Bernoulli **25**(1), 2019, 623–653
https://doi.org/10.3150/17-BEJ1000

Optimal rates of statistical seriation

NICOLAS FLAMMARION[1], CHENG MAO[2,*] and PHILIPPE RIGOLLET[2,**]

[1] *Department of EECS, University of California, 465 Soda Hall, MC-1776, Berkeley, CA 94720-1776, USA. E-mail: flammarion@berkeley.edu*

[2] *Mathematics Department, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139-4307, USA. E-mail: *maocheng@mit.edu; **rigollet@math.mit.edu*

Given a matrix, the seriation problem consists in permuting its rows in such way that all its columns have the same shape, for example, they are monotone increasing. We propose a statistical approach to this problem where the matrix of interest is observed with noise and study the corresponding minimax rate of estimation of the matrices. Specifically, when the columns are either unimodal or monotone, we show that the least squares estimator is optimal up to logarithmic factors and adapts to matrices with a certain natural structure. Finally, we propose a computationally efficient estimator in the monotonic case and study its performance both theoretically and experimentally. Our work is at the intersection of shape constrained estimation and recent work that involves permutation learning, such as graph denoising and ranking.

Keywords: adaptation; matrix estimation; minimax estimation; permutation learning; shape constraints; statistical seriation

1. Introduction

Seriation has been a central technique for data analysis for over a century. It has roots in archeology and especially *sequence dating* where the goal is to recover the chronological order of sepultures based on artifacts found in them [58]. Since then seriation has found applications in a variety of disciplines ranging from anthropology [26] to sociology [36], biology [64] and marketing [4]. More recently, it was proposed as a method in computational biology for de novo DNA assembly [2]. See [49] for a detailed account of seriation in data analysis. In modern language, seriation belongs to the class of *unsupervised learning* problems. Akin to clustering, it aims at rearranging heterogeneous data into a simple structure that is amenable to better interpretation and understanding. Actually, in his seminal work on clustering, Hartigan [42] advocates for a post-processing of direct clustering with seriation for better data visualization. However, unlike clustering methods that quantize the data into a pre-specified number of clusters, seriation methods are truly nonparametric and "non-destructive", a term coined by Murtagh [56], meaning that it does not discard information from the data. Perhaps one of the most spectacular successes of seriation was achieved in bioinformatics where it was used to display genome-wide expression patterns [32]. Despite its widespread use, seriation has not been the subject of statistical analysis. The main goal of this paper is to propose a new model that is amenable to a statistical analysis of seriation.

To describe seriation in further details, we begin with a canonical problem, the *consecutive 1's problem* (C1P) [38] that is defined as follows. Given a binary matrix A the goal is to permute its rows in such a way that the resulting matrix enjoys the *consecutive 1's property*: each of its

1350-7265 © 2019 ISI/BS

columns is a vector $v = (v_1, \dots, v_n)^\top$ where $v_j = 1$ if and only if $a \le j \le b$ for two integers a, b between 1 and n. This problem arises in the archeology where the entry $A_{i,j}$ of matrix A indicates the presence of an artifact of type j in sepulture i. In his seminal work, egyptologist Flinders Petrie [58] formulated the hypothesis that two sepultures should be close in the time domain if they present similar sets of artifacts, which indicate that the matrix A should be close to a matrix having the consecutive 1's property. In an influential follow-up work, Robinson [60] generalized this problem to the case where $A_{i,j}$ *counts* the number of artifacts of type j in sepulture i. Robinson argues that *"types come into and get out of general use"* so that it is reasonable to assume that the columns of A are, in fact unimodal: the count of a certain type of artifact increases as it comes into general use and decreases as it gets out. Note that matrices that satisfy the consecutive 1's property have, in particular, unimodal columns. More generally, seriation is used to rearrange matrices whose rows are permuted and whose columns satisfy a nonparametric *shape constraint*. For example, the case where A has monotone columns arises in bipartite ranking under the strong stochastic transitivity assumption (see Section 2.2.2). In the rest of this paper, we consider both the unimodal and the monotone setting.

Because of the presence of a latent permutation, the C1P exhibits interesting algorithmic challenges already in the noiseless case and that have motivated much of its study. In particular, it is reducible to the famous Traveling Salesman Problem [41] as observed by statistician David Kendall [43–46] who employed early tools from multidimensional scaling as a heuristic to solve it. The C1P belongs to a more general class of problems that consist in optimizing various criteria over the discrete set of permutations and that can be recast as examples of the notoriously hard *quadratic assignment problem* [51]. While such problems are NP-hard in general, some examples, including C1P, may be solved efficiently using either combinatorial optimization [38], spectral methods [5] or convex optimization [35,50]. However, little is known about the robustness to statistical noise of such methods.

In order to set the benchmark for the noisy case, we propose a *statistical seriation model* and study optimal rates of estimation for this model. Assume that we observe an $n \times m$ matrix $Y = \Pi A + Z$, where Π is an unknown $n \times n$ permutation matrix, Z is an $n \times m$ noise matrix and $A \in \mathbb{R}^{n \times m}$ is assumed to have columns that satisfy a certain shape constraint. Our goal is to give estimators $\hat{\Pi}$ and $\hat{A}$ so that $\hat{\Pi}\hat{A}$ is close to ΠA. The shape constraint can be the consecutive 1's property, but more generally, we consider the class of matrices that have unimodal columns, which also include monotone columns as a special case. These terms will be formally defined at the end of this section.

The rest of the paper is organized as follows. In Section 2, we formulate the model and discuss related work. Section 3 collects our main results, including uniform and adaptive upper bounds for the least squares estimator together with corresponding minimax lower bounds in the general unimodal case. In Section 4, for the special case of monotone columns, we propose a computationally efficient alternative to the least squares estimator and study its rates of convergence both theoretically and numerically. Section 5 presents new bounds for unimodal regression implied by our analysis, which are minimax optimal up to logarithmic factors. Section 6 is devoted to the proofs of the results. We conclude with a discussion in Section 7.

Notation. For a positive integer n, define $[n] = \{1, \dots, n\}$. For a matrix $A \in \mathbb{R}^{n \times m}$, let $\|A\|_F$ denote its Frobenius norm, and let $A_{i,\cdot}$ be its ith row and $A_{\cdot,j}$ be its jth column. Let $\mathcal{B}^n(a, t)$

denote the Euclidean ball of radius t centered at a in $\mathbb{R}^n$. We use C and c to denote positive constants that may change from line to line. For any two sequences $(u_n)_n$ and $(v_n)_n$, we write $u_n \lesssim v_n$ if there exists an absolute constant $C > 0$ such that $u_n \le C v_n$ for all n. We define $u_n \gtrsim v_n$ analogously. Given two real numbers a, b, define $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$.

Denote the closed convex cone of increasing[1] sequences in $\mathbb{R}^n$ by $\mathcal{S}_n = \{a \in \mathbb{R}^n : a_1 \le \cdots \le a_n\}$. We define $\mathcal{S}^m$ to be the Cartesian product of m copies of $\mathcal{S}_n$ and we identify $\mathcal{S}^m$ to the set of $n \times m$ matrices with increasing columns.

For any $l \in [n]$, define the closed convex cone $\mathcal{C}_l = \{a \in \mathbb{R}^n : a_1 \le \cdots \le a_l\} \cap \{a \in \mathbb{R}^n : a_l \ge \cdots \ge a_n\}$, which consists of vectors in $\mathbb{R}^n$ that increase up to the lth entry and then decrease. Define the set $\mathcal{U}$ of unimodal sequences in $\mathbb{R}^n$ by $\mathcal{U} = \bigcup_{l=1}^n \mathcal{C}_l$. We define $\mathcal{U}^m$ to be the Cartesian product of m copies of $\mathcal{U}$ and we identify $\mathcal{U}^m$ to the set of $n \times m$ matrices with unimodal columns. It is also convenient to write $\mathcal{U}^m$ as a union of closed convex cones as follows. For $\mathbf{l} = (l_1, \dots, l_m) \in [n]^m$, let $\mathcal{C}_{\mathbf{l}}^m = \mathcal{C}_{l_1} \times \cdots \times \mathcal{C}_{l_m}$. Then $\mathcal{U}^m$ is the union of the n^m closed convex cones $\mathcal{C}_{\mathbf{l}}^m, \mathbf{l} \in [n]^m$.

Finally, let $\mathfrak{S}_n$ be the set of $n \times n$ permutation matrices and define $\mathcal{M} = \bigcup_{\Pi \in \mathfrak{S}_n} \Pi \mathcal{U}^m$ where $\Pi \mathcal{U}^m = \{\Pi A : A \in \mathcal{U}^m\}$, so that $\mathcal{M}$ is the union of the $n!n^m$ closed convex cones $\Pi \mathcal{C}_{\mathbf{l}}^m, \Pi \in \mathfrak{S}_n, \mathbf{l} \in [n]^m$.

2. Problem setup and related work

In this section, we formally state the problem of interest and discuss several lines of related work.

2.1. The seriation model

Suppose that we observe a matrix $Y \in \mathbb{R}^{n \times m}$, $n \ge 2$ such that

$$Y = \Pi^* A^* + Z, \tag{2.1}$$

where $A^* \in \mathcal{U}^m$, $\Pi \in \mathfrak{S}_n$ and Z is a centered sub-Gaussian noise matrix with variance proxy $\sigma^2 > 0$. Specifically, Z is a matrix such that $\mathbb{E}[Z] = 0$ and, for any $M \in \mathbb{R}^{n \times m}$,

$$\mathbb{E}\big[\exp\big(\mathsf{Tr}\big(Z^\top M\big)\big)\big] \le \exp\bigg(\frac{\sigma^2 \|M\|_F^2}{2}\bigg),$$

where $\mathsf{Tr}(\cdot)$ is the trace operator. We write $Z \sim \text{subG}_{n,m}(\sigma^2)$ or simply $Z \sim \text{subG}(\sigma^2)$ when dimensions are clear from the context.

Given the observation Y, our goal is to estimate the unknown pair (Π^*, A^*). The performance of an estimator $(\hat{\Pi}, \hat{A}) \in \mathfrak{S}_n \times \mathcal{U}^m$, is measured by the quadratic loss:

$$\frac{1}{nm} \big\| \hat{\Pi} \hat{A} - \Pi^* A^* \big\|_F^2.$$

[1] Throughout the paper, we loosely use the terms “increasing” and “decreasing” to mean “monotonically non-decreasing” and “monotonically non-increasing” respectively.

In particular, its expectation is the mean squared error. Since we are interested in estimating $\Pi^* A^* \in \mathcal{M}$, we can also view $\mathcal{M}$ as the parameter space.

In the general unimodal case, upper bounds on the above quadratic loss do not imply individual upper bounds on estimation of the matrix Π^* or the matrix A^* due to lack of identifiability. Nevertheless, if we further assume that the columns of A^* are monotone increasing, that is $A^* \in \mathcal{S}^m$, then the following lemma holds.

Lemma 2.1. *If $A^*, \tilde{A} \in \mathcal{S}^m$, then for any $\Pi^*, \tilde{\Pi} \in \mathfrak{S}_n$, we have that*

$$\|\tilde{A} - A^*\|_F^2 \le \|\tilde{\Pi}\tilde{A} - \Pi^* A^*\|_F^2,$$

and that

$$\|\tilde{\Pi} A^* - \Pi^* A^*\|_F^2 \le 4\|\tilde{\Pi}\tilde{A} - \Pi^* A^*\|_F^2.$$

Proof. Let $a, b \in \mathcal{S}_n$ and $b_\pi = (b_{\pi(1)}, \dots, b_{\pi(n)})$ where $\pi : [n] \to [n]$ is a permutation. It is easy to check that $\sum_{i=1}^n a_i b_i \ge \sum_{i=1}^n a_i b_{\pi(i)}$, so $\|a - b\|_2^2 \le \|a - b_\pi\|_2^2$. Applying this inequality to columns of matrices, we see that

$$\|\tilde{A} - A^*\|_F^2 \le \|\tilde{A} - \tilde{\Pi}^{-1}\Pi^* A^*\|_F^2 = \|\tilde{\Pi}\tilde{A} - \Pi^* A^*\|_F^2,$$

since $A^*, \tilde{A} \in \mathcal{S}^m$. Moreover, $\|\tilde{\Pi} A^* - \tilde{\Pi}\tilde{A}\|_F = \|A^* - \tilde{A}\|_F$, so

$$\|\tilde{\Pi} A^* - \Pi^* A^*\|_F \le \|A^* - \tilde{A}\|_F + \|\tilde{\Pi}\tilde{A} - \Pi^* A^*\|_F \le 2\|\tilde{\Pi}\tilde{A} - \Pi^* A^*\|_F,$$

by the triangle inequality and the previous display. □

Lemma 2.1 guarantees that $\|\tilde{\Pi}\tilde{A} - \Pi^* A^*\|_F$ is a pertinent measure of the performance of both $\tilde{\Pi}$ and $\tilde{A}$. Note further that $\|\tilde{\Pi} A^* - \Pi^* A^*\|_F$ is large if $\tilde{\Pi}$ misplaces rows of A^* that have large differences, and is small if $\tilde{\Pi}$ only misplaces rows of A^* that are close to each other. We argue that, in the seriation context, this measure of distance between permutations is more natural than ad hoc choices such as the trivial 0/1 distance or popular choices such as Kendall's τ or Spearman's ρ.

Apart from Section 4 (and Section 6.4), the rest of this paper focuses on the least squares (LS) estimator defined by

$$(\hat{\Pi}, \hat{A}) \in \operatorname*{argmin}_{(\Pi, A) \in \mathfrak{S}_n \times \mathcal{U}^m} \|Y - \Pi A\|_F^2. \tag{2.2}$$

Taking $\hat{M} = \hat{\Pi}\hat{A}$, we see that it is equivalent to define the LS estimator by

$$\hat{M} \in \operatorname*{argmin}_{M \in \mathcal{M}} \|Y - M\|_F^2. \tag{2.3}$$

Note that in our case, the set of parameters $\mathcal{M}$ is a union of $n!n^m$ closed convex cones but is not convex itself. Thus it is not clear how to compute the LS estimator efficiently. We discuss this aspect in further details in the context of monotone columns in Section 4. Nevertheless, the main focus of this paper is the least squares estimator which, as we shall see, is near-optimal in a minimax sense and therefore serves as a benchmark for the statistical seriation model.

2.2. Related work

Our work falls broadly in the scope of statistical inference under shape constraints but presents a major twist: the unknown latent permutation Π^*.

2.2.1. *Shape constrained regression*

To set our goals, we first consider the case where the permutation is known and assume without loss of generality that $\Pi^* = I_n$. In this case, we can estimate individually each column $A^*_{\cdot,j}$ by an estimator $\hat{A}_{\cdot,j}$ and then obtain an estimator $\hat{A}$ for the whole matrix by concatenating the columns $\hat{A}_{\cdot,j}$. Thus, the task is reduced to estimation of a vector θ^* which satisfies a certain shape constraint from an observation $y = \theta^* + z$ where $z \sim \mathrm{subG}_{n,1}(\sigma^2)$.

When θ^* is assumed to be increasing we speak of isotonic regression [7]. The LS estimator defined by $\hat{\theta} = \mathrm{argmin}_{\theta \in \mathcal{S}_n} \|\theta - y\|_2^2$ can be computed in closed form in $O(n)$ using the Pool-Adjacent-Violators algorithm (PAVA) [6,7,59] and its statistical performance has been studied by Zhang [72] (see also [30,53,57,69,71] for similar bounds using empirical process theory) who showed in the Gaussian case $z \sim N(0, \sigma^2 I_n)$ that the mean squared error behaves like

$$\frac{1}{n}\mathbb{E}\|\hat{\theta} - \theta^*\|_2^2 \asymp \left(\frac{\sigma^2 V(\theta^*)}{n}\right)^{2/3}, \tag{2.4}$$

where $V(\theta) = \max_{i\in[n]} \theta_i - \min_{i\in[n]} \theta_i$ is the variation of $\theta \in \mathbb{R}^n$. Note that $2/3 = 2\beta/(2\beta+1)$ for $\beta = 1$ so that this is the minimax rate of estimation of Lipschitz functions (see, e.g., [66]).

The rate in (2.4) is said to be *global* as it holds uniformly over the set of monotone vectors with variation $V(\theta^*)$. Recently, [20] have initiated the study of *adaptive* bounds that may be better if θ^* has a simpler structure in some sense. To define this structure, let $k(\theta) = \mathsf{card}(\{\theta_1, \dots, \theta_n\})$ denote the cardinality of entries of $\theta \in \mathbb{R}^n$. In this context, [20] showed that the LS estimator satisfies the adaptive bound

$$\frac{1}{n}\mathbb{E}\|\hat{\theta} - \theta^*\|_2^2 \le C \inf_{\theta \in \mathcal{S}_n}\left(\frac{\|\theta - \theta^*\|^2}{n} + \frac{\sigma^2 k(\theta)}{n}\log\frac{en}{k(\theta)}\right). \tag{2.5}$$

This result was extended in [9] to a sharp oracle inequality where $C = 1$. This bound was also shown to be optimal in a minimax sense [8,20].

Unlike its monotone counterpart, unimodal regression where $\theta^* \in \mathcal{U}$ has received sporadic attention [22,47,63]. This state of affairs is all the more surprising given that unimodal density estimation has been the subject of much more research [12,13,27,28,31,67]. It was recently shown in [22] that the LS estimator also adapts to $V(\theta^*)$ and $k(\theta^*)$ for unimodal regression:

$$\frac{1}{n}\|\hat{\theta} - \theta^*\|_2^2 \lesssim \min\left(\sigma^{4/3}\left(\frac{V(\theta^*) + \sigma}{n}\right)^{2/3}, \frac{\sigma^2}{n}k\big(\theta^*\big)^{3/2}(\log n)^{3/2}\right) \tag{2.6}$$

with probability at least $1 - n^{-\alpha}$ for some $\alpha > 0$. The exponent 3/2 in the second term was improved to 1 in the new version of [22] after the first version of our current paper was posted. Note that the exponents in (2.6) are different from the isotonic case. Our results will imply that they are not optimal and in fact the LS estimator achieves the same rate as in isotonic regression. See Corollary 5.1 for more details. The algorithmic aspect of unimodal regression has received more attention [15,16,37,40] and [65] showed that the LS estimator can be computed with time complexity $O(n)$ using a modified version of PAVA. Hence there is little difference between isotonic and unimodal regressions from both computational and statistical points of views.

2.2.2. *Latent permutation learning*

When the permutation Π^* is unknown the estimation problem is more involved. Noisy permutation learning was explicitly addressed in [24] where the problem of matching two sets of noisy vectors was studied from a statistical point of view. Given $n \times m$ matrices $Y = A + Z$ and $\tilde{Y} = \Pi^* A + \tilde{Z}$, where $A \in \mathbb{R}^{n\times m}$ is an unknown matrix and $\Pi^* \in \mathbb{R}^{n\times n}$ is an unknown permutation matrix, the goal is to recover Π^*. It was shown in [24] that if $\min_{i\neq j} \|A_{i,\cdot} - A_{j,\cdot}\|_2 \geq c\sigma((\log n)^{1/2} \vee (m\log n)^{1/4})$, then the LS estimator defined by $\hat{\Pi} = \operatorname{argmin}_{\Pi\in\mathfrak{S}_n} \|\Pi Y - \tilde{Y}\|_F^2$ recovers the true permutation with high probability. However they did not directly study the behavior of $\|\hat{\Pi}A - \Pi^* A\|_F^2$.

In his celebrated paper on matrix estimation [19], Sourav Chatterjee describes several noisy matrix models involving unknown latent permutations. One is the *nonparametric Bradley–Terry–Luce* (NP-BTL) model where we observe a matrix $Y \in \mathbb{R}^{n\times n}$ with independent entries $Y_{i,j} \sim \mathrm{Ber}(P_{i,j})$ for some unknown parameters $P = \{P_{i,j}\}_{1\leq i,j\leq n}$ where $P_{i,j} \in [0,1]$ is equal to the probability that item i is preferred over item j and $P_{j,i} = 1 - P_{i,j}$. Crucially, the NP-BTL model assumes the so-called *strong stochastic transitivity (SST)* [29,33] assumption: there exists an unknown permutation matrix $\Pi \in \mathbb{R}^{n\times n}$ such that the ordered matrix $A = \Pi^\top P \Pi$ satisfies $A_{1,k} \leq \cdots \leq A_{n,k}$ for all $k \in [n]$. Note that the NP-BTL model is a special case of our model (2.1) where $m = n$ and $Z \sim \mathrm{subG}(1/4)$ is taken to be Bernoulli. Chatterjee proposed an estimator $\hat{P}$ that leverages the fact that any matrix P in the NP-BTL model can be approximated by a low rank matrix and proved [19], Theorem 2.11, that $n^{-2}\|\hat{P} - P\|_F^2 \lesssim n^{-1/4}$, which was improved to $n^{-1/2}$ by [61] for a variation of this estimator. This method does not yield individual estimators of Π or A. Instead [23] proposed estimators $\hat{\Pi}$ and $\hat{A}$ so that $\hat{\Pi}\hat{A}\hat{\Pi}^\top$ estimates P with the same rate $n^{-1/2}$ up to a logarithmic factor. The non-optimality of this rate has been observed in [61] who showed that the correct rate should be of order n^{-1} up to a possible $\log n$ factor. However, it is not known whether a computationally efficient estimator could achieve the fast rate. A recent work [62] explored a new notion of adaptivity for which the authors proved a computational lower bound, and also proposed an efficient estimator whose rate of estimation matches that lower bound.

Also mentioned in Chatterjee's paper is the so-called *stochastic block model* that has since received such extensive attention in various communities that it is futile to attempt to establish a comprehensive list of references. Instead, we refer the reader to [39] and references therein. This paper establishes the minimax rates for this problem and its continuous limit, the graphon estimation problem and, as such, constitutes the state-of-the-art in the statistical literature. In the stochastic block model with $k \geq 2$ blocks, we assume that we observe a matrix $Y = P + Z$ where $P = \Pi A \Pi^\top$, $\Pi \in \mathbb{R}^{n\times n}$ is an unknown permutation matrix and A has a block structure, namely, there exist positive integers $n_1 < \cdots < n_k < n_{k+1} := n$, and k^2 real numbers $a_{s,t}$, $(s,t) \in [k]^2$

such that A has entries

$$A_{i,j} = \sum_{(s,t)\in[k]^2} a_{s,t}\mathbb{I}\{n_s \le i \le n_{s+1}, n_t \le j \le n_{t+1}\}, \qquad i, j \in [n].$$

While traditionally, the stochastic block model is a network model and therefore pertains only to Bernoulli observations, the more general case of sub-Gaussian additive error is also explicitly handled in [39]. For this problem, Gao, Lu and Zhou have established that the least squares estimator $\hat{P}$ satisfies $n^{-2}\|\hat{P} - P\|_F^2 \lesssim k^2/n^2 + (\log k)/n$ together with a matching lower bound. Using piecewise constant approximation to bivariate Hölder functions, they also establish that this estimator with a correct choice of k leads to minimax optimal estimation of smooth graphons. Both results exploit extensively the fact that the matrix P is equal to or can be well approximated by a piecewise constant matrix and our results below take a similar route by observing that monotone and unimodal vectors are also well approximated by piecewise constant ones. In addition, we allow for rectangular matrices.

In fact, our result can be also formulated as a network estimation problem but on a bipartite graph, thus falling at the intersection of the above two examples. Assume that n left nodes represent items and that m right nodes represent users. Assume further that we observe the $n \times m$ adjacency matrix Y of a random graph where the presence of edge (i, j) indicates that user j has purchased or liked item i. Define $P = \mathbb{E}[Y]$ and assume SST across items in the sense that there exists an unknown $n \times n$ permutation matrix Π^* such that $P = \Pi^* A^*$ and A^* is such that $A^*_{1,j} \le \cdots \le A^*_{n,j}$ for all users $j \in [m]$. This model of bipartite ranking falls into the scope of the statistical seriation model (2.1).

3. Main results

3.1. Adaptive oracle inequalities

For a matrix $A \in \mathcal{U}^m$, let $k(A_{\cdot,j}) = \mathsf{card}(\{A_{1,j}, \dots, A_{n,j}\})$ be the number of values taken by the jth column of A and define $K(A) = \sum_{j=1}^m k(A_{\cdot,j})$. Observe that $K(A) \ge m$. The first theorem shows that the LS estimator adapts to the complexity K.

Theorem 3.1. *For $A^* \in \mathbb{R}^{n\times m}$ and $Y = \Pi^* A^* + Z$, let $(\hat{\Pi}, \hat{A})$ be the LS estimator defined in (2.2). Then the following oracle inequality holds*

$$\frac{1}{nm}\|\hat{\Pi}\hat{A} - \Pi^* A^*\|_F^2 \lesssim \min_{A\in\mathcal{U}^m}\left(\frac{1}{nm}\|A - A^*\|_F^2 + \sigma^2\frac{K(A)}{nm}\log\frac{enm}{K(A)}\right) + \sigma^2\frac{\log n}{m} \tag{3.1}$$

with probability at least $1 - e^{-c(n+m)}$, $c > 0$. Moreover,

$$\frac{1}{nm}\mathbb{E}\|\hat{\Pi}\hat{A} - \Pi^* A^*\|_F^2 \lesssim \min_{A\in\mathcal{U}^m}\left(\frac{1}{nm}\|A - A^*\|_F^2 + \sigma^2\frac{K(A)}{nm}\log\frac{enm}{K(A)}\right) + \sigma^2\frac{\log n}{m}. \tag{3.2}$$

Note that while we assume that $A^* \in \mathcal{U}^m$ in (2.1), the above oracle inequalities hold in fact for any $A^* \in \mathbb{R}^{n\times m}$ even if its columns are *not* assumed to be unimodal. The oracle inequalities indicate that the LS estimator automatically trades off the approximation error $\|A - A^*\|_F^2$ for the stochastic error $\sigma^2 K(A)\log(enm/K(A))$. Moreover, 3 is the best constant we can achieve before the oracle approximation term when the error is expressed in the Frobenius norm, that is,

$$\|\hat{\Pi}\hat{A} - \Pi^* A^*\|_F \le \min_{A\in\mathcal{U}^m}\bigl(3\|A - A^*\|_F + \text{stochastic error terms}\bigr).$$

This is the content of (6.6) in the proof of Theorem 3.1. Making (3.1) and (3.2) into sharp oracle inequalities remains an interesting open problem.

If A^* is assumed to have unimodal columns, then we can take $A = A^*$ in (3.1) and (3.2) to get the following corollary.

Corollary 3.2. *For $A^* \in \mathcal{U}^m$ and $Y = \Pi^* A^* + Z$, the LS estimator $(\hat{\Pi}, \hat{A})$ satisfies*

$$\frac{1}{nm}\|\hat{\Pi}\hat{A} - \Pi^* A^*\|_F^2 \lesssim \sigma^2\left(\frac{K(A^*)}{nm}\log\frac{enm}{K(A^*)} + \frac{\log n}{m}\right)$$

with probability at least $1 - e^{-c(n+m)}$, $c > 0$. Moreover, the corresponding bound with the same rate holds in expectation.

The two terms in the adaptive bound can be understood as follows. The first term corresponds to the estimation of the matrix A^* with unimodal columns if the permutation Π^* is known. It can be viewed as a matrix version of the adaptive bound (2.5) for the vector case. The LS estimator adapts to the cardinality of entries of A^* as it achieves a provably better rate if $K(A^*)$ is smaller while not requiring knowledge of $K(A^*)$. The second term corresponds to the error due to the unknown permutation Π^*. As m grows to infinity this second term vanishes, because we have more samples to estimate Π^* better. If $m \ge n$, it is easy to check that the permutation term is dominated by the first term, so the rate of estimation is the same as if the permutation is known.

3.2. Global oracle inequalities

The bounds in Theorem 3.1 adapt to the cardinality of the oracle. In this subsection, we state another type of upper bounds for the LS estimator $(\hat{\Pi}, \hat{A})$. They are called global bounds because they hold uniformly over the class of matrices whose columns are unimodal and that have bounded variation. Recall that we call *variation* of a vector $a \in \mathbb{R}^n$ the scalar $V(a) \ge 0$ defined by

$$V(a) = \max_{1\le i\le n} a_i - \min_{1\le i\le n} a_i.$$

We extend this notion to a matrix $A \in \mathbb{R}^{n\times m}$ by defining

$$V(A) = \left(\frac{1}{m}\sum_{j=1}^{m} V(A_{\cdot,j})^{2/3}\right)^{3/2}.$$

While this 2/3-norm may seem odd at first sight, it turns out to be the correct extrapolation from vectors to matrices, at least in the context under consideration here. Indeed, the following upper bound, in which this quantity naturally appears, is matched by the lower bound of Theorem 3.6 up to logarithmic terms.

Theorem 3.3. *For $A^* \in \mathbb{R}^{n\times m}$ and $Y = \Pi^* A^* + Z$, let $(\hat{\Pi}, \hat{A})$ be the LS estimator defined in* (2.2). *Then it holds that*

$$\frac{1}{nm}\big\|\hat{\Pi}\hat{A} - \Pi^* A^*\big\|_F^2 \lesssim \min_{A\in\mathcal{U}^m}\left[\frac{1}{nm}\big\|A - A^*\big\|_F^2 + \left(\frac{\sigma^2 V(A)\log n}{n}\right)^{2/3}\right] + \sigma^2\frac{\log n}{n\wedge m} \tag{3.3}$$

with probability at least $1 - e^{-c(n+m)}$, $c > 0$. Moreover, the corresponding bound with the same rate holds in expectation.

If $A^* \in \mathcal{U}^m$, then taking $A = A^*$ in Theorem 3.3 leads to the following corollary that indicates that the LS estimator is adaptive to the quantity $V(A^*)$.

Corollary 3.4. *For $A^* \in \mathcal{U}^m$ and $Y = \Pi^* A^* + Z$, the LS estimator $(\hat{\Pi}, \hat{A})$ satisfies*

$$\frac{1}{nm}\big\|\hat{\Pi}\hat{A} - \Pi^* A^*\big\|_F^2 \lesssim \left(\frac{\sigma^2 V(A^*)\log n}{n}\right)^{2/3} + \sigma^2\frac{\log n}{n\wedge m}$$

with probability at least $1 - e^{-c(n+m)}$, $c > 0$. Moreover, the corresponding bound with the same rate holds in expectation.

Akin to the adaptive bound, the above inequality can be viewed as a sum of a matrix version of (2.4) and an error due to estimation of the unknown permutation. Observe that if $\sigma = 1$, $m \geq n^{2/3}$ and all the entries are bounded by a universal constant, then the rate of estimation simplifies to $\tilde{O}(n^{-2/3})$. Since every monotone vector is unimodal, the rate $\tilde{O}(n^{-2/3})$ also holds for the case where columns of A^* are monotone, which will be discussed in detail in Section 4. Recently, rates of $\tilde{O}(n^{-1})$ have been established for bi-isotonic matrices with latent permutations [23,61], where bi-isotonicity means that the columns and the rows of the underlying matrix are both monotone. We emphasize that our rate is slower because only the columns of the matrix are assumed to be unimodal or monotone, while no constraints are imposed on the rows. The minimax lower bounds below in fact suggest that the rate $\tilde{O}(n^{-2/3})$ is optimal up to a logarithmic factor.

Having stated the main upper bounds, we digress a little to remark that the proofs of Theorem 3.1 and Theorem 3.3 also yield a minimax optimal rate of estimation (up to logarithmic factors) for unimodal regression, which improves the bound (2.6). We discuss the details in Section 5.

3.3. Minimax lower bounds

Given the model $Y = \Pi^* A^* + Z$ where entries of Z are i.i.d. $N(0, \sigma^2)$ random variables, let $(\hat{\Pi}, \hat{A})$ denote any estimator of (Π^*, A^*), i.e., any pair in $\mathfrak{S}_n \times \mathbb{R}^{n\times m}$ that is measurable with respect to the observation Y. We will prove lower bounds that match the rates of estimation in Corollary 3.2 and Corollary 3.4 up to logarithmic factors. The combination of upper and lower bounds, implies simultaneous near optimality of the least squares estimator over a large scale of matrix classes.

For $m \le K_0 \le nm$ and $V_0 > 0$, define $\mathcal{U}^m_{K_0} = \{A \in \mathcal{U}^m : K(A) \le K_0\}$ and $\mathcal{U}^m(V_0) = \{A \in \mathcal{U}^m : V(A) \le V_0\}$. We present below two lower bounds, one for the adaptive rate uniformly over $\mathcal{U}^m_{K_0}$ and one for the global rate uniformly over $\mathcal{U}^m(V_0)$. This splitting into two cases is solely justified by better readability but it is worth noting that a stronger lower bound that holds on the intersection $\mathcal{U}^m_{K_0} \cap \mathcal{U}^m(V_0)$ can also be proved and is presented as Proposition 6.9.

Theorem 3.5. *There exists a constant $c \in (0,1)$ such that for any $K_0 \ge m$, and any estimator $(\hat{\Pi}, \hat{A})$, it holds that*

$$\sup_{(\Pi,A)\in\mathfrak{S}_n\times\mathcal{U}^m_{K_0}} \mathbb{P}_{\Pi A}\left[\frac{1}{nm}\|\hat{\Pi}\hat{A} - \Pi A\|_F^2 \gtrsim \sigma^2\left(\frac{K_0}{nm} + \frac{\log l}{m}\right)\right] \ge c,$$

where $l = \min(K_0 - m, m) + 1$ and $\mathbb{P}_{\Pi A}$ is the probability distribution of $Y = \Pi A + Z$. It follows that the lower bound with the same rate holds in expectation.

In fact, the lower bound holds for any estimator of the matrix $\Pi^* A^*$, not only those of the form $\hat{\Pi}\hat{A}$ with $\hat{A} \in \mathcal{U}^m$. The above lower bound matches the upper bound in Corollary 3.2 up to logarithmic factors.

Note the presence of a $\log l$ factor in the second term. If $l = 1$, then $K_0 = m$ which means that each column of A is simply a constant block, so $\Pi A = A$ for any $\Pi \in \mathfrak{S}_n$. In this case, the second term vanishes because the permutation does not play a role. More generally, the number $l - 1$ can be understood as the maximal number of columns of A on which the permutation does have an effect. The larger l, the harder the estimation. It is easy to check that if $l \ge n$ the second term in the lower bound will be dominated by the first term in the upper bound.

A lower bound corresponding to Corollary 3.4 also holds.

Theorem 3.6. *There exists a constant $c \in (0,1)$ such that for any $V_0 \ge 0$, and any estimator $(\hat{\Pi}, \hat{A})$, it holds that*

$$\sup_{(\Pi,A)\in\mathfrak{S}_n\times\mathcal{U}^m(V_0)} \mathbb{P}_{\Pi A}\left[\frac{1}{nm}\|\hat{\Pi}\hat{A} - \Pi A\|_F^2 \gtrsim \left(\frac{\sigma^2 V_0}{n}\right)^{2/3} + \frac{\sigma^2}{n} + \frac{\sigma^2}{m} \wedge m^2 V_0^2\right] \ge c,$$

where $\mathbb{P}_{\Pi A}$ is the probability distribution of $Y = \Pi A + Z$. The lower bound with the same rate also holds in expectation.

There is a slight mismatch between the upper bound of Corollary 3.4 and the lower bound of Theorem 3.6 above. Indeed the lower bound features a term $\frac{\sigma^2}{m} \wedge m^2 V_0^2$ instead of just $\frac{\sigma^2}{m}$. In the regime $m^2 V_0^2 < \frac{\sigma^2}{m}$, where A has very small variation, the LS estimator may not be optimal. Proposition 3.7 below, whose proof can be found in the supplement [34], indicates that a matrix with constant columns obtained by averaging achieves optimality in this extreme regime.

Proposition 3.7. *For $Y = \Pi^* A^* + Z$ where $Z \sim \mathrm{subG}(\sigma^2)$, let $\hat{\Pi} = I_n$ and $\hat{A}$ be defined by $\hat{A}_{i,j} = \frac{1}{n} \sum_{k=1}^{n} Y_{k,j}$ for all $(i, j) \in [n] \times [m]$. Then,*

$$\frac{1}{nm} \left\| \hat{\Pi}\hat{A} - \Pi^* A^* \right\|_F^2 \lesssim \frac{\sigma^2}{n} + m^2 V(A)^2$$

with probability at least $1 - \exp(-m)$ and the corresponding bound with the same rate holds in expectation.

4. Further results in the monotone case

A particularly interesting subset of unimodal matrices is $\mathcal{S}^m$, the set of $n \times m$ matrices with monotonically increasing columns. While it does not amount to the seriation problem in its full generality, this special case is of prime importance in the context of shape constrained estimation as illustrated by the discussion and references in Section 2.2. In fact, it covers the example of bipartite ranking discussed at the end of Section 2.2. In the rest of this section, we devote further investigation to this important case. To that end, consider the model (2.1) where we further assume that $A^* \in \mathcal{S}^m$. We refer to this model as the *monotone seriation model*. In this context, define the LS estimator by

$$(\hat{\Pi}, \hat{A}) \in \operatorname*{argmin}_{(\Pi, A) \in \mathfrak{S}_n \times \mathcal{S}^m} \|Y - \Pi A\|_F^2.$$

Since $\mathcal{S}^m$ is a convex subset of $\mathcal{U}^m$, it is easily seen that the upper bounds in Theorem 3.1 and 3.3 remain valid in this case. The lower bounds of Theorem 3.5 (with $\log l$ replaced by 1) and Theorem 3.6 also extend to this case; see Section 6.3.

Although for unimodal matrices the established error bounds do not imply any bounds on estimation of A^* or Π^* in general, for the monotonic case, however, Lemma 2.1 yields that

$$\left\| \hat{A} - A^* \right\|_F^2 \vee \frac{1}{4} \left\| (\hat{\Pi} - \Pi^*) A^* \right\|_F^2 \leq \left\| \hat{\Pi}\hat{A} - \Pi^* A^* \right\|_F^2$$

so that the LS estimator $(\hat{\Pi}, \hat{A})$ also leads to good individual estimators of Π^* and A^* respectively.

4.1. RankScore: An efficient estimator and its performance

Because it requires optimizing over a union of $n!$ cones $\Pi\mathcal{S}^m$, no efficient way of computing the LS estimator is known since. As an alternative, we describe a simple and efficient algorithm to estimate (Π^*, A^*) and study its rate of estimation.

The main difficulty of the problem lies in providing an efficient estimator $\tilde{\Pi}$ of Π^*, because after determining $\tilde{\Pi}$ we may project Y onto the convex cone $\tilde{\Pi}\mathcal{S}^m$ efficiently to estimate A^*. Recovering the permutation Π^* is equivalent to sorting the rows of $\Pi^* A^*$ from their noisy version Y. One simple method to aggregate information across columns, which we call RankSum, is to sort the rows of Y so that they have increasing row sums. However, it is easy to observe that this method fails if

$$A^* = \begin{bmatrix} 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 \\ \sqrt{m} & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ \sqrt{m} & 0 & \dots & 0 \end{bmatrix}, \tag{4.1}$$

where the last $\lfloor \frac{n}{2} \rfloor$ entries in the first column of A^* are equal to $\sqrt{m}$ and the entries of Z are i.i.d. standard Gaussian variables. Because the sum of noise in each row is of order $\sqrt{m}$ which is no less than the gaps between row sums of A^*, RankSum will place a nonzero row before a zero row with a constant probability. Therefore, if $\tilde{\Pi}$ is the permutation given by RankSum, then $\|\tilde{\Pi}\tilde{A} - \Pi^* A^*\|_F^2$ will be of order nm regardless of the matrix $\tilde{A} \in \mathcal{S}^m$, so we have no hope of consistent estimation in general.

In fact, it is easy to distinguish the two types of rows of A^* even when noise is present, for example, by looking at the first entry of a row. To circumvent the issue raised by this A^*, we would like to combine the information from rows sums with that from each individual column. This motivates us to consider the following method called RankScore, which outperforms RankSum and yields consistent estimation.

For $A^* \in \mathbb{R}^{n \times m}$ and $i, i' \in [n]$, define

$$\Delta_{A^*}(i, i') = \max_{j \in [m]} (A^*_{i',j} - A^*_{i,j}) \vee \frac{1}{\sqrt{m}} \sum_{j=1}^{m} (A^*_{i',j} - A^*_{i,j})$$

and define $\Delta_Y(i, i')$ analogously. The quantity $\Delta_{A^*}(i, i')$ measures the difference between row i and row i' of A^* by either the largest difference between two corresponding entries, or the difference between the row sums scaled by the effective noise level $m^{-1/2}$, whichever is larger. If the noisy version $\Delta_Y(i, i')$ is larger than some threshold τ, then with high probability row i of Y should be placed after row i' in the original order. The procedure RankScore aggregates the comparison results between all pairs of rows of Y as follows:

1. For each $i \in [n]$, define the score s_i of the ith row of Y by

$$s_i = \sum_{l=1}^{n} \mathbb{I}(\Delta_Y(l, i) \geq 2\tau), \tag{4.2}$$

where $\tau := C\sigma\sqrt{\log(nm)}$ for some tuning constant C (see Section 6.4 for details).
2. Order the rows of Y so that their scores are increasing, with ties broken arbitrarily.

The score s_i is just the number of comparisons row i wins. Intuitively, rows with larger entries will win more comparisons and thus be placed after rows with smaller entries. Hence, RankScore can be viewed as a variant of the classical counting-based method for ranking, Copeland's method [25], with a counting rule designed specifically for the model under consideration.

The RankScore procedure recovers an order of the rows of Y, which leads to an estimator $\tilde{\Pi}$ of the permutation. Then we define $\tilde{A} \in \mathcal{S}^m$ so that $\tilde{\Pi}\tilde{A}$ is the projection of Y onto the convex cone $\tilde{\Pi}\mathcal{S}^m$.

To quantify the rate of estimation for the RankScore estimator $(\tilde{\Pi}, \tilde{A})$, we define a new quantity $R(A)$ for $A \in \mathcal{S}^m$ as follows:

$$R(A) = \frac{1}{n} \max_{\substack{\mathcal{I} \subset [n]^2 \\ |\mathcal{I}|=n}} \sum_{(i,j) \in \mathcal{I}} \left(\frac{\|A_{i,\cdot} - A_{j,\cdot}\|_2^2}{\|A_{i,\cdot} - A_{j,\cdot}\|_\infty^2} \wedge \frac{m\|A_{i,\cdot} - A_{j,\cdot}\|_2^2}{\|A_{i,\cdot} - A_{j,\cdot}\|_1^2} \right), \tag{4.3}$$

where the summand is understood to be 1 if the rows $A_{i,\cdot}$ and $A_{j,\cdot}$ are identical.

To understand what properties of A the quantity $R(A)$ captures, consider the difference between the rows $A_{i,\cdot}$ and $A_{j,\cdot}$, denoted by $u \in \mathbb{R}^m$. First, the quantity $\|u\|_2^2/\|u\|_\infty^2$ is small when u is sparse. We have $\|u\|_2^2/\|u\|_\infty^2 \geq 1$ with equality achieved when $\|u\|_0 = 1$. Second, the quantity $m\|u\|_2^2/\|u\|_1^2$ is small when u is dense. We have $m\|u\|_2^2/\|u\|_1^2 \geq 1$ with equality achieved when all entries of u are the same. In particular, it holds that $R(A) \geq 1$, and $R(A)$ is small when the differences between rows of A are either very sparse or very dense. For example, if A is the matrix in (4.1), then the difference between any two distinct rows is 1-sparse, so we have $R(A) = 1$. Another example is

$$A = \begin{bmatrix} \mathbf{0} \\ \mathbf{1} \end{bmatrix}, \tag{4.4}$$

where the lower $\lfloor n/2 \rfloor$ rows of A are all ones while the remaining entries are all zeros. For this matrix, the difference between any two distinct rows is the all ones vector, so again we have $R(A) = 1$.

Moreover, $\|u\|_2^2 \leq \|u\|_1\|u\|_\infty$ by Hölder's inequality, so $\frac{\|u\|_2^2}{\|u\|_\infty^2} \wedge \frac{m\|u\|_2^2}{\|u\|_1^2} \leq \sqrt{m}$ as the product of the two terms is no larger than m. The equality is achieved by $u = (1, \dots, 1, 0, \dots, 0)$ where the first $\sqrt{m}$ entries are equal to one. Therefore, we have

$$R(A) \in [1, \sqrt{m}]. \tag{4.5}$$

Roughly speaking, the quantity $R(A)$ is large if there exist $\Theta(n)$ pairs of rows for which the differences are $\sqrt{m}$-sparse. An example of such an A is the lower triangular matrix with all ones on the lower triangle. We can take pairs of rows that are $\sqrt{m}$ positions apart, and their differences are exactly $\sqrt{m}$-sparse binary vectors. Thus, we have $R(A) \asymp \sqrt{m}$.

Since RankScore makes use of entrywise differences between rows, together with the difference between row sums, we expect a better performance of RankScore when the differences between rows of A^* are either very sparse or very dense, which is exactly what captured by the quantity $R(A^*)$. Therefore, it is natural that the estimator $(\tilde{\Pi}, \tilde{A})$ enjoys the following rate of estimation, characterized by $R(A^*)$ together with $K(A)$ defined in the previous section.

Theorem 4.1. *For $A^* \in \mathcal{S}^m$ and $Y = \Pi^* A^* + Z$, let $(\tilde{\Pi}, \tilde{A})$ be the estimator defined above using the* RankScore *procedure with threshold $\tau = 3\sigma\sqrt{(C+1)\log(nm)}$, $C > 0$. Then it holds that*

$$\frac{1}{nm}\|\tilde{\Pi}\tilde{A} - \Pi^* A^*\|_F^2 \lesssim \min_{A \in \mathcal{S}^m}\left(\frac{1}{nm}\|A - A^*\|_F^2 + \sigma^2 \frac{K(A)}{nm}\log\frac{enm}{K(A)}\right) + (C+1)\sigma^2 \frac{R(A^*)\log(nm)}{m},$$

with probability at least $1 - e^{-c(n+m)} - (nm)^{-C}$ for some constant $c > 0$.

The quantity $R(A^*)$ only depends on the matrix A^*. If $R(A^*)$ is bounded logarithmically, the estimator $(\tilde{\Pi}, \tilde{A})$ achieves the minimax rate up to logarithmic factors. In any case, $R(A^*) \le \sqrt{m}$, so the estimator is still consistent with the permutation error (i.e., the last term) decaying at a rate $\tilde{O}(\frac{1}{\sqrt{m}})$. Furthermore, it is worth noting that $R(A^*)$ is not needed to construct $(\tilde{\Pi}, \tilde{A})$, so the estimator adapts to $R(A^*)$ automatically.

Remark 4.2. In the same way that Theorem 3.3 follows from Theorem 3.1, we can deduce from Theorem 4.1 a global bound for the estimator $(\tilde{\Pi}, \tilde{A})$ which has rate

$$\left(\frac{\sigma^2 V(A^*)\log n}{n}\right)^{2/3} + \sigma^2\left(\frac{\log n}{n} + R(A^*)\frac{\log(nm)}{m}\right).$$

4.2. Simulations

We corroborate the theoretical results above with a numerical comparison between the RankSum and RankScore procedures.

Consider the model (2.1) with $A^* \in \mathcal{S}^m$ and assume without loss of generality that $\Pi^* = I_n$. For various $n \times m$ matrices A^*, we generate observations $Y = A^* + Z$ where entries of Z are i.i.d. standard Gaussian variables. The performance of the estimators given by RankScore and RankSum defined above is compared to the performance of the oracle $\hat{A}^{\text{oracle}}$ defined by the projection of Y onto the cone $\mathcal{S}^m$. Note that we are not able to compute the LS estimator efficiently, so instead the oracle estimator is used as the benchmark. For the RankScore estimator, we take $\tau = 6$. The curves are generated based on 30 equally spaced points on the base-10 logarithmic scale, and all results are averaged over 10 replications. The vertical axis represents the estimation error of an estimator $\hat{\Pi}\hat{A}$, measured by the sample mean of $\log_{10}(\frac{1}{nm}\|\hat{\Pi}\hat{A} - A^*\|_F^2)$ unless otherwise specified.

We begin with a simple example for which we set $n = m$. For each $\alpha \in [0, 1]$, define a matrix $A^* = A^*(\alpha) \in \mathbb{R}^{n \times n}$ by $A^*_{i,j} = m^{(1-\alpha)/2}$ for $n/2 \le i \le n$, $1 \le j \le m^\alpha$ and $A^*_{i,j} = 0$ otherwise. Note that A^* is an interpolation between the matrix in (4.1) (where $\alpha = 0$) and the matrix in (4.4) (where $\alpha = 1$). The nonzero rows of A^* have ℓ_2-norm equal to $\sqrt{m}$ for any $\alpha \in [0, 1]$.

In Figure 1, we plot the estimation errors of the oracle, RankScore and RankSum estimators for this A^* in the three plots, respectively. As expected, RankSum has poor performance in estimating the true permutation when α is close to zero, because it fails to exploit the differences

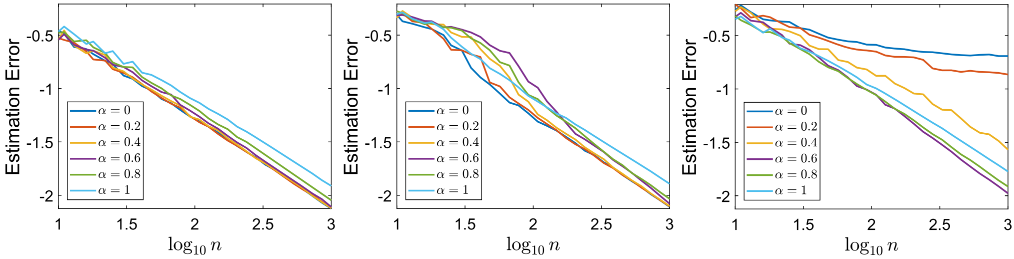

Figure 1. Estimation errors of the three estimators for $A^* = A^*(\alpha)$ where α ranges from 0 to 1. Left: the oracle estimator; Middle: the RankScore estimator; Right: the RankSum estimator.

between rows along individual columns. When α is close to one, the weight of a nonzero row of A^* is distributed evenly across the columns, so it is appropriate to only consider row sums and thus RankSum behaves well. On the other hand, RankScore outperforms RankSum in recovering the permutation for any $\alpha \in [0, 1]$ when n is large, and it has roughly the same performance as the oracle. According to the discussion after (4.3), we have $R(A^*) = 1$ for $\alpha = 0$ or 1. Thus, Theorem 4.1 predicts the fast rate, which is verified by the experiment. For α close to 1/2, however, Theorem 4.1 only guarantees a rate $\tilde{O}(m^{-1/2})$ while the experiment suggests that RankScore still behaves as well as the oracle. Hence, improving the adaptive bound in Theorem 4.1 remains an interesting problem for future research.

Note that the performance of each estimator for $\alpha = 0.6$ is slightly better than that for $\alpha = 1$. This is not inconsistent with our theoretical guarantees as the bounds we proved are up to logarithmic factors. Achieving sharper bounds to explain such a phenomenon also remains an interesting open question out of the scope of the present work.

In Figure 2, we compare the performance of RankScore to that of the oracle in three regimes of (n, m). The matrices A^* are randomly generated for different values of n and m as follows. For the right plot, A^* is generated so that $V(A^*) \leq 1$, by sorting the columns of a matrix with i.i.d. $U(0, 1)$ entries. For the left plot, we further require that $K(A^*) = 5m$ by uniformly partitioning

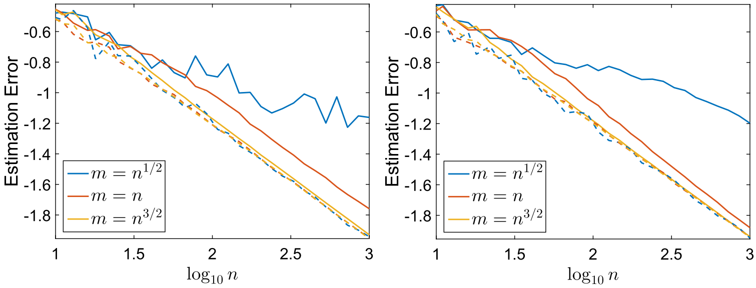

Figure 2. Estimation errors of the oracle (dashed lines) and RankScore (solid lines) for different regimes of (n, m) and randomly generated A^* of size $n \times m$. Left: $K(A^*) = 5m$; Right: $V(A^*) \leq 1$.

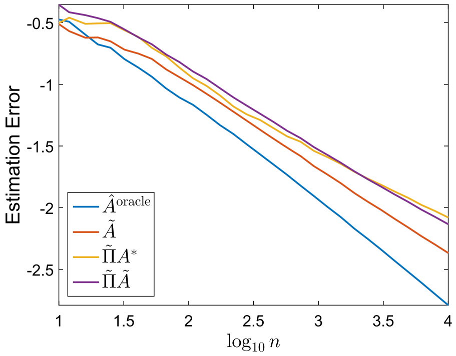

Figure 3. Various estimation errors of the oracle and RankScore for the triangular matrix.

each column of A^* into five blocks and assigning each block the corresponding value from a sorted sample of five i.i.d. $U(0,1)$ variables.

Since the oracle knows the true permutation, its behavior is independent of m, and its rates of estimation are bounded by $\frac{\log n}{n}$ for $K(A^*) = 5m$ and $(\frac{\log n}{n})^{\frac{2}{3}}$ for $V(A^*) = 1$ respectively, by Theorems 3.1 and 3.3. (The difference is minor in the plots as n is not sufficiently large). For RankScore, the permutation term dominates the estimation term when $m = n^{1/2}$ by Theorem 4.1. From the plots, the rates of estimation are better than $\tilde{O}(n^{-1/4})$ predicted by the worst-case analysis in both examples. For $m = n$, we also observe rates of estimation faster than the worst-case rate $\tilde{O}(n^{-1/2})$ and close to the oracle rates. We could explain this phenomenon by $R(A^*) < \sqrt{m}$, but such an interpretation may not be optimal since our analysis is based on worst-case deterministic A^*. Potential study of random designs of A^* is left open. Finally, for $m = n^{3/2}$, the permutation term is of order $\tilde{O}(n^{-3/4})$ theoretically, in between of the oracle rates for the two cases. Indeed RankScore has almost the same performance as the oracle experimentally. Overall Figure 2 illustrates the good behavior of RankScore in these random scenarios.

To conclude our numerical experiments, we consider the $n \times n$ lower triangular matrix A^* defined by $A^*_{i,j} = \mathbb{I}(i \geq j)$. For this matrix, it is easy to check that $K(A^*) = 2n - 1$ and $R(A^*) \approx \sqrt{n}$. We plot in Figure 3 the estimation errors of $\tilde{\Pi}\tilde{A}$, $\tilde{\Pi}A^*$ and $\tilde{A}$ given by RankScore, in addition to the oracle. By Theorem 4.1, the rate of estimation achieved by $\tilde{\Pi}\tilde{A}$ is of order $\tilde{O}(n^{-1/2})$, while that achieved by the oracle is of order $\tilde{O}(n^{-1})$ since there is no permutation term. The plot confirms this discrepancy. Moreover, $\frac{1}{n^2}\|\tilde{\Pi}A^* - A^*\|_F^2$ is an appropriate measure of the performance of $\tilde{\Pi}$ by Lemmas 6.13 and 2.1, and the plot suggests that the rates of estimation achieved by $\tilde{\Pi}A^*$ and $\tilde{\Pi}\tilde{A}$ are about the same order. Finally $\tilde{A}$ seems to have a slightly faster rate of estimation than $\tilde{\Pi}\tilde{A}$, so in practice $\tilde{A}$ could be used to estimate A. However, we refrain from making an explicit conjecture about the rate.

5. Unimodal regression

If the permutation in the main model (2.1) is known, then the estimation problem simply becomes a concatenation of m unimodal regressions. In fact, our proofs imply new oracle inequalities for unimodal regression. Recall that $\mathcal{U}$ denotes the cone of unimodal vectors in $\mathbb{R}^n$. Suppose that we observe

$$y = \theta^* + z,$$

where $\theta^* \in \mathbb{R}^n$ and z is a sub-Gaussian vector with variance proxy σ^2. Define the LS estimator $\hat{\theta}$ by

$$\hat{\theta} \in \underset{\theta \in \mathcal{U}}{\operatorname{argmin}} \|\theta - y\|_2^2.$$

Moreover let $k(\theta) = \mathsf{card}(\{\theta_1, \dots, \theta_n\})$ and $V(\theta) = \max_{i\in[n]} \theta_i - \min_{i\in[n]} \theta_i$.

Corollary 5.1. *There exists a constant $c > 0$ such that with probability at least $1 - n^{-\alpha}$, $\alpha \geq 1$,*

$$\frac{1}{n}\|\hat{\theta} - \theta^*\|_2^2 \lesssim \min_{\theta \in \mathcal{U}} \left(\frac{1}{n}\|\theta - \theta^*\|_2^2 + \sigma^2 \frac{k(\theta)}{n} \log \frac{en}{k(\theta)} \right) + \alpha\sigma^2 \frac{\log n}{n} \tag{5.1}$$

and

$$\frac{1}{n}\|\hat{\theta} - \theta^*\|_2^2 \lesssim \min_{\theta \in \mathcal{U}} \left[\frac{1}{n}\|\theta - \theta^*\|_2^2 + \left(\frac{\sigma^2 V(\theta) \log n}{n} \right)^{2/3} \right] + \alpha\sigma^2 \frac{\log n}{n}.$$

The corresponding bounds in expectation also hold.

The proof of Corollary 5.1 can be found in the supplement [34]. Note that the bounds above match the minimax lower bounds for isotonic regression in [8] up to logarithmic factors. Since every monotone vector is unimodal, lower bounds for isotonic regression automatically hold for unimodal regression. Therefore, we have proved that the LS estimator is minimax optimal up to logarithmic factors for unimodal regression.

A result similar to (5.1) was obtained by Pierre C. Bellec in the revision of [9] that was prepared independently and contemporaneously to this paper. Sabyasachi Chatterjee and John Lafferty also improved their bounds to having optimal exponents [22] after the first version of our current paper was posted. Interestingly Bellec employs bounds on the statistical dimension by leveraging results from [1], and Chatterjee and Lafferty use both the variational formula and the statistical dimension. Moreover, their results are presented in the well-specified case where $\theta^* \in \mathcal{U}$ and $\theta = \theta^*$.

6. Proofs

In this section, we provide the proofs of the main results.

6.1. Proof of the upper bounds

Before proving the main theorems, we discuss two methods adopted in recent works to bound the error of the LS estimator in shape constrained regression, in a general setting. Consider the least squares estimator $\hat{\theta}$ of the model $y = \theta^* + z$, where θ^* lies in a parameter space Θ and z is Gaussian noise. One way to study $\mathbb{E}\|\hat{\theta} - \theta^*\|_2^2$ is to use the *statistical dimension* [1] of a convex cone Θ defined by

$$\mathbb{E}\Big[\Big(\sup_{\theta\in\Theta,\,\|\theta\|_2\le 1} \langle\theta, z\rangle\Big)^2\Big].$$

This has been successfully applied to isotonic and more general shape constrained regression [9,20].

Another prominent approach is to express the error of the LS estimator via what is known as *Chatterjee's variational formula*, proved in [18] and given by

$$\big\|\hat{\theta} - \theta^*\big\|_2 = \operatorname*{argmax}_{t\ge 0}\Big(\sup_{\theta\in\Theta,\,\|\theta-\theta^*\|_2\le t} \big\langle\theta - \theta^*, z\big\rangle - \frac{t^2}{2}\Big). \tag{6.1}$$

Note that the first term is related to the *Gaussian width* (see, e.g., [17]) of Θ defined by $\mathbb{E}[\sup_{\theta\in\Theta}\langle\theta, z\rangle]$, whose connection to the statistical dimension was studied in [1]. The variational formula was first proposed for convex regression [18], and later exploited in several different settings, including matrix estimation with shape constraints [21] and unimodal regression [22]. Similar ideas have appeared in other works, for example, analysis of empirical risk minimization [55], ranking from pairwise comparison [61] and isotonic regression [9]. In this latter work, Bellec has used the statistical dimension approach to prove spectacularly sharp oracle inequalities that seem to be currently out of reach for methods based on Chatterjee's variational formula (6.1). On the other hand, Chatterjee's variational formula seems more flexible as computations of the statistical dimension based on [1] are currently limited to convex sets Θ with a polyhedral structure. In this paper, we use exclusively Chatterjee's variational formula.

6.1.1. *A variational formula for the error of the LS estimator*

We begin the proof by stating an extension of Chatterjee's variational formula. While we only need this lemma to hold for a union of closed convex sets, we present a version that holds for all closed sets. The latter extension was suggested to us by Pierre C. Bellec in a private communication [10].

Lemma 6.1. *Let $\mathcal{C}$ be a closed subset of $\mathbb{R}^d$. Suppose that $y = a^* + z$ where $a^* \in \mathcal{C}$ and $z \in \mathbb{R}^d$. Let $\hat{a} \in \operatorname{argmin}_{a\in\mathcal{C}} \|y - a\|_2^2$ be a projection of y onto $\mathcal{C}$. Define the function $f_{a^*} : \mathbb{R}_+ \to \mathbb{R}$ by*

$$f_{a^*}(t) = \sup_{a\in\mathcal{C}\cap\mathcal{B}^d(a^*,t)} \big\langle a - a^*, z\big\rangle - \frac{t^2}{2}.$$

Then we have

$$\big\|\hat{a} - a^*\big\|_2 \in \operatorname*{argmax}_{t\ge 0} f_{a^*}(t). \tag{6.2}$$

Moreover, if there exists $t^* > 0$ such that $f_{a^*}(t) < 0$ for all $t \ge t^*$, then $\|\hat{a} - a^*\|_2 \le t^*$.

Proof. By definition,

$$\hat{a} \in \operatorname*{argmin}_{a \in \mathcal{C}} \big(\|a - a^*\|_2^2 - 2\langle a - a^*, z \rangle + \|z\|_2^2 \big) = \operatorname*{argmax}_{a \in \mathcal{C}} \Big(\langle a - a^*, z \rangle - \frac{1}{2} \|a - a^*\|_2^2 \Big).$$

Together with the definition of f_{a^*}, this implies that

$$\begin{aligned} f_{a^*}\big(\|\hat{a} - a^*\|_2\big) &\geq \langle \hat{a} - a^*, z \rangle - \frac{1}{2} \|\hat{a} - a^*\|_2^2 \\ &\geq \sup_{a \in \mathcal{C} \cap \mathcal{B}^d(a^*, t)} \Big(\langle a - a^*, z \rangle - \frac{1}{2} \|a - a^*\|_2^2 \Big) \\ &\geq \sup_{a \in \mathcal{C} \cap \mathcal{B}^d(a^*, t)} \langle a - a^*, z \rangle - \frac{t^2}{2} = f_{a^*}(t). \end{aligned}$$

Therefore (6.2) follows.

Furthermore, suppose that there is $t^* > 0$ such that $f_{a^*}(t) < 0$ for all $t \geq t^*$. Since $f_{a^*}(\|\hat{a} - a^*\|_2) \geq f_{a^*}(0) = 0$, we have $\|\hat{a} - a^*\|_2 \leq t^*$. □

Note that this structural result holds for any error vector $z \in \mathbb{R}^d$ and any closed set $\mathcal{C}$ which is not necessarily convex. In particular, this extends the results in [18] and [22] which hold for convex sets and finite unions of convex sets respectively.

6.1.2. *Proof of Theorem* 3.1

For our purpose, we need a standard chaining bound on the supremum of a sub-Gaussian process that holds in high probability. The interested readers can find the proof, for example, in [68], Theorem 5.29, and refer to [48] for a more detailed account of the technique.

Lemma 6.2 (Chaining tail inequality). *Let* $\Theta \subset \mathbb{R}^d$ *and* $z \sim \text{subG}(\sigma^2)$ *in* $\mathbb{R}^d$. *For any* $\theta_0 \in \Theta$, *it holds that*

$$\sup_{\theta \in \Theta} \langle \theta - \theta_0, z \rangle \leq C\sigma \int_0^{\operatorname{diam}(\Theta)} \sqrt{\log N\big(\Theta, \|\cdot\|_2, \varepsilon\big)}\, d\varepsilon + s$$

with probability at least $1 - C \exp(-\frac{cs^2}{\sigma^2 \operatorname{diam}(\Theta)^2})$ *where* C *and* c *are positive constants.*

Let $\tilde{A} \in \mathcal{U}^m$. To lighten the notation, we define two rates of estimation:

$$R_1 = R_1(\tilde{A}, n) = \sigma \Bigg(\sqrt{K(\tilde{A}) \log \frac{enm}{K(\tilde{A})}} + \sqrt{n \log n} \Bigg) \tag{6.3}$$

and

$$R_2 = R_2(\tilde{A}, n) = \sigma^2 \bigg(K(\tilde{A}) \log \frac{enm}{K(\tilde{A})} + n \log n \bigg). \tag{6.4}$$

Note that $R_2 \leq R_1^2 \leq 2R_2$.

Lemma 6.3. *Suppose $Y = A^* + Z$ where $A^* \in \mathbb{R}^{n\times m}$ and $Z \sim \mathrm{subG}(\sigma^2)$. For $\tilde{A} \in \mathcal{U}^m$ and all $t > 0$, define*

$$f_{\tilde{A}}(t) = \sup_{A\in\mathcal{M}\cap\mathcal{B}^{nm}(\tilde{A},t)} \langle A-\tilde{A}, Y-\tilde{A}\rangle - \frac{t^2}{2}.$$

Then for any $s > 0$, it holds simultaneously for all $t > 0$ that

$$f_{\tilde{A}}(t) \le CR_1 t + t\|A^* - \tilde{A}\|_F - \frac{t^2}{2} + st \tag{6.5}$$

with probability at least $1 - C\exp(-\frac{cs^2}{\sigma^2})$, where C and c are positive constants.

Proof. Define $\Theta = \Theta_{\mathcal{M}}(\tilde{A},1) = \bigcup_{\lambda\ge 0}\{B - \lambda\tilde{A} : B \in \mathcal{M}\cap\mathcal{B}^{nm}(\lambda\tilde{A},1)\}$ (see also Definition (6.9)). In particular, $\Theta \subset \mathcal{B}^{nm}(0,1)$ and $0 \in \Theta$. Since $\mathcal{M}$ is a finite union of convex cones and thus is star-shaped, by scaling invariance,

$$\begin{aligned}\sup_{A\in\mathcal{M}\cap\mathcal{B}^{nm}(\tilde{A},t)} \langle A-\tilde{A}, Z\rangle &= t \sup_{B\in\mathcal{M}\cap\mathcal{B}^{nm}(t^{-1}\tilde{A},1)} \langle B - t^{-1}\tilde{A}, Z\rangle \\ &\le t\sup_{M\in\Theta}\langle M, Z\rangle.\end{aligned}$$

By Lemma 6.2, with probability at least $1 - C\exp(-\frac{cs^2}{\sigma^2})$,

$$\sup_{M\in\Theta}\langle M, Z\rangle \le C\sigma\int_0^2 \sqrt{\log N(\Theta, \|\cdot\|_F, \varepsilon)}\,d\varepsilon + s.$$

Moreover, it follows from Lemma 6.8 that

$$\log N(\Theta, \|\cdot\|_F, \varepsilon) \le C\varepsilon^{-1}K(\tilde{A})\log\frac{enm}{K(\tilde{A})} + n\log n.$$

Combining the previous three displays, we see that

$$\begin{aligned}&\sup_{A\in\mathcal{M}\cap\mathcal{B}^{nm}(\tilde{A},t)} \langle A-\tilde{A}, Z\rangle \\ &\quad\le C\sigma t\int_0^2\sqrt{C\varepsilon^{-1}K(\tilde{A})\log\frac{enm}{K(\tilde{A})} + n\log n}\,d\varepsilon + st \\ &\quad\le C\sigma t\sqrt{K(\tilde{A})\log\frac{enm}{K(\tilde{A})}} + C\sigma t\sqrt{n\log n} + st \\ &\quad= CR_1 t + st\end{aligned}$$

with probability at least $1 - C\exp(-\frac{cs^2}{\sigma^2})$. Therefore,

$$\begin{aligned} f_{\tilde{A}}(t) &= \sup_{A \in \mathcal{M} \cap \mathcal{B}^{nm}(\tilde{A},t)} \langle A - \tilde{A}, Y - \tilde{A}\rangle - \frac{t^2}{2} \\ &\leq \sup_{A \in \mathcal{M} \cap \mathcal{B}^{nm}(\tilde{A},t)} \langle A - \tilde{A}, Z\rangle + \sup_{A \in \mathcal{M} \cap \mathcal{B}^{nm}(\tilde{A},t)} \big\langle A - \tilde{A}, A^* - \tilde{A}\big\rangle - \frac{t^2}{2} \\ &\leq CR_1 t + st + t\big\|A^* - \tilde{A}\big\|_F - \frac{t^2}{2} \end{aligned}$$

with probability at least $1 - C\exp(-\frac{cs^2}{\sigma^2})$ simultaneously for all $t > 0$. □

We are now in a position to prove the adaptive oracle inequalities in Theorem 3.1. Recall that $(\hat{\Pi}, \hat{A})$ denotes the LS estimator defined in (2.2). Without loss of generality, assume that $\Pi^* = I_n$ and $Y = A^* + Z$.

Fix $\tilde{A} \in \mathcal{U}^m$ and define $f_{\tilde{A}}$ as in Lemma 6.3. We can apply Lemma 6.1 with $a^* = \tilde{A}$, $z = Y - \tilde{A}$, $y = Y$ and $\hat{a} = \hat{\Pi}\hat{A}$ to achieve an error bound on $\|\hat{\Pi}\hat{A} - \tilde{A}\|_F$, since $\hat{\Pi}\hat{A} \in \operatorname{argmin}_{M \in \mathcal{M}} \|Y - M\|_F^2$. To be more precise, for any $s > 0$ we define $t^* = 3C_1R_1 + 2\|A^* - \tilde{A}\|_F + 2s$ where C_1 is the constant in (6.5). Then it follows from Lemma 6.3 that with probability at least $1 - C\exp(-\frac{cs^2}{\sigma^2})$, it holds for all $t \geq t^*$ that

$$f_{\tilde{A}}(t) \leq C_1R_1t + t\big\|A^* - \tilde{A}\big\|_F - \frac{t^2}{2} + st < 0.$$

Therefore by Lemma 6.1,

$$\|\hat{\Pi}\hat{A} - \tilde{A}\|_F \leq t^* = 3C_1R_1 + 2\big\|A^* - \tilde{A}\big\|_F + 2s,$$

and thus

$$\big\|\hat{\Pi}\hat{A} - A^*\big\|_F \leq CR_1 + 3\big\|A^* - \tilde{A}\big\|_F + 2s \tag{6.6}$$

with probability at least $1 - C\exp(-\frac{cs^2}{\sigma^2})$.

In particular, if $s = R_1$, then $s \geq \sigma\sqrt{n+m}$ as $K(\tilde{A}) \geq m$. We see that with probability at least $1 - C\exp(-\frac{cs^2}{\sigma^2}) \geq 1 - e^{-c(n+m)}$,

$$\big\|\hat{\Pi}\hat{A} - A^*\big\|_F \lesssim R_1 + \big\|A^* - \tilde{A}\big\|_F$$

and thus

$$\big\|\hat{\Pi}\hat{A} - A^*\big\|_F^2 \lesssim \big\|A^* - \tilde{A}\big\|_F^2 + \sigma^2 K(\tilde{A}) \log\frac{enm}{K(\tilde{A})} + \sigma^2 n \log n.$$

Finally, (3.1) follows by taking the infimum over $\tilde{A} \in \mathcal{U}^m$ on the right-hand side and dividing both sides by nm.

Next, to prove the bound in expectation, observe that (6.6) yields

$$\mathbb{P}\big[\|\hat{\Pi}\hat{A}-A^*\|_F^2 - C\big(R_2+\|A^*-\tilde{A}\|_F^2\big)\ge s\big]\le C\exp\Big(-\frac{cs}{\sigma^2}\Big),$$

where R_2 is defined in (6.4). Integrating the tail probability, we get that

$$\mathbb{E}\|\hat{\Pi}\hat{A}-A^*\|_F^2 - C\big(R_2+\|A^*-\tilde{A}\|_F^2\big)\lesssim\int_0^\infty\exp\Big(-\frac{cs}{\sigma^2}\Big)\,ds=\frac{\sigma^2}{c}$$

and therefore

$$\mathbb{E}\|\hat{\Pi}\hat{A}-A^*\|_F^2\lesssim R_2+\|A^*-\tilde{A}\|_F^2.$$

Dividing both sides by nm and minimizing over $\tilde{A}\in\mathcal{U}^m$ yields (3.2).

6.1.3. *Proof of Theorem* 3.3

In the setting of isotonic regression, [8] derived global bounds from adaptive bounds by a block approximation method, which also applies to our setting. The lemma below is a generalization of [8], Lemma 2, to the case of unimodal matrices.

For $k\in[n]$, let

$$\mathcal{U}_k=\big\{a\in\mathcal{U}:\mathsf{card}\big(\{a_1,\dots,a_n\}\big)\le k\big\}.$$

Define $k^*=\lceil(\frac{V(a)^2n}{\sigma^2\log(en)})^{1/3}\rceil$. More generally, for $\mathbf{k}\in[n]^m$, we write $\mathbf{k}=(k_1,\dots,k_m)$ and let

$$\mathcal{U}^m_{\mathbf{k}}=\big\{A\in\mathcal{U}^m:\mathsf{card}\big(\{A_{1,j},\dots,A_{n,j}\}\big)=k_j\text{ for }1\le j\le m\big\}.$$

Then $K(A)=\sum_{j=1}^m k_j$ for $A\in\mathcal{U}^m_{\mathbf{k}}$. Define $\mathbf{k}^*$ by

$$k_j^*=\bigg\lceil\bigg(\frac{V(A_{\cdot,j})^2n}{\sigma^2\log(en)}\bigg)^{1/3}\bigg\rceil.$$

Lemma 6.4. *For $A\in\mathcal{U}^m$, there exists $\tilde{A}\in\mathcal{U}^m_{\mathbf{k}^*}$ such that*

$$\frac{1}{nm}\|\tilde{A}-A\|_F^2\le\frac{1}{4}\bigg(\frac{\sigma^2V(A)\log(en)}{n}\bigg)^{2/3}+\frac{\sigma^2}{4n}\log(en)$$

and

$$\frac{\sigma^2K(\tilde{A})}{nm}\log(en)\le2\bigg(\frac{\sigma^2V(A)\log(en)}{n}\bigg)^{2/3}+\frac{2\sigma^2}{n}\log(en).$$

The proof of the lemma is provided in the supplement [34]. To prove the theorem, for $A \in \mathcal{U}^m$, choose $\tilde{A} \in \mathcal{U}^m_{\mathbf{k}^*}$ according to Lemma 6.4. Then

$$\begin{aligned}\frac{1}{nm}\|\tilde{A} - A^*\|_F^2 &\le \frac{2}{nm}\|A - A^*\|_F^2 + \frac{2}{nm}\|\tilde{A} - A\|_F^2 \\ &\le \frac{2}{nm}\|A - A^*\|_F^2 + \frac{5}{4}\left(\frac{\sigma^2 V(A)\log n}{n}\right)^{2/3} + \frac{5\sigma^2}{4n}\log n \qquad (6.7)\end{aligned}$$

by noting that $\log(en) \le 2.5\log n$ for $n \ge 2$, and similarly

$$\frac{\sigma^2 K(\tilde{A})}{nm}\log(en) \le 5\left(\frac{\sigma^2 V(A)\log n}{n}\right)^{2/3} + \frac{5\sigma^2}{n}\log n. \qquad (6.8)$$

Plugging (6.7) and (6.8) into the right-hand side of (3.1) and (3.2), and then minimizing over $A \in \mathcal{U}^m$, we complete the proof.

6.2. Metric entropy

This section is devoted to studying various *covering numbers* or *metric entropy* related to the parameter space of the model (2.1). The proofs of the lemmas in this section are provided in the supplementary material [34].

Recall that an ε-net of a subset $G \subset \mathbb{R}^n$ with respect to a norm $\|\cdot\|$ is a set $\{w_1, \dots, w_N\} \subset G$ such that for any $w \in G$, there exists $i \in [N]$ for which $\|w - w_i\| \le \varepsilon$. The covering number $N(G, \|\cdot\|, \varepsilon)$ is the cardinality of the smallest ε-net with respect to the norm $\|\cdot\|$. Metric entropy is defined as the logarithm of a covering number. In the following, we will consider the Euclidean norm unless otherwise specified.

We start with a lemma bounding the metric entropy of a Cartesian product of convex cones. It is useful in later proofs and has its own interest. Let $\{I_i\}_{i=1}^m$ be a partition of $[n]$ with $|I_i| = n_i$ and $\sum_{i=1}^m n_i = n$. For $a \in \mathbb{R}^n$, the restriction of a to the coordinates in I_i is denoted by $a_{I_i} \in \mathbb{R}^{n_i}$. Let $\mathcal{C}_i$ be a convex cone in $\mathbb{R}^{n_i}$ and $\mathcal{C} = \mathcal{C}_1 \times \cdots \times \mathcal{C}_m$.

Lemma 6.5. *With the notation above, suppose that $a_{I_i} \in \mathcal{C}_i \cap (-\mathcal{C}_i)$. Then for any $t > 0$ and $\varepsilon \in (0, t]$,*

$$\log N\big(\mathcal{C} \cap \mathcal{B}^n(a, t), \|\cdot\|_2, \varepsilon\big) \le m \log\frac{Ct}{\varepsilon} + \sum_{i=1}^m \log N\left(\mathcal{C}_i \cap \mathcal{B}^{n_i}(a_{I_i}, t), \|\cdot\|_2, \frac{\varepsilon}{3}\right)$$

for some constant $C > 0$.

Recall that $\mathcal{S}_n$ denotes the closed convex cone of increasing vectors in $\mathbb{R}^n$. First, we give a result on the metric entropy of $\mathcal{S}_n$ intersecting with a ball.

Lemma 6.6. *Let $b \in \mathbb{R}^n$ be such that $b_1 = \cdots = b_n$. Then for any $t > 0$ and $\varepsilon > 0$,*

$$\log N\big(\mathcal{S}_n \cap \mathcal{B}^n(b, t), \|\cdot\|_2, \varepsilon\big) \le C\varepsilon^{-1} t \log(en).$$

Next, we study the metric entropy of the set of matrices with unimodal columns. Recall that $\mathcal{C}_l = \{a \in \mathbb{R}^n : a_1 \le \cdots \le a_l\} \cap \{a \in \mathbb{R}^n : a_l \ge \cdots \ge a_n\}$ for $l \in [n]$. For $\mathbf{l} = (l_1, \dots, l_m) \in [n]^m$, define $\mathcal{C}_{\mathbf{l}}^m = \mathcal{C}_{l_1} \times \cdots \times \mathcal{C}_{l_m}$. Moreover, for $A \in \mathbb{R}^{n\times m}$, $t > 0$ and $\mathcal{C} \subset \mathbb{R}^{n\times m}$, define

$$\begin{aligned}\Theta_{\mathcal{C}}(A,t) &= \bigcup_{\lambda\ge 0}\{B - \lambda A : B \in \mathcal{C} \cap \mathcal{B}^{nm}(\lambda A, t)\}\\ &= \bigcup_{\lambda\ge 0}\big(\mathcal{C} \cap \mathcal{B}^{nm}(\lambda A, t) - \lambda A\big).\end{aligned} \tag{6.9}$$

Note that in particular $\Theta_{\mathcal{C}}(A,t) \subset \mathcal{B}^{nm}(0,t)$.

Lemma 6.7. *Given $A \in \mathbb{R}^{n\times m}$ and $\mathbf{l} = (l_1, \dots, l_m) \in [n]^m$, we define the quantities $k(A_{\cdot,j}) =$ $\mathsf{card}(\{A_{1,j}, \dots, A_{n,j}\})$ and $K(A) = \sum_{j=1}^m k(A_{\cdot,j})$. Then for any $t > 0$ and $\varepsilon > 0$,*

$$\log N\big(\Theta_{\mathcal{C}_{\mathbf{l}}^m}(A,t), \|\cdot\|_F, \varepsilon\big) \le C\varepsilon^{-1} t K(A) \log \frac{enm}{K(A)}.$$

Finally, we consider the metric entropy of $\Theta_{\mathcal{M}}(A,t)$ for $A \in \mathbb{R}^{n\times m}$, $t > 0$ and $\mathcal{M} = \bigcup_{\Pi\in\mathfrak{S}_n} \Pi\mathcal{U}^m$. The above analysis culminates in the following lemma which we use to prove the main upper bounds.

Lemma 6.8. *Let $A \in \mathbb{R}^{n\times m}$ and $K(A)$ be defined as in the previous lemma. Then for any $\varepsilon > 0$ and $t > 0$,*

$$\log N\big(\Theta_{\mathcal{M}}(A,t), \|\cdot\|_F, \varepsilon\big) \le C\varepsilon^{-1} t K(A) \log \frac{enm}{K(A)} + n\log n.$$

6.3. Proof of the lower bounds

For minimax lower bounds, we consider the model $Y = \Pi^* A^* + Z$ where entries of Z are i.i.d. $N(0,\sigma^2)$. Define $\mathcal{U}_{K_0}^m(V_0) = \mathcal{U}_{K_0}^m \cap \mathcal{U}^m(V_0)$ and $\mathcal{M}_{K_0}(V_0) = \bigcup_{\Pi\in\mathfrak{S}_n} \Pi\mathcal{U}_{K_0}^m(V_0)$. Define the subset of $\mathcal{M}_{K_0}(V_0)$ containing permutations of monotone matrices by $\mathcal{M}_{K_0}^{\mathcal{S}}(V_0) = \{\Pi A \in \mathcal{M}_{K_0}(V_0) : \Pi \in \mathfrak{S}_n, A \in \mathcal{S}^m\}$. Since each estimator pair $(\hat\Pi, \hat A)$ gives an estimator $\hat M = \hat\Pi\hat A$ of $M = \Pi A$, it suffices to prove a lower bound on $\|\hat M - M\|_F^2$. In fact, we prove a lower bound stronger than the one in Theorem 3.5. The proofs of the lemmas below can be found in the supplement [34].

Proposition 6.9. *Suppose that $K_0 \le m(\frac{16n}{\sigma^2})^{1/3} V_0^{2/3} - m$. Then*

$$\begin{aligned}\inf_{\hat M}\sup_{M\in\mathcal{M}_{K_0}(V_0)} \mathbb{P}_M\bigg[\frac{1}{nm}\|\hat M - M\|_F^2 \ge c\sigma^2\frac{K_0}{nm}\\ + c\max_{1\le l\le\min(K_0-m,m)+1}\min\bigg(\frac{\sigma^2}{m}\log l, m^2 l^{-3} V_0^2\bigg)\bigg] \ge c'\end{aligned} \tag{6.10}$$

for some $c, c' > 0$, where $\mathbb{P}_M$ is the probability with respect to $Y = M + Z$. This bound remains valid for the parameter subset $\mathcal{M}_{K_0}^{S}(V_0)$ if $l = 1$ or 2.

Note that the bound also holds for the larger parameter space $\mathcal{M}_{K_0} = \bigcup_{\Pi \in \mathfrak{S}_n} \Pi \mathcal{U}_{K_0}^{m}$. By taking $l = \min(K_0 - m, m) + 1$ and V_0 large enough, we see that the assumption in Proposition 6.9 is satisfied and the second term becomes simply $\frac{\sigma^2}{m} \log l$, so Theorem 3.5 follows. In the monotonic case, by the last statement of the proposition, if $K_0 \geq m + 1$ then taking $l = 2$ and V_0 large enough yields a lower bound of rate $\sigma^2(\frac{K_0}{nm} + \frac{1}{m})$ for the set of matrices A with increasing columns and $K(A) \leq K_0$.

The proof of Proposition 6.9 has two parts which correspond to the two terms respectively. First, the term $\sigma^2 \frac{K_0}{nm}$ is derived from the proof of lower bounds for isotonic regression in [8]. Then we derive the other term $\frac{\sigma^2}{m} \log l$ for any $1 \leq l \leq \min(K_0 - m, m) + 1$, which is due to the unknown permutation.

Lemma 6.10. *Suppose that $K_0 \leq m(\frac{16n}{\sigma^2})^{1/3} V_0^{2/3} - m$. For some $c, c' > 0$,*

$$\inf_{\hat{M}} \sup_{M \in \mathcal{M}_{K_0}^{S}(V_0)} \mathbb{P}_M\big[\|\hat{M} - M\|_F^2 \geq c\sigma^2 K_0\big] \geq c,$$

where $\mathbb{P}_M$ is the probability with respect to $Y = M + Z$.

For the second term in (6.10), we first note that the bound is trivial for $l = 1$ since $\log l = 0$. The next lemma deals with the case $l = 2$.

Lemma 6.11. *There exist constants $c, c' > 0$ such that for any $K_0 \geq m + 1$ and $V_0 \geq 0$,*

$$\inf_{\hat{M}} \sup_{M \in \mathcal{M}_{K_0}^{S}(V_0)} \mathbb{P}_M\big[\|\hat{M} - M\|_F^2 \geq cn \min\big(\sigma^2, m^3 V_0^2\big)\big] \geq c',$$

where $\mathbb{P}_M$ is the probability with respect to $Y = M + Z$.

For the previous two lemmas, we have only used matrices with increasing columns. However, to achieve the second term in (6.10) for $l \geq 3$, we need matrices with unimodal columns.

Lemma 6.12. *There exist constants $c, c' > 0$ such that for any $K_0 \geq m$, $V_0 \geq 0$ and $3 \leq l \leq \min(K_0 - m, m) + 1$,*

$$\inf_{\hat{M}} \sup_{M \in \mathcal{M}_{K_0}(V_0)} \mathbb{P}_M\big[\|\hat{M} - M\|_F^2 \geq cn \min\big(\sigma^2 \log l, m^3 l^{-3} V_0^2\big)\big] \geq c',$$

where $\mathbb{P}_M$ is the probability with respect to $Y = M + Z$.

Proof of Proposition 6.9. Combining Lemmas 6.10, 6.11 and 6.12, and then dividing the bound by nm, we get (6.10) because the max of two terms is lower bounded by a half of their sum. The last statement in Proposition 6.9 holds since Lemmas 6.10 and 6.11 are proved for matrices with increasing columns. □

Furthermore, the proof of Theorem 3.6, provided in the supplement [34], only uses Lemmas 6.10 and 6.11, so the lower bound of rate $(\frac{\sigma^2 V_0}{n})^{2/3} + \frac{\sigma^2}{n} + \min(\frac{\sigma^2}{m}, m^2 V_0^2)$ holds even if the matrices are required to have increasing columns.

6.4. Matrices with increasing columns

For the model $Y = \Pi^* A^* + Z$ where $A^* \in \mathcal{S}^m$ and $Z \sim \mathrm{subG}(\sigma^2)$, a computationally efficient estimator $(\tilde{\Pi}, \tilde{A})$ has been constructed in Section 4 using the RankScore procedure. We will bound its rate of estimation in this section. Recall that the definition of $(\tilde{\Pi}, \tilde{A})$ consists of two steps. First, we recover an order (or a ranking) of the rows of Y, which leads to an estimator $\tilde{\Pi}$ of the permutation. Then define $\tilde{A} \in \mathcal{S}^m$ so that $\tilde{\Pi}\tilde{A}$ is the projection of Y onto the convex cone $\tilde{\Pi}\mathcal{S}^m$. For the analysis of the algorithm, we deal with the projection step first, and then turn to learning the permutation. The proofs of the results in the section can be found in the supplement [34].

In fact, for *any* estimator $\tilde{\Pi}$, if $\tilde{A}$ is defined as above by the projection corresponding to $\tilde{\Pi}$, then the error $\|\tilde{\Pi}\tilde{A} - \Pi^* A^*\|_F^2$ can be split into two parts: the permutation error $\|(\tilde{\Pi} - \Pi^*)A^*\|_F^2$ and the estimation error of order $\tilde{O}(\sigma^2 K(A^*))$.

Lemma 6.13. *Consider the model $Y = \Pi^* A^* + Z$ where $A^* \in \mathcal{S}^m$ and $Z \sim \mathrm{subG}(\sigma^2)$. For any $\tilde{\Pi} \in \mathfrak{S}_n$, define $\tilde{A} \in \mathcal{S}^m$ so that $\tilde{\Pi}\tilde{A}$ is the projection of Y onto $\tilde{\Pi}\mathcal{S}^m$. Then with probability at least $1 - e^{-c(n+m)}$, it holds simultaneously for all $\tilde{\Pi} \in \mathfrak{S}_n$ that*

$$\|\tilde{\Pi}\tilde{A} - \Pi^* A^*\|_F^2 \lesssim \min_{A \in \mathcal{S}^m}\left(\|A - A^*\|_F^2 + \sigma^2 K(A)\log\frac{enm}{K(A)}\right) + \sigma^2 n\log n + \|(\tilde{\Pi} - \Pi^*)A^*\|_F^2.$$

The idea of splitting the error into two terms as in Lemma 6.13 has appeared in [23,61].

By virtue of Lemma 6.13, it remains to control the permutation error $\|\tilde{\Pi}A^* - \Pi^* A^*\|_F^2$ where $\tilde{\Pi}$ is given by the RankScore procedure defined in Section 4. Recall that

$$\Delta_{A^*}(i, i') = \max_{j \in [m]}(A^*_{i',j} - A^*_{i,j}) \vee \frac{1}{\sqrt{m}}\sum_{j=1}^{m}(A^*_{i',j} - A^*_{i,j})$$

for $i, i' \in [n]$ and $\Delta_Y(i, i')$ is defined analogously. Since columns of A^* are increasing,

$$|\Delta_{A^*}(i, i')| = \|A^*_{i',\cdot} - A^*_{i,\cdot}\|_\infty \vee \frac{1}{\sqrt{m}}\|A^*_{i',\cdot} - A^*_{i,\cdot}\|_1. \tag{6.11}$$

Recall that the RankScore procedure is defined as follows. First, for $i \in [n]$, we associate with the ith row of Y a score s_i defined by $s_i = \sum_{l=1}^n \mathbb{I}(\Delta_Y(l, i) \ge 2\tau)$ for the threshold $\tau :=$

$3\sigma\sqrt{\log(nm\delta^{-1})}$ where δ is the probability of failure. Then we order the rows of Y so that the scores are increasing with ties broken arbitrarily. This is equivalent to requiring that the corresponding permutation $\tilde{\pi} : [n] \to [n]$ satisfies that if $s_i < s_{i'}$ then $\tilde{\pi}^{-1}(i) < \tilde{\pi}^{-1}(i')$. Define $\tilde{\Pi}$ to be the $n \times n$ permutation matrix corresponding to $\tilde{\pi}$ so that $\tilde{\Pi}_{\tilde{\pi}(i),i} = 1$ for $i \in [n]$ and all other entries of $\tilde{\Pi}$ are zero. Moreover, let $\pi^* : [n] \to [n]$ be the permutation corresponding to Π^*.

To control the permutation error, we first state a lemma which asserts that if the gap between two rows of A^* is sufficiently large, then the permutation defined above will recover their relative order with high probability.

Lemma 6.14. *There is an event $\mathcal{E}$ of probability at least $1 - \delta$ on which the following holds. For any $i, i' \in [n]$, if $\Delta_{A^*}(i, i') \geq 4\tau$, then $\tilde{\pi}^{-1} \circ \pi^*(i) < \tilde{\pi}^{-1} \circ \pi^*(i')$.*

Equipped with the above lemma, we are able to bound the permutation error in terms of the quantity $R(A^*)$ defined in (4.3).

Lemma 6.15. *There is an event $\mathcal{E}$ of probability at least $1 - \delta$ on which*

$$\|\tilde{\Pi}A^* - \Pi^* A^*\|_F^2 \lesssim \sigma^2 R(A^*) n \log(nm\delta^{-1}).$$

Finally, the bound of Theorem 4.1 is an immediate consequence of Lemma 6.13 and Lemma 6.15 with $\delta = (nm)^{-C}$ for $C > 0$.

7. Discussion

While computational aspects of the seriation problem have received significant attention, the robustness of this problem to noise was still unknown to date. To overcome this limitation, we have introduced in this paper the statistical seriation model and studied optimal rates of estimation by showing, in particular, that the least squares estimator enjoys several desirable statistical properties such as adaptivity and minimax optimality (up to logarithmic terms).

While this work paints a fairly complete statistical picture of the statistical seriation model, it also leaves many unanswered questions. There are several logarithmic gaps in the bounds. In the case of adaptive bounds, some logarithmic terms are unavoidable as illustrated by Theorem 3.5 (for the permutation term) and also by statistical dimension consideration explained in [9] (for the estimation term). However, a more refined argument for the uniform bound, namely one that uses covering in ℓ_2-norm rather than ℓ_∞-norm, would allow us to remove the $\log n$ factor from the estimation term in the upper bound of Corollary 3.4. Such an argument can be found in [3,14,70] for the larger class of vectors with bounded total variation (see [54]) but we do not pursue sharp logarithmic terms in this work. For the permutation term, $\log n$ in the upper bound of Corollary 3.2 and $\log l$ in the lower bound of Theorem 3.5 do not match if $l < n$. We do not seek answers to these questions in this paper but note that their answers may be different for the unimodal and the monotone case.

Perhaps the most pressing question is that of computationally efficient estimators. Indeed, while statistically optimal, the least squares estimator requires searching through $n!$ permutations, which is not realistic even for problems of moderate size, let alone genomics applications. We gave a partial answer to this question in the specific context of monotone columns by proposing and studying the performance of a simple and efficient estimator called RankScore. This study reveals the existence of a potentially intrinsic gap between the statistical performance achievable by efficient estimators and that achievable by estimators with access to unbounded computation. A similar gap is also observed in the SST model for pairwise comparisons [61]. We conjecture that achieving optimal rates of estimation in the seriation model is computationally hard in general but argue that the planted clique assumption that has been successfully used to establish statistical vs. computational gaps in [11,52,62] for example, is not the correct primitive. Instead, one has to seek for a primitive where hardness comes from searching through permutations rather than subsets.

Acknowledgments

P.R. is partially supported by NSF CAREER DMS-1541099, NSF DMS-1541100, NSF DMS-1712596, DARPA W911NF-16-1-0551, ONR N00014-17-1-2147 and a grant from the MIT NEC Corporation. N.F. gratefully acknowledges partial support by NSF DMS-1541100, the Institute for Data Systems and Society and the Mathematics Department during his visit at MIT. C.M. acknowledges partial support by NSF CAREER DMS-1541099. P.R. also thanks Enno Mammen for his help in retracing the literature on sharp entropy bounds for monotone classes and Ramon van Handel for a stimulating discussion on entropy numbers. We thank Alexandre Tsybakov for bringing concurrent work on unimodal regression by Pierre C. Bellec to their attention. We thank Pierre C. Bellec for pointing out an error in an earlier version of Lemma 6.1 and suggesting that it holds for all closed sets. Finally, we thank an anonymous reviewer for providing helpful comments.

Supplementary Material

Supplement to "Optimal Rates of Statistical Seriation" (DOI: 10.3150/17-BEJ1000SUPP; .pdf). We include additional technical details in this supplement.

References

[1] Amelunxen, D., Lotz, M., McCoy, M.B. and Tropp, J.A. (2014). Living on the edge: Phase transitions in convex programs with random data. *Inf. Inference* **3** 224–294. MR3311453

[2] Annexstein, F. and Swaminathan, R. (1998). On testing consecutive-ones property in parallel. *Discrete Appl. Math.* **88** 7–28. MR1658548

[3] Anuchina, N.N., Babenko, K.I., Godunov, S.K., Dmitriev, N.A., Dmitrieva, L.V., D'yachenko, V.F., Zabrodin, A.V., Lokutsievskiĭ, O.V., Malinovskaya, E.V., Podlivaev, I.F., Prokopov, G.P., Sofronov, I.D. and Fedorenko, R.P. (1979). *Teoreticheskie Osnovy i Konstruirovanie Chislennykh Algoritmov Zadach Matematicheskoĭ Fiziki*. Moscow: "Nauka". MR0568904

[4] Arabie, P., Schleutermann, S., Daws, J. and Hubert, L. (1988). Marketing applications of sequencing and partitioning of nonsymmetric and/or two-mode matrices. In *Data*, *Expert Knowledge and Decisions*: *An Interdisciplinary Approach with Emphasis on Marketing Applications* (W. Gaul and M. Schader, eds.) 215–224. Berlin, Heidelberg: Springer.

[5] Atkins, J.E., Boman, E.G. and Hendrickson, B. (1999). A spectral algorithm for seriation and the consecutive ones problem. *SIAM J*. *Comput*. **28** 297–310. MR1630473

[6] Ayer, M., Brunk, H.D., Ewing, G.M., Reid, W.T. and Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *Ann*. *Math*. *Stat*. **26** 641–647. MR0073895

[7] Barlow, R.E., Bartholomew, D.J., Bremner, J.M. and Brunk, H.D. (1972). *Statistical Inference Under Order Restrictions*. *The Theory and Application of Isotonic Regression*. London–New York–Sydney: Wiley. MR0326887

[8] Bellec, P. and Tsybakov, A.B. (2015). Sharp oracle bounds for monotone and convex regression through aggregation. *J*. *Mach*. *Learn*. *Res*. **16** 1879–1892.

[9] Bellec, P.C. (2015). Sharp oracle inequalities for least squares estimators in shape restricted regression. Preprint. Available at ArXiv:1510.08029.

[10] Bellec, P.C. (2016). Private communication.

[11] Berthet, Q. and Rigollet, P. (2013). Complexity theoretic lower bounds for sparse principal component detection. In *COLT* 2013 *– The* 26*th Conference on Learning Theory*, *Princeton*, *NJ*, *June* 12–14, 2013 (S. Shalev-Shwartz and I. Steinwart, eds.). *JMLR W&CP* **30** 1046–1066.

[12] Bickel, P.J. and Fan, J. (1996). Some problems on the estimation of unimodal densities. *Statist*. *Sinica* **6**.

[13] Birgé, L. (1997). Estimation of unimodal densities without smoothness assumptions. *Ann*. *Statist*. **25**.

[14] Birman, M.Š. and Solomjak, M.Z. (1967). Piecewise polynomial approximations of functions of classes W_p^α. *Mat*. *Sb*. **73 (115)** 331–355. MR0217487

[15] Boyarshinov, V. and Magdon-Ismail, M. (2006). Linear time isotonic and unimodal regression in the L_1 and L_∞ norms. *J*. *Discrete Algorithms* **4** 676–691. MR2577688

[16] Bro, R. and Sidiropoulos, N. (1998). Least squares algorithms under unimodality and non-negativity constraints. *J*. *Chemom*. **12** 223–247.

[17] Chandrasekaran, V., Recht, B., Parrilo, P.A. and Willsky, A.S. (2012). The convex geometry of linear inverse problems. *Found*. *Comput*. *Math*. **12** 805–849. MR2989474

[18] Chatterjee, S. (2014). A new perspective on least squares under convex constraint. *Ann*. *Statist*. **42** 2340–2381. MR3269982

[19] Chatterjee, S. (2015). Matrix estimation by universal singular value thresholding. *Ann*. *Statist*. **43** 177–214. MR3285604

[20] Chatterjee, S., Guntuboyina, A. and Sen, B. (2015). On risk bounds in isotonic and other shape restricted regression problems. *Ann*. *Statist*. **43** 1774–1800.

[21] Chatterjee, S., Guntuboyina, A. and Sen, B. (2018). On matrix estimation under monotonicity constraints. *Bernoulli* **24** 1072–1100.

[22] Chatterjee, S. and Lafferty, J. (2015). Adaptive risk bounds in unimodal regression. Preprint. Available at ArXiv:1512.02956.

[23] Chatterjee, S. and Mukherjee, S. (2016). On estimation in tournaments and graphs under monotonicity constraints. Preprint. Available at ArXiv:1603.04556.

[24] Collier, O. and Dalalyan, A.S. (2016). Minimax rates in permutation estimation for feature matching. *J*. *Mach*. *Learn*. *Res*. **17** 1–32.

[25] Copeland, A.H. (1951). A reasonable social welfare function. In *Mimeographed Notes from a Seminar on Applications of Mathematics to the Social Sciences*, *University of Michigan*.

[26] Czekanowski, J. (1909). Zur differential Diagnose der Neandertalgruppe. Korrespondenzblatt der deutschen Gesellschaft für Anthropologie. *Ethnologie und Urgeschichte* **40** 44–47.

[27] Daskalakis, C., Diakonikolas, I. and Servedio, R.A. (2012). Learning k-modal distributions via testing. In *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms* 1371–1382. ACM, New York. MR3205298

[28] Daskalakis, C., Diakonikolas, I., Servedio, R.A., Valiant, G. and Valiant, P. (2012). Testing k-modal distributions: Optimal algorithms via reductions. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms* 1833–1852. SIAM, Philadelphia, PA. MR3221282

[29] Davidson, D. and Marschak, J. (1959). Experimental tests of a stochastic decision theory. *Measurement*: *Definitions and Theories*.

[30] Donoho, D.L. (1990). Gel'fand n-widths and the method of least squares Statistics Technical Report No. 282, Univ. California, Berkeley.

[31] Eggermont, P.P.B. and LaRiccia, V.N. (2000). Maximum likelihood estimation of smooth monotone and unimodal densities. *Ann*. *Statist*. **28**.

[32] Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc*. *Natl*. *Acad*. *Sci*. *USA* **95** 14863–14868.

[33] Fishburn, P.C. (1973). Binary choice probabilities: On the varieties of stochastic transitivity. *J*. *Math*. *Psych*. **10**.

[34] Flammarion, N., Mao, C. and Rigollet, P. (2017). Supplement to "Optimal rates of statistical seriation." DOI:10.3150/17-BEJ1000SUPP.

[35] Fogel, F., Jenatton, R., Bach, F. and d'Aspremont, A. (2013). Convex Relaxations for Permutation Problems. In *Advances in Neural Information Processing Systems* 26 (C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger, eds.) 1016–1024. Curran Associates, Inc.

[36] Forsyth, E. and Katz, L. (1946). A matrix approach to the analysis of sociometric data: Preliminary report. *Sociometry* **9** 340–347.

[37] Frisen, M. (1986). Unimodal regression. *J*. *R*. *Stat*. *Soc*., *Ser*. *D Stat*. **35** 479–485.

[38] Fulkerson, D.R. and Gross, O.A. (1964). Incidence matrices with the consecutive 1's property. *Bull*. *Amer*. *Math*. *Soc*. **70** 681–684.

[39] Gao, C., Lu, Y. and Zhou, H.H. (2015). Rate-optimal graphon estimation. *Ann*. *Statist*. **43** 2624–2652. MR3405606

[40] Geng, Z. and Shi, N.Z. (1990). Algorithm AS 257: Isotonic regression for umbrella orderings. *J*. *R*. *Stat*. *Soc*. *Ser*. *C*. *Appl*. *Stat*. **39** 397–402.

[41] Gertzen, T.L. and Grötschel, M. (2012). Flinders Petrie, the travelling salesman problem, and the beginning of mathematical modeling in archaeology. *Doc*. *Math*. **X** 199–210.

[42] Hartigan, J.A. (1972). Direct clustering of a data matrix. *J*. *Amer*. *Statist*. *Assoc*. **67** 123–129.

[43] Kendall, D.G. (1963). A statistical approach to Flinders Petrie's sequence-dating. *Bull*. *Inst*. *Int*. *Stat*. **40** 657–681. MR0175234

[44] Kendall, D.G. (1969). Incidence matrices, interval graphs and seriation in archeology. *Pacific J*. *Math*. **28** 565–570. MR0239990

[45] Kendall, D.G. (1970). A mathematical approach to seriation. *Philos*. *Trans*. *R*. *Soc*. *Lond*. *Ser*. *A*, *Math*. *Phys*. *Sci*. **269** 125–134.

[46] Kendall, D.G. (1971). Abundance matrices and seriation in archaeology. *Z*. *Wahrsch*. *Verw*. *Gebiete* **17** 104–112. MR0289329

[47] Köllmann, C., Bornkamp, B. and Ickstadt, K. (2014). Unimodal regression using Bernstein–Schoenberg splines and penalties. *Biometrics* **70** 783–793. MR3295739

[48] Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces*. *Ergebnisse der Mathematik und Ihrer Grenzgebiete* (3) [*Results in Mathematics and Related Areas* (3)] **23**. Berlin: Springer.

[49] Liiv, I. (2010). Seriation and matrix reordering methods: An historical overview. *Stat*. *Anal*. *Data Min*. **3** 70–91. MR2608807

[50] Lim, C.H. and Wright, S. (2014). Beyond the Birkhoff polytope: Convex relaxations for vector permutation problems. In *Advances in Neural Information Processing Systems* 27 (Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence and K.Q. Weinberger, eds.) 2168–2176. Curran Associates, Inc.

[51] Loiola, E.M., Maia de Abreu, N.M., Boaventura-Netto, P.O., Hahn, P. and Querido, T. (2007). A survey for the quadratic assignment problem. *European J. Oper. Res.* **176** 657–690. MR2267435

[52] Ma, Z. and Wu, Y. (2015). Computational barriers in minimax submatrix detection. *Ann. Statist.* **43** 1089–1116.

[53] Mammen, E. (1991). Estimating a smooth monotone regression function. *Ann. Statist.* **19** 724–740.

[54] Mammen, E. and van de Geer, S. (1997). Locally adaptive regression splines. *Ann. Statist.* **25** 387–413.

[55] Mendelson, S. (2015). Learning without concentration. *J. ACM* **62**.

[56] Murtagh, F. (1989). Review of Book Data, Expert Knowledge and Decisions, W. Gaul and M. Schader (eds.), Springer-Verlag, 1988. *J. Classification* **6** 129–132.

[57] Nemirovskiĭ, A.S., Polyak, B.T. and Tsybakov, A.B. (1985). The rate of convergence of nonparametric estimates of maximum likelihood type. *Problemy Peredachi Informatsii* **21** 17–33.

[58] Petrie, W.M.F. (1899). Sequences in prehistoric remains. *J. Anthropol. Inst. G.B. Irel.* **29** 295–301.

[59] Robertson, T., Wright, F.T. and Dykstra, R. (1988). *Order Restricted Statistical Inference*. *Probability and Statistics Series*. New York: Wiley.

[60] Robinson, W.S. (1951). A method for chronologically ordering archaeological deposits. *Am. Antiq.* **16** 293–301.

[61] Shah, N.B., Balakrishnan, S., Guntuboyina, A. and Wainwright, M.J. (2017). Stochastically transitive models for pairwise comparisons: Statistical and computational issues. *IEEE Trans. Inform. Theory* **63** 934–959. MR3604649

[62] Shah, N.B., Balakrishnan, S. and Wainwright, M.J. (2016). Feeling the Bern: Adaptive estimators for Bernoulli probabilities of pairwise comparisons. Preprint. Available at ArXiv:1603.06881.

[63] Shoung, J.M. and Zhang, C.H. (2001). Least squares estimators of the mode of a unimodal regression function. *Ann. Statist.* **29**.

[64] Sokal, R.R. (1963). The principles and practice of numerical taxonomy. *Taxon* **12** 190–199.

[65] Stout, Q.F. (2008). Unimodal regression via prefix isotonic regression. *Comput. Statist. Data Anal.* **53** 289–297. MR2649085

[66] Tsybakov, A.B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. Berlin: Springer.

[67] Turnbull, B.C. and Ghosh, S.K. (2014). Unimodal density estimation using Bernstein polynomials. *Comput. Statist. Data Anal.* **72** 13–29. MR3139345

[68] van Handel, R. (2014). Probability in High Dimension. Lecture Notes (Princeton University).

[69] van de Geer, S. (1990). Estimating a regression function. *Ann. Statist.* **18**.

[70] van de Geer, S. (1991). The entropy bound for monotone functions, Technical Report No. 91-10, Leiden Univ.

[71] van de Geer, S. (1993). Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.* **21**.

[72] Zhang, C.-H. (2002). Risk bounds in isotonic regression. *Ann. Statist.* **30** 528–555.

Received January 2017 and revised August 2017