 *simulations with standard deviations in parentheses. For Bayesian models, results are assessed over various beta-binomial priors for balances' probability of inclusion,* $\boldsymbol{\xi}$

		Balances			
	Selection prior	Selected	Sens.	Spec.	MCC
JM	$a = 1, b = 9$	9.30 (0.99)	0.92 (0.09)	1.00 (0.00)	0.94 (0.06)
	$a = 1, b = 99$	3.87 (2.61)	0.38 (0.27)	1.00 (0.00)	0.55 (0.23)
	$a = 1, b = 999$	0.80 (1.10)	0.08 (0.11)	1.00 (0.00)	0.38 (0.11)
	$a = 1, b = 9$	10.33 (0.76)	0.97 (0.06)	1.00 (0.01)	0.95 (0.07)
DMLM-Bayes	$a = 1, b = 99$	6.87 (3.22)	0.87 (0.34)	1.00 (0.00)	0.79 (0.25)
	$a = 1, b = 999$	1.23 (1.52)	0.12 (0.15)	1.00 (0.00)	0.42 (0.15)
DMLM-Pen	–	1.17 (1.46)	0.04 (0.09)	0.99 (0.01)	0.11 (0.19)

In terms of model fit, the DMLM-Bayes two-step approach with hyperparameters $a = 1$ and $b = 9$ had the smallest MSE on average, as expected given its balance selection performance (Table 3). For both Bayesian approaches the average MSE increased with more informative priors. This is mainly due to diminished sensitivity for both covariates and balances. Our joint model with weakly-informative priors had the lowest PMSE on average, closely followed by the DMLM-Bayes approach with similar prior specification. Despite its improved prediction performance, the JM had relatively higher PMSE standard deviations compared to the DMLM-Bayes approach, as hypothesized. The DMLM-Pen approach had the largest MSE and PMSE overall, reflecting its relatively poor selection performance for both covariates and balances.

4.2. *Sensitivity analysis*. We investigated the model sensitivity to specification of hyperparameters b_0, a, and b. In each of the sensitivity analyses, replicate data generated from the model defined in the simulation section were used. We evaluated the number of covariates selected, sensitivity, specificity, MCC, MSE and PMSE for the scale parameter in the Inverse-gamma prior for the error variance, b_0, at values in the set $\{1, 2, 4, 8\}$ (on $\log_2$ scale), holding $a_0 = 2$. With this parameterization, b_0 can interpreted as the expectation of σ^2. Additionally,

TABLE 3

Simulation results for the proposed joint model (JM), the two-step Bayesian approach (DMLM-Bayes) and the two-step penalized approach (DMLM-Pen) in terms of mean squared error (MSE) and prediction mean squared error (PMSE) averaged over 30 *simulations with standard deviation in parentheses. For Bayesian models, results are assessed over various beta-binomial priors for a covariate's probability of inclusion*

	Selection prior	MSE	PMSE
JM	$a = 1, b = 9$	101.28 (31.74)	563.54 (226.62)
	$a = 1, b = 99$	1250.13 (734.24)	1953.09 (1019.78)
	$a = 1, b = 999$	3214.91 (974.96)	3494.93 (913.75)
	$a = 1, b = 9$	67.22 (16.93)	785.63 (327.73)
DMLM-Bayes	$a = 1, b = 99$	509.03 (580.00)	2527.30 (1220.18)
	$a = 1, b = 999$	2710.42 (1134.56)	3562.42 (819.85)
DMLM-Pen	–	3521.97 (863.37)	4267.24 (1206.15)

TABLE 4

Results of sensitivity analysis for hyperparameter b_0 in Inverse-gamma prior for total number of selected covariates across taxa and balances (#), sensitivity (Sens.), specificity (Spec.), Matthew's correlation coefficient (MCC) and mean squared error (MSE)

		Covariates				Balances					
b	b_0	#	Sens.	Spec.	MCC	#	Sens.	Spec.	MCC	MSE	PMSE
9	1	10	0.90	1.00	0.90	8	0.80	1.00	0.89	112.76	619.39
	2	10	0.90	1.00	0.90	8	0.80	1.00	0.89	111.84	616.63
	4	10	0.90	1.00	0.90	9	0.80	0.99	0.83	117.54	680.12
	8	10	0.90	1.00	0.90	5	0.50	1.00	0.69	510.74	1063.09
99	1	10	1.00	1.00	1.00	7	0.70	1.00	0.83	340.21	691.27
	2	10	1.00	1.00	1.00	7	0.70	1.00	0.83	340.25	694.94
	4	10	1.00	1.00	1.00	6	0.60	1.00	0.76	428.12	768.15
	8	10	1.00	1.00	1.00	1	0.10	1.00	0.31	2073.22	2125.34
999	1	9	0.80	1.00	0.84	1	0.10	1.00	0.31	2646.49	2427.75
	2	9	0.80	1.00	0.84	1	0.10	1.00	0.31	2646.49	2427.75
	4	9	0.80	1.00	0.84	1	0.10	1.00	0.31	2882.92	2568.60
	8	9	0.80	1.00	0.84	1	0.10	1.00	0.31	3156.92	2735.28

we assessed the model's sensitivity to different beta-binomial priors for the inclusion indicators. Specifically, we used a weakly ($a = 1, b = 9$), moderately ($a = 1, b = 99$) and highly ($a = 1, b = 999$) informative prior, with $E[\zeta_{jp}] = 0.1, 0.01$ and 0.001, respectively.

To assess the sensitivity of the model to the specification of the Inverse-gamma prior for the random error σ^2, we set $a_0 = 2$ and fit the model across a range of b_0. The results of our sensitivity analysis are presented in Table 4. As expected, selection performance for the covariates associated with taxa probabilities were unaffected by b_0. However, we observed a negative relation between b_0 and the number of balances selected as well as balance sensitivity, specificity and MCC. As a result, we observed a positive relation between b_0 and MSE/PMSE. Additionally, the number of selected covariates and balances decreased with the expected prior probability of inclusion.

5. Discussion. In this work we have presented a Bayesian model for jointly identifying dietary covariates that are associated with microbial data and predicting a continuous, phenotypic response using a set of balances constructed from the estimated compositional taxa probabilities. Our approach induces sparsity on both balances and covariates while incorporating the structure of the multivariate count data. In our application we found numerous associations between microbial taxa and dietary factors that may lead to a microbiome that is generally more hospitable to the development of chronic diseases, such as obesity. Additionally, we observed similar prediction performance of BMI for fecal and oral microbiome data. Through simulation we have demonstrated the benefits of jointly modeling these data in terms of covariate selection performance and prediction accuracy. Additionally, we show how the Bayesian two-step approach had lower prediction accuracy and may underestimate prediction uncertainty by treating the compositional count data as fixed. In clinical applications this may result in overconfident prediction estimates of the phenotypic response which may promote the implementation of ineffective treatments or intervention strategies. While designed to study microbial abundance data, our method can handle any research setting in which multivariate count data may mediate the relation between a set of risk factors and a continuous response. Thus, our proposed model is agnostic to the sequencing approach used to quantify microbial samples.

Our model provides an integrated analysis of the relations between behavioral, microbial and phenotypic measures collected on a cohort of healthy obese and lean individuals. Given the complexity of the model, full validation of clinical results requires the availability of data collected on dietary covariates, fecal and/or oral microbiome samples, BMI as well as potential confounders (i.e., age and sex). However, the conditional independence structure implied by the joint model allows researchers to validate key aspects separately. For example, the selected associations between individual dietary factors and microbial counts can be directly compared to other studies investigating these relations. In these settings reproducibility is primarily challenged by vast heterogeneity in microbial abundances found across individuals and populations (Falony et al. (2016), Huttenhower et al. (2012), Li et al. (2014), Takeshita et al. (2014)) as well as study design issues, including differences in food frequency questionnaires (Bowyer et al. (2018)). Another key aspect of our model is its ability to accommodate taxa heterogeneity when predicting phenotypic responses. While our case study was not large enough to justify out-of-sample validation, larger follow-up studies could assess predictive performance using the cross-validation approach described at the end of Section 2.1.

While our approach provides unique insights into the relation between modulating factors and phenotypic outcomes via microbial composition samples, it currently lacks the ability to accommodate repeated measures data collected in longitudinal studies. The ability to model both fixed and random effects would allow researchers to investigate how the relations between diet, microbiome and BMI vary over time and across subjects. Additionally, structural information on phylogenetic trees could be incorporated into the multinomial distribution used to model the relation between covariates and the multivariate count data using a Dirichlet-tree multinomial model which permits both positive and negative correlation structures among the count data (Tang, Ma and Nicolae (2018), Wang and Zhao (2017a)). Also, our approach is developed for exploratory data analysis settings designed to generate hypotheses regarding the relations among covariates, compositional data and a response. In more confirmatory settings researchers may aim to assess treatment effects on microbial composition as well as the phenotypic response, while controlling for a set of possible confounders. Oftentimes, the appropriate subset of confounders to control for may be unknown, and the space to search through is large compared to the number of observations. In this setting our approach could be extended to search the pool of potential confounders in human microbiome studies following the methods proposed in Antonelli, Parmigiani and Dominici (2019). In this analysis we construct balances using binary sequential separation and focus our inference on prediction, not explanation. Future studies could incorporate biological information when constructing balances, similar to Morton et al. (2017), Silverman et al. (2017), Washburne et al. (2017), and, additionally, investigate the relations between balances and phenotypic responses. Lastly, our approach is presented for continuous outcomes, but discrete as well as survival outcomes are often encountered in biomedical settings. To handle discrete phenotypic outcomes, such as disease onset, the joint model could easily be adjusted using data augmentation approaches (Albert and Chib (1993), Polson, Scott and Windle (2013)).

Acknowledgments. Matthew Koslovsky is supported by NSF via the Research Training Group award DMS-1547433. Data utilized from the diet and microbiome study was supported by a grant to Carrie Daniel from The University of Texas MD Anderson Cancer Center Duncan Family Institute for Cancer Prevention and Risk Assessment. Carrie Daniel's efforts on this project are further supported by the NIH/NCI Cancer Center Support Grant P30 CA016672.

SUPPLEMENTARY MATERIAL

Supplemental code and tutorial (DOI: 10.1214/20-AOAS1354SUPPA; .zip). To help researchers use our approach, we provide R code and an accompanying tutorial applying our approach to simulated data. To enhance the performance of our approach, we integrated C++ into our source code using Rcpp and RcppArmadillo (Eddelbuettel and Sanderson (2014), Eddelbuettel et al. (2011)). The code developed for this manuscript, simulated data, and a worked example are publicly available on GitHub: https://github.com/mkoslovsky/DMLMbvs

Supplemental simulations and results (DOI: 10.1214/20-AOAS1354SUPPB; .pdf). In this document, we provide an additional simulation study demonstrating the model's invariance to balance specification and how balance sparsity can improve prediction performance, as well as additional tables and figures containing results of our case study analysis.

REFERENCES

AGUS, A., DENIZOT, J., THEVENOT, J., MARTINEZ-MEDINA, M., MASSIER, S., SAUVANET, P., BERNALIER-DONADILLE, A., DENIS, S., HOFMAN, P. et al. (2016). Western diet induces a shift in microbiota composition enhancing susceptibility to adherent-invasive *E. coli* infection and intestinal inflammation. *Sci. Rep.* **6** Art. ID 19032.

AITCHISON, J. (1986). *The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability*. CRC Press, London. MR0865647 https://doi.org/10.1007/978-94-009-4109-0

ALBERT, J. H. and CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88** 669–679. MR1224394

ANTONELLI, J., PARMIGIANI, G. and DOMINICI, F. (2019). High-dimensional confounding adjustment using continuous spike and slab priors. *Bayesian Anal.* **14** 825–848. MR3960772 https://doi.org/10.1214/18-BA1131

ARGIENTO, R., BIANCHINI, I. and GUGLIELMI, A. (2015). A priori truncation method for posterior sampling from homogeneous normalized completely random measure mixture models. Preprint. Available at arXiv:1507.04528.

BARBIERI, M. M. and BERGER, J. O. (2004). Optimal predictive model selection. *Ann. Statist.* **32** 870–897. MR2065192 https://doi.org/10.1214/009053604000000238

BOWYER, R. C. E., JACKSON, M. A., PALLISTER, T., SKINNER, J., SPECTOR, T. D., WELCH, A. A. and STEVES, C. J. (2018). Use of dietary indices to control for diet in human gut microbiota studies. *Microbiome* **6** Art. ID 77. https://doi.org/10.1186/s40168-018-0455-y

BROWN, P. J., VANNUCCI, M. and FEARN, T. (1998). Multivariate Bayesian variable selection and prediction. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 627–641. MR1626005 https://doi.org/10.1111/1467-9868.00144

BRUNO, F., GRECO, F. and VENTRUCCI, M. (2016). Non-parametric regression on compositional covariates using Bayesian P-splines. *Stat. Methods Appl.* **25** 75–88. MR3460487 https://doi.org/10.1007/s10260-015-0339-2

CARROLL, R. J., RUPPERT, D., STEFANSKI, L. A. and CRAINICEANU, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*, 2nd ed. *Monographs on Statistics and Applied Probability* **105**. CRC Press/CRC, Boca Raton, FL. MR2243417 https://doi.org/10.1201/9781420010138

CHATFIELD, C. (1995). Model uncertainty, data mining and statistical inference. *J. R. Stat. Soc., Ser. A, Stat. Soc.* **158** 419–466.

CHEN, J. and LI, H. (2013). Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *Ann. Appl. Stat.* **7** 418–442. MR3086425 https://doi.org/10.1214/12-AOAS592

CHEN, J., ZHANG, X. and LI, S. (2017). Multiple linear regression with compositional response and covariates. *J. Appl. Stat.* **44** 2270–2285. MR3670303 https://doi.org/10.1080/02664763.2016.1157145

DAO, M. C., EVERARD, A., ARON-WISNEWSKY, J., SOKOLOVSKA, N., PRIFTI, E., VERGER, E. O., KAYSER, B. D., LEVENEZ, F., CHILLOUX, J. et al. (2016). *Akkermansia muciniphila* and improved metabolic health during a dietary intervention in obesity: Relationship with gut microbiome richness and ecology. *Gut* **65** 426–436.

DAVID, L. A., MAURICE, C. F., CARMODY, R. N., GOOTENBERG, D. B., BUTTON, J. E., WOLFE, B. E., LING, A. V., DEVLIN, A. S., VARMA, Y. et al. (2014). Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505** 559–563.

EDDELBUETTEL, D. and SANDERSON, C. (2014). RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Comput. Statist. Data Anal.* **71** 1054–1063. MR3132026 https://doi.org/10.1016/j.csda.2013.02.005

EDDELBUETTEL, D., FRANÇOIS, R., ALLAIRE, J., USHEY, K., KOU, Q., RUSSEL, N., CHAMBERS, J. and BATES, D. (2011). Rcpp: Seamless R and C++ integration. *J. Stat. Softw.* **40** Issue 8.

EDGAR, R. C. (2013). UPARSE: Highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods* **10** 996–998. https://doi.org/10.1038/nmeth.2604

EGOZCUE, J. J. and PAWLOWSKY-GLAHN, V. (2005). Groups of parts and their balances in compositional data analysis. *Math. Geol.* **37** 795–828. MR2183639 https://doi.org/10.1007/s11004-005-7381-9

GLOOR, G. B., MACKLAIM, J. M., PAWLOWSKY-GLAHN, V. and EGOZCUE, J. J. (2017). Microbiome datasets are compositional: And this is not optional. *Front. Microbiol.* **8** Art. ID 2224.

EGOZCUE, J. J., PAWLOWSKY-GLAHN, V., MATEU-FIGUERAS, G. and BARCELÓ-VIDAL, C. (2003). Isometric logratio transformations for compositional data analysis. *Math. Geol.* **35** 279–300. MR1986165 https://doi.org/10.1023/A:1023818214614

FALONY, G., JOOSSENS, M., VIEIRA-SILVA, S., WANG, J., DARZI, Y., FAUST, K., KURILSHIKOV, A., BONDER, M. J., VALLES-COLOMER, M. et al. (2016). Population-level analysis of gut microbiome variation. *Science* **352** 560–564.

FAN, X., PETERS, B. A., JACOBS, E. J., GAPSTUR, S. M., PURDUE, M. P., FREEDMAN, N. D., ALEKSEYENKO, A. V., WU, J., YANG, L. et al. (2018). Drinking alcohol is associated with variation in the human oral microbiome in a large study of American adults. *Microbiome* **6** Art. ID 59.

FIŠEROVÁ, E. and HRON, K. (2011). On the interpretation of orthonormal coordinates for compositional data. *Math. Geosci.* **43** Art. ID 455.

GARCIA, T. P., MÜLLER, S., CARROLL, R. J. and WALZEM, R. L. (2013). Identification of important regressor groups, subgroups and individuals via regularization methods: Application to gut microbiome data. *Bioinformatics* **30** 831–837.

GELFAND, A. E. (1996). Model determination using sampling-based methods. In *Markov Chain Monte Carlo in Practice. Interdiscip. Statist.* 145–161. CRC Press, London. MR1397969

GEORGE, E. I. and MCCULLOCH, R. E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* 339–373.

GOODSON, J. M., GROPPO, D., HALEM, S. and CARPINO, E. (2009). Is obesity an oral bacterial disease? *J. Dent. Res.* **88** 519–523.

GOPALAKRISHNAN, V., SPENCER, C. N., NEZI, L., REUBEN, A., ANDREWS, M. C., KARPINETS, T. V., PRIETO, P. A., VICENTE, D., HOFFMAN, K. et al. (2018). Gut microbiome modulates response to anti-PD-1 immunotherapy in melanoma patients. *Science* **359** 97–103.

GURRY, T., GIBBONS, S. M., KEARNEY, S. M., ANANTHAKRISHNAN, A., JIANG, X., DUVALLET, C., KASSAM, Z., ALM, E. J. et al. (2018). Predictability and persistence of prebiotic dietary supplementation in a healthy human cohort. *Sci. Rep.* **8** Art. ID 12699.

HAFFAJEE, A. D. and SOCRANSKY, S. S. (2009). Relation of body mass index, periodontitis and *Tannerella forsythia*. *J. Clin. Periodontol.* **36** 89–99. https://doi.org/10.1111/j.1600-051X.2008.01356.x

HANSEN, T. H., KERN, T., BAK, E. G., KASHANI, A., ALLIN, K. H., NIELSEN, T., HANSEN, T. and PEDERSEN, O. (2018). Impact of a vegan diet on the human salivary microbiota. *Sci. Rep.* **8** Art. ID 5847. https://doi.org/10.1038/s41598-018-24207-3

ERCOLINI, D., FRANCAVILLA, R., VANNINI, L., DE FILIPPIS, F., CAPRIATI, T., DI CAGNO, R., IACONO, G., DE ANGELIS, M. and GOBBETTI, M. (2015). From an imbalance to a new imbalance: Italian-style gluten-free diet alters the salivary microbiota and metabolome of African celiac children. *Sci. Rep.* **5** Art. ID 18571.

HOFFMAN, K. L., HUTCHINSON, D. S., FOWLER, J., SMITH, D. P., AJAMI, N. J., ZHAO, H., SCHEET, P., CHOW, W.-H., PETROSINO, J. F. et al. (2018). Oral microbiota reveals signs of acculturation in Mexican American women. *PLoS ONE* **13** Art. ID e0194100.

HRON, K., FILZMOSER, P. and THOMPSON, K. (2012). Linear regression with compositional explanatory variables. *J. Appl. Stat.* **39** 1115–1128. MR2914515 https://doi.org/10.1080/02664763.2011.644268

HUTTENHOWER, C., GEVERS, D., KNIGHT, R., ABUBUCKER, S., BADGER, J. H., CHINWALLA, A. T., CREASY, H. H., EARL, A. M., FITZGERALD, M. G. et al. (2012). Structure, function and diversity of the healthy human microbiome. *Nature* **486** 207–214.

JAMES, L. F., LIJOI, A. and PRÜNSTER, I. (2009). Posterior analysis for normalized random measures with independent increments. *Scand. J. Stat.* **36** 76–97. MR2508332 https://doi.org/10.1111/j.1467-9469.2008.00609.x

KATO, I., VASQUEZ, A., MOYERBRAILEAN, G., LAND, S., DJURIC, Z., SUN, J., LIN, H.-S. and RAM, J. L. (2017). Nutritional correlates of human oral microbiome. *J. Am. Coll. Nutr.* **36** 88–98.

KNIGHTS, D., PARFREY, L. W., ZANEVELD, J., LOZUPONE, C. and KNIGHT, R. (2011). Human-associated microbial signatures: Examining their predictive value. *Cell Host Microbe* **10** 292–296.

KOMAROFF, A. L. (2017). The microbiome and risk for obesity and diabetes. *J. Am. Med. Assoc.* **317** 355–356.

KOSLOVSKY, M. D., HOFFMAN, K. L., DANIEL, C. R. and VANNUCCI, M. (2020). Supplement to "A Bayesian model of microbiome data for simultaneous identification of covariate associations and prediction of phenotypic outcomes." https://doi.org/10.1214/20-AOAS1354SUPPA, https://doi.org/10.1214/20-AOAS1354SUPPB

KOVATCHEVA-DATCHARY, P., NILSSON, A., AKRAMI, R., LEE, Y. S., DE VADDER, F., ARORA, T., HALLEN, A., MARTENS, E., BJÖRCK, I. et al. (2015). Dietary fiber-induced improvement in glucose metabolism is associated with increased abundance of *Prevotella*. *Cell Metab.* **22** 971–982.

LA ROSA, P. S., BROOKS, J. P., DEYCH, E., BOONE, E. L., EDWARDS, D. J., WANG, Q., SODERGREN, E., WEINSTOCK, G. and SHANNON, W. D. (2012). Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PLoS ONE* **7** Art. ID e52078.

LEITE, A. Z., RODRIGUES, N. D. C., GONZAGA, M. I., PAIOLO, J. C. C., DE SOUZA, C. A., STEFANUTTO, N. A. V., OMORI, W. P., PINHEIRO, D. G., BRISOTTI, J. L. et al. (2017). Detection of increased plasma interleukin-6 levels and prevalence of *Prevotella copri* and *Bacteroides vulgatus* in the feces of type 2 diabetes patients. *Front. Immunol.* **8** Art. ID 1107.

LI, H. (2015). Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annu. Rev. Stat. Appl.* **2** 73–94.

LI, J., QUINQUE, D., HORZ, H.-P., LI, M., RZHETSKAYA, M., RAFF, J. A., HAYES, M. G. and STONEKING, M. (2014). Comparative analysis of the human saliva microbiome from different climate zones: Alaska, Germany, and Africa. *BMC Microbiol.* **14** Art. ID 316.

LIN, W., SHI, P., FENG, R. and LI, H. (2014). Variable selection in regression with compositional covariates. *Biometrika* **101** 785–797. MR3286917 https://doi.org/10.1093/biomet/asu031

LIN, D., PETERS, B. A., FRIEDLANDER, C., FREIMAN, H. J., GOEDERT, J. J., SINHA, R., MILLER, G., BERNSTEIN, M. A., HAYES, R. B. et al. (2018). Association of dietary fibre intake and gut microbiota in adults. *Br. J. Nutr.* **120** 1014–1022.

MARCOBAL, A., BARBOZA, M., SONNENBURG, E. D., PUDLO, N., MARTENS, E. C., DESAI, P., LEBRILLA, C. B., WEIMER, B. C., MILLS, D. A. et al. (2011). Bacteroides in the infant gut consume milk oligosaccharides via mucus-utilization pathways. *Cell Host Microbe* **10** 507–514.

MARTIN-FERNANDEZ, J. A., BARCELÓ-VIDAL, C. and PAWLOWSKY-GLAHN, V. (2000). Zero replacement in compositional data sets. In *Data Analysis*, *Classification*, *and Related Methods* 155–160. Springer, Berlin.

MARTÍN-FERNÁNDEZ, J.-A., HRON, K., TEMPL, M., FILZMOSER, P. and PALAREA-ALBALADEJO, J. (2015). Bayesian-multiplicative treatment of count zeros in compositional data sets. *Stat. Model.* **15** 134–158. MR3325750 https://doi.org/10.1177/1471082X14535524

MARTINEZ-MEDINA, M., DENIZOT, J., DREUX, N., ROBIN, F., BILLARD, E., BONNET, R., DARFEUILLE-MICHAUD, A. and BARNICH, N. (2014). Western diet induces dysbiosis with increased *E. coli* in CEABAC10 mice, alters host barrier function favouring AIEC colonisation. *Gut* **63** 116–124.

MARUVADA, P., LEONE, V., KAPLAN, L. M. and CHANG, E. B. (2017). The human microbiome and obesity: Moving beyond associations. *Cell Host Microbe* **22** 589–599.

MERT, M. C., FILZMOSER, P., ENDEL, G. and WILBACHER, I. (2018). Compositional data analysis in epidemiology. *Stat. Methods Med. Res.* **27** 1878–1891. MR3803272 https://doi.org/10.1177/0962280216671536

MILLEN, A. E., MIDTHUNE, D., THOMPSON, F. E., KIPNIS, V. and SUBAR, A. F. (2005). The National Cancer Institute diet history questionnaire: Validation of pyramid food servings. *Am. J. Epidemiol.* **163** 279–288.

MORTON, J. T., SANDERS, J., QUINN, R. A., MCDONALD, D., GONZALEZ, A., VÁZQUEZ-BAEZA, Y., NAVAS-MOLINA, J. A., SONG, S. J., METCALF, J. L. et al. (2017). Balance trees reveal microbial niche differentiation. *mSystems* **2** Art. ID e00162-16.

NEWTON, M. A., NOUEIRY, A., SARKAR, D. and AHLQUIST, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5** 155–176.

PETERS, B. A., MCCULLOUGH, M. L., PURDUE, M. P., FREEDMAN, N. D., UM, C. Y., GAPSTUR, S. M., HAYES, R. B. and AHN, J. (2018). Association of coffee and tea intake with the oral microbiome: Results from a large cross-sectional study. *Cancer Epidemiol. Biomark. Prev.* **27** 814–821.

PINTO, J. R., EGOZCUE, J. J., PAWLOWSKY-GLAHN, V., PAREDES, R., NOGUERA-JULIAN, M. and CALLE, M. L. (2017). Balances: A new perspective for microbiome analysis. bioRxiv 219386. https://doi.org/10.1101/219386

PIOMBINO, P., GENOVESE, A., ESPOSITO, S., MOIO, L., CUTOLO, P. P., CHAMBERY, A., SEVERINO, V., MONETA, E., SMITH, D. P. et al. (2014). Saliva from obese individuals suppresses the release of aroma compounds from wine. *PLoS ONE* **9** Art. ID e85611.

POLSON, N. G., SCOTT, J. G. and WINDLE, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *J. Amer. Statist. Assoc.* **108** 1339–1349. MR3174712 https://doi.org/10.1080/01621459.2013.829001

RICHARDSON, S., BOTTOLO, L. and ROSENTHAL, J. S. (2011). Bayesian models for sparse regression analysis of high dimensional data. In *Bayesian Statistics* 9 539–568. Oxford Univ. Press, Oxford. MR3204018 https://doi.org/10.1093/acprof:oso/9780199694587.003.0018

RIDAURA, V. K., FAITH, J. J., REY, F. E., CHENG, J., DUNCAN, A. E., KAU, A. L., GRIFFIN, N. W., LOMBARD, V., HENRISSAT, B. et al. (2013). Gut microbiota from twins discordant for obesity modulate metabolism in mice. *Science* **341** 1241214.

SAVITSKY, T., VANNUCCI, M. and SHA, N. (2011). Variable selection for nonparametric Gaussian process priors: Models and computational strategies. *Statist. Sci.* **26** 130–149. MR2849913 https://doi.org/10.1214/11-STS354

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

SHETTY, S. A., HUGENHOLTZ, F., LAHTI, L., SMIDT, H. and DE VOS, W. M. (2017). Intestinal microbiome landscaping: Insight in community assemblage and implications for microbial modulation strategies. *FEMS Microbiol. Rev.* **41** 182–199. https://doi.org/10.1093/femsre/fuw045

SHI, P., ZHANG, A. and LI, H. (2016). Regression analysis for microbiome compositional data. *Ann. Appl. Stat.* **10** 1019–1040. MR3528370 https://doi.org/10.1214/16-AOAS928

SILVERMAN, J. D., WASHBURNE, A. D., MUKHERJEE, S. and DAVID, L. A. (2017). A phylogenetic transform enhances analysis of compositional microbiota data. *eLife* **6** Art. ID e21887.

SONNENBURG, J. L. and BÄCKHED, F. (2016). Diet-microbiota interactions as moderators of human metabolism. *Nature* **535** 56–64. https://doi.org/10.1038/nature18846

SONNENBURG, E. D., ZHENG, H., JOGLEKAR, P., HIGGINBOTTOM, S. K., FIRBANK, S. J., BOLAM, D. N. and SONNENBURG, J. L. (2010). Specificity of polysaccharide use in intestinal bacteroides species determines diet-induced microbiota alterations. *Cell* **141** 1241–1252.

STINGO, F. C., CHEN, Y. A., VANNUCCI, M., BARRIER, M. and MIRKES, P. E. (2010). A Bayesian graphical modeling approach to microRNA regulatory network inference. *Ann. Appl. Stat.* **4** 2024–2048. MR2829945 https://doi.org/10.1214/10-AOAS360

SUBAR, A. F., THOMPSON, F. E., KIPNIS, V., MIDTHUNE, D., HURWITZ, P., MCNUTT, S., MCINTOSH, A. and ROSENFELD, S. (2001). Comparative validation of the Block, Willett, and National Cancer Institute food frequency questionnaires: The Eating at America's Table Study. *Am. J. Epidemiol.* **154** 1089–1099.

TADESSE, M. G., IBRAHIM, J. G., GENTLEMAN, R., CHIARETTI, S., RITZ, J. and FOA, R. (2005). Bayesian error-in-variable survival model for the analysis of GeneChip arrays. *Biometrics* **61** 488–497. MR2140921 https://doi.org/10.1111/j.1541-0420.2005.00313.x

TAKESHITA, T., MATSUO, K., FURUTA, M., SHIBATA, Y., FUKAMI, K., SHIMAZAKI, Y., AKIFUSA, S., HAN, D.-H., KIM, H.-D. et al. (2014). Distinct composition of the oral indigenous microbiota in South Korean and Japanese adults. *Sci. Rep.* **4** Art. ID 6990.

MAO, J., CHEN, Y. and MA, L. (2017). Bayesian graphical compositional regression for microbiome data. Preprint. Available at arXiv:1712.04723.

TANG, Y., MA, L. and NICOLAE, D. L. (2018). A phylogenetic scan test on a Dirichlet-tree multinomial model for microbiome data. *Ann. Appl. Stat.* **12** 1–26. MR3773384 https://doi.org/10.1214/17-AOAS1086

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

TURNBAUGH, P. J. (2017). Microbes and diet-induced obesity: Fast, cheap, and out of control. *Cell Host Microbe* **21** 278–281. https://doi.org/10.1016/j.chom.2017.02.021

VALDES, A. M., WALTER, J., SEGAL, E. and SPECTOR, T. D. (2018). Role of the gut microbiota in nutrition and health. *BMJ* **361** Art. ID k2179. https://doi.org/10.1136/bmj.k2179

VEHTARI, A., GELMAN, A. and GABRY, J. (2016). loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models. R package version 0.1.6.

VEHTARI, A., GELMAN, A. and GABRY, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* **27** 1413–1432. MR3647105 https://doi.org/10.1007/s11222-016-9696-4

VERSACE, F., KYPRIOTAKIS, G., BASEN-ENGQUIST, K. and SCHEMBRE, S. M. (2015). Heterogeneity in brain reactivity to pleasant and food cues: Evidence of sign-tracking in humans. *Soc. Cogn. Affect. Neurosci.* **11** 604–611.

WADSWORTH, W. D., ARGIENTO, R., GUINDANI, M., GALLOWAY-PENA, J., SHELBURNE, S. A. and VANNUCCI, M. (2017). An integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. *BMC Bioinform.* **18** Art. ID 94. https://doi.org/10.1186/s12859-017-1516-0

WANG, T. and ZHAO, H. (2017a). A Dirichlet-tree multinomial regression model for associating dietary nutrients with gut microorganisms. *Biometrics* **73** 792–801. MR3713113 https://doi.org/10.1111/biom.12654

WANG, T. and ZHAO, H. (2017b). Constructing predictive microbial signatures at multiple taxonomic levels. *J. Amer. Statist. Assoc.* **112** 1022–1031. MR3735357 https://doi.org/10.1080/01621459.2016.1270213

WASHBURNE, A. D., SILVERMAN, J. D., LEFF, J. W., BENNETT, D. J., DARCY, J. L., MUKHERJEE, S., FIERER, N. and DAVID, L. A. (2017). Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets. *PeerJ* **5** Art. ID e2969.

WILLETT, W. (1998). Implications of total energy intake for epidemiologic analyses. In *Nutritional Epidemiology* 273–301. Oxford Univ. Press, Oxford.

WU, G. D., CHEN, J., HOFFMANN, C., BITTINGER, K., CHEN, Y.-Y., KEILBAUGH, S. A., BEWTRA, M., KNIGHTS, D., WALTERS, W. A. et al. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334** 105–108.

XIA, Y. and SUN, J. (2017). Hypothesis testing and statistical analysis of microbiome. *Genes Dis*. **4** 138–148.

XU, Z. and KNIGHT, R. (2015). Dietary effects on human gut microbiome diversity. *Br. J. Nutr.* **113** S1–S5.

YANG, Y., CAI, Q., ZHENG, W., STEINWANDEL, M., BLOT, W. J., SHU, X.-O. and LONG, J. (2019). Oral microbiome and obesity in a large study of low-income and African-American populations. *J. Oral Microbiol*. **11** Art. ID 1650597.

ZEIGLER, C. C., PERSSON, G. R., WONDIMU, B., MARCUS, C., SOBKO, T. and MODÉER, T. (2012). Microbiota in the oral subgingival biofilm is associated with obesity in adolescence. *Obesity* **20** 157–164.

ZHANG, Z., SCHWARTZ, S., WAGNER, L. and MILLER, W. (2000). A greedy algorithm for aligning DNA sequences. *J. Comput. Biol*. **7** 203–214.

ZHANG, H., DIBAISE, J. K., ZUCCOLO, A., KUDRNA, D., BRAIDOTTI, M., YU, Y., PARAMESWARAN, P., CROWELL, M. D., WING, R. et al. (2009). Human gut microbiota in obesity and after gastric bypass. *Proc. Natl. Acad. Sci. USA* **106** 2365–2370.

ZHANG, Y., ZHOU, H., ZHOU, J. and SUN, W. (2017). Regression models for multivariate count data. *J. Comput. Graph. Statist*. **26** 1–13. MR3610402 https://doi.org/10.1080/10618600.2016.1154063