Biometrika (2020), **107**, 1, *pp.* 75–92
Printed in Great Britain
doi: 10.1093/biomet/asz062
Advance Access publication 6 December 2019

Multisample estimation of bacterial composition matrices in metagenomics data

By YUANPEI CAO

Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, U.S.A.
yuanpeic@sas.upenn.edu

ANRU ZHANG

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

and HONGZHE LI

Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, U.S.A.
hongzhe@upenn.edu

Summary

Metagenomics sequencing is routinely applied to quantify bacterial abundances in microbiome studies, where bacterial composition is estimated based on the sequencing read counts. Due to limited sequencing depth and DNA dropouts, many rare bacterial taxa might not be captured in the final sequencing reads, which results in many zero counts. Naive composition estimation using count normalization leads to many zero proportions, which tend to result in inaccurate estimates of bacterial abundance and diversity. This paper takes a multisample approach to estimation of bacterial abundances in order to borrow information across samples and across species. Empirical results from real datasets suggest that the composition matrix over multiple samples is approximately low rank, which motivates a regularized maximum likelihood estimation with a nuclear norm penalty. An efficient optimization algorithm using the generalized accelerated proximal gradient and Euclidean projection onto simplex space is developed. Theoretical upper bounds and the minimax lower bounds of the estimation errors, measured by the Kullback–Leibler divergence and the Frobenius norm, are established. Simulation studies demonstrate that the proposed estimator outperforms the naive estimators. The method is applied to an analysis of a human gut microbiome dataset.

Some key words: Microbiome; Nuclear norm penalty; Poisson-multinomial distribution; Proximal gradient descent.

1. Introduction

The human microbiome is the totality of all microbes at different body sites, whose contribution to human health and disease has increasingly been recognized. Recent studies have demonstrated that the microbiome composition varies across individuals due to different health and environmental conditions (Human Microbiome Project Consortium, 2012), and may be associated with

complex diseases such as obesity, atherosclerosis and Crohn's disease (Turnbaugh et al., 2009; Koeth et al., 2013; Lewis et al., 2015). With the development of next-generation sequencing technologies, the human microbiome can be quantified by using direct DNA sequencing of either marker genes or the whole metagenomes. After aligning the sequence reads to the reference microbial genomes, one obtains counts of sequencing reads that can be assigned to a set of bacterial taxa observed in the samples. Such count data provide information about the relative abundance of different bacteria in different samples.

In order to account for the large variability in the total number of reads obtained, the sequencing count data are often normalized into a relative measure of abundance of the taxa observed. Such relative abundances provide information about the bacterial composition. However, due to limited sequencing depth, undersampling and DNA dropouts, some rare microbial taxa might not be captured in the metagenomic sequencing, which results in zero read counts assigned to these taxa. Naive estimation of taxon composition using count normalization leads to excessive zeros, especially for rare taxa. Such a naive estimate can be inaccurate and leads to a suboptimal estimate of taxa diversity. It also causes difficulty in downstream data analysis for compositional data. These zero counts are regarded as rounded zeros, which are not truly zeros, but rather represent observed values due to undersampling or dropouts.

Since the pioneering work of Aitchison (2003), several techniques have been proposed to deal with such rounded zeros (Martín-Fernández et al., 2011) in count data. One approach is to estimate nonzero compositions through a Bayesian-multiplicative model (Martín-Fernández et al., 2014) from the counts. Such a Bayesian method involves a Dirichlet prior distribution as the conjugate prior distribution of a multinomial distribution and a multiplicative modification of the nonzero counts. In fact, the Bayesian-multiplicative method is essentially equivalent to nonparametric imputation, where the zero-replacement values were determined by the parameterizations of the prior distribution. In compositional data analysis, these zero-replacement values are usually chosen as half of the minimum nonzero values. For more details, see Aitchison (2003), Lin et al. (2014), Shi et al. (2016) and Cao et al. (2018a, 2018b). In addition, Cai et al. (2019) recently studied the detection of differential microbial community networks by discretizing the data into a binary Markov random field based on a prespecified abundance threshold.

This paper addresses the problem of estimating microbial composition in positive simplex space from a high-dimensional sparse count table. The observed counts are assumed to follow a Poisson-multinomial model, where (i) the total number of read counts for each individual is a Poisson random variable; (ii) given the total count for each individual, the stratified read counts over different taxa follow a multinomial distribution with the underlying parameters given by a positive composition.

If the compositions across different individuals are combined into a matrix, an approximately low-rank structure on this matrix is indicated by recent observations on the co-occurrence pattern (Faust et al., 2012) and various symbiotic relationships in microbial communities (Woyke et al., 2006; Horner-Devine et al., 2007; Chaffron et al., 2010).

Motivated by the nuclear norm minimization used in the noisy matrix completion problem (Negahban & Wainwright, 2012; Klopp et al., 2015), this paper solves the problem of composition estimation using a nuclear norm regularized maximum likelihood approach. However, it should be emphasized that our approach is very different from the matrix completion problem because the missing mechanism and data generation models are different. The observed zero counts are the result of undersampling or dropouts, rather than the random missingness assumed in the matrix completion literature. Besides, the sparse counts are assumed to be generated from a Poisson-multinomial model, and the focus of this paper is to estimate the underlying composition,

rather than to recover the zero counts. In this framework, the asymptotic upper and minimax lower bounds of the resulting regularized estimator are obtained. Simulations show that the estimator recovers the low-rank composition matrix accurately. Although the observed composition can be seen as the true composition plus noise, and the problem can be roughly framed as a version of matrix denoising, the classic methods in the literature such as singular value thresholding (Candès et al., 2013; Donoho & Gavish, 2014) may not be suitable here due to the heteroscedasticity of different observations in the Poisson-multinomial data.

Our work can be seen as a variant of low-rank Poisson matrix recovery. Salmon et al. (2014) studied nonlocal principal component analysis for Poisson matrix data. A two-step procedure was proposed: after achieving a warm start via regular singular value decomposition, the iterative Newton steps were applied until convergence. Soni & Haupt (2014) considered the Poisson denoising problem with sparse and structured dictionary models. A constrained maximum likelihood method was proposed and the ℓ_2 risk upper bound was developed using complexity-penalized maximum likelihood analyses. Cao & Xie (2016) introduced penalized and constrained likelihood methods for Poisson matrix recovery and Poisson matrix completion, respectively. Theoretical guarantees were developed, including near-matching minimax-optimal bounds for Frobenius norm loss in Poisson matrix completion. However, these results are not directly applicable to our problem. In microbiome 16S rRNA sequencing data analysis, our goal is to estimate the microbial composition rather than their absolute values for each individual. In addition, the hidden sparse dictionary structure imposed by Soni & Haupt (2014) is not likely to hold in our applications. The zero counts in our problem are due to undersampling, which is different from the missing entries in Poisson matrix completion (Cao & Xie, 2016). Theoretically, our proposed penalized nuclear norm minimization estimator is convex and is proved to achieve near-optimal rates of estimation risks in both Kullback–Leibler divergence and the Frobenius norm.

2. A Poisson-multinomial model for microbiome count data

For any integer $n > 0$, we write $[n] = \{1, \ldots, n\}$ and denote $e_i(n)$ as the canonical basis in $\mathbb{R}^n$ with the ith entry as one and the others as zero. We refer to any $u \in \mathbb{R}^p$ as a composition vector if $u \geqslant 0$ and $\sum_{i=1}^p u_i = 1$. For any two composition vectors $u, v \in \mathbb{R}^p$, the Kullback–Leibler divergence is defined as $\mathrm{D_{KL}}(u, v) = \sum_{i=1}^p u_i \log(u_i/v_i)$. For two composition matrices X^* and $\hat{X}$ with each row being a composition vector, let $\mathrm{D}(X^*, \hat{X})$ denote the sum of the Kullback–Leibler divergence between rows of X^* and $\hat{X}$,

$$\mathrm{D}(X^*, \hat{X}) = \sum_{i=1}^n \mathrm{D_{KL}}(X_i^*, \hat{X}_i) = \sum_{i=1}^n \sum_{j=1}^p X_{ij}^* \log(X_{ij}^*/\hat{X}_{ij}). \tag{1}$$

A common amplicon-based sequencing method used to identify and compare bacteria present within a given sample is 16S ribosomal RNA, rRNA, sequencing. In such studies, the sequencing reads are mapped to a set of p known bacterial taxa and the resulting data are summarized as a count matrix $W \in \mathbb{R}^{n \times p}$, where W_{ij}, the (i,j)th entry of W represents the observed read count of taxon j in individual i. For the ith individual, the total count of all taxa, N_i, is determined by the sequencing depth and DNA materials that are modelled as a Poisson random variable as $N_i \sim \mathrm{Po}(\nu_i)$, where ν_i is an unknown positive parameter. Given N_i, it is natural to model the stratified count data over p taxa as a multinomial distribution. Therefore, the proposed Poisson-multinomial model for count-compositional data can be written as

$$N_i \sim \text{Po}(\nu_i), \quad i \in [n],$$

$$f_{X_i^*}(W_{i1}, \ldots, W_{ip} \mid N_i) = \frac{N_i!}{\prod_{j=1}^p W_{ij}!} \prod_{j=1}^p X_{ij}^{*W_{ij}}, \quad i \in [n].$$

Here, $X^* = (X_{ij}^*) \in \mathbb{R}^{n \times p}$ is the unknown taxon composition matrix lying in the positive simplex space $\mathcal{S} = \{X \in \mathbb{R}^{n \times p} \mid X 1_p = 1_n, X > 0\}$, where 1_p is the p-vector of 1s.

Our goal is to estimate X^* based on W. One might attempt to consider the maximum likelihood estimate $\hat{X}^{\text{mle}}$. Conditioning on a fixed number of total count N and ignoring the terms that do not depend on X^*, the negative loglikelihood of the observations is given as

$$\mathcal{L}_N\left(X^*\right) = -N^{-1} \sum_{i=1}^n \sum_{j=1}^p W_{ij} \log X_{ij}^*, \tag{2}$$

where $N = \sum_{i=1}^n N_i = \sum_{i=1}^n \sum_{j=1}^p W_{ij}$ is the total number of observed counts, which follows $\text{Po}(\sum_{i=1}^n \nu_i)$. Without further constraints, minimizing (2) leads to $\hat{X}^{\text{mle}}$, which is the naive count normalization:

$$\hat{X}_{ij}^{\text{mle}} = W_{ij} / \sum_{k=1}^p W_{ik}, \quad i \in [n], \, j \in [p]. \tag{3}$$

Due to dropouts in sample preparation or N not being sufficiently large, $\hat{X}^{\text{mle}}$ often contains a large number of zeros. In microbiome studies these zero counts are treated as rounded zeros, which means that their corresponding compositions are below the detection lower limit. However, the zero counts yield zero estimates of these compositions and cause difficulty in downstream log-ratio-based compositional data analysis (Aitchison, 2003; Lin et al., 2014; Shi et al., 2016; Cao et al., 2018b).

To overcome this difficulty, replacing the zero counts by a below-detection value through either the Bayesian-multiplicative model (Martín-Fernández et al., 2014) or nonparametric imputation (Martín-Fernández et al., 2003) is commonly seen in the literature. These two methods are essentially equivalent, and are widely used in compositional data analysis by replacing the zero counts by 0.5 in the data (Aitchison, 2003; Lin et al., 2014; Shi et al., 2016):

$$\hat{X}^{\text{zr}} \in \mathbb{R}^{n \times p}, \quad \hat{X}_{ij}^{\text{zr}} = \left(W_{ij} \vee 0.5\right) / \sum_{l=1}^p (W_{il} \vee 0.5).$$

However, the pseudo-count 0.5 is chosen arbitrarily and the downstream analysis might be highly sensitive to this value.

On the other hand, under the Poisson-multinomial model, $W_{ij} = EW_{ij} + (W_{ij} - EW_{ij}) = \nu_i X_{ij}^* + (W_{ij} - EW_{ij})$, where $\{\nu_i X_{ij}^*\}_{i,j=1}^{n,p}$ is a low-rank matrix and $(W_{ij} - EW_{ij})$ can be regarded as the noise. Thus, estimating W can be seen as a version of matrix denoising. Singular value thresholding (Donoho & Gavish, 2014; Gavish & Donoho, 2014; Chatterjee, 2015) provides an alternative method for composition estimation. Such an estimator $\hat{X}^{\text{svt}}$ is given as

$$\hat{X}^{\text{svt}} \in \mathbb{R}^{n\times p}, \quad \hat{X}^{\text{svt}}_{ij} = \left(\hat{W}_{ij} \vee 0.5\right) / \sum_{l=1}^{p} (\hat{W}_{il} \vee 0.5), \tag{4}$$

where $W = \sum_k \sigma_k u_k v_k^{\mathrm{T}}$ is the singular value decomposition and

$$\hat{W} = \sum_k I_{\{\sigma_k \geqslant \lambda\}} \cdot \sigma_k u_k v_k^{\mathrm{T}}. \tag{5}$$

However, this singular value thresholding method may not be suitable for our Poisson-multinomial model for the following reasons. First, the Poisson distribution is heteroscedastic according to the values of the means, but $\hat{X}^{\text{svt}}$ is most efficient for homoscedastic noisy data; see, e.g., Donoho & Gavish (2014). Second, since there is no guarantee of positivity in the singular value decomposition, $\hat{W}$ in (5) may contain a large number of negative values. Third, singular value thresholding does not guarantee correct normalization in the sense that the row sums of $\hat{X}^{\text{svt}}$ are typically not 1. More comparisons and discussions are given in § 5.

3. Regularized estimation of the compositional matrix and algorithm

3.1. *Regularized estimation of the compositional matrix*

In order to improve the composition estimate, the approximate low-rank structure of the compositional matrix X^* is explored. The co-occurrence patterns (Faust et al., 2012), various symbiotic relationships in microbial communities (Woyke et al., 2006; Horner-Devine et al., 2007; Chaffron et al., 2010) and samples in similar microbial communities are expected to lead to an approximately low-rank structure of the composition matrix in the sense that the singular values of X^* decay to zero at a fast rate. Such a low-rank structure is further investigated in our real data analysis in § 6, showing empirical evidence for the low-rank compositional matrix. This motivates us to propose a nuclear norm regularized maximum likelihood approach to estimate the composition matrix:

$$\begin{aligned} \hat{X} &= \underset{X \in \mathcal{S}(\alpha_X, \beta_X)}{\arg\min} \; \mathcal{L}_N(X) + \lambda \|X\|_*, \\ \mathcal{S}(\alpha_X, \beta_X) &= \left\{ X \in \mathbb{R}^{n\times p} \mid X 1_p = 1_n, \; \alpha_X/p \leqslant X_{ij} \leqslant \beta_X/p, \text{ for any } (i,j) \in [n] \times [p] \right\}, \end{aligned} \tag{6}$$

where $\lambda > 0$ and $\mathcal{S}(\alpha_X, \beta_X)$ is a bounded simplex space with tuning parameters $0 < \alpha_X \leqslant \beta_X$. The constrained elementwise lower bound, $X_{ij} \geqslant \alpha_X/p$, guarantees the positive sign of the estimator. The elementwise upper bound constraint, $X_{ij} \leqslant \beta_X/p$, is only needed for theoretical analysis.

The proposed estimator (6) is essentially a regularized nuclear norm minimization, which can be solved by either semidefinite programming via an interior-point semidefinite programming solver (Liu & Vandenberghe, 2009; Recht et al., 2010) or a first-order method via templates for first-order conic solvers (Becker et al., 2011). However, the interior-point semidefinite programming solver computes the nuclear norm via a less efficient eigenvalue decomposition, which does not scale well with large n and p. Templates for first-order conic solvers, on the other hand, often result in oscillations or overshoots along the trajectory of the iterations (Su et al., 2016). To achieve a stable and efficient optimization for (6) in the high-dimensional setting, we propose an algorithm based on the generalized accelerated proximal gradient method and Nesterov's scheme (Su et al., 2016).

3.2. *Generalized accelerated proximal gradient method*

We introduce an optimization algorithm for (6) based on Nesterov's generalized accelerated scheme, which follows the formulation of Beck & Teboulle (2009) and the spirit of Su et al. (2016). First, based on the count matrix W, we initialize X_0 and Y_0 as

$$X_0, Y_0 \in \mathbb{R}^{n\times p}, \quad \text{where } (X_0)_{ij} = (Y_0)_{ij} = W_{ij} / \sum_{l=1}^{p} W_{il}.$$

Then X_0, Y_0 are essentially the row-wise normalization of W. Next, we update X_k and Y_k as

$$\begin{aligned} X_k &= \underset{X\in\mathcal{S}(\alpha_X,\beta_X)}{\arg\min}\ 2^{-1}L_k\|X - Y_{k-1} + L_k^{-1}\nabla\mathcal{L}_N(Y_{k-1})\|_{\mathrm{F}}^2 + \lambda\|X\|_*, \\ Y_k &= X_k + (k+\rho-1)^{-1}(k-1)(X_k - X_{k-1}), \end{aligned} \tag{7}$$

until convergence or a maximum number of iterations is reached. Here, $\nabla\mathcal{L}_N$ is the gradient function of $\mathcal{L}_N(X)$:

$$\nabla\mathcal{L}_N(X) \in \mathbb{R}^{n\times p}, \quad \text{with } \{\nabla\mathcal{L}_N(X)\}_{ij} = -(NX_{ij})^{-1}W_{ij},$$

where we treat possible $0/0$ as zero and L_k is the reciprocal of the step size in the kth iteration, which can be chosen by the following line search strategy. Denote

$$\mathcal{F}_L(X,Y) = \mathcal{L}_N(X) - \mathcal{L}_N(Y) - \langle X - Y, \nabla\mathcal{L}_N(Y)\rangle - 2^{-1}L\|X - Y\|_{\mathrm{F}}^2$$

as the error of approximating $\mathcal{L}_N(X)$ by a second-order Taylor expansion with the second-order coefficient set to L. In the kth iteration, we start with the integer $n_k = 1$ and let $L_k = \gamma^{n_k}L_{k-1}$ for some scale parameter $\gamma > 1$, then repeatedly increase $n_k = 1, 2, \ldots$ until $\mathcal{F}_{L_k}(X_k, Y_{k-1}) \leqslant 0$. In the optimization literature $(k-1)/(k+\rho-1)$ and ρ are, respectively, referred to as the momentum term and the friction parameter. We follow the suggestions by Su et al. (2016) and set a high friction rate of $\rho \geqslant 9/2$.

Optimization (7) is the proximal mapping of the nuclear norm function, and it can be solved by singular value thresholding (Cai et al., 2010),

$$X_k = \Pi_{\mathcal{S}(\alpha_X,\beta_X)}\left[\mathcal{D}_{\lambda L_k^{-1}}\left\{Y_{k-1} - L_k^{-1}\nabla\mathcal{L}_N(Y_{k-1})\right\}\right].$$

Here, $\Pi_{\mathcal{S}(\alpha_X,\beta_X)}(X)$ is the Euclidean projection of X onto the positive simplex space $\mathcal{S}(\alpha_X,\beta_X)$, and we postpone a detailed discussion until § 3.3. Provided that $X = U\Sigma V^{\mathrm{T}}$ is the singular value decomposition, $\mathcal{D}_\tau$, the soft-thresholding operator, is defined as

$$\mathcal{D}_\tau(X) = U\mathcal{D}_\tau(\Sigma)V^{\mathrm{T}}, \quad \mathcal{D}_\tau\{\Sigma\} = \mathrm{diag}\{(\sigma_i - \tau)\wedge 0\}.$$

3.3. *Euclidean projection onto the simplex space*

The final step of the algorithm involves Euclidean projection onto the simplex space $\mathcal{S}(\alpha_X,\beta_X)$, a key step in the proposed generalized accelerated proximal gradient method. An efficient algorithm, summarized as Algorithm 1, is used to perform such a projection onto compositional space.

Algorithm 1. Euclidean projection onto compositional space with upper lower bounds.

Input: To-be-projected vector $x \in \mathbb{R}^p$; simplex constraint parameters α_X and β_X.
Output: $\hat{x} = \Pi_{\mathcal{S}(\alpha_X, \beta_X)}(x)$.
Calculate $v = \{x_i - \alpha_X/p, x_i - \beta_X/p\}_{i=1}^p \in \mathbb{R}^{2p}$, and sort it as $v_{(1)} \leqslant \cdots \leqslant v_{(2p)}$.
For $1 \leqslant j \leqslant 2p$, calculate

$$d_j = \sum_{i=1}^{p} \left\{ (x_i - v_{(j)}) \wedge (\beta_X/p) \right\} \vee (\alpha_X/p) - 1.$$

Naturally d_j is a decreasing sequence from nonnegative values to nonpositive values. Find $1 \leqslant j^* \leqslant 2p-1$ such that $d_{j^*} \geqslant 0$ and $d_{j^*+1} \leqslant 0$.
Calculate the final estimator $\hat{x}_i$ as follows:

$$\hat{x}_i = \begin{cases} \beta_X/p & x_i - v_{(j^*)} > \beta_X/p, \\ \alpha_X/p & x_i - v_{(j^*)} \leqslant \alpha_X/p, \\ x_i - v_{j^*} - \dfrac{d_j^*}{|\{i | \alpha_X/p \leqslant x_i - v_{(j^*)} \leqslant \beta_X/p\}|} & \alpha_X/p \leqslant x_i - v_{(j^*)} \leqslant \beta_X/p. \end{cases}$$

Return $\hat{x}$

Proposition 1 provides theoretical guarantees for the performance of Algorithm 1. The central idea of the proof of Proposition 1, see the Supplementary Material, lies in the Karush–Kuhn–Tucker conditions for the optimization problem (8).

PROPOSITION 1. *Recall*

$$\Pi_{\mathcal{S}(\alpha_X, \beta_X)}(x) = \arg\min_{\hat{x}} \|\hat{x} - x\|_2^2 \quad \textit{subject to} \quad \sum_{i=1}^{p} \hat{x}_i = 1,\ \alpha_X/p \leqslant \hat{x}_i \leqslant \beta_X/p. \tag{8}$$

Then $\hat{x}$ calculated from Algorithm 1 exactly equals $\Pi_{\mathcal{S}(\alpha_X, \beta_X)}(x)$.

3.4. *Selection of the tuning parameters*

We propose a variation of K-fold cross validation to select the tuning parameters λ and α_X. We set $\beta_X = p$ to remove the elementwise upper bound constraint.

Let W be the observed count matrix and T_1, T_2 be two sets of grids of positive values. We first randomly split the rows of W into two groups of sizes $n_1 \sim (K-1)n/K$ and $n_2 \sim n/K$ for a total of L times. For the lth split, denote by $I_l, I_l^c \subseteq [n]$ the row index sets of the two groups, respectively. For each $i \in I_l^c$, we further randomly select a subset $J_{i,l} \subseteq [p]$ with cardinality $p_1 \sim p(K-1)/K$. For the lth split, the training set is defined as $\Omega_l = \{(i,j) : (i,j) \in I_l \times [p] \text{ or } i \in I_l^c, j \in J_{i,l}\} \subseteq [n] \times [p]$, which contains both complete and incomplete rows of $[n] \times [p]$. Denote by W_{Ω_l} the training matrix W with all entries in Ω_l^c being set to zero. Next, for each $(\lambda, \alpha_X) \in T_1 \times T_2$, we apply the proposed estimator to W_{Ω_l} with tuning parameters (λ, α_X) and obtain the estimates $\hat{X}^{(l)}(\lambda, \alpha_X)$ $(l = 1, \ldots, L)$. We use the Kullback–Leibler divergence defined in (1) to evaluate the prediction error on the rows of I_l^c,

$$\hat{R}(\lambda, \alpha_X) = \sum_{l=1}^{I} \sum_{i \in I_l^c} \mathrm{D_{KL}} \left\{ \hat{X}_{i\cdot}^{\mathrm{mle}}, \hat{X}_{i\cdot}^{(l)}(\lambda, \alpha_X) \right\},$$

where $\hat{X}^{\mathrm{mle}}$ is the maximum likelihood estimator defined in (3). We choose

$$(\lambda^*, \alpha_X^*) = \underset{\lambda \in T_1, \alpha_X \in T_2}{\arg\min}\ \hat{R}(\lambda, \alpha_X)$$

as the final tuning parameters and obtain the final estimate $\hat{X}$ based on the full dataset. The performance of this tuning parameter selection procedure is verified through numerical studies on both simulated and real datasets in § 5 and § 6.

4. THEORETICAL PROPERTIES

4.1. *Theoretical property under the low-rank matrix assumption*

We investigate the theoretical properties of $\hat{X}$ proposed in § 3; in particular, the upper bounds of the estimation accuracy for the whole composition matrix are provided in Theorems 1 and 3, and the lower bound results are given in Theorem 2. These results establish the optimal recovery rate over a certain class of low-rank composition matrices. Additionally, we study the diversity index estimation and present the upper bound results in Corollary 1.

Let $R_i = \nu_i / \sum_{j=1}^n \nu_j$ for $i \in [n]$, which quantifies the proportion of the total count for the ith subject. We establish the upper bound for $\hat{X}$ in the Frobenius norm error and the average Kullback–Leibler divergence. The following theorem gives an upper bound result over a class of bounded low-rank composition matrices:

$$\mathbb{B}_0(r, \alpha_X, \beta_X) = \{X \in \mathcal{S}(\alpha_X, \beta_X) \mid \mathrm{rank}(X) \leqslant r\},$$

where $\mathcal{S}(\alpha_X, \beta_X)$ is the set of bounded composition matrices defined in (6).

THEOREM 1. *Assume there exist constants α_R, β_R, α_X and β_X such that, for any $i \in [n]$, $\alpha_R/n \leqslant R_i \leqslant \beta_R/n$. Suppose $X^* \in \mathbb{B}_0(r, \alpha_X, \beta_X)$. Conditioning on fixed N, suppose that $N \geqslant (n+p)\log(n+p)$ and the tuning parameter is selected as*

$$\lambda = \delta \left\{ \frac{\beta_R}{\alpha_X^2} \frac{p(n \vee p)\log(n+p)}{nN} \right\}^{1/2} \tag{9}$$

with some constant $\delta \geqslant 7$. Then, there exist constants $C_1(\alpha_X, \alpha_R, \beta_X, \beta_R)$ and $C_2(\alpha_X, \alpha_R, \beta_X, \beta_R)$ that only depend on α_X, α_R, β_X and β_R such that $\hat{X}$ in (6) satisfies

$$\frac{p}{n} \left\| \hat{X} - X^* \right\|_{\mathrm{F}}^2 \leqslant C_1(\alpha_X, \alpha_R, \beta_X, \beta_R) \frac{(n+p) r \log(n+p)}{N}, \tag{10}$$

$$\frac{1}{n} \mathrm{D}(X^*, \hat{X}) \leqslant C_2(\alpha_X, \alpha_R, \beta_X, \beta_R) \frac{(n+p) r \log(n+p)}{N} \tag{11}$$

with probability at least $1 - 3(n+p)^{-1}$. In particular, $C_1(\alpha_X, \alpha_R, \beta_X, \beta_R)$ and $C_2(\alpha_X, \alpha_R, \beta_X, \beta_R)$ satisfy

$$C_1(\alpha_X, \alpha_R, \beta_X, \beta_R) = C\beta_X^4(\beta_X \vee \beta_R)/(\alpha_R^2 \alpha_X^4),$$

$$C_2(\alpha_X, \alpha_R, \beta_X, \beta_R) = \begin{cases} C\beta_X^2(\beta_X \vee \beta_R)/(\alpha_R^2 \alpha_X^3) & \text{if } N < 6(n+p)^2 \log(n+p)/(\alpha_X \alpha_R); \\ C\beta_X^5 \beta_R/(\alpha_R^2 \alpha_X^6) & \text{if } N \geqslant 6(n+p)^2 \log(n+p)/(\alpha_X \alpha_R), \end{cases}$$

where $C > 0$ is a uniform constant that does not depend on X^*, r, N, p, n, α_X, α_R, β_X or β_R.

Remark 1. In contrast to its population counterpart $\nu := \sum_{i=1}^n \nu_i$, the total count N is an observable value, we therefore choose to present the results of Theorem 1 conditioning on the fixed number N. If one replaces all N by ν in the conditions and conclusions of this theorem, the unconditional results hold similarly.

Remark 2. The coefficient p/n in the Frobenius norm error in (10) is used to calibrate the rate effect from $\mathcal{S}(\alpha_X, \beta_X)$. For any X^* and $\hat{X} \in \mathcal{S}(\alpha_X, \beta_X)$, $\|\hat{X} - X^*\|_{\mathrm{F}}^2 \leqslant n(\beta_X - \alpha_X)/p$.

Remark 3. For technical purposes, we have imposed the entrywise upper and lower bounds of α_X, β_X, α_R and β_R in Theorem 1. These conditions are mainly for regularizing the gradient of the likelihood function $\mathcal{L}_N$ and facilitate the follow-up analysis; in particular, see (S28) and the proof of Theorem 1 in the Supplementary Material. In fact, the entrywise upper and lower bounds appear widely in theoretical works for a wide range of Poisson inverse problems, especially for ones with minimax optimality. Examples include, but are not limited to, Poisson sparse regression (Li & Raskutti, 2018, Assumption 2.1; Jiang et al., 2015, Assumption 2.1), Poisson matrix completion (Cao & Xie, 2016, Equation (10)) and the point autoregressive model (Hall et al., 2017, Definition of $\mathcal{A}_s$ on p. 4).

Conditioning on the fixed N, the count matrix W follows a multinomial distribution: $(W_{ij}, 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant p) \sim \mathrm{Mu}\{N, (R_i X^*_{ij}, 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant p)\}$, where $R_i = \nu_i / \sum_{j=1}^n \nu_j$ represents the row probability and the composition X^*_{ij} represents the column probability. Defining a probability matrix by the Hadamard, or entrywise, product $\Pi = (R1_p^{\mathrm{T}}) \circ X^*$, we rewrite $W = \sum_{k=1}^N E_k$, where E_k are independent and identically distributed copies of a Bernoulli random matrix E that satisfies $\mathrm{pr}\{E = e_i(n)e_j^{\mathrm{T}}(p)\} = \Pi_{ij}$ and the total count N represents the number of copies. This product-type sampling distribution Π is widely used in the matrix completion literature (Negahban & Wainwright, 2012; Lafond et al., 2014; Klopp et al., 2014, 2015). A key step in the proof of Theorem 1 is to bound the weighted Kullback–Leibler divergence $\sum_{i=1}^n \sum_{j=1}^p R_i X^*_{ij} \log(X^*_{ij}/\hat{X}_{ij})$. We apply a peeling scheme by partitioning the set of all possible values of $\hat{X}$, and then derive estimation loss upper bounds for each of these subsets based on concentration inequalities, including the matrix Bernstein inequality (Lemma S6) and an empirical process version of Hoeffding's inequality (Bühlmann & Van De Geer, 2011, Theorem 14.2). The techniques are related to recent work on matrix completion (Negahban & Wainwright, 2012), although our problem set-up, method and sampling procedure are all distinct from matrix completion.

Theorem 2 on the minimax lower bounds shows that the upper bound in Theorem 1 is nearly rate optimal.

THEOREM 2. *Conditioning on fixed N, if $2 \leqslant r \leqslant p/2$, there exist constants C_1 and C_2 which only depend on α_X, β_X, α_R and β_R such that*

$$\inf_{\hat{X}} \sup_{\substack{X^* \in \mathbb{B}_0(r, \alpha_X, \beta_X) \\ \alpha_R/n \leqslant R_i \leqslant \beta_R/n}} \frac{p}{n} E\left(\|\hat{X} - X^*\|_{\mathrm{F}}^2\right) \geqslant C_1 \frac{(n+p)r}{N},$$

$$\inf_{\hat{X}} \sup_{\substack{X^* \in \mathbb{B}_0(r, \alpha_X, \beta_X) \\ \alpha_R/n \leqslant R_i \leqslant \beta_R/n}} \frac{1}{n} E\left\{D(X^*, \hat{X})\right\} \geqslant C_2 \frac{(n+p)r}{N}.$$

4.2. *Theoretical property under the approximate low-rank matrix assumption*

We consider the following class of approximately low-rank composition matrices with singular values of X^* belonging to an ℓ_q ball,

$$\mathbb{B}_q\left(\rho_q, \alpha_X, \beta_X\right) = \left\{X \in \mathcal{S}(\alpha_X, \beta_X) \mid \sum_{i=1}^{n\wedge p} |\sigma_i(X)|^q \leqslant \rho_q\right\},$$

where $0 \leqslant q \leqslant 1$. In particular, if $q = 0$ the ℓ_0 ball $\mathbb{B}_0(\rho_0, \alpha_X, \beta_X)$ corresponds to the set of bounded composition matrices with rank at most ρ_0. In general, we have the following upper bound result.

THEOREM 3. *Assume there exist constants α_R and β_R such that, for any $i \in [n]$, $\alpha_R/n \leqslant R_i \leqslant \beta_R/n$. The tuning parameter is selected by (9). Conditioning on fixed N, if $N \geqslant (n+p)\log(n+p)$ and $N = O\left\{\rho_q p^{q/2}(n+p)^{2+q/2}\log(n+p)/n^{q/2}\right\}$, then for any composition $X^* \in \mathbb{B}_q\left(\rho_q, \alpha_X, \beta_X\right)$ the estimator $\hat{X}$ in (6) satisfies*

$$\frac{p}{n}E\left\{\|\hat{X} - X^*\|_{\mathrm{F}}^2\right\} \leqslant C_1\frac{\rho_q p^{q/2}}{n^{q/2}}\left\{\frac{(n+p)\log(n+p)}{N}\right\}^{1-q/2}, \tag{12}$$

$$\frac{1}{n}E\left\{\mathrm{D}(X^*, \hat{X})\right\} \leqslant C_2\frac{\rho_q p^{q/2}}{n^{q/2}}\left\{\frac{(n+p)\log(n+p)}{N}\right\}^{1-q/2}. \tag{13}$$

Remark 4. The rates of convergence of (13) and (12) reduce to the exact low-rank case when $q = 0$ and $\rho_0 = r$.

4.3. *Estimation of the diversity index*

Various microbial diversity measures are often used to quantify the composition of microbial communities; see, e.g., Haegeman et al. (2013). Given $X \in \mathbb{R}^{n\times p}$ representing p-taxa compositions across n individuals, two widely used measurements of microbial community diversity include Shannon's index, $H_{\mathrm{sh}}(X_i) = -\sum_{j=1}^{p} X_{ij}\log X_{ij}$, $1 \leqslant i \leqslant n$, and Simpson's index, $H_{\mathrm{sp}}(X_i) = \sum_{j=1}^{p} X_{ij}^2$, $1 \leqslant i \leqslant n$, where $\{H_{\mathrm{sh}}(X_i)\}_{i=1}^n$ and $\{H_{\mathrm{sp}}(X_i)\}_{i=1}^n$ are n-dimensional vectors, each component measuring the richness and evenness of the microbial community in an individual. A higher value of Shannon's index, or a lower value of Simpson's index, reflects a more even distribution among different taxa.

We estimate various diversity indices by plugging the proposed estimator $\hat{X}$ into the indices defined above. The following corollary provides the upper bounds of the mean squared errors of these estimators when $X^* \in \mathbb{B}_q\left(\rho_q, \alpha_X, \beta_X\right)$.

COROLLARY 1. *Assume that the assumptions in Theorem 3 hold and that the tuning parameter is selected by (9). For any constants $0 < \alpha_X < 1 < \beta_X$ and $X^* \in \mathbb{B}_q\left(\rho_q, \alpha_X, \beta_X\right)$, there exist constants C_1 and C_2 that do not depend on n, N, p or r such that the estimate $\hat{X}$ in (6) satisfies*

$$\frac{1}{n}\sum_{i=1}^{n} E\left\{H_{sh}(\hat{X}_i) - H_{sh}(X_i^*)\right\}^2 \leqslant C_1\frac{\rho_q(\log p)^2 p^{q/2}}{n^{q/2}}\left\{\frac{(n+p)\log(n+p)}{N}\right\}^{1-q/2}, \tag{14}$$

$$\frac{1}{n}\sum_{i=1}^{n} E\left\{H_{sp}(\hat{X}_i) - H_{sp}(X_i^*)\right\}^2 \leqslant C_2\frac{\rho_q}{p^{2-q/2}n^{q/2}}\left\{\frac{(n+p)\log(n+p)}{N}\right\}^{1-q/2}. \tag{15}$$

Table 1. *Means of various performance measures for* $\hat{X}$, $\hat{X}^{\mathrm{zr}}$ *and* $\hat{X}^{\mathrm{svt}}$ *in the low-rank model over 100 replications*

	$p = 50$			$p = 100$			$p = 200$		
γ	$\hat{X}$	$\hat{X}^{\mathrm{zr}}$	$\hat{X}^{\mathrm{svt}}$	$\hat{X}$	$\hat{X}^{\mathrm{zr}}$	$\hat{X}^{\mathrm{svt}}$	$\hat{X}$	$\hat{X}^{\mathrm{zr}}$	$\hat{X}^{\mathrm{svt}}$
				Squared Frobenius norm error $(\times 10^{-2})$					
1	40.70	95.01	84.33	26.74	68.65	55.20	19.00	48.98	36.24
2	35.08	87.79	75.11	26.49	63.29	48.97	18.38	44.63	30.10
3	35.37	80.91	67.98	27.77	56.80	42.22	18.99	40.62	25.72
4	37.09	74.25	61.75	24.65	53.11	37.44	18.24	36.87	22.92
5	36.47	67.76	55.81	25.22	49.20	34.03	18.22	34.26	20.54
				Average Kullback–Leibler divergence $(\times 10^{-2})$					
1	4.31	19.04	16.03	3.77	19.68	14.16	3.82	20.02	12.73
2	3.25	18.65	14.73	3.78	19.15	12.47	3.53	18.90	9.87
3	3.33	16.47	12.36	4.04	16.56	9.80	3.82	16.59	7.31
4	3.62	14.41	10.45	3.27	14.59	7.71	3.51	14.39	5.85
5	3.58	12.36	8.69	3.40	12.77	6.42	3.54	12.57	4.69
				Shannon index mean squared errors $(\times 10^{-3})$					
1	3.92	19.90	13.50	3.25	19.82	8.88	3.00	18.63	6.35
2	2.21	23.31	12.52	4.19	18.34	5.99	2.25	21.67	4.24
3	2.66	18.96	8.78	3.56	16.91	4.07	2.54	17.84	2.18
4	2.83	14.72	6.03	2.31	14.65	2.67	2.00	14.41	1.43
5	2.60	11.78	4.58	2.20	12.46	1.95	2.12	11.60	0.89
				Simpson index mean squared errors $(\times 10^{-6})$					
1	5.93	55.73	35.62	1.21	14.27	5.71	0.28	3.45	0.94
2	3.25	51.24	25.20	1.50	11.70	3.57	0.21	3.10	0.47
3	3.94	40.86	17.77	1.35	9.12	2.01	0.23	2.40	0.24
4	4.12	31.49	12.31	0.82	7.87	1.31	0.18	1.81	0.15
5	3.68	23.85	8.96	0.77	6.36	0.89	0.19	1.45	0.10

$\hat{X}$, proposed estimator; $\hat{X}^{\mathrm{zr}}$, zero-replacement estimator; $\hat{X}^{\mathrm{svt}}$, singular value thresholding estimator.

In addition, results for the exact low-rank case correspond to $q = 0$ *and* $\rho_0 = r$.

Remark 5. Jiao et al. (2017) considered the maximum likelihood estimation of functionals, particularly Shannon's and Simpson's indices, for discrete distributions. According to their results, if $R_i \in [\alpha_R/n, \beta_R/n]$ for any $i \in [n]$, conditioning on fixed N, we can derive the following rate of convergence for $\hat{X}^{\mathrm{mle}}$:

$$\frac{1}{n}\sum_{i=1}^{n} E\left\{H_{sh}(\hat{X}_i^{\mathrm{mle}}) - H_{sh}(X_i^*)\right\}^2 \asymp \frac{n^2p^2}{N^2} + \frac{n(\log p)^2}{N}, \tag{16}$$

$$\frac{1}{n}\sum_{i=1}^{n} E\left\{H_{\mathrm{sp}}(\hat{X}_i^{\mathrm{mle}}) - H_{\mathrm{sp}}(X_i^*)\right\}^2 \asymp \frac{n}{N}. \tag{17}$$

Wu & Yang (2016) studied the minimax-optimal estimation of Shannon's index. They showed that a best polynomial approximation estimator $\hat{X}^{\mathrm{bpa}}$ achieves the following sharper rate than the maximum likelihood estimator:

$$\frac{1}{n}\sum_{i=1}^{n} E\left\{H_{sh}(\hat{X}^{\mathrm{bpa}} - \hat{X}^*)\right\}^2 \asymp \frac{n^2p^2}{N^2(\log p)^2} + \frac{n(\log p)^2}{N}.$$

Table 2. *Means of various performance measures for $\hat{X}$ and $\hat{X}^{\mathrm{zr}}$ in the full-rank model over 100 replications*

	$p = 50$		$p = 100$		$p = 200$	
γ	$\hat{X}$	$\hat{X}^{\mathrm{zr}}$	$\hat{X}$	$\hat{X}^{\mathrm{zr}}$	$\hat{X}$	$\hat{X}^{\mathrm{zr}}$
	Squared Frobenius norm error $(\times 10^{-2})$					
1	26.60	94.74	14.91	65.64	10.42	46.65
2	25.60	87.48	14.35	62.72	9.82	44.13
3	25.09	80.43	13.24	57.11	9.47	40.50
4	24.12	73.73	12.64	52.58	8.79	37.15
5	23.59	67.64	12.41	49.11	8.65	34.73
	Average Kullback–Leibler divergence $(\times 10^{-2})$					
1	1.82	18.67	1.14	18.04	1.11	18.17
2	1.71	18.46	1.07	19.51	0.99	19.20
3	1.66	16.58	0.90	17.03	0.91	16.71
4	1.52	14.44	0.82	14.97	0.78	14.99
5	1.43	12.66	0.78	13.09	0.76	13.23
	Shannon index mean squared errors $(\times 10^{-3})$					
1	0.66	29.13	0.15	31.15	0.12	31.04
2	0.55	27.22	0.16	31.35	0.13	29.97
3	0.80	21.03	0.17	23.89	0.14	23.66
4	0.65	17.20	0.15	19.20	0.08	19.22
5	0.47	14.49	0.13	15.90	0.08	16.00
	Simpson index mean squared errors $(\times 10^{-6})$					
1	1.02	71.18	0.06	17.77	0.01	4.43
2	0.83	56.82	0.06	15.40	0.01	3.70
3	1.14	42.53	0.06	11.31	0.01	2.88
4	0.96	33.12	0.06	8.72	0.01	2.17
5	0.58	25.82	0.05	7.21	0.01	1.78

$\hat{X}$, proposed estimator; $\hat{X}^{\mathrm{zr}}$, zero-replacement estimator.

Compared with diversity estimation via $\hat{X}^{\mathrm{mle}}$ or $\hat{X}^{\mathrm{bpa}}$, our proposed diversity estimator achieves a sharper bound in the estimation error when the number of total counts increases at a small rate. In particular, when $n = p = d$ and rank $(X^*) = r$, the rate of convergence of both Shannon's and Simpson's index provided by Corollary 1 is sharper than (16) and (17), respectively, if the number of total counts increases under a certain rate: $N = O\{d^2 r \log d \vee d^3 r^{-1} (\log d)^{-5}\}$, which is close to the required condition in Corollary 1.

5. SIMULATION STUDIES

We now evaluate the numerical performances of the proposed estimator $\hat{X}$ under various settings. Since the Poisson-multinomial model is equivalent to the multinomial model when the total count is fixed, the count matrix W is generated as follows. Let $U \in \mathbb{R}^{n \times r}$ be the absolute values of an independent and identically distributed standard normal matrix. In order to simulate correlated compositional data arising from metagenomics, let $V = V_1 + V_2 \in \mathbb{R}^{p \times r}$, where

$$(V_1)_{ij} = \begin{cases} 1 & i = j, \\ 1 & i \neq j \text{ with probability } 0.3, \\ 0 & i \neq j \text{ with probability } 0.7; \end{cases} \qquad (V_2)_{ij} \sim N(0, 10^{-3}).$$

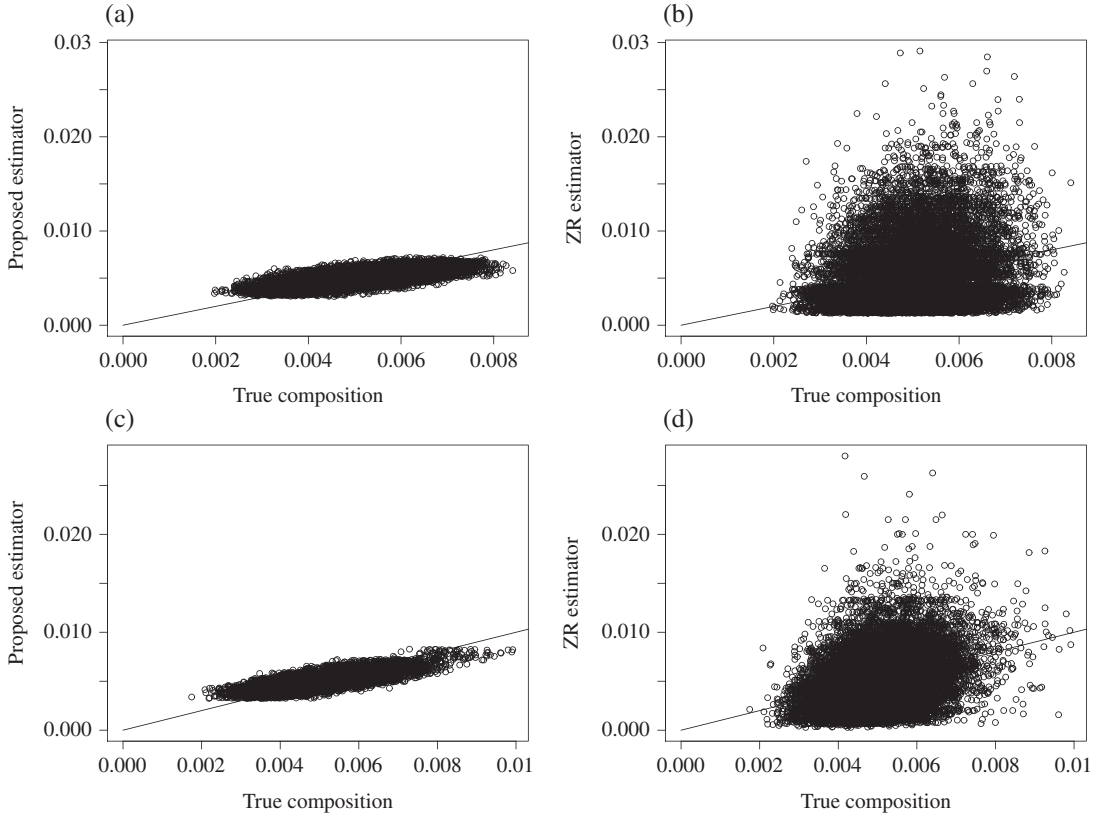

Fig. 1. Scatterplot comparing true composition and estimated composition in the full-rank model with $p = 200$ for different sequencing depths. (a) and (c): the proposed estimator $\hat{X}$. (b) and (d): zero-replacement estimator. Top: $\gamma = 1$. Bottom: $\gamma = 5$. Line: $y = x$.

The true composition matrix is generated as $X^*_{ij} = Z_{ij} / \sum_{k=1}^{p} Z_{ik}$, where $Z = UV^{\mathrm{T}}$. Since this procedure may produce nonpositive values in X^* by a small chance, this is repeated until a positive matrix X^* is generated. In order to account for the heterogeneity of the total count across different samples, we generate $R_i = P_i / \sum_{k=1}^{n} P_k$ with $P_i \sim \mathrm{Un}[1, 10]$ for each individual $i \in [n]$. Based on R_i and X^*, the read counts W are generated from the multinomial model, i.e., $W^*_i \sim \mathrm{Mu}(N_i; X^*_i)$, where $N_i = \gamma npR_i$, $\gamma = 1, 2, 3, 4, 5$. The sample size and the number of taxa are set as $n = 100$, $p \in \{50, 100, 200\}$ and $r = 20$, low-rank model, or $r = n \wedge p$, the full-rank model. These parameters are chosen to mimic the data dimensions of typical microbiome studies.

The proposed nuclear norm regularized maximum likelihood estimator $\hat{X}$ is applied to recover X^*. The simulations are repeated 100 times, and the tuning parameters (λ, α_X) are selected based on the data-driven procedure. The estimation performances are evaluated by the means of average loss in squared Frobenius norm $\|\hat{X} - X^*\|_{\mathrm{F}}^2$, average Kullback–Leibler divergence $n^{-1}\,\mathrm{D}(X^*, \hat{X})$ and the mean squared errors for the estimates of Shannon's and Simpson's indices. The results are compared with the standard zero-replacement estimator $\hat{X}$, in both exact low-rank and full-rank models, and the singular value thresholding estimator $\hat{X}^{\mathrm{svt}}$ in (4), only in the exact low-rank model due to the difficulties in selecting r.

The results are summarized in Tables 1 and 2 for the low-rank and full-rank compositional matrix, respectively. The proposed estimator $\hat{X}$ outperforms the zero-replacement estimator $\hat{X}^{\mathrm{zr}}$ and singular value thresholding estimator $\hat{X}^{\mathrm{svt}}$ in almost all the settings. In particular, the diversity

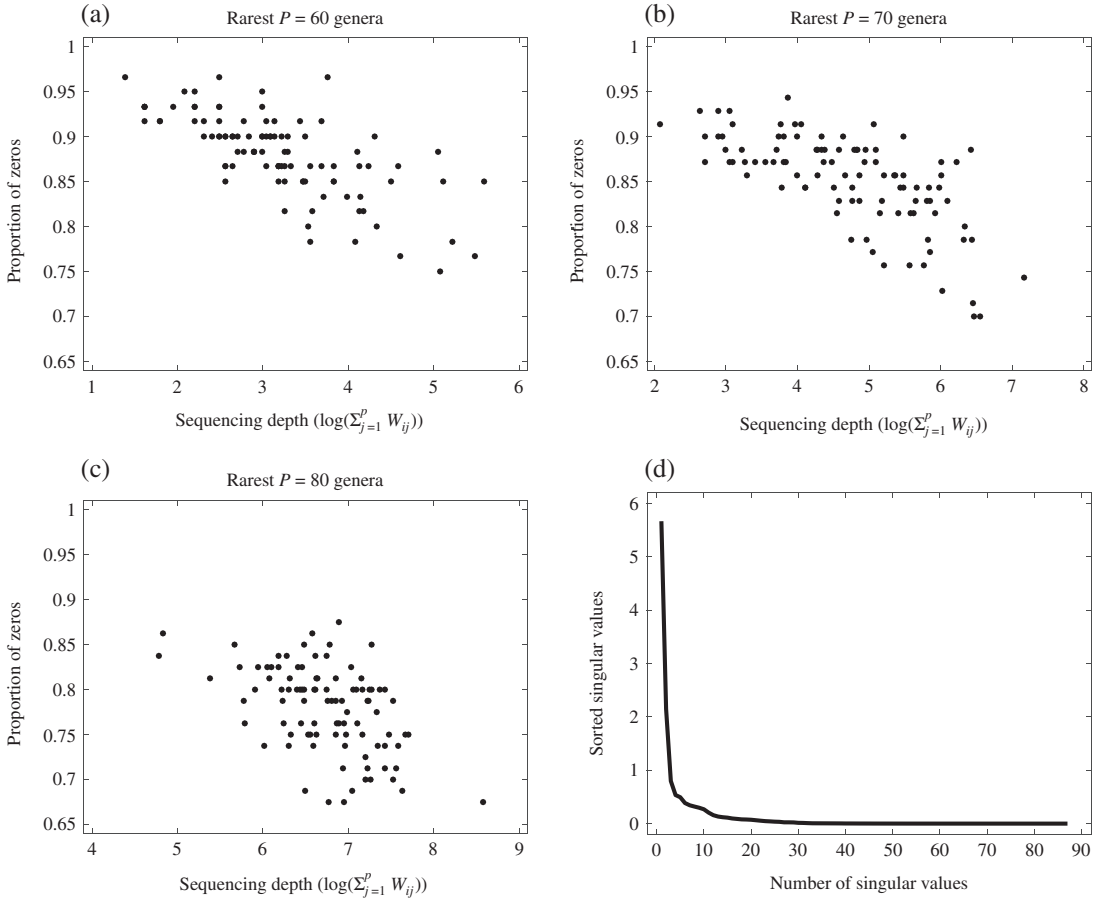

Fig. 2. Analysis of the gut microbiome dataset. (a)–(c): Proportion of zero count versus the total reads count for each individual, indicating that many observed zeros are due to undersampling. (d): Decay of singular values d_{ii} based on the singular value decomposition of $\hat{X}^{\text{mle}} = UDV^{\text{T}}$, indicating the low-rank structure of the compositional matrix.

index estimates based on the proposed estimator uniformly outperform other methods by a large margin. These results are consistent across different model dimensions even when X^* is full rank. In addition, the difference between the loss of $\hat{X}$ and $\hat{X}^{\text{zr}}$ becomes more significant for smaller γ, i.e., when the number of total read counts is small. Therefore, our method enjoys greater improvement than the traditional methods, especially when the sequencing depth is limited.

To further compare the resulting estimates, Fig. 1 shows a scatterplot comparing the true composition matrix X^* and the estimated composition matrix $\hat{X}$ for a randomly chosen simulated dataset in the low-rank setting with $p = 200$, and $\gamma = 1$ and $\gamma = 5$, respectively. Although $\hat{X}$ is slightly biased due to the nuclear norm penalty in the estimation, it still greatly outperforms the commonly used zero-replacement estimator $\hat{X}^{\text{zr}}$. Estimates from singular value thresholding are not compared since it can result in negative estimates.

6. GUT MICROBIOME DATA ANALYSIS

The gut microbiome plays an important role in regulating metabolic functions and influences human health and disease (Human Microbiome Project Consortium, 2012). We apply the proposed method to the cross-sectional study of diet and stool microbiome composition (Wu et al., 2011).

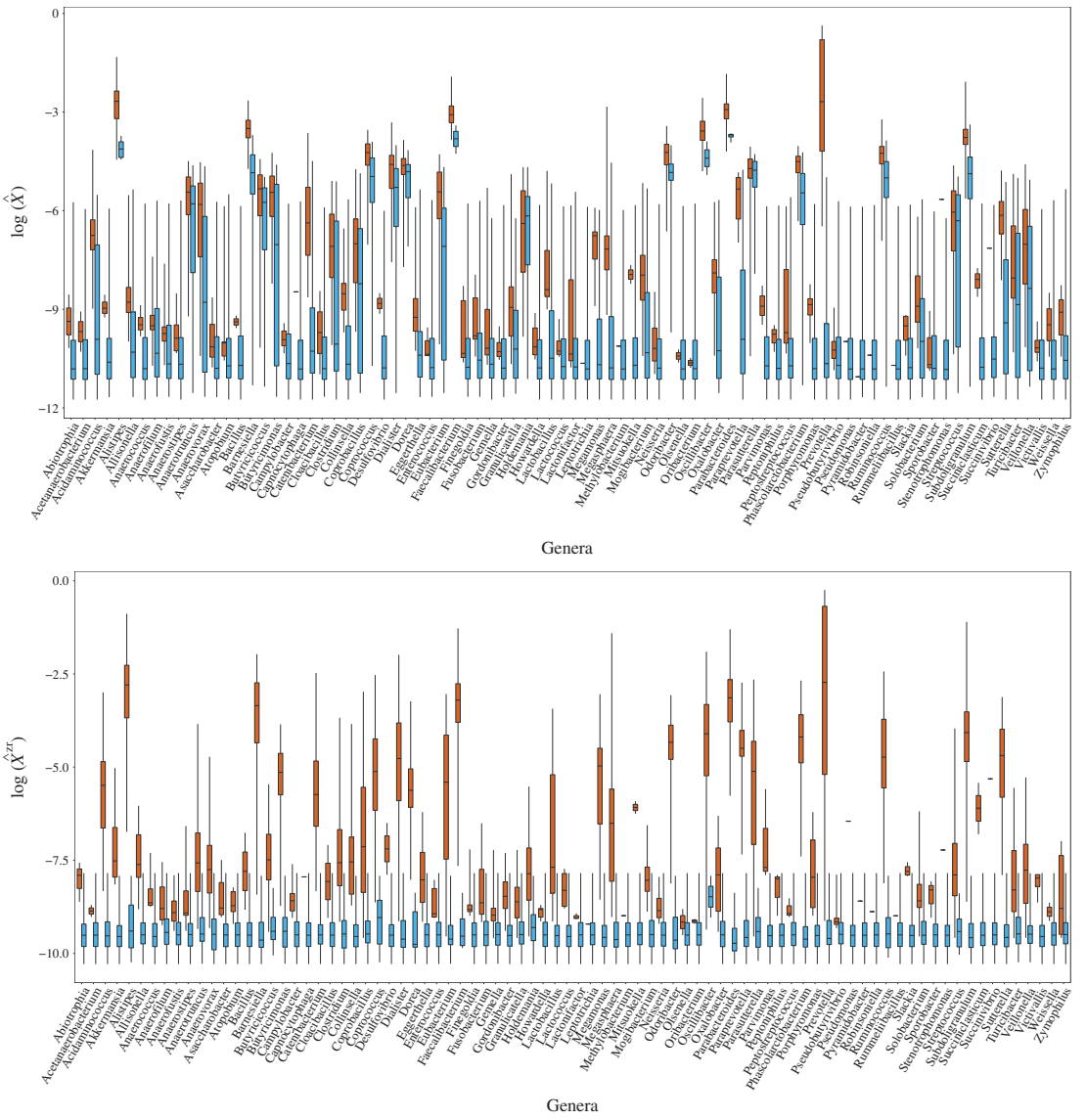

Fig. 3. Boxplots of the estimated compositions for the genera corresponding to nonzero observations (red) and zero observations (blue) in the gut microbiome dataset. Top panel: The proposed estimator $\hat{X}$. Bottom panel: The zero-replacement estimator $\hat{X}^{\mathrm{zr}}$.

In this study, DNA from stool samples of 98 healthy volunteers was analysed by 454/Roche pyrosequencing of 16S rRNA gene segments and yielded an average of 9265 reads per sample, with a standard deviation of 386, which led to identification of 3068 operational taxonomic units and 87 bacterial genera that were presented in at least one sample. Figures 2(a)–(c) show the proportion of zero counts versus total number of sequencing reads for each sample. It is clear that the samples with a smaller number of read counts often produced more zeros in the genus counts, indicating that many observed zeros are likely due to undersampling. It is therefore reasonable to assume that the true compositions of these rare genera are positive. Figure 2(d) shows the decay of singular values of $\hat{X}^{\mathrm{mle}}$, indicating an approximate low-rank composition matrix.

The proposed regularized maximum likelihood estimator $\hat{X}$ is applied to the count matrix of these $p = 87$ bacterial genera over $n = 98$ samples. As a comparison, the traditional

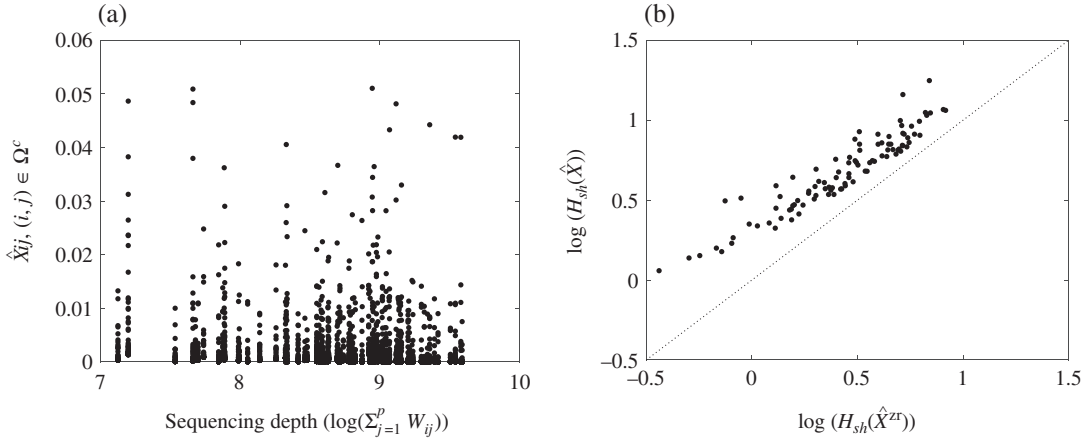

Fig. 4. Analysis of the gut microbiome dataset. (a) The estimated compositions $\hat{X}$ for genera corresponding to zero observations (Ω^c) versus the sequencing depth. (b) Logarithm of the estimated Shannon index from regularized estimator $\hat{X}$ versus zero-replacement estimator $\hat{X}^{zr}$, where each dot represents one sample.

zero-replacement estimator $\hat{X}^{zr}$ is also calculated and compared. To compare the results, define $\Omega = \{(i,j) \in [n] \times [p] \mid W_{ij} > 0\}$ and Ω^c as the support and the zero count indices set of W, respectively. The top panel of Fig. 3 shows boxplots of the estimated compositions $\hat{X}$ excluding three common genera *Bacteroides*, *Blautia* and *Roseburia* that have been observed in all individuals. For $\hat{X}$, the observed nonzero compositions have an effect on estimating the compositions of the genera that were observed as zeros. The estimated compositions of $\hat{X}$ in Ω^c tend to shrink towards those in Ω. In contrast, the zero-replacement estimator $\hat{X}^{zr}$, see the bottom panel of Fig. 3, provides almost the same estimates for all samples/taxa in Ω^c, and $\{W_{ij}\}_{(i,j)\in\Omega}$, i.e., the nonzero counts, have little effect on $\{\hat{X}_{ij}^{zr}\}_{(i,j)\in\Omega^c}$.

Furthermore, as shown in Fig. 4(a), $\{\hat{X}_{ij}\}_{(i,j)\in\Omega^c}$ tends to decrease as the total number of counts for each individual, i.e., N_i, increases. This is reasonable, as the zero counts are more likely to correspond to the very rare taxa when the sequencing gets deeper. However, in contrast to the simple zero-replacement estimates, sequencing depth is not the only factor that determines the compositions of the taxa with zero counts. The compositional data observed in samples with nonzero counts also contribute to the final estimates.

Figure 4(b) shows the estimates of Shannon's index for each individual using the proposed estimator $\hat{X}$ versus the index based on the zero-replacement estimator $\hat{X}^{zr}$, indicating that $\hat{X}^{zr}$ produces a uniformly smaller Shannon index than $\hat{X}$. This is mainly due to the fact that $\hat{X}^{zr}$ replaces all nonpositive counts with the small value 0.5, which yields an uneven distribution between taxa in Ω and Ω^c, resulting in lower diversity among all taxa and a smaller Shannon index.

ACKNOWLEDGEMENT

We thank the editor, associate editor and two referees for their insightful comments. This research was supported by the National Institutes of Health and the National Science Foundation.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes proofs of all the lemmas and theorems.

References

Aitchison, J. (2003). *The Statistical Analysis of Compositional Data*. Caldwell, NJ: Blackburn Press.

Beck, A. & Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imag. Sci.* **2**, 183–202.

Becker, S. R., Candès, E. J. & Grant, M. C. (2011). Templates for convex cone problems with applications to sparse signal recovery. *Math. Prog. Comput.* **3**, 165–218.

Bühlmann, P. & Van De Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. New York: Springer Science & Business Media.

Cai, J.-F., Candès, E. J. & Shen, Z. (2010). A singular value thresholding algorithm for matrix completion. *SIAM J. Optim.* **20**, 1956–82.

Cai, T. T., Li, H., Ma, J. & Xia, Y. (2019). Differential Markov random field analysis with an application to detecting differential microbial community networks. *Biometrika* **106**, 401–16.

Candès, E. J., Sing-Long, C. A. & Trzasko, J. D. (2013). Unbiased risk estimates for singular value thresholding and spectral estimators. *IEEE Trans. Sig. Proces.* **61**, 4643–57.

Cao, Y., Lin, W. & Li, H. (2018a). Large covariance estimation for compositional data via composition-adjusted thresholding. *J. Am. Statist. Assoc.* **114**, 1–14.

Cao, Y., Lin, W. & Li, H. (2018b). Two-sample tests of high-dimensional means for compositional data. *Biometrika* **105**, 115–32.

Cao, Y. & Xie, Y. (2016). Poisson matrix recovery and completion. *IEEE Trans. Sig. Proces.* **64**, 1609–20.

Chaffron, S., Rehrauer, H., Pernthaler, J. & von Mering, C. (2010). A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Res.* **20**, 947–59.

Chatterjee, S. (2015). Matrix estimation by universal singular value thresholding. *Ann. Statist.* **43**, 177–214.

Donoho, D. & Gavish, M. (2014). Minimax risk of matrix denoising by singular value thresholding. *Ann. Statist.* **42**, 2413–40.

Faust, K., Sathirapongsasuti, J. F., Izard, J., Segata, N., Gevers, D., Raes, J. & Huttenhower, C. (2012). Microbial co-occurrence relationships in the human microbiome. *PLoS Comput. Biol.* **8**, e1002606.

Gavish, M. & Donoho, D. L. (2014). The optimal hard threshold for singular values is $4/\sqrt{(3)}$. *IEEE Trans. Info. Theory* **60**, 5040–53.

Haegeman, B., Hamelin, J., Moriarty, J., Neal, P., Dushoff, J. & Weitz, J. S. (2013). Robust estimation of microbial diversity in theory and in practice. *ISME J.* **7**, 1092–101.

Hall, E. C., Raskutti, G. & Willett, R. (2017). Inference of high-dimensional autoregressive generalized linear models. *arXiv:*1605.02693v2.

Horner-Devine, M. C., Silver, J. M., Leibold, M. A., Bohannan, B. J., Colwell, R. K., Fuhrman, J. A., Green, J. L., Kuske, C. R., Martiny, J. B., Muyzer, G. et al. (2007). A comparison of taxon co-occurrence patterns for macro- and microorganisms. *Ecology* **88**, 1345–53.

Human Microbiome Project Consortium (2012). A framework for human microbiome research. *Nature* **486**, 215–21.

Jiang, X., Raskutti, G. & Willett, R. (2015). Minimax optimal rates for Poisson inverse problems with physical constraints. *IEEE Trans. Info. Theory* **61**, 4458–74.

Jiao, J., Venkat, K., Han, Y. & Weissman, T. (2017). Maximum likelihood estimation of functionals of discrete distributions. *IEEE Trans. Info. Theory* **63**, 6774–98.

Klopp, O. (2014). Noisy low-rank matrix completion with general sampling distribution. *Bernoulli* **20**, 282–303.

Klopp, O., Lafond, J., Moulines, R. & Salmon, J. (2015). Adaptive multinomial matrix completion. *Electron. J. Statist.* **9**, 2950–75.

Koeth, R. A., Wang, Z., Levison, B. S., Buffa, J. A., Org, E., Sheehy, B. T., Britt, E. B., Fu, X., Wu, Y., Li, L. et al. (2013). Intestinal microbiota metabolism of L-carnitine, a nutrient in red meat, promotes atherosclerosis. *Nature Med.* **19**, 576–85.

Lafond, J., Klopp, O., Moulines, E. & Salmon, J. (2014). Probabilistic low-rank matrix completion on finite alphabets. In *Proc. Adv. Neural Information Processing Syst. 27*, Ed. Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence & K. Weinberger. pp. 1727–35. Red Hook, NY: Curran Associates, Inc.

Lewis, J. D., Chen, E. Z., Baldassano, R. N., Otley, A. R., Griffiths, A. M., Lee, D., Bittinger, K., Bailey, A., Friedman, E. S., Hoffmann, C. et al. (2015). Inflammation, antibiotics, and diet as environmental stressors of the gut microbiome in pediatric Crohn's disease. *Cell Host & Microbe* **18**, 489–500.

Li, Y. & Raskutti, G. (2018). Minimax optimal convex methods for Poisson inverse problems under ℓ_q-ball sparsity. *IEEE Trans. Info. Theory* **64**, 5498–512.

Lin, W., Shi, P., Feng, R. & Li, H. (2014). Variable selection in regression with compositional covariates. *Biometrika* **101**, 785–97.

Liu, Z. & Vandenberghe, L. (2009). Interior-point method for nuclear norm approximation with application to system identification. *SIAM J. Matrix Anal. Appl.* **31**, 1235–56.

Martín-Fernández, J. A., Barceló-Vidal, C. & Pawlowsky-Glahn, V. (2003). Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Math. Geol.* **35**, 253–78.

MARTÍN-FERNÁNDEZ, J.-A., HRON, K., TEMPL, M., FILZMOSER, P. & PALAREA-ALBALADEJO, J. (2014). Bayesian-multiplicative treatment of count zeros in compositional data sets. *Statist. Mod.* **15**, 134–58.

MARTÍN-FERNÁNDEZ, J. A., PALAREA-ALBALADEJO, J. & OLEA, R. A. (2011). Dealing with zeros. In *Compositional Data Analysis: Theory and Applications*, Ed. V. Pawlowsky-Glahn and A. Buccianti, pp. 43–58. New York: J. Wiley and Sons.

NEGAHBAN, S. & WAINWRIGHT, M. J. (2012). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *J. Mach. Learn. Res.* **13**, 1665–97.

RECHT, B., FAZEL, M. & PARRILO, P. A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.* **52**, 471–501.

SALMON, J., HARMANY, Z., DELEDALLE, C.-A. & WILLETT, R. (2014). Poisson noise reduction with non-local PCA. *J. Math. Imag. Vis.* **48**, 279–94.

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

SONI, A. & HAUPT, J. (2014). Estimation error guarantees for Poisson denoising with sparse and structured dictionary models. In *Proc. 2014 IEEE Int. Symp. Information Theory*, pp. 2002–6. New York: IEEE.

SU, W., BOYD, S. & CANDES, E. J. (2016). A differential equation for modeling Nesterovs accelerated gradient method: Theory and insights. *J. Mach. Learn. Res.* **17**, 1–43.

TURNBAUGH, P. J., HAMADY, M., YATSUNENKO, T., CANTAREL, B. L., DUNCAN, A., LEY, R. E., SOGIN, M. L., JONES, W. J., ROE, B. A., AFFOURTIT, J. P. et al. (2009). A core gut microbiome in obese and lean twins. *Nature* **457**, 480–4.

WOYKE, T., TEELING, H., IVANOVA, N. N., HUNTEMANN, M., RICHTER, M., GLOECKNER, F. O., BOFFELLI, D., ANDERSON, I. J., BARRY, K. W., SHAPIRO, H. J. et al. (2006). Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature* **443**, 950–5.

WU, G. D., CHEN, J., HOFFMANN, C., BITTINGER, K., CHEN, Y. Y., KEILBAUGH, S. A., BEWTRA, M., KNIGHTS, D., WALTERS, W. A., KNIGHT, R. et al. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334**, 105–8.

WU, Y. & YANG, P. (2016). Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Trans. Info. Theory* **62**, 3702–20.

[*Received on* 8 *February* 2018. *Editorial decision on* 24 *April* 2019]