Received: 12 April 2021 | Revised: 28 February 2022 | Accepted: 7 March 2022

DOI: 10.1002/sim.9384

RESEARCH ARTICLE

Statistics in Medicine WILEY

Negative binomial factor regression with application to microbiome data analysis

Aditya K. Mishra[1] | **Christian L. Müller**[1,2,3]

[1]Center for Computational Mathematics, Flatiron Institute, Simons Foundation, New York, New York, USA

[2]Department of Statistics, LMU München, Munich, Germany

[3]Institute of Computational Biology, Helmholtz Zentrum München, Munich, Germany

Correspondence
Aditya K. Mishra, Center for Computational Mathematics, Flatiron Institute, Simons Foundation, New York, NY 10010, USA.
Email: amishra@flatironinstitute.org

Abstract

The human microbiome provides essential physiological functions and helps maintain host homeostasis via the formation of intricate ecological host-microbiome relationships. While it is well established that the lifestyle of the host, dietary preferences, demographic background, and health status can influence microbial community composition and dynamics, robust generalizable associations between specific host-associated factors and specific microbial taxa have remained largely elusive. Here, we propose factor regression models that allow the estimation of structured parsimonious associations between host-related features and amplicon-derived microbial taxa. To account for the overdispersed nature of the amplicon sequencing count data, we propose negative binomial reduced rank regression (NB-RRR) and negative binomial co-sparse factor regression (NB-FAR). While NB-RRR encodes the underlying dependency among the microbial abundances as outcomes and the host-associated features as predictors through a rank-constrained coefficient matrix, NB-FAR uses a sparse singular value decomposition of the coefficient matrix. The latter approach avoids the notoriously difficult joint parameter estimation by extracting sparse unit-rank components of the coefficient matrix sequentially, effectively delivering interpretable bi-clusters of taxa and host-associated factors. To solve the nonconvex optimization problems associated with these factor regression models, we present a novel iterative block-wise majorization procedure. Extensive simulation studies and an application to the microbial abundance data from the American Gut Project (AGP) demonstrate the efficacy of the proposed procedure. In the AGP data, we identify several factors that strongly link dietary habits and host life style to specific microbial families.

KEYWORDS

American Gut Project, microbiome, multivariate analysis, overdispersed count data, reduced rank regression, sparse singular value decomposition

1 INTRODUCTION

The human microbiome, the collection of microbes that reside on or within human tissues and fluids, has formed intricate ecological relationships with the host over the course of co-evolution.[1] Advances in next-generation amplicon sequencing technology and analysis techniques have enabled the direct identification of microbial species compositions and abundances in their natural habitat. These approaches have revealed considerable variability in both composition and diversity across different body sites[2] and allowed the estimation of potential associations between the microbiome and the underlying health condition of the subject.[3] For instance, differential abundances of the gut microbiome have been linked to medical conditions such as inflammatory bowel disease (IBD), irritable bowel syndrome (IBS), type 2 diabetes, obesity, and neurological disorders.[4] The gut microbiome also makes considerable contribution to metabolic functioning, for example, by breaking down specific food components.[5] Dietary changes can thus induce considerable shifts in gut microbial compositions.[6] Similar intricate relationships between the environment and the microbiome are also known in other ecosystems. For instance, the soil microbiome plays a significant role in the cycle of carbon and nitrogen fixation, thus having direct implications for plant growth.[7] The soil microbiome also shows large variability with respect to soil conditions such as pH, temperature, moisture, and spatial location. Likewise, cyanobacteria in the marine ecosystems contribute to a large extent to the ocean's primary productivity, yet exhibit considerable abundance variability across location, season, and water conditions.[8,9]

Amplicon-based microbiome survey data are derived from samples of the habitat of interest, for example, the human gut, where variable regions of the bacterial and archaeal 16S ribosomal RNA are experimentally extracted and sequenced. These marker gene sequences serve as a proxy to the underlying bacterial taxon abundances and are summarized in operational taxonomic units (OTUs) or amplicon sequence variants (ASVs).[10,11] Reference databases are used to identify the (approximate) taxonomy of the representative microbial sequences. Bioinformatic workflows and databases, such as, for example, QIIME-2[10] or the Qiita framework,[12] allow standardized processing of and access to these OTU/species counts. In addition, large-scale microbiome survey studies such as the American Gut Project (AGP),[13] the Human Microbiome Project (HMP),[14] and the Earth Microbiome Project[15] also collect large/high-dimensional host- or environment-associated covariate data. These survey data reach a level of scale and completeness that, in principle, allows to make quantitative predictions about the relationship between host-associated factors and microbial abundance patterns. For example, the AGP data comprises hundreds of host-associated features, including variables indicating dietary intake, medical conditions, medication use, participants' demography, and life style.

Using the AGP data as representative microbiome data resource, we here introduce a statistical factor regression framework that allows the identification of key associations between host-related features and microbial taxa. While recent work[16] has already identified individual host factors that confound microbial abundance patterns in relation to specific disease phenotypes, we propose a general factor model that simultaneously takes into account all relevant host-associated covariates and links them to the observed microbial abundances, independent of a specific downstream task. Since the observed microbial abundances across all levels of taxonomic aggregation come in form of overdispersed count data, we base our model on the classical negative binomial (NB) distribution (see Figure 1). NB models are common place in genomics and microbiome data analysis. For example, the popular DESeq2 package,[17] used extensively in differential expression testing in bulk RNA-Seq data, uses the NB model as underlying model for total mRNA transcript abundance. Due to technical and experimental limitations, microbial count data, however, carry only relative information and show varied sequencing depth across samples. To mitigate these limitations, microbiome data require transformation/normalization approaches prior to statistical modeling.[18] Important examples include rarefying samples to a common sequencing depth or scaling using factors such as a cumulative sum, median, upper quartile, or the total sum, the latter of which leading to compositional or relative abundance data. A particularly popular approach for NB modeling of microbial count data is common sum scaling, as put forward by McMurdie and Holmes.[19] When modeling microbial count data on the OTU/ASV level, zero-inflated extensions of the NB model have been proposed to account for the excess number of zeros in the data[20] (see, eg, Reference 21 for a critical assessment).

Alternative generative statistical modeling approaches include the Dirichlet multinomial (mixture) framework,[22] latent Dirichlet allocation,[23,24] and Poisson distribution models[25] (including their respective zero-inflated extensions). Several models also allow the inclusion of host or environmental covariate data in generative modeling, including Poisson factor models,[26-28] latent Dirichlet allocation,[29] and Bayesian Dirichlet multinomial models.[30-32] Due to the abundance of excess zeros at the amplicon or species level, some of these models also include a zero-inflation modeling component. This and the high dimensionality of the data at the OTU/ASV level make both estimation and biological interpretability challenging.

Our modeling framework follows a different objective and focuses on accurate and parsimonious estimation of key host-associated factors that influence microbial taxa at *higher* taxonomic ranks. This approach facilitates crisp, biologically interpretable statements about the underlying potential role of host features on broad microbial abundance patterns at the expense of taxonomic resolution. When summarizing microbiome survey data on a higher taxonomic level, the data remains overdispersed, yet contains no excess zeros. This makes the addition of zero-inflated components in the model superfluous (see Figure S5 of the supplementary materials). Prior work[33] already established that NB regression (NB-GLM) is capable of relating individual taxa to the host-associated features (see Figure 1A). This marginal model, however, ignores the fact that some of the taxa are likely influenced by a set of common factors, for example, age, diet, or life style.

Here, we alleviate this shortcoming and introduce a NB factor regression framework that models microbial abundance data as outcome and covariates as predictors jointly while encoding the underlying dependencies in a parsimonious fashion. In the high-dimensional multivariate linear regression setting, it is common practice to model the underlying dependency for dependent outcomes via a structured coefficient matrix[34] (see Figure 1B,C). The model parameters are estimated by solving a regularized optimization problem. For example, reduced-rank regression[35-37] promotes information-sharing among response and predictors through a low-rank coefficient matrix. When covariates are high-dimensional, sparsity is known to facilitate identifiability and better model interpretation.[38] In the multivariate setting, this has been achieved via a sparse factorization of the model coefficient matrix.[34,39-41] When the outcome matrix comprises non-Gaussian or mixed type variables, for example, Bernoulli-type for binary outcomes and Poisson-type for counts, Luo et al[42] proposed *mixed-outcome reduced-rank regression*. Mishra et al[43] proposed *generalized co-sparse factor regression* (GO-FAR) to model the outcome jointly under sparsity constraints. As we will show in the remainder of the article, these existing models are inappropriate for microbial taxon data due to the overdispersed nature of the counts.

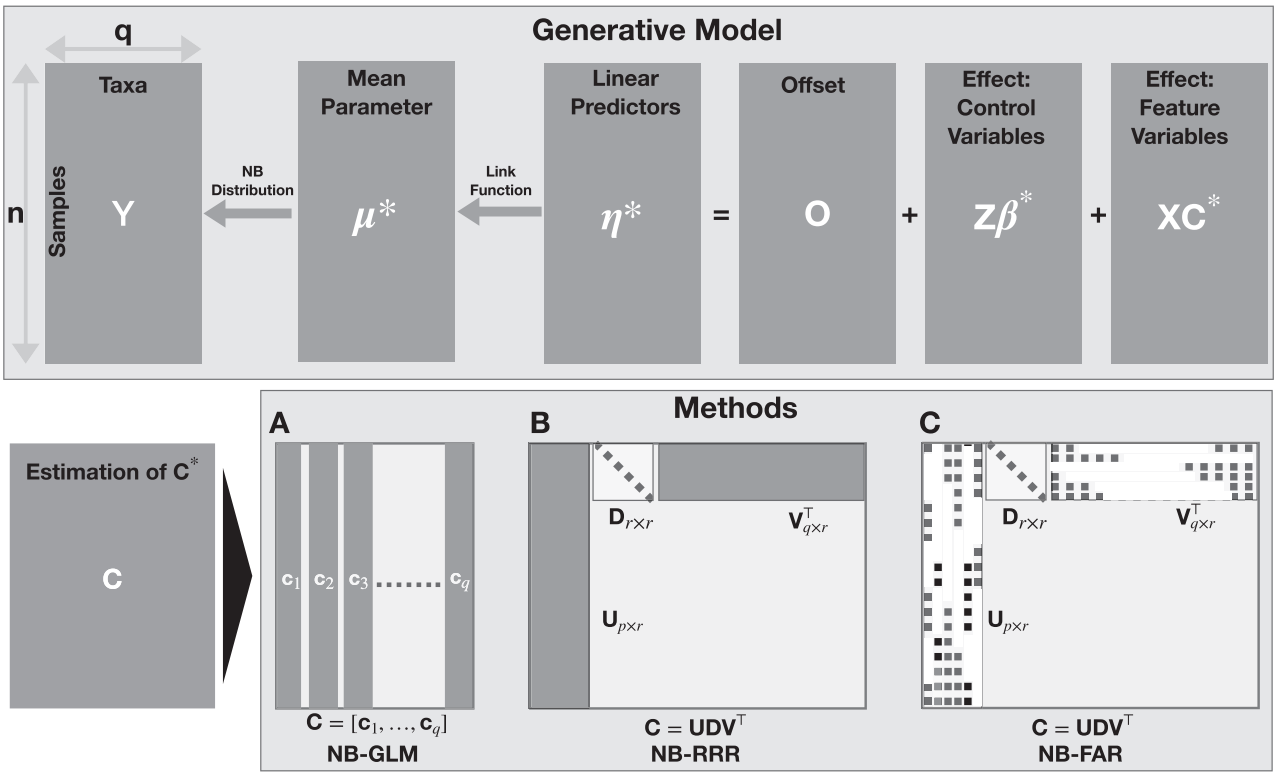

FIGURE 1 Regression model for the overdispersed microbial abundance data $\mathbf{Y}$ of count types in terms of the covariates $\mathbf{X}$ and the control variables $\mathbf{Z}$. The upper panel presents the true generative model with parameters $\{\boldsymbol{\beta}^*, \mathbf{C}^*\}$ and the lower panel presents the three approaches for estimating the parameters. (A) The generalized linear model of negative binomial regression (NB-GLM) estimates each of the columns of $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_q]$ separately; (B) negative binomial reduced rank regression (NB-RRR) jointly estimates a low-rank coefficient matrix as $\mathbf{C} = \mathbf{UDV}^\mathrm{T}$; (C) negative binomial co-sparse factor regression (NB-FAR) jointly estimates a low-rank coefficient matrix as $\mathbf{C} = \mathbf{UDV}^\mathrm{T}$ with sparse singular vectors $\{\mathbf{U}, \mathbf{V}\}$

Instead, we propose *negative binomial reduced rank regression* (NB-RRR) and *negative binomial co-sparse factor regression* (NB-FAR) to jointly model the microbial abundance data using the host-associated features as covariates (see Figure 1B,C for details). NB-RRR follows previous reduced rank regression frameworks by capturing the underlying dependencies among response and predictors via a low-rank coefficient matrix. NB-FAR extends the GO-FAR framework and encodes the underlying dependency via a sparse singular value decomposition (SSVD) of the coefficient matrix. Following the estimation strategy of GO-FAR, we extract unit-rank components of the coefficient matrix sequentially,[34] thus alleviating the challenging problem of joint estimation. There, each sequential step solves a co-sparse unit rank estimation problem with a suitably adjusted offset term that accounts for the effects of previous steps. NB-FAR thus models the associations of microbial abundance and host-associated features via a few latent factors that comprise only a subset of predictors. Both NB-RRR and NB-FAR estimation procedures are implemented, tested, validated, and made publicly available in the R package nbfar, available on GitHub at https://github.com/amishra-stats/nbfar.

The remainder of the article is organized as follows. Section 2 provides the details of the NB-RRR and NB-FAR framework. Section 3 provides the details of the parameter estimation procedure. In Section 4, we present simulation studies to demonstrate the efficacy of the estimation procedures. Section 5 provides a detailed analysis of the AGP taxon data on the family level and an extensive set of host-associated covariates using our NB factor regression methods. Section 6 discusses the findings and provides future research directions. Additional data analysis plots and the details of the estimation procedures are provided in the supplementary materials.

2 | FACTOR MODELS FOR MICROBIOME DATA

As motivating example we consider the data from the AGP[13] where samples from thousands of participants have been collected, sequenced, and processed to obtain microbial abundances. Each sample is associated with participant-specific covariates that are related to diet, heath, and lifestyle. Our overall goal is to understand the associations of these covariates with the observed microbial abundance patterns. Let us denote the abundance/count data of q taxa from n samples as $\mathbf{Y} = [y_{ik}]_{n\times q} = [\mathbf{y}_1, \ \dots \ \mathbf{y}_n]^{\mathrm{T}} \in \mathbb{N}_0^{n\times q}$, the associated predictors/covariates as $\mathbf{X} = [\mathbf{x}_1, \ \dots \ , \mathbf{x}_n]^{\mathrm{T}} \in \mathbb{R}^{n\times p}$, and the control variables as $\mathbf{Z} = [\mathbf{z}_1, \ \dots \ , \mathbf{z}_n]^{\mathrm{T}} \in \mathbb{R}^{n\times c}$. $\mathbf{Z}$ comprises variables, such as age and gender, that are held constant in an experiment and are thus fully adjusted for in the model.

The observed taxon abundance data are overdispersed, that is, the variance of the taxa tends to be considerably larger than their mean.[33] This fact motivates the use of a parametric framework based on the NB distribution to model the underlying associations between multivariate count outcome $\mathbf{Y}$ and factors $\{\mathbf{X}, \mathbf{Z}\}$. Using the alternative parameterization of the NB distribution,[44] the generative model for the abundance of the jth taxon in the ith sample is given by

$$p(y_{ij}; \mu_{ij}, \phi_j) = \mathrm{NB}(y_{ij}; \mu^*_{ij}, \phi^*_j) = \binom{y_{ij} + \phi^*_j - 1}{y_{ij}} \frac{\mu^{*y_{ij}}_{ij} \phi^{*\phi^*_j}_j}{(\mu^*_{ij} + \phi^*_j)^{y_{ij}+\phi^*_j}}, \tag{1}$$

where $\mu^*_{ij} > 0$ is the entry-specific mean parameter and $\phi^*_j \in \mathbb{R}^+$ is the taxon-specific shape parameter. Let us jointly represent the shape parameters of q taxa by $\mathbf{\Phi}^* = [\phi^*_1, \ \dots \ , \phi^*_q]$ and the entry-specific mean parameters by $\boldsymbol{\mu}^* = [\mu^*_{ij}]_{n\times q}$. For the generative model (1), $\mathbb{E}(y_{ij}) = \mu^*_{ij}$ and $\mathrm{var}(y_{ij}) = \mu^*_{ij} + \frac{\mu^{*2}_{ij}}{\phi^*_j}$, that is, $\mathrm{var}(y_{ij}) \geq \mathbb{E}(y_{ij})$, making the model suitable for overdispersed count data. Then, the joint negative log-likelihood is given by

$$\mathcal{L}(\boldsymbol{\mu}^*, \mathbf{\Phi}^*) = -\sum_{i=1}^{n}\sum_{k=1}^{q} \ell_k(\mu^*_{ik}, \phi^*_k), \tag{2}$$

where $\ell_k(\mu^*_{ik}, \phi^*_k) = \log p(y_{ik}; \mu^*_{ik}, \phi^*_k)$.

To associate the participant-specific covariates to the microbial abundance, we link entry-specific mean parameters $\boldsymbol{\mu}^*$ to the linear predictors as

$$g(\boldsymbol{\mu}^*) = \boldsymbol{\eta}^*(\mathbf{C}^*, \boldsymbol{\beta}^*, \mathbf{O}) = \mathbf{O} + \mathbf{Z}\boldsymbol{\beta}^* + \mathbf{X}\mathbf{C}^*, \tag{3}$$

where $g(\cdot)$ is a suitable link function that satisfies $\boldsymbol{\mu}^* > 0$, $\mathbf{O} = [o_{ik}]_{n\times q} \in \mathbb{R}^{n\times q}$ is a fixed offset term, $\mathbf{C}^* = [\mathbf{c}^*_1, \ \dots \ , \mathbf{c}^*_q] \in \mathbb{R}^{p\times q}$ is the coefficient matrix corresponding to the predictors $\mathbf{X}$, and $\boldsymbol{\beta} = [\boldsymbol{\beta}^*_1, \ \dots \ , \boldsymbol{\beta}^*_q] \in \mathbb{R}^{c\times q}$ is the coefficient matrix

corresponding to the control variables $\mathbf{Z}$. The formulation includes an intercept in the model by setting the first column of $\mathbf{Z}$ to be $\mathbf{1}_n$, the $n \times 1$ vector of ones. Following the work of Zeileis et al[44] and Anders and Huber,[45] we choose $g(x) = \log x$ as the link function so that any $\mu^*_{ij} > 0$. Depending on the problem, one may choose another link function that satisfies $\boldsymbol{\mu}^* > 0$. Unless otherwise stated, we write $\boldsymbol{\eta}^*(\mathbf{C}^*, \boldsymbol{\beta}^*, \mathbf{O})$ as $\boldsymbol{\eta}^*$.

With the shape parameter fixed, the NB distribution (1) belongs to the exponential dispersion family.[46] To utilize the form of this family, we define the corresponding natural parameter matrix $\boldsymbol{\Theta}^* = [\theta^*_{ij}]_{n\times q} \in \mathbb{R}^{n\times q}$ as

$$\boldsymbol{\Theta}^*(\mathbf{C}^*, \boldsymbol{\beta}^*, \mathbf{O}) = h(\boldsymbol{\mu}^*, \boldsymbol{\Phi}^*) = [h(\mu^*_{ij}, \phi^*_j)]_{n\times q} = h(g^{-1}(\boldsymbol{\eta}^*(\mathbf{C}^*, \boldsymbol{\beta}^*, \mathbf{O}), \boldsymbol{\Phi}^*), \tag{4}$$

where $h(\mu^*_{ij}, \phi^*_j) = \log \frac{\mu^*_{ij}}{\mu^*_{ij}+\phi^*_j}$. Again, unless otherwise stated, we will conveniently express $\boldsymbol{\Theta}^*(\mathbf{C}^*, \boldsymbol{\beta}^*, \mathbf{O})$ as $\boldsymbol{\Theta}^*$. We assume the outcomes to be conditionally independent given $\mathbf{X}$ and $\mathbf{Z}$. In terms of the natural parameter $\boldsymbol{\Theta}^*$, we rewrite the negative log-likelihood function (2) as

$$\mathcal{L}(\boldsymbol{\Theta}^*, \boldsymbol{\Phi}^*) = -\text{tr}(\mathbf{Y}^{\mathrm{T}}\boldsymbol{\Theta}^*) + \text{tr}(\mathbf{J}^{\mathrm{T}}\mathbf{B}(\boldsymbol{\Theta}^*)) + \sum_{i,\,j} \log \binom{y_{ij} + \phi^*_j - 1}{y_{ij}}, \tag{5}$$

where $\mathbf{J} = \mathbf{1}_{n\times q}$, $\text{tr}(\mathbf{A})$ is the *trace* of a square matrix $\mathbf{A}$ and $\mathbf{B}(\boldsymbol{\Theta}^*) = [b(\theta^*_{ij})]_{n\times q}$ such that $b(\theta^*_{ij}) = -\phi^*_j \log(1 - e^{\theta^*_{ij}})$. For fixed shape parameter $\boldsymbol{\Phi}^*$, it is straightforward to show that $\mathbf{E}(y_{ij}) = \mu^*_{ij} = b'(\theta^*_{ij})$. This results in linking $\boldsymbol{\Theta}^*$ to the linear predictor $\boldsymbol{\eta}^*$ via $g(b'(\theta^*_{ij}) = \eta^*_{ij}$.

To obtain an estimate of the model parameters, we minimize the objective function $\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\Phi})$ with respect to $\{\mathbf{C}, \boldsymbol{\beta}, \boldsymbol{\Phi}\}$, where $g(b'(\boldsymbol{\Theta}(\mathbf{C}, \boldsymbol{\beta}, \mathbf{O}))) = \boldsymbol{\eta}(\mathbf{C}, \boldsymbol{\beta}, \mathbf{O}) = \mathbf{O} + \mathbf{X}\mathbf{C} + \mathbf{Z}\boldsymbol{\beta}$ and $\boldsymbol{\Phi} = [\phi_1, \ \dots \ , \phi_j]$. Minimizing $\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\Phi})$ with respect to $\{\mathbf{C}, \boldsymbol{\beta}, \boldsymbol{\Phi}\}$ is equivalent to separately fitting a NB regression model for each outcome, ignoring potential dependencies among covariates and responses.

In microbiome data analysis, however, we often observe correlated microbial taxa abundances. Moreover, it is biologically plausible that certain groups of host covariates (eg, diet components, participants' life style) only influence specific subsets of bacterial clades. To capture such response-predictors dependencies, we follow recent advances in multivariate mixed outcomes modeling[42,43] and introduce multivariate models for correlated predictors and interrelated responses via structured low-rank and sparse coefficient matrices.

The microbial abundance data model (1) with the rank constraint

$$\text{rank}(\mathbf{C}^*) \leq r^* \tag{6}$$

is referred to as negative binomial reduced rank regression, denoted by NB-RRR. The *low-rank* coefficient matrix $\mathbf{C}^*$ implies a significantly lower number of effective model parameters, thus enabling, under certain assumptions, better estimation in high-dimensional data problems.[34,39,43] Any low-rank coefficient matrix $\mathbf{C}^*$ can be expressed as the product of any two low-rank matrices. Based on the formulation of the linear predictor $\boldsymbol{\eta}^*$ (3), we associate the responses to latent factors that are constructed as linear combinations of predictors $\mathbf{X}$. To ensure identifiability in the large-dimensional setting, we may additionally assume that only a subset of predictors are relevant, and that the latent factors may be associated with only a subset of responses. This can be achieved by expressing $\mathbf{C}^*$ as a product of two unique and identifiable low-rank matrices that are *entry-wise sparse*. Motivated by the recent work of Mishra et al,[34,43] we assume that the singular value decomposition (SVD) of the coefficient matrix $\mathbf{C}^*$ in (1) is *co-sparse*[34] and decompose $\mathbf{C}^*$ as

$$\mathbf{C}^* = \mathbf{U}^*\mathbf{D}^*\mathbf{V}^{*^{\mathrm{T}}}, \quad \text{s.t.} \quad \mathbf{U}^{*^{\mathrm{T}}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{U}^*/n = \mathbf{V}^{*^{\mathrm{T}}}\mathbf{V}^* = \mathbf{I}_{r^*}, \tag{7}$$

where both the left singular vector matrix $\mathbf{U}^* = [\mathbf{u}^*_1, \ \dots \ , \mathbf{u}^*_{r^*}] \in \mathbb{R}^{p\times r^*}$ and the right singular vector matrix $\mathbf{V}^* = [\mathbf{v}^*_1, \ \dots \ , \mathbf{v}^*_{r^*}] \in \mathbb{R}^{q\times r}$ are assumed to be *sparse*, and $\mathbf{D} = \text{diag}\{d^*_1, \ \dots \ , d^*_{r^*}\} \in \mathbb{R}^{r^*\times r^*}$ is the diagonal matrix with the nonzero singular values on its diagonal. In the high-dimensional setting, this formulation facilitates better interpretation. Additional orthogonality constraints in the formulation ensure that the SVD of $\mathbf{C}^*$ is identifiable and the latent factors $(1/\sqrt{n})\mathbf{X}\mathbf{u}^*_k$ for $k = 1, \ \dots \ , r^*$ are uncorrelated. Each of the right singular vectors in $\mathbf{V}^*$ associates the corresponding latent factor to the microbial abundance outcome $\mathbf{Y}$, and the singular values $\{d^*_1, \ \dots \ , d^*_{r^*}\}$ denote the strength of the association. We term the proposed model negative binomial *co-sparse* factor regression, denoted by NB-FAR.

Since the majority of the available microbial count data only provides relative abundance information, it is common practice to scale or transform the data prior to statistical analysis.[19] Finding a suitable normalization/transformation approach remains an active area of research in microbiome data analysis.[18,19,47] Since both NB-RRR and NB-FAR are tailored toward modeling microbial *count* data $\mathbf{Y}$, we facilitate scaling/normalization of the data via specifying the offset matrix $\mathbf{O} \in \mathbb{R}^{n\times q}$ in (3). Let o_{ik} denotes sample i's and taxon k's entry of $\mathbf{O}$. Similar to DESeq2,[17] the R package nbfar offers several normalization schemes:

(i) $o_{ik} = \log \frac{m_i}{m_{\min}}$ where $m_i = \sum_{j=1}^{q} y_{ij}$ and $m_{\min} = \min_{1\le i\le n} m_i$, related common sum scaling.[18,19]
(ii) $o_{ik} = \log \sum_{j=1}^{q} y_{ij}$, related to total sum scaling.
(iii) $o_{ik} = \log \underset{j:G_j\neq 0}{\text{median}} \frac{y_{ij}}{G_j}$ where $G_j = \left(\prod_{i=1}^{n} y_{ij}\right)^{1/n}$ (similar to DESeq's size factors[45]).
(iv) $o_{ik} = \log \left(\prod_{j=1}^{q} y_{ij}\right)^{1/q}$, that is, sample-wise geometric mean scaling.

We emphasize that each normalization approach can significantly influence the outcome of the analysis.[18] Following McMurdie and Holmes 19, we use normalization (*i*) as default setting throughout this study. For convenience, the package also enables the inclusion of a fully user-defined offset matrix (eg, when data-specific prior knowledge is available).

3 | ESTIMATION PROCEDURES

Parameter estimation in both models requires minimizing a constrained, nonconvex negative log-likelihood function $\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\Phi})$ with respect to $\{\mathbf{C}, \boldsymbol{\beta}, \boldsymbol{\Phi}\}$. Joint estimation of the parameters $\{\mathbf{C}, \boldsymbol{\beta}, \boldsymbol{\Phi}\}$ satisfying the rank constraint (6) in the case of NB-RRR and the orthogonality constraints (7) in the case of NB-FAR is a notoriously difficult problem. In both cases, we solve the optimization problem using the majorization-minimization (MM) approach,[48] an alternating procedure that updates the blocks of parameters in cyclic order until convergence. In an update step, we minimize a convex surrogate that majorizes the objective function of the optimization problem.

3.1 | Negative binomial reduced rank regression

The optimization problem to estimate the parameters of NB-RRR is given by

$$(\hat{\mathbf{C}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Phi}}) \equiv \underset{\mathbf{C},\boldsymbol{\beta},\boldsymbol{\Phi}}{\arg\min}\, \mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\Phi}) \quad \text{s.t.} \quad \text{rank}(\mathbf{C}) \le r, \tag{8}$$

where $g(b'(\boldsymbol{\Theta}(\mathbf{C}, \boldsymbol{\beta}, \mathbf{O}))) = \boldsymbol{\eta}(\mathbf{C}, \boldsymbol{\beta}, \mathbf{O}) = \mathbf{O} + \mathbf{XC} + \mathbf{Z}\boldsymbol{\beta}$. Let us denote the problem as NB-RRR($\mathbf{C}, \boldsymbol{\beta}, \boldsymbol{\Phi}$; $\mathbf{W}, \mathbf{Z}, \mathbf{X}, \mathbf{O}, r$). Unless otherwise stated, we will write $\boldsymbol{\Theta}(\mathbf{C}, \boldsymbol{\beta}, \mathbf{O})$ as $\boldsymbol{\Theta}$ and $\boldsymbol{\eta}(\mathbf{C}, \boldsymbol{\beta}, \mathbf{O})$ as $\boldsymbol{\eta}$. Using the framework of MM, we minimize the objective function using an iterative procedure that cycles between $\mathbf{C}$-step, $\boldsymbol{\beta}$-step, and $\boldsymbol{\Phi}$-step to update $\mathbf{C}$, $\boldsymbol{\beta}$, and $\boldsymbol{\Phi}$, respectively, until convergence.

In the $\mathbf{C}$-step, for fixed $\boldsymbol{\beta}$ and $\boldsymbol{\Phi}$, let us denote the natural parameter $\boldsymbol{\Theta}$ and $\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\Phi})$ as $\boldsymbol{\Theta}(\mathbf{C})$ and $\mathcal{L}(\boldsymbol{\Theta}(\mathbf{C}))$, respectively. Suppose differentiable $\mathcal{L}(\boldsymbol{\Theta}(\mathbf{C}))$ is L-Lipschitz continuous gradient function for some constant L_c that is, $\|\nabla\mathcal{L}(\boldsymbol{\Theta}(\check{\mathbf{C}})) - \nabla\mathcal{L}(\boldsymbol{\Theta}(\mathbf{C}))\| \le L_c\|\check{\mathbf{C}} - \mathbf{C}\|$ for any conformable $\check{\mathbf{C}}$. The statement holds for any L_c such that $\sup_{\mathbf{C}} \|\nabla^2\mathcal{L}(\boldsymbol{\Theta}(\mathbf{C}))\| \le L_c = \max_{1\le j\le q} \|\mathbf{X}^{\mathrm{T}}\text{diag}(\mathbf{Y}_j + 1)\mathbf{X}\|/2$; see Section 1.1 of the supplementary material for details. Using the result, we majorize $\mathcal{L}(\boldsymbol{\Theta}(\mathbf{C}), \boldsymbol{\Phi})$ by a convex surrogate at a given $\check{\mathbf{C}}$ and update the parameter $\mathbf{C}$ as $\overline{\mathbf{C}} = \mathbb{T}^{(r)}(\check{\mathbf{C}} - \nabla\mathcal{L}(\boldsymbol{\Theta}(\check{\mathbf{C}}))/L_c)$, where $\mathbb{T}^{(r)}(\mathbf{M})$ extracts r SVD components of matrix $\mathbf{M}$. Similarly, in the $\boldsymbol{\beta}$-step, for fixed $\mathbf{C}$ and $\boldsymbol{\Phi}$, we denote $\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\Phi})$ as $\mathcal{L}(\boldsymbol{\Theta}(\boldsymbol{\beta}))$. Following the $\mathbf{C}$-step procedure, $\mathcal{L}(\boldsymbol{\Theta}(\boldsymbol{\beta}))$ is also L-Lipschitz continuous gradient function for some constant $L_b = \max_{1\le j\le q} \|\mathbf{Z}^{\mathrm{T}}\text{diag}(\mathbf{Y}_j + 1)\mathbf{Z}\|/2$; see Section 1.1 of the supplementary material for details. At any $\check{\boldsymbol{\beta}}$, we majorize $\mathcal{L}(\boldsymbol{\Theta}(\boldsymbol{\beta}))$ and then update the parameter $\boldsymbol{\beta}$ as $\overline{\boldsymbol{\beta}} = \check{\boldsymbol{\beta}} - \nabla\mathcal{L}(\boldsymbol{\Theta}(\check{\boldsymbol{\beta}}), \boldsymbol{\Phi})/L_b$. Finally, in the $\boldsymbol{\Phi}$-step, we follow Zeileis et al[44] to update each shape parameter ϕ_j using the Newton-Raphson method for fixed $\mathbf{C}$ and $\boldsymbol{\beta}$; see Section 1.6 of the supplementary material for details. We compute the parameter estimates for $1 \le r \le \tilde{r}$ where $\tilde{r}$ is the user-specified conservative maximum rank, and select a rank r using K-fold cross-validation.[49] The iterative procedure is summarized in Algorithm 1.

Algorithm 1. Negative binomial reduced rank regression (NB-RRR)

Input: $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{O}$ and rank $r \geq 1$; Set: $\mathbf{C}^{(0)} = \mathbf{0}$, $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{\Phi}^{(0)}$, $t \leftarrow 0$.
Set $L_b = \max_{1\leq j\leq q} \|\mathbf{Z}^{\mathrm{T}}\mathrm{diag}(\mathbf{Y}_{\cdot j} + 1)\mathbf{Z}\|/2$, $L_c = \max_{1\leq j\leq q} \|\mathbf{X}^{\mathrm{T}}\mathrm{diag}(\mathbf{Y}_{\cdot j} + 1)\mathbf{X}\|/2$
repeat
(1) $\mathbf{C}$-step: $\mathbf{C}^{(t+1)} = \mathbb{T}^{(r)}(\mathbf{C}^{(t)} - \nabla\mathcal{L}(\boldsymbol{\Theta}(\mathbf{C}^{(t)}))/L_c)$ where $g(b'(\boldsymbol{\Theta}(\mathbf{C}^{(t)}))) = \boldsymbol{\eta}(\mathbf{C}^{(t)}, \boldsymbol{\beta}^{(t)}, \mathbf{O})$, and $\mathbb{T}^{(r)}(\mathbf{M})$ extracts r SVD components of matrix $\mathbf{M}$.
(2) $\boldsymbol{\beta}$-step:$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \nabla\mathcal{L}(\boldsymbol{\Theta}(\boldsymbol{\beta}^{(t)}))/L_b$ where $g(b'(\boldsymbol{\Theta}(\boldsymbol{\beta}^{(t)}))) = \boldsymbol{\eta}(\mathbf{C}^{(t+1)}, \boldsymbol{\beta}^{(t)}, \mathbf{O})$.

(3) $\boldsymbol{\Phi}$-step: $\boldsymbol{\Phi}^{(t+1)} \leftarrow$ Apply Newton-Raphson to increase $\mathcal{L}(\boldsymbol{\Theta}^{(t+1)}, \boldsymbol{\Phi})$ at $\boldsymbol{\Phi}^{(t)}$ such that $g(b'(\boldsymbol{\Theta}^{(t+1)})) = \boldsymbol{\eta}(\mathbf{O}, \mathbf{C}^{(t+1)}, \boldsymbol{\beta}^{(t+1)})$ $t \leftarrow t + 1$.
until convergence, for example, $\|[\mathbf{C}^{(t+1)}\ \boldsymbol{\beta}^{(t+1)}] - [\mathbf{C}^{(t)}\ \boldsymbol{\beta}^{(t)}]\|_F / \|[\mathbf{C}^{(t)}\ \boldsymbol{\beta}^{(t)}]\|_F \leq \epsilon$ with $\epsilon = 10^{-6}$.
return $\widehat{\mathbf{C}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Phi}}$.

3.1.1 | Monotonically decreasing property of NB-RRR

Using Algorithm 1, we estimate the parameters $\{\mathbf{C}, \boldsymbol{\beta}, \boldsymbol{\Phi}\}$ of NB-RRR. The iterative procedure consists of minimizing several convex surrogates of the objective function with fixed Lipschitz constants $\{L_b, L_c\}$. Let us jointly denote the updated parameters from $\mathbf{C}$-step, $\boldsymbol{\beta}$-step and $\boldsymbol{\Phi}$-step after tth iteration by $\{\mathbf{C}^{(t)}, \boldsymbol{\beta}^{(t)}, \Phi^{(t)}\}$.

Theorem 1. *The sequence of parameter estimates* $\{\mathbf{C}^{(t)}, \boldsymbol{\beta}^{(t)}, \Phi^{(t)}\}_{t\in\mathbb{N}}$ *obtained using Algorithm 1 satisfies*

$$\mathcal{L}(\boldsymbol{\Theta}(\mathbf{C}^{(t+1)}, \boldsymbol{\beta}^{(t+1)}, \mathbf{O}), \boldsymbol{\Phi}^{(t+1)}) \leq \mathcal{L}(\boldsymbol{\Theta}(\mathbf{C}^{(t)}, \boldsymbol{\beta}^{(t)}, \mathbf{O}), \boldsymbol{\Phi}^{(t)})$$

for the constants $L_b = \max_{1\leq j\leq q} \|\mathbf{Z}^{\mathrm{T}}\mathrm{diag}(\mathbf{Y}_{\cdot j} + 1)\mathbf{Z}\|/2$ *and* $L_c = \max_{1\leq j\leq q} \|\mathbf{X}^{\mathrm{T}}\mathrm{diag}(\mathbf{Y}_{\cdot j} + 1)\mathbf{X}\|/2$.

We have relegated the proof of Theorem 1 to Section 1.2 of the supplementary material. In extensive simulation studies, we have found that the sequence always converges in practice.

3.2 | Negative binomial co-sparse factor regression

Joint estimation of the model parameters $\{\mathbf{U}, \mathbf{D}, \mathbf{V}, \boldsymbol{\beta}, \boldsymbol{\Phi}\}$ requires solving an optimization problem that minimizes $\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\Phi})$ in the presence of a sparsity-inducing penalty on $\{\mathbf{U}, \mathbf{V}\}$ and the orthogonality constraint (7).[43] The optimization problem requires the rank r to be specified. Since existing optimization tools are computationally inefficient for the task, we extend the sequential extraction procedure, proposed by Mishra et al,[34,43] for NB-FAR and estimate the unit-rank components of $\mathbf{C} = \sum_{k=1}^{r}\mathbf{C}_k = \sum_{k=1}^{r} d_k\mathbf{u}_k\mathbf{v}_k^{\mathrm{T}}$, that is, $(d_k, \mathbf{u}_k, \mathbf{v}_k)$, for $k = 1, \ldots, r$. Let $\hat{\mathbf{C}}_i$ for $i = 1, \ldots, k-1$ be the estimate of the unit-rank components. Then, to extract the kth unit-rank component in the kth step of the sequential procedure, we solve the optimization problem

$$\begin{aligned}(\hat{d}_k, \hat{\mathbf{u}}_k, \hat{\mathbf{v}}_k, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Phi}}) \equiv \underset{\mathbf{u},\mathbf{d},\mathbf{v},\boldsymbol{\beta},\boldsymbol{\Phi}}{\arg\min}\ & \mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\Phi}) + \rho_\lambda(\mathbf{C}), \\ \text{s.t.}\quad & \mathbf{C} = d\mathbf{u}\mathbf{v}^{\mathrm{T}},\ \mathbf{u}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{u}/n = \mathbf{v}^{\mathrm{T}}\mathbf{v} = 1, g(b'(\boldsymbol{\Theta})) = \boldsymbol{\eta}(\mathbf{C}, \boldsymbol{\beta}, \mathbf{O}^{(k)}),\end{aligned} \tag{9}$$

where $\rho_\lambda(\mathbf{C})$ is a sparsity-inducing penalty function with tuning parameter λ and $\mathbf{O}^{(k)} = \mathbf{O} + \mathbf{X}\sum_{i=2}^{k}\hat{\mathbf{C}}_{i-1}$ with $\mathbf{O}^{(1)} = \mathbf{O}$ is the offset term. This problem is referred to as negative binomial co-sparse unit-rank estimation (NB-CURE) with input parameters $\mathbf{C}, \boldsymbol{\beta}, \boldsymbol{\Phi}; \mathbf{Y}, \mathbf{X}, \mathbf{Z}, \mathbf{O}^{(k)}$ and penalty function ρ, in short, NB-CURE($\mathbf{C}, \boldsymbol{\beta}, \boldsymbol{\Phi}; \mathbf{Y}, \mathbf{X}, \mathbf{Z}, \mathbf{O}^{(k)}, \rho$).

Following Mishra et al,[43] we use the elastic net penalty and its adaptive version[50] for the kth step as

$$\rho_\lambda(\mathbf{C}) = \rho_\lambda(\mathbf{C}; \mathbf{W}, \alpha) = \alpha\lambda\|\mathbf{W}\circ\mathbf{C}\|_1 + (1-\alpha)\lambda\|\mathbf{C}\|_F^2, \tag{10}$$

where the operator "$\circ$" stands for the Hadamard product, $\mathbf{W} = [w_{ij}]_{p\times q}$ is a prespecified weighting matrix, λ is a tuning parameter controlling the overall amount of regularization and $\alpha \in (0, 1)$ controls the relative weights between the two penalty terms. In the kth step of NB-FAR, we let $\mathbf{W}_k = |\tilde{\mathbf{C}}_k|^{-\gamma}$, where $\gamma = 1$ and $\tilde{\mathbf{C}}_k = \tilde{d}_k\tilde{\mathbf{u}}_k\tilde{\mathbf{v}}_k^{\mathrm{T}}$ is an initial estimate of $\mathbf{C}_k$.

Here, we solve NB-RRR($\mathbf{C}, \boldsymbol{\beta}, \boldsymbol{\Phi}; \mathbf{Y}, \mathbf{Z}, \mathbf{X}, \mathbf{O}^{(k)}, 1$) to obtain this initial estimate $\tilde{\mathbf{C}}_k$ and extract $\{\tilde{d}_k, \tilde{\mathbf{u}}_k, \tilde{\mathbf{v}}_k\}$. Assuming that the NB-CURE problem can be solved for a suitable tuning parameter λ (see Section 3.2.1), NB-FAR's estimation procedure is summarized in Algorithm 2.

Algorithm 2. Negative binomial co-sparse factor regression (NB-FAR)

Input: $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{O}$ and rank $r \leq \text{rank}(\mathbf{X})$; Set: $\mathbf{C}^{(0)} = \mathbf{0}$, $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{\Phi}^{(0)}$, $t \leftarrow 0$.
for $k \leftarrow 1$ to r **do**
(1) Update offset: $\mathbf{O}^{(k)} = \mathbf{O} + \mathbf{X}\sum_{i=2}^{k} \widehat{\mathbf{C}}_{i-1}$
(2) Initialize: $\tilde{\mathbf{C}}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\Phi}}$ = NB-RRR($\mathbf{C}, \boldsymbol{\beta}, \boldsymbol{\Phi}; \mathbf{Y}, \mathbf{Z}, \mathbf{X}, \mathbf{O}^{(k)}, 1$) with $\tilde{\mathbf{C}} = \tilde{d}\tilde{\mathbf{u}}\tilde{\mathbf{v}}^{\mathrm{T}}$.
(3) Set $\mathbf{u}^{(0)} = \tilde{\mathbf{u}}$, $\mathbf{v}^{(0)} = \tilde{\mathbf{v}}$, $d^{(0)} = \tilde{d}$, $\boldsymbol{\beta}^{(0)} = \tilde{\boldsymbol{\beta}}$, $\boldsymbol{\Phi}^{(0)} = \tilde{\boldsymbol{\Phi}}$ and $\mathbf{W} = |\tilde{\mathbf{C}}_k|^{-\gamma}$ where $\gamma = 1$.
(4) Solve NB-CURE($\mathbf{C}, \boldsymbol{\beta}, \boldsymbol{\Phi}; \mathbf{Y}, \mathbf{X}, \mathbf{Z}, \mathbf{O}^{(k)}, \rho$) such that $\mathbf{C} = d\mathbf{u}\mathbf{v}^{\mathrm{T}}$.
(5) $\widehat{\mathbf{u}}_k = \widehat{\mathbf{u}}$, $\widehat{d}_k = \widehat{d}$, $\widehat{\mathbf{v}}_k = \widehat{\mathbf{v}}$, $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\Phi}} = \widehat{\boldsymbol{\Phi}}$ and $\widehat{\mathbf{C}}_k = \widehat{d}_k\widehat{\mathbf{u}}_k\widehat{\mathbf{v}}_k^{\mathrm{T}}$.
if $\widehat{d}_k = 0$ **then** Set $\widehat{r} = k$ **end if**
end for
return $\widehat{\mathbf{C}} = \sum_{k=1}^{\widehat{r}} \widehat{\mathbf{C}}_k$, $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\Phi}}$.

3.2.1 | Computation of negative binomial constrained unit-rank regression (NB-CURE)

The general form of the optimization problem for the kth step of the sequential procedure, that is, NB-CURE($\mathbf{C}, \boldsymbol{\beta}, \boldsymbol{\Phi}; \mathbf{Y}, \mathbf{Z}, \mathbf{X}, \mathbf{O}, \rho$), is given by

$$(\hat{d}, \hat{\mathbf{u}}, \hat{\mathbf{v}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Phi}}) \equiv \underset{\mathbf{u},\mathbf{d},\mathbf{v},\boldsymbol{\beta},\boldsymbol{\Phi}}{\arg\min} \{F_\lambda(d, \mathbf{u}, \mathbf{v}, \boldsymbol{\beta}, \boldsymbol{\Phi}) = \mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\Phi}) + \rho_\lambda(\mathbf{C}; \mathbf{W})\},$$
$$\text{s.t.} \quad \mathbf{C} = d\mathbf{u}\mathbf{v}^{\mathrm{T}}, \ \mathbf{u}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{u}/n = \mathbf{v}^{\mathrm{T}}\mathbf{v} = 1, g(b'(\boldsymbol{\Theta})) = \boldsymbol{\eta}(\mathbf{C}, \boldsymbol{\beta}, \mathbf{O}), \tag{11}$$

where $\mathbf{W} = w^{(d)}\mathbf{w}^{(u)}\mathbf{w}^{(v)^{\mathrm{T}}}$. In practice, we fix $\alpha = 0.95$ and denote $\rho_\lambda(\mathbf{C}; \mathbf{W}, \alpha)$ as $\rho_\lambda(\mathbf{C}; \mathbf{W})$. Similar to NB-RRR model estimation, we use the MM framework and solve the optimization problem using an iterative procedure that cycles between $\mathbf{u}$-step, $\mathbf{v}$-step, $\boldsymbol{\beta}$-step, and $\boldsymbol{\Phi}$-step to update the parameters in blocks of $(\mathbf{u}, d)$, $(\mathbf{v}, d)$, $\boldsymbol{\beta}$, and $\boldsymbol{\Phi}$, respectively, until convergence.

Let us represent $\boldsymbol{\Theta}$ and $\mathcal{L}(\boldsymbol{\Theta}(\check{\mathbf{u}}\mathbf{v}^{\mathrm{T}}), \boldsymbol{\Phi})$, which are functions of $\mathbf{C}$, as $\boldsymbol{\Theta}(\mathbf{C})$ and $\mathcal{L}(\boldsymbol{\Theta}(\check{\mathbf{u}}\mathbf{v}^{\mathrm{T}}))$, respectively. In the $\mathbf{u}$-step, for fixed $\{\mathbf{v}, \boldsymbol{\beta}, \boldsymbol{\Phi}\}$, $\mathcal{L}(\boldsymbol{\Theta}(\check{\mathbf{u}}\mathbf{v}^{\mathrm{T}}))$ has L-Lipschitz continuous gradients for some L_u where $\check{\mathbf{u}} = d\mathbf{u}$. Again, using this fact and following the NB-RRR estimation procedure, we majorize $\mathcal{L}(\boldsymbol{\Theta}(\check{\mathbf{u}}\mathbf{v}^{\mathrm{T}}))$ by a convex surrogate and then minimize it to update the block variable $\check{\mathbf{u}}$ as

$$\hat{\check{\mathbf{u}}} = \mathbf{S}(\check{\mathbf{u}} - \nabla\mathcal{L}(\boldsymbol{\Theta}(\check{\mathbf{u}}\mathbf{v}^{\mathrm{T}}))/L_u; \alpha\lambda\mathbf{v}^{\mathrm{T}}\mathbf{w}^{(v)}w^{(d)}\mathbf{w}^{(u)}/L_u)/\{1 + 2\lambda(1-\alpha)\|\mathbf{v}\|_2^2/L_u\}, \tag{12}$$

where $\mathbf{S}(\mathbf{t}; \tilde{\lambda}) = \text{sign}(\mathbf{t})(|\mathbf{t}| - \tilde{\lambda})_+$ is the element-wise soft-thresholding operator on any $\mathbf{t} \in \mathbb{R}^p$; see Section 1.3 of the supplementary material for details. Similarly, in the $\mathbf{v}$-step, for fixed $\{\mathbf{u}, \boldsymbol{\beta}, \boldsymbol{\Phi}\}$ with $\check{\mathbf{v}} = d\mathbf{v}$, we show that $\mathcal{L}(\boldsymbol{\Theta}(\mathbf{u}\check{\mathbf{v}}^{\mathrm{T}}))$ is a L-Lipschitz continuous gradient function for some L_v and update the block variable $\check{\mathbf{v}}$ as

$$\hat{\check{\mathbf{v}}} = \mathbf{S}(\check{\mathbf{v}} - \nabla\mathcal{L}(\boldsymbol{\Theta}(\mathbf{u}\check{\mathbf{v}}^{\mathrm{T}}))/L_v; \alpha\lambda\mathbf{u}^{\mathrm{T}}\mathbf{w}^{(u)}w^{(d)}\mathbf{w}^{(v)}/L_v)/\{1 + 2\lambda(1-\alpha)\|\mathbf{u}\|_2^2/L_v\}. \tag{13}$$

We apply the equality constraints in (9) to recover the estimate of $\{d, \mathbf{u}, \mathbf{v}\}$ from the estimate of the block variables $\{\check{\mathbf{u}}, \check{\mathbf{v}}\}$. In the $\boldsymbol{\beta}$-step and the $\boldsymbol{\Phi}$-step, we follow the corresponding step of NB-RRR parameter estimation (see Algorithm 1) to update $\boldsymbol{\beta}$ and $\boldsymbol{\Phi}$, respectively. We have relegated the details of the convex surrogate function that majorizes the objective function, the computation of the constants (L_u, L_v, L_b), and the update steps to Section 1.3 of the supplementary material.

We compute the parameter estimates for several λ values (the default is 50) in the range of $\lambda_{\max}$ to $\lambda_{\min}$ that are equi-spaced on a log scale, where $\lambda_{\max} = 2\|\mathbf{X}^{\mathrm{T}}(\mathbf{Y} - g^{-1}(\mathbf{O}))\|_\infty$ and $\lambda_{\min} = 1e^{-6} \times \lambda_{\max}$. We apply K-fold cross-validation[49] to select a tuning parameter λ. The iterative procedure is summarized in Algorithm 3.

Algorithm 3. Negative binomial constrained unit-rank regression (NB-CURE)

Input: $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{O}, \lambda, \mathbf{W}$; Set $L_b = \max_{1\le j\le q} \|\mathbf{Z}^{\mathrm{T}}\mathrm{diag}(\mathbf{Y}_{\cdot j}+1)\mathbf{Z}\|/2$.
Initialize $\mathbf{u}^{(0)} = \tilde{\mathbf{u}}$, $\mathbf{v}^{(0)} = \tilde{\mathbf{v}}$, $d^{(0)} = \tilde{d}$, $\boldsymbol{\beta}^{(0)} = \tilde{\boldsymbol{\beta}}$, $\boldsymbol{\Phi}^{(0)} = \tilde{\boldsymbol{\Phi}}$
repeat
$L_u = \|\mathbf{X}^{\mathrm{T}}\mathbf{X} + \sum_{i=1}^{n}\mathbf{x}_i\Big(\sum_{k=1}^{q} y_{ik}v_k^{(t)2}\Big)\mathbf{x}_i^{\mathrm{T}}\|/2,\ L_v = \frac{\max_{1\le j\le q}\mathbf{u}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathrm{diag}(\mathbf{Y}_{\cdot j}+1)\mathbf{X}\mathbf{u}}{2}$
(I) $\mathbf{u}$-step: Set $\check{\mathbf{u}} = d^{(t)}\mathbf{u}^{(t)}$ and $\mathbf{v} = \mathbf{v}^{(t)}$. Update $\check{\mathbf{u}}^{(t+1)}$ using (12).Recover block variable $(\tilde{d}^{(t+1)}, \mathbf{u}^{(t+1)})$ using equality constraint in (11).
(II) $\mathbf{v}$-step: Set $\check{\mathbf{v}} = \tilde{d}^{(t+1)}\mathbf{v}^{(t)}$ and $\mathbf{u} = \mathbf{u}^{(t+1)}$. Update $\check{\mathbf{v}}^{(t+1)}$ using (13).Recover block variable $(d^{(t+1)}, \mathbf{v}^{(t+1)})$ using equality constraint in (11).
(III) $\boldsymbol{\beta}$-step: $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \frac{1}{L_b}\nabla\mathcal{L}(\boldsymbol{\Theta}(\mathbf{C}^{(t+1)}, \boldsymbol{\beta}^{(t)}), \boldsymbol{\Phi})$ where $\mathbf{C}^{(t+1)} = d^{(t+1)}\mathbf{u}^{(t+1)}\mathbf{v}^{(t+1)\mathrm{T}}$.
(IV) $\boldsymbol{\Phi}$-step: $\boldsymbol{\Phi}^{(t+1)} \leftarrow$ Apply Newton-Raphson to increase $\mathcal{L}(\boldsymbol{\Theta}^{(t+1)}, \boldsymbol{\Phi})$ at $\boldsymbol{\Phi}^{(t)}$ such that $g(b'(\boldsymbol{\Theta}^{(t+1)})) = \boldsymbol{\eta}(\mathbf{O}, \mathbf{C}^{(t+1)}, \boldsymbol{\beta}^{(t+1)})$
$t \leftarrow t+1$.
until convergence, for example, the relative ℓ_2 change in parameters is less than $\epsilon = 10^{-6}$.
return $\hat{\mathbf{u}}, \hat{d}, \hat{\mathbf{v}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Phi}}$

3.2.2 | Monotonically decreasing property of NB-CURE

Using Algorithm 3, we solve the optimization problem of NB-CURE to estimate the parameters $\{d, \mathbf{u}, \mathbf{v}, \boldsymbol{\beta}, \boldsymbol{\Phi}\}$. In the iterative procedure, the objective function (11) is majorized by a convex surrogate in each of the $\mathbf{u}$-step, $\mathbf{v}$-step, and $\boldsymbol{\beta}$-step, and then minimized. Let us jointly denote the updated parameters from $\mathbf{u}$-step, $\mathbf{v}$-step, $\boldsymbol{\beta}$-step, and $\boldsymbol{\Phi}$-step after tth iteration by $\{d^{(t)}, \mathbf{u}^{(t)}, \mathbf{v}^{(t)}, \boldsymbol{\beta}^{(t)}, \Phi^{(t)}\}$.

Theorem 2. *The sequence of parameters estimate* $\{d^{(t)}, \mathbf{u}^{(t)}, \mathbf{v}^{(t)}, \boldsymbol{\beta}^{(t)}, \Phi^{(t)}\}_{t\in\mathbb{N}}$ *obtained from Algorithm 3 satisfies*

$$F_\lambda(d^{(t+1)}, \mathbf{u}^{(t+1)}, \mathbf{v}^{(t+1)}, \boldsymbol{\beta}^{(t+1)}, \Phi^{(t+1)}) \le F_\lambda(d^{(t)}, \mathbf{u}^{(t)}, \mathbf{v}^{(t)}, \boldsymbol{\beta}^{(t)}, \Phi^{(t)})$$

for $L_u = \|\mathbf{X}^{\mathrm{T}}\mathbf{X} + \sum_{i=1}^{n}\mathbf{x}_i\left(\sum_{k=1}^{q} y_{ik}v_k^{(t)2}\right)\mathbf{x}_i^{\mathrm{T}}\|/2$, $L_v = \frac{\max_{1\le j\le q}\mathbf{u}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathrm{diag}(\mathbf{Y}_{\cdot j}+1)\mathbf{X}\mathbf{u}}{2}$ *and* $L_b = \max_{1\le j\le q}\|\mathbf{Z}^{\mathrm{T}}\mathrm{diag}(\mathbf{Y}_{\cdot j}+1)\mathbf{Z}\|/2$.

We have relegated the proof of Theorem 2 to Section 1.4 of the supplementary material. Similar to NB-RRR, we have found in extensive simulation studies that the sequence always converges in practice.

3.2.3 | Handling missing outcome values in NB-FAR

Besides microbiome data, the NB factor models may also prove useful for multivariate count data in other domains, including genomics, sports, image analysis, and text mining. A common scenario in these domains is the presence of missing entries in the outcome matrix $\mathbf{Y}$. To highlight NB-FAR's ability to account for missing entries, we can extend the framework of (5) by calculating the negative log-likelihood as follows.

Let us define an index set of the observed entries in $\mathbf{Y}$ as $\boldsymbol{\Omega} = \{(i,k); y_{ik} \text{ is observed}, i = 1, \dots, n, k = 1, \dots, q\}$, and denote the projection of $\mathbf{Y}$ onto $\boldsymbol{\Omega}$ by $\tilde{\mathbf{Y}} = \mathcal{P}_{\boldsymbol{\Omega}}(\mathbf{Y})$, where $\tilde{y}_{ik} = y_{ik}$ for any $(i,k) \in \boldsymbol{\Omega}$ and $\tilde{y}_{ik} = 0$ otherwise. Following Mishra et al,[43] we write the negative log-likelihood function with incomplete data as

$$\mathcal{L}(\boldsymbol{\Theta}^*, \boldsymbol{\Phi}^*) = -\mathrm{tr}(\tilde{\mathbf{Y}}^{\mathrm{T}}\boldsymbol{\Theta}^*) + \mathrm{tr}(\tilde{\mathbf{J}}^{\mathrm{T}}\mathbf{B}(\boldsymbol{\Theta}^*)) + \sum_{i,\, j\in\boldsymbol{\Omega}} \log\binom{y_{ij} + \phi_j^* - 1}{y_{ij}}, \tag{14}$$

where $\tilde{\mathbf{J}} = \mathcal{P}_{\boldsymbol{\Omega}}(\mathbf{J})$ and $g(b'(\boldsymbol{\Theta}^*) = \boldsymbol{\eta}^*$. In case of missing entries in the outcome matrix $\mathbf{Y}$, one should replace $\mathbf{Y}$ with $\tilde{\mathbf{Y}}$ and $\mathbf{J}$ with $\tilde{\mathbf{J}}$ and apply our proposed procedure to estimate the parameters. The same approach is applicable in the NB-RRR model.

4 | SIMULATION STUDIES

4.1 | Setup

We compare the performance of NB-RRR and NB-FAR with GO-FAR and NB-GLM to showcase the efficacy of the proposed procedures in modeling multivariate overdispersed count data in the high/large-dimensional settings. We evaluate the performance of the methods in terms of estimation error, prediction accuracy, sparsity recovery, rank identification, and shape error. GO-FAR (implemented in the R package gofar) assumes that the underlying distribution of the count outcomes is Poisson. The specific comparison with GO-FAR allows us to probe the effect of overdispersion in the data on model quality. The comparison with NB-GLM highlights the potential limitations of marginal approaches in modeling dependent variables.

We followed Mishra et al[43] to simulate the predictor matrix $\mathbf{X}$ and sparse SVD components of the coefficient matrix $\mathbf{C}^*$, that is, $\{\mathbf{U}^*, \mathbf{D}^*, \mathbf{V}^*\}$. Our setup considers the true rank $r^* = 3$ with $\mathbf{U}^* = [\mathbf{u}_1^*, \mathbf{u}_2^*, \mathbf{u}_3^*]$, $\mathbf{V}^* = [\mathbf{v}_1^*, \mathbf{v}_2^*, \mathbf{v}_3^*]$, and $\mathbf{D}^* = \text{diag}[d_1^*, d_2^*, d_3^*]$ such that $d_1^* = 6$, $d_2^* = 5$, $d_3^* = 4$. The notations $\text{unif}(\mathcal{A}, b)$ denote a vector of length b whose entries are uniformly distributed on the set $\mathcal{A}$ and $\text{rep}(a, b)$ denote the vector of length b with all entries equal to a. We generate $\mathbf{u}_k^*$ as $\mathbf{u}_k^* = \check{\mathbf{u}}_k / \|\check{\mathbf{u}}_k\|$, where $\check{\mathbf{u}}_1 = [\text{unif}(\mathcal{A}_u, 8), \text{rep}(0, p-8)]^{\mathrm{T}}$, $\check{\mathbf{u}}_2 = [\text{rep}(0, 5), \text{unif}(\mathcal{A}_u, 9), \text{rep}(0, p-14)]^{\mathrm{T}}$, and $\check{\mathbf{u}}_3 = [\text{rep}(0, 11), \text{unif}(\mathcal{A}_u, 9), \text{rep}(0, p-20)]^{\mathrm{T}}$. Similarly, we generate $\mathbf{v}_k^*$ as $\mathbf{v}_k^* = \check{\mathbf{v}}_k / \|\check{\mathbf{v}}_k\|$, where $\check{\mathbf{v}}_1 = [\text{unif}(\mathcal{A}_v, 5), \text{rep}(0, q-5)]^{\mathrm{T}}$, $\check{\mathbf{v}}_2 = [\text{rep}(0, 5), \text{unif}(\mathcal{A}_v, 5), \text{rep}(0, q-10)]^{\mathrm{T}}$, and $\check{\mathbf{v}}_3 = [\text{rep}(0, 10), \text{unif}(\mathcal{A}_v, 5), \text{rep}(0, q-15)]^{\mathrm{T}}$. Here we set $\mathcal{A}_u = \pm 1$ and $\mathcal{A}_v = [-1, -0.3] \cup [0.3, 1]$. An intercept is included in the model by setting $\mathbf{Z} = \mathbf{1}_n$ with $\boldsymbol{\beta}^* = [\text{rep}(0.5, q)]^{\mathrm{T}}$. We have considered simulation settings with $p = 100$ and $p = 300$ to demonstrate the efficacy of the proposed procedure in large/high-dimensional examples.

We simulate the predictor matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ from a multivariate normal distribution with some rotations such that the latent factors $\mathbf{X}\mathbf{U}^*/\sqrt{n}$ satisfy the orthogonality constraint (7); we refer to the simulation study of Mishra et al[34] for details on the formulation. At the OTU/ASV level in the taxonomy, typical microbial abundance observations are excessively sparse. Since our factor models are tailored toward modeling taxa aggregated on a higher taxonomic rank, for example, the family level, we first estimate the level of sparsity in the observed AGP data. We found that, on the family level, 20% of the entries AGP data are zeros. In the simulation setting, we thus set the shape parameters to $\phi_k^* = 0.5$ for $k = 1, \ldots, q$, resulting in 20% zero entries in the simulated outcome matrix $\mathbf{Y}$. Based on the model suggested in (1), we simulate $\mathbf{Y}$ such that $g(\boldsymbol{\mu}^*) = \boldsymbol{\eta}^* = \mathbf{Z}\boldsymbol{\beta}^* + \mathbf{X}\mathbf{C}^*$. Finally, we also include a simulation scenario where 20% of entries in the response matrix $\mathbf{Y}$ are missing at random. The latter scenario showcases the ability of NB-FAR and NB-RRR to handle missing values in $\mathbf{Y}$.

We evaluate model performance in terms of (a) the estimation error $\text{Er}(\mathbf{C}) = \|\hat{\mathbf{C}} - \mathbf{C}^*\|_F/(pq)$, (b) the prediction error $\text{Er}(\boldsymbol{\eta}) = \|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*\|_F/(nq)$, (c) sparsity recovery using the false positive rate (FPR) and the false negative rate (FNR), and (d) rank estimation $\hat{r}$. FNR is computed by comparing the support (nonzero entries) of $(\mathbf{u}_k^*, \mathbf{v}_k^*)$ with corresponding entries in its estimate $(\hat{\mathbf{u}}_k, \hat{\mathbf{v}}_k^*)$ for $k = 1, \ldots, r^*$. The FPR, on the other hand, compares zero entries in $(\mathbf{u}_k^*, \mathbf{v}_k^*)$ with corresponding entries in its estimate $(\hat{\mathbf{u}}_k, \hat{\mathbf{v}}_k^*)$ for $k = 1, \ldots, r^*$. In case of overestimated rank, we report the relative residual signal in the excessive components as $\text{R}\% = 100(\sum_{i=r^*+1}^{\hat{r}} \hat{d}_i^2)/(\sum_{i=1}^{\hat{r}} \hat{d}_i^2)$. The error in the shape parameter estimate is reported as $\text{Er}(\Phi) = \|\hat{\boldsymbol{\Phi}} - \boldsymbol{\Phi}^*\|/\sqrt{q}$.

4.2 | Results

Since we observed similar model performances in the $p = 100$ and $p = 300$ scenarios, we report the model evaluation statistics only for the latter case in Table 1. The results are average performances over 100 replicates. The model comparison for the $p = 100$ case is available in Table S1 of the supplementary material. The boxplot in Figure 2 compares the models in terms of prediction error $\text{Er}(\boldsymbol{\eta})$ (see Figure S1 of the supplementary material for comparison on the basis of estimation error $\text{Er}(\mathbf{C})$). Compared to the standard approach of modeling overdispersed count outcome using the marginal negative binomial regression model (NB-GLM) or the Poisson counterpart of NB-FAR, that is, GO-FAR, both NB factor models show superior performance in terms of parameter estimation, prediction, support identification, and rank estimation. In the present model scenario, NB-FAR outperforms NB-RRR at the expense of a higher computational cost. Since the true parameters of the underlying model are sparse, we expect and confirm this superior behavior of NB-FAR. The performance decrease of the GO-FAR model highlights the effect of the misspecification in the model with respect to the overdispersed data. In particular, the performance of GO-FAR considerably deteriorates in terms of false negative rate.

TABLE 1 Simulation: Model evaluation based on 100 replications using various performance measures (standard deviations are shown in parentheses) in case of $p = 300$ with negative binomial responses

	M%	Er(C)	Er(Θ)	FPR	FNR	*R%*	*r*	Er(Φ)	Time (seconds)
NB-FAR	0	2.68 (1.10)	20.70 (3.32)	5.10 (1.60)	1.42 (1.44)	0.00 (0.00)	3.00 (0.00)	0.10 (0.07)	263.20 (28.01)
NB-RRR	0	14.49 (1.14)	40.99 (1.71)	100.00 (0.00)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)	0.63 (0.14)	95.63 (8.43)
GO-FAR	0	10.79 (2.17)	47.58 (5.77)	4.56 (3.58)	49.80 (30.21)	0.00 (0.00)	2.51 (0.98)	1.37 (0.00)	60.49 (27.45)
NB-GLM	0	276.31 (10.27)	163.75 (2.15)	100.00 (0.00)	0.00 (0.00)	69.66 (1.29)	29.58 (0.56)	3741.83 (149.34)	1376.36 (35.38)
NB-FAR	20	3.36 (1.35)	22.98 (3.58)	5.63 (2.13)	2.51 (2.29)	0.00 (0.00)	3.00 (0.00)	0.12 (0.08)	273.12 (26.96)
NB-RRR	20	15.15 (1.35)	43.77 (2.07)	100.00 (0.00)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)	0.74 (0.17)	92.31 (8.80)
GO-FAR	20	11.78 (2.59)	50.37 (7.45)	4.89 (3.79)	54.17 (28.03)	0.00 (0.00)	2.54 (0.99)	1.37 (0.00)	79.46 (20.03)
NB-GLM	20	228.27 (7.35)	172.59 (2.53)	100.00 (0.00)	0.00 (0.00)	71.53 (1.24)	29.68 (0.47)	3773.07 (135.35)	1774.59 (41.64)

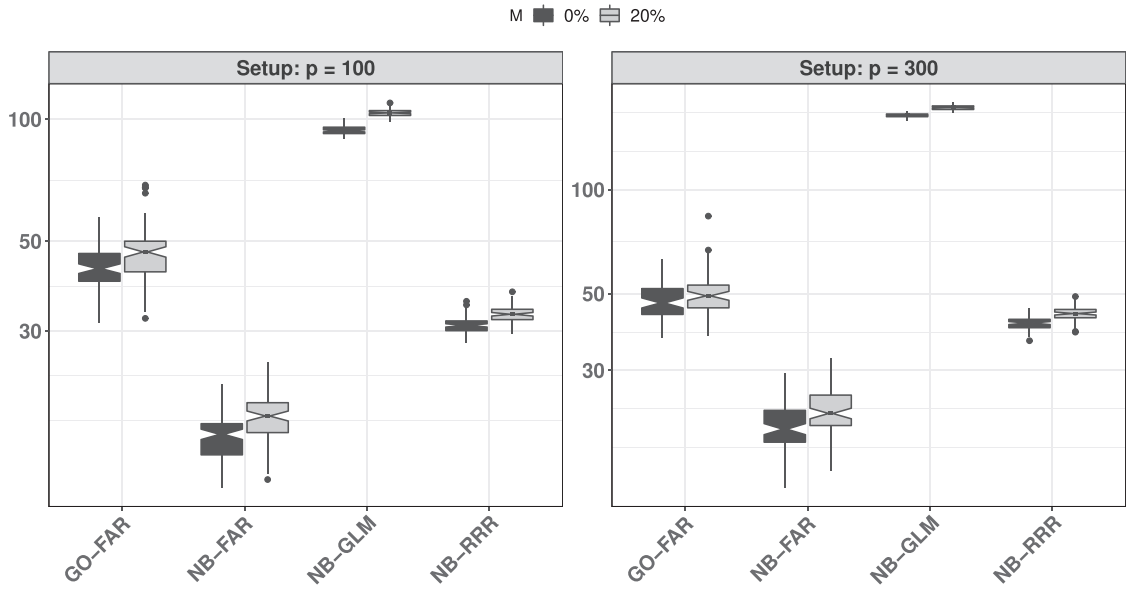

FIGURE 2 Notched boxplots of the prediction error Er($\boldsymbol{\eta}$) on the simulated count data under Setup I ($p = 100$) and II ($p = 300$), respectively. The results are based on 100 replications

Based on the results in the simulation examples where 20% of entries in $\mathbf{Y}$ are missing (*M%*20 rows in Table 1), we observe that NB-FAR can efficiently estimate the model parameters with slight deterioration compared to the full data model.

5 | APPLICATION

We now illustrate the performance of the NB factor models using the microbial abundance data from the AGP.[13] AGP comprises around 30 000 fecal, oral, hand, skin, and other body site samples. We used Qiita,[12] an open-source platform for microbial study, to access the raw data. We selected and downloaded the microbial abundance data that were obtained after processing 150-nucleotides-long trimmed sequences from the 16S V4 region and used the provided Greengenes reference database for taxonomic annotation. The AGP study also records metadata/covariates related to each participant's health condition, diet, demography, nutrient intake, and habit. Here, we focused on a subset of $n = 627$ participants where both fecal amplicon data (with sufficient sequencing depth >2000) and VioScreen variables were available. The VioScreen variables provide a detailed account of the dietary habits of the participants.

We aggregated the microbial count data to the family level using the available taxonomic annotation and performed sample-wise geometric mean scaling[18,19] at the minimum sequencing depth. After dropping taxa observed in less than

10% of samples, we arrive at $q = 39$ microbial families as multivariate outcome $\mathbf{Y}$. We curated the metadata as follows. First, we dropped several descriptive variables, such as, for example, sample name and sample identifier. We then removed all variables that were missing in more than 50 out of $n = 627$ samples. The final covariate matrix $\mathbf{X}$ comprises $p = 357$ predictors. Each of the continuous predictors in $\mathbf{X}$ are both centered and scaled. We manually assigned each of the different predictors to high-level categories, such as, for example, diet, habit, health-related, nutrient-related, etc. The curated AGP dataset with the assigned categories is available on the GitHub page of the project (see also agAnalysis.R in the supplementary material). Finally, we considered gender, body mass index (BMI), and age as control variables $\mathbf{Z}$ and included an intercept, leading to $\mathbf{Z} \in 627 \times 4$, and sample-wise geometric mean scaling in the offset specification.

We used NB-FAR and NB-RRR to learn about the underlying association between the set of $p = 357$ covariates and the observed microbial family abundances with a special focus on the patterns in the low-rank and sparse coefficient matrix estimates $\mathbf{C}$. For comparison, we also ran NB-GLM and GO-FAR and assessed model quality based on AIC, BIC, and log-likelihood. We also report rank estimates r and model size. Table 2 summarizes model performance results from 100 replications with 80% of data used for training and 20% for testing. Compared to the marginal approach of NB-GLM and the GO-FAR model, NB-FAR and NB-RRR achieve considerably lower AIC, BIC, and log-likelihood. Both NB-FAR and NB-RRR have comparable prediction errors on the test data and choose approximately rank-3 models. Interestingly, GO-FAR estimates a rank-1 model, though at the expense of decreased predictive performance.

Since NB-FAR achieved comparable performance to NB-RRR in terms of predictive log-likelihood with considerably reduced model complexity, we re-estimated the model parameters of NB-FAR on the full data set using Algorithm 2. NB-FAR again identified a rank($\mathbf{C}$) = 3 solution, enabling a parsimonious and interpretable description of host covariates—microbial family associations with only three sparse latent factors, given by sparse left and right singular vectors $\hat{\mathbf{U}}$ and $\mathbf{V}$. The support size (% of nonzero entries) of the estimates of the singular vectors are supp($\mathbf{U}$) = $\{16\%, 17\%, 20\%\}$ and supp($\mathbf{V}$) = $\{28\%, 74\%, 59\%\}$. Using the union of the support of the estimated $\mathbf{U}$ and $\mathbf{V}$, we can visualize all associations with the block-sparse coefficient matrix $\mathbf{C}$, shown in Figure 3(left panel). The other panels in Figure 3 display the individual unit-rank components $\mathbf{C}_1$, $\mathbf{C}_2$, and $\mathbf{C}_3$. Note that each of the unit-rank components is orthogonal to one another. We highlight the high-level categories of the covariates, including diet, habit, and health, by vertical lines. Horizontal lines delineate the different microbial phyla to which the families belong to. We observed that each of the singular vectors induced a different sparse pattern of positive (red) and negative (blue) blocks of associations between covariates and taxa. Overall, we found that 30 out of the 39 microbial families were found to be associated with host-associated covariates. The families *Pseudomonadaceae*, *Barnesiellaceae*, *Paraprevotellaceae*, *Christensenellaceae*, *Coriobacteriaceae*, *ML615J-28*, *Mogibacteriaceae*, *Ys2*, and *Oxalobacteraceae* were not associated with any of the covariates.

As an example for how to read and interpret the coefficient matrices, consider the top-left sub-matrix of $\mathbf{C}$ which relates the two Actinobacteria (red row label) *Bifidobacteriaceae* and *Corynebacteriaceae* to diet covariates (green column label). Here, we observe an almost disjoint association pattern of positive and negative factors with diet covariates for these two families. This pattern arises from the first two latent factor $\mathbf{C}_1$ and $\mathbf{C}_2$ (Figure 3, middle panels). There, we observe a unique nonzero association pattern of diet variables with *Corynebacteriaceae* (second row in $\mathbf{C}_1$) and a unique nonzero association pattern of diet variables with *Bifidobacteriaceae* (first row in $\mathbf{C}_2$). The third latent component does not contribute any additional associations.

We next focused on the analysis of the overall most important covariates-taxa associations. The left panel of Figure 4 shows the top 25 covariates based on the row sum $\sum_{j=1}^{q} |\hat{\mathbf{C}}|_{ij}$ of the absolute values of the estimated coefficient matrix; the

TABLE 2 Summary of average model performances on the AGP data in terms of AIC, BIC, and log-likelihood (second to fourth column) on the test data

Model	AIC	BIC	Log-likelihood	r	Supp(U)	Supp(V)
NB-FAR	1.37	1.71	−9.44	3.15	7.97	16.80
NB-RRR	1.86	3.69	−9.23	3.15	100.00	100.00
GO-FAR	34.90	35.20	−19.00	1.00	17.90	3.59
NB-GLM	36.40	37.10	−19.80	27.20	55.10	86.00

Note: The other columns summarize the average rank estimate r and support of the singular vectors of $\mathbf{C}$ as {Supp($\mathbf{U}$), Supp($\mathbf{V}$)}.

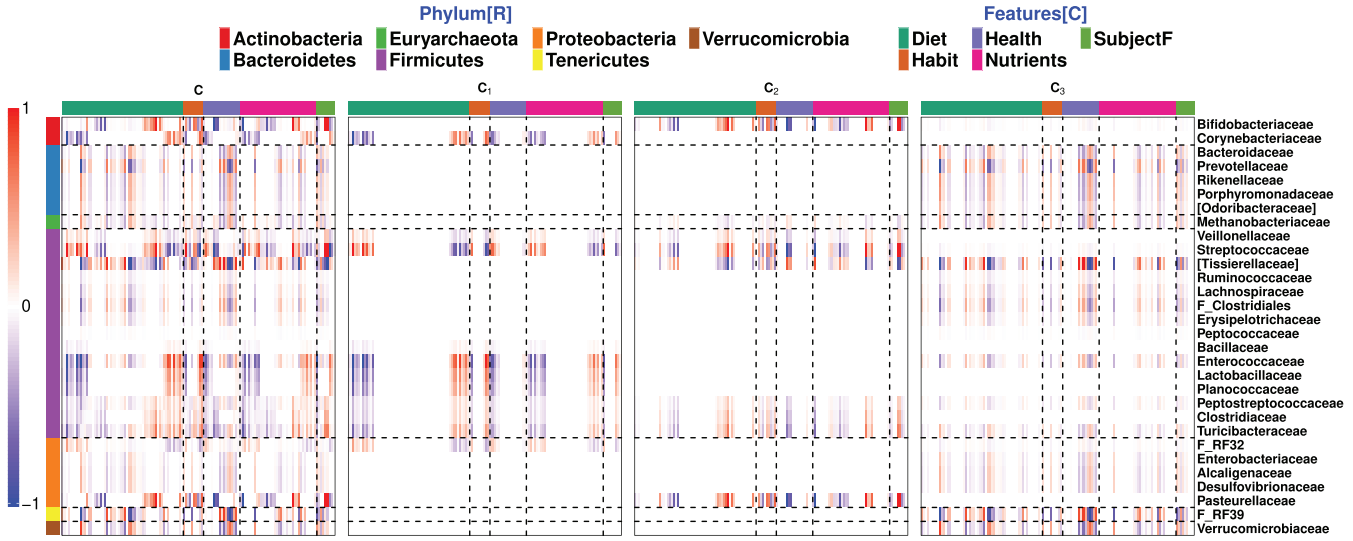

FIGURE 3 Application—AGP: The sparse estimate of the selected rows and columns of the coefficient matrix $\hat{\mathbf{C}}$ with its corresponding unit-rank components using NB-FAR. Based on the phylum of the taxon, horizontal lines separate the response into seven ranks: Actinobacteria, Bacteroidetes, Euryarchaeota, Firmicutes, Proteobacteria, Tenericutes, and Verrucomicrobia (top to bottom). Based on the type of the covariates, vertical lines (left to right) separate the selected predictors into five categories: diet, habit, health, nutrients, and subject features

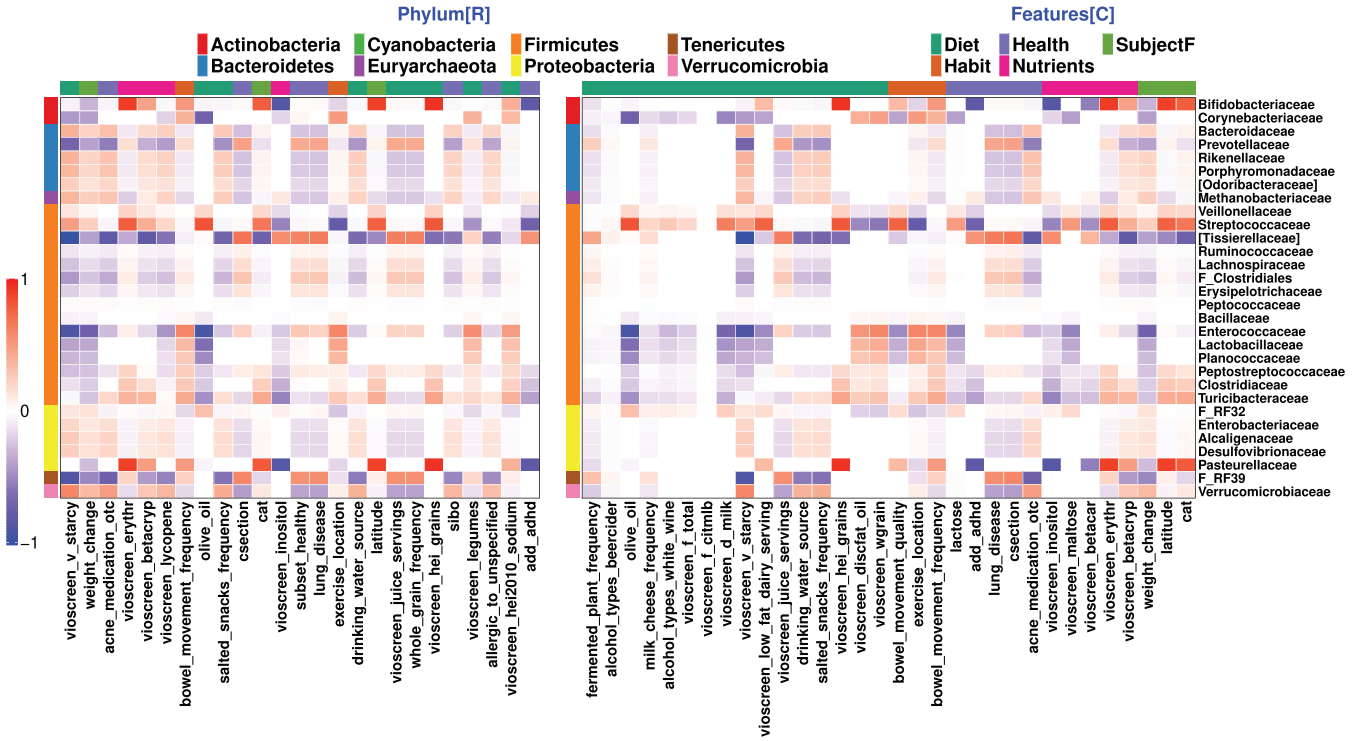

FIGURE 4 Application—AGP: Plots show the selected rows and columns of the coefficient matrix $\mathbf{C}$ based on the estimated effect size. The left plot selects the top 25 covariates based on the row sum of the absolute values of the estimated coefficient matrix, $\sum_{j=1}^{q} |\hat{\mathbf{C}}|_{ij}$. The right plot shows the union of the top 10 covariates selected by each of the three unit-rank components of the estimated $\mathbf{C}$

right plot shows the union of the top 10 covariates selected by each of the three unit-rank components of the estimated $\mathbf{C}$. The color intensity in the plot reflects the effect of covariates on the abundance (red/blue for positive/negative effect). For instance, our analysis suggests that latitude (an indicator of geography) or cat ownership significantly impact the abundance of *Bifidobacteriaceae*, *Pasteurellaceae*, and *Streptococcaceae*. This finding is supported by several studies that also report a strong role of geography[51-53] and pet ownership[54,55] on microbial abundance patterns. Likewise, the consumption of olive oil (unsaturated fatty acids) negatively impacts the abundance of *Corynebacteriaceae* and of several families in the Firmicutes phylum while showing no influence on Bacteroidetes families. Several studies such as De Wit et al,[56] Zhao et al,[57] and Farràs et al[58] have reported associations between olive oil intake and a reduction in *Firmicutes*/*Bacteroidetes* ratio, consistent with the observations here. The NB-FAR model also suggests that *Bifidobacteriaceae* abundances are positively associated with grain intake (*vioscreen_hei_grain*: Healthy Eating Index [HEI] score of total grain), as previously observed.[59] Finally, we also observed several associations between an underlying medical condition and microbial taxa. For instance, attention deficit hyperactivity disorder (ADHD) (last column in left panel, *add_adhd*) negatively impacts the abundance of *Streptococcaceae*, *Bifidobacteriaceae*, and *Pasteurellaceae*. A similar observation about *Streptococcaceae* has been reported in a separate study on children with Autism Spectrum Disorder (showing signs of ADHD).[60]

The orthogonality property of the three latent factors also allows a unique factor-by-factor analysis of the estimated associations. Since each unit-rank component estimate divides response-predictor pairs into two groups, that is, positive and negative, we can cluster each component into exactly four quadrants, essentially providing disentangled bi-clustering of microbial families and host-associated covariates. Figure 5 illustrates these biclusters for each of the three unit-rank components $\{\mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3\}$. For instance, in the $\mathbf{C}_1$ estimate plot, the upper left quadrant shows the negative associations between subsets of covariates and subsets of families. The covariate olive oil intake (column A1) is significantly negatively associated with seven out of the eight taxa, including *Corynebacteriaceae* and *Enterococcaceae*. On the other hand, the covariate exercise location (column A3) is positively associated with these (and other) taxa. Similarly, from the $\mathbf{C}_2$ estimate, we observed that the taxa *Bifidobacteriaceae*, *Pasteurellaceae* and *Streptococcaceae* are positively associated with latitude/geography, Erythritol intake, and grain (columns B1-B3), and negatively associated with inositol level (nutrients) and ADHD (columns B4 and B5). Finally, from the $\mathbf{C}_3$ estimate, we observe that the covariates sets {starchy vegetable, acne medication, salted snack frequency, drinking water source} (columns C1-C3 and C6) and {cesarean birth, lung diseases, juice serving} (columns C4, C5, and C7) have significantly opposite effects on the $\mathbf{C}_3$-associated microbial families.

Overall, these results highlight a strong influence of specific host-associated features on specific family abundance pattern which should be taken into account in downstream analysis whenever these features are available in a microbiome study.

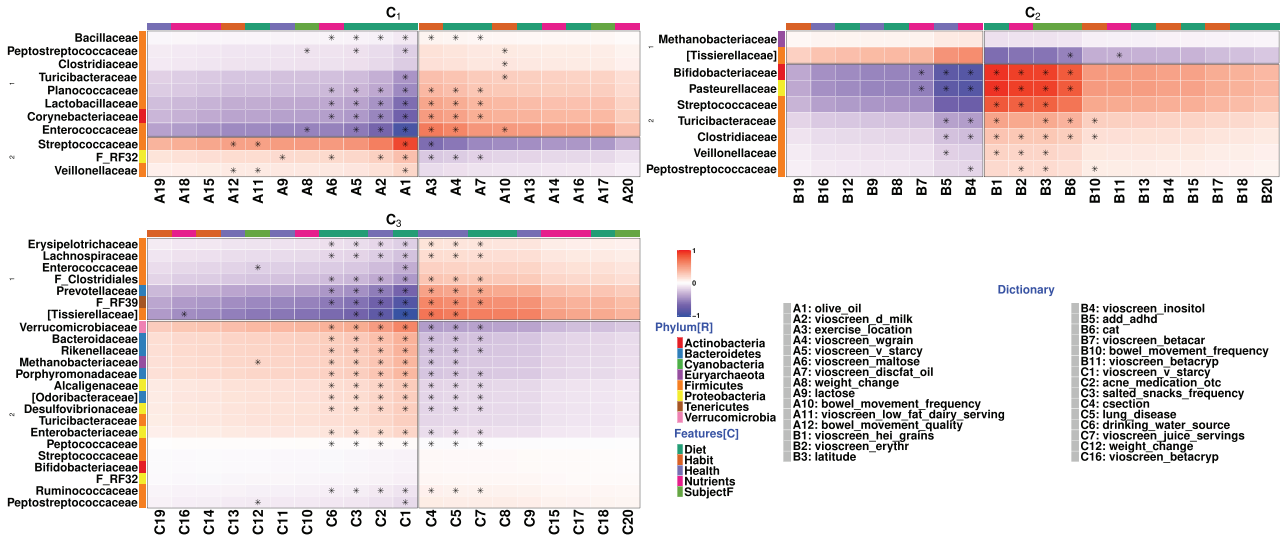

FIGURE 5 Application—AGP: Plots show the selected rows and columns of the estimated unit-rank components, that is, $\{\mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3\}$, of the coefficient matrix $\mathbf{C}$. Covariates selected in the right plot of Figure 4 are marked $\star$ in each of the three components. Each quadrant in the subplots shows the underlying associations

6 | DISCUSSION

In this contribution, we have presented two novel NB factor regression models, NB-FAR and NB-RRR, for the analysis of microbial abundance data. The models have been tailored toward modeling overdispersed count outcome data in the context of amplicon-derived microbiome data. However, we posit that the models may prove useful in other application areas, including clinical trials,[61] sports,[62] and single-cell genomics.[63] The key novelty of the models is to express underlying dependencies between responses (eg, microbial counts) and predictors (eg, host or environmental covariates) by assuming either a dense low-rank or a co-sparse low-rank representation of the coefficient matrix. These structural assumptions appear realistic in the context of microbiome data where certain bacterial taxa are likely specialized in metabolizing specific food ingredients and hence show a concerted diet dependence. Compared to marginal approaches where each of the families is separately modeled using a Poisson or NB regression, we have shown that our models are both computationally more efficient and simultaneously achieve better estimation performance on simulated data and the large-scale AGP data compendium. These results, in turn, challenge recent efforts in establishing host-microbiome relationships using marginal approaches.[33,64] In particular, the study by Manor et al[64] attempted to estimate large-scale association patterns between microbial genera and host features across thousands of participants using marginal logistic or Poisson regression. There, the authors used microbial abundances as predictors and reported the most significant association patterns with individual host covariates as outcome. In particular, they identified the genera Ys2, Ml615j-28, Coriobacteriaceae, Christensenellaceae, Mogibacteriaceae, and Oxalobacter to be significantly associated with host covariates all of which were among the few families that were *not* associated with host covariates in our analysis. These discrepancies highlight the fact that, despite reasonably large sample sizes, there still remain considerable inconsistencies across microbiome studies that require further statistical (meta-)analysis. As we have shown in our application on the AGP data, the ability of the NB-FAR model to deliver crisp bi-clustering of the underlying host-microbiome associations makes them an ideal tool to perform such future analysis and generate testable biological hypotheses.

On the statistical side, potentially fruitful extensions of the NB factor models include the handling of excess zeros in the outcome data using, for example, zero-inflated components or hurdle models. Since there are two types of zeros in the microbial abundance data, true (structural) zeros and experimental (measurement) zeros, one potential path forward is to identify structural zeros a priori[65] and treat the remaining zeros as missing values the latter of which can already be efficiently handled by NB-FAR and NB-RRR. Furthermore, our current approach is restricted to using the *log* link function associating the mean to the linear predictor. Introducing alternative link functions that satisfy the positive mean constraint would add to the flexibility and generality of the current modeling framework. Finally, in our current framework we select the tuning parameter λ via K-fold cross-validation, making the parameter estimation procedure computationally intensive. This can potentially be alleviated by developing a stage-wise algorithm[66] for parameter estimation since such a strategy has been proven to be computationally efficient in the multivariate linear regression setting with normally distributed response matrix $\mathbf{Y}$.

Going forward, we also posit that NB-FAR and NB-RRR may serve as useful sub-routines in more complex statistical analysis workflows, including causal inference. For example, consider a typical randomized clinical trials experiment that aims at understanding the causal effect of a treatment on a phenotype of interest. In a diet intervention study, for instance, it is not unlikely that the intended direct effect on host health is mediated or confounded by the presence of certain microbes in the microbiome. While the instrumental variable (IV) approach[67] provides a powerful framework to uncover causal effects (see also Ailer et al[68] in the context of microbiome data), it requires that the instruments are strong and not confounded. A standard IV approach for continuous data estimates the parameters using two-stage least square. For the high-dimensional data problem, Lin et al[69] proposed a regularized two-stage framework that solves a penalized multivariate linear regression in the first stage and a Lasso problem[38] in the second stage. This framework can be likely extended by using the NB-FAR/NB-RRR methodology in the first stage when overdispersed count data serve as the independent variables.

Taken together, we believe that the introduced NB factor regression models and their efficient implementation in the R package `nbfar` provide a useful statistical framework for analyzing overdispersed count data in medicine, biology, and other scientific disciplines.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Andreas Buja for his comments and suggestions in developing the proposed procedure in the article.

DATA AVAILABILITY STATEMENT

R code for the simulation study and the American Gut application is available as part of the supplementary material. All the supporting files, including the specific subset of the American Gut Project data, are available on GitHub at https://github.com/amishra-stats/nbfar.

ORCID

Aditya K. Mishra https://orcid.org/0000-0003-0775-3219

REFERENCES

1. Ley RE, Lozupone CA, Hamady M, Knight R, Gordon JI. Worlds within worlds: evolution of the vertebrate gut microbiota. *Nat Rev Microbiol*. 2008;6(10):776-788.
2. Ding T, Schloss PD. Dynamics and associations of microbial community types across the human body. *Nature*. 2014;509(7500):357-360.
3. Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. *Nat Rev Genet*. 2012;13(4):260-270.
4. Shreiner AB, Kao JY, Young VB. The gut microbiome in health and in disease. *Curr Opin Gastroenterol*. 2015;31(1):69.
5. Vyas U, Ranganathan N. Probiotics, prebiotics, and synbiotics: gut and beyond. *Gastroenterol Res Pract*. 2012;2012:872716.
6. David LA, Maurice CF, Carmody RN, et al. Diet rapidly and reproducibly alters the human gut microbiome. *Nature*. 2014;505(7484):559-563.
7. Fierer N. Embracing the unknown: disentangling the complexities of the soil microbiome. *Nat Rev Microbiol*. 2017;15(10):579-590.
8. Sunagawa S, Coelho LP, Chaffron S, et al. Structure and function of the global ocean microbiome. *Science*. 2015;348(6237):1261359.
9. Fuhrman JA, Cram JA, Needham DM. Marine microbial community dynamics and their ecological interpretation. *Nat Rev Microbiol*. 2015;13(3):133-146.
10. Bolyen E, Rideout JR, Dillon MR, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol*. 2019;37(8):852-857.
11. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: high-resolution sample inference from Illumina amplicon data. *Nat Methods*. 2016;13(7):581-583.
12. Gonzalez A, Navas-Molina JA, Kosciolek T, et al. Qiita: rapid, web-enabled microbiome meta-analysis. *Nat Methods*. 2018;15(10):796-798.
13. McDonald D, Hyde E, Debelius JW, et al. American Gut: an open platform for citizen science microbiome research. *Msystems*. 2018;3(3):e00031-e00018.
14. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The human microbiome project. *Nature*. 2007;449(7164):804-810.
15. Gilbert JA, Jansson JK, Knight R. The earth microbiome project: successes and aspirations. *BMC Biol*. 2014;12(1):69.
16. Vujkovic-Cvijin I, Sklar J, Jiang L, Natarajan L, Knight R, Belkaid Y. Host variables confound gut microbiota studies of human disease. *Nature*. 2020;587(7834):448-454. doi:10.1038/s41586-020-2881-9
17. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):1-21.
18. Badri M, Kurtz ZD, Bonneau R, Müller CL. Shrinkage improves estimation of microbial associations under different normalization methods. *NAR Genom Bioinform*. 2020;2(4):lqaa100.
19. McMurdie PJ, Holmes S. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput Biol*. 2014;10(4):e1003531.
20. Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput Biol*. 2015;11(5):e1004226.
21. Xu L, Paterson AD, Turpin W, Xu W. Assessment and selection of competing models for zero-inflated microbiome data. *PLoS One*. 2015;10(7):1-30. doi:10.1371/journal.pone.0129606
22. Holmes I, Harris K, Quince C. Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PLoS One*. 2012;7(2):e30126. doi:10.1371/journal.pone.0030126
23. Sankaran K, Holmes SP. Latent variable modeling for the microbiome. *Biostatistics*. 2019;20(4):599-614.
24. Deek RA, Li H. A zero-inflated latent dirichlet allocation model for microbiome studies. *Front Genet*. 2021;11(January):1-10. doi:10.3389/fgene.2020.602594
25. Xia Y, Sun J, Chen DG. Modeling zero-inflated microbiome data. *Statistical Analysis of Microbiome Data with R*. Springer; 2018:847.
26. Lee S, Chugh PE, Shen H, Eberle R, Dittmer DP. Poisson factor models with applications to non-normalized microRNA profiling. *Bioinformatics*. 2013;29(9):1105-1111.
27. Sohn MB, Li H. A GLM-based latent variable ordination method for microbiome samples. *Biometrics*. 2018;74(2):448-457.
28. Xu T, Demmer RT, Li G. Zero-inflated Poisson factor model with application to microbiome read counts. *Biometrics*. 2021;77(1):91-101.
29. Breuninger TA, Wawro N, Breuninger J, et al. Associations between habitual diet, metabolic disease, and the gut microbiota using latent Dirichlet allocation. *Microbiome*. 2021;9(1):1-18. doi:10.1186/s40168-020-00969-9
30. Wadsworth WD, Argiento R, Guindani M, Galloway-Pena J, Shelbourne SA, Vannucci M. An integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. *BMC Bioinform*. 2017;18(1):1-12. doi:10.1186/s12859-017-1516-0

31. Wadsworth WD, Argiento R, Guindani M, Galloway-Pena J, Shelbourne SA, Vannucci M. scCODA is a Bayesian model for compositional single-cell data analysis. *Nat Commun*. 2021;12(1):1-10.
32. Ostner J, Carcy S, Müller CL. tascCODA: Bayesian tree-aggregated analysis of compositional amplicon and single-cell data. *Front Genet*. 2021;12:766405.
33. Zhang X, Mallick H, Tang Z, et al. Negative binomial mixed models for analyzing microbiome count data. *BMC Bioinform*. 2017;18(1):1-10. doi:10.1186/s12859-016-1441-7
34. Mishra A, Dey DK, Chen K. Sequential co-sparse factor regression. *J Comput Graph Stat*. 2017;26(4):814-825.
35. Anderson TW. Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Ann Math Stat*. 1951;22(3):327-351.
36. Reinsel GC, Velu P. *Multivariate Reduced-Rank Regression: Theory and Applications*. Springer Science & Business Media; 2013:136.
37. Bunea F, She Y, Wegkamp M. Optimal selection of reduced rank estimators of high-dimensional matrices. *Ann Stat*. 2011;39(2):1282-1309.
38. Tibshirani RJ. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B*. 1996;58:267-288.
39. Chen K, Chan KS, Stenseth NC. Reduced rank stochastic regression with a sparse singular value decomposition. *J Royal Stat Soc Ser B (Stat Methodol)*. 2012;74(2):203-221.
40. Chen L, Huang JZ. Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *J Am Stat Assoc*. 2012;107(500):1533-1545.
41. Bunea F, She Y, Wegkamp M. Joint variable and rank selection for parsimonious estimation of high dimensional matrices. *Ann Stat*. 2012;40(5):2359-2388.
42. Luo C, Liang J, Li G, et al. Leveraging mixed and incomplete outcomes via reduced-rank modeling. *J Multivar Anal*. 2018;167:378-394.
43. Mishra A, Dey DK, Chen Y, Chen K. Generalized co-sparse factor regression. *Comput Stat Data Anal*. 2020;157:107127.
44. Zeileis A, Kleiber C, Jackman S. Regression models for count data in R. *J Stat Softw*. 2008;27(8):1-25.
45. Anders S, Huber W. Differential expression analysis for sequence count data. *Nat Preced*. 2010:1-1.
46. Jorgensen B. Exponential dispersion models. *J R Stat Soc Ser B (Methodol)*. 1987;49(2):127-145.
47. Lin H, Peddada SD. Analysis of microbial compositions: a review of normalization and differential abundance analysis. *NPJ Biofilms Microbiomes*. 2020;6(1):1-13.
48. Sun Y, Babu P, Palomar DP. Majorization-minimization algorithms in signal processing, communications, and machine learning. *IEEE Trans Signal Process*. 2016;65(3):794-816.
49. Stone M. Cross-validation and multinomial prediction. *Biometrika*. 1974;61(3):509-515. doi:10.1093/biomet/61.3.509
50. Zou H, Zhang HH. On the adaptive elastic-net with a diverging number of parameters. *Ann Stat*. 2009;37(4):1733.
51. Yatsunenko T, Rey FE, Manary MJ, et al. Human gut microbiome viewed across age and geography. *Nature*. 2012;486(7402):222-227.
52. Gupta VK, Paul S, Dutta C. Geography, ethnicity or subsistence-specific variations in human microbiome composition and diversity. *Front Microbiol*. 2017;8:1162.
53. He Y, Wu W, Zheng HM, et al. Regional variation limits applications of healthy gut microbiome reference ranges and disease models. *Nat Med*. 2018;24(10):1532-1535.
54. Nermes M, Niinivirta K, Nylund L, et al. Perinatal pet exposure, faecal microbiota, and wheezy bronchitis: is there a connection? *Int Sch Res Not*. 2013;2013:827934.
55. Song SJ, Lauber C, Costello EK, et al. Cohabiting family members share microbiota with one another and with their dogs. *elife*. 2013;2:e00458.
56. De Wit N, Derrien M, Bosch-Vermeulen H, et al. Saturated fat stimulates obesity and hepatic steatosis and affects gut microbiota composition by an enhanced overflow of dietary fat to the distal intestine. *Am J Physiol Gastrointestinal Liver Physiol*. 2012;303(5):G589-G599.
57. Zhao Z, Shi A, Wang Q, Zhou J. High oleic acid peanut oil and extra virgin olive oil supplementation attenuate metabolic syndrome in rats by modulating the gut microbiota. *Nutrients*. 2019;11(12):3005.
58. Farràs M, Martinez-Gili L, Portune K, et al. Modulation of the gut microbiota by olive oil phenolic compounds: implications for lipid metabolism, immune system, and obesity. *Nutrients*. 2020;12(8):2200.
59. Jang HB, Choi MK, Kang JH, Park SI, Lee HJ. Association of dietary patterns with the fecal microbiota in Korean adolescents. *BMC Nutrition*. 2017;3(1):20.
60. Liu S, Li E, Sun Z, et al. Altered gut microbiota and short chain fatty acids in Chinese children with autism spectrum disorder. *Sci Rep*. 2019;9(1):1-9.
61. Byers AL, Allore H, Gill TM, Peduzzi PN. Application of negative binomial modeling for discrete outcomes: a case study in aging research. *J Clin Epidemiol*. 2003;56(6):559-564. doi:10.1016/S0895-4356(03)00028-3
62. Bittner E, Nußbaumer A, Janke W, Weigel M. Football fever: goal distributions and non-Gaussian statistics. *Eur Phys J B*. 2009;67(3):459-471. doi:10.1140/epjb/e2008-00396-1
63. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol*. 2019;20(1):1-15.
64. Manor O, Dai CL, Kornilov SA, et al. Health and disease markers correlate with gut microbiome composition across thousands of people. *Nat Commun*. 2020;11(1):1-12. doi:10.1038/s41467-020-18871-1
65. Kaul A, Davidov O, Peddada SD. Structural zeros in high-dimensional data with applications to microbiome studies. *Biostatistics*. 2017;18(3):422-433. doi:10.1093/biostatistics/kxw053
66. Chen K, Dong R, Xu W, Zheng Z. Statistically guided divide-and-conquer for sparse factorization of large matrix; 2020. arXiv preprint arXiv:2003.07898.

67. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *J Am Stat Assoc*. 1996;91(434):444-455.
68. Ailer E, Müller CL, Kilbertus N. A causal view on compositional data; 2021. arXiv preprint arXiv:2106.11234.
69. Lin W, Feng R, Li H. Regularization methods for high-dimensional instrumental variables regression with an application to genetical genomics. *J Am Stat Assoc*. 2015;110(509):270-288.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Mishra AK, Müller CL. Negative binomial factor regression with application to microbiome data analysis. *Statistics in Medicine*. 2022;1-18. doi: 10.1002/sim.9384