Bernoulli **28**(4), 2022, 2578–2605
https://doi.org/10.3150/21-BEJ1430

Matrix means and a novel high-dimensional shrinkage phenomenon

ASAD LODHIA[1,a], KEITH LEVIN[2,b] and ELIZAVETA LEVINA[3,c]

[1]*Broad Institute of MIT and Harvard, 415 Main Street, Cambridge, MA 02142-1027.* [a]*asad.i.lodhia@gmail.com*
[2]*Department of Statistics, University of Wisconsin-Madison, 1220 Medical Sciences Center, 1300 University Ave, Madison, WI 53706-1510.* [b]*kdlevin@wisc.edu*
[3]*Department of Statistics, University of Michigan, 323 West Hall, 1085 S. University Ave, Ann Arbor, MI 48109-1107.* [c]*elevina@umich.edu*

Many statistical settings call for estimating a population parameter, most typically the population mean, based on a sample of matrices. The most natural estimate of the population mean is the arithmetic mean, but there are many other matrix means that may behave differently, especially in high dimensions. Here we consider the matrix harmonic mean as an alternative to the arithmetic matrix mean. We show that in certain high-dimensional regimes, the harmonic mean yields an improvement over the arithmetic mean in estimation error as measured by the operator norm. Counter-intuitively, studying the asymptotic behavior of these two matrix means in a spiked covariance estimation problem, we find that this improvement in operator norm error does not imply better recovery of the leading eigenvector. We also show that a Rao-Blackwellized version of the harmonic mean is equivalent to a linear shrinkage estimator studied previously in the high-dimensional covariance estimation literature, while applying a similar Rao-Blackwellization to regularized sample covariance matrices yields a novel nonlinear shrinkage estimator. Simulations complement the theoretical results, illustrating the conditions under which the harmonic matrix mean yields an empirically better estimate.

Keywords: Matrix means; positive definite cone; covariance estimation; shrinkage Estimators

1. Introduction

Matrix estimation problems arise in statistics in a number of areas, most prominently in covariance estimation, but also in network analysis, low-rank recovery and time series analysis. Typically, the focus is on estimating a matrix based on a single noisy realization of that matrix. For example, the problem of covariance estimation [25] focuses on estimating a population covariance matrix based on a sample of vectors, which are usually combined to form a sample covariance matrix or another estimate of the population covariance. In network analysis and matrix completion problems, the goal is typically to estimate the expectation of a matrix-valued random variable based on a single observation under suitable structural assumptions (see, e.g., [17] and citations therein). A related setting that has received less attention is the case where a sample of matrices is available and the goal is to estimate an underlying population mean or other parameter. This arises frequently in neuroimaging data analysis, where each matrix represents connectivity within a particular subject's brain and the goal is to estimate a population brain connectivity pattern [10,55].

The most direct approach to estimating the underlying population mean from a sample of matrices is to take the arithmetic (sample) mean, perhaps with some regularization to ensure the stability of the resulting estimate. The arithmetic matrix mean is the mean with respect to Euclidean geometry on the space of matrices, which is often not the most suitable average for a given matrix model. A simple example can be found in our recent work [39], where we showed that in the problem of estimating a low-rank expectation of a collection of independent random matrices with different variances, a

weighted average improves upon the naïve arithmetic matrix mean, analogously to the scalar analogue, in which a weighted average can improve upon the unweighted mean in the presence of heterogeneous variances. Somewhat surprisingly in light of the rich geometry of matrices, fairly little attention has been paid in the literature to other matrix geometries and their associated means. An exception is work by Schwartzman [53], who argued for using the intrinsic geometry of the positive definite cone [7] in the problem of covariance estimation, and showed that a mean with respect to this different matrix geometry can, under certain models, yield an appreciably better estimate. Recent work has also considered Fréchet means in the context of multiple-network analysis [42]. Continuing in this vein, the current work aims to better understand the sampling distribution of matrix means other than the arithmetic mean under different matrix geometries.

Computing means with respect to different geometries has been studied at some length in the signal processing and computer vision communities, mostly in the context of the Grassmann and Stiefel manifolds [1,23,45]. See [54] for a good discussion of how taking intrinsic geometry into account leads to estimators other than the arithmetic mean. Recent work has considered similar geometric concerns in the context of network data [31,42]. Kolaczyk and coauthors [31] considered the problem of averaging multiple networks on the same number of vertices, developed a novel geometry for this setting and derived a Fréchet mean for that geometry. Recent work by Lunagómez, Olhede and Wolfe [42] considered a similar network averaging problem, and presented a framework for both specifying distributions of networks and deriving corresponding sample means. Unfortunately, most of these matrix means are not amenable to the currently available tools from random matrix theory that could help analyze their properties.

In this paper, we consider the behavior of the harmonic mean of a collection of random matrices, a matrix mean that arises from a markedly different eometry than the matrix arithmetic mean. The harmonic mean turns out to be well-suited to analysis using techniques in random matrix theory, and it is our hope that results established here will be extended to other related matrix means in the future. Building on random matrix results developed by the first author [41], we show how the harmonic matrix mean can, in certain regimes, yield a better estimate of the population mean matrix in spectral norm compared to the arithmetic mean. We also show that this improvement does not carry over to recovery of the top population eigenvector in a spiked covariance model, making an important distinction between two measures of estimation performance that are often assumed to behave similarly. We characterize the settings in which the harmonic matrix mean improves upon the arithmetic matrix mean as well as the settings in which it does not, and show the implications of these results for covariance estimation.

Our focus in this work is on estimating the population mean of a collection of Wishart matrices, which can be thought of as sample covariance matrices. There is an extensive literature on estimating a population covariance matrix on p variables from n observations based on a single covariance matrix. The sample covariance matrix is the maximum likelihood estimator for Gaussian data, and when the dimension p is fixed, classical results fully describe its behavior [2,47]. In the high-dimensional regime, where p is allowed to grow with n, the sample covariance matrix is not well-behaved, and in particular becomes singular as soon as $p \geq n$. There has been extensive work on understanding this phenomenon in random matrix theory, starting from the pioneering work of [44] and followed by numerous more recent results, especially focusing on estimating the spectrum [24,60]. Much work in random matrix theory has focused on the spiked model, in which the population covariance is the sum of the identity matrix and a low-rank signal matrix. [6,12,26,30].

The problem of covariance estimation is now several decades old (see, e.g., [56], and refer to [22] for a thorough overview of early work). In the past two decades, literature on covariance estimation in high dimensions (see [25] for a review) has focused on addressing the shortcomings of the sample covariance, mainly by applying regularization. James-Stein type shrinkage was considered in early

Bayesian approaches [19] and in the Ledoit-Wolf estimator [36], which shrinks the sample covariance matrix towards the identity matrix using coefficients optimized with respect to a Frobenius norm loss. Subsequent papers presented variants with different estimates of the optimal coefficients and different choices of the shrinkage target matrix for normal data [18,28], as well as for the more general case of finite fourth moments [57]. A related line of work has focused on estimation with respect to Stein's loss [21,38,56].

More recent work introduced the class of orthogonally invariant estimators [22], which unifies many of the regularization approaches outlined above. Given a covariance matrix, an orthogonally invariant estimator produces an estimate of the population covariance matrix with eigenspace identical to that of the sample covariance, and spectrum that is a function Φ of the spectrum of the sample covariance. Donoho and coauthors [22] showed how choices of loss function yield different (asymptotically) optimal choices of Φ. Nonlinear eigenvalue regularization approaches in this vein have been explored extensively [24,32,35,37,40,51,58]. Typically, the eigenvalues of the sample covariance matrix are adjusted either according to the asymptotic relation between the limiting spectral distribution of a spiked covariance and its Stieltjes transform or based on the method of moments applied to the limiting spectral distribution. Approaches such as these tend to improve upon the linear shrinkage estimates introduced in [36] by accounting for nonlinear dispersion of the sample eigenvalues with respect to their population analogues.

One shortcoming of orthogonally invariant estimators is that they do not in any way change the estimated eigenvectors, which are not consistent in the high-dimensional regime [5,30]. An alternative approach that overcomes this limitation is to regularize the sample covariance matrix by imposing structural constraints. This class of methods includes banding or tapering of the covariance matrix, suitable when the variables have a natural ordering [8,59], and thresholding when the variables are not ordered [9,11,52]. Minimax rates for many structured covariance estimation problems are now known [14–16]; see [13] for an overview of these results.

The remainder of this paper is laid out as follows. In Section 2 we establish notation and introduce the random matrix models under study. In Section 3, we establish the asymptotic behavior of the harmonic matrix mean under these models. In Section 4 we compute a Rao-Blackwellized version of the harmonic mean of two random covariance matrices, illuminating connections between the harmonic mean and a family of regularized covariance estimators. In Section 5, we analyze a spiked covariance model and the behavior of the top eigenvector of the harmonic mean estimator under that model. Finally, Section 6 briefly presents numerical simulations highlighting the settings in which the harmonic matrix mean does and does not have an advantage in covariance estimation. Section 7 concludes with discussion.

2. Problem setup

We begin by establishing notation. We denote the identity matrix by $\mathbb{I}$, with its dimension clear from context. For a $p \times p$ matrix M, $\|M\|$ denotes its operator norm and $\|M\|_F$ denotes its Frobenius norm. For a set A, let $\mathbf{1}_A(x) = 1$ if $x \in A$ and $\mathbf{1}_A(x) = 0$ otherwise. We denote by $\mathcal{S}_p(\mathbb{R})$ and $\mathcal{S}_p(\mathbb{C})$ the spaces of $p \times p$ symmetric and Hermitian positive definite matrices, respectively. For a $p \times p$ symmetric or Hermitian matrix M, the eigenvalues of M are denoted $\lambda_1(M) \geq \lambda_2(M) \geq \cdots \geq \lambda_p(M)$ and their corresponding eigenvectors are denoted $v_1(M), \ldots, v_p(M)$. We use $\preceq$ for the positive semidefinite ordering, so that $M_1 \preceq M_2$ if and only if $M_2 - M_1$ is positive semidefinite.

Suppose that we wish to estimate the population mean Σ of a collection of N independent identically distributed self-adjoint positive definite p-by-p random matrices. The most commonly used model for positive (semi)definite random matrices is the Wishart distribution, which arises in covariance estimation and is well-studied in the random matrix theory literature.

Definition 2.1 (Wishart Random Matrix: Real Observation Model). Let X be a random $p \times n$ matrix with columns drawn i.i.d. from a centered normal with covariance $\Sigma \in \mathbb{R}^{p \times p}$. Then

$$W = \frac{XX^*}{n}$$

is a real-valued random Wishart matrix with parameters Σ and n.

Many of our results are also true for the complex-valued version of the Wishart distribution, which we define here for the special case of identity covariance.

Definition 2.2 (Wishart Random Matrix: Complex Observation Model). Let X be a random $p \times n$ matrix with i.i.d. complex standard Gaussian random entries, i.e., entries of the form

$$\frac{Z_1 + \sqrt{-1}Z_2}{\sqrt{2}},$$

where Z_1 and Z_2 are independent standard real Gaussian random variables. Then $W = XX^*/n$ is a random matrix following the complex Wishart distribution with parameters $\mathbb{I}$ and n.

Let $\{X_i\}_{i=1}^N$ be a sequence of independent identically distributed $p \times n$ matrices with columns drawn i.i.d. from a centered normal with covariance Σ. Then for each $i = 1, 2, \ldots, N$,

$$W_i := \frac{X_i X_i^*}{n} \tag{2.1}$$

is the sample covariance matrix, which follows the real-valued Wishart distribution with parameters Σ and n. The aim of covariance estimation is to recover the population covariance Σ, with estimation error most commonly measured in Frobenius norm or operator norm, the latter of which is more relevant in some applications since, by the Davis-Kahan theorem [20], small operator norm error implies that one can recover the leading eigenvectors of Σ. This is of particular interest in covariance estimation when the task at hand is principal component analysis (see Section 5), but is also relevant in other problems when Σ is low-rank. For example, in the case of network analysis [39], the eigenvectors of Σ encode community structure.

Even in the modestly high-dimensional regime of $p/n \to \gamma \in (0, 1)$, estimating Σ is more challenging. When $\Sigma = \mathbb{I}$, the spectral measure of each W_i satisfies the Marčenko-Pastur law with parameter γ in the large-n limit. In fact, we have the stronger result (see Proposition 3.1 below) that

$$\|W_i - \mathbb{I}\| \to \gamma + 2\sqrt{\gamma} \quad \text{a.s.}$$

A straightforward estimator of Σ in this setting is the arithmetic mean of the N matrices,

$$\mathbf{A} := \frac{\sum_{i=1}^N W_i}{N},$$

which can be equivalently expressed as

$$\mathbf{A} = \frac{\left[X_1, \cdots, X_N\right]\left[X_1, \cdots, X_N\right]^*}{Nn}.$$

The arithmetic mean is a sample covariance based on a total of $T = nN$ observations in this case, since we center by the known rather than estimated observation mean, and every covariance matrix is based

on the same number of observations. Note that in the present work we assume that the observed data are mean-0, and thus there is no need to center the observations about a sample mean. This assumption comes with minimal loss of generality, since the centered sample covariance matrix of a collection of normal random variables is Wishart distributed with parameter $n-1$ in place of n, which has no effect on the asymptotic analyses below.

Remark 2.3. In practical applications, there are situations where pooling observations is not appropriate, and the arithmetic mean may be ill-suited to estimating Σ as a result. For example, in resting state fMRI data, pooling observations from different subjects at a given brain region is infeasible, as the response signals at a particular brain location are not time-aligned across subjects. Nonetheless, combining sample covariance or correlation matrices across subjects via some other procedure may still be valid for estimating the population covariance or correlation matrix.

Throughout this paper, T will denote the total number of observations of points in p-dimensional space. The regime of interest is that in which $p/T \to \Gamma$, and we will consider $T = nN$ where N is a fixed number of matrices, and n will be tending to infinity with p. It will be convenient to define $\gamma = \lim p/n$, which satisfies $\Gamma = \gamma/N$. In this setting (see Proposition 3.1),

$$\|\mathbf{A} - \mathbb{I}\| \to \Gamma + 2\sqrt{\Gamma} \qquad \text{a.s.}$$

That is, the arithmetic mean is not a consistent estimator of Σ, even in the simple case where $\Sigma = \mathbb{I}$.

As an alternative to the arithmetic mean $\mathbf{A}$, we can consider the matrix harmonic mean

$$\mathbf{H} := N\left(\sum_{i=1}^{N} W_i^{-1}\right)^{-1}, \tag{2.2}$$

provided that $n > p$ (so that the W_i are invertible almost surely). In past work [41], the first author analyzed the behavior of the harmonic mean of a collection of independent Wishart random matrices in the regime $p/n \to \gamma \in (0,1)$. While the harmonic mean is also inconsistent as an estimator for Σ, we will see below that its operator-norm bias is, under certain conditions, smaller than that of the arithmetic mean seen above.

3. Improved operator norm error of the harmonic mean

When the W_i are drawn from the same underlying population, the harmonic mean can be a better estimate of the population mean Σ in operator norm than the arithmetic mean [41]. This improvement is best understood as a data-splitting result,in which we partition a sample of p-dimensional observations and compute the harmonic mean of the covariance estimates computed from each part. This is certainly counter-intuitive, but we remind the reader that our intuitions are often wrong in the high-dimensional regime.

Let D be a set of T points in $\mathbb{R}^p$ and let $\mathcal{P}$ be a partition of D into $N \geq 2$ disjoint subsets D_i,

$$\mathcal{P} := \{D_i\}_{i=1}^N \quad \text{such that} \quad D = \bigcup_{i=1}^{N} D_i.$$

Define the Wishart random matrix associated with each sujbset D_i as

$$W(D_i) := \frac{1}{|D_i|} \sum_{x \in D_i} xx^*,$$

and define the arithmetic and harmonic means associated with $\mathcal{P}$ as, respectively,

$$\mathbf{A}(\mathcal{P}) := \frac{1}{N}\sum_{D_i\in\mathcal{P}} W(D_i) \text{ and } \mathbf{H}(\mathcal{P}) := N\Big(\sum_{D_i\in\mathcal{P}} W(D_i)^{-1}\Big)^{-1},$$

provided that $W(D_i)$ is invertible for all i.

If the sets making up the partition $\mathcal{P}$ are all of the same size, then $\mathbf{A}(\mathcal{P})$ is in fact simply the sample covariance of the vectors in D and does not depend on $\mathcal{P}$. The convergence of the spectrum of $\mathbf{A}(\mathcal{P})$ is classical [44]. The statement as given here can be found in [3, Theorems 3.6–7 and Theorems 5.9–10].

Proposition 3.1 (Marčenko-Pastur law). *Suppose D is a collection of T i.i.d. p-dimensional real or complex Gaussians with zero mean and covariance $\mathbb{I}$, with p and T tending to infinity in such a way that $p/T \to \Gamma \in (0, 1/2)$, and let $\mathcal{P}$ be a deterministic sequence of partitions of D such that $|D_i|$ are equal for all $D_i \in \mathcal{P}$. The spectral measure of $\mathbf{A}(\mathcal{P})$ converges weakly almost surely to the measure with density*

$$\frac{1}{2\pi\Gamma x}\sqrt{(S_+ - x)(x - S_-)}\mathbf{1}_{[S_-,S_+]}(x),$$

where

$$S_\pm = \left(1 \pm \sqrt{\Gamma}\right)^2. \tag{3.1}$$

Further, we have the convergence

$$\|\mathbf{A}(\mathcal{P}) - \mathbb{I}\| \to \Gamma + 2\sqrt{\Gamma} \quad \text{a.s.}$$

The following result describes the limiting behaviour of $\mathbf{H}(\mathcal{P})$ under similar conditions.

Proposition 3.2. *Let D be a collection of T i.i.d. p-dimensional real or complex Gaussians with zero mean and covariance $\mathbb{I}$, with p and T tending to infinity in such a way that*

$$\left|\frac{p}{T} - \Gamma\right| \le \frac{K}{p^2}$$

where $K > 0$ is a constant and $\Gamma \in (0, 1/2)$. Let $N \ge 2$ be fixed with T divisible by N, and let $\mathcal{P}$ be a deterministic sequence of partitions of D of size $N \le \lfloor \Gamma^{-1} \rfloor$ such that $|D_i|$ are equal for all $D_i \in \mathcal{P}$. The spectral measure of $\mathbf{H}(\mathcal{P})$ converges weakly almost surely to the measure with density

$$\frac{1}{2\pi\Gamma x}\sqrt{(E_+ - x)(x - E_-)}\mathbf{1}_{[E_-,E_+]}(x),$$

where

$$E_\pm := 1 - (N-2)\Gamma \pm 2\sqrt{\Gamma}\sqrt{1 - (N-1)\Gamma}. \tag{3.2}$$

Further, we have the convergence

$$\|\mathbf{H}(\mathcal{P}) - \mathbb{I}\| \to (N-2)\Gamma + 2\sqrt{\Gamma}\sqrt{1 - (N-1)\Gamma} \quad \text{a.s.}$$

Proof. For the complex Gaussian case, the above result is a restatement of [41, Theorem 2.1], since if the $D_1, D_2, \ldots, D_N$ are disjoint and equal sized, then $W(D_i)$ are a collection of N growing Wishart matrices of the same dimension p and shared parameter n, with $p/n \to \gamma = N\Gamma$. As discussed in [41,

Remark 3], the extension of this result to the real Gaussian setting requires the strong asymptotic freeness of real Wishart random matrices, which was established in recent work [27]. Details are supplied in Appendix B. □

We remark that the case where $|D_i|$ are permitted to vary in i, while still feasibly handled by the tools of the paper [41], is more complicated. The limiting spectrum of the harmonic mean in this setting depends on the roots of a high-degree polynomial, whence comparison of the harmonic and arithmetic mean requires a more subtle analysis. In contrast, when the cells of the partition are of equal size, the limiting spectral measure of the harmonic mean is characterized by the roots of a quadratic and thus admits an explicit solution. Thus, for the sake of concreteness and simplicity, we will assume that $\mathcal{P}$ is a partition with cells of equal size for the remainder of the paper. With the interpretation of $\mathbf{H}(\mathcal{P})$ as a mean formed by splitting D into equal parts, we have the following Theorem.

Theorem 3.3. *Under the assumptions of Proposition 3.2, the operator norm $\|\mathbf{H}(\mathcal{P}) - \mathbb{I}\|$ is minimized for a partition $\mathcal{P}$ of size $N = 2$. Further, for such a partition,*

$$\lim_{p,T\to\infty} \|\mathbf{H}(\mathcal{P}) - \mathbb{I}\| = 2\sqrt{\Gamma}\sqrt{1-\Gamma} < \lim_{p,T\to\infty} \|\mathbf{A}(\mathcal{P}) - \mathbb{I}\| = \Gamma + 2\sqrt{\Gamma}.$$

Proof. The function

$$f(x) = (x-2)\sqrt{\Gamma} + 2\sqrt{\Gamma}\sqrt{1-(x-1)\Gamma}, \qquad x < 1 + \frac{1}{\Gamma}$$

has derivative

$$f'(x) = \sqrt{\Gamma} - \frac{\Gamma\sqrt{\Gamma}}{\sqrt{1+(x-1)\Gamma}}$$

which is greater than 0 whenever

$$x > 1 + \Gamma - \frac{1}{\Gamma}.$$

For $\Gamma \in (0, 1/2)$, this region includes the point $x = 2$ so that the minimizer of $f(x)$ on the interval $[2, \infty)$ is 2. □

We note that for $N = 2$, $E_+ = 1 + 2\sqrt{\Gamma}\sqrt{1-\Gamma} < S_+$, so that E_+ is closer to 1. That is, at least in the case where the true covariance matrix is the identity, the harmonic mean is shrunk toward the true population covariance when compared with the arithmetic mean. The above result suggests that given a collection D of $T \geq 2p$ observations, it is better asymptotically (as measured in operator norm error) to estimate the covariance by splitting D into two equal parts D_1 and D_2 and computing the harmonic mean of $W(D_1)$ and $W(D_2)$ than it is to directly compute the sample covariance matrix of D. The requirement that the vectors have identity covariance is partially addressed by [41, Corollary 2.1.1], which we restate here.

Proposition 3.4. *Under the same assumptions as Proposition 3.2, let $N = 2$ and suppose Σ is a positive definite matrix such that*

$$\limsup_{p,T\to\infty} \frac{\|\Sigma\|\|\Sigma^{-1}\|\|\mathbf{H}(\mathcal{P}) - \mathbb{I}\|}{\|\mathbf{A}(\mathcal{P}) - \mathbb{I}\|} < 1 \text{ a.s.}$$

Then

$$\limsup_{p,T\to\infty} \frac{\|\sqrt{\Sigma}\mathbf{H}(\mathcal{P})\sqrt{\Sigma} - \Sigma\|}{\|\sqrt{\Sigma}\mathbf{A}(\mathcal{P})\sqrt{\Sigma} - \Sigma\|} < 1 \text{ a.s.}$$

Since multiplying $\mathbf{H}(\mathcal{P})$ by $\sqrt{\Sigma}$ on both sides gives a Wishart model with population covariance Σ (see Remark 5.2 below), the bound in Proposition 3.4 holds so long as the condition number of Σ lies in a certain range. For example, when $\Gamma = 1/4$, Proposition 3.4 requires that the condition number be (asymptotically) smaller than $5/2\sqrt{3} \approx 1.44$ (see Remark 2 in [41] for further details). Thus, in a certain sense, Proposition 3.4 applies in the setting that is the opposite of most results on the spiked covariance model. Namely, the harmonic mean is best suited to the case where the signal is spread over many eigenvectors, with the extreme case of this being the setting where $\Sigma = \mathbb{I}$.

This leaves open the question of whether or not it is reasonable to assume that such a bound holds, given real data. One heuristic for checking this assumption is to simply examine the spectra of $\mathbf{H}(\mathcal{P})$ and $\mathbf{A}(\mathcal{P})$, but of course this runs into a circular problem, in which one must appeal to concentration of eigenvalues in order to motivate a result on operator-norm concentration. Ultimately, the decision as to whether or not the condition number bound required by Proposition 3.4 is reasonable lies with the practitioner. Nonetheless, an appealing approach would be to develop a method for estimating the population condition number rather than the full spectrum. This would in turn give a good indication of which of the arithmetic or harmonic matrix means are better suited to the observed data. We leave the further exploration of this line to future work.

Beyond the above operator norm results, the harmonic mean has an interesting additional property with respect to the Frobenius norm. The arithmetic matrix mean is usually motivated as minimizing the squared Frobenius norm error. A similar objective motivates many existing shrinkage estimators for the covariance matrix [28,36]. Under the setting considered above, the harmonic matrix mean, despite not being optimized for this loss, *matches* the Frobenius norm error of the arithmetic mean asymptotically.

Lemma 3.5. *Under the conditions of Proposition 3.2, when $N = 2$ we have*

$$\lim_{p,T\to\infty} \frac{1}{p}\|\mathbf{H}(\mathcal{P}) - \mathbb{I}\|_F^2 = \lim_{p,T\to\infty} \frac{1}{p}\|\mathbf{A}(\mathcal{P}) - \mathbb{I}\|_F^2 = \Gamma \text{ a.s.}$$

Proof. Since $\mathbf{H}(\mathcal{P}) - \mathbb{I}$ is symmetric, by the almost sure weak convergence of $\mathbf{H}(\mathcal{P})$, it suffices to show

$$\lim_{p,T\to\infty} \frac{1}{p} \operatorname{Tr}\left[(\mathbf{H}(\mathcal{P}) - \mathbb{I})^2\right] \to \int_{E_-}^{E_+} (x-1)^2 \frac{\sqrt{(E_+ - x)(x - E_-)}}{2\pi\Gamma x}\, \mathrm{d}x,$$

where $E_\pm$ are defined in Equation (3.2), and compare with

$$\lim_{p,T\to\infty} \frac{1}{p} \operatorname{Tr}\left[(\mathbf{A}(\mathcal{P}) - \mathbb{I})^2\right] \to \int_{S_-}^{S_+} (x-1)^2 \frac{\sqrt{(S_+ - x)(x - S_-)}}{2\pi\Gamma x}\, \mathrm{d}x,$$

where $S_\pm$ are defined in Equation (3.1). Note that

$$\frac{E_+ - E_-}{E_+ + E_-} = 2\sqrt{\Gamma}\sqrt{1-\Gamma}, \qquad \frac{E_+ + E_-}{2} = 1$$

and

$$\frac{S_+ - S_-}{S_+ + S_-} = \frac{2\sqrt{\Gamma}}{1+\Gamma}, \qquad \frac{S_+ + S_-}{2} = 1 + \Gamma.$$

Using Lemma A.1 in the Appendix,

$$\int_{E_-}^{E_+} (x^2 - 2x + 1) \frac{\sqrt{(E_+ - x)(x - E_-)}}{2\pi\Gamma x} \, \mathrm{d}x = 1 - \Gamma - 2(1 - \Gamma) + 1 = \Gamma,$$

while

$$\int_{S_-}^{S_+} (x^2 - 2x + 1) \frac{\sqrt{(S_+ - x)(x - S_-)}}{2\pi\Gamma x} \, \mathrm{d}x = 1 + \Gamma - 2 + 1 = \Gamma. \qquad \square$$

4. Rao-Blackwell improvement of the harmonic mean

The results in Section 3 are somewhat unexpected, and raise the question of whether other matrix means have similar properties. Analyzing other matrix means such as the geometric mean or more complicated Frechét means [7,53] under the high-dimensional regime poses a significant challenge since these operations currently fall outside the scope of known results in free probability theory. Random matrix techniques do, however, allow us to extend our analysis of the harmonic mean by computing its expectation conditioned on the arithmetic mean. By the Rao-Blackwell Theorem, using this conditional expectation as an estimator yields an expected spectral norm error no worse than the unconditioned harmonic mean.

In this section we restrict ourselves to the model of Definition 2.1, so as to ensure the availability of explicit integrals for our quantities of interest. We expect the same results can be established for the complex model in Definition 2.2, but doing so would require reworking the results of [33] (restated in the Appendix for ease of reference) for the complex Wishart ensemble, which is outside the scope of the present article. Further, we restrict our attention to $T = 2n \geq p$ to facilitate comparison with the $N = 2$ case studied in the previous section. To this end, let $\mathcal{P} = D_1 \cup D_2$ where D_1 and D_2 are disjoint, and

$$W_1 := W(D_1) = \frac{1}{|D_1|} \sum_{x \in D_1} xx^* = \frac{1}{n} \sum_{x \in D_1} xx^*,$$

$$W_2 := W(D_2) = \frac{1}{|D_2|} \sum_{x \in D_2} xx^* = \frac{1}{n} \sum_{x \in D_2} xx^*.$$

The matrices W_1 and W_2 have densities [2, Theorem 7.2.2]

$$f_{W_i}(w_i) = C_{n,p} \det(w_i)^{\frac{1}{2}(n-p-1)} \exp\left(-\frac{n}{2} \operatorname{Tr} \Sigma^{-1} w_i \right),$$

where

$$C_{n,p} = \frac{n^{\frac{pn}{2}}}{2^{\frac{1}{2}np} \det(\Sigma)^{\frac{n}{2}} \Gamma_p\left(\frac{n}{2}\right)}, \; \Gamma_p(x) = \pi^{\frac{p(p-1)}{4}} \prod_{i=1}^{p} \Gamma\left(x - \frac{i-1}{2}\right).$$

These densities are supported on the space $\mathcal{S}_p(\mathbb{R})$ of $p \times p$ symmetric positive definite real matrices. As before, define

$$\mathbf{A} = \frac{W_1 + W_2}{2},$$

$$\mathbf{H} = 2\left(W_1^{-1} + W_2^{-1}\right)^{-1}$$

and note that

$$\mathbf{A} = \frac{1}{2n} \sum_{x \in D} xx^*$$

is a Wishart random matrix with parameter Σ and $2n$, and thus

$$f_{\mathbf{A}}(a) := C_{2n,p} \det(a)^{\frac{1}{2}(2n-p-1)} \exp(-n \operatorname{Tr} \Sigma^{-1} a), \tag{4.1}$$

where a takes values in all of $\mathcal{S}_p(\mathbb{R})$.

Recall that the matrix $\mathbf{A}$ is a sufficient statistic for the covariance matrix Σ, and note that the loss function

$$\ell(M, \Sigma) := \|M - \Sigma\|,$$

is convex in the variable M. By the Rao-Blackwell Theorem [50, 5a.2 (ii)], we have

$$\mathbb{E}\, \ell\big(\mathbb{E}[\mathbf{H}|\mathbf{A}], \Sigma\big) \leq \mathbb{E}\, \ell\big(\mathbf{H}, \Sigma\big),$$

which is to say that as an estimator, $\mathbf{H}$ is outperformed by the conditional expectation $\mathbb{E}[\mathbf{H}|\mathbf{A}]$, which we now compute.

Observe that the harmonic mean satisfies [7, Section 4.1]

$$\mathbf{H} = 2W_1 - W_1 \mathbf{A}^{-1} W_1,$$
$$\mathbf{H} = 2W_2 - W_2 \mathbf{A}^{-1} W_2.$$

Averaging these two equations gives

$$\mathbf{H} = 2\mathbf{A} - \frac{1}{2} W_1 \mathbf{A}^{-1} W_1 - \frac{1}{2} W_2 \mathbf{A}^{-1} W_2,$$

and taking the conditional expectation yields

$$\mathbb{E}[\mathbf{H}|\mathbf{A}] = 2\mathbf{A} - \frac{1}{2} \mathbb{E}[W_1 \mathbf{A}^{-1} W_1 | \mathbf{A}] - \frac{1}{2} \mathbb{E}[W_2 \mathbf{A}^{-1} W_2 | \mathbf{A}].$$

To compute the matrix-valued integrals

$$\mathbb{E}[W_1 \mathbf{A}^{-1} W_1 | \mathbf{A}] \quad \text{and} \quad \mathbb{E}[W_2 \mathbf{A}^{-1} W_2 | \mathbf{A}],$$

we proceed by directly computing the conditional density of W_i given $\mathbf{A}$.

We begin with the joint density of W_1 and W_2:

$$f_{W_1}(w_1) f_{W_2}(w_2) = C_{n,p}^2 \det(w_1)^{\frac{1}{2}(n-p-1)} \det(w_2)^{\frac{1}{2}(n-p-1)} \exp\left[-\frac{n}{2} \operatorname{Tr}\left\{ \Sigma^{-1}(w_1 + w_2) \right\} \right].$$

We will use this formula to obtain an expression for the joint density of W_1 and $\mathbf{A}$. For a symmetric matrix M with entries $m_{i,j}$, let $\mathrm{d}m_{i,j}$ denote Lebesgue measure over that entry and define

$$(\mathrm{d}M) := \bigwedge_{1 \leq i \leq j \leq p} \mathrm{d}m_{i,j},$$

that is $(\mathrm{d}M)$ is the volume form of the matrix M. The "shear" transformation

$$(w_1, w_2) \mapsto (w_1, a), \quad a := \frac{w_1 + w_2}{2},$$

maps the domain $\mathcal{S}_p(\mathbb{R}) \times \mathcal{S}_p(\mathbb{R})$ to the region

$$\{M \in \mathcal{S}_p(\mathbb{R}) : 0 \preceq M \preceq 2a\} \times \mathcal{S}_p(\mathbb{R}),$$

where we remind the reader that $\preceq$ denotes the positive semidefinite ordering. The Jacobian of this mapping is

$$(\mathrm{d}w_1) \wedge (\mathrm{d}w_2) = 2^{\frac{p(p+1)}{2}} (\mathrm{d}w_1) \wedge (\mathrm{d}a),$$

hence the joint density is

$$f_{W_1,\mathbf{A}}(w_1, a) = C_{n,p}^2 2^{\frac{p(p+1)}{2}} \det(w_1)^{\frac{1}{2}(n-p-1)} \det(2a - w_1)^{\frac{1}{2}(n-p-1)} \exp[-n \operatorname{Tr} \Sigma^{-1} a].$$

To obtain the conditional distribution, we divide by (4.1), yielding

$$f_{W_1|\mathbf{A}}(w_1|a) = \frac{C_{n,p}^2}{C_{2n,p}} \frac{\det(w_1)^{\frac{n-p-1}{2}} \det(2a - w_1)^{\frac{n-p-1}{2}}}{\det(2a)^{n - \frac{p+1}{2}}}, \tag{4.2}$$

where w_1 is supported on the region

$$\mathcal{D}(a) := \{m \in \mathcal{S}_p(\mathbb{R}) : 0 \preceq m \preceq 2a\}.$$

Evaluating this density at $a = \mathbf{A}$, for $w_1 \in \mathcal{D}(\mathbf{A})$ yields

$$f_{W_1|\mathbf{A}}(w_1|\mathbf{A}) := \frac{C_{n,p}^2}{C_{2n,p}} \frac{\det(w_1)^{\frac{n-p-1}{2}} \det(2\mathbf{A} - w_1)^{\frac{n-p-1}{2}}}{\det(2\mathbf{A})^{\frac{2n-p-1}{2}}},$$

a multivariate Beta distribution $B(p; n, n; 2\mathbf{A})$ (see Definition A.2 in the Appendix). With this notation, we have

$$\mathbb{E}[W_1 \mathbf{A}^{-1} W_1 | \mathbf{A}] = \int_{\mathcal{D}(\mathbf{A})} w_1 \mathbf{A}^{-1} w_1 \, f_{W_1|\mathbf{A}}(w_1|\mathbf{A}) (\mathrm{d}w_1)$$

the integration over w_1 can be done using Theorem A.3 in the Appendix, with $n_1 = n$, $n_2 = n$, and setting $\Delta = 2\mathbf{A}$ yields the following Lemma.

Lemma 4.1. *For any F that is a function of $\mathbf{A}$ taking value in the space of $p \times p$ matrices,*

$$\mathbb{E}[W_1 F W_1 | \mathbf{A}] = \frac{\{n(2n+1) - 2\}\mathbf{A}F\mathbf{A} + n\{(\mathbf{A}F\mathbf{A})^\top + \operatorname{Tr}(\mathbf{A}F)\mathbf{A}\}}{(2n-1)(n+1)}.$$

Setting $F = \mathbf{A}^{-1}$ yields

$$\mathbb{E}[W_1 \mathbf{A}^{-1} W_1 | \mathbf{A}] = \frac{2n(n+1) - 2 + pn}{(2n-1)(n+1)} \mathbf{A}.$$

The same calculation can be carried out for $\mathbb{E}[W_2\mathbf{A}^{-1}W_2|\mathbf{A}]$ to give

$$\mathbb{E}[\mathbf{H}|\mathbf{A}] = 2\mathbf{A} - \left\{\frac{2n(n+1)-2+pn}{(2n-1)(n+1)}\right\}\mathbf{A} = \frac{n(2n-p)}{(2n-1)(n+1)}\mathbf{A},$$

which is simply a rescaling of $\mathbf{A}$ by a deterministic constant. We summarize this result as a theorem.

Theorem 4.2. *Let $T = 2n$ and D be as in Definition 2.1. If $\mathcal{P}$ is a partition of size 2 with $|D_1| = |D_2| = n$, then*

$$\mathbb{E}[\mathbf{H}(\mathcal{P})|\mathbf{A}(\mathcal{P})] = \frac{n(2n-p)}{(2n-1)(n+1)}\mathbf{A}(\mathcal{P}).$$

Note that as $p/T = p/(2n) \to \Gamma \in (0,1/2)$, the limiting spectral measure of the above conditional expectation converges to

$$(1-\Gamma)Z,$$

where Z is a random variable distributed according to the limiting spectral distribution of $\mathbf{A}$.

The above calculations can be further extended by making a few adjustments to the matrices W_1, W_2. A number of matrix estimators take the form

$$\tilde{\mathbf{A}} := c(\mathbf{A} + d\hat{\Lambda}),$$

where c, d are positive scalars and $\hat{\Lambda}$ is a positive semidefinite matrix, all depending only on $\mathbf{A}$. Estimators of this form have been extensively studied in the covariance estimation literature [28,34,36]. One could take the extra step of applying the same regularization procedure to the matrices W_1 and W_2 before computing their (Rao-Blackwellized) harmonic mean. Suppose we replace W_1 and W_2 with

$$\tilde{W}_1 := c(W_1 + d\hat{\Lambda}) \quad \text{and} \quad \tilde{W}_2 := c(W_2 + d\hat{\Lambda}),$$

respectively. Letting $\tilde{\mathbf{H}}$ be the harmonic mean of $\tilde{W}_1$ and $\tilde{W}_2$, we can compute a Rao-Blackwell improvement of $\tilde{\mathbf{H}}$ in much the same way that we did for $\mathbf{H}$ above. Indeed, we still have

$$\tilde{\mathbf{H}} = 2\tilde{\mathbf{A}} - \frac{\tilde{W}_1\tilde{\mathbf{A}}^{-1}\tilde{W}_1}{2} - \frac{\tilde{W}_2\tilde{\mathbf{A}}^{-1}\tilde{W}_2}{2}.$$

We can compute the conditional expectation with respect to $\mathbf{A}$ as follows

$$\begin{aligned}\mathbb{E}[\tilde{\mathbf{H}}|\mathbf{A}] = 2\tilde{\mathbf{A}} &- \frac{c^2}{2}\big(\mathbb{E}[W_1\tilde{\mathbf{A}}^{-1}W_1|\mathbf{A}] + \mathbb{E}[W_2\tilde{\mathbf{A}}^{-1}W_2|\mathbf{A}]\big) \\ &- c^2 d\big(\mathbf{A}\tilde{\mathbf{A}}^{-1}\hat{\Lambda} + \hat{\Lambda}\tilde{\mathbf{A}}^{-1}\mathbf{A}\big) - (cd)^2\hat{\Lambda}\tilde{\mathbf{A}}^{-1}\hat{\Lambda}.\end{aligned} \tag{4.3}$$

Using Lemma 4.1, we have

$$\begin{aligned}&\frac{c^2}{2}\big(\mathbb{E}[W_1\tilde{\mathbf{A}}^{-1}W_1|\mathbf{A}] + \mathbb{E}[W_2\tilde{\mathbf{A}}^{-1}W_2|\mathbf{A}]\big) \\ &\qquad = c^2\left[\frac{\{2n(n+1)-2\}\mathbf{A}\tilde{\mathbf{A}}^{-1}\mathbf{A} + n\operatorname{Tr}(\mathbf{A}\tilde{\mathbf{A}}^{-1})\mathbf{A}}{(2n-1)(n+1)}\right].\end{aligned} \tag{4.4}$$

Combining Equations (4.3) and (4.4), we have

$$\mathbb{E}[\tilde{\mathbf{H}}|\mathbf{A}] = 2\tilde{\mathbf{A}} - c^2\left[\frac{\{2n(n+1)-2\}\mathbf{A}\tilde{\mathbf{A}}^{-1}\mathbf{A} + n\,\mathrm{Tr}(\mathbf{A}\tilde{\mathbf{A}}^{-1})\mathbf{A}}{(2n-1)(n+1)}\right] - c^2 d\left(\mathbf{A}\tilde{\mathbf{A}}^{-1}\hat{\Lambda} + \hat{\Lambda}\tilde{\mathbf{A}}^{-1}\mathbf{A}\right) - (cd)^2\hat{\Lambda}\tilde{\mathbf{A}}^{-1}\hat{\Lambda},$$

which we can write solely in terms of $\tilde{\mathbf{A}}$ and $\hat{\Lambda}$ by substituting $c\mathbf{A}$ with $\tilde{\mathbf{A}} - cd\hat{\Lambda}$, obtaining

$$\mathbb{E}[\tilde{\mathbf{H}}|\mathbf{A}] = 2\tilde{\mathbf{A}} - \frac{[2n(n+1)-2][\tilde{\mathbf{A}} - 2cd\hat{\Lambda} + (cd\hat{\Lambda})\tilde{\mathbf{A}}^{-1}(cd\hat{\Lambda})]}{(2n-1)(n+1)} - \frac{[np - ncd\,\mathrm{Tr}(\hat{\Lambda}\tilde{\mathbf{A}}^{-1})](\tilde{\mathbf{A}} - cd\hat{\Lambda})}{(2n-1)(n+1)} - 2cd\hat{\Lambda} + (cd\hat{\Lambda})\tilde{\mathbf{A}}^{-1}(cd\hat{\Lambda}).$$

We summarize the above results in the following Theorem.

Theorem 4.3. *The Rao-Blackwell improvement of* $\tilde{\mathbf{H}}$*, the harmonic mean of the regularized matrices* $c(W_1 + d\hat{\Lambda})$ *and* $c(W_2 + d\hat{\Lambda})$*, where* c *and* d *are positive constants that only depend on* $\mathbf{A}$ *and* $\hat{\Lambda}$ *is a positive definite matrix that only depends on* $\mathbf{A}$ *given by*

$$\mathbb{E}[\tilde{\mathbf{H}}|\mathbf{A}] = n\left[\frac{2n - p + cd\,\mathrm{Tr}(\hat{\Lambda}\tilde{\mathbf{A}}^{-1})}{(2n-1)(n+1)}\right]\tilde{\mathbf{A}} + \left[\frac{-n+1}{(2n-1)(n+1)}\right](cd\hat{\Lambda})\tilde{\mathbf{A}}^{-1}(cd\hat{\Lambda}) + \left[\frac{2n + np - 2 - ncd\,\mathrm{Tr}(\hat{\Lambda}\tilde{\mathbf{A}}^{-1})}{(2n-1)(n+1)}\right]cd\hat{\Lambda}.$$

Remark 4.4. For linear shrinkage estimators of the form

$$\tilde{\mathbf{A}} = (1-\lambda)\mathbf{A} + \lambda\mathbb{I},$$

as in [36], setting $\hat{\Lambda} = \mathbb{I}$, $c = (1-\lambda)$, and $cd = \lambda$ in our formula gives

$$\mathbb{E}[\tilde{\mathbf{H}}|\mathbf{A}] = n\left[\frac{2n - p + \lambda\,\mathrm{Tr}(\tilde{\mathbf{A}}^{-1})\}}{(2n-1)(n+1)}\right]\tilde{\mathbf{A}} + \left[\frac{-n+1}{(2n-1)(n+1)}\right]\lambda^2\tilde{\mathbf{A}}^{-1} + \left[\frac{2n + np - 2 - n\lambda\,\mathrm{Tr}(\tilde{\mathbf{A}}^{-1})}{(2n-1)(n+1)}\right]\lambda\mathbb{I}. \tag{4.5}$$

The results outlined above are unexpected, and somewhat odd. The implication of Theorem 4.2 is that the Rao-Blackwellization of $\mathbf{H}(\mathcal{P})$ is a deterministic constant multiple of $\mathbf{A}(\mathcal{P})$. This suggests that the expense of computing $\mathbf{H}(\mathcal{P})$ is not warranted, since while $\mathbf{H}(\mathcal{P})$ may improve upon $\mathbf{A}(\mathcal{P})$ as an estimator, a scalar multiple of $\mathbf{A}(\mathcal{P})$ improves still further upon $\mathbf{H}(\mathcal{P})$. Figure 4 in Section 6 explores this point empirically in the finite-sample regime. On the other hand, the form of the Rao-Blackwellized estimator in Equation (4.5), obtained from Theorem 4.3, bears noting. In contrast to the Rao-Blackwellized version of $\mathbf{H}$ considered in Theorem 4.2, this estimator involves a linear combination of $\tilde{\mathbf{A}}$, $\tilde{\mathbf{A}}^{-1}$ and $\mathbb{I}$. As a result, this estimator differs from the linear shrinkage estimators considered elsewhere in the literature [28,36,57]. The estimator in Theorem 4.3, which has not been proposed previously to the best of our knowledge, falls under the heading of orthogonally invariant estimators, as discussed in [22]. Thus, in particular, there should exist some orthogonally invariant loss function for which the estimator in Equation (4.5) is the optimal estimator. We leave further study of this estimator and its properties to future work.

5. Eigenvector recovery

A major motivation for working with the operator norm is to obtain guarantees on convergence of eigenvectors, which are often the main object of interest in covariance estimation, as in when the covariance is used for principal component analysis. This is done via the Davis-Kahan theorem, which bounds the distance between the leading eigenvectors $v_1(\hat{\Sigma})$ and $v_1(\Sigma)$ in terms of $\|\hat{\Sigma} - \Sigma\|$. For example, it can be shown [61, Corollary 1] that

$$\|v_1(\hat{\Sigma}) - v_1(\Sigma)\| \leq \frac{2^{\frac{3}{2}}\|\hat{\Sigma} - \Sigma\|}{\lambda_1(\Sigma) - \lambda_2(\Sigma)} \quad \text{if} \quad \langle v_1(\hat{\Sigma}), v_1(\Sigma)\rangle > 0 \tag{5.1}$$

In this section, we show that under a spiked covariance model, the leading eigenvector of $\mathbf{H}(\mathcal{P})$ carries information about the leading eigenvector of the population covariance matrix Σ in the regime $p/T \to \Gamma \in (0, 1/2)$, and we compare its performance with that of the leading eigenvector of $\mathbf{A}(\mathcal{P})$.

Definition 5.1. Let D be a set of T i.i.d. p-dimensional centered multivariate real or complex Gaussians with population covariance matrix

$$\Sigma = \mathbb{I} + \theta v v^*, \tag{5.2}$$

where $\theta > 0$ and v is a p-dimensional (real or complex) unit-norm vector. As in Proposition 3.2 assume that

$$\left|\frac{p}{T} - \Gamma\right| \leq \frac{K}{p^2},$$

where $\Gamma \in (0, 1/2)$ and $K > 0$ is a constant that does not depend on p or T.

Remark 5.2. Let D^0 be a collection of multivariate real or complex Gaussians with zero mean and covariance $\mathbb{I}$. If we define

$$D = \{\sqrt{\Sigma}x : x \in D^0\} =: \sqrt{\Sigma}D^0,$$

where Σ is given in Definition 5.1, then D has the same distribution as the model in Equation (5.2). Moreover, by this same transformation, we may take a partition $\mathcal{P}^0$ of D^0 and generate a partition $\mathcal{P}$ of D by replacing each D_i^0 in $\mathcal{P}^0$ by $D_i := \sqrt{\Sigma}D_i^0$. With this definition we have the equality

$$\mathbf{H}(\mathcal{P}) = \sqrt{\Sigma}\mathbf{H}(\mathcal{P}^0)\sqrt{\Sigma} \quad \text{and} \quad \mathbf{A}(\mathcal{P}) = \sqrt{\Sigma}\mathbf{A}(\mathcal{P}^0)\sqrt{\Sigma}.$$

The Theorem below follows from well-established results in the literature and can be generalized without any change to higher-rank perturbations of the identity. We focus here on the simple case of one spike in order to get more explicit insight into the performance of $\mathbf{H}(\mathcal{P})$.

Theorem 5.3. *Let D be a spiked model as in Definition 5.1, and suppose $\mathcal{P}$ is a partition satisfying the conditions of Proposition 3.2, with $N = 2$. Then we have the almost sure convergence*

$$\lambda_1\{\mathbf{H}(\mathcal{P})\} \to \begin{cases} 1 + \dfrac{\Gamma}{\theta} + (1-\Gamma)\theta & \text{if } \theta > \sqrt{\dfrac{\Gamma}{1-\Gamma}} \\ 1 + 2\sqrt{\Gamma}\sqrt{1-\Gamma} & \text{otherwise,} \end{cases}$$

and

$$\left|\left\langle v_1\{\mathbf{H}(\mathcal{P})\}, v\right\rangle\right|^2 \to \begin{cases} \dfrac{\theta+1}{\theta}\dfrac{\theta^2(1-\Gamma)-\Gamma}{\theta^2(1-\Gamma)+\theta+\Gamma} & \text{if } \theta > \sqrt{\dfrac{\Gamma}{1-\Gamma}}, \\ 0 & \text{otherwise.} \end{cases}$$

Proof. To prove this result we will use the general framework [5], which considers multiplicative spikes of the form

$$\tilde{M} := \sqrt{\mathbb{I} + \theta v v^*} M \sqrt{\mathbb{I} + \theta v v^*}.$$

Here θ and v are as in Definition 5.1 and M is a symmetric (or Hermitian if M has complex entries) matrix whose eigenvalue distribution converges weakly to a spectral measure ν almost surely, and ν is supported on the interval $[a, b]$. Assume further that the convergence of the largest (smallest) eigenvalue of M is the right (left) edge of the support of ν and that the distribution of M is invariant under conjugation by an orthogonal matrix (unitary if M has complex entries). Recall that this implies that the matrix of eigenvectors of M is Haar distributed on the orthogonal group (unitary group for the complex case). Define for $z \in \mathbb{C}\backslash[a, b]$

$$m_\nu(z) := \int_{\mathbb{R}} \frac{\nu(\mathrm{d}x)}{z - x},$$
$$t_\nu(z) := \int_{\mathbb{R}} \frac{x\nu(\mathrm{d}x)}{z - x} = -1 + z m_\nu(z),$$

and let $t_\nu^{-1}(z)$ be the functional inverse of t_ν. By [5, Theorem 2.7] we have the almost sure convergence

$$\lambda_1(\tilde{M}) \to \begin{cases} t_\nu^{-1}\left(\dfrac{1}{\theta}\right) & \theta > \dfrac{1}{t_\nu(b^+)}, \\ b & \text{otherwise.} \end{cases}$$

Here $t_\nu(b^+)$ is the limit as $z \to b$ of $t_\nu(z)$, well defined when ν has a density with square root decay near b [5, Proposition 2.10]. Furthermore, by [5, Remark 2.11] we have the almost sure convergence

$$|\langle v_1(\tilde{M}), v\rangle|^2 \to \begin{cases} -\dfrac{\theta+1}{\theta^2 \rho t_\nu'(\rho)} & \text{if } \theta > 1/t_\nu(b), \\ 0 & \text{otherwise,} \end{cases}$$

where

$$\rho := t_\nu^{-1}\left(\frac{1}{\theta}\right).$$

Applying these results using Remark 5.2 and taking $M = \mathbf{H}(\mathcal{P}^0)$ where $\mathcal{P}^0$ is the partition of a data set D^0 with population covariance $\mathbb{I}$, we see that M satisfies the required convergence properties by Proposition 3.2 and is unitarily invariant. Letting ν equal to the limiting spectral measure of $\mathbf{H}(\mathcal{P}^0)$, and noting for $N = 2$

$$E_\pm = 1 \pm 2\sqrt{\Gamma}\sqrt{1-\Gamma},$$

the proof now proceeds by calculation. From the results of [41, Equation (18)], $m_\nu(z)$ satisfies the fixed point equation

$$\Gamma z m_\nu(z)^2 + (1 - 2\Gamma - z) m_\nu(z) + 1 = 0.$$

Inserting the definition of $t_\nu(z)$ and simplifying yields

$$\Gamma t_\nu(z)^2 + (1-z)t_\nu(z) + 1 - \Gamma = 0.$$

Taking the limit as z goes to E_+ and utilizing the square root decay of ν at E_+ yields

$$\left(t_\nu(E_+) - \sqrt{\frac{1-\Gamma}{\Gamma}}\right)^2 = 0 \implies t_\nu(E_+) = \sqrt{\frac{1-\Gamma}{\Gamma}}.$$

Hence, a phase transition in the largest eigenvalue of $\mathbf{H}(\mathcal{P})$ occurs for $\theta > \sqrt{\Gamma/(1-\Gamma)}$. We can solve for the inverse of $t_\nu(z)$ by substituting $z = t_\nu^{-1}(w)$ into the polynomial fixed point equation for $t_\nu(z)$,

$$t_\nu^{-1}(w) = 1 + \Gamma w + \frac{1-\Gamma}{w}.$$

Assuming $\theta > \sqrt{\Gamma/(1-\Gamma)}$ and inserting $w = 1/\theta$ gives the location of the spiked eigenvalue of $\mathbf{H}(\mathcal{P})$,

$$\rho = t_\nu^{-1}\left(\frac{1}{\theta}\right) = 1 + \frac{\Gamma}{\theta} + (1-\Gamma)\theta.$$

Differentiating the fixed point equation of $t_\nu(z)$ gives the following fixed point equation for $t_\nu'(z)$

$$(2\Gamma t_\nu(z) + 1 - z)t_\nu'(z) - t_\nu(z) = 0.$$

Substituting ρ yields

$$t_\nu'(\rho) = \frac{1}{2\Gamma + \theta(1-\rho)} = \frac{1}{\Gamma - \theta^2(1-\Gamma)},$$

and hence

$$-\frac{\theta+1}{\theta^2 \rho t_\nu'(\rho)} = \frac{\theta+1}{\theta^2}\frac{\theta^2(1-\Gamma)-\Gamma}{1+\frac{\Gamma}{\theta}+(1-\Gamma)\theta},$$

which concludes the proof. □

Remark 5.4. The same analysis has been performed on $\mathbf{A}(\mathcal{P})$ [4,29,48,49]. Under this setting, we have the almost sure convergence [5, Section 3.2]

$$\lambda_1\{\mathbf{A}(\mathcal{P})\} \to \begin{cases} (\theta+1)\left(1+\frac{\Gamma}{\theta}\right) & \text{if } \theta > \sqrt{\Gamma}, \\ \left(1+\sqrt{\Gamma}\right)^2 & \text{otherwise,} \end{cases}$$

and

$$\left|\left\langle v_1\{\mathbf{A}(\mathcal{P})\}, v\right\rangle\right|^2 \to \begin{cases} \frac{1-\frac{\Gamma}{\theta^2}}{1+\frac{\Gamma}{\theta}} & \text{if } \theta > \sqrt{\Gamma}, \\ 0 & \text{otherwise.} \end{cases}$$

When $0 < \Gamma < \frac{1}{2}$, we have $\sqrt{\Gamma/(1-\Gamma)} > \sqrt{\Gamma}$, and it is possible to choose θ such that

$$\sqrt{\Gamma} < \theta \le \sqrt{\frac{\Gamma}{1-\Gamma}}. \tag{5.3}$$

Comparing the phase transition for the harmonic mean given in Theorem 5.3 and that for the arithmetic mean given in Remark 5.4, we see that when θ satisfies Equation (5.3), Theorem 5.3 and Remark 5.4 imply the almost sure convergence

$$\left|\left\langle v_1\left\{\mathbf{H}(\mathcal{P})\right\}, v\right\rangle\right|^2 \to 0,$$

$$\left|\left\langle v_1\left\{\mathbf{A}(\mathcal{P})\right\}, v\right\rangle\right|^2 \to \frac{1-\frac{\Gamma}{\theta^2}}{1+\frac{\Gamma}{\theta}}.$$

This means that for low signal strength θ, $v_1\left\{\mathbf{H}(\mathcal{P})\right\}$ fails to have any relationship with v.

On the other hand, when $\theta > \sqrt{\Gamma/(1-\Gamma)}$ the leading eigenvectors of both $\mathbf{H}(\mathcal{P})$ and $\mathbf{A}(\mathcal{P})$ have some relationship with v in the limit as $p/T \to \Gamma$. We observe that

$$\lim_{p,T\to\infty}\left(\left|\left\langle v_1\left\{\mathbf{A}(\mathcal{P})\right\}, v\right\rangle\right|^2 - \left|\left\langle v_1\left\{\mathbf{H}(\mathcal{P})\right\}, v\right\rangle\right|^2\right)$$

$$= \frac{1-\frac{\Gamma}{\theta^2}}{1+\frac{\Gamma}{\theta}} - \frac{\theta+1}{\theta}\frac{\theta^2(1-\Gamma)-\Gamma}{\theta^2(1-\Gamma)+\theta+\Gamma} = \frac{\Gamma^2(1+\theta)^2}{\left(1+\frac{\Gamma}{\theta}\right)\theta\left\{\theta^2(1-\Gamma)+\theta+\Gamma\right\}} > 0,$$

so that the leading eigenvector of $\mathbf{A}(\mathcal{P})$ functions as a better estimator, asymptotically, for all possible values of θ.

Compare this result with the bound predicted by solely analyzing the upper bounds obtained from the Davis-Kahan Theorem. That is, taking $\lambda_1(\Sigma) - \lambda_2(\Sigma) = \theta$ in Equation (5.1), we have

$$\left\|v_1\left\{\mathbf{H}(\mathcal{P})\right\} - v\right\| \le \frac{2^{\frac{3}{2}}\|\mathbf{H}(\mathcal{P}) - \Sigma\|}{\theta} \quad \text{when} \quad \left\langle v_1\left\{\mathbf{H}(\mathcal{P})\right\}, v\right\rangle > 0, \text{ and}$$

$$\left\|v_1\left\{\mathbf{A}(\mathcal{P})\right\} - v\right\| \le \frac{2^{\frac{3}{2}}\|\mathbf{A}(\mathcal{P}) - \Sigma\|}{\theta} \quad \text{when} \quad \left\langle v_1\left\{\mathbf{A}(\mathcal{P})\right\}, v\right\rangle > 0.$$

Since the condition number of Σ is $1+\theta$, by Proposition 3.4 we have

$$\limsup_{p,T\to\infty} \frac{\|\mathbf{H}(\mathcal{P}) - \Sigma\|}{\|\mathbf{A}(\mathcal{P}) - \Sigma\|} < 1 \text{ whenever } \theta < \frac{2+\sqrt{\Gamma}}{2\sqrt{1-\Gamma}} - 1.$$

In order for $|\langle v_1\{\mathbf{A}(\mathcal{P})\}, v\rangle|^2 \not\to 0$, we would need $\theta > \sqrt{\Gamma}$. Notice that as $\Gamma \to 1/2$,

$$\lim_{\Gamma\to\frac{1}{2}} \frac{2+\sqrt{\Gamma}}{2\sqrt{1-\Gamma}} - 1 = \sqrt{2} - \frac{1}{2} > \lim_{\Gamma\to\frac{1}{2}} \sqrt{\Gamma} = \frac{1}{\sqrt{2}}.$$

It follows that there exist values of θ and Γ for which the Davis-Kahan theorem predicts that the harmonic mean yields better eigenvector recovery than the arithmetic mean, even though $v_1\{\mathbf{H}(\mathcal{P})\}$ has no asymptotic relationship to v while $v_1\{\mathbf{A}(\mathcal{P})\}$ does.

6. Experiments

Our theoretical results presented above are asymptotic, so we complement them here with a brief investigation of the empirical finite-sample performance of the harmonic and arithmetic matrix means

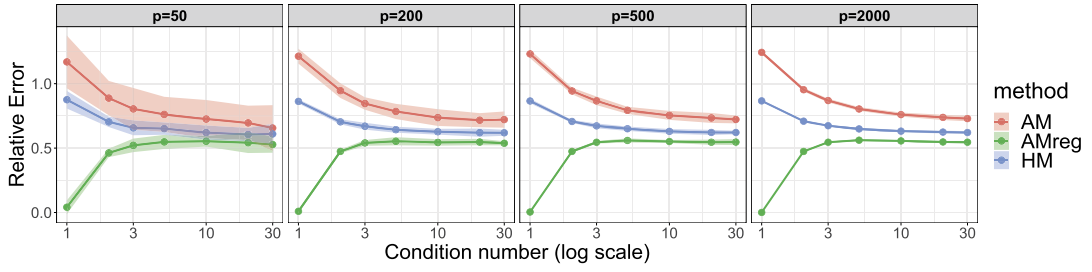

Figure 1. Operator norm relative error $\|\hat{\Sigma} - \Sigma\|/\|\Sigma\|$ as a function of the condition number parameter b for different choices of the data dimension p. The plot compares the arithmetic matrix mean (red), the Fisher-Sun estimator (green) and the harmonic matrix mean (blue). Each point is the mean of 20 independent trials, with the shaded regions indicating two standard deviations.

and related estimators. We begin by comparing the operator norm error of the arithmetic and harmonic matrix means in recovering the true covariance matrix. We generate a random covariance matrix $\Sigma = UDU^T \in \mathbb{R}^{p\times p}$ by choosing $U \in \mathbb{R}^{p\times p}$ according to Haar measure on the orthogonal matrices and populating the entries of the diagonal matrix D with i.i.d. draws from the uniform distribution on the interval $[1, b]$. The parameter $b \geq 1$ serves as a proxy for the condition number of Σ (e.g., $b = 1$ corresponds to $\Sigma = \mathbb{I}$). We then draw $T = 4p$ independent mean 0 normal random vectors with covariance matrix Σ. Splitting these T random vectors into two equal size samples of size $n = T/2$, we compute the harmonic mean of the two resulting sample covariances. We compare the performance of this estimator to the sample covariance matrix of the full sample (i.e., the arithmetic mean of the two splits). We also include, for the sake of comparison, the shrinkage estimator proposed by Fisher and Sun [28] specifically for normal data in the high-dimensional setting, which should improve on the sample covariance matrix. Similar to the scheme originally proposed by Ledoit and Wolf [36], this estimator is a convex combination of the sample covariance matrix and a target matrix, which we take here to be the identity.

Figures 1 and 2 display, for various choices of the dimension p and the condition number b, the operator norm relative error in recovering Σ for these three estimators. Over a wide range of condition numbers, the harmonic mean yields a better estimate of the population covariance Σ than does the arithmetic mean, but does not manage to match the performance of the Fisher-Sun regularized estimator. Unsurprisingly, when the regularization target matrix is close to the truth, as is the case when the con-

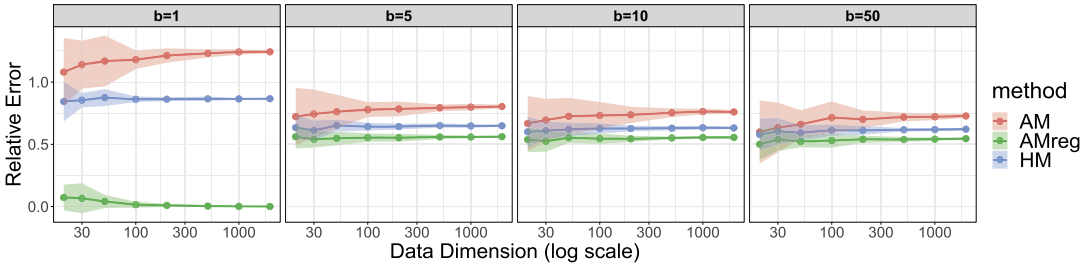

Figure 2. Operator norm relative error in recovering the true covariance matrix Σ as a function of the data dimension for different choices of condition number parameter b for the arithmetic mean (red), Fisher-Sun regularized sample covariance (green) and harmonic mean (blue). Each point is the mean of 20 independent trials, with the shaded regions indicating two standard deviations.

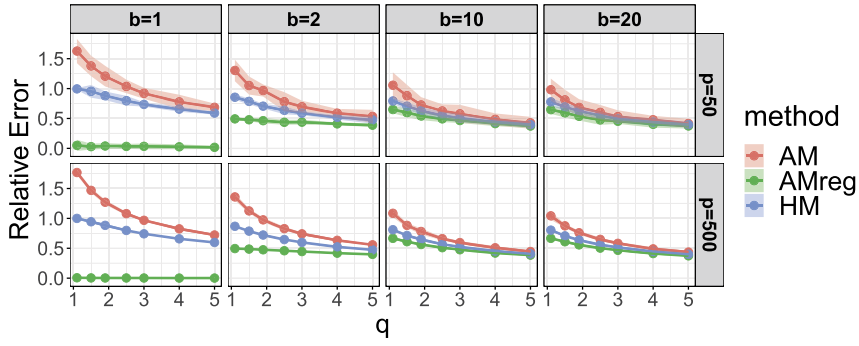

Figure 3. Operator norm relative error in recovering the true covariance matrix Σ as a function of the number of normal samples, parameterized by q, for different choices of the data dimension p and the condition number b. for the arithmetic mean (red), Fisher-Sun regularization of the arithmetic mean (green), and harmonic mean (blue). Each point is the mean of 20 independent trials, with the shaded regions indicating two standard deviations.

dition number b is close to 1, regularization yields an especially large improvement in estimation error, but the gap between the harmonic mean and the Fisher-Sun estimator narrows substantially as soon as the condition number becomes even moderately large, corresponding to the population covariance matrix being far from the identity. In keeping with Proposition 3.4, which suggests that we should only expect the harmonic mean to yield improvement over the arithmetic mean for suitably small condition numbers, the size of the improvement of harmonic mean over the (unregularized) arithmetic mean does appear to shrink as the condition number increases. Note that performance stabilizes as b and p increase because we are assessing the estimators according to relative error, not because the problem is necessarily becoming easier for larger values of these parameters.

It is natural to ask how these estimators compare as the number of observations vary. Toward that end, consider the same experimental setup discussed above, but taking $\lceil 2qp \rceil$ samples, with $q > 1$ to ensure that splitting the observations into two samples yields two invertible sample covariances. Larger values of q correspond, roughly, to better-conditioned sample covariance matrices. Figure 3 shows how the three estimators of the population matrix mean compare as a function of this parameter q for different choices of the dimension p and condition number b. We see that as the number of samples increases (i.e., as q increases), the improvement of the harmonic mean over the arithmetic mean decreases. This is in keeping with Proposition 3.2 as well as the intuition that as the number of samples increases, the sample covariance becomes a more stable (though still not consistent) estimate of the population covariance.

In Section 4, we saw that in the $N = 2$ case, the matrix harmonic mean $\mathbf{H}$ could be further improved by Rao-Blackwellization. Thus, we have four possible estimates of the population covariance: the arithmetic mean, the Fisher-Sun regularization of the arithmetic mean, the harmonic mean and the Rao-Blackwellization of the harmonic mean. Figure 4 compares these four different estimators, under the same experimental setup as the one described above. The plot shows that Rao-Blackwellizing the harmonic mean does little to change its behavior. Indeed, the performance of the harmonic mean and its Rao-Blackwellization are so similar that their lines overlap in the plot. As in the plots above, the arithmetic mean performs more poorly as an estimator compared to the harmonic mean, but regularization using the method of Fisher and Sun improves its performance over that of the harmonic mean, if only slightly.

We close by briefly exploring the eigenvector recovery results discussed in Section 5. Recall that the spiked eigenvector estimation problem considered in that section concerns Wishart matrices with co-

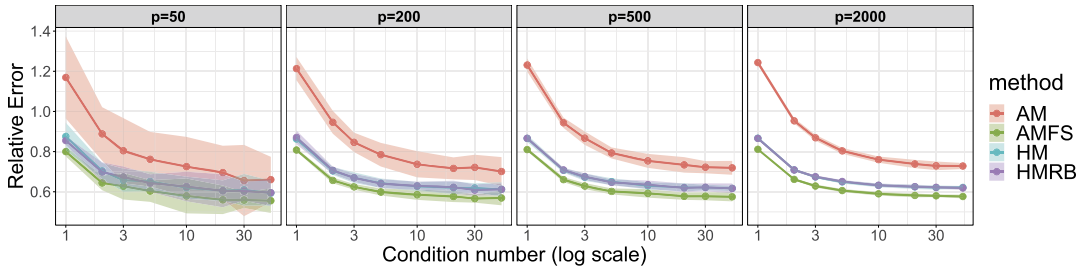

Figure 4. Operator norm relative error in recovering the true covariance matrix Σ as a function of the condition number b for different choices the data dimension p for the arithmetic mean (red), Fisher-Sun regularization of the arithmetic mean (green), harmonic mean (blue) and Rao-Blackwellized harmonic mean (purple). Each point is the mean of 20 independent trials, with the shaded regions indicating two standard deviations. Note that the lines corresponding to the harmonic mean and its Rao-Blackwellization overlap.

variance $\Sigma = \mathbb{I} + \theta vv^*$, where $\theta > 0$ and $v \in \mathbb{R}^p$ has unit norm, and the goal is to recover the spike eigenvector v based on $N = 2$ Wishart matrices, each constructed on n independent mean-0 covariance-Σ normals. Theorem 5.3 and Remark 5.4 predict that (in the large-p limit) as θ increases, the absolute value of the inner products between v and the leading eigenvector of both the arithmetic and the harmonic mean increase to 1. Further, this behavior undergoes a phase transition-like change, the location of which is determined by $\Gamma = \lim p/Nn$. In particular the arithmetic mean undergoes its phase transition at $\theta = \sqrt{\Gamma}$, and the harmonic mean undergoing its phase transition at $\theta = \sqrt{\Gamma/(1-\Gamma)}$. Figure 5 examines this behavior in the finite-sample regime.

Consider a pair of independent Wishart matrices, each based on $n = 4000$ independent normals of dimension $p = 2000$ with mean 0 and covariance $\Sigma = \mathbb{I} + \theta vv^*$ where $\theta > 0$ and $v \in \mathbb{R}^p$ has unit norm. That is, we are under the setting of Theorem 5.3, with $N = 2$ and $\Gamma = p/Nn = 1/4$. Having generated two such Wisharts, we can compute the arithmetic and harmonic means of these two covariance matrices and compare, for each of these two different means, how well the leading eigenvector $\hat{v}$ estimates the true spike eigenvector v, as measured by $|\langle \hat{v}, v \rangle|$. Figure 5 summarizes this experiment. The plot shows, for both the arithmetic (red) and the harmonic (blue) matrix means, the recovery of the spike v by the leading eigenvector of the matrix mean, as a function of the signal strength $\theta > 0$. Each data point is the mean of twenty independent trials, with error bars indicating two standard errors of

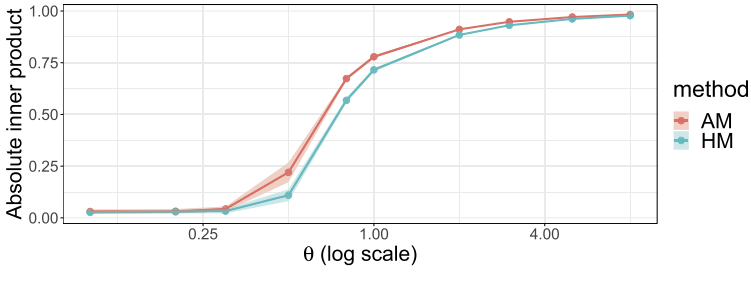

Figure 5. Recovery of the spike eigenvector v, as measured by the absolute value of the inner product between v and the leading eigenvector of the arithmetic matrix mean (red) and the harmonic matrix mean (blue), as a function of the signal strength θ. Each data point is the mean of twenty independent trials, with error bars indicating two standard errors of the mean.

the mean. With $\Gamma = 1/4$, the theory presented in Section 5 predicts that in the large-p limit, below $\theta = 1/\sqrt{3} \approx 0.577$, the inner product between the leading eigenvector of the harmonic mean with v should be close to zero. Similarly, in the case of the arithmetic mean, the inner product between v and the leading eigenvector of the arithmetic mean should be close to zero below $\theta = 1/2$. Of course, Figure 5 reflects this inner product for the finite-sample setting $p = 2000$. Nonetheless, examining the plot, we see that the predictions of Section 5 are borne out. Both the arithmetic and harmonic mean improve in their detection of the spike as θ increases, with the arithmetic mean appearing to detect the spike eigenvector before the harmonic mean does. Also as predicted by the theory, the arithmetic mean maintains better estimation of the leading eigenvector at all values of θ.

7. Discussion

The results presented here seem to contradict our intuition about matrix estimation under various matrix norms. It is common in the literature to measure the quality of an estimator $\hat{\Sigma}$ by the limit of the Frobenius norm error $\|\hat{\Sigma} - \Sigma\|_F$, and shrinkage estimators such as [36] are designed to minimize this quantity. Since the operator norm is bounded above by the Frobenius norm, controlling the Frobenius norm error is sufficient to control the operator norm. However, in the high-dimensional regime, these arguments become more subtle. One must often normalize the Frobenius norm by the matrix dimension to ensure convergence and obtain a sensible asymptotic analysis. Typically, this normalization takes the form $p^{-\frac{1}{2}}\|\hat{\Sigma} - \Sigma\|_F$. The analogous operator norm quantity, $p^{-\frac{1}{2}}\|\hat{\Sigma} - \Sigma\|$, converges to zero in many common settings, rendering the upper bound in terms of $p^{-\frac{1}{2}}\|\hat{\Sigma} - \Sigma\|_F$ (asymptotically) trivial. For example, when $\hat{\Sigma}$ is Wishart-distributed (as happens when $\hat{\Sigma}$ is a sample covariance), $\|\hat{\Sigma} - \Sigma\| = O(1)$. Thus, obtaining non-trivial bounds on $\|\hat{\Sigma} - \Sigma\|$ requires direct analysis rather than a Frobenius norm bound.

Ultimately, the choice to analyze the operator norm error as opposed to the Frobenius norm error or some other matrix norm is guided by the inference task at hand, and the available information about the population covariance Σ that one wishes to capture in the estimator $\hat{\Sigma}$. For the task of recovering the leading eigenvector(s) of Σ, the operator norm takes a more prominent role due to the Davis-Kahan inequality, but here again the resulting bound need not be tight, and the resulting bound on eigenvector recovery may be trivial.

In summary, our results highlight the ways in which the harmonic mean $\mathbf{H}(\mathcal{P})$ may outperform the arithmetic mean $\mathbf{A}(\mathcal{P})$ as an estimator of the population covariance Σ:

- Under certain conditions on Σ, the operator norm error of $\mathbf{H}(\mathcal{P})$ is *better than* that of $\mathbf{A}(\mathcal{P})$. In particular, when $\Sigma = \mathbb{I}$, the normalized Frobenius norm error of $\mathbf{H}(\mathcal{P})$ *matches* that of $\mathbf{A}(\mathcal{P})$, despite $\mathbf{H}(\mathcal{P})$ not being optimized for the Frobenius norm loss. We have observed similar behavior empirically when Σ is close but not equal to the identity.
- For a spiked model $\Sigma = \mathbb{I} + \theta vv^*$, there is a range of values of θ for which the eigenvector recovery using $\mathbf{H}(\mathcal{P})$ is always *worse than* that obtained using $\mathbf{A}(\mathcal{P})$, even though $\mathbf{H}(\mathcal{P})$ provides a *better* operator norm error than $\mathbf{A}(\mathcal{P})$.

We are not aware of any other estimators with these two properties, let alone of one that has the interpretation of being a mean with respect to a different geometry. Moreover, the fact that $\mathbf{H}(\mathcal{P})$ can be interpreted as the result of a data-splitting procedure suggests the possibility of other procedures with similar interpretations that improve over classical estimators in the high-dimensional regime. The Rao-Blackwellization of $\mathbf{H}(\mathcal{P})$ shows that, for the purpose of covariance estimation, a similar or better improvement over $\mathbf{A}(\mathcal{P})$ can also be achieved by a suitably-chosen scalar multiple of $\mathbf{A}(\mathcal{P})$. While this shows that $\mathbf{H}(\mathcal{P})$ has better-performing alternatives in practice, our results shed light on the behavior of

different means in high dimensions, opening the door to future work and a better understanding other measures of matrix estimation error.

Our original reason for investigating other matrix means was the problem of misaligned observations, as happens in brain imaging data. In such settings, pooling the raw data is not feasible, (e.g., the underlying time series of resting state fMRI imaging), since observations cannot be aligned across samples. We initially believed that $\mathbf{H}(\mathcal{P})$ outperforming $\mathbf{A}(\mathcal{P})$ empirically was a consequence of this setting, but surprisingly found it to be the case even in the classic covariance estimation problem with no misalignment. The results presented here are only a partial explanation of this phenomenon, and a better understanding of the various matrix means in both the aligned and the misaligned settings in high dimensions warrants further study.

Appendix A: Technical results

Lemma A.1. *For every $0 < a < b$ and $k \geq 1$,*

$$\int_a^b x^{k-1}\sqrt{(b-x)(x-a)}\,\mathrm{d}x = \frac{\pi}{2}\left(\frac{a+b}{2}\right)^{k+1}\sum_{j=0}^{\lfloor\frac{k-1}{2}\rfloor}\frac{1}{j+1}\binom{k-1}{2j}\binom{2j}{j}\left(\frac{b-a}{b+a}\right)^{2j+2}\frac{1}{2^{2j}}.$$

Proof. This identity is a well-known expression for the moments of the Marčenko-Pastur law, aside from changing the support of the distribution. We include a proof for the sake of completeness. The change of variables

$$u = \frac{2}{b-a}\left(x - \frac{a+b}{2}\right)$$

makes the above integral equal to

$$\left(\frac{b-a}{2}\right)^2\int_{-1}^1\left\{\left(\frac{b-a}{2}\right)u + \left(\frac{a+b}{2}\right)\right\}^{k-1}\sqrt{1-u^2}\,\mathrm{d}u,$$

and expanding, this is equal to

$$\sum_{j=0}^{k-1}\binom{k-1}{j}\left(\frac{b-a}{2}\right)^{j+2}\left(\frac{a+b}{2}\right)^{k-1-j}\int_{-1}^1 u^j\sqrt{1-u^2}\,\mathrm{d}u.$$

Each odd j vanishes by symmetry, yielding

$$\sum_{j=0}^{\lfloor\frac{k-1}{2}\rfloor}\binom{k-1}{2j}\left(\frac{b-a}{2}\right)^{2j+2}\left(\frac{a+b}{2}\right)^{k-1-2j}\int_{-1}^1 u^{2j}\sqrt{1-u^2}\,\mathrm{d}u.$$

This integral with respect to u is well-known and can be evaluated either through trigonometric substitution $u = \sin\theta$ and repeatedly integrating by parts, or by using evenness, changing variables to $u = \sqrt{v}$ and relating the resulting integral to the Beta function. We obtain

$$\int_{-1}^1 u^{2j}\sqrt{1-u^2}\,\mathrm{d}u = \frac{\pi}{2^{2j+1}}\binom{2j}{j}\frac{1}{j+1}.$$

Inserting this into the previous expression, the quantity of interest becomes

$$\frac{\pi}{2}\sum_{j=0}^{\lfloor\frac{k-1}{2}\rfloor}\binom{k-1}{2j}\left(\frac{b-a}{2}\right)^{2j+2}\left(\frac{a+b}{2}\right)^{k-1-2j}\binom{2j}{j}\frac{1}{j+1}\frac{1}{2^{2j}}. \qquad \square$$

The following are results from [33]. Let Δ be a fixed $p\times p$ positive definite matrix.

Definition A.2 (Multivariate Beta). A random $p\times p$ positive definite random matrix L is said to follow a multivariate Beta distribution with parameters p, n_1, n_2 and Δ if its density obeys

$$\begin{aligned} f_L(\ell) &:= K_{n_1,n_2}\frac{\det(\ell)^{\frac{n_1-p-1}{2}}\det(\Delta-\ell)^{\frac{n_2-p-1}{2}}}{\det(\Delta)^{\frac{(N-p-1)}{2}}}, \quad 0\le\ell\le\Delta,\\ K_{n_1,n_2} &:= \frac{\Gamma_p\left(\frac{N}{2}\right)}{\Gamma_p\left(\frac{n_1}{2}\right)\Gamma_p\left(\frac{n_1}{2}\right)},\\ \Gamma_p(x) &:= \pi^{\frac{p(p-1)}{4}}\prod_{i=1}^{p}\Gamma\Big(x-\frac{i-1}{2}\Big),\\ N &:= n_1+n_2, \end{aligned} \tag{A.1}$$

we denote this distribution as $B(p;n_1,n_2;\Delta)$.

The following result appears as Corollary 3.3 (ii) in [33]. We restate it here for ease of reference.

Theorem A.3. *Let T be an arbitrary $p\times p$ deterministic matrix. Suppose the matrix L is distributed as $B(p;n_1,n_2;\Delta)$ then*

$$\mathbb{E}[LTL]=\frac{n_1}{N(N-1)(N+2)}\left[\{n_1(N+1)-2\}\Delta T\Delta+n_2\{(\Delta T\Delta)^\top+\mathrm{Tr}(\Delta T)\Delta\}\right], \tag{A.2}$$

where $N=n_1+n_2$.

Appendix B: Strong asymptotic freeness for real Wishart matrices

Let $X_i\in\mathbb{R}^{p\times n_i}$ with $1\le i\le N$ be a sequence of random matrices with independent standard Gaussian entries and define $W_i=X_iX_i^\top/n_i$ as a sequence of i.i.d Wishart random matrices. Assume that $p/n_i\to\gamma_i>0$ as $p\to\infty$. We make use of free probability for what follows, for an overview see [46]. Let $(\mathcal{A},\|\cdot\|_{\mathcal{A}},*,\phi)$ be a non-commutative C^*-probability space with faithful tracial state ϕ such that there is a sequence of non-commutative freely independent free Poisson random variables $\{\mathfrak{p}_i\}_{1\le i\le N}\subset\mathcal{A}$ whose marginal distributions are

$$\phi\left(\mathfrak{p}_j^k\right)=\int_{\mathbb{R}}x^k\mu_{\mathrm{MP},\gamma_j}(\mathrm{d}x), \qquad 1\le j\le N, \quad k\ge 1,$$

where $\mu_{\mathrm{MP},\gamma_j}$ is the Marčenko-Pastur law with parameter γ_j. The space of non-commutative $*$-polynomials of N variables are non-commutative polynomials with complex coefficients in variables

$z_1, \ldots, z_N$ and $z_1^*, \ldots, z_N^*$. We wish to show that for any non-commutative $*$-polynomial Q in we have the following almost sure convergence

$$\lim_{n_i \to \infty} \|Q(W_1, \ldots, W_N)\| = \|Q(\mathfrak{p}_1, \ldots, \mathfrak{p}_N)\|_{\mathcal{A}},$$

as discussed in the proof of Proposition 3.2, this is the only result needed to extend the result of Proposition 3.2 from complex Wishart random matrices to real Wishart random matrices, the details of which are in [41, Remark 3].

This result will the follow from the recent result [27, Theorem 3.2], which we restate here. Suppose $\mathfrak{s}_1, \ldots, \mathfrak{s}_N \in \mathcal{A}$ are a sequence of freely independent free semicircular elements, that is,

$$\phi(\mathfrak{s}_j^k) = \int_{\mathbb{R}} \frac{x^k \sqrt{4 - x^2}}{2\pi} \, \mathrm{d}x,$$

and $y_1, \ldots, y_q \in \mathcal{A}$ are fixed with $\{\mathfrak{s}_i\}_{1 \le i \le N}$ free from $\{y_i\}_{1 \le i \le q}$ and $G_1, \ldots G_N$ are an i.i.d sequence of $M \times M$ GOE random matrices (real symmetric matrices with centered Gaussian entries with variance $1/M$ off the diagonal and $2/M$ on the diagonal) and let $Y_1, \ldots Y_q$ be a sequence of $M \times M$ deterministic matrices such that $\sup_{M \ge 1} \sup_{j \le q} \|Y_j\| \le C$ for some constant $C > 0$ and for any $*$-polynomial P in q non-commutative variables, we have

$$\lim_{M \to \infty} \frac{1}{M} \operatorname{Tr} P(Y_1, \ldots, Y_q) = \phi\big[P(y_1, \ldots, y_q)\big],$$
$$\lim_{M \to \infty} \|P(Y_1, \ldots, Y_q)\| = \|P(y_1, \ldots, y_q)\|_{\mathcal{A}},$$

then for any non-commutative polynomial Q in $N + q$ variables, we have the almost sure convergence

$$\lim_{M \to \infty} \frac{1}{M} \operatorname{Tr} Q(G_1, \ldots, G_N, Y_1, \ldots Y_q) = \phi\big[Q(\mathfrak{s}_1, \ldots, \mathfrak{s}_N, y_1, \ldots, y_q)\big],$$
$$\lim_{M \to \infty} \|Q(G_1, \ldots, G_N, Y_1, \ldots, Y_q)\| = \big\|Q(\mathfrak{s}_1, \ldots, \mathfrak{s}_N, y_1, \ldots, y_q)\big\|_{\mathcal{A}}.$$

We now follow the method outlined in [43, Section 9.2] to obtain the result we need from the result of [27] as follows. Let $\mathbf{0}_s$ be an $s \times s$ matrix of 0's and let $\mathbf{0}_{s,t}$ be an $s \times t$ matrix of 0's and let the parameter $M = p + \sum_{j=1}^N n_j$ and define the sequence of $M \times M$ block matrices

$$\mathbf{e}_0^{(M)} = \begin{bmatrix} \mathbb{I}_p & \mathbf{0}_{p,M-p} \\ \mathbf{0}_{p,M-p} & \mathbf{0}_{M-p} \end{bmatrix},$$
$$\mathbf{e}_j^{(M)} = \begin{bmatrix} \mathbf{0}_p & & & \\ & \mathbf{0}_{\sum_{k=1}^{j-1} n_k} & & \\ & & \mathbb{I}_{n_j} & \\ & & & \mathbf{0}_{\sum_{k=j+1}^N n_k} \end{bmatrix} \qquad 1 \le j \le N,$$

where in the definition of $\mathbf{e}_j^{(M)}$, $1 \le j \le N$, the omitted entries are all 0. Let $G_1, \ldots, G_N$ be a sequence of i.i.d $M \times M$ GOE matrices. Now consider the matrices

$$\frac{1}{\sqrt{n_j}} \tilde{X}_j = \sqrt{\frac{M}{n_j}} e_0^{(M)} G_j e_j^{(M)}, \qquad 1 \le j \le N,$$

note that the only non-zero entries of the matrix $\tilde{X}_j$ is a $p \times n_j$ block matrix in the first p rows and the columns from $p+1+\sum_{k=1}^{j-1} n_k$ to $p+\sum_{k=1}^{j} n_k$ and the distribution of this block matrix is identical to that of $\frac{1}{\sqrt{n_j}} X_j$. We now use the result of the previous paragraph to show convergence of $*$-polynomials of $\frac{1}{\sqrt{n_j}} \tilde{X}_j$ (with the matrices $\{\mathbf{e}_j^{(M)}\}_{1 \le j \le N}$ playing the role of the deterministic matrices Y_j). The remaining part of the proof of strong freeness of $W_1, \ldots, W_N$ follows from block matrix manipulations, where the polynomials of $\frac{1}{\sqrt{n_j}} \tilde{X}_j$ are used to produce polynomials in W_j. These calculations are essentially the same as those described in [43, Lemmas 9.3–5] which did these manipulations for GUE (complex Hermitian Gaussian matrices) matrices instead of GOE matrices and proved the strong freeness for complex Wisharts as a Corollary to their main result [43, Theorem 1.6]. One additional adjustment is needed, since the results of [43] concern Wishart matrices of the form $p = rd$ and $n_j = s_j d$, where $d \to \infty$ and r and s_j are fixed positive integers. These can be adapted to the present setting in the same way that we adjusted the block matrices $\{\mathbf{e}_j^{(M)}\}_{0 \le j \le N}$ above, and the remaining arguments go through similarly with only cosmetic changes to the original proof. Details are omitted for the sake of brevity.

Acknowledgments

The authors would like to thank the referees for their comments and suggestions.

Funding

K. Levin and A. Lodhia are supported by a NSF DMS Research Training Grant 1646108. E. Levina's research is partially supported by NSF DMS grants 1521551 and 1916222.

References

[1] Absil, P.-A., Mahony, R. and Sepulchre, R. (2004). Riemannian geometry of Grassmann manifolds with a view on algorithmic computation. *Acta Appl. Math.* **80** 199–220. MR2035507 https://doi.org/10.1023/B:ACAP.0000013855.14971.91

[2] Anderson, T.W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. *Wiley Series in Probability and Statistics*. Hoboken, NJ: Wiley Interscience. MR1990662

[3] Bai, Z. and Silverstein, J.W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*, 2nd ed. *Springer Series in Statistics*. New York: Springer. MR2567175 https://doi.org/10.1007/978-1-4419-0661-8

[4] Baik, J., Ben Arous, G. and Péché, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.* **33** 1643–1697. MR2165575 https://doi.org/10.1214/009117905000000233

[5] Benaych-Georges, F. and Nadakuditi, R.R. (2011). The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Adv. Math.* **227** 494–521. MR2782201 https://doi.org/10.1016/j.aim.2011.02.007

[6] Berthet, Q. and Rigollet, P. (2013). Optimal detection of sparse principal components in high dimension. *Ann. Statist.* **41** 1780–1815. MR3127849 https://doi.org/10.1214/13-AOS1127

[7] Bhatia, R. (2007). *Positive Definite Matrices*. *Princeton Series in Applied Mathematics*. Princeton, NJ: Princeton Univ. Press. MR2284176

[8] Bickel, P.J. and Levina, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008 https://doi.org/10.1214/08-AOS600

[9] Bickel, P.J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969 https://doi.org/10.1214/009053607000000758

[10] Bullmore, E. and Sporns, O. (2009). Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* **10** 186–198.
[11] Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Amer. Statist. Assoc.* **106** 672–684. MR2847949 https://doi.org/10.1198/jasa.2011.tm10560
[12] Cai, T., Ma, Z. and Wu, Y. (2015). Optimal estimation and rank detection for sparse spiked covariance matrices. *Probab. Theory Related Fields* **161** 781–815. MR3334281 https://doi.org/10.1007/s00440-014-0562-z
[13] Cai, T.T., Ren, Z. and Zhou, H.H. (2016). Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electron. J. Stat.* **10** 1–59. MR3466172 https://doi.org/10.1214/15-EJS1081
[14] Cai, T.T., Zhang, C.-H. and Zhou, H.H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.* **38** 2118–2144. MR2676885 https://doi.org/10.1214/09-AOS752
[15] Cai, T.T. and Zhou, H.H. (2012). Minimax estimation of large covariance matrices under ℓ_1-norm. *Statist. Sinica* **22** 1319–1349. MR3027084
[16] Cai, T.T. and Zhou, H.H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *Ann. Statist.* **40** 2389–2420. MR3097607 https://doi.org/10.1214/12-AOS998
[17] Chatterjee, S. (2015). Matrix estimation by universal singular value thresholding. *Ann. Statist.* **43** 177–214. MR3285604 https://doi.org/10.1214/14-AOS1272
[18] Chen, Y., Wiesel, A., Eldar, Y.C. and Hero, A.O. (2010). Shrinkage algorithms for MMSE covariance estimation. *IEEE Trans. Signal Process.* **58** 5016–5029. MR2722661 https://doi.org/10.1109/TSP.2010.2053029
[19] Daniels, M.J. and Kass, R.E. (2001). Shrinkage estimators for covariance matrices. *Biometrics* **57** 1173–1184. MR1950425 https://doi.org/10.1111/j.0006-341X.2001.01173.x
[20] Davis, C. and Kahan, W.M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.* **7** 1–46. MR0264450 https://doi.org/10.1137/0707001
[21] Dey, D.K. and Srinivasan, C. (1985). Estimation of a covariance matrix under Stein's loss. *Ann. Statist.* **13** 1581–1591. MR0811511 https://doi.org/10.1214/aos/1176349756
[22] Donoho, D., Gavish, M. and Johnstone, I. (2018). Optimal shrinkage of eigenvalues in the spiked covariance model. *Ann. Statist.* **46** 1742–1778. MR3819116 https://doi.org/10.1214/17-AOS1601
[23] Edelman, A., Arias, T.A. and Smith, S.T. (1999). The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.* **20** 303–353. MR1646856 https://doi.org/10.1137/S0895479895290954
[24] El Karoui, N. (2008). Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Ann. Statist.* **36** 2757–2790. MR2485012 https://doi.org/10.1214/07-AOS581
[25] Fan, J., Liao, Y. and Liu, H. (2016). An overview of the estimation of large covariance and precision matrices. *Econom. J.* **19** C1–C32. MR3501529 https://doi.org/10.1111/ectj.12061
[26] Fan, Z., Johnstone, I.M. and Sun, Y. (2018). Spiked covariances and principal components analysis in high-dimensional random effects models. Available at arXiv:1806.09529.
[27] Fan, Z., Sun, Y. and Wang, Z. (2021). Principal components in linear mixed models with general bulk. *Ann. Statist.* **49** 1489–1513. MR4302572 https://doi.org/10.1214/20-aos2010
[28] Fisher, T.J. and Sun, X. (2011). Improved Stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix. *Comput. Statist. Data Anal.* **55** 1909–1918. MR2765053 https://doi.org/10.1016/j.csda.2010.12.006
[29] Hoyle, D.C. and Rattray, M. (2007). Statistical mechanics of learning multiple orthogonal signals: Asymptotic theory and fluctuation effects. *Phys. Rev. E (3)* **75** 016101, 13. MR2324655 https://doi.org/10.1103/PhysRevE.75.016101
[30] Johnstone, I.M. and Lu, A.Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104** 682–693. MR2751448 https://doi.org/10.1198/jasa.2009.0121
[31] Kolaczyk, E.D., Lin, L., Rosenberg, S., Walters, J. and Xu, J. (2020). Averages of unlabeled networks: Geometric characterization and asymptotic behavior. *Ann. Statist.* **48** 514–538. MR4065172 https://doi.org/10.1214/19-AOS1820
[32] Kong, W. and Valiant, G. (2017). Spectrum estimation from samples. *Ann. Statist.* **45** 2218–2247. MR3718167 https://doi.org/10.1214/16-AOS1525
[33] Konno, Y. (1988). Exact moments of the multivariate F and beta distributions. *J. Japan Statist. Soc.* **18** 123–130. MR0980810

[34] Kubokawa, T. and Inoue, A. (2014). Estimation of covariance and precision matrices under scale-invariant quadratic loss in high dimension. *Electron. J. Stat.* **8** 130–158. MR3165436 https://doi.org/10.1214/14-EJS878
[35] Lam, C. (2016). Nonparametric eigenvalue-regularized precision or covariance matrix estimator. *Ann. Statist.* **44** 928–953. MR3485949 https://doi.org/10.1214/15-AOS1393
[36] Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.* **88** 365–411. MR2026339 https://doi.org/10.1016/S0047-259X(03)00096-4
[37] Ledoit, O. and Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Ann. Statist.* **40** 1024–1060. MR2985942 https://doi.org/10.1214/12-AOS989
[38] Ledoit, O. and Wolf, M. (2018). Optimal estimation of a large-dimensional covariance matrix under Stein's loss. *Bernoulli* **24** 3791–3832. MR3788189 https://doi.org/10.3150/17-BEJ979
[39] Levin, K., Lodhia, A. and Levina, E. (2019). Recovering low-rank structure from multiple networks with unknown edge distributions. *J. Mach. Learn. Res.* To appear. arXiv:1906.07265.
[40] Li, W., Chen, J., Qin, Y., Bai, Z. and Yao, J. (2013). Estimation of the population spectral distribution from a large dimensional sample covariance matrix. *J. Statist. Plann. Inference* **143** 1887–1897. MR3095079 https://doi.org/10.1016/j.jspi.2013.06.017
[41] Lodhia, A. (2021). Harmonic means of Wishart random matrices. *Random Matrices Theory Appl.* **10** Paper No. 2150016, 24. MR4260211 https://doi.org/10.1142/S2010326321500167
[42] Lunagómez, S., Olhede, S.C. and Wolfe, P.J. (2020). Modeling network populations via graph distances. *J. Amer. Statist. Assoc.* https://doi.org/10.1080/01621459.2020.1763803.
[43] Male, C. (2012). The norm of polynomials in large random and deterministic matrices. *Probab. Theory Related Fields* **154** 477–532. With an appendix by Dimitri Shlyakhtenko. MR3000553 https://doi.org/10.1007/s00440-011-0375-2
[44] Marčenko, V.A. and Pastur, L.A. (1967). Distribution of eigenvalues in certain sets of random matrices. *Mat. Sb. (N.S.)* **72** 507–536. MR0208649
[45] Marrinan, T., Beveridge, J.R., Draper, B., Kirby, M. and Peterson, C. (2014). Finding the subspace mean or median to fit your need. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 1082–1089.
[46] Mingo, J.A. and Speicher, R. (2017). *Free Probability and Random Matrices*. *Fields Institute Monographs* **35**. New York: Springer; Toronto, ON: Fields Institute for Research in Mathematical Sciences. MR3585560 https://doi.org/10.1007/978-1-4939-6942-5
[47] Muirhead, R.J. (1982). *Aspects of Multivariate Statistical Theory*. *Wiley Series in Probability and Mathematical Statistics*. New York: Wiley. MR0652932
[48] Nadler, B. (2008). Finite sample approximation results for principal component analysis: A matrix perturbation approach. *Ann. Statist.* **36** 2791–2817. MR2485013 https://doi.org/10.1214/08-AOS618
[49] Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica* **17** 1617–1642. MR2399865
[50] Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed. *Wiley Series in Probability and Mathematical Statistics*. New York: Wiley. MR0346957
[51] Rao, N.R., Mingo, J.A., Speicher, R. and Edelman, A. (2008). Statistical eigen-inference from large Wishart matrices. *Ann. Statist.* **36** 2850–2885. MR2485015 https://doi.org/10.1214/07-AOS583
[52] Rothman, A.J., Levina, E. and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *J. Amer. Statist. Assoc.* **104** 177–186. MR2504372 https://doi.org/10.1198/jasa.2009.0101
[53] Schwartzman, A. (2016). Lognormal distributions and geometric averages of symmetric positive definite matrices. *Int. Stat. Rev.* **84** 456–486. MR3580425 https://doi.org/10.1111/insr.12113
[54] Smith, S.T. (2005). Covariance, subspace, and intrinsic Cramér-Rao bounds. *IEEE Trans. Signal Process.* **53** 1610–1630. MR2148599 https://doi.org/10.1109/TSP.2005.845428
[55] Sporns, O. (2012). *Discovering the Human Connectome*. Cambridge: MIT Press.
[56] Stein, C. (1986). Lectures on the theory of estimation of many parameters. *J. Sov. Math.* **34** 1373–1403.
[57] Touloumis, A. (2015). Nonparametric Stein-type shrinkage covariance matrix estimators in high-dimensional settings. *Comput. Statist. Data Anal.* **83** 251–261. MR3281809 https://doi.org/10.1016/j.csda.2014.10.018
[58] Won, J.-H., Lim, J., Kim, S.-J. and Rajaratnam, B. (2013). Condition-number-regularized covariance estimation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 427–450. MR3065474 https://doi.org/10.1111/j.1467-9868.2012.01049.x

[59] Wu, W.B. and Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika* **90** 831–844. MR2024760 https://doi.org/10.1093/biomet/90.4.831

[60] Yao, J., Zheng, S. and Bai, Z. (2015). *Large Sample Covariance Matrices and High-Dimensional Data Analysis*. *Cambridge Series in Statistical and Probabilistic Mathematics* **39**. New York: Cambridge Univ. Press. MR3468554 https://doi.org/10.1017/CBO9781107588080

[61] Yu, Y., Wang, T. and Samworth, R.J. (2015). A useful variant of the Davis-Kahan theorem for statisticians. *Biometrika* **102** 315–323. MR3371006 https://doi.org/10.1093/biomet/asv008

Received October 2019 and revised July 2021