Biometrika (2022), **109**, 2, *pp.* 405–420 doi: 10.1093/biomet/asab020
Advance Access publication 19 August 2021

High-dimensional log-error-in-variable regression with applications to microbial compositional data analysis

By PIXU SHI

Department of Biostatistics & Bioinformatics, Duke University, 2424 Erwin Road, Durham, North Carolina 27710, U.S.A.
pixu.shi@duke.edu

YUCHEN ZHOU and ANRU R. ZHANG

Department of Statistics, University of Wisconsin-Madison, 1300 University Avenue, Madison, Wisconsin 53706, U.S.A.
yuchenzhou@stat.wisc.edu anruzhang@stat.wisc.edu

Summary

In microbiome and genomic studies, the regression of compositional data has been a crucial tool for identifying microbial taxa or genes that are associated with clinical phenotypes. To account for the variation in sequencing depth, the classic log-contrast model is often used where read counts are normalized into compositions. However, zero read counts and the randomness in covariates remain critical issues. We introduce a surprisingly simple, interpretable and efficient method for the estimation of compositional data regression through the lens of a novel high-dimensional log-error-in-variable regression model. The proposed method provides corrections on sequencing data with possible overdispersion and simultaneously avoids any subjective imputation of zero read counts. We provide theoretical justifications with matching upper and lower bounds for the estimation error. The merit of the procedure is illustrated through real data analysis and simulation studies.

Some key words: Compositional data; Error-in-variable; High-dimensional regression; Microbiome study.

1. Introduction

High-dimensional regression has attracted enormous attention in contemporary statistical research. The canonical model of high-dimensional regression can be written as $y = X\beta^* + \varepsilon$, where $y = (y_1, \ldots, y_n)^{\mathrm{T}}$ is the response vector, $X \in \mathbb{R}^{n\times p}$ is the covariate matrix and $\beta^* \in \mathbb{R}^p$ is the unknown coefficient vector of interest. Most prior work has focused on the clean data case where the covariates are accurately observed. However, in applications of econometrics, genomics and engineering, we also frequently see covariates corrupted with noise. Previous literature referred to such scenarios as error-in-variable, and showed that performing standard regression methods directly on the corrupted covariates may yield inaccurate inference results (Hausman, 2001). When the observable covariates are corrupted by additive Gaussian or sub-Gaussian noise, the methods and theories for error-in-variable regression have been widely considered in both the classic low-dimensional setting (Deming, 1943), and more recently in the high-dimensional

setting (Rosenbaum & Tsybakov, 2010; Loh & Wainwright, 2012; Belloni et al., 2017; Datta & Zou, 2017).

The focus of high-dimensional error-in-variable regression has so far mainly been on homoscedastic Gaussian, sub-Gaussian or bounded corruption settings. Motivated by applications in high-throughput sequencing in microbiome studies, we consider here a new framework of high-dimensional error-in-variable regression that adapts to compositional covariates.

The human microbiome is the aggregate of all microbes that reside on human bodies. It has attracted enormous recent attention due to its strong tie with human health (The Human Microbiome Project Consortium, 2012). For example, recent studies found that the human microbiome may be closely related with various diseases, such as cancer (Schwabe & Jobin, 2013) and obesity (Turnbaugh et al., 2006). Modern next-generation sequencing technologies, such as 16S ribosomal RNA and shotgun metagenomics sequencing, provide quantification of the human microbiome by performing direct sequencing on either whole metagenomes or individual marker genes. By aligning sequencing reads to referential microbial genomes, we can organize the sequencing data into a count matrix with rows representing samples and columns representing microbial taxa or genes. Such data can be seen as the random realization of the relative abundance of bacteria in each sample.

To account for the difference in sequencing depth across samples, the read counts are often normalized into compositions; see (Li, 2015) for a survey, and the references herein. The resulting data, also called compositional data, pose statistical challenges due to the collinearity and nonnormality that come from their compositional nature. To address these issues, Aitchison & Bacon-Shone (1984) introduced the log-contrast model:

$$y_i = \sum_{j=1}^{p-1} \log(Z_{ij}/Z_{ip})\beta_j^* + \varepsilon_i \qquad (i = 1, \ldots, n). \tag{1}$$

Here, W_{ij} and $Z_{ij} = W_{ij}/(\sum_{j'=1}^{p} W_{ij'})$ are respectively the absolute count and the relative abundance of the jth component, e.g., bacterial gene or taxon, in the ith sample. The analysis of the log-contrast model (1) is often dependent on the choice of reference component Z_{ip}, especially in high-dimensional settings. Thus, Lin et al. (2014) reformulated (1) by introducing $\beta_p^* = -\sum_{j=1}^{p-1} \beta_j^*$,

$$y_i = \sum_{j=1}^{p} \log(Z_{ij})\beta_j^* + \varepsilon_i \quad (i = 1, \ldots, n) \qquad \text{subject to } \sum_{j=1}^{p} \beta_j^* = 0, \tag{2}$$

and proposed estimating β^* through the constrained l_1 regularized estimator. More recently, Shi et al. (2016) studied the statistical inference and confidence intervals for β^*, and Wang & Zhao (2017) considered the subcomposition selection in compositional data regression via a tree-guided regularization method.

The direct application of (1) and (2) by normalizing sequencing read counts, i.e., using $Z_{ij} = W_{ij}/(\sum_{j=1}^{p} W_{ij})$ as covariates, has several drawbacks. First, it ignores the fact that the Z_{ij} are random realizations rather than true compositions of the components. In next-generation sequencing data, Z_{ij} is the proportion of the read count of component j among all components in sample i, and is thus a transformation of discrete random variables that reflect the underlying true compositions with measurement errors. As mentioned earlier, overlooking the measurement error in regressors may lead to inaccurate results. By treating Z_{ij} as the true compositions, it is also overlooking the heteroskedasticity or overdispersion of Z_{ij} caused by enormous uncontrollable factors of variation in sequencing, e.g., time, sampling location or technical variability (Chen & Li, 2013). Second, the procedure requires $Z_{ij} > 0$, while in reality compositional data

from next-generation sequencing often contain a lot of zeros due to the rarity of certain components. Strategies to deal with the zeros include replacing zero counts by a subjectively chosen small number, such as 0.5, before normalizing counts into compositions (Martin-Fernandez et al., 2000), or imputing the entire composition matrix (Cao et al., 2020) based on a low-rank assumption. However, to the best of our knowledge, there is still no consensus on the best approach to deal with zero read counts in compositional data regression.

To address the aforementioned challenges in compositional data regression, we introduce a high-dimensional log-error-in-variable regression model that directly handles count covariates without normalization into compositions or imputation of zeros. Recall that W_{ij} is the count of the jth component in the ith sample. We assume $W_i = (W_{i1}, \ldots, W_{ip})^{\mathrm{T}}$ follows the Dirichlet-multinomial distribution (Mosimann, 1962) given the total count $N_i = \sum_{j=1}^{p} W_{ij}$ in the ith sample, and $N_i \sim \mathrm{Po}(\nu_i)$ to account for the randomness of sequencing depth. That is,

$$\begin{aligned} &\mathrm{pr}\left\{(W_{i1}, \ldots, W_{ip}) = (k_{i1}, \ldots, k_{ip}) \mid (N_i, q_{i1}, \ldots, q_{ip})\right\} = N_i!(k_{i1}! \cdots k_{ip}!)^{-1} \textstyle\prod_{j=1}^{p} q_{ij}^{k_{ij}}, \\ &f(q_{i1}, \ldots, q_{ip} \mid N_i) = B(\alpha_i X_{i1}, \ldots, \alpha_i X_{ip})^{-1} \textstyle\prod_{j=1}^{p} q_{ij}^{\alpha_i X_{ij} - 1}, \end{aligned} \tag{3}$$

where $\sum_{j=1}^{p} k_{ij} = N_i$, $k_{i1}, \ldots, k_{ip} \in \{0, 1, 2, \ldots\}$, and $\sum_{j=1}^{p} q_{ij} = 1$, $q_{i1}, \ldots, q_{ip} \geqslant 0$. Here, $X_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}}$ is the underlying true composition of the p components, $B(\alpha_i X_{i1}, \ldots, \alpha_i X_{ip}) = \{\prod_{j=1}^{p} \Gamma(\alpha_i X_{ij})\}/\Gamma(\alpha_i)$ is the Beta function and α_i is the overdispersion parameter of the subject from which the ith sample is measured. The Dirichlet-multinomial distribution is a standard assumption and has been commonly used to model the multivariate count datasets with overdispersion. See, for example, Holmes et al. (2012), La Rosa et al. (2012), Chen & Li (2013), Mandal et al. (2015), Wadsworth et al. (2017) and Dai et al. (2019). The Dirichlet-multinomial model has also been used in applications of econometrics (Guimaraes & Lindrooth, 2007), single-cell mRNA studies (Qiu et al., 2017) and text mining (Yin & Wang, 2014), amongst others. When $\alpha_i \to +\infty$, the Dirichlet-multinomial distribution degenerates to the regular multinomial distribution.

Since the observable count, W_i, is merely a realization of the underlying composition, X_i, it is more reasonable to assume association between y_i and X_i rather than between y_i and W_i. We thus assume the regression response y_i to be dependent on X_i through the following log-contrast model:

$$y_i = \textstyle\sum_{j=1}^{p} \log(X_{ij})\beta_j^* + \varepsilon_i \quad (i = 1, \ldots, n) \qquad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^* = 0. \tag{4}$$

We refer to (3) together with (4) as the log-error-in-variable regression model. Our aim is to estimate β^* based on responses $y \in \mathbb{R}^n$ and error-in-covariates $W \in \mathbb{R}^{n \times p}$. Most of the results on error-in-variables regression deal with homoscedastic continuous variables, and may not be directly applied here since the W_i are discrete random variables with heteroscedasticity depending on X_i and α_i. Therefore, new methods are required.

In this paper we propose a surprisingly simple and straightforward estimation scheme, named variable correction regularized estimator, for high-dimensional log-error-in-variable regression. In particular, when the count observations are without overdispersion, we propose to add 0.5 to all counts W_{ij}, then estimate the regression parameters using constrained lasso with $\log(W_{ij} + 0.5)$ as predictors; for overdispersed data, we propose to add an amount related to the overdispersion level to W_{ij} to alleviate the effect of any overly large or small counts due to overdispersion. In practice, we recommend adding 0.5 to all counts if quantification of the overdispersion level is difficult, for example when there is no repeated measurement or the Dirichlet-multinomial distribution is largely violated.

In addition, we further generalize the proposed variable correction scheme in a more general log-error-in-variable regression model in § 4:

$$y_i = \sum_{j=1}^{p} \log(\nu_{ij})\beta_j^* + \varepsilon_i \quad (i = 1, \ldots, n) \qquad \text{subject to} \qquad \sum_{j=1}^{p} \beta_j^* = 0. \tag{5}$$

Here, the observed y_i and independent random covariates $W_{ij} \geqslant 0$ are linked through $\log(\nu_{ij})$, where $\nu_{ij} = \mathbb{E}W_{ij}$ does not need to form compositions, β^* is the parameter of interest and ε_i is independent and identically distributed sub-Gaussian noise with mean zero and variance σ^2. We prove that the +0.5 variable correction regularized estimator for β^* in (5) achieves good performance.

2. METHODS FOR LOG-ERROR-IN-VARIABLE REGRESSION

To estimate β^* in (3) and (4), one classic method is simple normalization, i.e., using W_{ij}/N_i as a surrogate for X_{ij} and implementing the classic high-dimensional regularized estimators with $\log(W_{ij}/N_i)$ as covariates. As discussed earlier, this idea has two critical issues. First, the zero-valued W_{ij} need to be replaced by a small value to make them positive in the log transformation. The choice of this value is often difficult, but critical to the performance of the final estimates. Second, even though $\mathbb{E}(W_{ij}/N_i \mid N_i) = X_{ij}$, $\log(W_{ij}/N_i)$ may be a biased estimator for $\log(X_{ij})$, which can cause additional inaccuracy in the regression analysis. To further illustrate the biased nature of $\log(W_{ij}/N_i)$ and to introduce our fixing plan, we first focus on the non-overdispersion case, i.e., $\alpha_i = +\infty$ in (3), or equivalently $(W_{i1}, \ldots, W_{ip}) \mid N_i \sim \mathrm{Mu}(N_i, X_{i1}, \ldots, X_{ip})$. In this case, W_{ij} follows $\mathrm{Po}(\nu_i X_{ij})$ and $E(W_{ij}) = \mathrm{var}(W_{ij}) = \nu_i X_{ij}$. For any $z_i \geqslant 0$, the Taylor expansion of $\log(W_{ij} + z_i)$ at $\nu_i X_{ij}$ yields the following approximation:

$$\begin{aligned} E\{\log(W_{ij} + z_i)\} &\approx \log(\nu_i X_{ij}) + \frac{E(W_{ij} - \nu_i X_{ij} + z_i)}{\nu_i X_{ij}} \\ &\quad - \frac{\mathrm{var}(W_{ij}) + 2z_i E(W_{ij} - \nu_i X_{ij}) + z_i^2}{2\nu_i^2 X_{ij}^2} \\ &= \log(\nu_i X_{ij}) + \frac{z_i - 1/2}{\nu_i X_{ij}} - \frac{z_i^2}{2\nu_i^2 X_{ij}^2}. \end{aligned}$$

Since N_i is the total number of reads in sample i and is generally large in practice, e.g., around 10^4 to 10^5 in our real data example, we can assume $\nu_i = EN_i$ to be large. Then,

$$\begin{aligned} E\left\{\log\left(\frac{W_{ij} + z_i}{N_i}\right)\right\} - \log(X_{ij}) &\approx \frac{z_i - 1/2}{\nu_i X_{ij}} - \frac{z_i^2}{2\nu_i^2 X_{ij}^2} + \log(\nu_i) - E\log(N_i) \\ &\approx \frac{z_i - 1/2}{\nu_i X_{ij}} - \frac{z_i^2}{2\nu_i^2 X_{ij}^2}. \end{aligned}$$

We can see that the bias of $\log\{(W_{ij}+z_i)/N_i\}$ for estimating $\log(X_{ij})$ is approximately $-(1/2)\nu_i X_{ij}$ and $-(1/8)\nu_i^2 X_{ij}^2$ when $z_i = 0$ and $z_i = 1/2$, respectively. For large ν_i, one has $(1/8)\nu_i^2 X_{ij}^2 \ll (1/2)\nu_i X_{ij}$. Therefore, heuristically $\log\{(W_{ij} + 1/2)/N_i\}$ is a significantly less biased estimator

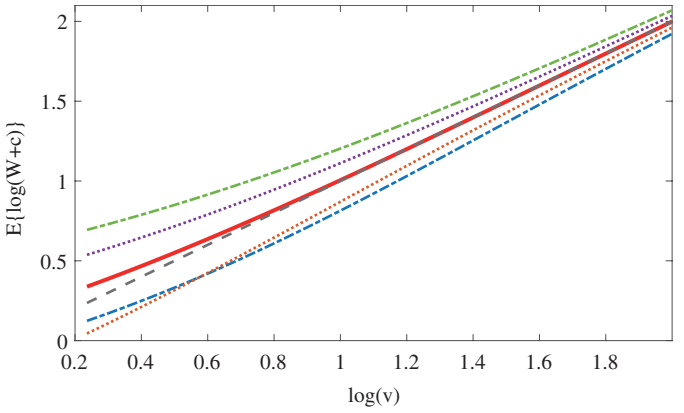

Fig. 1. $\log(\nu)$ versus $E\log(W \vee 0.5)$ (zero-replace, blue dashed) and $E\log(W + c)$ for $c = 1/4$ (red dotted), $1/2$ (red solid), $3/4$ (purple dotted), 1 (green dashed). Here, $W \sim \text{Po}(\nu)$.

for $\log(X_{ij})$ compared to $\log(W_{ij}/N_i)$, or $\log\{(W_{ij} + c)/N_i\}$ for any $c \neq 1/2$. Figure 1 illustrates the bias of $\log(W \vee 0.5)$, i.e., replacing zeros by $1/2$, and $\log(W + c), c = 1/4, 1/2, 3/4, 1$ for estimating $\log(\nu)$ when W follows $\text{Po}(\nu)$. The plot suggests that $\log(W + 1/2)$ achieves the minimum bias among these choices. In addition, by adding the positive value $1/2$ to all W_{ij}, the previously mentioned zero-replacement issue is simultaneously solved!

To account for higher variability in the count data, we also consider the overdispersed case where $\alpha_i < \infty$ in (3). In this case, we have

$$E(W_{ij}) = \nu_i X_{ij}, \qquad \text{var}(W_{ij}) = \nu_i X_{ij}(1 - X_{ij})(\nu_i + \alpha_i + 1)/(\alpha_i + 1) + \nu_i X_{ij}^2.$$

Similarly, by investigating the Taylor expansion of $E[\log\{(W_{ij} + z_i)/N_i\}]$, it can be shown that taking $z_i = (N_i + \alpha_i + 1)/\{2(\alpha_i + 1)\}$ will make $\log\{(W_{ij} + z_i)/N_i\}$ a better estimator for $\log(X_{ij})$; a more rigorous argument is postponed to the Supplementary Material. It is noteworthy that z_i is an estimate for half of $(\nu_i + \alpha_i + 1)/(\alpha_i + 1)$, which quantifies the overdispersion rate of W_{ij} compared with the multinomial distribution.

These heuristic arguments inspire us to the following variable correction regularized estimator for the log-error-in-variable regression in (3) and (4):

$$\hat{\beta} = \arg\min_{\beta} \left\{ \|y - B_W \beta\|_2^2/(2n) + \lambda \|\beta\|_1 \right\} \qquad \text{subject to} \quad \textstyle\sum_{j=1}^{p} \beta_j = 0, \tag{6}$$

where $B_W \in \mathbb{R}^{n \times p}$ and $(B_W)_{ij} = \log\left\{W_{ij} + (N_i + \alpha_i + 1)/(2\alpha_i + 2)\right\}$. Particularly if $\alpha_i = \infty$, i.e., $W_{i\cdot} \mid N_i$ satisfies the regular multinomial, $(B_W)_{ij} = \log(W_{ij} + 1/2)$.

Remark 1. Different from the classic zero-replacement scheme that replaces only the zero covariates by a fixed value, we propose to add $1/2$ to all covariates in the non-overdispersion case. For an overdispersed sample, we propose to correct W_{ij} with a larger value: $z_i = (N_i + \alpha_i + 1)/\{2(\alpha_i + 1)\}$. In particular, with larger total count N_i or larger degree of overdispersion, i.e., smaller α_i, the observable count covariate W_i contains noisier information about the true underlying composition X_i. The larger added values can alleviate the effect of overly large or small counts due to overdispersion.

When there is evidence of overdispersion and there are multiple samples W_i that share the same X_i and α_i in practice, α_i can be estimated by the method of moments (La Rosa et al., 2012)

or the maximum likelihood estimator (Tvedebrink, 2010). Otherwise, $\alpha_i = +\infty$ and $z_i = 1/2$ are suggested. More detailed discussions on the method of moment estimator of α_i are given in the Supplementary Material.

Remark 2. In the existing methods for high-dimensional error-in-variable regression, it is often a key step to construct good estimators for both the covariate V and the Gram matrix $V^{\mathrm{T}}V$, e.g., Loh & Wainwright (2012), Rosenbaum & Tsybakov (2013), Belloni et al. (2017), Datta & Zou (2017) and Rudelson & Zhou (2017). Even though our construction targets $(B_W)_{ij}$, a nearly unbiased estimator for $V_{ij} = \log(\nu_{ij})$, we can further show that $B_W^{\mathrm{T}}B_W$ is also a good estimator of $V^{\mathrm{T}}V$ based on some key properties of the log-error-in-variable model.

Remark 3. The proposed variable correction scheme requires knowledge of N_i, i.e., the total count from each subject. Due to the mechanism of sequencing techniques in microbiome studies, the raw data are usually counts as opposed to the normalized compositions, and the total number of sequencing reads N_i is usually available; N_i is also available in a wide range of applications, e.g., single-cell sequencing data analysis (Navin et al., 2011) and text mining. In addition, if the observation comes as a composition, but the total count N_i is not available, as long as the distribution of error-in-variable, i.e., $W_{ij} \mid X_{ij}$, can be parameterized as $\mathbb{P}_{X_{ij}}$ and $\mathbb{P}_{X_{ij}}$ is known for given X_{ij}, the compositional regression problem may be addressed by the general high-dimensional log-error-in-variable regression model discussed in § 5.

3. THEORETICAL ANALYSIS

We now investigate the theoretical performance of the proposed variable correction regularized estimator for the log-error-in-variable regression model. Denote $\bar{\nu} = \sum_{i=1}^{n} \nu_i/n$ and $\nu = (\nu_1, \ldots, \nu_n)^{\mathrm{T}}$. We say a matrix M satisfies the restricted isometry property (Candès & Tao, 2007) with constant $\delta_s(M) \in (0, 1)$ if

$$n\{1 - \delta_s(M)\}\|\beta\|_2^2 \leqslant \|M\beta\|_2^2 \leqslant n\{1 + \delta_s(M)\}\|\beta\|_2^2 \qquad \forall\, s\text{-sparse vectors } \beta. \tag{7}$$

The restricted isometry property (7) is commonly used in the high-dimensional regression literature. Recall the corrected design matrix $B_W \in \mathbb{R}^{n\times p}$ and $(B_W)_{ij} = \log\big\{W_{ij} + (N_i + \alpha_i + 1)/(2\alpha_i + 2)\}$. We assume the centralized $\bar{B}_W = B_W\{I_p - (1/p)1_p 1_p^{\mathrm{T}}\}$ satisfies the following condition.

Condition 1. The centralized design matrix $\bar{B}_W$ satisfies the restricted isometry property with constant $\delta_{2s}(\bar{B}_W) < 1/10$ with probability $1 - \epsilon'$ for some small quantity $\epsilon' > 0$.

We first consider the case where the observable counts have no overdispersion, i.e., $\alpha = \infty$. We show that the proposed variable correction regularized estimator (6) satisfies the following upper bound.

THEOREM 1 (NO OVERDISPERSION, UPPER BOUND). *Consider* (3) *and* (4) *with* $\alpha_i = \infty$, *i.e., W has no overdispersion. Suppose Condition* 1 *holds,* $n \geqslant Cs\log(p)$, *and* $a\bar{\nu} \leqslant \nu_i \leqslant b\bar{\nu}$, $a/p \leqslant X_{ij} \leqslant b/p$ *for constants* $0 < a < 1 < b$. *If, for some large constant* $C > 0$, *some* $\epsilon > 0$ *and a constant* C_ϵ *that only depends on* ϵ, *we have* $(\bar{\nu}/p) \geqslant C(s + \log(np) + C_\epsilon)$, *then, by choosing* $\lambda = C[\{\log(p)/n\}\{\sigma^2 + (p/\bar{\nu})\|\beta^*\|_2^2\} + s\,(p/\bar{\nu})^{3-2\epsilon}\,\|\beta^*\|_2^2]^{1/2}$ *for some large constant*

$C > 0$, the variable correction regularized estimator (6) satisfies

$$\|\hat{\beta} - \beta^*\|_2^2 \leqslant C\left[(s\log p/n)\left\{\sigma^2 + (p/\bar{\nu})\|\beta^*\|_2^2\right\} + s^2\,(p/\bar{\nu})^{3-2\epsilon}\,\|\beta^*\|_2^2\right] \tag{8}$$

with probability at least $1 - 4p^{-C'} - \epsilon'$. Moreover, if $\bar{\nu} \geqslant p(sn/\log p)^{1/(2-2\epsilon)}$, then, by choosing $\lambda = C\left[\{\log(p)/n\}\left\{\sigma^2 + (p/\bar{\nu})\|\beta^*\|_2^2\right\}\right]^{1/2}$ for constant $C > 0$, with probability at least $1 - 4p^{-C'} - \epsilon'$,

$$\|\hat{\beta} - \beta^*\|_2^2 \leqslant C(s\log p/n)\left\{\sigma^2 + (p/\bar{\nu})\|\beta^*\|_2^2\right\}. \tag{9}$$

Proofs of all the theorems are provided in the Supplementary Material.

Remark 4. Theorem 1 shows that the estimation error gets smaller with larger sample size n, smaller dimension p, smaller noise variance σ^2, higher $\bar{\nu}$, smaller signal amplitude $\|\beta^*\|_2^2$ or smaller sparsity level s. If $\bar{\nu}$ is large, a sufficient sample size to ensure consistency of β is $n \geqslant C\max\{\sigma^2, 1\}s\log p$, which matches the classic result in high-dimensional sparse linear regression (Bickel et al., 2009). In addition, the error bound (8) includes two components: $C(s\log p/n)\sigma^2$ corresponds to the error of ε_i and $C\{(s\log p/n)(p/\bar{\nu})\|\beta^*\|_2^2 + s^2\,(p/\bar{\nu})^{3-2\epsilon}\,\|\beta^*\|_2^2\}$ is incurred by the error in covariates.

An important factor in both the condition and upper bound in Theorem 1 is $\bar{\nu}$, as $\bar{\nu} = \sum_{i=1}^n \nu_i/n$, $\nu_i = \mathbb{E}\sum_{j=1}^p W_{ij}$, $(\bar{\nu}/p)$ quantifies the average count of all components among all subjects. In practice, such as in microbiome studies, it is reasonable to assume $(\bar{\nu}/p)$ is a constant or logarithmic in scale as p increases, because when we investigate bacteria with higher resolution at lower taxonomic levels, the average sequencing depth quantified by $\bar{\nu}$ should also be increased accordingly to maintain the accuracy of the analysis.

Let $\bar{V} = \left\{\log(\nu_i X_{ij}) - p^{-1}\sum_{l=1}^p \log(\nu_i X_{il})\right\}_{1\leqslant i\leqslant n; 1\leqslant j\leqslant p}$ be the centralized population design matrix. We further consider the following class of covariate matrices and parameter vectors,

$$\mathcal{F}_{p,n,s}(R,Q) = \big\{(\nu, X, \beta) : a\bar{\nu} \leqslant \nu_i \leqslant b\bar{\nu}, a/p \leqslant X_{ij} \leqslant b/p \text{ for constants } 0 < a < 1 < b;$$
$$\delta_{2s}(\bar{V}) < 1/20, ||\beta||_2 \leqslant R, 1_p^{\mathrm{T}}\beta = 0, e^{-3/2}Q \leqslant \bar{\nu} \leqslant e^{3/2}Q\big\}. \tag{10}$$

The constraints in $\mathcal{F}_{p,n,s}(R,Q)$ correspond to the regularization assumptions in Theorem 1. The upper bound in Theorem 1 turns out to match the minimax lower bound in $\mathcal{F}_{p,n,s}(R,Q)$.

THEOREM 2 (LOWER BOUND). *Suppose $\varepsilon_i \overset{\text{iid}}{\sim} N(0,\sigma^2)$. If we have $n \geqslant Cs\log p$ for some large constant $C > 0$, $R \geqslant \bar{c}\sqrt{\{s\log(p/s)\,\sigma^2/n\}}$ for some constant $\bar{c} > 0$, $Q \geqslant p$ and $s \geqslant 4$, then*

$$\inf_{\hat{\beta}} \sup_{(\nu,X,\beta)\in\mathcal{F}_{p,n,s}(R,Q)} E(\|\hat{\beta} - \beta\|_2^2) \geqslant c\{s\log(p/s)/n\}\left\{\sigma^2 + (p/Q)R^2\right\}. \tag{11}$$

Next, we consider the overdispersed case where $\alpha_i < \infty$ in (3) and (4). We have the following upper bound for the estimation accuracy of the proposed variable correction regularized estimator (6).

THEOREM 3 (WITH OVERDISPERSION, UPPER BOUND). *Suppose Condition 1 holds and $a\bar{\nu} \leqslant \nu_i \leqslant b\bar{\nu}$, $a/p \leqslant X_{ij} \leqslant b/p$ for constants $0 < a < 1 < b$. Set $\zeta_{\max} = \max_i \zeta_i$, where $\zeta_i = (\nu_i + \alpha_i + 1)/\{2(\alpha_i + 1)\}$ represents the level of overdispersion for the ith sample. If, for some $\delta > 0$, some large constant C and a large constant $C(\delta)$ that only depends on δ, we have*

$\nu_{ij} \geqslant \zeta_i^{1+\delta}$, $n \geqslant Cs\log p$ and $\bar{\nu}/(p\zeta_{\max}) \geqslant \max[C\log(np), C(\delta)]$, then, by choosing $\lambda = C\left(\{\log(p)/n\}^{1/2}[\sigma + \{(p/\bar{\nu})\zeta_{\max}\}^{\frac{1}{2}}\|\beta^*\|_1] + \log(\bar{\nu}/p)(p/\bar{\nu})\zeta_{\max}\|\beta^*\|_1\right)$*, we have*

$$\|\hat{\beta} - \beta\|_2^2 \leqslant C\left[(s\log p/n)\left\{\sigma^2 + (p/\bar{\nu})\zeta_{\max}\|\beta^*\|_1^2\right\} + s\log^2(\bar{\nu}/p)(p/\bar{\nu})^2\zeta_{\max}^2\|\beta^*\|_1^2\right] \quad (12)$$

with probability at least $1 - 6p^{-C'} - \epsilon'$*, where* C' *is a constant.*

4. General high-dimensional log-error-in-variable regression

We extend the discussion to general high-dimensional log-error-in-variable regression that accommodates broader scenarios. This will also justify the +0.5 variable correction rule under a broader range of misspecified models. Specifically, let

$$y = V\beta^* + \varepsilon, \quad \text{or equivalently } y_i = \sum_{j=1}^{p}\log(\nu_{ij})\beta_j^* + \varepsilon_i, \qquad \text{subject to} \quad \sum_{j=1}^{p}\beta_j^* = 0,$$
$$\text{where } W = (W_{ij}),\ \mathbb{E}W_{ij} = \nu_{ij},\ W_{ij} \text{ are independent} \qquad (i = 1,\ldots,n, j = 1,\ldots,p). \quad (13)$$

Here, $V = \{\log(\nu_{ij})\}_{1\leqslant i\leqslant n, 1\leqslant j\leqslant p}$ are unknown underlying covariates, the ε_i are independent and identically distributed sub-Gaussian noise with mean zero and variance σ^2, β^* is the sparse parameter of interest and W_{ij} satisfies the following sub-exponential tail condition:

$$\mathbb{P}(|W_{ij} - \nu_{ij}| \geqslant t) \leqslant C\exp\left\{-c(t^2/\nu_{ij}) \wedge t\right\}, \qquad \forall t > 0. \quad (14)$$

In particular, the Poisson distribution satisfies (14), and hence our log error-in-variable regression model without overdispersion can be viewed as a special case of (13).

We consider the following +0.5 variable correction regularized estimator for β^* in the general high-dimensional log-error-in-variable regression model (13):

$$\hat{\beta} = \arg\min_{\beta}\left(\frac{1}{2n}\|y - B_W\beta\|_2^2 + \lambda\|\beta\|_1\right) \qquad \text{subject to } \sum_{j=1}^{p}\beta_j = 0. \quad (15)$$

Here, $B_W \in \mathbb{R}^{n\times p}$ with $(B_W)_{ij} = \log(W_{ij} + 0.5)$, and λ is some tuning parameter.

The following theorem provides an upper bound for the variable correction regularized estimator (15) in the general high-dimensional log-error-in-variable regression model (13).

Theorem 4 (General upper bound). *Suppose Condition 1 holds, $n \geqslant Cs\log(p)$ and $|\log(\nu_{ij}) - \log(\nu_{kl})| \leqslant a$ for some constant $a > 0$ and for all $1 \leqslant i, k \leqslant n$, $1 \leqslant j, l \leqslant p$. Denote $\bar{\nu} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}\nu_{ij}$ and $F = \max_{ij}|\mathrm{var}(W_{ij})/\nu_{ij} - 1|$. If, for some uniform constant $C > 0$, some $\epsilon > 0$ and a constant C_ϵ that only relies on ϵ, we have $\bar{\nu} \geqslant Cp\{S + Fs\log(s) + \log(np) + C_\epsilon\}$, then, by choosing $\lambda = C[\{\log(p)/n\}\{\sigma^2 + (p/\bar{\nu})\|\beta^*\|_2^2\} + s\{\min\{F^2, C\}(p/\bar{\nu})^2 + (p/\bar{\nu})^{3-2\epsilon}\}\|\beta^*\|_2^2]^{1/2}$ for some large constant $C > 0$, we have*

$$\|\hat{\beta} - \beta^*\|_2^2 \leqslant \frac{Cs\log p}{n}\left\{\sigma^2 + (p/\bar{\nu})\|\beta^*\|_2^2\right\} + Cs^2\left\{(p/\bar{\nu})^{3-2\epsilon} + \min\{F^2, C\}(p/\bar{\nu})^2\right\}\|\beta^*\|_2^2 \quad (16)$$

with probability $1 - 3p^{-C'} - \epsilon'$.

Remark 5. Theorem 4 shows that for the log-error-in-variable model, when the observed variables W_{ij} with exponential tail probability are linked with the response y in the form of (13) through its first moment and $\mathrm{var}(W_{ij})$ is close to ν_{ij}, like the Poisson case, the +0.5 correction rule can achieve a reasonable estimation error under proper conditions. If this is violated, such as $\mathrm{var}(W_{ij})$ being much larger than ν_{ij}, the error upper bound for this correction can be large, which can be a potential limit of the method.

Moreover, the estimation error upper bound of (16) includes two parts: (a) $C(s\log p/n)\sigma^2$, which corresponds to the error of ε_i and also appears in Theorem 1; (b) $C\left[(s\log p/n)(p/\bar{\nu})\|\beta^*\|_2^2 + Cs^2\left\{(p/\bar{\nu})^{3-2\epsilon} + \min\{F^2, C\}\,(p/\bar{\nu})^2\right\}\|\beta^*\|_2^2\right]$, which originates from the error-in-variable and is no smaller than the one in Theorem 1. When W_{ij} is Poisson distributed, we have $\mathrm{var}(W_{ij}) = \mathbb{E}W_{ij}$, $F = 0$, and the upper bound (16) reduces to (8) in Theorem 1.

5. Simulation studies

Now we evaluate the performance of our method using three simulation schemes for different purposes. In the first simulation scheme, we compare the prediction and estimation performance of the proposed variable correction regularized estimator with the classic method of zero-replacement under our model assumption with different parameter settings. To simulate the count matrix W with $n = 50, 100$ samples and $p = 100, 200, 400$ covariates, we first generate N_i from a negative binomial distribution with mean 3×10^4 and variance 3×10^6. Here, the main purpose of choosing negative binomial instead of Poisson as in the theoretical analysis is to show that the Poisson assumption on N_i is not crucial in practice. Then we set $X_{ij} = X_{i+n/2,j} = \exp(\Phi_{ij})/\{\sum_{k=1}^{p}\exp(\Phi_{ij})\}$ for $j = 1, \ldots, p, i = 1, \ldots, n/2$, with $\{\Phi_{ij}\}$ generated independently from $N(\mu_j, 1.5^2)$, where μ_1, μ_2, μ_3 are drawn from Un[1,3], $\mu_4, \ldots, \mu_7$ are drawn from Un[2,4] and μ_j for $j = 8, \ldots, p$ are drawn from Un[0,2]. With this setting, the average count of covariates will have a reasonable variation, with causal covariates slightly more abundant than noncausal ones. Then we generate $(W_{i1}, \ldots, W_{ip})$ from Dir-Mu$(N_i, \alpha X_{i1}, \ldots, \alpha X_{ip})$, where the overdispersion parameter $\alpha = 200, 1000, 5000$. The ith and $(i + n/2)$th samples are designed to be from the same subject so they share the same X_{ij} and can be used to estimate their shared overdispersion parameter. The response y is generated as $y_i = \sum_{j=1}^{p}\log(X_{ij})\beta_j + \varepsilon_i$, where $\beta = (1, -0.8, -1.5, 0.6, -0.9, 1.2, 0.4, 0, \ldots, 0)$ is the deterministic coefficient vector, and the ε_i are independent and identically distributed noise generated from $N(0, 0.5^2)$. The results are aggregated in Fig. 2. We can see that variable correction significantly outperforms zero-replacement by 0.5 in all parameter configurations.

To evaluate the performance of the proposed method when the response variable y is shared by samples from the same subject, like we have in real data analysis, we repeat the aforementioned simulation with one change: y is generated with $\varepsilon_i = \varepsilon_{i+n/2}$ and ε_i, $i = 1, \ldots, n/2$, are independent and identically distributed from $N(0, 0.5^2)$. The results are summarized in Fig. 3. We can see that the pattern of performance is similar to Fig. 2 and the variable correction method still significantly achieves smaller estimation and prediction errors.

In the second simulation scheme, we compare different methods under misspecified models, i.e., when the data-generation mechanism deviates from our model assumption. We generate N_i and X_{ij} in the same way as the first simulation scheme, and y_i is drawn in the same way as the first simulation scheme with independent error ε_i. To simulate W_{ij}, we first set $\alpha = 1000$ and generate $(\tilde{Q}_{i1}^{(1)}, \ldots, \tilde{Q}_{ip}^{(1)})$ from a log-normal distribution such that

$$\log(\tilde{Q}_{i1}^{(1)}, \ldots, Q_{ip}^{(1)}) \sim N\left\{\log(\alpha X_{i1}, \ldots, \alpha X_{ip}) - 1/8, \Sigma\right\}, \qquad \Sigma_{ij} = 0.5^{|i-j|}/4, 1 \leqslant i, j \leqslant p.$$

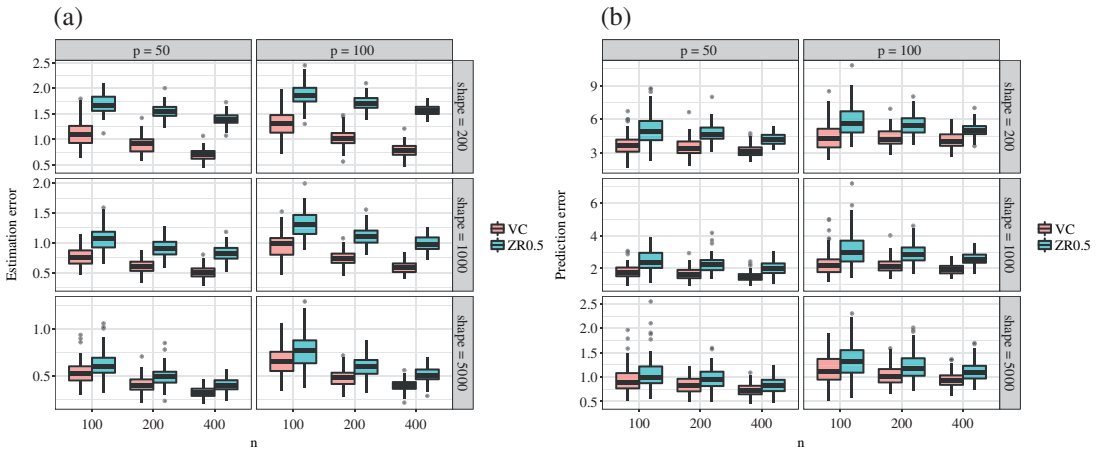

Fig. 2. Comparison between variable correction estimator (pink) and zero-replacement estimator by 0.5 (blue) in simulation analysis. (a) Estimation error (b) Prediction error. The noise terms ε_i are all independent.

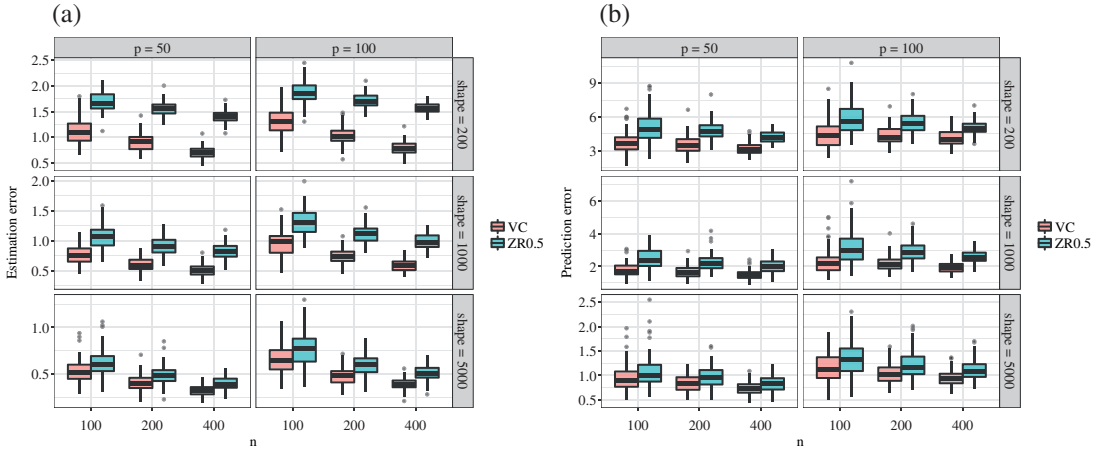

Fig. 3. Comparison between the variable correction estimator (pink) and the zero-replacement estimator by 0.5 (blue) in simulation analysis. (a) Estimation error (b) Prediction error. Here, y is shared by samples from the same subject.

Then $(\tilde{Q}_{i1}^{(1)}, \ldots, \tilde{Q}_{ip}^{(1)})$ are normalized into proportions:

$$(Q_{i1}^{(1)}, \ldots, Q_{ip}^{(1)}) = (\tilde{Q}_{i1}^{(1)}, \ldots, \tilde{Q}_{ip}^{(1)}) / \textstyle\sum_{j=1}^{p} \tilde{Q}_{ij}^{(1)}.$$

Next, we generate $(Q_{i1}^{(2)}, \ldots, Q_{ip}^{(2)})$ from $\mathrm{Dir}(\alpha X_{i1}, \ldots, \alpha X_{ip})$ and take $Q_{ij} = wQ_{ij}^{(1)} + (1 - w)Q_{ij}^{(2)}$, where w takes a series of values between 0 and 0.5. Finally, $(W_{i1}, \ldots, W_{ip})$ is drawn from $\mathrm{Mu}(N_i, Q_{i1}, \ldots, Q_{ip})$. When $w = 0$, this simulation scheme is exactly Dirichlet-multinomial. The larger w is, the more misspecified the model is.

We compare the performance of (a) VC_MOM, the variable correction estimator with the overdispersion parameters estimated via the method of moment described in the Supplementary Material; (b) VC_AH, the variable correction estimator with the overdispersion parameters set to ∞, which means adding all counts W_{ij} by half; (c) ZR0.5, zero-replacement by 0.5. We use ZR0.5

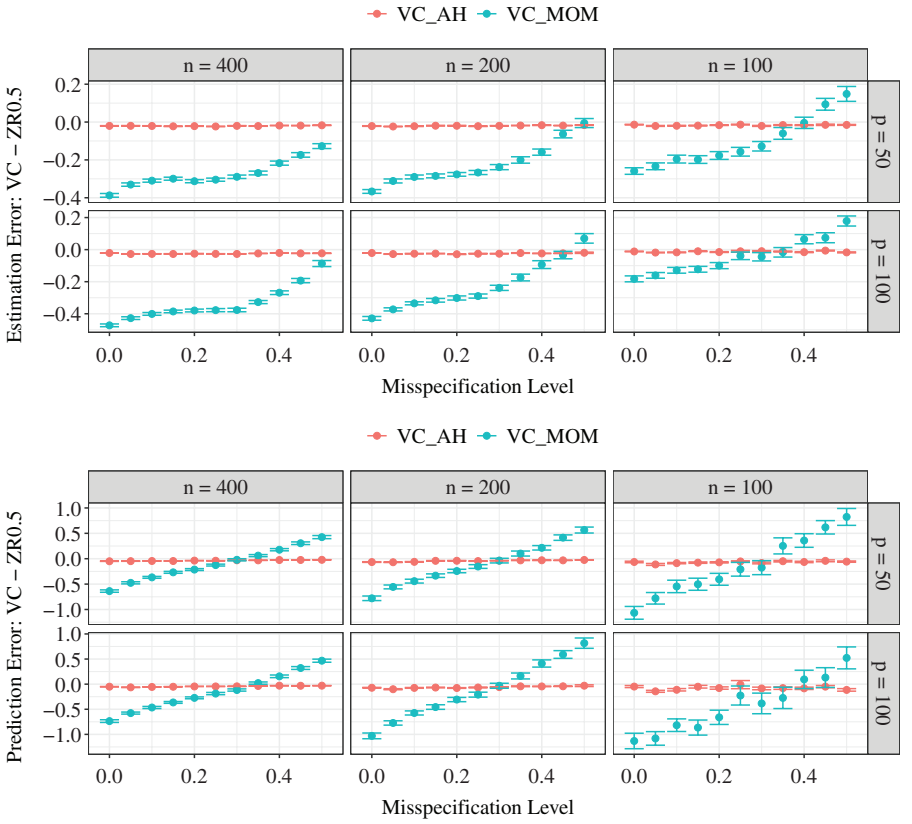

Fig. 4. Estimation (upper panel) and prediction (lower panel) performance of VC_AH (orange) and VC_MOM (blue) in reference to ZR0.5 under the misspecified settings. The dots represent mean difference in performance, and the error bars represent mean ± standard deviation.

as the baseline and summarize the difference in performance between each method and ZR0.5 for each round of simulation in Fig. 4. We can see that the proposed VC_MOM has significantly better performance than the classic ZR0.5 when the misspecification is moderate; VR_AH is always slightly better than the classic ZR0.5.

Finally, in the third simulation scheme we consider a setting where only one measurement is available for each subject. We generate N_i, y_i and W_{ij} in the same way as the second simulation scheme, but we no longer require $X_{ij} = X_{i+n/2,j}$ so that all the samples are from different subjects. We can see from Fig. 5 that the proposed procedure VC_AH performs better than the classic zero-replacement ZR0.5 scheme in all settings.

6. Regression analysis for longitudinal microbiome studies

We apply the proposed procedure to a longitudinal microbiome study reported by Flores et al. (2014). We focus on the association between body mass index, BMI, and gut microbiome composition at the genus level for healthy adults by excluding subjects with missing BMI, antibiotic disturbance, or other medication use. For the remaining 40 subjects, each having 4 samples, 92 bacteria genera appear in more than 10% of the samples and will be used for the analysis here. We also adjust for gender, age, race/ethnicity (caucasian, asian/pacific islander, hispanic/half-white hispanic, or other), dietary preference (vegan, omnivore, but no red meat, or omnivore),

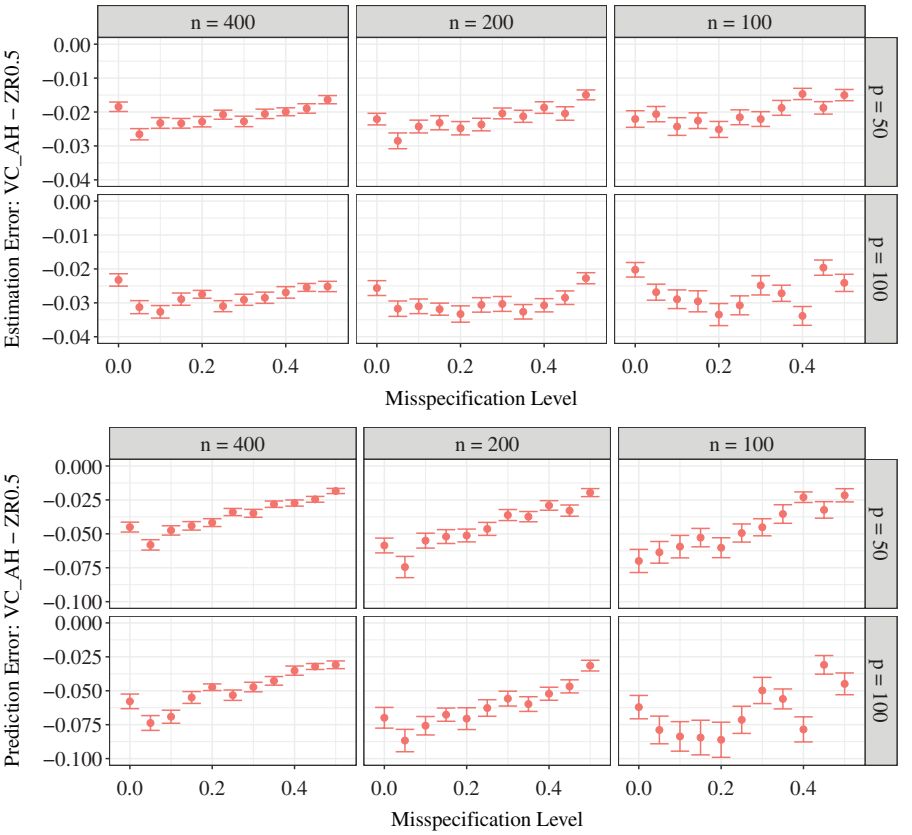

Fig. 5. Estimation (upper panel) and prediction (lower panel) performance of VC_AH in reference to ZR0.5 under the misspecified settings when no repeated measurement is available. The dots represent mean difference in performance, and the error bars represent mean ± standard deviation.

vitamin intake (yes or no) and exercise frequency (daily/regularly or occasionally/rarely) of the subjects in the regression analysis.

We implement the proposed variable correction estimator. Specifically, we assume the samples of the same subject share the same unobserved composition X_{ij} and overdispersion parameter α_i, and estimate α_i for each subject respectively using the method of moments estimator $\alpha_{i,\mathrm{MOM}}$ described in the Supplementary Material. Then we apply the regression model (6) with y representing BMI, and W_{ij} representing the read count of the ith sample and jth genus. For comparison, we also perform the classic zero-replacement method in the literature with zero counts changed to $c = 0.1$ and 0.5, respectively:

$$\hat{\beta}^{\mathrm{ZR}} = \underset{\beta:1_p^{\mathrm{T}}\beta=0}{\arg\min} \left[\sum_{i=1}^{n} \left\{ y_i - \sum_{j=1}^{p} \log\left(W_{ij} \vee c\right) \beta_j \right\}^2 / (2n) + \lambda \|\beta\|_1 \right].$$

To obtain stable variable selection, we generate 100 bootstrap samples of size $n/2$, repeat all methods with five-fold cross-validation choosing the tuning parameter λ on each subsample, and record the frequency of each variable being selected among the 100 bootstrap fittings. For illustration purposes, we consider a variable to be selected if its selection frequency is no less than 0.7.

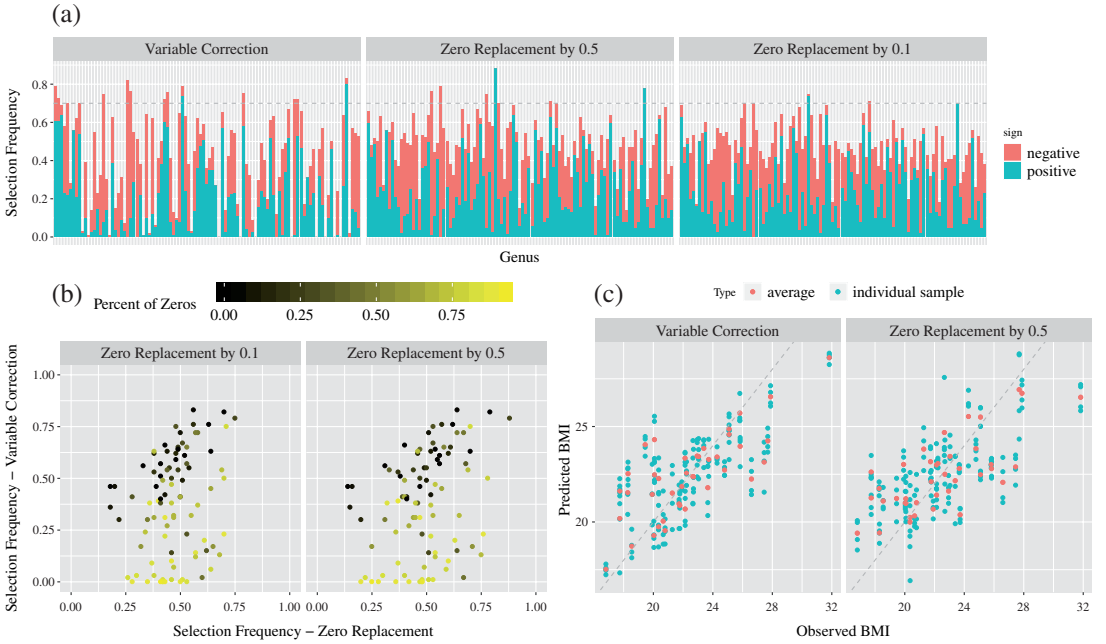

Fig. 6. (a) Selection frequency of 92 genera using different methods. (b) Comparison of selection frequency with regard to proportion of zero counts. (c) Comparison of prediction performance.

Figure 6(a) illustrates the selection frequency of the nine covariates we adjusted for and 92 genera for each method. The dashed line corresponds to the selection frequency of 0.7. It can be observed that the variable correction estimator selects variables with either very high or very low frequency, while the zero-replacement estimator has many more variables selected with a midrange frequency. This comparison indicates that the regularized estimator has much better stability in variable selection than zero-replacement. The variables selected by all three methods are *Bacteroides*(−), *Dialister*(+) and *Megamonas*(+). Variables selected by the variable correction method only are being male(+), age(+), frequent exercise(+), being hispanic/half-white hispanic(−), *Akkermansia*(−), *Bifidobacterium*(−), *Coprobacillus*(+), *Coprococcus*(+), *Porphyromonas*(−), *Prevotella*(+) and *Sutterella*(+). Variables selected by zero-replacement with $c = 0.1$ form a subset of that with $c = 0.5$, where *Arcanobacterium*(−) and *Slackia*(+) are selected by both, and *Dehalobacterium*(−), *Dorea*(−) and *Lactococcus*(+) are selected by $c = 0.5$ only. Here, (+) and (−) are the signs of regression coefficients by plurality vote in the 100 bootstrap fittings. These selected genera correspond to the bars exceeding the dashed line in the top panel of Fig. 6. Among the genera selected by variable correction, but missed by zero-replacement with $c = 0.5$, *Bifidobacterium* has been widely studied for its lipid-lowering effect and negative association with obesity (An et al., 2011; Million et al., 2012). *Akkermansia* has been reported to be negatively related to obesity extensively in the literature (Everard et al., 2013; Dao et al., 2016; Derrien et al., 2017). *Coprococcus* has also been reported to be positively related to obesity (Kasai et al., 2015) and negatively related to weight loss induced by diet or gastric bypass surgery, as indicated by Damms-Machado et al. (2015).

Figure 6(b) offers a closer look at the selection frequency with regard to the proportion of $W_{ij} = 0$ for each variable. Compared to the zero-replacement method, the variable correction method tends to select variables with fewer zeros. This makes the variable correction method more desirable since the bacteria with large proportions of zeros are often possessed by only a

few subjects, are far less reliable for prediction and interpretation purposes, and are difficult to generalize to a larger population. Figure 6(c) compares the prediction performance of variable correction and zero-replacement with $c = 0.5$, where the predicted BMI for each sample is obtained using refitted coefficients of the genera that have selection frequency no less than 0.7. When calculated using all the individual samples, R^2 is 0.50 for variable correction and 0.42 for zero-replacement with $c = 0.5$. Since each subject has multiple samples, we also provide the average predicted BMI of each subject in the figure. Using average predicted BMI, R^2 is 0.56 for variable correction and 0.51 for zero-replacement. We can see that the proposed method achieves much better prediction compared to zero-replacement using both the individual and average predicted BMI.

7. DISCUSSION

The proposed log-error-in-variable regression model method provides a new solution to deal with zero read counts in high-dimensional regression analysis of microbiome studies. In contrast, many existing methods, such as EdgeR (Robinson et al., 2010), DESeq2 (Love et al., 2014) and metagenomeSeq (Paulson et al., 2013), see McMurdie & Holmes (2014) for an overview of procedures, model the zero read counts through negative binomial distribution or zero-inflated distributions. These methods can draw conclusions about the marginal effect of each component one at a time, while our method aims at association between regression response and a component when other components are adjusted for. In another related line of work, de la Cruz & Kreft (2019) introduced a method to find a modified version of the geometric mean that is close to the traditional geometric mean while being able to handle zeros. Their method can be modified to find a good alternative for a given linear combination of the log read counts, but not for unknown linear combinations like our regression equation.

One limitation of our regression model is that it does not discriminate between zero counts due to undersampling and actual zeros due to absence of the component. It assumes the underlying true composition to be positive, although it can be infinitely close to zero. Another limitation of our method is its ability to deal with rare components. Our variable correction is less accurate when the true bacterial abundance X_{ij} is close to zero. This limitation is intrinsic to the log-error-in-variable model because of the large derivative of the log function around zero. However, as shown in our real data analysis, the ℓ_1 penalized regression we used for variable selection tends to select the more abundant bacteria, possibly because the small read counts are overshadowed by the correction added to it.

In addition to the aforementioned microbiome study, the proposed framework can be used for other applications on regression with count covariates. For example, in single-cell RNA-seq data analysis, a high-throughput sequencing technique was performed on each of the single cells, and the gene expressions can be measured as the total number of reads mapped to exonic regions. This resulting count matrix has rows and columns representing single cells and gene expressions, respectively. With the proposed method, we can perform regression analysis to study the association among the gene expressions of single cells and clinical phenotypes. Another potential application is in text mining, where one central task of topic modelling is to learn the topics of various documents when they share the same vocabulary of words. By counting the number of words or n-grams in each document, one can obtain count matrix data. Compared to the absolute counts of these words and n-grams, the relative abundances may be more predictive on the topic. Thus, our proposed method can be useful for building classifiers for topics of documents.

Acknowledgement

We thank the editor, the associate editor and two referees for their helpful comments that helped improve the presentation of this paper. We thank Sündüz Keleş for helpful discussions. The research of Zhou and Zhang was supported in part by the National Science Foundation (CAREER-1944904, DMS-1811868) and the National Institutes of Health (R01 GM131399). Shi was supported in part by the National Institutes of Health (R01 HG003747 and R21 HG009744).

Supplementary material

Supplementary Material available at *Biometrika online* includes the additional estimation procedure of the overdispersion parameter and proofs of all lemmas and theorems.

References

Aitchison, J. & Bacon-Shone, J. (1984). Log contrast models for experiments with mixtures. *Biometrika* **71**, 323–30.

An, H. M., Park, S. Y., Lee, D. K., Kim, J. R., Cha, M. K., Lee, S. W., Lim, H. T., Kim, K. J. & Ha, N. J. (2011). Antiobesity and lipid-lowering effects of *Bifidobacterium* spp. in high fat diet-induced obese rats. *Lipids Health Dis.* **10**, 116.

Belloni, A., Rosenbaum, M. & Tsybakov, A. B. (2017). Linear and conic programming estimators in high dimensional errors-in-variables models. *J. R. Statist. Soc.* B **79**, 939–56.

Bickel, P. J., Ritov, Y. & Tsybakov, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37**, 1705–32.

Candès, E. & Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35**, 2313–51.

Cao, Y., Zhang, A. & Li, H. (2020). Multisample estimation of bacterial composition matrices in metagenomics data. *Biometrika* **107**, 75–92.

Chen, J. & Li, H. (2013). Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *Ann. Appl. Statist.* **7**, 418–42.

Dai, Z., Wong, S. H., Yu, J. & Wei, Y. (2019). Batch effects correction for microbiome data with Dirichlet-multinomial regression. *Bioinformatics* **35**, 807–14.

Damms-Machado, A., Mitra, S., Schollenberger, A. E., Kramer, K. M., Meile, T., Königsrainer, A., Huson, D. H. & Bischoff, S. C. (2015). Effects of surgical and dietary weight loss therapy for obesity on gut microbiota composition and nutrient absorption. *BioMed Res. Int.* **2015**, 806248.

Dao, M. C., Everard, A., Aron-Wisnewsky, J., Sokolovska, N., Prifti, E., Verger, E. O., Kayser, B. D., Levenez, F., Chilloux, J., Hoyles, L. et al. (2016). *Akkermansia muciniphila* and improved metabolic health during a dietary intervention in obesity: Relationship with gut microbiome richness and ecology. *Gut* **65**, 426–36.

Datta, A. & Zou, H. (2017). Cocolasso for high-dimensional error-in-variables regression. *Ann. Statist.* **45**, 2400–26.

de la Cruz, R. & Kreft, J.-U. (2019). Geometric mean extension for data sets with zeros. *arXiv:*1806.06403v2.

Deming, W. E. (1943). *Statistical Adjustment of Data*. New York: Wiley.

Derrien, M., Belzer, C. & de Vos, W. M. (2017). *Akkermansia muciniphila* and its role in regulating host functions. *Microb. Pathog.* **106**, 171–81.

Everard, A., Belzer, C., Geurts, L., Ouwerkerk, J. P., Druart, C., Bindels, L. B., Guiot, Y., Derrien, M., Muccioli, G. G., Delzenne, N. M., de Vos, W. M. & Cani, P. D. (2013). Cross-talk between *akkermansia muciniphila* and intestinal epithelium controls diet-induced obesity. *Proc. Nat. Acad. Sci.* **110**, 9066–71.

Flores, G. E., Caporaso, J. G., Henley, J. B., Rideout, J. R., Domogala, D., Chase, J., Leff, J. W., Vázquez-Baeza, Y., Gonzalez, A., Knight, R., Dunn, R. R. & Fierer, N. (2014). Temporal variability is a personalized feature of the human microbiome. *Genome Biol.* **15**, 531.

Guimaraes, P. & Lindrooth, R. C. (2007). Controlling for overdispersion in grouped conditional logit models: A computationally simple application of Dirichlet-multinomial regression. *Economet. J.* **10**, 439–52.

Hausman, J. (2001). Mismeasured variables in econometric analysis: Problems from the right and problems from the left. *J. Econ. Persp.* **15**, 57–67.

Holmes, I., Harris, K. & Quince, C. (2012). Dirichlet multinomial mixtures: Generative models for microbial metagenomics. *PloS ONE* **7**, e30126.

Kasai, C., Sugimoto, K., Moritani, I., Tanaka, J., Oya, Y., Inoue, H., Tameda, M., Shiraki, K., Ito, M., Takei, Y. & Takase, K. (2015). Comparison of the gut microbiota composition between obese and non-obese individuals in a Japanese population, as analyzed by terminal restriction fragment length polymorphism and next-generation sequencing. *BMC Gastroenterology* **15**, 100.

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

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

LIN, W., SHI, P., FENG, R. & LI, H. (2014). Variable selection in regression with compositional covariates. *Biometrika* **101**, 785–97.

LOH, P.-L. & WAINWRIGHT, M. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Ann. Statist.* **40**, 1637–64.

LOVE, M. I., HUBER, W. & ANDERS, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550.

MANDAL, S., VAN TREUREN, W., WHITE, R. A., EGGESBØ, M., KNIGHT, R. & PEDDADA, S. D. (2015). Analysis of composition of microbiomes: A novel method for studying microbial composition. *Microbial Ecol. Health Dis.* **26**, 27663.

MARTIN-FERNANDEZ, J., BARCELÓ-VIDAL, C. & PAWLOWSKY-GLAHN, V. (2000). Zero replacement in compositional data sets. In *Data Analysis, Classification, and Related Methods*. H. A. L. Kiers, J.-P. Rasson, P. J. F. Groenen & M. Schader, eds. New York: Springer, pp. 155–60.

MCMURDIE, P. J. & HOLMES, S. (2014). Waste not, want not: Why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.* **10**, e1003531.

MILLION, M., MARANINCHI, M., HENRY, M., ARMOUGOM, F., RICHET, H., CARRIERI, P., VALERO, R., RACCAH, D., VIALETTES, B. & RAOULT, D. (2012). Obesity-associated gut microbiota is enriched in *lactobacillus reuteri* and depleted in *bifidobacterium animalis* and *methanobrevibacter smithii*. *Int. J. Obesity* **36**, 817–25.

MOSIMANN, J. E. (1962). On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika* **49**, 65–82.

NAVIN, N., KENDALL, J., TROGE, J., ANDREWS, P., RODGERS, L., MCINDOO, J., COOK, K., STEPANSKY, A., LEVY, D., ESPOSITO, D. et al. (2011). Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90.

PAULSON, J. N., POP, M. & BRAVO, H. C. (2013). *Metagenomeseq: Statistical analysis for sparse high-throughput sequencing*. Bioconductor package. http://www.cbcb.umd.edu/software/metagenomeSeq [last accessed 20 April 2021].

QIU, X., HILL, A., PACKER, J., LIN, D., MA, Y.-A. & TRAPNELL, C. (2017). Single-cell mRNA quantification and differential analysis with census. *Nature Methods* **14**, 309.

ROBINSON, M. D., MCCARTHY, D. J. & SMYTH, G. K. (2010). Edger: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–40.

ROSENBAUM, M. & TSYBAKOV, A. B. (2010). Sparse recovery under matrix uncertainty. *Ann. Statist.* **38**, 2620–51.

ROSENBAUM, M. & TSYBAKOV, A. B. (2013). Improved matrix uncertainty selector. In *From Probability to Statistics and Back: High-Dimensional Models and Processes – A Festschrift in Honor of Jon A. Wellner*. Institute of Mathematical Statistics, pp. 276–90.

RUDELSON, M. & ZHOU, S. (2017). Errors-in-variables models with dependent measurements. *Electron. J. Statist.* **11**, 1699–797.

SCHWABE, R. F. & JOBIN, C. (2013). The microbiome and cancer. *Nature Rev. Cancer* **13**, 800.

SHI, P., ZHANG, A. & LI, H. (2016). Regression analysis for microbiome compositional data. *Ann. Appl. Statist.* **10**, 1019–40.

THE HUMAN MICROBIOME PROJECT CONSORTIUM (2012). A framework for human microbiome research. *Nature* **486**, 215.

TURNBAUGH, P. J., LEY, R. E., MAHOWALD, M. A., MAGRINI, V., MARDIS, E. R. & GORDON, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444**, 1027.

TVEDEBRINK, T. (2010). Overdispersion in allelic counts and θ-correction in forensic genetics. *Theoret. Pop. Biol.* **78**, 200–10.

WADSWORTH, W. D., ARGIENTO, R., GUINDANI, M., GALLOWAY-PENA, J., SHELBURNE, S. A. & VANNUCCI, M. (2017). An integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. *BMC Bioinform.* **18**, 94.

WANG, T. & ZHAO, H. (2017). Structured subcomposition selection in regression and its application to microbiome data analysis. *Ann. Appl. Statist.* **11**, 771–91.

YIN, J. & WANG, J. (2014). A Dirichlet multinomial mixture model-based approach for short text clustering. In *Proc. 20th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pp. 233–42.

[*Received on* 12 *April* 2019. *Editorial decision on* 8 *March* 2021]