Volume Ratio, Sparsity, and Minimaxity under Unitarily Invariant Norms

Zongming Ma and Yihong Wu

***Abstract*—This paper presents a non-asymptotic study of the minimax estimation of high-dimensional mean and covariance matrices. Based on the convex geometry of finite-dimensional Banach spaces, we develop a unified volume ratio approach for determining minimax estimation rates of unconstrained mean and covariance matrices under all unitarily invariant norms. We also establish the rate for estimating mean matrices with group sparsity, where the sparsity constraint introduces an additional term in the rate whose dependence on the norm differs completely from the rate of the unconstrained counterpart .**

I. Introduction

Driven by contemporary applications such as functional genomics, network analysis, machine learning, etc., there has been a recent surge in the study of estimating large-scale mean matrices and covariance matrices. See, for instance, [1, 2, 3, 4]. The difficulty of a high-dimensional estimation problem can be captured by the *minimax rate*, a function of the model parameters which is within absolute constant factors of the exact minimax risk. There are two major challenges arising from high-dimensional matrix estimation problems:

1) The matrix to be estimated is a *finite* but *high dimensional* object. In many contexts the size of the matrix can far exceed the sample size or the signal-to-noise ratio.
2) Different applications require the use of different *matrix norms as the loss function* other than the traditional quadratic loss (i.e., Frobenius norm loss). For example, Bickel and Levina [1, 2] considered spectral norm loss for covariance matrix estimation; Rohde and Tsybakov [4] used Schatten norm loss in the study of trace regression.

As pointed out in [5], the minimax rates of these matrix problems can depend critically on the choice of norms in the loss function. In the literature, such dependence has so far been explored mostly on a case-by-case basis. Determining the minimax rates under general matrix norm losses calls for new machinery. It turns out that many of the commonly used norms in loss functions are *unitarily invariant* (see Section II for definition), e.g., Frobenius norm, spectral norm, and, more generally, the classes of Ky Fan norms and Schatten norms [6]. In this paper, we establish minimax rates in several matrix estimation problems for *all unitarily invariant norm losses* via a *unified* approach. The results depend crucially on the convex geometry of finite-dimensional Banach spaces.

Z. Ma is with the Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA. Email: zongming@wharton.upenn.edu. Y. Wu is with Department of Electrical and Computer Engineering, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA. Email: yihongwu@illinois.edu.

In many matrix estimation problems, the parameter of interest belongs to (or is well approximable by) a space of much lower dimension than the size of the matrix. Examples include bandable matrices [1], sparse matrices [2], (nearly) low rank matrices [3], etc. We call this lower-dimensional space the *support* of the parameter. The minimax rates in such structured problems can usually be expressed as the sum (or equivalently, the maximum) of two terms (e.g., [7, 8, 9]): one *oracle* risk term arising from estimation error of an oracle estimator which knows the support, and the other *excess* risk term originating from the uncertainty about the support and approximation. In some cases, one term can dominate the other in certain regime of the parameter space.

As a prelude to studying structured problems, we first focus on the minimax estimation of a mean or covariance matrix without structural assumptions. Such investigation yields a legitimate lower bound to the related structured problem via an oracle argument by assuming knowledge of the support, and provides insights on how the statistical difficulty depends on the interplay between the metric and the statistical structures. These "unstructured" problems are non-trivial due to the generality of loss functions induced by unitarily invariant norms. To the best of our knowledge, even for estimating a covariance matrix with independent normal samples, the minimax rate under the squared Frobenius norm is not known for all sample size, dimension, and spectral radius.

The oracle lower bounds are obtained by an application of Fano's lemma to a local Kullback-Leibler (KL) neighborhood, followed by an estimate of the packing number via volume estimates. The standard strategy (see, e.g., [10, 11, 12]) is to turn the estimation problem into a multiple hypothesis testing problem by choosing an ϵ-packing set (with respect to the loss function) of the parameter space. If the log-cardinality of the set is sufficiently larger than the maximal mutual information, then the hypotheses cannot be discriminated reliably, which then incurs an estimation error at least ϵ. Capitalizing on the finite-dimensionality and the volume measure on the Euclidean space, we take this standard approach one step further by lower bounding the packing number using the *volume ratio*:

$$\frac{\mathrm{vol}(\text{KL neighborhood})}{\mathrm{vol}(\text{norm ball})}.$$

This abstract approach allows us to sidestep the explicit construction of packing sets used in Fano's inequality. Exploiting the connections between Gaussian measures and volume estimates in convex geometry, we further bound the volume of the KL neighborhood and the norm ball from below and above

using Urysohn's inequality and inverse Santaló's inequality [13], respectively. Consequently, the *Gaussian width* of the norm ball plays a key role in the oracle lower bounds. Surprisingly, the oracle minimax rates in both mean and covariance estimation depend on the norm only through the value of the norm on the identity matrix.

For structured problem, we need to further determine the excess risk, which can depend on the norm in a different way from the oracle risk. For the problem of orthogonal regression with group sparsity [8], we show that the excess risk depends on any unitarily invariant norm only through its (restricted) *Lipschitz constant* defined in Section II. In contrast, the oracle risk only depends on the norm of the identity matrix. See Section V and [14] for details.

Our lower bound techniques are closely related to the minimax results of Yang and Barron [12] and Birgé [11], which are obtained for general models under conditions on the loss function and the growth of the metric entropy. In this paper, we impose minimal technical conditions since we focus on concrete matrix models. Moreover, we note the following distinctions which render the results from [12, 11] not directly applicable: First, [12] gives the optimal rate for minimax estimation over massive parameter sets, whose metric entropy (with respect to the KL divergence) grows super polynomially, such as those infinite-dimensional spaces used in nonparametric function estimation. However, their lower bound is known to be loose for finite-dimensional spaces [12, Section 7], while the matrices we are interested in are *finite-but-high-dimensional* objects. Second, while the minimax lower bound in [12, Theorem 1] applies to arbitrary losses satisfying a weak triangle inequality, it is only shown to be tight for the KL loss $L(\theta,\theta') = D(P_\theta \,||\, P_{\theta'})$ or its equivalent under suitable entropy growth conditions. On the other hand, the results in [11] are dedicated to squared Hellinger loss. In contrast, our method is applicable to any norm loss under the matrix models considered in the current paper, and, in particular, optimal for all unitarily invariant norm losses.

II. Preliminaries

For a positive integer p, $[p]$ denotes the set $\{1,2,...,p\}$. For any set I, $|I|$ denotes its cardinality. For any square matrix $A = (a_{ij})$, denote its trace by $\mathsf{Tr}(A) = \sum_i a_{ii}$. Denote by S_k (resp. S_k^+) the set of $k \times k$ symmetric (resp. positive semi-definite) matrices. We use $\mathbf{1}$ to denote the all-one vector.

For any real number a and b, set $a \vee b = \max\{a,b\}$, $a \wedge b = \min\{a,b\}$ and $a_+ = a \vee 0$. For any sequences $\{a_n\}$ and $\{b_n\}$ of positive numbers, we write $a_n \asymp b_n$ if $\frac{a_n}{b_n}$ are bounded from below and above by absolute positive constants.

A. Unitarily invariant norms

On a Hilbert space, the *dual norm* of a norm $\|\cdot\|$ is defined as $\|x\|_* = \sup_{\|y\|\le 1} \langle x, y\rangle$. Two Hilbert spaces are of interest: 1) the Euclidean space $\mathbb{R}^d$ with inner product $\langle x, y\rangle = x'y$, and 2) the space $\mathbb{R}^{k\times m}$ of $k \times m$ matrices, with inner product $\langle A, B\rangle = \mathsf{Tr}(A'B)$. By definition, we have $\langle x, y\rangle \le \|x\|\|y\|_*$.

We need the notion of symmetric gauge to define unitarily invariant norms. A function $\tau : \mathbb{R}^d \to [0,\infty)$ is called a *symmetric gauge function* if it is a norm on $\mathbb{R}^d$ which is invariant with respect to sign changes and permutations [6].

Lemma 1. *Let τ be a symmetric gauge function on $\mathbb{R}^d$. Then*

1) *τ is monotone: $\tau(x_1, x_2, \ldots, x_d) \ge \tau(x_1', x_2, \ldots, x_d)$ for any $|x_1| \ge |x_1'|$ and any $x_2, \ldots, x_d$;*
2) *The dual norm τ_* is also a symmetric gauge function and satisfies $\tau_*(\mathbf{1})\tau(\mathbf{1}) = d$.*

A matrix norm $\|\cdot\|$ is called a *unitarily invariant norm* if for any $A \in \mathbb{R}^{k\times m}$ and any orthogonal matrices U and V, $\|A\| = \|UAV\|$. A fundamental result due to von Neumann states that for any unitarily invariant norm $\|\cdot\|$ on $\mathbb{R}^{k\times m}$, there exists a symmetric gauge function τ on $\mathbb{R}^{k\wedge m}$ such that $\|A\| = \tau(\sigma(A))$, where $\sigma(A) = (\sigma_1(A), \ldots, \sigma_{k\wedge m}(A))'$ consists of singular values of A in decreasing order. Henceforth we denote a unitarily invariant norm by $\|\cdot\|_\tau$ with τ the associated symmetric gauge. On the space of $k \times m$ matrices, the dual norm of $\|\cdot\|_\tau$ is the unitarily invariant norm $\|\cdot\|_{\tau_*}$, induced by τ_*, the dual of τ on $\mathbb{R}^{k\wedge m}$ [6, Proposition IV.2.11]. The Lipschitz constant of τ is

$$L_\tau = \sup_{x\neq y} \frac{|\tau(x) - \tau(y)|}{\|x - y\|_2} = \sup_{x\neq 0} \frac{\tau(x)}{\|x\|_2}, \tag{1}$$

where $\|x\|_2$ is the Euclidean norm of x.

We now introduce two important classes of unitarily invariant matrix norms: *Schatten q-norms* and *Ky Fan ℓ-norms*. For any $q \in [1,\infty]$, the Schatten q-norm of $A = (a_{ij}) \in \mathbb{R}^{k\times m}$ is

$$\|A\|_{\mathrm{S}_q} = (\textstyle\sum_{i=1}^{k\wedge m} \sigma_i^q(A))^{1/q}. \tag{2}$$

The dual norm of $\|\cdot\|_{\mathrm{S}_q}$ is $\|\cdot\|_{\mathrm{S}_{q^*}}$, where $1/q + 1/q^* = 1$. For any $\ell \in [k \wedge m]$, the Ky Fan ℓ-norm of A is

$$\|A\|_{(\ell)} = \textstyle\sum_{i=1}^{\ell} \sigma_i(A). \tag{3}$$

We note the following special cases: 1) Frobenius norm: $\|A\|_{\mathrm{S}_2} = (\sum_i \sigma_i^2(A))^{1/2} = (\sum_{i,j} a_{ij}^2)^{1/2}$, also denoted by $\|A\|_{\mathrm{F}}$; 2) Spectral (operator) norm: $\|A\|_{\mathrm{S}_\infty} = \|A\|_{(1)} = \sigma_1(A)$, also denoted by $\|A\|_{\mathrm{op}}$; 3) Nuclear norm: $\|A\|_{\mathrm{S}_1} = \|A\|_{(k\wedge m)} = \sum_{i=1}^{k\wedge m} \sigma_i(A)$.

B. Volume ratio of convex bodies

Recall that K is a *symmetric convex body* in $\mathbb{R}^d$ if K is a compact convex set with non-empty interior such that $K = -K$. In particular, norm balls are symmetric convex bodies. Let $B^d_{\|\cdot\|}(\epsilon) = \{x \in \mathbb{R}^d : \|x\| \le \epsilon\}$ be the norm ball of radius ϵ centered at zero, and B_2^d and $B_2^{k\times m}$ be the unit Euclidean ball and Frobenius ball at zero in $\mathbb{R}^d$ and $\mathbb{R}^{k\times m}$, respectively. We omit the superscript of dimension when no confusion arises.

For a convex body K, $K^\circ = \left\{y \in \mathbb{R}^d : \sup_{x\in K} \langle x, y\rangle \le 1\right\}$ is its *polar*, which is also convex. The *Minkowski functional* of a symmetric convex body K is defined as $\|x\|_K = \inf\{r > 0 : x \in rK\}$. Note that if $K = \{x : \|x\| \le 1\}$ is some unit norm ball, then $\|\cdot\|_K = \|\cdot\|$. Moreover, $\|\cdot\|_{K^\circ} = \sup_{x\in K} \langle x, \cdot\rangle$.

The following result reveals a deep connection between the volume of a convex body and the Gaussian measure.

Lemma 2 (Urysohn's Inequality [13, p.7])**.** *Let K be a symmetric convex body in $\mathbb{R}^d$. Then*

$$\left(\frac{\mathrm{vol}(K)}{\mathrm{vol}(B_2)}\right)^{\frac{1}{d}} \leq \frac{1}{\sqrt{d}}\,\mathbb{E}\sup_{y\in K}\langle G, y\rangle\,, \tag{4}$$

where $G \sim N(0, I_d)$ is standard Gaussian. The expectation of the supremum on the right-hand side of (4) is called the Gaussian width *of K.*

Moreover, for any symmetric convex body $K \subset \mathbb{R}^d$,

$$\frac{1}{2} \leq \left(\frac{\mathrm{vol}(K)\mathrm{vol}(K^\circ)}{\mathrm{vol}(B_2^d)^2}\right)^{\frac{1}{d}} \leq 1. \tag{5}$$

The upper bound is known as Santaló's inequality [13, p. 100]. The lower bound is due to [15]. The product $\mathrm{vol}(K)\mathrm{vol}(K^\circ)$ is called the *Mahler volume* of K. In view of (5) and the fact that $\mathrm{vol}(B_2^d)^{\frac{1}{d}} = \frac{\sqrt{\pi}}{\Gamma(\frac{d}{2}+1)^{\frac{1}{d}}} \asymp \frac{1}{\sqrt{d}}$, applying Lemma 2 to the polar K° yields the following lemma which is useful in lower bounding the volume of a convex body.

Lemma 3 (Inverse Santaló's inequality)**.** *There exists a universal constant c_0, such that for any symmetric convex body K in $\mathbb{R}^d$,*

$$\mathrm{vol}(K)^{\frac{1}{d}} \geq \frac{c_0}{\mathbb{E}\,\|G\|_K}. \tag{6}$$

For the space of $k \times m$ matrices, Lemmas 2 and 3 hold with $d = km$ and G a $k \times m$ Gaussian random matrix with iid $N(0,1)$ entries. Both lemmas can be further generalized to the convex bodies in the space of symmetric matrices, S_k, with d replaced by the dimension $d_k = k(k+1)/2$, and G replaced by $\frac{G+G'}{2}$, the Gaussian Orthogonal Ensemble GOE(k). Note that the volume on the linear subspace S_k is defined by the usual Jacobian determinant formula.

III. Oracle minimax rates for mean matrices

As pointed out in the introduction, understanding the minimax rates in unconstrained matrix estimation problems is the first step toward deriving the rates in those with structural constraints. In this section, we derive tight minimax rates for the unconstrained mean matrix estimation problem under all unitarily invariant norms.

Consider the following Gaussian mean problem, where we observe the $k \times m$ matrix

$$Y = M + Z. \tag{7}$$

Here $M \in \mathbb{R}^{k\times m}$ is the mean matrix we want to estimate, and Z is the noise matrix with i.i.d. $N(0,1)$ entries.

Note that we can always vectorize the Y, M and Z matrices in (7), reducing the model to a d-dimensional Gaussian mean problem with $d = km$. In addition, any matrix norm on $\mathbb{R}^{k\times m}$ induces a vector norm on $\mathbb{R}^d$. In view of such a connection, we derive below a general lower bound for estimating a mean vector in $\mathbb{R}^d$ from observations contaminated by Gaussian white noises.

First, we establish the connection between minimax lower bounds and volume ratios in the following proposition, which is a variant of Fano's lemma [10, Lemma 5.1, p. 356].

Proposition 1. *Let $\Theta \subset \mathbb{R}^d$ and $\|\cdot\|$ be a norm on $\mathbb{R}^d$. Let $\{P_\theta : \theta \in \Theta\}$ a collection of probability measures. Define the* Kullback-Leibler diameter *of T by $d_{\mathrm{KL}}(T) \triangleq \sup_{\theta,\theta'\in T} D(P_\theta \,||\, P_{\theta'})$. Then*

$$\inf_{\hat\theta}\sup_{\theta\in\Theta}\mathbb{E}_\theta\|\hat\theta(X)-\theta\|^2 \geq \sup_{T\subset\Theta}\sup_{\epsilon>0}\frac{\epsilon^2}{4}\left(1-\frac{d_{\mathrm{KL}}(T)+\log 2}{\log\frac{\mathrm{vol}(T)}{\mathrm{vol}(B_{\|\cdot\|}(\epsilon))}}\right).$$

The specialization of Proposition 1 to Gaussian measures, together with Lemma 2, leads to the following result for Gaussian location model.

Theorem 1 (General norm)**.** *For any $d \in \mathbb{N}$, consider the Gaussian location model $Y = \theta + Z$, where $\theta \in \mathbb{R}^d$ and $Z \sim N(0, I_d)$ is a d-dimensional white noise vector. Then for any norm $\|\cdot\|$ on $\mathbb{R}^d$,*

$$\frac{\log 2}{2048}\frac{d^2}{(\mathbb{E}\|Z\|_*)^2} \leq \inf_{\hat\theta}\sup_{\theta\in\mathbb{R}^d}\mathbb{E}_\theta\|\hat\theta(Y)-\theta\|^2 \leq \mathbb{E}\|Z\|^2, \tag{8}$$

where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$.

Remark 1. Recall from Lemma 2 that the Gaussian width of a symmetric convex body $K \subset \mathbb{R}^d$ is $\mathbb{E}\max_{x\in K}\langle x, Z\rangle$. By the definition of the dual norm, the quantity $\mathbb{E}\|Z\|_*$ in the lower bound (8) is equal to the Gaussian width of the unit ball in $\mathbb{R}^d$ equipped with the norm $\|\cdot\|$ used in the loss function.

Proof: The upper bound is obtained by taking the specific estimator $\hat\theta = Y$ and the triangle inequality. To prove the lower bound, note that the Kullback-Leibler divergence of the normal mean model is given by

$$D(N(\theta, I_d)\,||\,N(\theta', I_d)) = \frac{1}{2}\|\theta-\theta'\|_2^2, \tag{9}$$

where $\|\cdot\|_2$ denotes the ℓ_2-norm on $\mathbb{R}^d$. Let $T = B_2(\delta) = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq \delta\}$ denote the Euclidean ball of radius δ centered at the origin. Then $d_{\mathrm{KL}}(T) \leq 4\delta^2$. Moreover,

$$\frac{\mathrm{vol}(B_2(\delta))}{\mathrm{vol}(B_{\|\cdot\|}(\epsilon))} = \frac{\delta^d\mathrm{vol}(B_2(1))}{\epsilon^d\mathrm{vol}(B_{\|\cdot\|}(1))} \geq \left(\frac{\delta\sqrt{d}}{\epsilon\,\mathbb{E}\|Z\|_*}\right)^d, \tag{10}$$

where the last inequality follows from Lemma 2. Now we choose $\delta = \frac{\sqrt{d\log 2}}{2\sqrt{2}}$ and $\epsilon = \frac{\delta\sqrt{d}}{4\,\mathbb{E}\|Z\|_*} = \frac{\sqrt{\log 2}d}{8\sqrt{2}\,\mathbb{E}\|Z\|_*}$. Applying Proposition 1 and using $d \geq 1$, we obtain the following lower bound

$$\inf_{\hat\theta}\sup_{\theta\in\Theta}\mathbb{E}_\theta\|\hat\theta(Y)-\theta\|^2 \geq \left(\frac{3}{4}-\frac{1}{2d}\right)\frac{\epsilon^2}{4} = \frac{d^2\log 2}{2048(\mathbb{E}\|Z\|_*)^2}.$$

■

Turning back to the matrix Gaussian location model (7), we are now in the position of establishing the minimax rates for estimating M with respect to all unitarily invariant norms.

Note that any matrix norm on the space $\mathbb{R}^{k\times m}$ induces a vector norm on $\mathbb{R}^d$ for $d = km$. In view of Theorem 1, it suffices to upper bound both $\mathbb{E}\|Z\|_*$ and $\mathbb{E}\|Z\|^2$, provided that the resulting lower and upper bounds agree up to a constant factor. It turns out that this can indeed be achieved, resulting in the following theorem.

Theorem 2. *For any $k, m \in \mathbb{N}$ and any unitarily invariant norm $\|\cdot\|_\tau$, where τ is a symmetric gauge function on $\mathbb{R}^{k\wedge m}$, the minimax rate for estimating M under* (7) *with respect to the loss $\|\cdot\|_\tau^2$ satisfies*

$$\inf_{\widehat{M}} \sup_{M\in\mathbb{R}^{k\times m}} \mathbb{E}\|\widehat{M} - M\|_\tau^2 \asymp (k\vee m)\tau^2(\mathbf{1}), \qquad (11)$$

where $\mathbf{1}$ denotes the all-one vector in $\mathbb{R}^{k\wedge m}$.

Remark 2 (Dependence on τ)**.** Theorem 2 reveals the following remarkable fact: The minimax rate under the unitarily invariant norm $\|\cdot\|_\tau$ depends on the symmetric gauge function τ *only* through its value at the all-one vector. On the one hand, $\tau(\mathbf{1})$ appears in the lower bound because it governs the volume asymptotics of a unit ball under the $\|\cdot\|_\tau$ norm in $\mathbb{R}^{k\times m}$. On the other hand, since the noise matrix has i.i.d. entries, all of its singular values scale with the dimensions at the same rate. Hence, the risk achieved by the observation is also proportional to $\tau^2(\mathbf{1})$. In addition, such a dependence pattern also suggests that the least-favorable prior on M should concentrate on those matrices in general position, i.e., having full rank and bounded condition number. This is intuitively natural because neither the unitarily invariant norm nor the noise singular value spectrum favor any specific direction.

Remark 3. Theorem 2 also provides a rigorous justification of the following intuitive fact: If both the noise and the loss function are sufficiently symmetric, then there is nothing significantly better than estimating by the raw observation, which is the maximum likelihood estimator under the Gaussian assumption. Of course, the caveat is that such a claim crucially depends on the choice of the loss function. For example, if the loss function is given by $L(\widehat{M}, M) = \rho(\|\widehat{M} - M\|_{\mathrm{F}})$, where $\rho(x) = x^2 + (k\vee m)^4\mathbf{1}_{\{x\le 1\}}$, then estimating by the observation is clearly rate-suboptimal. Instead, the minimax estimator can be obtained by shrinkage toward zero.

Proof of Theorem 2: Note that τ_* is also a symmetric gauge function. By the monotonicity of symmetric gauge functions (cf. Lemma 1), we have for $\eta = \tau$ or τ_*,

$$\|Z\|_\eta = \eta(\sigma(Z)) \le \eta(\sigma_1(Z)\mathbf{1}) = \sigma_1(Z)\eta(\mathbf{1}). \qquad (12)$$

For the lower bound, (12) leads to $\|Z\|_{\tau_*} \le \sigma_1(Z)\tau_*(\mathbf{1}) = \frac{\sigma_1(Z)(k\wedge m)}{\tau(\mathbf{1})}$, where the last equality is due to the second claim of Lemma 1. Applying Theorem 1, we have

$$\begin{aligned}\inf_{\widehat{M}} \sup_{M\in\mathbb{R}^{k\times m}} \mathbb{E}\|\widehat{M} - M\|^2 &\ge \frac{ck^2m^2}{(\mathbb{E}\|Z\|_{\tau_*})^2} \ge \frac{c(k\vee m)^2\tau^2(\mathbf{1})}{(\mathbb{E}\sigma_1(Z))^2} \\ &\ge \frac{c(k\vee m)^2\tau^2(\mathbf{1})}{(\sqrt{k}+\sqrt{m})^2} \ge c(k\vee m)\tau^2(\mathbf{1}),\end{aligned}$$

where we have used Gordon's inequality $\mathbb{E}\sigma_1(Z) \le \sqrt{k}+\sqrt{m}$; cf. [16].

For the upper bound, in view of (12), it suffices to bound $\mathbb{E}\sigma_1(Z)^2$. To this end, note that the Davidson–Szarek bound [16] implies that for any $a > 1$,

$$\mathbb{P}(\sigma_1(Z) > a(\sqrt{k}+\sqrt{m})) \le \mathrm{e}^{-(a-1)^2(\sqrt{k}+\sqrt{m})^2/2}.$$

Integrating the above with respect to a leads to $\mathbb{E}\sigma_1(Z)^2 \le (5+\sqrt{\pi})(k\vee m)$. In view of Theorem 1 and (12), we obtain the desired upper bound. ∎

IV. Oracle minimax rates for covariance matrices

In this section we switch to the problem of estimating covariance matrices and show that the volume approach developed in Section III can be successfully imported here to derive optimal minimax rates. Let X denote the observed $n\times k$ data matrix, whose rows $X_{1*}, \ldots, X_{n*}$ are independently drawn from $N(0,\Sigma)$. A sufficient statistic for Σ is the sample covariance matrix $S = \frac{1}{n}X'X$.

Without assuming additional covariance structure, we consider the following parameter space for Σ:

$$\Xi(k,\lambda) = \{\Sigma \in \mathsf{S}_k^+ : \|\Sigma\|_{\mathrm{op}} \le \lambda\}, \qquad (13)$$

which is simply the operator norm ball of radius λ in the space of $k\times k$ positive semi-definite matrices.

We have the following analogous result to Theorem 2 for covariance matrices. The main difference is that instead of (9), the KL divergence in the covariance model is given by

$$D(N(0,\widehat{\Sigma}) \,||\, N(0,\Sigma)) = \frac{1}{2}\mathsf{Tr}(\Sigma^{-1}\widehat{\Sigma} - I) - \frac{1}{2}\log\frac{\det\widehat{\Sigma}}{\det\Sigma}.$$

Therefore the KL neighborhood in the covariance model is not a Frobenius ball, which requires additional volume estimates via the inverse Santaló inequality in the lower bound argument. The detailed proof can be found in [14, Section 5].

Theorem 3. *For any $n, k \in \mathbb{N}$, any $\lambda > 0$, and any unitarily invariant norm $\|\cdot\|_\tau$, where τ is a symmetric gauge function on $\mathbb{R}^k$,*

$$\inf_{\widehat{\Sigma}} \sup_{\Sigma\in\Xi(k,\lambda)} \mathbb{E}\|\widehat{\Sigma} - \Sigma\|_\tau^2 \asymp \left(\frac{k}{n}\wedge 1\right)\lambda^2\tau^2(\mathbf{1}). \qquad (14)$$

Analogous to the discussion of Theorem 2 in Remark 2, the minimax rate in Theorem 3 is also proportional to $\tau^2(\mathbf{1})$, which suggests that the worst-case prior are in general position. Note that it is natural that the minimax rate in Theorem 3 is proportional to the squared spectral radius λ^2. The reasons are two-fold: First, since the the covariance model is a scale model, the Kullback-Leibler divergence is scaling invariant. On the other hand, the loss in terms of squared norm scales quadratically with λ^2. Second, the magnitude of the "effective noise" matrix $S - \Sigma$ also scales with the spectral norm of Σ.

It is interesting to compare Theorem 3 to the classical results focusing on the *exact* minimax risk of estimating the covariance matrices in the low-dimensional regime. Using invariance

theory, Stein [17] proved that if $k \le n$, any constant multiple of the sample covariance matrix is not minimax with respect to the KL loss; He also obtained the exact minimax estimator for this problem. In contrast, our focus here is to investigate the minimax *rate*, the non-asymptotic characterization of the minimax risk modulo constants. In particular, we see that the sample covariance matrix is minimax *rate*-optimal for all triples (k, n, λ) and all unitarily invariant norms. This conclusion, even in the simplest setting of quadratic loss (squared Frobenius norm), seems to be new in the literature.

V. Mean matrix estimation under group sparsity

To illustrate how the choice of norm influences the minimax rates in structured problem, consider the following mean matrix estimation problem under group sparsity [8]. For any matrix M, let $\mathrm{supp}(M)$ be the index set of nonzero rows of M. Suppose we observe a $p \times m$ matrix

$$Y = M + Z, \tag{15}$$

where $Z = (z_{ij})$ with $z_{ij} \overset{\text{i.i.d.}}{\sim} N(0,1)$ and M belongs to the parameter space $\mathcal{F}_0(k,p,m) = \{M \in \mathbb{R}^{p\times m} : |\mathrm{supp}(M)| \le k\}$, that is, at most $k(\le p)$ rows of M contain nonzero entries. We are interested in estimating M. If the oracle knowledge of $\mathrm{supp}(M)$ were given, the problem would reduce to the unstructured mean estimation problem studied in Section III.

To describe the minimax rate of (15), we introduce the following notations. Let $\|\cdot\|_\tau$ be an arbitrary unitarily invariant norm on $\mathbb{R}^{p\times m}$, given by a symmetric gauge τ on $\mathbb{R}^{p\wedge m}$. Denote by $\tau|_k$ the restriction of τ on $\mathbb{R}^{k\wedge m}$, i.e., $\tau|_k(x_1,\dots,x_{k\wedge m}) = \tau(x_1,\dots,x_{k\wedge m},0,\dots,0)$. Then $\tau|_k$ is a symmetric gauge on $\mathbb{R}^{k\wedge m}$, whose Lipschitz constant $L_{\tau|_k}$ is defined according to (1).

Theorem 4. *For any unitarily invariant norm $\|\cdot\|_\tau$, the minimax rate for estimating M under model (15) is*

$$\inf_{\widehat{M}} \sup_{\mathcal{F}_0(k,p,m)} \mathbb{E}\|\widehat{M}-M\|_\tau^2 \asymp (\tau|_k)^2(\mathbf{1})(k\vee m) + L_{\tau|_k}^2\, k\log\frac{\mathrm{e}p}{k}.$$

Example 1 (Schatten norm)**.** For the Schatten q-norm (2) with $q \in [1,\infty]$, $\tau|_k(\mathbf{1}) = r^{1/q}$ and $L_{\tau|_k} = r^{(1/q-1/2)_+}$, where $r = k\wedge m$. Then, Theorem 4 gives the rate

$$(k\wedge m)^{2/q}(k\vee m) + (k\wedge m)^{(2/q-1)_+}\, k\log\frac{\mathrm{e}p}{k}.$$

Example 2 (Ky Fan norm)**.** For the Ky Fan ℓ-norm (3) with $l \in [k\wedge m]$, $\tau|_k(\mathbf{1}) = \ell, L_{\tau|_k} = \sqrt{\ell}$, and so the rate is

$$\ell^2(k\vee m) + \ell k\log\frac{\mathrm{e}p}{k}.$$

The minimax rate in Theorem 4 consists of two parts: the *oracle risk*, which is the minimax risk if one knows $\mathrm{supp}(M)$ *a priori*, and the *excess risk*, which is due to the combinatorial uncertainty of the support set. The two terms depend on τ and hence the norm in very different ways: the oracle risk depends on τ via $\tau|_k(\mathbf{1})$ and the excess risk via $L_{\tau|_k}$. The oracle risk follows from Theorem 2. To lower bound the excess risk, we construct a least favorable configuration from the worst-case matrix attaining the Lipschitz constant $L_{\tau|_k}$. For details of the lower bound and the estimating procedure, see [14].

References

[1] P. Bickel and E. Levina, "Regularized estimation of large covariance matrices," *The Annals of Statistics*, vol. 36, no. 1, pp. 199–227, 2008.

[2] ——, "Covariance regularization by thresholding," *The Annals of Statistics*, vol. 36, no. 6, pp. 2577–2604, 2008.

[3] S. Negahban and M. Wainwright, "Estimation of (near) low-rank matrices with noise and high-dimensional scaling," *The Annals of Statistics*, vol. 39, no. 2, pp. 1069–1097, 2011.

[4] A. Rohde and A. Tsybakov, "Estimation of high-dimensional low-rank matrices," *The Annals of Statistics*, vol. 39, no. 2, pp. 887–930, 2011.

[5] T. T. Cai, C.-H. Zhang, and H. H. Zhou, "Optimal rates of convergence for covariance matrix estimation," *The Annals of Statistics*, vol. 38, no. 4, pp. 2118–2144, 2010.

[6] R. Bhatia, *Matrix analysis*. New York, NY: Springer Verlag, 1997.

[7] T. T. Cai, Z. Ma, and Y. Wu, "Sparse PCA: Optimal rates and adaptive estimation," 2012, preprint. [Online]. Available: http://arxiv.org/abs/1211.1309

[8] K. Lounici, M. Pontil, S. Van De Geer, and A. B. Tsybakov, "Oracle inequalities and optimal inference under group sparsity," *The Annals of Statistics*, vol. 39, no. 4, pp. 2164–2204, 2011.

[9] G. Raskutti, M. Wainwright, and B. Yu, "Minimax-optimal rates for sparse additive models over kernel classes via convex programming," *The Journal of Machine Learning Research*, vol. 13, pp. 389–427, 2012.

[10] I. Ibragimov and R. Has'minskii, *Statistical Estimation: Asymptotic Theory*. Springer, 1981.

[11] L. Birgé, "Approximation dans les espaces métriques et théorie de l'estimation," *Z. für Wahrscheinlichkeitstheorie und Verw. Geb.*, vol. 65, no. 2, pp. 181–237, 1983.

[12] Y. Yang and A. R. Barron, "Information-theoretic determination of minimax rates of convergence," *The Annals of Statistics*, vol. 27, no. 5, pp. 1564–1599, 1999.

[13] G. Pisier, *The volume of convex bodies and Banach space geometry*. Cambridge University Press, 1999.

[14] Z. Ma and Y. Wu, "Volume ratio, sparsity, and minimaxity under unitarily invariant norms," preprint. [Online]. Available: http://www.ifp.illinois.edu/ yihongwu/volume.pdf

[15] G. Kuperberg, "From the Mahler conjecture to Gauss linking integrals," *Geometric And Functional Analysis*, vol. 18, no. 3, pp. 870–892, 2008.

[16] K. Davidson and S. Szarek, *Handbook on the Geometry of Banach Spaces*. Elsevier Science, 2001, vol. 1, ch. Local operator theory, random matrices and Banach spaces, pp. 317–366.

[17] C. Stein, "Some problems in multivariate analysis, Part I," Stanford University, Department of Statistics, Tech. Rep. 6, Oct. 1956.