Statistica Sinica Preprint No: SS-2025-0223	
Title	High-Dimensional Log Contrast Models with Measurement Errors
Manuscript ID	SS-2025-0223
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0223
Complete List of Authors	Wenxi Tan, Lingzhou Xue, Songshan Yang and Xiang Zhan
Corresponding Authors	Lingzhou Xue
E-mails	lzxue@psu.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0223

High-dimensional log contrast models with measurement errors

Wenxi Tan[1], Lingzhou Xue[1], Songshan Yang[2], and Xiang Zhan[3]

[1]*The Pennsylvania State University,* [2]*Renmin University of China, and* [3]*Southeast University*

Abstract: High-dimensional compositional data are increasingly prevalent across diverse fields of modern scientific research. Regression analysis involving compositional data presents unique challenges, particularly when covariate measurement errors are present. These errors can propagate across composition components due to their inherent dependency structure, complicating the application of conventional error-in-variables regression techniques. To simultaneously address the compositional nature and measurement errors in the high-dimensional design matrix of compositional covariates, we propose the **Er**ror-**i**n-**C**omposition (Eric) Lasso, a novel method for regression analysis with high-dimensional compositional covariates subject to measurement error. We establish theoretical guarantees for Eric Lasso, including estimation error bounds and asymptotic sign-consistent variable selection properties. The finite-sample performance of the method is demonstrated through simulation studies and a real-world appli-

The authors are listed in alphabetical order. Correspondence should be addressed to Xiang Zhan (zhanx@seu.edu.cn) and Lingzhou Xue (lzxue@psu.edu)

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0223

cation.

Key words and phrases: Compositional data, Error-in-variable, High-dimensional regression, Log contrast models, Lasso.

1. Introduction

Compositional data, representing relative proportions or percentages of different parts that make up a whole, have a wide range of applications in many fields, including geology, ecology, social sciences, and biology. In biological and biomedical research, compositional data primarily arise from high-throughput sequencing technologies-based profiling experiments, which share a similar measurement process in which the total abundance information is lost and most sequence counts reflect only the relative abundances (i.e., compositional) information of unique sequences of interest (**?**). Regression analysis with these compositional covariates is essential to disentangle the relationships between compositions and an outcome of interest. Due to compositionality, traditional linear regression models fail for regression analysis with compositional predictors. To address the "curse of compositionality" in regression analysis, the log contrast model was proposed in the context of experiments with mixtures (**?**). Since then, multiple extensions have been developed for regression analysis with compositional

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0223

data, for instance, ??????????;? and ?.

Much of the existing work on high-dimensional log contrast regression has focused on the clean data case. However, measurement errors are ubiquitous in many scientific endeavors. Taking the compositional sequence count data in biomedical research that motivates our study as an example, measurement errors may occur at any stage of the experimental workflow, such as DNA extraction, PCR amplification, sequencing process, and even bioinformatics preprocessing procedure (?). Moreover, studies have reported that the problem of data contamination may also be related to genome databases with a large number of mislabeled sequences, which could potentially lead to wrong sequencing read counts inflated by orders of magnitude (?). These measurement errors need to be well accommodated in statistical analysis in order to avoid potential misleading or invalid scientific findings (?).

These measurement errors in sequencing studies are also referred to as the sequence bias (??). In those sequencing experiments, when the research of interest is an individual count variable, the issue of sequencing bias can be appropriately addressed or attenuated by multilevel modeling techniques, such as the Beta-Binomial regression (?) or the Poisson-Gamma model (?) to account for skewness or overdispersion observed in the contami-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0223

nated data. However, beyond marginal analysis, it is much more difficult to analyze contaminated compositional predictors normalized from multiple count variables, since measurement error in one component has a ripple effect on other components due to the compositional constraint. Similar to the regression analysis, we are facing the "curse of compositionality" in measurement error modeling for compositional data.

The canonical model for high-dimensional regression analysis is expressed as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$, where $\mathbf{y} = (y_1, \ldots, y_n)^T$ is the response vector, $\mathbf{X} \in R^{n\times p}$ is the design matrix of high-dimensional covariates, and $\boldsymbol{\beta}^* \in R^p$ is the regression coefficient vector of interest. In many applications, $\mathbf{X}$ may not be accurately measured, and a corrupted version $\mathbf{Z}$ of $\mathbf{X}$ is often available. In the literature of high-dimensional statistics, many versatile methods and theories have been developed for inference of $\boldsymbol{\beta}^*$ based on $\mathbf{Z}$ (**???**). In terms of compositional data, each row of $\mathbf{X}$ and $\mathbf{Z}$ belongs to simplex $\mathcal{S}^p = \{(x_1, \ldots, x_p) : x_j > 0, \sum_{j=1}^p x_j = 1\}$. Clearly, measurement error in one component has a ripple effect on other components due to the compositional constraint, and thus, existing statistical methods and theories are not directly applicable to measurement error problems of compositional data.

A more recent paper considered the measurement error problem in the

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0223

framework of log contrast models for regression analysis with compositional predictors (**?**). In particular, the variable correction regularized estimator proposed by **?** requires the knowledge of observed counts and leverages the Dirichlet-Multinomial distribution to correct observed counts towards unobservable underlying compositions. However, as pointed out in a recent paper (**?**), sequence read counts could be inflated by many orders of magnitude due to potential contamination issues of draft reference genomes. Assumptions of the variable correction regularized estimator no longer hold when measurement errors in counts are extremely large. Moreover, it is sometimes less meaningful to analyze counts from different platforms than compositions (**?**), which limits the applicability of the variable correction regularized estimator to large cohort studies where samples are typically sequenced at different locations/batches using different platforms. These potential limitations motivate our investigation to take a different approach to fill this research gap.

To develop a new method addressing the aforementioned limitations, we utilize the recently proposed mathematical model to characterize the issue of sequence bias in next-generation sequencing experiments (**??**). To further accommodate compositionality in the design matrix, we build our regression analysis framework upon the Aitchison log contrast model (**??**)

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0223

and then propose a new method named **Er**ror-**i**n-**c**ompositional (Eric) Lasso to handle corrupted high-dimensional compositional predictors. The error bound of our Eric Lasso estimator, along with its selection sign consistency property, is established. In summary, we propose a new method to simultaneously address both compositional nature and measurement errors in regressors, which distinguishes it from existing ones. The novelty of our work lies in both a new methodology with desirable statistical properties and interpretations, but also the particular application, which is an important and timely problem for which no satisfactory analysis methods exist so far.

The rest of this article is organized as follows. In Section **??**, we first introduce some background on regression analysis with compositional covariates and bias correction for error-in-variable in log contrast regression. Then, we propose the Eric Lasso method to handle regression analysis with contaminated compositional covariates. The estimation error bounds of Eric Lasso and its asymptotic sign consistency selection properties are established in Section **??**. We next demonstrate the superior performance of our method both using simulation studies and real data application examples in Section **??** and Section **??**, respectively. This article concludes with a discussion in Section **??**. Theoretical proofs are provided in the appendix.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0223

2. Methods

2.1 Preliminaries

Let y_i and $(U_{i1}, \ldots, U_{ip})$ denotes response of interest and p compositional covariates (i.e., $\sum_{j=1}^{p} U_{ij}$=1 and $U_{ij} > 0, \forall j$) measured from the ith sampling unit. To study relationships between response and compositional predictors, the following linear log contrast model (**?**) has been widely used:

$$y_i = \sum_{j=1}^{p} \log(U_{ij})\beta_j^* + \epsilon_i, \; s.t., \; \sum_{j=1}^{p} \beta_j^* = 0, \tag{2.1}$$

where the intercept term is omitted if both the outcome and predictors are centered. One remarkable feature in Model (**??**) is the zero-sum constraint on regression coefficients, which is essential to guarantee some basic principles in compositional data analysis: scale invariance, permutation invariance, and subcompositional coherence (**??**), which are often necessary for statistically meaningful interpretations of analysis results on compositional data. To understand this, one can first see that each component j of the compositional vector only carries relative information, and thus β_j^* itself is less meaningful in compositional data analysis. On the other hand, the contrast between two coefficients $\beta_j^* - \beta_l^*$ is meaningful in that it can measure the relative importance of components j and l. To get rid of potential bias in selecting a specific reference level l, one can use the quantity

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0223

$\frac{1}{p}\sum_{l=1}^{p}(\beta_j^* - \beta_l^*)$ to measure the relative importance of component j compared to the remaining ones in the compositional vector. This quantity reduces to β_j^* if and only if $\sum_{l=1}^{p}\beta_l^* = 0$ holds. In other words, the coefficient β_j^* can measure the relative importance of the jth component of the compositional vector under this zero-sum constraint. Therefore, it is crucial to develop interpretable regression analysis methods for compositional data under this zero-sum constraint in practice (**????**).

In many scenarios, compositions U_{ij} are not observable or measured with errors. For example, in a study on associations between gut microbial compositions and body mass index (**?**), the gut microbial compositions are not measurable and are approximated to the microbiota compositions in stool samples. In a typical sequencing study (e.g., 16S rRNA microbiome surveys or single-cell RNA-seq studies) that motivates our research, the observed compositions O_{ij}'s are often calculated from observed sequence counts W_{ij}'s. That is, $O_{ij} = W_{ij}/N_i$, where $N_i = \sum_{j=1}^{p} W_{ij}$ is the total counts (or sequencing depth) of sample i. These $W_{ij}'s$ are often called the original scale of measurements for compositions. By treating these original scales of measurements as random realizations of a certain distribution with compositions being its parameters, it has been argued that modeling these original scales of measurements has advantages over Aitchison's log-ratio

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0223

approaches (**?**). However, this approach (**?**) often treats compositional measurements as response variables, which does not apply to our setting of regression with corrupted compositional predictors as explanatory variables in log contrast regression models considered in the current article. Since true compositions U_{ij}'s may not be directly measurable, a naive idea is to run a surrogate linear log-contrast regression model with observed measurements W_{ij}'s:

$$y_i = \sum_{j=1}^{p} log(W_{ij})\beta_j + \epsilon_i,\ s.t.,\ \sum_{j=1}^{p} \beta_j = 0. \tag{2.2}$$

The existence of measurement errors will cause the departure of the estimated coefficients $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$ of the model (**??**) from the true values $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)^T$ of interest in model (**??**). Using a bias correction approach, **?** has shown that the bias term $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ can be improved if W_{ij} in model (**??**) is replaced by the corrected variables $W_{ij}^c = W_{ij} + \frac{N_i + \alpha_i + 1}{2\alpha_i + 1}$, where α_i is the over-dispersion parameter associated with sample i in the Dirichlet-Multinominal distribution (**?**). In other words, the authors recommend using the following model (**??**) to get a more accurate inference on $\boldsymbol{\beta}^*$:

$$y_i = \sum_{j=1}^{p} log(W_{ij}^c)\beta_j + \epsilon_i,\ s.t.,\ \sum_{j=1}^{p} \beta_j = 0. \tag{2.3}$$

Solving (**??**) requires the knowledge of observed counts and leverages the Dirichlet-Multinomial distribution to correct observed counts towards

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0223

unobservable underlying compositions. However, sequence read counts could be inflated by many orders of magnitude due to potential contamination issues of draft reference genomes (**?**). Consequently, the assumptions of the previous method (**?**) may no longer hold when measurement errors in counts are extremely large. Moreover, it is sometimes less meaningful to analyze counts from different platforms than compositions (**?**), which limits the applicability of model (**??**) to large cohort studies where samples are typically sequenced at different locations/batches using different platforms.

2.2 Log contrast regression with measurement errors

A notable nature of measurement errors in compositional covariates is the ripple effect, that is, measurement error in one component has an impact on at least one of the other components due to the unit-sum constraint. Unfortunately, most existing high-dimensional error-in-variable regression methods and theories have focused on the unconstrained data (**??**), which are not directly applicable to compositional data. In a typical sequence count technologies-based microbiome study, discrepancies between U_{ij} and O_{ij} (or simply sequencing bias) are thought to approximately act multiplicatively on the taxon abundances (**?**). Under this assumption, the following mathematical model has been proposed by computational biologists

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0223

to characterize the sequence bias issue of microbiome compositional data collected from sequencing experiments (**??**):

$$O_{ij} = U_{ij} \cdot \frac{E_{ij}}{\sum_{j=1}^{p} U_{ij} E_{ij}}, \tag{2.4}$$

where E_{ij} denotes the measurement error term for sample i and taxon j. That is, measurement errors are both sample-specific and feature-specific. On the one hand, features are not detected equally well, and measurement error is determined by the interaction between experimental protocols and the biological/chemical/physical state of each feature. On the other hand, samples might come from different sources and be treated by different technicians, and thus measurement error may also depend on the sampling process (**?**).

The compositional nature of measurements has been addressed in model (**??**) since $\sum_{j=1}^{p} O_{ij} = 1$. That is, ripple effects of measurement errors in one compositional component to another are well accommodated in model (**??**). Also we have $\sum_{j=1}^{p} \beta_j log(O_{ij}) = \sum_{j=1}^{p} \beta_j log(U_{ij}) + \sum_{j=1}^{p} \beta_j log(E_{ij})$ holds for any $\beta_1, \ldots, \beta_p$ with $\sum_{j=1}^{p} \beta_j = 0$. For ease of presentation, we define $X_{ij} = log(U_{ij}), Z_{ij} = log(O_{ij}), B_{ij} = log(E_{ij})$. Then, in the scale of log contrasts, we have $\mathbf{Z}\boldsymbol{\beta} = (\mathbf{X} + \mathbf{B})\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is an arbitrary zero-sum vector. We emphasize that the formulation $\mathbf{Z} = \mathbf{X} + \mathbf{B}$ is only true under the log contrast regression model, and this formulation will largely

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0223

facilitate measurement error modeling for regression analysis with contaminated compositional predictors. In other words, we do have the "blessing of compositionality" in measurement error modeling for log contrast regression with compositional covariates. We further assume that rows of $\mathbf{B}$ are independent and identically distributed with zero mean, finite covariance $\mathbf{\Sigma}_B$, and sub-Gaussian parameter τ^2 to develop further statistical inference.

Let $S = \{j : \beta_j^* \neq 0\}$ denote the support of $\boldsymbol{\beta}^*$ and $s = |S|$ is the cardinality of S. For any subset $J \subset \{1, \ldots, p\}$ and any index $j \in J$, let J^c denote the complement of J and define $J - j = J \backslash \{j\}$. For any matrix $K \in \mathbb{R}^{a \times b}$, K_J denotes the submatrix of the jth column for $j \in J \cap \{1, \ldots, b\}$ of K, $K_{J_1 J_2}$ is the submatrix formed by (i, j)th entries for $i \in J_1 \cap \{1, \ldots, a\}, j \in J_2 \cap \{1, \ldots, b\}$. $\|\cdot\|_\infty$ and $\|\cdot\|_1$ denote the l_∞ norm and l_1 norm respectively. Then, our log contrast model is expressed as: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon} = \mathbf{X}_S \boldsymbol{\beta}_S^* + \boldsymbol{\epsilon}$, $s.t.$, $\sum_{j=1}^p \beta_j^* = \sum_{j \in S} \beta_j^* = 0$, where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$ are independent and identically distributed random errors that follow $N(0, \sigma^2)$. While predictors $\mathbf{X}$ may not be observable, we propose a new method that uses their surrogate $\mathbf{Z}$ to perform inference on $\boldsymbol{\beta}^*$. We further introduce some necessary notation before introducing our new method. Let $\mathbf{X}^p = \{log(U_{ij}/U_{ip})\} \in R^{n \times (p-1)}$ and $\mathbf{Z}^p = \{log(O_{ij}/O_{ip})\} \in R^{n \times (p-1)}$. Without loss of generality, we pick the last component as the reference level when

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0223

defining $\mathbf{X}^p$ and $\mathbf{Z}^p$ to facilitate the development and presentation of our method. However, in the subsequent methodological development, we will guarantee that the proposed method is invariant in the selection of the reference component. Taking into account these log ratios, we can divide the p-dimensional regression coefficients as $\boldsymbol{\beta} = (\boldsymbol{\beta}_{-p}^T, \beta_p)^T$, where $\boldsymbol{\beta}_{-p} = (\beta_1, \ldots, \beta_{p-1})^T$ is free from any constraint. Assume all columns in $\mathbf{X}$ are centered and define $\boldsymbol{\Sigma} = \frac{1}{n}\mathbf{X}^T\mathbf{X}$. Then, one way to fit a sparse log contrast model in the high-dimensional setting is via the following L_1-regularization:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \arg\min_{\boldsymbol{\beta}} \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \,, s.t., \sum_{j=1}^{p} \beta_j = 0 \\ &= \arg\min_{\boldsymbol{\beta}=\mathbf{D}^p\boldsymbol{\beta}_{-p}} \frac{1}{2n}\|\mathbf{y} - \mathbf{X}^p\boldsymbol{\beta}_{-p}\|_2^2 + \lambda\|\mathbf{D}^p\boldsymbol{\beta}_{-p}\|_1 \\ &= \arg\min_{\boldsymbol{\beta}=\mathbf{D}^p\boldsymbol{\beta}_{-p}} \frac{1}{2}\boldsymbol{\beta}_{-p}^T\boldsymbol{\Sigma}^p\boldsymbol{\beta}_{-p} - (\boldsymbol{\rho}^p)^T\boldsymbol{\beta}_{-p} + \lambda\|\mathbf{D}^p\boldsymbol{\beta}_{-p}\|_1, \qquad (2.5)\end{aligned}$$

where $\boldsymbol{\Sigma}^p = \frac{1}{n}(\mathbf{X}^p)^T\mathbf{X}^p$, $\boldsymbol{\rho}^p = \frac{1}{n}(\mathbf{X}^p)^T\mathbf{y}$ and $\mathbf{D}^P = (\mathbf{I}_{p-1}, -\mathbf{1}_{p-1})^T$.

Since $\mathbf{X}$ is unobserved, we have to choose surrogates $\tilde{\boldsymbol{\Sigma}}^p$ and $\tilde{\boldsymbol{\rho}}^p$ to replace $\boldsymbol{\Sigma}^p$ and $\boldsymbol{\rho}^p$ in (??). We begin by first constructing $\hat{\boldsymbol{\Sigma}}$, an unbiased estimator of $\boldsymbol{\Sigma}$. As we specify the observed design matrix $\mathbf{Z}$ being contaminated by additive measurement error $\mathbf{B}$, where the rows of $\mathbf{B}$ are independent and identically distributed with zero mean and finite covariance $\boldsymbol{\Sigma}_B$, then $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\mathbf{Z}^T\mathbf{Z} - \boldsymbol{\Sigma}_B$ is an unbiased estimator of $\boldsymbol{\Sigma}$. Then, following the suggestion of CoCoLasso (?), the surrogate $\tilde{\boldsymbol{\Sigma}}$ and $\tilde{\boldsymbol{\rho}}$ for the

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0223

unobserved $\boldsymbol{\Sigma} = \frac{1}{n}\mathbf{X}^T\mathbf{X}$ and $\boldsymbol{\rho} = \frac{1}{n}\mathbf{X}^T\mathbf{y}$ are defined as:

$$\tilde{\boldsymbol{\Sigma}} = \arg\min_{\mathbf{K}\geq 0} \|\mathbf{K} - \hat{\boldsymbol{\Sigma}}\|_{\max} \ , \ \tilde{\boldsymbol{\rho}} = \frac{1}{n}\mathbf{Z}^T\mathbf{y},$$

where $\|\mathbf{K}\|_{\max}$ denotes the element-wise maximum norm of $\mathbf{K}$ and $\tilde{\boldsymbol{\Sigma}}$ is obtained through ADMM algorithm. Through matrix transformation, we have $\boldsymbol{\Sigma}^p = \frac{1}{n}(\mathbf{X}^p)^T\mathbf{X}^p = (\mathbf{D}^p)^T\boldsymbol{\Sigma}\mathbf{D}^p$. Then, surrogates $\tilde{\boldsymbol{\Sigma}}^p$ and $\tilde{\boldsymbol{\rho}}^p$ is calculated from $\tilde{\boldsymbol{\Sigma}}^p = (\mathbf{D}^p)^T\tilde{\boldsymbol{\Sigma}}\mathbf{D}^p$, $\tilde{\boldsymbol{\rho}}^p = (\mathbf{D}^p)^T\tilde{\boldsymbol{\rho}}$. Finally, plugging these quantities calculated from observed measurements $\mathbf{Z}$ into (**??**), we obtain the following estimator

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\, s.t. \boldsymbol{\beta}=\mathbf{D}^p\boldsymbol{\beta}_{-p}} \frac{1}{2}\boldsymbol{\beta}_{-p}^T\tilde{\boldsymbol{\Sigma}}^p\boldsymbol{\beta}_{-p} - (\tilde{\boldsymbol{\rho}}^p)^T\boldsymbol{\beta}_{-p} + \lambda\|\mathbf{D}^p\boldsymbol{\beta}_{-p}\|_1. \tag{2.6}$$

We now call this new estimator (**??**) the Eric Lasso estimator hereafter. The estimator (**??**) can be equivalently expressed as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\, s.t. \sum_j \beta_j = 0} \frac{1}{2}\boldsymbol{\beta}^T\tilde{\boldsymbol{\Sigma}}\boldsymbol{\beta} - \tilde{\boldsymbol{\rho}}^T\boldsymbol{\beta} + \lambda\|\boldsymbol{\beta}\|_1.$$

We then solve the minimization problem using the coordinate descent method with Lagrangian multipliers, following **?**. Clearly, our algorithm is independent of the choice of reference taxon.

Remark 1: The core of CoCoLasso lies in the two-step procedure of constructing a good estimator for matrix $\boldsymbol{\Sigma}$: first getting an unbiased estimation $\hat{\boldsymbol{\Sigma}}$ and then conducting projection to obtain a positive semi-definite matrix $\tilde{\Sigma}$. In regression analysis with compositional predictors, it

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0223

is tempting to directly apply this two-step CoCoLasso procedure to log-ratio transformed measurements $\mathbf{Z}^p$ to obtain surrogates $\tilde{\mathbf{\Sigma}}^p$ and $\tilde{\boldsymbol{\rho}}^p$ of $\mathbf{X}^p$, which unfortunately leads to analysis results that are sensitive to selection of the reference component and thus lacks valid interpretation for compositional data analysis. In contrast, the proposed Eric Lasso method is permutation invariant, which gives the same result under any permutation of the p components. We assume without loss of generality that $p \in S$ is such that $S - p$ is well-defined. Otherwise, we could permute the compositional covariates to ensure it. This assumption is reasonable as we can guarantee all conditions and proofs are invariant to permutations of covariate indices, which has been well-checked throughout this article. Also, we assume that measurement error does not depend on its relative abundance, that is, E_{ij} and U_{ij} are independent, which is used solely to facilitate theoretical analysis of our method, a common practice in the literature (e.g., **??**). Crucially, our method does not rely on this assumption for practical applicability, as shown in Scenario 3 of Simulation I reported later in this article, where the measurement error is explicitly modeled as heteroscedastic and dependent on U_{ij}.

Remark 2: The error covariance $\mathbf{\Sigma}_B$ might be unknown in practice and must be obtained through estimation. Suppose we can borrow informa-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0223

tion from either independent external data or replicated data to calculate an error matrix $B_o \in \mathbb{R}^{n_o \times p}$, where $(B_o)_{ij}$ denotes the additive measurement error for the jth covariate from the ith sample. Correspondingly, $\hat{\mathbf{\Sigma}}_B = \frac{1}{n} B_o^T B_o$ is an estimate of $\mathbf{\Sigma}_B$, which has also been widely assumed in literature (**??**). We can show that our theoretical analysis still holds when replacing $\mathbf{\Sigma}_B$ with $\hat{\mathbf{\Sigma}}_B$ in Lemma 1 in Section A of the online supplementary materials if $n/n_o = O(1)$.

3. Theoretical analysis

To establish the theoretical results of Eric Lasso, we need the following two regularity conditions:

Condition 1. The matrix Σ_{SS} satisfies $\Lambda_{\min}(\Sigma_{SS}) \geq C_{\min} > 0$, where $\Lambda_{\min}(K)$ denotes the minimal eigenvalue of matrix K.

Condition 2. There exists some $\xi \in (0, 1]$ such that

$$\|\Sigma^p_{S^cS}(\Sigma^p_{SS})^{-1}\{\mathrm{sgn}(\beta^*_{S-p}) - \mathrm{sgn}(\beta^*_p)1_{s-1}\} + \mathrm{sgn}(\beta^*_p)1_{p-s}\|_\infty \leq 1 - \xi. \quad (3.7)$$

Condition **??** is a common assumption in high-dimensional statistics, and some implications to guarantee the permutation invariance of our analysis under this condition are derived in Lemmas 3 and 4 in the appendix. Condition **??** is taken from **?**, which is central to guaranteed support recov-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0223

ery of L_1 regularization. An important quantity associated with Condition **??** is $\phi = \|D^p_{SS}(\Sigma^p_{SS})^{-1}(D^p_{SS})^T\|_\infty$, which is permutation invariant according to Proposition 2 of **?**. That is, the left-hand part of Equation (**??**) is independent of the selection of reference component p. We further assume that $\max_j \|X_j\|_2^2 \leq n$. With all these regulatory conditions, we establish the main theoretical results of Eric Lasso.

Theorem 1 (error bound and sign consistency)**.** *Under Condition* **??** *and Condition* **??**, *let* $\zeta = \max(\tau^4, \sigma^4, 1)$, *for* $\lambda \leq \tau^2$ *and* $\epsilon \leq \min(\epsilon_1, \lambda/(\lambda\epsilon_2 + \epsilon_3))$ *where* ϵ_i*'s are bounded positive constants depending of* Σ, β^*_S, θ *and* ϕ, *then there exists universal constants* C *and* c, *with probability at least* $1 - \delta_1$ *where* $\delta_1 = p^2 C \exp\{-cn(s-1)^{-4}\epsilon^2\zeta^{-1}\} + pC\exp(-cns^{-2}\lambda^2\xi^2\zeta^{-1})$, *problem (***??***) has an optimal solution* $\hat{\boldsymbol{\beta}}$ *that satisfies the following properties: (a)* l_∞*-loss:* $\|\hat{\beta}_S - \beta^*_S\|_\infty \leq 9\phi\lambda/2$, *(b) sign consistency: if* $\min_{j\in S}|\beta^*_j| > 9\phi\lambda/2$, *then* $sgn(\hat{\beta}) = sgn(\beta^*)$.

Remark 3: To further understand the implications of Theorem 1, we assume for simplicity that ϕ is constant. From the expression of δ_1, if $(s-1)^4 \log p/n \to 0$ as $n, p \to \infty$ and $\min_{j\in S}|\beta^*_j| \gg s(\zeta \log p/n)^{1/2}$, we can choose a λ satisfying that $\lambda \gg s(\zeta \log p/n)^{1/2}$ and $\min_{j\in S}|\beta^*_j| > 9\phi\lambda/2$ such that δ_1 goes to zero and hence the sign-consistency of Eric Lasso is achieved.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0223

4. Simulation Studies

4.1 Simulation setup

We used three different approaches to generate compositional predictors. In Scenario 1, we followed the simulation design of **?** to simulate $\mathbf{X}$ using the logistic normal distribution. In particular, we first simulated a $n \times p$ latent matrix denoted as W from a multivariate normal distribution $N(\boldsymbol{\theta}, \boldsymbol{\Sigma}_W)$ where $(\boldsymbol{\Sigma}_W)_{ij} = \rho^{|i-j|}$, where $\rho = 0.5$ and $\boldsymbol{\theta}$ was set in the following way: $\theta_j = \log(0.2p)$ for j = 1 to 5 and $\theta_j = 0$ for other indices. The unobserved compositions were further calculated as $U_{ij} = \exp(W_{ij}) / \sum_{k=1}^{p} \exp(W_{ik})$ and $X_{ij} = log(U_{ij})$. Compositions generated under this scheme were very heterogeneous in that the first five components dominated the remaining components. The true coefficient vector $\boldsymbol{\beta}^*$ were specified as (1.2, -0.8, 0.7, 0, 0, -1.5, -1, 1.4, 0, $\cdots$, 0)^T. The response was then generated as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$, where $\epsilon_1, \ldots, \epsilon_n$ are iid errors simulated from $N(0, 0.5^2)$. To simulate a corrupted composition, we first generated an error matrix $\mathbf{B}$, whose rows were independently simulated from $N(\mathbf{0}, \tau^2\mathbf{I})$ and then generated the corrupted compositions O_{ij} according to model (**??**), which were further transformed into $Z_{ij} = log(O_{ij})$. In Scenario 2, we considered more homogeneous community compositions by using the Dirichlet distribution

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0223

to generate community compositions (**?**). In particular, each row of $\mathbf{U}$ was independently simulated from the Dirichlet $(\frac{1}{p}\mathbf{1}_p)$, where $\mathbf{1}_p$ denotes the p-dimensional vector of all ones, and we kept all other settings the same as the first scenario. In Scenario 3, we examined method performance under misspecified models violating our core model assumptions as stated in Equation (**??**). In particular, we followed the exact design in **?** to generate latent counts. We first generated underlying compositions $(U_{i1}, \ldots, U_{ip})$ from the above logistic normal distribution and total sequence count N_i of sample i from the negative binomial distribution with mean 3×10^4 and variance 3×10^6. Then, observed counts $(W_{i1}, \ldots, W_{ip})$ were simulated from DirMult$(N_i, \alpha U_{i1}, \ldots, \alpha U_{ip})$ with $\alpha = 5000$, where DirMult denotes Dirichlet-Multinomial distribution. Following **?** on handling zero counts, we calculated the observed design matrix $\mathbf{Z}$ as $Z_{ij} = log\{(W_{ij} + 0.5)/\sum_j (W_{ij} + 0.5)\}$. In all of the following experiments, we adopt the same strategy for zero counts in the data.

In Simulation 1, we fixed variance of measurement errors at $\tau = 0.5$ and varied (n, p) to evaluate the performance of our method under different data sizes. This type of setting was considered in **?**. In Simulation II, we fixed (n,p) $= (100, 100)$ to mimic a real data set analyzed later and varied τ to evaluate the robustness of methods with respect to noise levels

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0223

of measurement errors following **?**. Scenario 3, whose misspecified model lacks τ, was included only in Simulation I.

After each dataset was simulated under a particular scenario and setting. We applied five different methods to data $(\mathbf{y}, \mathbf{Z})$ to obtain the regression coefficient estimator $\hat{\boldsymbol{\beta}}$ and then compared it to true regression coefficients $\boldsymbol{\beta}^* = (1.2, -0.8, 0.7, 0, 0, -1.5, -1, 1.4, 0, \ldots, 0)$. The five methods are Eric Lasso, compositional Lasso (**?**), CoCoLasso (**?**), the vanilla Lasso (**?**), and the compositional data analysis version of the Debiased Lasso estimator (**?**). For ease of presentation, we refer to these Lasso methods as Eric, Coda, CoCo, Vani, and Debi, respectively, hereafter in this article. Among these methods, Coda Lasso incorporates the compositional nature in data but ignores measurement errors. CoCoLasso accommodates measurement errors yet fails to model the compositional constraint in the regression covariates space. The Vani Lasso fails to accommodate either characteristic of the data. The proposed Eric Lasso and Debi Lasso take both aspects into account. The tuning parameter selection in Eric and CoCo was done by calibrated cross-validation (**?**) and that in Coda. Vani and Debi Lasso were done using cross-validation. To compare the estimation accuracy of different methods, we calculated three metrics including squared error (SE), prediction error (PE) and l_∞ loss:

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0223

$SE(\hat{\boldsymbol{\beta}}) = \|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_2^2$, $PE(\hat{\boldsymbol{\beta}}) = (\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}})^T \boldsymbol{\Sigma} (\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}})$, $l_\infty(\hat{\boldsymbol{\beta}}) = \max |\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}|$. To compare the selection sign consistency of different methods, we calculated the False Positive Rate(FPR) and False Negative Rate(FNR). For the Debiased Lasso method, the FPR and FNR were computed using the estimated p-values, with a significance threshold of 0.05, while for the other methods, the selection was obtained from the estimated coefficients, where positives refer to nonzero regression coefficients. For each specific simulation scenario, we repeated it 100 times to obtain multiple values of these metrics and reported their mean values along with the standard errors of the mean.

4.2 Simulation results

We first compare the estimation and selection performance of different Lasso methods under Simulation I. Results under Scenarios 1 are reported in Table **??**. Results under Scenarios 2 are reported in Table 2 of the online supplementary material. The results under Scenario 3 (model misspecification) are reported in Table **??**. The original Debi Lasso method (**?**) performs poorly in prediction when p increases, as reported in the following Table **??**-**??**. The issue is likely attributable to the step of calculating the sample inverse covariance matrix $\boldsymbol{\Sigma}^{-1}$ in equation (9) of **?**. The direct in-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0223

version used in the original implementation of **?** yields unstable results, as evidenced by the ultra-high variance in performance reported in Table **??**. We therefore introduce a refined version and denote it with Debi+. The Debi+ estimator is obtained by replacing $\mathbf{\Sigma}^{-1}$ in Debi with the Tikhonov regularized inverse $(\mathbf{\Sigma} + 10^{-4}I)^{-1}$. As shown in the tables, Eric Lasso and CoCo Lasso consistently have better estimation performance than those of Coda, Vanilla and Debiased-type Lasso. Although regularization mitigates the degraded predictive performance of Debi, the standard error (SE) and prediction error (PE) of Debi+ remain substantially higher than Eric Lasso. Under all three scenarios, we observe that estimation errors of Eric Lasso tend to decrease as the sample size n increases, with the only exception being the PE of Eric Lasso under Scenario 3 (model misspecification). For selection accuracy, the FPRs of Coda and Vani Lasso are significantly higher than those of Eric, CoCo, and Debi Lasso under Scenario 1. A similar phenomenon has been observed in the literature that Coda Lasso tends to select more unnecessary false positives to recover the true model (**??**). On the other hand, the FNRs of Eric and Debi lasso are comparable at $(n, p) \in \{(100, 200), (250, 400)\}$, but Eric has significantly lower FNR and FPR than both Debi and Debi+ as the dimensionality increases. Under Scenario 2, covariates $\mathbf{X}$ do not follow the logistic-normal distribution,

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0223

which violates Assumption 1 of the Debi Lasso method (**?**). Consequently, the estimation and selection performances of Debi Lasso are worse than those of the other four models. Notably, when $(n, p) = (100, 200)$, its FPR hits 1, suggesting that Debi Lasso selects nearly all features regardless of their relevance. Patterns under Scenario 2 are similar to those under Scenario 1 for Eric, Coda, CoCo, and Vani Lasso. Because the compositional design matrix generated under Scenario 2 is more homogeneous and the magnitudes of six true signals are relatively large, all methods successfully identified the signals, leading to zero FNR for all methods under Scenario 2. Overall, Eric Lasso has the best selection performance under these two scenarios. When models are misspecified under Scenario 3, a remarkable change is that Eric and CoCoLasso have much worse FNR compared to Coda and Vani Lasso, which is not surprising given that the FPR of Coda and Vani are two to five times of those of Eric and CoCo. Though the selection performance of Debi is significantly worse than Eric Lasso when $p > 400$. There are no methods that are uniformly better than others in terms of both FPR and FNR under this scenario. Additional numerical results on model interpretation of different methods under Simulation I were also compared and reported in the online supplementary materials.

We next evaluate the robustness of different methods against different

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0223

Table 1: Comparison of different Lasso estimators under Scenario 1 of Simulation I. Mean values (standard errors) of different evaluation metrics are based on 100 simulation replicates.

(n,p)	Model	SE	PE	FPR	FNR
(100,200)	Eric	2.16(0.08)	0.77(0.03)	0.08(0)	0.17(0.01)
	Coda	2.91(0.08)	1.10(0.04)	0.11(0.01)	0.16(0.01)
	CoCo	2.22(0.10)	0.75(0.03)	0.09(0)	0.16(0.02)
	Vani	2.96(0.09)	1.03(0.04)	0.13(0.01)	0.17(0.02)
	Debi	3.30(0.07)	1.25(0.03)	0.05(0)	0.16(0.01)
	Debi+	3.30(0.07)	1.25(0.03)	0.05(0)	0.16(0.01)
(250,400)	Eric	1.06(0.03)	0.43(0.01)	0.05(0)	0.03(0.01)
	Coda	1.95(0.03)	0.83(0.01)	0.09(0)	0.03(0.01)
	CoCo	1.03(0.04)	0.42(0.01)	0.06(0)	0.02(0.01)
	Vani	1.92(0.03)	0.78(0.02)	0.11(0.01)	0.03(0.01)
	Debi	3.01(0.05)	1.24(0.02)	0.04(0)	0.05(0.01)
	Debi+	3.01(0.05)	1.24(0.02)	0.04(0)	0.05(0.01)
(500,500)	Eric	0.59(0.02)	0.26(0.01)	0.03(0)	0(0)
	Coda	1.54(0.02)	0.72(0.01)	0.07(0)	0(0)
	CoCo	0.57(0.01)	0.26(0.01)	0.03(0)	0(0)
	Vani	1.54(0.02)	0.70(0.01)	0.08(0)	0(0)
	Debi	$\gg 10$	$\gg 10$	0.04(0)	0.17(0.03)
	Debi+	9.08(0.71)	2.64(0.15)	0.04(0)	0.05(0.01)
(550,700)	Eric	0.69(0.02)	0.31(0.01)	0.02(0)	0(0)
	Coda	1.57(0.02)	0.74(0.01)	0.05(0)	0(0)
	CoCo	0.92(0.02)	0.41(0.01)	0.03(0)	0(0)
	Vani	1.56(0.02)	0.72(0.01)	0.06(0)	0(0)
	Debi	$\gg 10$	$\gg 10$	0.05(0.01)	0.31(0.03)
	Debi+	7.19(0.22)	2.4(0.07)	0.04(0)	0.04(0.01)

noise levels of measurement errors, measured by the τ parameter. ROC curves of different Lasso methods are displayed in Figure **??**. Debi Lasso is not included in this figure, as its selection is calculated with a fixed p-value

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0223

Table 2: Comparison of different Lasso estimators under Scenario 3 of Simulation I. Mean values (standard errors) of different evaluation metrics are calculated based on 100 simulation replicates.

(n,p)	Model	SE	PE	FPR	FNR
(100,200)	Eric	2.27(0.05)	0.86(0.02)	0.08(0)	0.17(0.01)
	Coda	2.44(0.07)	0.92(0.03)	0.14(0.01)	0.11(0.01)
	CoCo	2.45(0.05)	0.86(0.02)	0.08(0)	0.22(0.01)
	Vani	2.43(0.07)	0.87(0.03)	0.16(0.01)	0.12(0.01)
	Debi	2.37(0.07)	0.87(0.02)	0.05(0)	0.05(0.01)
	Debi+	2.37(0.07)	0.87(0.02)	0.05(0)	0.05(0.01)
(250,400)	Eric	2.12(0.04)	0.99(0.02)	0.03(0)	0.15(0.01)
	Coda	2.23(0.04)	1.03(0.02)	0.13(0.01)	0.02(0.01)
	CoCo	2.22(0.04)	0.98(0.02)	0.03(0)	0.18(0.01)
	Vani	2.15(0.04)	0.97(0.01)	0.15(0.01)	0.01(0)
	Debi	2.73(0.05)	1.12(0.01)	0.04(0)	0.01(0)
	Debi+	2.73(0.05)	1.12(0.01)	0.04(0)	0.01(0)
(500,500)	Eric	1.51(0.04)	0.73(0.02)	0.02(0.01)	0.04(0.01)
	Coda	1.90(0.02)	1.02(0.01)	0.12(0)	0(0)
	CoCo	1.54(0.04)	0.72(0.02)	0.03(0)	0.03(0.01)
	Vani	1.88(0.02)	1.00(0.01)	0.13(0.01)	0(0)
	Debi	$\gg 10$	$\gg 10$	0.04(0.01)	0.62(0.04)
	Debi+	9.33(0.34)	3.47(0.12)	0.05(0)	0.01(0)
(550,700)	Eric	2.07(0.04)	1.04(0.02)	0.02(0)	0.09(0.01)
	Coda	2.24(0.02)	1.20(0.01)	0.10(0)	0(0)
	CoCo	1.92(0.03)	0.97(0.02)	0.04(0)	0.01(0.01)
	Vani	2.21(0.02)	1.18(0.01)	0.11(0)	0(0)
	Debi	$\gg 10$	$\gg 10$	0.03(0.01)	0.79(0.02)
	Debi+	12.88(0.41)	5.08(0.15)	0.05(0)	0.02(0.01)

threshold of 0.05. Consequently, the FPR of the Debi Lasso remains near 0.05, and a direct ROC curve comparison between Debi Lasso with other methods is inappropriate. As τ increases, it is more difficult to detect find-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0223

ings for all methods, such that both FPR and TPR reduce. Clearly, the left panel shows that Eric Lasso and CoCoLasso have larger areas under the curve (AUC) than Coda and Vanilla Lasso, which is consistent with our conclusions in Simulation I. Furthermore, the right panel of Figure **??** shows that Eric has a larger AUC than others. The AUC values are reported in the figure, confirming that our method outperforms the others. ROC curves under Scenario 2 show the same pattern and are displayed in Section B of the online supplementary materials. Therefore, the special consideration in Eric Lasso to accommodate compositionality does improve its discriminative power over the classic CoCoLasso method developed for high-dimensional Gaussian data. Combining all numerical results in Simulation I and II, the proposed Eric Lasso method stands out in obtaining a more robust regression model with more accurate coefficient estimation and selection even under a misspecified model.

5. A case study

The human gut microbiota has been shown to play a very important role in nutrient digestion and absorption (**?**), and most existing analyses have successfully shown that there exists a significant association between the overall gut microbiome community and body mass index (BMI) using per-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0223

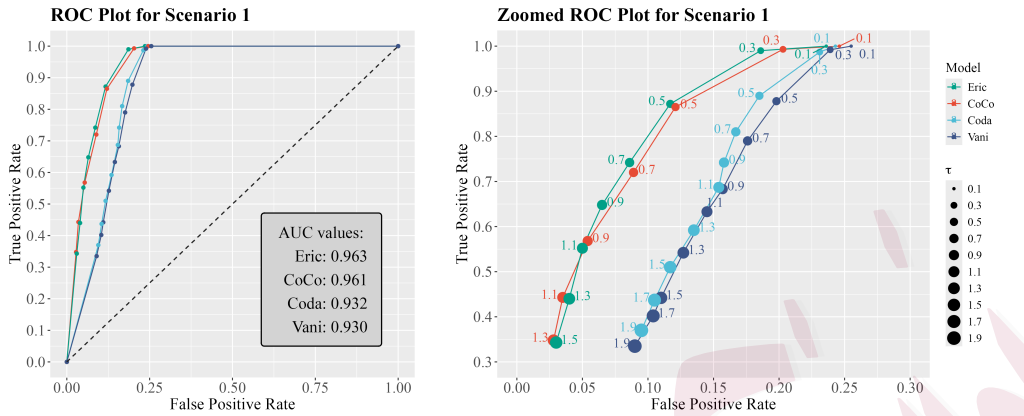

Figure 1: The ROC curve with different τ's under Scenario 1. The left panel is at the original scale, and the right panel zooms in on specific regions to better distinguish different methods. The corresponding ROC-AUC values are displayed in the left panel.

mutational multivariate analysis of variance type of approaches (**?**). However, identifying specific microbial taxa associated with the outcome is more challenging than detecting an existing overall community-level association, partially due to the compositional effect of microbiome data (individual taxa are closely related to or affected by each other). To illustrate the potential usefulness of Eric Lasso, we applied it to investigate BMI-associated gut microbial taxa using data collected in the COMBO study (**?**). A total of 3068 non-singleton operational taxonomic units (OTUs) were detected in the COMBO study. We first aggregated these OTUs into the genus level and then deleted genera that appeared in less than 2 samples, ending with

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0223

up to $p = 80$ genera. After data filtering and quality control, a total of $n = 96$ samples were kept for further analysis, and we followed the previous suggestion (**?**) to transfer these sequence counts into relative abundances for further analysis. To adjust for potential confounding effects, we first regressed BMI on total fat and caloric intake and then took residuals as outcomes for association analysis with gut microbial compositions.

Let $\mathbf{y} = (y_1, \ldots, y_n)$ denote the covariate-adjusted BMI values, U_{ij} denote the relative abundance of the jth genus in the ith sample, $i = 1, \ldots, n, j = 1, \ldots, p$. To mimic potential measurement error or bias in the sequencing procedure, we reserved U_{ij} as the true abundances and used perturbation to generate corrupted abundances. In particular, for each sample i, we first independently simulated each measurement error factor $e_{ij} \sim Unif(0.1, 10)$ and then calculated corrupted abundances as $U_{ij}e_{ij}$, which were further normalized into a compositional vector $(O_{i1}, \ldots, O_{ip})$ for downstream analysis. Let $X_{ij} = log(U_{ij}), Z_{ij} = log(O_{ij})$ and $\mathbf{X} = \{X_{ij}\}, \mathbf{Z} = \{Z_{ij}\}$ be the corresponding design matrix. Our goal is to use the noisy version $(\mathbf{y}, \mathbf{Z})$ to infer the relation of $(\mathbf{y}, \mathbf{X})$. To achieve this goal, we reserved $\mathbf{X}$ from model fitting and used it for evaluation purposes only. Following previous analyses (**??**) on this dataset, we generated bootstrap samples $(\mathbf{y}^b, \mathbf{X}^b, \mathbf{Z}^b)$ of size $n/2$ from the full dataset $(\mathbf{y}, \mathbf{X}, \mathbf{Z})$, and then

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0223

used observations in the bootstrap sample for model training and the other observations not selected in the bootstrap sample for prediction evaluation. This whole procedure (including generating $\mathbf{Z}$ matrix) was repeated for N=100 times and let $(\mathbf{y}^b, \mathbf{X}^b, \mathbf{Z}^b), b = 1, \ldots, N$ denote the bth bootstrap sample. For each bootstrap sample, we fit the model using $(\mathbf{y}^b, \mathbf{Z}^b)$ to obtain $\hat{\boldsymbol{\beta}}^b$ and keep track of observations used for prediction evaluation to make sure they are not contained in the bootstrap samples used for model training. Let $C_{-i} := \{b \in 1, \ldots, N | x_i \notin \mathbf{X}^b\}$ denote the indices of bootstrap samples that do not contain observation i. Then, the average (over N bootstrap samples) leave-one-out (LOO) squared prediction error on observation i is calculated as $\sum_{b \in C_{-i}} (y_i - X_i \hat{\beta}^b)^2 / |C_{-i}|$, where $|C_{-i}|$ denotes the cardinality of set C_{-i}. The mean squared error over all observations is given by: $MSE_{LOO} = n^{-1} \sum_{i=1}^{n} \frac{1}{|C_{-i}|} \sum_{b \in C_{-i}} (y_i - X_i \hat{\beta}^b)^2$. MSE_{LOO} values of Eric, Coda, CoCo, Vani, Debi, and Debi+ are 32.8, 34.9, 35.6, 36.0, 6467.1, and 212.9 respectively. Hence, the proposed Eric Lasso has the best prediction performance among all four Lasso methods being considered.

Recall that the results established in Theorem 1 guarantee performance accuracy in both estimation and selection. We thus compare the selection results of different Lasso methods. Out of the $N = 100$ bootstrap replicates, the frequency of each genus taxon being selected by each Lasso method was

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0223

calculated and presented in Figure **??**. As can be seen in the figure, Coda and Vani Lasso tend to select more taxa than Eric and CoCoLasso, while Debi Lasso exhibits more uniform selection across genera with a sharp peak at the reference genus. For example, 1, 6, 1, 4 taxa are selected by Eric, Coda, CoCo, and Vani Lasso, respectively, over 50 times out of the 100 bootstrap replicates. The reference taxon is selected by Debi for 97% of the replicates, since it calculates the corresponding coefficient by summing up all the other $p-1$ coefficients, i.e. $\hat{\beta}_p = -\sum_{j=1}^{p-1} \hat{\beta}_j$. Consequently, the reference taxon is usually selected even when it may not be associated with the outcome, underscoring the need for permutation-invariant methods. The distribution of taxa relative abundances in this COMBO data is very heterogeneous in that the top 5 most abundant taxa account for 80% of the total abundances of all p taxa. Hence, the unbalanced compositions generated under Scenarios 1 and 3 of Simulation I better mimic the distribution of taxa relative abundances in this COMBO dataset than the more homogeneous case used in Scenario 2 of Simulation I. As observed in Table **??** and Table **??** presented in the previous section, the FPR of Coda and Vani under Scenarios 1 and 3 are 2-5 times of those of Eric and CoCoLasso, which may explain why Coda and Vani Lasso have more findings in this data. In other words, our previous experience in numerical simu-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0223

lations implies that many of these additional findings of Coda and Vani in this COMBO dataset might be false positives. The same phenomenon of spurious findings caused by ignoring measurement errors in microbiome compositional data in statistical analysis has also been observed in the literature (**??**). Finally, the common taxon that is selected more than 50% of time by all methods, excluding Debiased Lasso, is genus *Acidaminococcus* of the *Firmicutes* phylum, which was implicated as important for gut dysbiosis in obese patients (**?**). In summary, the proposed Eric Lasso method can overall provide the best model prediction and variable selection performance in this gut microbiome data analysis collected from the COMBO study.

6. Discussion

Motivated by the ongoing debate on how data contamination can undermine scientific findings in microbiome research (**??**), we propose the Eric Lasso approach to mitigate the effects of measurement errors in compositional data analysis, providing more accurate and reliable statistical results. While the issue of measurement errors or sequence bias in microbiome compositional data is well-documented (**???**), statistical methods addressing this in a high-dimensional regression framework remain underdeveloped.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0223

Figure 2: Selection frequencies of each taxon under N=100 bootstrap replicates. The color indicates how each taxon is identified to be positively related with the outcome (red) or negatively related (green), represented by the sign of the estimated regression coefficient.

The variable correction regularized estimator (**?**) is one of the few existing methods, but it relies on the assumption of accurate total count measurements, which can be significantly biased in practice (**?**). To address this,

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0223

our Eric Lasso method directly targets compositions rather than counts, demonstrating better performance in reducing false positives when true compositions are heterogeneous, as shown in our numerical studies and case analysis. Although this article focuses on microbiome compositional data, the Eric Lasso methodology is broadly applicable to other types of high-dimensional compositional data.

As noted by a reviewer, the theoretical guarantees in Theorem 1 are derived under an i.i.d. measurement-error assumption. Relaxing this assumption is possible in principle but nontrivial, and is therefore beyond the scope of the present paper. Empirically, Eric Lasso displays some robustness to departures from this assumption: in Simulation I (Scenario 3), where we do not impose any specific correlation structure on the measurement errors, the method continues to perform well. The horizontal integration of sequencing datasets by combining samples from multiple studies has received increasing attention and has led to important biological discoveries (**??**). Extending Eric Lasso to that multi-study setting would naturally require introducing a block-structured model for the measurement-error covariance (for example, a block-diagonal covariance with study-specific blocks) and modifying both the estimator and its numerical implementation. While such an extension would likely be useful in practice, it would break the i.i.d. assumption

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0223

and bring substantial theoretical and computational challenges (e.g., new identifiability conditions, concentration bounds for correlated errors, and scalable estimation). This would be a promising direction for future work.

Supplementary Material

The online supplementary materials consist of technical proofs of theorems and additional numerical results.

Acknowledgments

The authors would like to thank the Co-Editor, the Associate Editor, and the anonymous referees for their helpful suggestions and constructive comments. The research of Tan and Xue was supported by the U.S. National Science Foundation (NSF) grant DMS-2210775 and the U.S. National Institutes of Health (NIH) grant 1R01GM152812. The research of Yang was supported by the National Key R&D Program of China 2023YFA1008702 and the National Natural Foundation of China (NSFC) 12301389. The research of Zhan was supported by the National Natural Science Foundation of China (grant no. 12371287).

Penn State University

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0223

E-mail: wkt5100@psu.edu Penn State University

E-mail: lzxue@psu.edu

Renmin University of China

E-mail: yangss@ruc.edu.cn

Southeast University

E-mail: zhanx@seu.edu.cn