Biometrika (2021), **108**, 3, *pp.* 693–708
Printed in Great Britain
doi: 10.1093/biomet/asaa082
Advance Access publication 4 January 2021

Optimal estimation of bacterial growth rates based on a permuted monotone matrix

By RONG MA

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

T. TONY CAI

Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104, U.S.A.
tcai@wharton.upenn.edu

and HONGZHE LI

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

Summary

Motivated by the problem of estimating bacterial growth rates for genome assemblies from shotgun metagenomic data, we consider the permuted monotone matrix model $Y = \Theta\Pi + Z$ where $Y \in \mathbb{R}^{n\times p}$ is observed, $\Theta \in \mathbb{R}^{n\times p}$ is an unknown approximately rank-one signal matrix with monotone rows, $\Pi \in \mathbb{R}^{p\times p}$ is an unknown permutation matrix, and $Z \in \mathbb{R}^{n\times p}$ is the noise matrix. In this article we study estimation of the extreme values associated with the signal matrix Θ, including its first and last columns and their difference. Treating these estimation problems as compound decision problems, minimax rate-optimal estimators are constructed using the spectral column-sorting method. Numerical experiments on simulated and synthetic microbiome metagenomic data are conducted, demonstrating the superiority of the proposed methods over existing alternatives. The methods are illustrated by comparing the growth rates of gut bacteria in inflammatory bowel disease patients and control subjects.

Some key words: Extreme value; Metagenomics; Minimax lower bound; Permutation; Spectral method.

1. Introduction

This paper is motivated by the problem of estimating bacterial growth dynamics using shotgun metagenomics data. Several methods have been developed to quantify bacterial growth dynamics based on shotgun metagenomics data by extrapolating particular patterns in the sequencing read coverages that result from bidirectional microbial DNA replications (Abel et al., 2015; Korem et al., 2015; Myhrvold et al., 2015; Brown et al., 2016). For bacterial species with known complete genome sequences, Korem et al. (2015) proposed using the peak-to-trough ratio, PTR, of read coverages to quantify the bacterial growth rates after aligning the sequencing reads to the bacterial

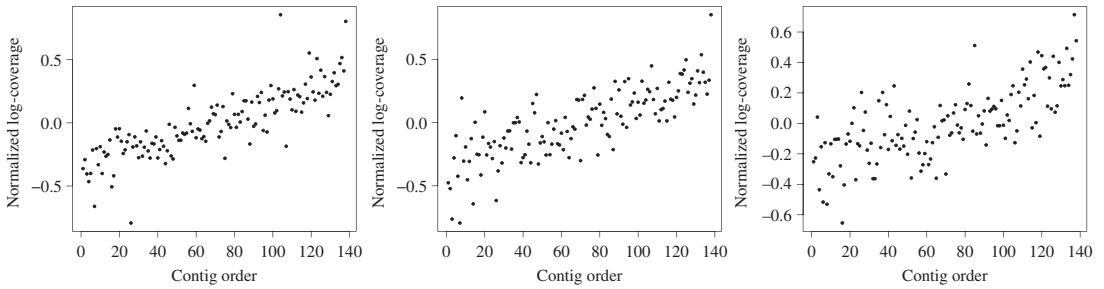

Fig. 1. The log-coverages of ordered contigs of an assembled bacterial species from three individuals with inflammatory bowel disease, detailed in § 5.3.

genomes. Besides quantifying the growth rates of the bacteria with complete genome sequences, it is also of great importance to estimate the growth rates of incomplete genome assemblies, where the coverages of contigs are observed in multiple samples. However, the order of the contigs is known only up to an unknown permutation.

Recently, Gao & Li (2018) developed a computational algorithm that can accurately estimate the growth dynamics of a given assembled species by taking advantage of highly fragmented contigs assembled from multiple samples. The algorithm is based on the following permuted monotone matrix model:

$$Y = \Theta\Pi + Z, \tag{1}$$

where the observed data $Y \in \mathbb{R}^{n\times p}$ are represented by the matrix of pre-processed contig coverages for a given bacterial species. Specifically, the entry Y_{ij} contains the log-transformed averaged read counts of the jth contig of the bacterial species for the ith sample after the pre-processing steps, including genome assemblies, GC adjustment of read counts and outlier filtering. In practice, the dataset is usually high-dimensional in the sense that the number of contigs p far exceeds the sample size n, so throughout we assume $p \gg n$. The signal matrix $\Theta \in \mathbb{R}^{n\times p}$ is the true log-transformed coverage matrix of n samples and p contigs, where each row is monotone because of the bidirectional DNA replication mechanism (Brown et al., 2016; Gao & Li, 2018), $Z \in \mathbb{R}^{n\times p}$ is the noise matrix, and $\Pi \in \mathbb{R}^{p\times p}$ is a permutation matrix that corresponds to some permutation π from the symmetric group $\mathcal{S}_p$. Ma et al. (2020) developed methods for optimally recovering the underlying permutation π from Y. In particular, with the loss function taken to be either the 0-1 loss or the normalized Kendall's τ distance, a minimax optimal permutation estimator was proposed and theoretically analysed under various parameter spaces.

In addition to the monotonicity constraint on the rows of Θ, real metagenomic datasets suggest an approximately linear relationship between the contig positions and their log-coverages for each sample, which indicates an approximately rank-one structure of Θ, after a certain normalization. As an example, Fig. 1 plots the normalized log-contig counts of an assembled bacterial genome from three individuals against the estimated contig orders, revealing the aforementioned approximately linear or rank-one structure; see § 3.1 for details.

Under the permuted monotone matrix model, one can relate the two extreme columns Θ_L and Θ_R, i.e., the first and last columns of Θ, to the log-transformed true peak and trough coverages of a given bacterial species, and define their difference $R(\Theta) = \Theta_R - \Theta_L$ to be the true log-PTRs that characterize the bacterial growth rates over n samples. The goal of this paper is to provide a rigorous statistical framework for optimal estimation of the extreme values in the approximately rank-one permuted monotone matrix model, including Θ_R and Θ_L and the range

vector $R(\Theta)$. Based on the idea of spectral column sorting and the theory of low-rank matrix estimation, we develop computationally efficient estimators for the extreme columns and the range vector. In particular, the minimax optimality of the proposed methods is theoretically justified and empirically illustrated with numerical experiments, which also demonstrate the applicability of the proposed methods to real data such as the microbiome metagenomics data.

Throughout the paper, we define the permutation π as a bijection from the set $\{1, 2, \ldots, p\}$ onto itself. For simplicity, we write $\pi = \{\pi(1), \pi(2), \ldots, \pi(p)\}$. The set of all permutations of $\{1, 2, \ldots, p\}$, equipped with the function composition operation $\circ$, is a symmetric group, denoted by $\mathcal{S}_p$. For any $\pi \in \mathcal{S}_p$, we let $\pi^{-1} \in \mathcal{S}_p$ denote its group inverse, so that $\pi \circ \pi^{-1} = \pi^{-1} \circ \pi = \mathrm{id}$. In particular, we may use π and its corresponding permutation matrix $\Pi \in \mathbb{R}^{p \times p}$ interchangeably, depending on the context. For a vector $a = (a_1, \ldots, a_n)^{\mathrm{T}} \in \mathbb{R}^n$, the ℓ_p-norm is defined by $\|a\|_p = (\sum_{i=1}^n a_i^p)^{1/p}$. For a matrix $\Theta \in \mathbb{R}^{p_1 \times p_2}$, we write $\Theta_{\cdot i} \in \mathbb{R}^{p_1}$ for its ith column and $\Theta_{i\cdot} \in \mathbb{R}^{p_2}$ for its ith row. We use the notation $a \wedge b = \min\{a, b\}$ and $ab = \max\{a, b\}$. Furthermore, for sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n = o(b_n)$ if $\lim_n a_n/b_n = 0$ and write $a_n = O(b_n)$, $a_n \lesssim b_n$ or $b_n \gtrsim a_n$ if there exists a constant C such that $a_n \leqslant C b_n$ for all n. We write $a_n \asymp b_n$ if both $a_n \lesssim b_n$ and $a_n \gtrsim b_n$.

2. Extreme value estimation via spectral sorting

2.1. *Spectral sorting and extreme column localization*

A crucial step in estimating the extreme columns is to sort the permuted columns in order to identify the extreme ones. In this section, we introduce a spectral approach to localizing the permuted columns. To this end, for any Θ with monotone rows, we consider the row-centred matrix

$$\Theta' = \Theta\left(I_p - \frac{1}{p} ee^{\mathrm{T}}\right) \in \mathbb{R}^{n \times p}, \tag{2}$$

where $e = (1, \ldots, 1)^{\mathrm{T}} \in \mathbb{R}^p$. Intuitively, Θ' is invariant with respect to the row averages of Θ and preserves the row-monotonicity structure as well as the distances between the columns of Θ. The singular value decomposition of Θ' can be written as

$$\Theta' = \sum_{i=1}^{r} \lambda_i u_i v_i^{\mathrm{T}} \tag{3}$$

for some $r \leqslant \min\{n, p\}$, where $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_r$ are the ordered singular values of Θ' and u_i and v_i are respectively the left and right singular vectors corresponding to λ_i. To overcome the identifiability issue, we assume the following condition.

Assumption 1. The leading singular value λ_1 has multiplicity one, and the first nonzero component of v_1 is negative.

The following proposition provides an important insight, namely that the row-monotonicity of a matrix actually implies the monotonicity of the components of its leading right singular vector v_1. This property plays a fundamental role in analysing the permuted monotone matrix model.

Proposition 1. *Let Θ be a row-monotone matrix whose row-centred version Θ' defined in* (2) *satisfies Assumption* 1. *Then its first right singular vector $v_1 = (v_{11}, \ldots, v_{1p})^{\mathrm{T}}$ is a centred*

monotone vector, i.e., $\sum_{i=1}^{p} v_{1i} = 0$ *and* $v_{11} \leqslant v_{12} \leqslant \cdots \leqslant v_{1p}$. *In addition, the sign vector* $\mathrm{sgn}(u_1)$ *indicates the direction of monotonicity of the rows of* Θ', *or* Θ.

From the above proposition, the relative orders of the columns of Θ' and Θ are qualitatively preserved by the leading right singular vector v_1, whereas the directions of monotonicity for different rows are coded by the leading left singular vector u_1. As a result, given a column-permuted and noisy matrix Y in (1), one can localize the extreme columns Θ_R and Θ_L in $\Theta\Pi$ by considering the row-normalized observation matrix $X = Y(I_p - p^{-1}ee^{\mathrm{T}})$ and its first right singular vector, i.e.,

$$\hat{v} = (\hat{v}_1, \ldots, \hat{v}_p)^{\mathrm{T}} = \underset{v \in \mathbb{R}^p:\, \|v\|_2 = 1}{\arg\max} \; v^{\mathrm{T}} X^{\mathrm{T}} X v. \tag{4}$$

In accordance with Proposition 1, it was shown by Ma et al. (2020) that the order statistics $\{\hat{v}_{(1)}, \ldots, \hat{v}_{(p)}\}$ can be used to optimally recover the permutation π, or the original column orders, by tracing back the permutation map between the elements of $\hat{v}$ and their order statistics. Clearly, for extreme column localization, the extreme value statistics $\hat{v}_{(1)}$ and $\hat{v}_{(p)}$ are more relevant. In fact, it is shown in the next subsection that minimax optimal estimators can be constructed using such spectral extreme value estimates.

2.2. *Compound decision problem and the proposed estimators*

The problem of estimating Θ_R, Θ_L or R consists of n individual subproblems, namely estimating each of the n coordinates. Following the concept proposed by Robbins (1951, 1964) and further elaborated by Samuel (1967), Copas (1969), Zhang (2003) and Brown & Greenshtein (2009), among many others, we observe that the problem of finding the minimax optimal estimator for Θ_R, Θ_L or R is a compound statistical decision problem, as the n individual subproblems are amalgamated into one larger problem through the combined risk shown in (6). Moreover, although the observations over n samples are independent, it has been argued that, in general, for a compound decision problem, the simple estimators where only the ith sample is used to estimate the ith coordinate are usually suboptimal; in contrast, a minimax optimal estimator should be compound in the sense that multiple samples are used for the estimation of each coordinate.

In light of our discussion in § 2.1 about the fundamental role of (λ_1, u_1, v_1), we propose the following estimators for the extreme columns:

$$\hat{\Theta}_R^* = \hat{v}_{(p)} X \hat{v} + \frac{1}{p} Ye \in \mathbb{R}^n, \quad \hat{\Theta}_L^* = \hat{v}_{(1)} X \hat{v} + \frac{1}{p} Ye \in \mathbb{R}^n. \tag{5}$$

Our proposed range estimator is

$$\hat{R}^* = \hat{\Theta}_R^* - \hat{\Theta}_L^* = (\hat{v}_{(p)} - \hat{v}_{(1)}) X \hat{v}.$$

We recall that $\hat{v}$ is defined in (4) and $\hat{v}_{(i)}$ is the ith smallest order statistic among $\{\hat{v}_1, \ldots, \hat{v}_p\}$. By construction, the proposed extreme column estimators in (5) are compound estimators, and each consists of two parts, with the first part estimating the extreme columns of the row-centred matrix Θ' and the second part compensating for the row-specific mean effects. In particular, in accordance with the observations made in § 1, to construct the first parts of $\hat{\Theta}_R^*$ and $\hat{\Theta}_L^*$, the approximately rank-one structure $\Theta'_\ell \approx \lambda_1 v_{1\ell} u$ for $\ell \in \{1, p\}$ is incorporated with $v_{1\ell}$ estimated by $\hat{v}_{(\ell)}$ and $\lambda_1 u_1$ estimated by $X\hat{v}$.

Ma et al. (2020) developed an optimal estimator for the permutation π under the model (1). Specifically, let $\mathfrak{r} : \mathbb{R}^p \to \mathcal{S}_p$ be the ranking operator, which is defined such that for any vector $\mathbb{X} \in \mathbb{R}^p$, $\mathfrak{r}(\mathbb{X})$ is the vector of ranks of the components of $\mathbb{X}$ in increasing order; whenever there are ties, increasing orders are assigned from left to right. The best linear projection estimator of π was defined as $\hat{\pi} = \{\mathfrak{r}(\hat{v})\}^{-1}$. This permutation estimator can be used to construct a natural two-step estimator of the two extreme columns. In the first step, we recover/sort the columns of Y to obtain the sorted matrix $\check{Y} = \begin{bmatrix} Y_{\cdot\hat{\pi}(1)} & Y_{\cdot\hat{\pi}(2)} & \dots & Y_{\cdot\hat{\pi}(p)} \end{bmatrix}$. Intuitively, the column-sorted matrix $\check{Y}$ is expected to be close to Θ. In the second step, we fit a simple linear regression between each row of $\check{Y}$ and the sorted projection scores $(\hat{v}_{(1)}, \hat{v}_{(2)}, \dots, \hat{v}_{(p)})$, which characterize the column relative locations. Denote the fitted intercepts by $\alpha = (\alpha_1, \dots, \alpha_n)^{\mathrm{T}}$ and the slopes by $\beta = (\beta_1, \dots, \beta_n)^{\mathrm{T}}$. We define the two-step regression estimators as

$$\hat{\Theta}_L^{\mathrm{Reg}} = \alpha + \beta \hat{v}_{(1)}, \quad \hat{\Theta}_R^{\mathrm{Reg}} = \alpha + \beta \hat{v}_{(p)}, \quad \hat{R}^{\mathrm{Reg}} = \beta(\hat{v}_{(p)} - \hat{v}_{(1)}).$$

It is easy to check, as shown in the Supplementary Material, that under the conditions of Proposition 1 we have that

$$\hat{\Theta}_L^{\mathrm{Reg}} = \hat{\Theta}_L^*, \quad \hat{\Theta}_R^{\mathrm{Reg}} = \hat{\Theta}_R^*, \quad \hat{R}^{\mathrm{Reg}} = \hat{R}^*.$$

Intuitively, the extreme columns of the sorted matrix $\check{Y}$ could be suboptimal as it does not make use of the rank-one structure. A better way is to project rows of $\check{Y}$ onto the eigenspace spanned by $\hat{v}$, which is equivalent to regressing rows of $\check{Y}$ on $\hat{v}$. This interesting observation provides another way of understanding our proposed estimators.

3. Theoretical properties

3.1. *Risk upper bounds for the extreme column estimators*

In this section we study the theoretical properties of the proposed estimator $\hat{\Theta}_R^*$; the results for Θ_L are parallel. We consider the normalized ℓ_2 distances $\|\hat{\Theta}_R - \Theta_R\|_2/\sqrt{n}$ and the corresponding estimation risk

$$\mathcal{R}_R(\hat{\Theta}_R) = \frac{1}{\sqrt{n}} E(\|\hat{\Theta}_R - \Theta_R\|_2). \tag{6}$$

We first define the set of monotone matrices

$$\mathcal{D} = \left\{ \begin{array}{c} \Theta = (\theta_{ij}) \in \mathbb{R}^{n \times p} : \text{for each } 1 \leqslant i \leqslant n, \text{either } \theta_{i,j} \leqslant \theta_{i,j+1} \text{ for all } j \\ \text{or } \theta_{i,j} \geqslant \theta_{i,j+1} \text{ for all } j \end{array} \right\}.$$

Recall that the row-centred version of Θ, namely Θ', has the singular value decomposition (3). Consequently, throughout this article we consider the following parameter space for (Θ, π):

$$\mathcal{D}_R(t, \beta) = \left\{ \begin{array}{c} (\Theta, \pi) \in \mathcal{D} \times \mathcal{S}_p : \text{Assumption 1 holds, } 0 \leqslant v_{1p} \leqslant \beta, \\ \lambda_1 \in [t/8, 8t], \ \sum_{i=2}^r \lambda_i \leqslant \sigma (\log p)^{1/2} \end{array} \right\}, \tag{7}$$

with $t \geqslant 0$ and $p^{-1/2} \leqslant \beta \leqslant 1$. Here the constraint on β is natural since v_1 is a unit vector and β is no less than the order of its largest component. Intuitively, the hyperparameters (t, β) characterize the global signal strength as well as the relative position of the extreme column Θ_R shared by the signal matrices in $\mathcal{D}_R(t, \beta)$, while the condition $\sum_{i=2}^r \lambda_i \leqslant \sigma (\log p)^{1/2}$ imposes a

strong approximately rank-one structure on the row-centred Θ. As our proposed estimators are not intended for estimating the possible additional structures over the leading rank-one structure, such an approximate rank-one condition is in this sense intrinsic to the problem.

To simplify notation, we define the rate function $\psi = \psi(n,p) = (\log p/n)^{1/2}$. The following theorem provides a uniform risk upper bound for the proposed estimator $\hat{\Theta}_R^*$ over $\mathcal{D}_R(t,\beta)$.

THEOREM 1 (Uniform upper bound). *Suppose that the pair* (t,β_R) *satisfies*

$$p^{-1/2} \leqslant \beta_R \leqslant 1, \quad t^2 \gtrsim \sigma^2 \left[\frac{1}{\beta_R^2} \wedge \left\{ \frac{1}{\psi^2} + \frac{1}{\psi} \left(\frac{p}{n \log p} \right)^{1/2} \right\} \right] n \log p$$

and that the noise matrix Z *has independent sub-Gaussian entries* Z_{ij} *with parameter* σ^2. *Then*

$$\sup_{\mathcal{D}_R(t,\beta_R)} \mathcal{R}_R(\hat{\Theta}_R^*) \lesssim \frac{\beta_R t}{\sqrt{n}} \left[\frac{\sigma\{(t^2+\sigma^2 p)n\}^{1/2}}{t^2} \wedge 1 \right] + \sigma\psi. \tag{8}$$

The risk upper bound (8) consists of two components. In the first component, the factor $[\sigma\{(t^2+\sigma^2 p)n\}^{1/2}/t^2 \wedge 1]$ is the error from estimating the leading left singular vector u_1 by its sample counterpart, whereas the factor $\beta_R t/\sqrt{n}$ reflects the overall magnitude of the extreme column Θ_R of the matrices in $\mathcal{D}_R(t,\beta_R)$. As for the second component $\sigma\psi(n,p)$, it comes from using the order statistic $\hat{v}_{(p)}$ to estimate the largest component of v_1.

Interestingly, regarding the first component, we observe two phase transitions when t^2 passes $\sigma^2(np)^{1/2}$ and $\sigma^2 p$. Specifically, in (8), we have

$$\frac{\beta_R t}{\sqrt{n}} \left[\frac{\sigma\{(t^2+\sigma^2 p)n\}^{1/2}}{t^2} \wedge 1 \right] \asymp \begin{cases} \dfrac{\beta_R t}{\sqrt{n}}, & t^2 \lesssim \sigma^2 (np)^{1/2}, \\ \dfrac{\beta_R \sigma^2 \sqrt{p}}{t}, & \sigma^2(np)^{1/2} \lesssim t^2 \lesssim \sigma^2 p, \\ \beta_R \sigma, & t^2 \gtrsim \sigma^2 p. \end{cases}$$

From the theory of low-rank matrix estimation (Cai & Zhang, 2018), the quantity $\sigma^2(np)^{1/2}$ is the critical point below which it is impossible to estimate the singular vector u_1. Hereafter, we refer to the parameter spaces $\{\mathcal{D}_R(t,\beta_R) : t^2 \leqslant \sigma^2(np)^{1/2}\}$, $\{\mathcal{D}_R(t,\beta_R) : \sigma^2(np)^{1/2} \lesssim t^2 \lesssim \sigma^2 p\}$ and $\{\mathcal{D}_R(t,\beta_R) : t^2 \gtrsim \sigma^2 p\}$ as the weak-, intermediate- and strong-SNR regimes, respectively, where SNR stands for signal-to-noise ratio.

To see the implications of the condition

$$t^2 \gtrsim \sigma^2 \left[\frac{1}{\beta_R^2} \wedge \left\{ \frac{1}{\psi^2} + \frac{1}{\psi} \left(\frac{p}{n \log p} \right)^{1/2} \right\} \right] n \log p \tag{9}$$

in Theorem 1 for the critical events $t^2 \asymp \sigma^2(np)^{1/2}$ and $t^2 \asymp \sigma^2 p$, observe that as long as $\beta_R \gg (n/p)^{1/4}$, upon ignoring the logarithmic factors the right-hand side of (9) is asymptotically smaller than both of the critical points $\sigma^2(np)^{1/2}$ and $\sigma^2 p$, so both phase transitions exist under the condition of Theorem 1.

3.2. *Optimality of the extreme column estimators and minimax rates*

Now we establish the minimax rate of convergence and the optimality of the proposed extreme column estimator $\hat{\Theta}_R^*$ over the parameter space $\mathcal{D}_R(t,\beta_R)$. Specifically, for some given (t,β), we

define the minimax risks over $\mathcal{D}_R(t,\beta_R)$ as $\inf_{\hat{\Theta}_R}\sup_{\mathcal{D}_R(t,\beta_R)}\mathcal{R}_R(\hat{\Theta}_R)$, where the infimum is over all possible estimators obtained from the data. The following theorem provides the minimax lower bound for the estimation risk under Gaussian noise.

THEOREM 2 (Minimax lower bound). *Suppose that Z in model (1) has independent and identically distributed entries $Z_{ij}\sim N(0,\sigma^2)$. Then, for any $\mathcal{D}_R(t,\beta)$ such that*

$$t^2 \geqslant c_0\left(\frac{1-\beta_R^2}{\beta_R^2}\sigma^2\log p+\frac{\beta_R^2}{1-\beta_R^2}\sigma^2 p\right),\quad c_1p^{-1/2}(\log p)^{1/2}\leqslant\beta_R\leqslant c_2,$$

for sufficiently large (n,p) and some constants $c_0,c_1>0$ and $0<c_2<1$, we have that

$$\inf_{\hat{\Theta}_R}\sup_{\mathcal{D}_R(t,\beta_R)}\mathcal{R}_R(\hat{\Theta}_R)\gtrsim\frac{\beta_R t}{\sqrt{n}}\left[\frac{\sigma\{(t^2+\sigma^2p)n\}^{1/2}}{t^2}\wedge 1\right]+\sigma\psi.$$

The proof of Theorem 2 is quite involved. The main difficulty lies in the nonlinearity and multi-dimensionality of the maps from the original parameter Θ to its extreme columns of interest. As the lower bound contains several components, we essentially derive three distinct minimax lower bounds corresponding to different worst-case scenarios. In addition to using existing techniques such as sphere packing of Grassmannian manifolds, we have developed two novel lower-bound techniques to facilitate the proof of the minimax lower bound. The details can be found in the Supplementary Material. The different conditions in Theorem 2 are due to the specific constructions in the lower-bound argument of the proof. However, from a broader perspective, the conditions in Theorem 2 agree with those of Theorem 1 in the sense that the first condition, a lower bound on t^2, ensures that the global signal strength is sufficiently large, while the second condition is a mild restriction, up to a logarithmic factor, on the range of β_R.

By combining the upper and lower bounds, we obtain the exact minimax rate for estimating Θ_R. Specifically, under the conditions of Theorems 1 and 2, i.e., for independent and identically distributed $Z_{ij}\sim N(0,\sigma^2)$ and for

$$t^2\gtrsim\sigma^2\left[\frac{1}{\beta_R^2}\wedge\left\{\frac{1}{\psi^2}+\frac{1}{\psi}\left(\frac{p}{n\log p}\right)^{1/2}\right\}\right]n\log p+\left(\frac{1-\beta_R^2}{\beta_R^2}\sigma^2\log p+\frac{\beta_R^2\sigma^2p}{1-\beta_R^2}\right),\qquad(10)$$

we have

$$\inf_{\hat{\Theta}_R}\sup_{\mathcal{D}_R(t,\beta_R)}\mathcal{R}_R(\hat{\Theta}_R)\asymp\frac{\beta_R t}{\sqrt{n}}\left[\frac{\sigma\{(t^2+\sigma^2p)n\}^{1/2}}{t^2}\wedge 1\right]+\sigma\psi,$$

where the optimal rate is attained by our proposed estimator $\hat{\Theta}_R^*$. To make better sense of condition (10), observe that as long as $\beta_R\gtrsim(n/p)^{1/4}$, upon ignoring the logarithmic factors (10) becomes equivalent to $t^2\gtrsim\sigma^2(np)^{1/2}$, which means that the minimax rate

$$\inf_{\hat{\Theta}_R}\sup_{\mathcal{D}_R(t,\beta_R)}\mathcal{R}_R(\hat{\Theta}_R)\asymp\frac{\sigma n^{1/4}(t^2+\sigma^2p)^{1/2}}{p^{1/4}t}+\sigma\psi$$

can essentially be achieved over the intermediate- and strong-SNR regimes. As a consequence of the phase transition phenomena pointed out earlier, some interesting insights into the interplay between the global signal strength t^2, the dimensionality of the problem, the difficulty of estimating Θ_R, and the difficulty of estimating the leading left singular vector u_1 can be gained.

Specifically, we observe that in the intermediate-SNR regime, $\sigma^2(np)^{1/2} \lesssim t^2 \lesssim \sigma^2 p$, increasing the signal strength t^2 will reduce the difficulty of estimating u_1, and therefore the rate for estimating Θ_R, and in the strong-SNR regime, $t^2 \gtrsim \sigma^2 p$, the difficulty of estimating Θ_R no longer depends on t^2, as in this case the improved estimation of u_1 is neutralized by the increased magnitude of Θ_R. In particular, all of the above rate analysis is subject to a possible lower bound of $\psi(n,p)$.

Moreover, since the above minimax optimal rates are simultaneously attained by the proposed estimator $\hat{\Theta}_R^*$ regardless of the specific values of the underlying indices (t,β_R), under sub-Gaussian noise $\hat{\Theta}_R^*$ is minimax rate-adaptive over the collection of parameter spaces $\mathcal{C} = \{\mathcal{D}_R(t,\beta_R) : p^{-1/2}c_1(\log p)^{1/2} \leqslant \beta_R \leqslant c_2 < 1,\ (10) \text{ holds}\}$. In particular, whenever $\beta_R \gtrsim (n/p)^{1/4}$, upon ignoring the logarithmic factors our proposed estimator is rate-optimally adaptive over the collection of parameter spaces in the intermediate- and strong-SNR regimes, namely $\mathcal{C}_{\text{Adap}} = \{\mathcal{D}_R(t,\beta_R) : t^2 \gtrsim \sigma^2(np)^{1/2}\}$.

3.3. *Optimality of the range estimator and minimax rates*

Based on our previous results on extreme column estimation, the theoretical properties of the range estimator $\hat{R}^*$ can be derived in the same manner. Again, we consider the normalized ℓ_2 distances $\|\hat{R} - R(\Theta)\|_2/\sqrt{n}$ and denote the corresponding estimation risk by $\mathcal{R}_W(\hat{R}) = n^{1/2}E(\|\hat{R} - R(\Theta)\|_2)$. Define the parameter space

$$\mathcal{D}_W(t,\beta_R,\beta_L) = \left\{ \begin{array}{l} (\Theta,\pi) \in \mathcal{D} \times \mathcal{S}_p : \text{Assumption 1 holds, } \lambda_1 \in [t/8, 8t], \\ \qquad\qquad \sum_{i=2}^r \lambda_i \leqslant \sigma(\log p)^{1/2},\ -\beta_L \leqslant v_{11} \leqslant 0 \leqslant v_{1p} \leqslant \beta_R \end{array} \right\} \tag{11}$$

with $t \geqslant 0$ and $p^{-1/2} \leqslant \beta_R, \beta_L \leqslant 1$. Also, define the function

$$q'(x,y,n,p) = \sigma^2 n \log p \left[\frac{1}{x^2} \wedge \left\{ \frac{1}{\psi^2} + \frac{1}{\psi}\left(\frac{p}{n\log p}\right)^{1/2} \right\} \right] + \left(\frac{1-x^2}{x^2}\sigma^2 \log p + \frac{y^2\sigma^2 p}{1-y^2} \right).$$

The following theorem establishes the minimax rate of convergence for estimating $R(\Theta)$, and the minimax optimality and adaptivity of our proposed estimator $\hat{R}^*$.

THEOREM 3 (Minimax rates). *Let $\beta_W = \beta_R + \beta_L$. Suppose that $t^2 \geqslant c_0 q'(\beta_R \wedge \beta_L, \beta_R\beta_L, n, p)$ and $c_1 p^{-1/2}(\log p)^{1/2} \leqslant \{\beta_R, \beta_L\} \leqslant c_2$ for sufficiently large (n,p) and some constants $c_0, c_1 > 0$ and $0 < c_2 < 1$, and suppose that Z has independent sub-Gaussian entries Z_{ij} with parameter σ^2. Then*

$$\inf_{\hat{R}} \sup_{\mathcal{D}_W(t,\beta_R,\beta_L)} \mathcal{R}_W(\hat{R}) \asymp \frac{\beta_W t}{\sqrt{n}} \left[\frac{\sigma\{(t^2+\sigma^2 p)n\}^{1/2}}{t^2} \wedge 1 \right] + \sigma\psi.$$

In particular, the minimax rates are simultaneously attained by the estimator $\hat{R}^*$.

4. A SPECIAL CASE: PERMUTED LINEAR GROWTH MODEL

In the preceding sections, theoretical results were obtained for the general approximately rank-one matrices characterized by (7) and (11) together with the conditions of Theorems 1–3. One advantage of the parameter spaces we consider is the rich row-monotonicity structures they contain, which adapt well to real applications such as our motivating example from microbiome studies, where the noisy datasets are generated from shotgun metagenomic sequencing; see Boulund et al. (2018), Gao & Li (2018) and Fig. 1. However, in many cases, such as in

classical theories of bacterial growth dynamics, an important subclass of the general permuted monotone matrix model has commonly been used because of its heuristic simplicity and explanatory power. We refer to this submodel as the permuted linear growth model, where (1) holds over the restricted set

$$\mathcal{D}_0 = \left\{ \begin{array}{l} (\Theta, \pi) \in \mathcal{D} \times \mathcal{S}_p : \theta_{ij} = a_i \eta_j + b_i \text{ where } a_i, b_i \in \mathbb{R} \text{ for } 1 \leqslant i \leqslant n, \\ \qquad\qquad\qquad\qquad \eta_j \leqslant \eta_{j+1} \text{ for } 1 \leqslant j \leqslant p-1 \text{ and } \sum_{j=1}^{p} \eta_j = 0 \end{array} \right\}.$$

In other words, each row of Θ has a linear growth pattern, and the different rows have possibly differing intercepts and slopes. Write $a = (a_i)_{1 \leqslant i \leqslant n}$, $\eta = (\eta_j)_{1 \leqslant j \leqslant p}$ and $b = (b_i)_{1 \leqslant i \leqslant n}$. In this case, the parameters of interest have the expressions $\Theta_R = a\eta_p$, $\Theta_L = a\eta_1$ and $R = a(\eta_p - \eta_1)$.

In the context of bacterial growth dynamics, the above model is commonly referred as the Cooper–Helmstetter model (Cooper & Helmstetter, 1968; Bremer & Churchward, 1977), which associates the copy numbers of genes with their relative distances from the replication origin. Specifically, a_i is the ratio of the genome replication time to the doubling time for the ith sample, η_j is the distance from the replication origin of the jth contig, and b_i is related to the read counts at the replication origin and the sequencing depth. Consequently, the extreme columns $a\eta_p$ and $a\eta_1$ correspond to the true log-transformed peak, and trough coverages that are used to quantify the bacterial growth dynamics across the samples; see § § 5.2 and 5.3 for more details.

In the following, we discuss the estimation of Θ_R under this special linear growth model; the results for estimating Θ_L and R follow similarly. By definition, the singular value decomposition (3) for $\Theta \in \mathcal{D}_0$ has a reduced form. Specifically, the row-centred matrix Θ' is exactly of rank one, where the leading right singular vector v_1 has components

$$v_{1j} = \frac{\eta_j}{\|\eta\|_2} \quad (j = 1, \ldots, p) \tag{12}$$

and the largest singular value admits the expression

$$\lambda_1 = \|a\|_2 \|\eta\|_2. \tag{13}$$

Intuitively, the set $\{v_{1j}\}_{1 \leqslant i < j \leqslant p}$ characterizes the exact normalized column positions of Θ', and of Θ, while λ_1 summarizes the slope magnitude of the rows and the overall separateness of the columns. Consequently, the risk upper bound in Theorem 1 has a reduced form, which admits simpler and more intuitive interpretations. Specifically, for any given $\Theta \in \mathcal{D}_R(t, \beta_R)$, we consider the pointwise risk upper bound

$$\mathcal{R}_R(\hat{\Theta}_R^*) \lesssim \frac{v_{1p}\lambda_1(\Theta)}{\sqrt{n}} \left[\frac{\sigma\{(\lambda_1^2(\Theta) + \sigma^2 p)n\}^{1/2}}{\lambda_1^2(\Theta)} \wedge 1 \right] + \sigma\psi \tag{14}$$

induced by (8) in Theorem 1. With the reparametrizations (12) and (13), we can rewrite (14) as

$$\mathcal{R}_R(\hat{\Theta}_R^*) \lesssim \frac{\eta_p \|a\|_2}{\sqrt{n}} \left[\frac{\sigma\{(\|a\|_2^2 \|\eta\|_2^2 + \sigma^2 p)n\}^{1/2}}{\|a\|_2^2 \|\eta\|_2^2} \wedge 1 \right] + \sigma\psi. \tag{15}$$

Some observations about this risk upper bound are in order.

(i) In the low-SNR regime, where $\|a\|_2^2 \|\eta\|_2^2 \lesssim \sigma^2 (np)^{1/2}$, (15) becomes

$$\mathcal{R}_R(\hat{\Theta}_R^*) \lesssim \frac{\|a\|_2 \eta_p}{\sqrt{n}} + \sigma\psi, \tag{16}$$

where the first term is proportional to the overall slope magnitude $\|a\|_2$, but does not depend on the locations of the other columns, i.e., η_j for $1 \leqslant j \leqslant p-1$. In this case, since the signal changes across different columns are so vague, $\hat{\Theta}_R^*$ fails to implement a good estimate for the slopes a, and the estimation error can only decrease when the extreme column $\Theta_R = a\eta_p$ itself, and its norm $\|a\|_2\eta_p$, is close to zero.

(ii) In the intermediate-SNR regime, where $\sigma^2(np)^{1/2} \lesssim \|a\|_2^2\|\eta\|_2^2 \lesssim \sigma^2 p$, (15) becomes

$$\mathcal{R}_R(\hat{\Theta}_R^*) \lesssim \frac{\sigma\eta_p}{\|\eta\|_2}\left(1 + \frac{\sigma^2 p}{\|a\|_2^2\|\eta\|_2^2}\right)^{1/2} + \sigma\psi. \tag{17}$$

In this case, as the signal differences between consecutive pairs of columns are steep enough that the slopes a can be well estimated, increasing $\|\eta\|_2$ or $\|a\|_2$ would enhance the advantage and therefore lead to a better estimate.

(iii) In the strong-SNR regime, where $\|a\|_2^2\|\eta\|_2^2 \gtrsim \sigma^2 p$, the upper bound (15) becomes

$$\mathcal{R}_R(\hat{\Theta}_R^*) \lesssim \frac{\sigma\eta_p}{\|\eta\|_2} + \sigma\psi. \tag{18}$$

In this case, the advantage of large $\|a\|_2$ has been exploited to the extreme so that increasing $\|a\|_2$ will no longer yield further improvements in the performance of $\hat{\Theta}_R^*$.

Comparing the rates (16), (17) and (18), an interesting difference in the role played by the overall slope magnitude $\|a\|_2$ can be observed. In general, the theoretical performance of $\hat{\Theta}_R^*$ is clearly driven by the global SNR, $\|a\|_2^2\|\eta\|_2^2/\sigma^2$, which measures the magnitude of the signal changes and the degree of monotonicity relative to the noise level.

Following the same argument as that used in the proof of Theorem 2, the minimax optimality of our proposed estimator $\hat{\Theta}_R^*$ can also be established under the permuted linear growth model. Specifically, if we define the indexed parameter space $\mathcal{D}_{0,R}(t,\beta) = \{(\Theta,\pi) \in \mathcal{D}_0 : 0 \leqslant \eta_p/\|\eta\|_2 \leqslant \beta,\ \|a\|_2\|\eta\|_2 \in [t/8, 8t]\}$, then it can be shown that for any pair (t,β_R) such that (10) holds and $p^{-1/2}(\log p)^{1/2} \lesssim \beta_R \leqslant c < 1$,

$$\inf_{\hat{\Theta}_R} \sup_{\mathcal{D}_{0,R}(t,\beta_R)} \mathcal{R}_R(\hat{\Theta}_R) \asymp \frac{\beta_R t}{\sqrt{n}}\left[\frac{\sigma\{(t^2+\sigma^2 p)n\}^{1/2}}{t^2} \wedge 1\right] + \sigma\psi,$$

where the optimal rate is simultaneously attained by the proposed estimator $\hat{\Theta}_R^*$.

5. NUMERICAL STUDIES

5.1. *Simulation with model-generated data*

To illustrate our theoretical results and compare the proposed approach with alternative methods, we generate data from model (1) with various configurations of the signal matrix Θ. Specifically, the signal matrix $\Theta = (\theta_{ij}) \in \mathbb{R}^{n\times p}$ is generated under the following two regimes: $S_1(n,p,\alpha)$, where for any $1 \leqslant i \leqslant n$, $\theta_{ij} = a_i\eta_j + b_i$ $(1 \leqslant j \leqslant p)$ with $a_i \sim \mathrm{Un}(0,\alpha)$, $b_i \sim \mathrm{Un}(0,6)$ and $(\eta_1,\ldots,\eta_p) = (-1,0,0,\ldots,0,1)$; and $S_2(n,p,\alpha)$, where for any $1 \leqslant i \leqslant n$, $\theta_{ij} = \log(1 + a_i j + \beta_i)$ $(1 \leqslant j \leqslant p)$ with $a_i \sim \mathrm{Un}(0,\alpha)$ and $b_i \sim \mathrm{Un}(0,6)$. By construction, $S_1(\alpha,n,p)$ belongs to the linear growth model whereas $S_2(\alpha,n,p)$ does not. The elements of Z are drawn from independent and identically distributed standard normal distributions, and without loss of generality we set $\Pi = I_p$.

For the extreme column Θ_R, we compare the empirical performance of our proposed estimator $\hat{\Theta}_R^*$ with that of the direct sorting estimator $\tilde{\Theta}_R = Y_{\cdot\hat{\pi}(p)}$, where $\hat{\pi}$ is as defined in § 2.2, and the

order statistic estimator $\check{\Theta}_R = (Y_{i,(p)})_{1\leqslant i\leqslant n}$, as all the rows of Θ are monotone increasing. For the range vector $R(\Theta)$, we compare our proposed estimator $\hat{R}^*$ with the direct sorting estimator $\tilde{R}_{\mathrm{DS}} = Y_{\cdot\hat{\pi}(p)} - Y_{\cdot\hat{\pi}(1)}$ and the order statistic estimator $\tilde{R}_{\mathrm{OS}} = (Y_{i,(p)} - Y_{i,(1)})_{1\leqslant i\leqslant n}$. We use the empirical risk, or the averaged normalized ℓ_2 distance, to compare these methods. For each setting, we evaluate the empirical performance of each method over a range of n, p and α values. Each setting is repeated 200 times.

The results are summarized as boxplots in Figs. 2 and 3. The empirical results agree with our theory in the following ways: (i) our proposed estimators $\hat{\Theta}^*_R$ and $\hat{R}^*$ perform the best in all the settings; (ii) in panels (c) and (d) of Figs. 2 and 3, the risks of the proposed estimators decrease as n grows, in agreement with our theorems. In addition, in panel (a) of Figs. 2 and 3 we can see that the risks of the order statistic estimator decrease as α increases. This is because under $S_1(\alpha, p, n)$, the parameter α characterizes the separateness of the two extreme columns from the other columns. The order statistic estimators apparently favour the cases where the separation is more significant. Our proposed estimators and the direct sorting estimators outperform the order statistic estimators, demonstrating the superiority of the compound estimators.

5.2. *Simulation with synthetic microbiome metagenomic data*

We now evaluate the empirical performance of our proposed method on a synthetic metagenomic sequencing dataset (Gao & Li, 2018), by generating sequencing reads based on 45 closely related bacterial genomes in 50 independent samples. In particular, Gao & Li (2018) presented a synthetic shotgun metagenomic sequencing dataset for a community of 45 phylogenetically related species from 15 genera of five different phyla with known RefSeq ID, taxonomy and replication origin (Gao et al., 2013). To generate the metagenomic reads, reference genome sequences of three randomly selected species in each genus were downloaded. Read coverages were generated along the genome based on an exponential distribution with a specified PTR, and a function for the cumulative distribution of read coverages along the genome was calculated. Sequencing reads were then generated using these cumulative distribution functions and a random location for each read on the genome, until the total read number achieved a randomly assigned average coverage between 0.5 and 10 folds for the species in a sample. Sequencing errors, including substitution, insertion and deletion, were simulated in a position- and nucleotide-specific pattern according to the metagenomic sequencing error profile of Illumina.

For the final dataset, the average nucleotide identities between species within each genus ranged from 66.6% to 91.2%. The probability of one species existing in each of the 50 simulated samples was set to 0.6, and a total of 1336 average coverages and the corresponding PTRs were randomly and independently assigned. After the same processing, filtering and CG-adjustment steps as in Gao & Li (2018), the final dataset included genome assemblies of 41 species. For each species, we obtained the permuted matrix of log-contig coverage with the number of samples ranging from 29 to 46, and the number of contigs ranging from 47 to 482.

We provide estimates of the log-PTRs of the assembled species for all the samples, or the range vector R, using our previous notation. As a comparison, alongside our proposed method, $\hat{R}^*$, we consider the iRep estimator proposed by Brown et al. (2016), where the contigs of a given species were ordered for each sample separately based on the observed read counts before fitting a piecewise-linear regression function. We evaluate the methods by considering the ℓ_2 distance between the vectors of the true log-PTRs and their estimates. To generalize our evaluation to diverse metagenomic datasets, we also evaluate the effects of the sample size n and the number of contigs p by randomly selecting subsets of samples or contigs from each dataset. The selection was made with replacement.

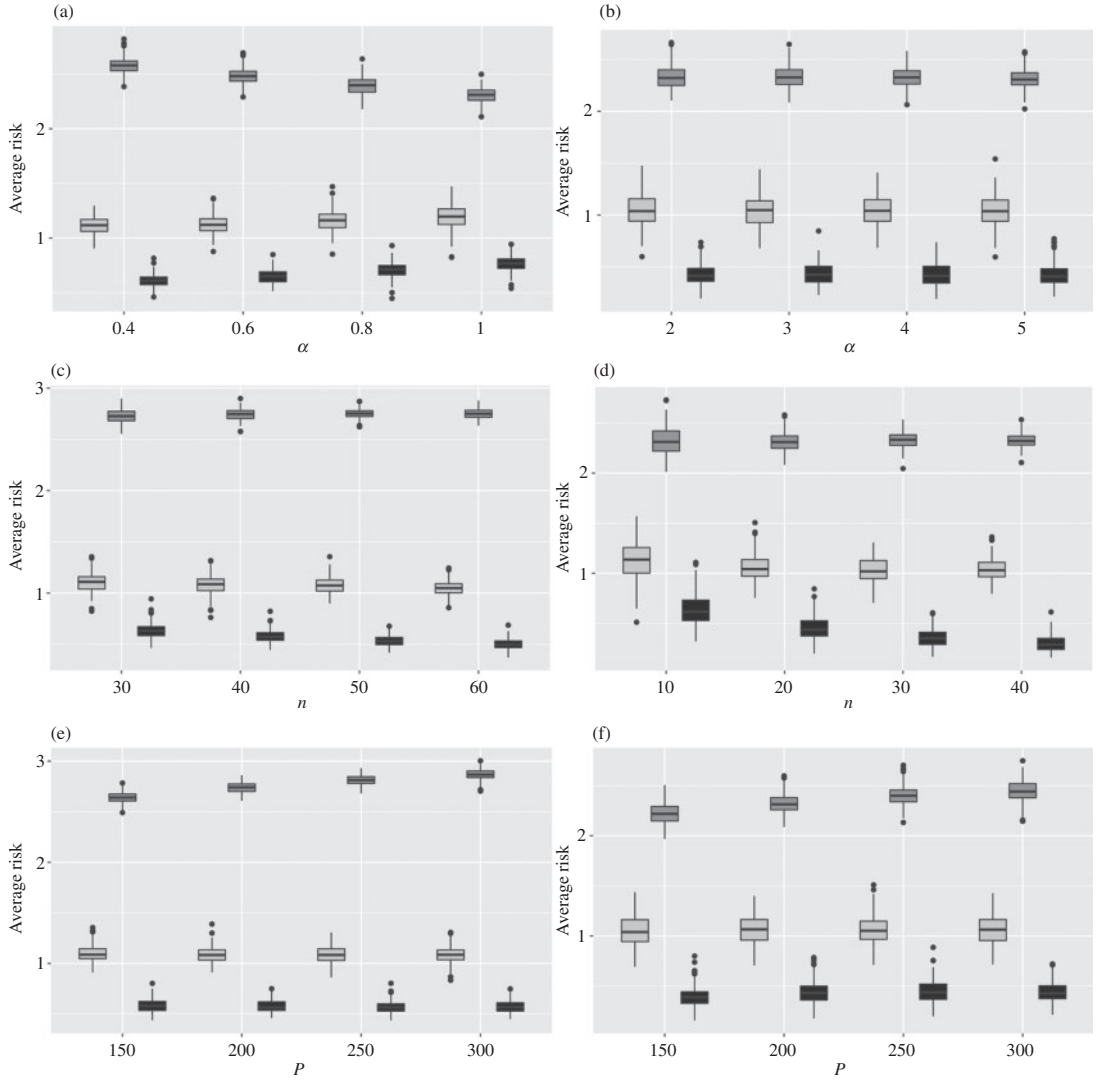

Fig. 2. Boxplots of the empirical risks in estimating Θ_R, for the proposed estimator $\hat{\Theta}_R^*$ (dark grey), the order statistic estimator $\check{\Theta}_R$ (medium grey) and the direct sorting estimator $\tilde{\Theta}_R$ (light grey) in the following settings: (a) regime S_1 with $n = 40$ and $p = 200$; (b) regime S_2 with $n = 20$ and $p = 200$; (c) regime S_1 with $p = 200$ and $\alpha = 0.6$; (d) regime S_2 with $p = 200$ and $\alpha = 4$; (e) regime S_1 with $n = 40$ and $\alpha = 0.6$; (f) regime S_1 with $n = 20$ and $\alpha = 4$.

The results are summarized in Fig. 4. As n or p varies, the proposed estimator consistently performs better than iRep. Moreover, the performance of the proposed method is not sensitive to the sample size, the number of contigs from the genome assemblies or the underlying true PTRs. These results partially explain why the algorithm of Gao & Li (2018) performs better than existing competitors.

5.3. *Analysis of a real microbiome metagenomic dataset*

We complete our numerical study by analysing a real metagenomic dataset from the NIH Integrative Human Microbiome Project, which includes the Inflammatory Bowel Disease Multi'omics Data project that investigates the differences in gut microbiome communities

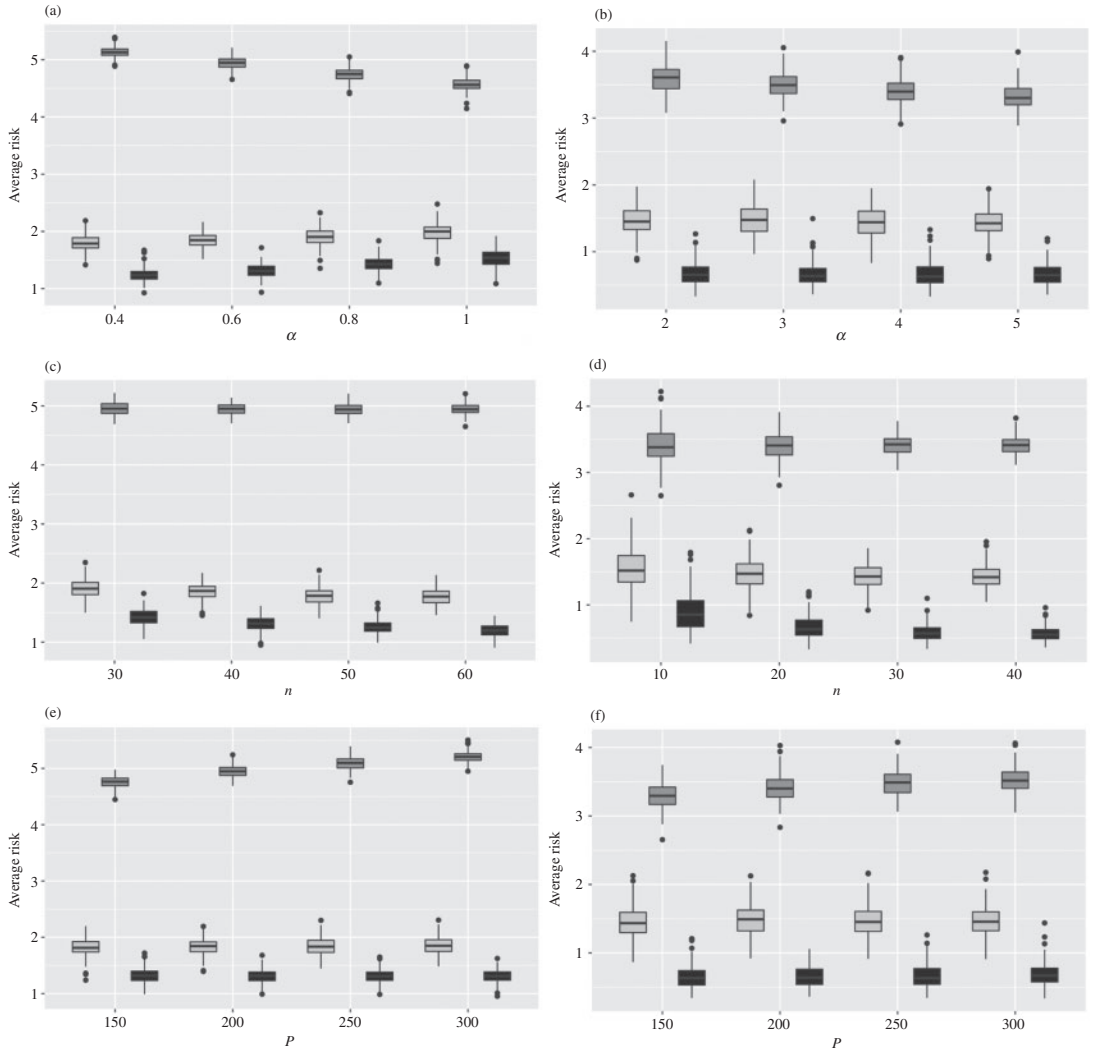

Fig. 3. Boxplots of the empirical risks in estimating R, for the proposed estimator $\hat{R}^*$ (dark grey), the order statistic estimator $\tilde{R}_{\mathrm{OS}}$ (medium grey) and the direct sorting estimator $\tilde{R}_{\mathrm{DS}}$ (light grey) in the following settings: (a) regime S_1 with $n = 40$ and $p = 200$; (b) regime S_2 with $n = 20$ and $p = 200$; (c) regime S_1 with $p = 200$ and $\alpha = 0.6$; (d) regime S_2 with $p = 200$ and $\alpha = 4$; (e) regime S_1 with $n = 40$ and $\alpha = 0.6$; (f) regime S_1 with $n = 20$ and $\alpha = 4$.

between adults and children with inflammatory bowel disease, IBD, and normal non-IBD controls (Lloyd-Price et al., 2019). Many studies have reported strong associations between gut microbiota composition and IBD, including Crohn's disease, CD, and ulcerative colitis, UC. Here we instead focus on comparing the bacterial growth rates in UC, CD and non-IBD individuals using the proposed methods.

The metagenomic datasets, including 300 samples of CD, UC and non-IBD subjects, were downloaded from `https://www.ibdmdb.org`. Specifically, we randomly select 100 samples each of UC, CD and non-IBD subjects. For each sample, the sequencing data were obtained from a stool sample by Illumina shotgun sequencing. We first use MEGAHIT (Li et al., 2015) version 1.1.1 to perform metagenomic co-assembly. The co-assembled contigs were then clustered into metagenomic bins or genome assemblies using MaxBin (Wu et al., 2015) version 2.2.4. Finally, Bowtie 2 (Langmead & Salzberg, 2012) version 2.3.2 was used to align reads back to

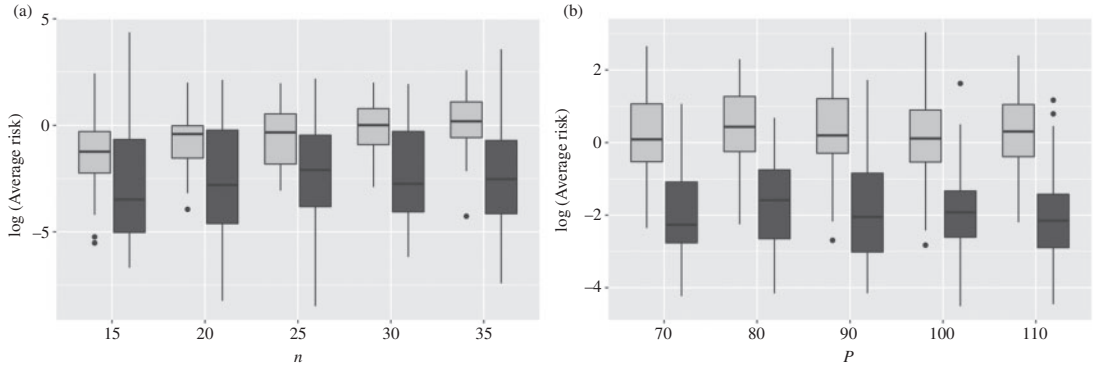

Fig. 4. Boxplots of the ℓ_2 distances between the estimated and true log-PTRs for the proposed method (dark grey) and the iRep estimation method (light grey), versus (a) the sample size n and (b) the number of contigs p.

Table 1. *Comparison of bacterial growth rates in CD, UC and non-IBD samples: bins that show significantly different growth rates are reported along with their taxonomic annotations; n_1, n_2 and n_3 refer to the numbers of CD, UC and non-IBD samples corresponding to each bin*

Bin	Taxonomic annotation	(n_1, n_2, n_3)	p-value
bin.054	Roseburia (genus)	(54, 32, 54)	0.015
bin.090	Faecalibacterium (genus)	(38, 41, 52)	0.005
bin.091	Clostridiales (order)	(26, 40, 52)	0.016
bin.099	Subdoligranulum (genus)	(30, 32, 49)	<0.001
bin.465	Dialister (genus)	(36, 41, 33)	0.043

UC, ulcerative colitis; CD, Crohn's disease; IBD, inflammatory bowel disease.

the assembled contigs for each of the samples, and the output alignments were then sorted by SAMtools (Li et al., 2009) version 0.1.19.

After these preparations, the algorithm of Gao & Li (2018), combined with our proposed methods, was applied to obtain the estimated PTRs, called ePTRs, of a given species, represented by a contig cluster, bin, for each sample. As a result, ePTRs of 25 bins were obtained for subsets of the UC, CD and non-IBD samples of sizes n_1, n_2 and n_3, respectively, with $n_1 + n_2 + n_3 \geqslant 100$, as some contig clusters may not be carried or abundant enough among many samples. For each bin, we compare the ePTRs of the UC, CD and non-IBD samples using an F-test. We apply the CAT/BAT algorithm (von Meijenfeldt et al., 2019), which compares the metagenomic assembled bins to a taxonomy database to obtain the taxonomic annotations of the 25 bins. We observe that only a few bins can be annotated at the species level, and many of the bins can be annotated only to genera or orders, suggesting that many of the assembled contig bins may correspond to new species. This agrees with the recent work of Almeida et al. (2020), which found that more than 70% of the assembled genomes lack cultured representatives.

Interestingly, based on the F-test, of the 25 contig clusters, five show significant differences in the ePTRs of the UC, CD and non-IBD samples; see Table 1. Because of space limitations, Table 1 shows only the taxonomic annotations of the bins at the genus level, except for bin.091, which could be determined only up to the order; see the Supplementary Material for the complete annotations. We also performed pairwise comparisons using the two-sample t-test for the five differential bins; see Table 2. We found that the difference in the growth rates of bin.054, *Roseburia*, bin.090, *Faecalibacterium*, and bin.099, *Subdoligranulum*, are more significant between IBD and non-IBD samples. In particular, boxplots in the Supplementary Material indicate higher

Table 2. *The p-values obtained from pairwise t-tests of differential growth rates between different groups for five genome assembly bins*

Bin	Taxonomic annotation	UC v. CD	UC v. non-IBD	CD v. non-IBD
bin.054	Roseburia (genus)	0.525	0.004	0.081
bin.090	Faecalibacterium (genus)	0.392	0.016	0.004
bin.091	Clostridiales (order)	0.012	0.054	0.335
bin.099	Subdoligranulum (genus)	0.960	<0.001	<0.001
bin.465	Dialister (genus)	0.042	0.818	0.026

UC, ulcerative colitis, CD, Crohn's disease; IBD, inflammatory bowel disease.

growth rates of bin.054, *Roseburia*, and bin.090, *Faecalibacterium*, and a lower growth rate of bin.099, *Subdoligranulum*, in IBD samples than in non-IBD samples. Moreover, the growth rate of bin.091, *Clostridiales*, is significantly higher in the UC samples, whereas the growth rate of bin.465, *Dialister*, is significantly higher in the CD samples, than in the samples of the other two categories. These results show that the gut microbiome communities in CD and UC patients or in IBD and non-IBD patients differ, not only in relative abundance, but also in the growth rates of certain bacterial species, an important insight gained from our data analysis.

6. Discussion

This paper focuses on the permuted monotone matrix model with homoscedastic noise. If the noise is heteroscedastic, for example when the columns of the noise matrix are not independent or when the variances of the noise matrix entries are not identical, we argue that, as long as the marginal distributions of the noise matrix remain sub-Gaussian, the framework developed in this paper still applies. Specifically, in light of the recent work of Zhang et al. (2019) on heteroscedastic principal component analysis and singular value decomposition, the key analytical tools that parallel those used in the present work, such as concentration and perturbation inequalities associated with heteroscedastic random matrices, can be obtained by generalizing the results of Zhang et al. (2019). Such extensions are complicated and are left for future research.

The current theoretical framework was developed based on the approximately rank-one structure suggested by our specific metagenomic applications. Extensions to other settings are possible by modifying the proposed methods. In particular, the key observations made in § 2.1 apply to any monotone matrix satisfying Assumption 1. When the approximate rank-one assumption is violated, say if the monotone signal matrix is of rank $r > 1$, one could construct estimators based on the leading r singular values and singular vectors by following the same idea as in § 2.2, although the theoretical analysis may be technically challenging.

Acknowledgement

We thank the editor, associate editor and two referees for their constructive suggestions that have helped to improve the presentation of the paper. This research was supported by the U.S. National Institutes of Health and National Science Foundation. Ma would like to thank Rui Duan, Yuan Gao and Shulei Wang for stimulating discussions at various stages of the project.

Supplementary material

Supplementary Material available at *Biometrika* online includes detailed proofs of the main theorems and additional results of the simulation studies.

REFERENCES

ABEL, S., ZUR WIESCH, P. A., CHANG, H.-H., DAVIS, B. M., LIPSITCH, M. & WALDOR, M. K. (2015). Sequence tag–based analysis of microbial population dynamics. *Nature Meth.* **12**, 223–6.

ALMEIDA, A., NAYFACH, S., BOLAND, M., STROZZI, F., BERACOCHEA, M., SHI, Z. J., POLLARD, K. S., SAKHAROVA, E., PARKS, D. H., HUGENHOLTZ, P. et al. (2020). A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature Biotechnol.* **38**. DOI: 10.1038/s41587-020-0603-3.

BOULUND, F., PEREIRA, M. B., JONSSON, V. & KRISTIANSSON, E. (2018). Computational and statistical considerations in the analysis of metagenomic data. In *Metagenomics*. London: Elsevier, pp. 81–102.

BREMER, H. & CHURCHWARD, G. (1977). An examination of the Cooper-Helmstetter theory of DNA replication in bacteria and its underlying assumptions. *J. Theor. Biol.* **69**, 645–54.

BROWN, C. T., OLM, M. R., THOMAS, B. C. & BANFIELD, J. F. (2016). Measurement of bacterial replication rates in microbial communities. *Nature Biotechnol.* **34**, 1256–63.

BROWN, L. D. & GREENSHTEIN, E. (2009). Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *Ann. Statist.* **37**, 1685–704.

CAI, T. T. & ZHANG, A. (2018). Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *Ann. Statist.* **46**, 60–89.

COOPER, S. & HELMSTETTER, C. E. (1968). Chromosome replication and the division cycle of *Escherichia coli* B/r. *J. Molec. Biol.* **31**, 519–40.

COPAS, J. (1969). Compound decisions and empirical Bayes. *J. R. Statist. Soc.* B **31**, 397–417.

GAO, F., LUO, H. & ZHANG, C.-T. (2013). Doric 5.0: An updated database of oriC regions in both bacterial and archaeal genomes. *Nucleic Acids Res.* **41**, D90.

GAO, Y. & LI, H. (2018). Quantifying and comparing bacterial growth dynamics in multiple metagenomic samples. *Nature Meth.* **15**, 1041–4.

KOREM, T., ZEEVI, D., SUEZ, J., WEINBERGER, A., AVNIT-SAGI, T., POMPAN-LOTAN, M., MATOT, E., JONA, G., HARMELIN, A. & COHEN, N. (2015). Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. *Science* **249**, 1101–6.

LANGMEAD, B. & SALZBERG, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Meth.* **9**, 357–9.

LI, D., LIU, C.-M., LUO, R., SADAKANE, K. & LAM, T.-W. (2015). MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–6.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G. & DURBIN, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–9.

LLOYD-PRICE, J., ARZE, C., ANANTHAKRISHNAN, A. N., SCHIRMER, M., AVILA-PACHECO, J., POON, T. W., ANDREWS, E., AJAMI, N. J., BONHAM, K. S., BRISLAWN, C. J. et al. (2019). Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–62.

MA, R., CAI, T. T. & LI, H. (2020). Optimal permutation recovery in permuted monotone matrix model. *J. Am. Statist. Assoc.* to appear, DOI: 10.1080/01621459.2020.1713794.

MYHRVOLD, C., KOTULA, J. W., HICKS, W. M., CONWAY, N. J. & SILVER, P. A. (2015). A distributed cell division counter reveals growth dynamics in the gut microbiota. *Nature Commun.* **6**, article no. 10039.

ROBBINS, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. In *Proc. 2nd Berkeley Sympos. Mathematical Statistics and Probability*. Berkeley, California: University of California Press, pp. 131–49.

ROBBINS, H. (1964). The empirical Bayes approach to statistical decision problems. *Ann. Math. Statist.* **35**, 1–20.

SAMUEL, E. (1967). The compound statistical decision problem. *Sankhyā* A **29**, 123–40.

VON MEIJENFELDT, F. B., ARKHIPOVA, K., CAMBUY, D. D., COUTINHO, F. H. & DUTILH, B. E. (2019). Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biol.* **20**, article no. 217.

WU, Y.-W., SIMMONS, B. A. & SINGER, S. W. (2015). MaxBin 2.0: An automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–7.

ZHANG, A., CAI, T. T. & WU, Y. (2019). Heteroskedastic PCA: Algorithm, optimality, and applications. *arXiv:* 1810.08316v3.

ZHANG, C.-H. (2003). Compound decision theory and empirical Bayes methods. *Ann. Statist.* **31**, 379–90.

[*Received on* 16 *June* 2020. *Editorial decision on* 26 *August* 2020]