New Methods for Handling Singular Sample Covariance Matrices

Gabriel H. Tucci and Ke Wang

***Abstract*—The estimation of a covariance matrix from an insufficient amount of data is one of the most common problems in fields as diverse as multivariate statistics, wireless communications, signal processing, biology, learning theory and finance. In a joint work of Marzetta, Tucci and Simon, a new approach to handle singular covariance matrices was suggested. The main idea was to use dimensionality reduction in conjunction with an average over the Stiefel manifold. In this paper we continue with this research and we consider some new approaches to handle this problem. One of the methods is called the mean conjugate estimator under Ewens measure and uses a randomization of the sample covariance matrix over all the permutation matrices with respect to the Ewens measure. The techniques used to attack this problem are broad and run from random matrix theory to combinatorics.**

***Index Terms*—sample covariance matrix, random matrices, Stiefel manifold, Haar measure, Ewens measure.**

I. Introduction

THE estimation of a covariance matrix from an insufficient amount of data is one of the most common problems in fields as diverse as multivariate statistics, wireless communications, signal processing, biology, learning theory and finance. For instance, the covariation between asset returns plays a crucial role in modern finance. The covariance matrix and its inverse are the key statistics in portfolio optimization and risk management. Many recent financial innovations involve complex derivatives, like exotic options written on the minimum, maximum or difference of two assets, or some structured financial products, such as CDOs. All of these innovations are built upon, or in order to exploit, the correlation structure of two or more assets. In the field of wireless communications, covariance estimates allows us to compute the direction of arrival (DOA), which is a critical task in smart antenna systems since it enables accurate mobile location (see [30], [31]). Another application is in the field of biology and involves the interactions between proteins or genes in an organism and the joint time evolution of their interactions (see [27] for instance).

Typically the covariance matrix of a multivariate random variable is not known but has to be estimated from the data. Estimation of covariance matrices then deals with the question of how to approximate the actual covariance matrix on the basis of samples from the multivariate distribution. Simple cases, where the number of observations is much greater than the number of variables, can be dealt with by using the sample covariance matrix. In this case, the sample covariance matrix is an unbiased and efficient estimator of the *true* covariance matrix. However, in many practical situations we would like to estimate the covariance matrix of a set of variables from an insufficient amount of data. In this case the sample covariance matrix is singular (non–invertible) and therefore a fundamentally bad estimate. More specifically, let X be a random vector $X = (X_1, \ldots, X_m)^T \in \mathbb{C}^{m\times 1}$ and assume for simplicity that X is centered. Then the true covariance matrix is given by

$$\Sigma = \mathbb{E}(XX^*) = (\mathrm{cov}(X_i, X_j))_{1\leq i,j\leq m}. \tag{1}$$

Consider n independent samples or realizations $x_1, \ldots, x_n \in \mathbb{C}^m$ and form the $m \times n$ data matrix $M = (x_1, \ldots, x_n)$. Then the sample covariance matrix is an $m \times m$ non–negative definite matrix defined as

$$K = \frac{1}{n} MM^*. \tag{2}$$

If $n \to +\infty$ and m is fixed, then the sample covariance matrix K converges (entrywise) to Σ almost surely. Whereas, as we mentioned before, in many empirical problems, the number of measurements is less than the dimension $(n < m)$, and thus the sample covariance matrix is singular. Our objective in this paper is to recover the true covariance matrix Σ from K under the condition $n < m$.

The conventional treatment of covariance singularity artificially converts the singular sample covariance matrix into an invertible (positive definite) covariance by the simple expedient of adding a positive diagonal matrix, or more generally, by taking a linear combination of the sample covariance and the identity matrix. This procedure is variously called "diagonal loading" or "ridge regression" [9], [24]. This one is defined as $\alpha K + \beta I_m$ where α and β are called loading parameters. The resulting matrix is positive definite, invertible and preserves the eigenvectors of the sample covariance. The eigenvalues of $\alpha K + \beta I_m$ are a uniform rescaling and shift of the eigenvalues of K. There are many methods in choosing the optimum loading parameters, see [17], [21] and [22]. On the other hand, if the true covariance matrix is assumed to have some level of sparsity, several works have been established, such as the banding and thresholding methods studied by Bickel and Levina [3], [4], Wu and Pourahmadi [35], El Karoui [10] and Rothman et al. [25], to mention a few. In more recent works, Cai, Zhang and Zhou [7] and Cai and Zhou [8] derive the

Gabriel H. Tucci is the global head of Central Risk and Cash Trading in the Equities division at Citi, 388 Greenwich Street. New York, NY 10013, USA. Email: gabrieltucci@gmail.com.

Ke Wang is with the Department of Mathematics, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong. Email: kewang@ust.hk. Ke Wang is supported by HKUST Initiation Grant IGN16SC05.

This paper was presented in part at the International Symposium on Information Theory, Boston, 2012.

optimal rate of convergence for estimating the true covariance matrix and its inverse under operator norm, Frobenius norm and l_1 norm, for a large range of sparse covariance matrices.

In Marzetta, Tucci and Simon's paper [20] a new approach to handle singular covariance matrices was suggested. They use the idea of *random dimension reduction*. Let $p \leq n$ be a parameter, to be estimated later, and consider the set of all $p \times m$ one-sided unitary matrices

$$\Omega_{p,m} = \{\Phi \in \mathbb{C}^{p\times m} \ : \ \Phi\Phi^* = I_p\}. \tag{3}$$

This set has a manifold structure and is called the Stiefel manifold. Note that ΦM, that is the multiplication of the one-sided unitary matrix Φ with the data matrix M, results in a new data matrix with reduced dimension. And

$$\frac{1}{n}(\Phi M)(\Phi M)^* = \Phi K \Phi^* \tag{4}$$

can be viewed as a new sample covariance matrix of size p. Then $\Phi^*(\Phi K\Phi^*)\Phi$ will project the data back to n-dimensional space. In [20], they endow the Stiefel manifold with the *Haar measure*, that is, the uniform distribution on the set $\Omega_{p,m}$. Further, they define the operators

$$\mathrm{cov}_p(K) = \mathbb{E}(\Phi^*(\Phi K\Phi^*)\Phi),$$

$$\mathrm{invcov}_p(K) = \mathbb{E}(\Phi^*(\Phi K\Phi^*)^{-1}\Phi),$$

where the expectation is taken with respect to the Haar measure. The operators $\mathrm{cov}_p(K)$ and $\mathrm{invcov}_p(K)$ are used to estimate the true covariance matrix Σ and its inverse Σ^{-1} respectively. It was found that

$$\mathrm{cov}_p(K) = \frac{p}{(m^2-1)m}\Big((mp-1)K + (m-p)\mathrm{Tr}(K)I_m\Big),$$

which is the same as diagonal loading. Moreover, they investigated the properties of $\mathrm{invcov}_p(K)$. If K is decomposed as $K = UDU^*$, with $D = \mathrm{diag}(d_1, \ldots, d_n, 0, \ldots, 0)$, then

$$\mathrm{invcov}_p(K) = U\mathrm{invcov}_p(D)U^*,$$

and

$$\mathrm{invcov}_p(D) = \mathrm{diag}(\lambda_1, \ldots, \lambda_n, \mu, \ldots, \mu). \tag{5}$$

In other words, $\mathrm{invcov}_p(K)$ preserves the eigenvectors of K, and transforms all the zero eigenvalues to a non–zero constant value. They also provided formulas to compute the values of λ_i and μ, and studied their asymptotic behavior using techniques from free probability.

The explicit formula of λ_i's of $\mathrm{invcov}_p(D)$ in (5) is derived in [20] as a partial derivative of a rather complicated integral (see (11) and Theorem 1 in [20]). In this paper, we further investigate the properties of the $\mathrm{invcov}_p(K)$ or equivalently the $\mathrm{invcov}_p(D)$ operators. These results are presented in Section II. We first show that $\mathrm{invcov}_p(D)$ has a surprisingly simple algebraic structure, i.e. it is a polynomial of the diagonal matrix D. We also provide formulas to compute the coefficients of the polynomial and illustrate the computation through a small dimensional example in Appendix A. The formulas involve complicated combinatorial subjects and thus make further investigation on the performance, i.e. optimize the error functions with respect to the parameters, rather difficult.

Therefore, it is natural to look for alternative random operators that are easy to compute, analyze and implement. It is known that a random unitary matrix with Haar measure behaves asymptotically like a random uniform permutation matrix (see [33] and [34]). Our first attempt is to conjugate the sample covariance matrix K with a permutation matrix M_σ. In [32], the mean conjugate $K_1 = \mathbb{E}(M_\sigma K M_\sigma^T)$ of a square matrix K averaging over uniform permutation matrix M_σ is studied. It is found in [32] that K_1 is always a scalar multiple of identity matrix plus a rank-one matrix (see Remark III-B), which is a well-conditioned matrix in most cases.

Now we investigate the mean conjugate of a matrix K under a generalized measure on the permutation group, called the Ewens measure with parameter $\theta > 0$ (see (14) below). We obtain a closed form expression for the estimator $K_\theta = \mathbb{E}(M_\sigma K M_\sigma^T)$ in Theorem III-A using combinatorial techniques. We find that the averaging operation on diagonal matrices is equivalent to the conventional diagonal loading (see Remark III-C). For the matrix K with certain structures, the averaging over all permutation matrices under Ewens measure by choosing θ propositional to the dimension m, is asymptotically equivalent to *linear shrinkage estimator* proposed by Lenoit and Wolf [18]. This result is proved in Section V-A. We propose this new method to estimate the covariance matrices and call it the *mean conjugate estimator under Ewens measure*.

In Section IV, we extend the ideas of constructing the $\mathrm{cov}_p(K)$ and $\mathrm{invcov}_p(K)$ operators by replacing random unitary matrices with random permutation matrices. We first extend the definition of permutation matrices to get $p \times m$ unitary matrices V_σ and use the Ewens measure in Section III. Then we define two new operators

$$K_{\theta,m,p} := \mathbb{E}\big(V_\sigma^T(V_\sigma K V_\sigma^T)V_\sigma\big)$$

$$\tilde{K}_{\theta,m,p} := \mathbb{E}\big(V_\sigma^T(V_\sigma K V_\sigma^T)^+V_\sigma\big)$$

to estimate Σ and Σ^{-1} respectively. Here A^+ is the *Moore-Penrose pseudo inverse* of the A. If A is an $m \times n$ complex or real matrix, then A^+ is an $n \times m$ complex or real matrix that satisfies AA^+ and A^+A are both Hermitian or symmetric, $AA^+A = A$ and $A^+AA^+ = A$. For any matrix A, the pseudo inverse A^+ always exists. We provide an explicit formula for $K_{\theta,m,p}$ and an inductive formula to compute $\tilde{K}_{\theta,m,p}$.

In Section V, we first study the asymptotic behavior for certain matrices with the mean conjugate estimator under Ewens measure. We conduct some simulation study focusing on the mean conjugate estimator under Ewens measure. However, we do not include the simulations on the hybrid operators $K_{\theta,m,p}$ and $\tilde{K}_{\theta,m,p}$ since currently we do not have adequate understanding on them from explicit formulas obtained in Section IV.

Notation: Throughout this paper, $\mathbf{1}_S$ is the indicator function of a set S. We sometimes use $[n]$ to present the set $\{1, 2, \ldots, n\}$, and $\mathrm{Tr}(A)$ is the trace of a matrix A. For an $m \times m$ matrix A, we use the (normalized) Frobenius norm $\|A\|_F = \frac{1}{\sqrt{m}}\sqrt{\mathrm{Tr}(AA^*)}$. We denote A^+ the Moore-Penrose

pseudo inverse of the matrix A. For a vector $v = (v_1, \ldots, v_m)$, we use the Euclidean norm $\|v\|_2 = \sqrt{\sum_{i=1}^m |v_i|^2}$. We use $v(k)$ to denote the kth entry of v. We use $\mathbf{e} = (1, \ldots, 1)^T$ to represent the all-one vector and e_i are the standard basis vectors. We use the notation $\kappa \vdash n$ to indicate that κ is an integer partition of the positive integer n.

II. Some Properties of the invcov_p Estimator

We first collect some preliminaries about Schur polynomials that will be needed later in studying the properties of the invcov_p estimator.

A. Preliminaries of Schur polynomials

A symmetric polynomial is a polynomial $P(x_1, x_2, \ldots, x_n)$ in n variables such that if any of the variables are interchanged one obtains the same polynomial. Formally, P is a symmetric polynomial if for any permutation σ of the set $\{1, 2, \ldots, n\}$ one has that

$$P(x_{\sigma(1)}, x_{\sigma(2)}, \ldots, x_{\sigma(n)}) = P(x_1, x_2, \ldots, x_n).$$

Symmetric polynomials arise naturally in the study of the relation between the roots of a polynomial in one variable and its coefficients, since the coefficients can be given by a symmetric polynomial expressions in the roots. Symmetric polynomials also form an interesting structure by themselves. The resulting structures, and in particular the ring of symmetric functions, are of great importance in combinatorics and in representation theory (see for instance [13], [19], [23], [26] for more on details on this topic).

The Schur polynomials are certain symmetric polynomials in n variables. This class of polynomials is also very important in representation theory since they are the characters of irreducible representations of the general linear groups. The Schur polynomials are indexed by partitions. A partition of a positive integer n, also called an integer partition, is a way of writing n as a sum of positive integers. Two partitions that differ only in the order of their summands are considered to be the same partition. Therefore, $\kappa = (\kappa_1, \ldots, \kappa_n) \vdash n$ is a partition of a positive integer of n if

$$\sum_{i=1}^n \kappa_i = n \quad \text{with} \quad \kappa_1 \geq \kappa_2 \geq \ldots \geq \kappa_n \geq 0.$$

The κ_i's are called the *parts* of κ. Notice that some of the κ_i could be zero. Sometimes, we use another equivalent way to represent a partition. We write $\kappa = (1^{r_1}, 2^{r_2}, \ldots, n^{r_n}) \vdash n$ where r_i is the number of i appearing as parts in κ. Thus $\sum_{i=1}^n i \cdot r_i = n$. Integer partitions are usually represented by the so called Young's diagrams (also known as Ferrers' diagrams). A Young diagram is a finite collection of boxes, or cells, arranged in left–justified rows, with the row lengths weakly decreasing (each row has the same or shorter length than its predecessor). Listing the number of boxes on each row gives a partition κ of a non-negative integer n, the total number of boxes of the diagram. The Young diagram is said to be of shape κ, and it carries the same information as that of partition. For instance, in Figure II-A we can see the Young diagram corresponding to the partition $(5, 4, 1)$ of the number 10. Given a partition κ of m. Assume $m \geq n$. The Schur polynomial of shape κ in the variables $(d_1, \ldots, d_n)$ is defined as

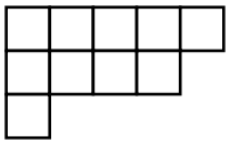


Fig. 1. Young digram representation of the partition $(5, 4, 1)$.

$$s_\kappa(d_1, \ldots, d_n) = \frac{\det(d_i^{n+\kappa_j - j})_{i,j=1}^n}{\det(d_i^{n-j})_{i,j=1}^n}.$$

Indeed the denominator $\det(d_i^{n-j})_{i,j=1}^n$ is the determinant of the Vandermonde matrix

$$\Delta(d_1, \ldots, d_n) = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ d_1 & d_2 & \cdots & d_n \\ \vdots & \vdots & \ddots & \vdots \\ d_1^{n-1} & d_2^{n-1} & \cdots & d_n^{n-1} \end{pmatrix}. \quad (6)$$

The numerator $\det(d_i^{n+\kappa_j - j})_{i,j=1}^n$ is an alternating polynomial (in other words it changes sign under any transposition of the variables):

$$\det(d_i^{n+\kappa_j - j})_{i,j=1}^n = \sum_{\sigma \in S_n} \epsilon(\sigma) d_{\sigma(1)}^{\kappa_1} \cdots d_{\sigma(n)}^{\kappa_n},$$

where S_n is the permutation group of the set $\{1, 2, \ldots, n\}$ and $\epsilon(\sigma)$ is the sign of the permutation σ.

Thus $s_\kappa(d_1, \ldots, d_n)$ is a symmetric function because the numerator and denominator are both alternating, and is a polynomial since all alternating polynomials are divisible by the Vandermonde determinant (see [13], [19], [26] for more details here). For instance, $s_{(2,1,1)}(x_1, x_2, x_3) = x_1\, x_2\, x_3\, (x_1 + x_2 + x_3)$ and

$$s_{(2,2,0)}(x_1, x_2, x_3) = x_1^2\, x_2^2 + x_1^2\, x_3^2 + x_2^2\, x_3^2 + x_1^2\, x_2\, x_3 + x_1\, x_2^2\, x_3 + x_1\, x_2\, x_3^2.$$

Another related definition is the *Hook length*, $\mathrm{hook}(x)$, of a box x in Young diagram of shape κ. This is defined as the number of boxes that are in the same row to the right of it plus those boxes in the same column below it, plus one (for the box itself). For instance, in Figure II-A, the hook length of the top-left corner box is $4 + 2 + 1 = 7$. The product of the hook's length of a partition is the product of the hook lengths of all the boxes in the partition.

Next, we collect a few properties of Schur polynomials $s_\kappa(d_1, \ldots, d_n)$ used in later proofs. For an $n \times n$ matrix A with eigenvalues $\alpha_1, \ldots, \alpha_n$, we use $s_\kappa(A) = s_\kappa(\alpha_1, \ldots, \alpha_n)$. Denote by $(n-k, 1^k)$ the partition $(n-k, 1, 1, \ldots, 1)$ with k ones. One of the basic properties of Schur polynomials is that for any integer $l \geq 1$,

$$\mathrm{Tr}(A^l) = \sum_{k=0}^{n-1} (-1)^k s_{(n-l, 1^l)}(A). \quad (7)$$

Let D_n be a diagonal matrix of size $n \times n$. Consider $\Omega_{p,n}$, the Stiefel manifold defined in (3), associated with the Haar

measure $d\phi$. For any $\Phi \in \Omega_{p,n}$, it is proved in [12, equation (18)] that

$$\int_{\Omega_{p,n}} s_\kappa(\Phi D_n \Phi^*)\, d\phi = \frac{s_\kappa(D_n)s_\kappa(I_n)}{s_\kappa(I_p)}. \tag{8}$$

Schur polynomials have a close connection with the border strips of partitions. We follow the definitions in Stanley's book [28, Chapter 7.17]. A *border strip* is a set of boxes in the Young diagram that forms a contiguous strip and has at most one box on each diagonal. The *height* of a border strip is one less than its number of rows. Given a partition $\lambda \vdash n$ and a decomposition $\rho = (\rho_1, \ldots, \rho_l)$ of n. A *border strip tableau* $\chi^\kappa(\rho)$ of shape κ and type ρ is obtained by replacing each box in the Young diagram of κ by one of the integers $\{1, 2, \ldots, l\}$ so that the boxes replaced by i form a ρ_i border strip in the diagram which consists of all boxes replaced by $\{1, 2, \ldots, i\}$.

By the celebrated Murnaghan–Nakayama rule (see Corollary 7.17.5 in [28]),

$$s_{(n-j,1^j)}(D) = \sum_{\rho=(1^{r_1},2^{r_2},\ldots,n^{r_n})\vdash n} \chi^{(n-j,1^j)}(\rho) \prod_{l=1}^{n} \frac{\mathrm{Tr}(D^l)^{r_l}}{l^{r_l} r_l!}, \tag{9}$$

where $\chi^\kappa(\rho) = \sum_T (-1)^{\mathrm{ht}(T)}$ sums over all border-strip tableaux of shape κ and type ρ. Here $\mathrm{ht}(T)$ is the *height* of a border-strip tableaux (see Section 7.17 in [28] for more details).

B. A new property of the $invcov_p$ estimator

Recall $\mathrm{invcov}_p(K) = \mathbb{E}(\Phi^*(\Phi K \Phi^*)^{-1}\Phi)$. We first collect the properties of the $\mathrm{invcov}_p(K)$ estimator obtained in the previous work of Marzetta, Tucci and Simon [20, Section IV and VI].

Proposition II-B1. *For a positive semi-definite matrix K of size m, one can decompose $K = UDU^*$ where U is unitary and $D = \mathrm{diag}(d_1, \ldots, d_m)$.*

1) *The eigenvectors of K are preserved under the invcov_p operatoration. More precisely, $\mathrm{invcov}_p(K) = U\mathrm{invcov}_p(D)U^*$ and $\mathrm{invcov}_p(D)$ is diagonal.*
2) *The zero-eigenvalues of K are converted to equal positive values. If $D = diag(D_n, 0_{m-n})$ where $D_n = (d_1, \ldots, d_n)$ is of full rank, then*

$$\mathrm{invcov}_p(D) = diag(\Lambda_L(D_n), \mu I_{m-n}),$$

where $\Lambda_L(D_n) = diag(\lambda_1, \ldots, \lambda_n)$. Besides, for any $1 \le k \le n$,

$$\lambda_k = \frac{\partial}{\partial d_k} \int_{\Omega_{p,n}} \mathrm{Tr}(\log(\Phi D_n \Phi^*))\, d\phi,$$
$$\mu = \frac{\det(G)}{\det(\Delta(d_1, \ldots, d_n))}. \tag{10}$$

Here $\Delta(d_1, \ldots, d_n)$ is the Vandermonde matrix in (6) and G is the matrix constructed by replacing the pth row of $\Delta(d_1, \ldots, d_n)$ by the row

$$(d_1^{n-(p+1)} \log(d_1), \cdots, d_n^{n-(p+1)} \log(d_n)).$$

We prove a new property of the $\mathrm{invcov}_p(K)$ estimator. We will show that $\mathrm{invcov}_p(K)$ has a surprisingly simple algebraic structure despite its rather complicated expression. Assume $K = UDU^*$ where U is unitary and $D = \mathrm{diag}(d_1, \ldots, d_m)$. By Proposition II-B1, it is enough to study the properties of $\mathrm{invcov}_p(D)$.

Let $\mathcal{A}(D)$ be the algebra generated by the matrices D and the $m \times m$ identity matrix I_m. By the Cayley–Hamilton Theorem, it is clear that

$$\mathcal{A}(D) = \Big\{ \alpha_{m-1} D^{m-1} + \ldots + \alpha_1 D + \alpha_0 I_m \ : \ \alpha_i \in \mathbb{C} \Big\}.$$

We define $\mathcal{D}_m$ as the set of all $m \times m$ diagonal matrices.

Lemma II-C. *Let $D = \mathrm{diag}(d_1, \ldots, d_m)$ be an $m \times m$ diagonal matrix. If $d_i \neq d_j$ for $i \neq j$ then $\mathcal{A}(D) = \mathcal{D}_m$. If $d_i = d_j$ for some $i \neq j$ then*

$$\mathcal{A}(D) = \{\mathrm{diag}(b_1, \ldots, b_i, \ldots, b_i, \ldots, b_m) \ : \ b_k \in \mathbb{C}\},$$

the set of all diagonal matrices where the ith and jth entries are equal.

Proof. First assume $d_i \neq d_j$ for all $i \neq j$. It is clear to see $\mathcal{A}(D) \subset \mathcal{D}_m$. On the other hand, for any $B = \mathrm{diag}(b_1, \ldots, b_m) \in \mathcal{D}_m$, we form a system of linear equations,

$$\begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} = V \begin{pmatrix} \alpha_0 \\ \vdots \\ \alpha_{m-1} \end{pmatrix},$$

where

$$V := \begin{pmatrix} 1 & d_1 & d_1^2 & \ldots & d_1^{m-1} \\ \vdots & & \ldots & & \vdots \\ 1 & d_m & d_m^2 & \ldots & d_m^{m-1} \end{pmatrix}.$$

The matrix V is a Vandermonde matrix with $\det(V) = \prod_{i<j}(d_i - d_j)$. The matrix V is invertible by our assumption. Thus we can find a vector $(\alpha_0, \ldots, \alpha_{m-1})$ such that

$$B = \alpha_0 I_m + \alpha_1 D + \ldots + \alpha_{m-1} D^{m-1} \in \mathcal{A}(D).$$

This completes the proof. To prove the second part we use essentially the same approach as before. □

Theorem II-D. *The matrix $\mathrm{invcov}_p(D)$ belongs to the algebra $\mathcal{A}(D)$.*

Proof. By Proposition II-B1, if the matrix D is equal to $D = \mathrm{diag}(D_n, 0_{m-n})$ where $D_n = (d_1, \ldots, d_n)$ is of full rank, then $\mathrm{invcov}_p(D) = \mathrm{diag}(\Lambda_L(D_n), \mu I_{m-n})$ where $\Lambda_L(D_n) = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. And

$$\lambda_k = \frac{\partial F(d_1, \ldots, d_n)}{\partial d_k},$$

where we define $F(d_1, \ldots, d_n) := \int_{\Omega_{p,n}} \mathrm{Tr}(\log(\Phi D_n \Phi^*)) d\phi$ for brevity. Recall $\Phi \in \Omega_{p,n}$ defined in (3). By (7) and (8), for any integer $l \ge 1$

$$\int_{\Omega_{p,n}} \mathrm{Tr}\big((\Phi D_n \Phi^*)\big)^l\, d\phi = \sum_{k=0}^{p-1} (-1)^k c_k^{(n,p)} s_{(l-k,1^k)}(D_n),$$

where $s_{(l-k,1^k)}(D_n)$ are the Schur polynomials and $c_k^{(n,p)}$ are explicit constants (see (78) in [20]). From Lemma II-C, it is enough to show that if $d_i = d_j$ for some $i \neq j$, then $\lambda_i =$

λ_j. By linearity and continuity, $F(d_1,\dots,d_n)$ is symmetric. Hence assuming $d_i = d_j$, $\partial F/\partial d_i = \partial F/\partial d_j$, which implies $\lambda_i = \lambda_j$. This completes the proof. □

E. Formulas for computing $\mathbb{E}(\Phi^*(\Phi D_n\Phi^*)^l\Phi)$.

In order to obtain the explicit formulas of cov_p and invcov_p in [20], it involves computing $\mathbb{E}(\Phi^* f(\Phi D_n\Phi^*)\Phi)$ for a differentiable function $f(x)$ (see parts A and B in section VI in [20]) and a diagonal matrix $D_n = \mathrm{diag}(d_1,\dots,d_n)$ with all d_i's positive. For instance, [20, Lemma 1] asserts that if f is differentiable on the interval $[\min\{d_i\},\max\{d_i\}]$, then

$$\frac{\partial}{\partial d_k}\int_{\Omega_{p,n}} \mathrm{Tr}(f(\Phi D_n\Phi^*))\,d\phi$$
$$= \Big(\int_{\Omega_{p,n}} \Phi^* f'(\Phi D_n\Phi^*)\Phi\,d\phi\Big)_{kk} = \mathbb{E}\big(\Phi^* f'(\Phi D_n\Phi^*)\Phi\big)_{kk}.$$

Note the eigenvalue λ_k of $\mathrm{invcov}_p(D)$ given in (10) is the left hand side of above identity with $f(x) = \log x$. To further understand the invcov_p operator, it is helpful to have the explicit formula for the eigenvalues λ_k's. By continuity and linearity, it is enough to provide formulas for computing $\mathbb{E}(\Phi^*(\Phi D_n\Phi^*)^l\Phi)$. In this subsection, we derive such formulas.

First, we observe that $\mathbb{E}(\Phi^*(\Phi D_n\Phi^*)^l\Phi)$ is still a diagonal matrix. The idea of proof is exactly the same as the proof of Proposition II-B1. We recall a fact that a matrix A is diagonal if and only if $\Omega^* A\Omega = A$ for any diagonal unitary matrix Ω. Note that

$$\Omega^*\mathbb{E}(\Phi^*(\Phi D_n\Phi^*)^l\Phi)\Omega$$
$$= \mathbb{E}\Big((\Phi\Omega)^*\big(\Phi\Omega(\Omega^* D_n\Omega)(\Phi\Omega)^*\big)^l\Phi\Omega\Big) = \mathbb{E}(\Phi^*(\Phi D_n\Phi^*)^l\Phi),$$

where we use that $\Phi\Omega$ has the same distribution as Ω, and $\Omega^* D_n\Omega = D_n$.

To compute the diagonal entries of $\mathbb{E}(\Phi^*(\Phi D_n\Phi^*)^l\Phi)$, using Lemma 1 in [20], we have

$$\big(\mathbb{E}(\Phi^*(\Phi D_n\Phi^*)^l\Phi)\big)_{ii} = \Big(\int_{\Omega_{p,n}} \Phi^*(\Phi D_n\Phi^*)^l\Phi\,d\phi\Big)_{ii}$$
$$= \frac{\partial}{\partial d_i}\int_{\Omega_{p,n}} \frac{1}{l+1}\mathrm{Tr}(\Phi D_n\Phi^*)^{l+1}\,d\phi. \tag{11}$$

Denote $N = l+1$ for convenience. By (7) and (8), we see that

$$\int_{\Omega_{p,n}} \mathrm{Tr}\big((\Phi D_n\Phi^*)^N\big)\,d\phi$$
$$= \sum_{j=0}^{p-1}(-1)^j \frac{s_{(N-j,1^j)}(I_p)}{s_{(N-j,1^j)}(I_n)} s_{(N-j,1^j)}(D_n)$$
$$= \sum_{j=0}^{p-1}(-1)^j \frac{(N+p-(j+1))!(n-(j+1))!}{(N+n-(j+1))!(p-(j+1))!} s_{(N-j,1^j)}(D_n). \tag{12}$$

Using the formula (9), one has

$$\frac{\partial s_{(N-j,1^j)}(D_n)}{\partial d_i} = \sum_{k=1}^{N} d_i^{k-1}\cdot\tilde{c}_{k-1} = \sum_{k=0}^{N-1}\tilde{c}_k d_i^k, \tag{13}$$

where we define

$$\tilde{c}_{k-1} := \sum_{\substack{\rho=(1^{r_1},2^{r_2},\dots,N^{r_N})\\ \rho\vdash N}} \chi^{(N-j,1^j)}(\rho)\frac{r_k\mathrm{Tr}(D^k)^{r_k-1}}{k^{r_k-1}r_k!}\prod_{l\neq k}\frac{\mathrm{Tr}(D^l)^{r_l}}{l^{r_l}r_l!}.$$

Therefore, combining (11) and (12), we obtain

$$\big(\mathbb{E}(\Phi(\Phi^* D\Phi)^l\Phi^*)\big)_{ii}$$
$$= \frac{1}{l+1}\sum_{j=0}^{p-1}(-1)^j\frac{(l+1+p-(j+1))!(n-(j+1))!}{(l+1+n(j+1))!(p-(j+1))!}$$
$$\cdot\frac{\partial s_{(N-j,1^j)}(D_n)}{\partial d_i}$$
$$= \sum_{k=0}^{l}\Big(\frac{\tilde{c}_k}{l+1}\sum_{j=0}^{p-1}(-1)^j\frac{(l+p-j)!(n-j-1))!}{(l+n-j)!(p-j-1)!}\Big)d_i^k$$
$$:= \sum_{k=0}^{l} a_k d_i^k.$$

The coefficients a_k depend only on D_n, p and l. Thus we are able to show $\mathbb{E}(\Phi^*(\Phi D_n\Phi^*)^l\Phi)$ is a polynomial in D_n of degree l,

$$\mathbb{E}\big(\Phi^*(\Phi D_n\Phi^*)^l\Phi\big) = \sum_{k=0}^{l} a_k D_n^k$$

where the coefficients are

$$a_k = \frac{1}{l+1}\Big(\sum_{j=0}^{p-1}(-1)^j\frac{(l+p-j)!(n-j-1))!}{(l+n-j)!(p-j-1)!}\Big)$$
$$\times\sum_{\substack{\rho=(1^{r_1},\dots,(l+1)^{r_{l+1}})\\ \rho\vdash l+1}}\chi^{(l+1-j,1^j)}(\rho)\frac{r_{k+1}\mathrm{Tr}(D^{k+1})^{r_{k+1}-1}}{(k+1)^{r_{k+1}-1}r_{k+1}!}$$
$$\cdot\prod_{l\neq k+1}\frac{\mathrm{Tr}(D^l)^{r_l}}{l^{r_l}r_l!}.$$

In the Appendix A, we provide a small dimensional example to show how to apply the derived formula for computation.

III. The mean conjugate estimator under Ewens measure

Let S_m be the set of permutations of the set $[m] := \{1,2,\dots,m\}$. For each permutation $\sigma\in S_m$, by cycle decomposition, σ can be viewed as the disjoint union of cycles of varying lengths. The Ewens measure is a probability measure on the set of permutations that depends on a parameter $\theta > 0$. In this measure, each permutation has a weight proportional to its total number of cycles. More specifically, for each permutation σ in S_m its probability is equal to

$$p_{\theta,m}(\sigma) = \frac{\theta^{\#\mathrm{cycl}(\sigma)}}{\theta(\theta+1)\dots(\theta+m-1)}, \tag{14}$$

where $\theta > 0$ and $\#\mathrm{cycl}(\sigma)$ is the number of cycles in σ. The case $\theta = 1$ corresponds to the uniform measure. This measure has recently appeared in mathematical physics models (see e.g. [2] and [11]) and one has only recently started to gain insight into the cycle structures of such random permutations.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2888734, IEEE Transactions on Information Theory

Let σ be a permutation in S_m, the corresponding permutation matrix M_σ is the $m \times m$ matrix defined as $M_\sigma(i,j) = \mathbf{1}_{\sigma(i)}(j)$. If we denote e_i to be a $1 \times m$ vector such that the i–th entry is equal to 1 and all the others entries are 0, then

$$M_\sigma = \begin{pmatrix} e_{\sigma(1)} \\ \vdots \\ e_{\sigma(m)} \end{pmatrix},$$

which is, of course, a unitary matrix. Given the sample covariance matrix K we define the new estimator for Σ as

$$K_\theta := \mathbb{E}(M_\sigma K M_\sigma^*), \tag{15}$$

where the expectation is taken with respect to the Ewens measure of parameter θ.

Theorem III-A. *Let $K = (a_{ij})$ be an $m \times m$ matrix in $\mathbb{C}^{m\times m}$. Then $K_\theta = \mathbb{E}(M_\sigma K M_\sigma^*)$ is an $m \times m$ matrix such that the diagonal terms satisfy*

$$(K_\theta)_{ii} = \frac{\theta - 1}{\theta + m - 1} a_{ii} + \frac{1}{\theta + m - 1}\mathrm{Tr}(K), \tag{16}$$

and the non–diagonal terms $(i \neq j)$ satisfy

$$\begin{aligned}(K_\theta)_{ij} &= \frac{1}{(\theta+m-2)(\theta+m-1)}\Big(\theta^2 a_{ij} + (\theta-1)a_{ji} \\ &\quad + \theta \sum_{k\neq i,j}(a_{ik}+a_{kj}) + \sum_{\substack{l\neq i, k\neq j \\ k\neq l}} a_{lk}\Big) \\ &= \frac{1}{(\theta+m-2)(\theta+m-1)}\Big((\theta^2-1)a_{ij} + (\theta-1)a_{ji} \\ &\quad + (\theta-1)\sum_{k\neq i,j}(a_{ik}+a_{kj}) + \sum_{l\neq k} a_{lk}\Big).\end{aligned} \tag{17}$$

Remark III-B. If $\theta = 1$, then

$$K_1 = \alpha \frac{\mathbf{e}\mathbf{e}^T}{m} + \beta(I_m - \frac{\mathbf{e}\mathbf{e}^T}{m}), \tag{18}$$

where

$$\alpha = \frac{\mathbf{e}^T K \mathbf{e}}{m} = \frac{\sum_{i,j} a_{ij}}{m}, \quad \beta = \frac{\mathrm{Tr}(K) - \alpha}{m-1}$$

and $\mathbf{e} = (1, 1, \ldots, 1)^T$. This result has already been shown in Prop. 2.2 of [32].

Remark III-C. If $K = D = \mathrm{diag}(d_1, \ldots, d_m)$, then

$$K_\theta = \frac{\theta - 1}{\theta + m - 1} D + \frac{\mathrm{Tr}(D)}{\theta + m - 1} I_m,$$

which corresponds to the diagonal loading.

Proof. First,

$$\begin{aligned} M_\sigma K M^* &= \begin{pmatrix} e_{\sigma(1)} \\ \vdots \\ e_{\sigma(m)} \end{pmatrix} K \begin{pmatrix} e^*_{\sigma(1)} & \cdots & e^*_{\sigma(m)} \end{pmatrix} \\ &= \Big(\sum_{l=1}^m \sum_{k=1}^m a_{kl} e_{\sigma(i)}(k) e_{\sigma(j)}(l)\Big) = (a_{\sigma(i)\sigma(j)})_{1\leq i,j\leq m}.\end{aligned}$$

For diagonal terms, recall the probability measure $p_{\theta,m}$ in (14),

$$\begin{aligned}(K_\theta)_{ii} &= \big(\mathbb{E}(M_\sigma K M_\sigma^*)\big)_{ii} = \sum_{\sigma\in S_m} p_{\theta,m}(\sigma) a_{\sigma(i)\sigma(i)} \\ &= a_{ii} \sum_{\substack{\sigma\in S_m \\ \sigma(i)=i}} p_{\theta,m}(\sigma) + \sum_{l\neq i} a_{ll} \sum_{\substack{\sigma\in S_m \\ \sigma(i)=l}} p_{\theta,m}(\sigma) \\ &= a_{ii}\frac{\theta}{\theta+m-1}\sum_{\tilde\sigma\in S_{m-1}} p_{\theta,m-1}(\tilde\sigma) \\ &\quad + \sum_{l\neq i}\frac{a_{ll}}{\theta+m-1}\sum_{\hat\sigma(l)} p_{\theta,m-1}(\hat\sigma(l)) \\ &= \frac{\theta}{\theta+m-1}a_{ii} + \frac{1}{\theta+m-1}\sum_{l\neq i} a_{ll} \\ &= \frac{\theta-1}{\theta+m-1}a_{ii} + \frac{1}{\theta+m-1}\mathrm{Tr}(K).\end{aligned}$$

Now we compute the off–diagonal terms $(K_\theta)_{ij}$ $(i \neq j)$. For $\sigma \in S_m$, if $\sigma(i) = i$ and $\sigma(j) = j$ then $\sigma = (i)(j)\sigma_1$ with $\sigma_1 \in S_{m-2}$, $\#\mathrm{cycl}(\sigma) = \#\mathrm{cycl}(\sigma_1) + 2$ and

$$p_{\theta,m}(\sigma) = \frac{\theta^2}{(\theta+m-2)(\theta+m-1)} p_{\theta,m-2}(\sigma_1).$$

If $\sigma(i) = j$ and $\sigma(j) = i$ we erase i and j from σ to obtain $\sigma_2 \in S_{m-2}$, and

$$p_{\theta,m}(\sigma) = \frac{\theta}{(\theta+m-2)(\theta+m-1)} p_{\theta,m-2}(\sigma_2).$$

If $\sigma(i) = i$ and $\sigma(j) = k \neq i, j$ then $\sigma = (i)\hat\sigma$ with $\hat\sigma \in S_{m-1}$ and $\#\mathrm{cycl}(\sigma) = \#\mathrm{cycl}(\hat\sigma) + 1$. Furthermore, we can erase j from $\hat\sigma$ to get a new permutation $\sigma_3(k) \in S_{m-2}$ such that $\#\mathrm{cycl}(\sigma_3(k)) = \#\mathrm{cycl}(\hat\sigma)$ and finally

$$p_{\theta,m}(\sigma) = \frac{\theta}{(\theta+m-2)(\theta+m-1)} p_{\theta,m-2}(\sigma_3(k)).$$

Notice that $\sum_{\sigma_3(k)} p_{\theta,m-2}(\sigma_3(k)) = 1$.

If $\sigma(i) = l \neq i, j$ and $\sigma(j) = j$ then as above we can have $\sigma_4(l) \in S_{m-2}$ such that

$$p_{\theta,m}(\sigma) = \frac{\theta}{(\theta+m-2)(\theta+m-1)} p_{\theta,m-2}(\sigma_4(l))$$

and again $\sum_{\sigma_4(l)} p_{\theta,m-2}(\sigma_4(l)) = 1$.

If $\sigma(i) = l \neq i$ and $\sigma(j) = k \neq j$ $(k \neq l)$ we exclude the case that $\sigma(i) = j, \sigma(j) = i$ and we erase i and j from σ to obtain $\sigma_5(l,k) \in S_{m-2}$. Thus

$$p_{\theta,m}(\sigma) = \frac{1}{(\theta+m-2)(\theta+m-1)} p_{\theta,m-2}(\sigma_5(l,k))$$

and $\sum_{\sigma_5(l,k)} p_{\theta,m-2}(\sigma_5(l,k)) = 1$.

Therefore, for $i \neq j$

$$
\begin{aligned}
(K_\theta)_{ij} &= \sum_{\sigma\in S_m} p_{\sigma,m}(\sigma) a_{\sigma(i)\sigma(j)} \\
&= a_{ij}\frac{\theta^2}{(\theta+m-2)(\theta+m-1)} \sum_{\sigma_1\in S_{m-2}} p_{\theta,m-2}(\sigma_1) \\
&+ a_{ji}\frac{\theta}{(\theta+m-2)(\theta+m-1)} \sum_{\sigma_2\in S_{m-2}} p_{\theta,m-2}(\sigma_2) \\
&+ \sum_{k\neq i,j} \frac{a_{ik}\cdot\theta}{(\theta+m-2)(\theta+m-1)} \sum_{\sigma_3(k)\in S_{m-2}} p_{\theta,m-2}(\sigma_3(k)) \\
&+ \sum_{l\neq i,j} \frac{a_{lj}\cdot\theta}{(\theta+m-2)(\theta+m-1)} \sum_{\sigma_4(l)\in S_{m-2}} p_{\theta,m-2}(\sigma_4(l)) \\
&+ \sum_{k\neq i,j \text{ and } k\neq l} \sum_{l\neq i,j} \sum_{\sigma_5(k,l)\in S_{m-2}} \frac{a_{lk} p_{\theta,m-2}(\sigma_5(k,l))}{(\theta+m-2)(\theta+m-1)} \\
&= \frac{1}{(\theta+m-2)(\theta+m-1)}\Big(\theta^2 a_{ij} + (\theta-1)a_{ji} \\
&\quad + \theta\sum_{k\neq i,j}(a_{ik}+a_{kj}) + \sum_{k\neq i,j \text{ and } k\neq l} a_{lk}\Big).
\end{aligned}
$$

□

IV. Hybrid Method

In this section, we combine the ideas of the first two methods to create a third hybrid method. First, we extend the definition of a permutation. For an integer $p \leq m$, let

$$
\begin{aligned}
&S_{p,m} \\
&:= \{\sigma \ : \sigma \text{ an injection from } \{1,2,\ldots,p\} \text{ to } \{1,2,\ldots m\}\}.
\end{aligned}
$$

The size of the set $S_{p,m}$ is $\frac{m!}{(m-p)!}$ and it is clear that $S_{m,m}$ is the set of all permutations on $[m]$. For $\sigma \in S_{p,m}$, the associated $p \times m$ matrix takes the form

$$
V_\sigma := \begin{pmatrix} e_{\sigma(1)} \\ e_{\sigma(2)} \\ \vdots \\ e_{\sigma(p)} \end{pmatrix},
$$

where $e_{\sigma(i)} = (e^1_{\sigma(i)}, e^2_{\sigma(i)}, \ldots, e^m_{\sigma(i)})$ is a $1 \times m$ row vector with the $\sigma(i)$–th entry 1 and all others 0. Notice

$$
V_\sigma V_\sigma^{\mathrm{T}} = I_p, \tag{19}
$$

and

$$
P_\sigma := V_\sigma^{\mathrm{T}} V_\sigma = \mathrm{diag}(b^\sigma_1, \ldots, b^\sigma_m), \tag{20}
$$

where

$$
b^\sigma_i = \sum_{l=1}^{p} (e_{\sigma(l)}(i))^2 = \begin{cases} 1 \text{ if } i \in \{\sigma(1),\ldots,\sigma(p)\}, \\ 0 \text{ otherwise.} \end{cases}
$$

Next, we use the *Ewens measure* on the permutation sets to define a probability on the set $S_{p,m}$. For each $\sigma \in S_{p,m}$, consider the set

$$
\Omega_\sigma := \Big\{\tilde{\sigma} \in S_m \ : \ \tilde{\sigma}_{\{1,\ldots,p\}} = \sigma\Big\}.
$$

In other words, Ω_σ is the set of all permutations in S_m whose restriction to the set $\{1,2,\ldots,p\}$ is equal to σ. Recall that $p_{\theta,m}$ is the *Ewens measure* on S_m with parameter θ. Define the probability measure on $S_{p,m}$ for $\sigma \in S_{p,m}$ as

$$
\mu_{\theta,m,p}(\sigma) := p_{\theta,m}(\Omega_\sigma) = \sum_{\tilde{\sigma}\in\Omega_\sigma} p_{\theta,m}(\tilde{\sigma}). \tag{21}
$$

Now we are ready to introduce two new operators

$$
K_{\theta,m,p} := \mathbb{E}\Big(V_\sigma^T (V_\sigma K V_\sigma^T) V_\sigma\Big) \tag{22}
$$

$$
\tilde{K}_{\theta,m,p} := \mathbb{E}\Big(V_\sigma^T (V_\sigma K V_\sigma^T)^+ V_\sigma\Big), \tag{23}
$$

where $(V_\sigma K V_\sigma^T)^+$ is the Moore–Penrose pseudo inverse of the matrix $V_\sigma K V_\sigma^T$. Recall the Moore–Penrose pseudo inverse of a square matrix A is a matrix A^+ of the same size and satisfies AA^+ and A^+A are both Hermitian, $AA^+A = A$ and $A^+AA^+ = A$. We use $K_{\theta,m,p}$ as an estimate for Σ and $\tilde{K}_{\theta,m,p}$ for Σ^{-1}. Now we show a few results on these new estimators.

Theorem IV-A. *Let $K = (a_{ij})$ be an $m \times m$ complex matrix. Then $K_{\theta,m,p}$ as in* (22) *is an $m \times m$ matrix such that the diagonal entries are equal to*

$$
(K_{\theta,m,p})_{ii} = \begin{cases} \frac{\theta+p-1}{\theta+m-1} a_{ii}, \text{ if } 1 \leq i \leq p, \\ \\ \frac{p}{\theta+m-1} a_{ii}, \text{ if } p+1 \leq i \leq m, \end{cases}
$$

and the non–diagonal entries, assuming $i < j$ (if $j < i$ then exchange i and j in the following expression) are equal to

$$
(K_{\theta,m,p})_{ij} = \begin{cases} \frac{(\theta+p-1)(\theta+p-2)}{(\theta+m-1)(\theta+m-2)} a_{ij}, \text{ if } 1 \leq i < j \leq p, \\ \\ \frac{(p-1)(\theta+p-1)}{(\theta+m-1)(\theta+m-2)} a_{ij}, \text{ if } 1 \leq i \leq p < j \leq m, \\ \\ \frac{p(p-1)}{(\theta+m-1)(\theta+m-2)} a_{ij}, \text{ if } p < i < j \leq m. \end{cases}
$$

Remark IV-B. In the special case that $K = \mathrm{diag}(d_1,\ldots,d_m)$ is a diagonal matrix , then

$$
K_{\theta,m,p} = \frac{p}{\theta+m-1} K + \frac{\theta-1}{\theta+m-1}\mathrm{diag}(d_1,\ldots,d_p,0,\ldots,0).
$$

For instance, if $p = 1$ and $m = 3$ then

$$
K_{\theta,3,1} = \frac{1}{\theta+2}\mathrm{diag}(\theta a_{11}, a_{22}, a_{33}).
$$

Remark IV-C. In the general case with $p = 2$ and $m = 3$ then

$$
K_{\theta,3,2} = \frac{1}{\theta+2}\begin{pmatrix} (\theta+1)a_{11} & \theta a_{12} & a_{13} \\ \theta a_{21} & (\theta+1)a_{22} & a_{23} \\ a_{31} & a_{32} & 2a_{33} \end{pmatrix}.
$$

Proof. Recall from Equation (20) that

$$
P_\sigma = V_\sigma^{\mathrm{T}} V_\sigma = \mathrm{diag}(b^\sigma_1,\ldots,b^\sigma_m),
$$

thus $V_\sigma^T (V_\sigma K V_\sigma^T) V_\sigma = (b^\sigma_i b^\sigma_j a_{ij})_{1\leq i,j\leq m}$, where

$$
b^\sigma_i = \sum_{l=1}^{p} (e_{\sigma(l)}(i))^2 = \begin{cases} 1 \text{ if } i \in \{\sigma(1),\ldots,\sigma(p)\}, \\ 0 \text{ otherwise.} \end{cases}
$$

For the diagonal entries, if $1 \le i \le p$,

$$\begin{aligned}
&(K_{\theta,m,p})_{ii} \\
&= \sum_{\sigma\in S_{m,p}} \mu_{\theta,m,p}(\sigma)(b_i^\sigma)^2 a_{ii} \\
&= a_{ii}\sum_{l=1}^{p}\sum_{\sigma\in S_{m,p},\sigma(l)=i} \mu_{\theta,m,p}(\sigma) \\
&= a_{ii}\Big(\sum_{\sigma\in S_{m,p},\sigma(i)=i}\mu_{\theta,m,p} + \sum_{l\neq i}\sum_{\sigma\in S_{m,p},\sigma(l)=i}\mu_{\theta,m,p}\Big) \\
&= a_{ii}\Big(\frac{\theta}{\theta+m-1}\sum_{\sigma'\in S_{m-1,p-1}}\mu_{\theta,m-1,p-1} \\
&\qquad + \frac{p-1}{\theta+m-1}\sum_{\sigma'\in S_{m-1,p-1}}\mu_{\theta,m-1,p-1}\Big) \\
&= \frac{\theta+p-1}{\theta+m-1}a_{ii}.
\end{aligned}$$

If $p+1 \le i \le m$,

$$\begin{aligned}
(K_{\theta,m,p})_{ii} &= \sum_{\sigma\in S_{m,p}} \mu_{\theta,m,p}(\sigma)(b_i^\sigma)^2 a_{ii} \\
&= a_{ii}\sum_{l=1}^{p}\sum_{\sigma\in S_{m,p},\sigma(l)=i}\mu_{\theta,m,p}(\sigma) \\
&= a_{ii}\Big(\frac{p}{\theta+m-1}\sum_{\sigma'\in S_{m-1,p-1}}\mu_{\theta,m-1,p-1}\Big) \\
&= \frac{p}{\theta+m-1}a_{ii}.
\end{aligned}$$

For non-diagonal entries, if $1 \le i < j \le p$, which turns out to be the most complicated case, $b_i^\sigma b_j^\sigma a_{ij}$ is non zero if $i,j \in \{\sigma(1),\dots,\sigma(p)\}$. Thus

$$(K_{\theta,m,p})_{ij} = a_{ij}\sum_{s,t\in[p],s\neq t}\ \sum_{\substack{\sigma\in S_{m,p},\\ \sigma(s)=i,\sigma(t)=j}} \mu_{\theta,m,p}(\sigma).$$

We divide the previous sum into five parts:

1) If $\sigma(i)=i$ and $\sigma(j)=j$ we "erase" i and j from the sets $[p]$ and $[m]$ to get a new injection σ_1 from $[p]\backslash\{i,j\}$ to $[m]\backslash\{i,j\}$ with $\#\text{cycl}(\sigma) = \#\text{cycl}(\sigma_1)+2$;
2) If $\sigma(s)=i$ for some $s\in[p]\backslash\{i,j\}$ and $\sigma(j)=j$ we "erase" j from the sets $[p]$ and $[m]$ and consider s and i as one number $\tilde{s}$. Then we get a new injection $\sigma_2 : [p]\cup\tilde{s}\backslash\{i,j,s\} \to [m]\cup\tilde{s}\backslash\{i,j,s\}$ with $\#\text{cycl}(\sigma) = \#\text{cycl}(\sigma_2)+1$;
3) If $\sigma(t)=j$ for some $t\in[p]\backslash\{i,j\}$ and $\sigma(i)=i$ then, similarly to case (2), by exchanging the roles of i and j we can get a new injection σ_3 with $\#\text{cycl}(\sigma) = \#\text{cycl}(\sigma_3)+1$;
4) If $\sigma(s)=i$ and $\sigma(t)=j$ with $s\neq t$ for some $s\in[p]\backslash\{i\}$ and $t\in[p]\backslash\{j\}$ then we consider s and i as a new number $\tilde{s}$ and t and j as a new number $\tilde{t}$ to get a new injection $\sigma_4 : [p]\cup\tilde{s},\tilde{t}\backslash\{i,j,s,t\} \to [m]\cup\tilde{s},\tilde{t}\backslash\{i,j,s,t\}$ with $\#\text{cycl}(\sigma) = \#\text{cycl}(\sigma_4)$;
5) If $\sigma(i)=j$ and $\sigma(j)=i$ we "erase" i and j to get a new injection $\sigma_5 : [p]\backslash\{i,j\} \to [m]\backslash\{i,j\}$ with $\#\text{cycl}(\sigma) = \#\text{cycl}(\sigma_5)+1$.

$$\begin{aligned}
&(K_{\theta,m,p})_{ij} \\
&= \frac{a_{ij}\theta^2}{(\theta+m-1)(\theta+m-2)}\sum_{\sigma_1\in S_{m-2,p-2}}\mu_{\theta,m-2,p-2}(\sigma_1) \\
&+ \frac{a_{ij}\theta(p-2)}{(\theta+m-1)(\theta+m-2)}\sum_{\sigma_2\in S_{m-2,p-2}}\mu_{\theta,m-2,p-2}(\sigma_2) \\
&+ \frac{a_{ij}\theta(p-2)}{(\theta+m-1)(\theta+m-2)}\sum_{\sigma_3\in S_{m-2,p-2}}\mu_{\theta,m-2,p-2}(\sigma_3) \\
&+ \frac{a_{ij}[(p-2)^2+(p-2)]}{(\theta+m-1)(\theta+m-2)}\sum_{\sigma_4\in S_{m-2,p-2}}\mu_{\theta,m-2,p-2}(\sigma_4) \\
&+ \frac{a_{ij}\theta}{(\theta+m-1)(\theta+m-2)}\sum_{\sigma_5\in S_{m-2,p-2}}\mu_{\theta,m-2,p-2}(\sigma_5) \\
&= \frac{(\theta+p-1)(\theta+p-2)}{(\theta+m-1)(\theta+m-2)}a_{ij}.
\end{aligned}$$

For $1 \le i \le p < j \le m$ we only need consider two cases: $s=i$ and $s\neq i$,

$$\begin{aligned}
&(K_{\theta,m,p})_{ij} \\
&= a_{ij}\frac{\theta(p-1)}{(\theta+m-1)(\theta+m-2)}\sum_{\sigma_1\in S_{m-2,p-2}}\mu_{\theta,m-2,p-2}(\sigma_1) \\
&+ a_{ij}\frac{(p-1)^2}{(\theta+m-1)(\theta+m-2)}\sum_{\sigma_2\in S_{m-2,p-2}}\mu_{\theta,m-2,p-2}(\sigma_2) \\
&= a_{ij}\frac{(p-1)(p+\theta-1)}{(\theta+m-1)(\theta+m-2)}.
\end{aligned}$$

For $p < i < j \le m$,

$$(K_{\theta,m,p})_{ij} = a_{ij}\frac{p(p-1)}{(\theta+m-1)(\theta+m-2)}.$$

□

Now we consider the estimate $\tilde{K}_{\theta,m,p}$ as in Equation (23). First we analyze the case when K is diagonal.

Theorem IV-D. *Let $D = D_m = \text{diag}(d_1,\dots,d_n,0,\dots,0)$, then for $p \le n$,*

$$\begin{aligned}
&\tilde{K}_{\theta,m,p} = \mathbb{E}\Big(V_\sigma^T(V_\sigma D V_\sigma^T)^+ V_\sigma\Big) \\
&= \frac{\theta+p-1}{\theta+m-1}D^+ - \frac{\theta-1}{\theta+m-1}\text{diag}(d_1^{-1},\dots,d_p^{-1},0,\dots,0),
\end{aligned}$$

where $D^+ = \text{diag}(d_1^{-1},\dots,d_n^{-1},0,\dots,0)$ by definition.

Proof. First we notice that

$$W_\sigma := V_\sigma D V_\sigma^T = (\sum_{i=1}^{n} d_l e_{\sigma(i)}(l)e_{\sigma(j)}(l))_{1\le i,j\le p}$$

is a diagonal matrix. For $1 \le i \le p$,

$$(W_\sigma)_{ii} = \sum_{l=1}^{n} d_l (e_{\sigma(i)}(l))^2 = \begin{cases} d_{\sigma(i)} & \text{if } \sigma(i)\in[n], \\ 0 & \text{otherwise.}\end{cases}$$

Thus

$$W_\sigma = \text{diag}(d_{\sigma(1)}\mathbf{1}_{\sigma(1)\in[n]},\dots,d_{\sigma(p)}\mathbf{1}_{\sigma(p)\in[n]})$$

and

$$W_\sigma^+ = \text{diag}\left((d_{\sigma(1)}\mathbf{1}_{\sigma(1)\in[n]})^+,\dots,(d_{\sigma(p)}\mathbf{1}_{\sigma(p)\in[n]})^+\right).$$

Next $V_\sigma^T W^+ V_\sigma = \sum_{l=1}^p (d_{\sigma(l)}\mathbf{1}_{\sigma(l)\in[n]})^+$ is still a diagonal matrix where for $1 \le i \le m$

$$(V_\sigma^T W^+ V_\sigma)_{ii} = \begin{cases} (d_{\sigma(l)}\mathbf{1}_{\sigma(l)\in[n]})^+ & \text{if } i \in \{\sigma(1),\dots,\sigma(p)\}, \\ 0 \quad \text{otherwise.} \end{cases}$$

Therefore $\tilde{K}_{\theta,m,p}$ is also diagonal and

$$(\tilde{K}_{\theta,m,p})_{ii} = \sum_{l=1}^p \sum_{\substack{\sigma\in S_{m,p},\\ \sigma(l)=i}} \mu_{\theta,m,p}(\sigma)(d_i\mathbf{1}_{i\in[n]})^+.$$

For $1 \le i \le n$,

$$(\tilde{K}_{\theta,m,p})_{ii} = d_i^{-1} \sum_{\substack{\sigma\in S_{m,p},\\ \sigma(l)=i}} \mu_{\theta,m,p}(\sigma) = \begin{cases} d_i^{-1}\frac{p}{\theta+m-1}, & \text{if } 1 \le i \le p, \\ d_i^{-1}\frac{\theta+p-1}{\theta+m-1}, & \text{if } p+1 \le i \le n. \end{cases}$$

For $n+1 \le i \le m$, $(\tilde{K}_{\theta,m,p})_{ii} = 0$. □

Obtaining a close form expression for Equation (23) in the general case seems to be much more challenging. However, we are able to obtain an inductive formula with the help of a result of Kurmayya and Sivakumar's result [16].

Theorem IV-E (Theorem 3.2, [16])**.** *Let $M = [A\ \ a] \in \mathbb{R}^{m\times n}$ be a block matrix, with $A \in \mathbb{C}^{m\times(n-1)}$ and $a \in \mathbb{C}^m$ being written as a column vector. Let $B = M^*M$ and $s = \|a\|^2 - a^*AA^+a$. Then if $s \neq 0$*

$$B^+ = \begin{pmatrix} (AA^*)^+ + s^{-1}(A^+a)(A^+a)^* & -s^{-1}(A^+a) \\ -s^{-1}(A^+a)^* & s^{-1} \end{pmatrix},$$

and if $s = 0$,

$$B^+ = \begin{pmatrix} \mathcal{B} & -\|b\|^2A^+a + A^+b \\ -\|b\|^2(A^+a)^* + (A^+b)^* & \|b\|^2 \end{pmatrix},$$

where

$$\mathcal{B} = \|b\|^2(M_1^+a)(M_1^+a)^* - (M_1^+a)(M_1^+b)^* - (M_1^+b)(M_1^+a)^*$$

and

$$b = (A^*)^+(I + A^+a(A^+a)^*)^{-1}A^+a.$$

For a non–negative definite matrix K, one can decompose

$$K = UDU^* = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_m \end{pmatrix} \begin{pmatrix} d_1 & & & \\ & d_1 & & \\ & & \ddots & \\ & & & d_m \end{pmatrix} \begin{pmatrix} u_1^* & u_2^* & \dots & u_m^* \end{pmatrix},$$

where U is a unitary matrix and $D = \mathrm{diag}(d_1,\dots,d_m)$. Then

$$W_\sigma = V_\sigma K V_\sigma^T = \begin{pmatrix} u_{\sigma(1)} \\ u_{\sigma(2)} \\ \vdots \\ u_{\sigma(p)} \end{pmatrix} D \begin{pmatrix} u_{\sigma(1)}^* & u_{\sigma(2)}^* & \dots & u_{\sigma(p)}^* \end{pmatrix} = \begin{pmatrix} \tilde{u}_{\sigma(1)} \\ \tilde{u}_{\sigma(2)} \\ \vdots \\ \tilde{u}_{\sigma(p)} \end{pmatrix} \begin{pmatrix} \tilde{u}_{\sigma(1)}^* & \tilde{u}_{\sigma(2)}^* & \dots & \tilde{u}_{\sigma(p)}^* \end{pmatrix} := M^*M,$$

where we denote

$$\tilde{u}_i = (\sqrt{d_1}u_i^1, \dots, \sqrt{d_m}u_i^m).$$

Let $M = [M_1\ a]$ with $M_1 = \begin{pmatrix} \tilde{u}_{\sigma(1)}^* & \tilde{u}_{\sigma(2)}^* & \dots & \tilde{u}_{\sigma(p-1)}^* \end{pmatrix}$ and $a = \tilde{u}_{\sigma(p)}^*$. Let $s = \|a\|^2 - a^*M_1M_1^+a$ and $b = (M_1^*)^+(I + M_1^+a(M_1^+a)^*)^{-1}M_1^+a$. By Theorem IV-E,

$$(M^*M)^+ = \begin{pmatrix} (M_1M_1^*)^+ & 0 \\ 0 & 0 \end{pmatrix} + E_\sigma$$

where the matrix E_σ is equal to

$$\begin{pmatrix} s^{-1}(M_1^+a)(M_1^+a)^* & -s^{-1}(M_1^+a) \\ -s^{-1}(M_1^+a)^* & s^{-1} \end{pmatrix}$$

if $s \neq 0$, and is equal to

$$\begin{pmatrix} \mathcal{E} & -\|b\|^2M_1^+a + M_1^+b \\ -\|b\|^2(A^+a)^* + (A^+b)^* & \|b\|^2 \end{pmatrix}$$

if $s = 0$. Here $\mathcal{E}$ represents the following expression

$$\|b\|^2(M_1^+a)(M_1^+a)^* - (M_1^+a)(M_1^+b)^* - (M_1^+b)(M_1^+a)^*.$$

Therefore,

$$\tilde{K}_{\theta,m,p} = \mathbb{E}(V_\sigma^T \begin{pmatrix} (M_1M_1^*)^+ & 0 \\ 0 & 0 \end{pmatrix} V_\sigma) + \mathbb{E}(V_\sigma^T E_\sigma V_\sigma) = \tilde{K}_{\theta,m,p-1} + \mathbb{E}(V_\sigma^T E_\sigma V_\sigma).$$

V. Performance and Simulations

In this section, we study the performance of our estimators and we compare them with other traditional methods. We focus on two types of true covariance matrix Σ of size $m \times m$. In the first example, $\Sigma = A_\alpha$ is an $m \times m$ Toeplitz covariance matrix with entries $\Sigma_{ij} = \alpha^{|i-j|}$. Here $0 < \alpha < 1$. Note that $\det(A_\alpha) = (1-\alpha^2)^{m-1}$ and thus A_α is positive semi-definite if and only if $|\alpha| \le 1$. We call A_α the *power Toeplitz matrix*. We observe that A_α is sparse in the sense that its entries decay in an exponential rate as they move away from the diagonal. In our experiment, we take $\alpha = 0.5$.

In the other example, we take $\Sigma = B_H$ to be the *long-range dependence matrix* of the form

$$\Sigma_{ij} = \frac{1}{2}[(|i-j|+1)^{2H} - 2|i-j|^{2H} + (|i-j|-1)^{2H})]$$

with $H \in [0.5, 1]$. This kind of covariance matrix presents a process exhibiting long-range dependence, for example, the increment process of fractional Brownian motion (see [3] for

instance). Contrary to the power Toepltiz matrix A_α, the off-diagonal entries of B_H (even far away from the diagonal) show long-range dependence and have non-negligible effort to the whole matrix. We choose $H = 0.9$ in the simulation.

A. Asymptotic behavior of the mean conjugate estimator under Ewens measure

In this subsection, we study the asymptotic behavior for some covariance matrices using the mean conjugate estimator under Ewens measure. For an $m \times m$ symmetric matrix K, denote the eigenvalues $\lambda_1(K) \leq \ldots \leq \lambda_m(K)$. The simplest statistic of the eigenvalues is the *empirical spectral measure*

$$\mu_m^K = \frac{1}{m}\sum_{j=1}^{m} \delta_{\lambda_j(K)}.$$

That is, for any set $E \subset \mathbb{R}$, $\mu_m(E)$ counts the proportion of eigenvalues of K that lie in E.

We show that if the diagonal entries of K are all equal to 1 and the off-diagonal entries are not too big, then by choosing θ proportional to the dimension in the Ewens measure, $K_\theta = \mathbb{E}(M_\sigma K M_\sigma^*)$ is asymptotically equivalent to a convex combination of K and the identity matrix I.

For two positive functions $f(n), g(n)$, denote $f(n) = o(g(n))$ if $f(n)/g(n) \to 0$ as $n \to \infty$ and $f(n) = O(g(n))$ if $f(n) \leq Cg(n)$ for some $C > 0$ for n sufficiently large.

Theorem V-B. *For an $m \times m$ symmetric matrix $K = (a_{ij})$, assume $a_{ii} = 1$ for all $1 \leq i \leq m$,*

$$\sum_{i\neq j} a_{ij}^2 = O(m), \quad |\sum_{l\neq k} a_{lk}| = o(m^{3/2})$$
$$\text{and} \quad \sum_{i\neq j}\big[\sum_{k\neq i,j}(a_{ik}+a_{kj})\big]^2 = o(m^3). \tag{24}$$

Then for the mean conjugate estimator K_θ as in (15) *with $\theta = \beta m$, we have*

$$\lim_{m\to\infty} \mu_m^{K_\theta} = \lim_{m\to\infty} \mu_m^{\frac{\beta^2}{(\beta+1)^2}K+(1-\frac{\beta^2}{(\beta+1)^2})I_m}.$$

Proof. By Lemma 2.3 in [1] the Levy metric of the empirical distributions of two $m \times m$ Hermitian matrix A, B satisfies

$$L(\mu_m^A, \mu_m^B) \leq \Big(\frac{1}{m}\mathrm{Tr}(A-B)(A-B)^*\Big)^{1/3}.$$

It is known (see Theorem 6, Section 4.3, [14]) that the distribution functions μ_m converges weakly to μ if and only if the Levy metric $L(\mu_m, \mu) \to 0$. Let

$$E = K_\theta - \big(I_m + \frac{\beta^2}{(\beta+1)^2}(K - I_m)\big).$$

Thus it is enough to check that

$$\frac{1}{m}\mathrm{Tr}(EE^T) = \frac{1}{m}\sum_{i,j} E_{ij}^2 \to 0$$

as $m \to \infty$.

Note that $a_{ii} = 1$ and $\theta = \beta m$. Applying Theorem III-A, we obtain $E_{ii} = 0$ and for $i \neq j$,

$$\begin{aligned}
E_{ij} &= (K_\theta)_{ij} - \frac{\beta^2}{(\beta+1)^2}a_{ij} \\
&= \Big[\frac{\beta^2 m^2 - \beta m - 2}{(\beta m + m - 2)(\beta m + m - 1)} - \frac{\beta^2}{(\beta+1)^2}\Big]a_{ij} \\
&\quad + \frac{\beta m - 1}{(\beta m + m - 2)(\beta m + m - 1)}\sum_{k\neq i,j}(a_{ik}+a_{kj}) \\
&\quad + \frac{1}{(\beta m + m - 2)(\beta m + m - 1)}\sum_{l\neq k} a_{lk}.
\end{aligned}$$

Therefore, using the basic inequality $(a+b+c)^2 \leq 3a^2 + 3b^2 + 3c^2$, we have

$$\begin{aligned}
&\frac{1}{m}\mathrm{Tr}(EE^T) = \frac{1}{m}\sum_{i\neq j} E_{ij}^2 \\
&\leq \frac{3}{m}\left[\frac{\beta^2 m^2 - \beta m - 2}{(\beta m + m - 2)(\beta m + m - 1)} - \frac{\beta^2}{(\beta+1)^2}\right]^2 \sum_{i\neq j} a_{ij}^2 \\
&\quad + \frac{3}{m}\frac{\beta^2 m^2}{(\beta m + m - 2)^4}\sum_{i\neq j}\big[\sum_{k\neq i,j}(a_{ik}+a_{kj})\big]^2 \\
&\quad + \frac{3}{m}\frac{m^2}{(\beta m + m - 2)^4}\big(\sum_{l\neq k} a_{lk}\big)^2 \\
&= o\Big(\frac{\sum_{i\neq j} a_{ij}^2}{m}\Big) + O\Big(\frac{1}{m^3}\sum_{i\neq j}\big[\sum_{k\neq i,j}(a_{ik}+a_{kj})\big]^2\Big) \\
&\quad + O\Big(\frac{1}{m^3}\big(\sum_{l\neq k} a_{lk}\big)^2\Big) = o(1)
\end{aligned}$$

by the assumption. This completes the proof. □

Remark V-C. Theorem V-B asserts if K possesses some level of sparsity in terms of (24), then asymptotically K_θ behaves like a linear convex combination of I_m and the sample covariance matrix K. We only show the convergence of the overall behavior of the eigenvalues. Indeed, if we impose stronger conditions on the entries of K, i.e.

$$\sum_{i\neq j} a_{ij}^2 = O(1), \quad |\sum_{l\neq k} a_{lk}| = o(m^{1/2})$$

and

$$\sum_{i\neq j}\big[\sum_{k\neq i,j}(a_{ik}+a_{kj})\big]^2 = o(m^2),$$

then the matrix E in the proof of Theorem V-B satisfies $\|E\|_F = o(1)$. By Weyl's inequality, one gets the individual eigenvalue of K_θ is close to that of $\frac{\beta^2}{(\beta+1)^2}K + (1-\frac{\beta^2}{(\beta+1)^2})I_m$. Similarly, by imposing extra conditions on the eigenvalues of K, one can obtain results on the perturbation of eigenvectors using the classical Davis-Kahan theorem (see for instance [29, Section V]). However, we found these imposed conditions are rather restrictive. It is an intriguing question to investigate the optimal conditions to guarantee the closeness of K_θ and $\frac{\beta^2}{(\beta+1)^2}K + (1-\frac{\beta^2}{(\beta+1)^2})I_m$.

Remark V-D. In [18], Ledoit and Wolf introduce the *linear shrinkage estimator* or the *LW estimator*

$$K_{LW} = \rho_1 I_m + \rho_2 K$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2888734, IEEE Transactions on Information Theory

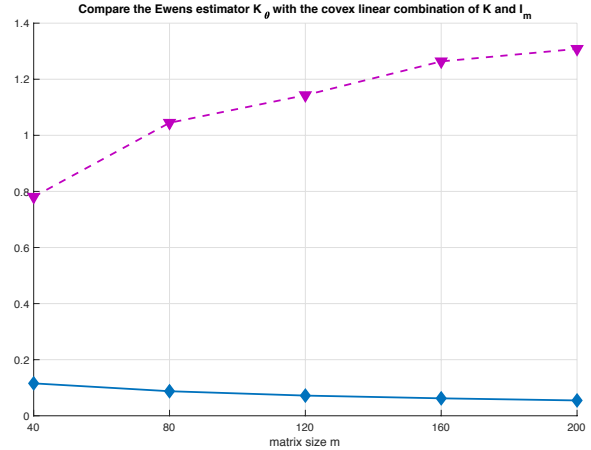
to estimate the true covariance matrix Σ. They provide the optimal parameter ρ_1^* and ρ_2^* to minimize the error $\mathbb{E}\|K_{LW} - \Sigma\|_F$ in the space of $\{\rho_1 I_m + \rho_2 K : \rho_1, \rho_1 \text{ non-random}\}$. The values of ρ_1^* and ρ_2^* actually depend on the true covariance matrix Σ. Specially, if $\Sigma_{ii} = 1$ for all i, then $\rho_1^* + \rho_2^* = 1$ and K_{LW} is the linear convex combination of I_m and K. They suggest consistent estimators $\hat{\rho}_1$ and $\hat{\rho}_2$ (see Section 3.2 in [18]) without prior knowledge of Σ. We will use the LW estimator K_{LW} with parameters $\hat{\rho}_1$ and $\hat{\rho}_2$ for performance comparison.

Remark V-E. For the power Toeplitz matrix $A_\alpha = (\alpha^{|i-j|})_{1\le i,j\le m}$. Assume $0 < \alpha < 1$, it is easy to verify that A_α satisfies (24) and thus the conclusion of Theorem V-B holds for A_α. Next let $K = (a_{ij})_{1\le i,j\le m}$ be the sample covariance matrix generated using Gaussian random variables. If the off-diagonal entries are not prominent (with high probability) in the sense of (24), then the effect of the Ewens estimator with parameter $\theta = \beta m$ is asymptotically the same as the linear shrinkage estimator. Set $\beta = 5$ and denote $\rho = \frac{\beta^2}{(\beta+1)^2}$. In Figure 2, we plot the difference

$$\|K_\theta - \big(\rho I_m + (1-\rho)K\big)\|_{NF}$$

for $m = 40, 80, 120, 160, 200$ and $n = m/2$, averaged over 50 repetitions. The blue line corresponds to the power Toeplitz matrix and the red dashed line is for the long-range dependence matrix. If the true covariance matrix Σ is the power Toeplitz matrix, then the difference between K_θ and $\rho I_m + (1-\rho)K$ under the normalized Frobenius norm is getting smaller as m, n getting larger. However, if Σ is the long-range dependence matrix, the difference between the Ewens estimator and the linear shrinkage estimator is getting bigger with the matrix size. This suggests the Ewens estimator has rather different behavior from the linear shrinkage estimator for the long-range dependence matrix.

Fig. 2. Difference between the Ewens and linear shrinkage estimators for $\Sigma = A_\alpha$ (the blue diamonds) and $\Sigma = B_H$ (the red triangles).



F. Simulation study: finite sample

In this subsection, we present some simulations to test the performance of our estimators. Let the random vector $X = (X^1, \ldots, X^m)^T$ have multivariate normal distribution $N(0, \Sigma)$. Now we have n measurements $(x_1, \ldots, x_n)$ where x_i's are independent copies of X. Let $M = (x_1, \ldots, x_n)$ and form the sample covariance matrix $K = MM^T/n$. Assume $n < m$, we want to recover Σ to the best of our knowledge.

For brevity, we call the mean conjugate estimator under Ewens measure the *Ewens estimator*, and the linear shrinkage estimator by Ledoit and Wolf [18] (see Remark V-D above) the *LW estimator*. We will compare the performance of the estimators K_{LW}, $\text{invcov}_p(K)$ and $K_\theta = \mathbb{E}(M_\sigma K M_\sigma^*)$ as well as the sample covariance matrix K itself. We will consider the error function

$$\|K - \Sigma\|_{NF} = \big(\frac{1}{m}\sum_{i,j=1}^{m}(K_{ij} - \Sigma_{ij})^2\big)^{1/2}$$

in terms of the *normalized Frobenius norm* for an estimator K of Σ for performance comparison.

Choosing the parameter θ for Ewens estimator. We first suggest how to choose the parameter θ for the Ewens estimator K_θ. Given the sample covariance matrix K, the explicit formula of K_θ is provided in Theorem III-A. We compute the formula of $\mathbb{E}\|K_\theta - \Sigma\|_{NF}^2$ in (33) in Appendix B, which is denoted by $\mathcal{G}_\Sigma(\theta)$ for brevity. Note that $\mathcal{G}_\Sigma(\theta)$ in (33) is a rational function of the form

$$\mathcal{G}_\Sigma(\theta) = \frac{a_4\theta^4 + a_3\theta^3 + a_2\theta^2 + a_1\theta + a_0}{(\theta+m-1)^2(\theta+m-2)^2},$$

where the coefficients a_i's depend on m, n and the matrix Σ. An intuitive way to choose θ is to set

$$\theta_0 = \text{argmin}_{\theta>0}\mathcal{G}_\Sigma(\theta),$$

which is the best choice under the expected quadratic normalized Frobenius loss function. We call this θ_0 the *oracle* parameter. If one has access to Σ (or a few quantities of Σ appearing in the formula (33)), then θ_0 is obtained by minimizing a rational function given m and n, and we simply take $\theta = \theta_0$ in the Ewens estimator. However, in application, it is rare that any information of Σ is known beforehand and only the sample covariance matrix K is available. To choose θ, we suggest the following method.

Since the coefficients a_i's in $\mathcal{G}_\Sigma(\theta)$ depend smoothly on m, n and the matrix Σ, a small perturbation of a_i's only leads to a small perturbation of the minimum value of $\mathcal{G}_\Sigma(\theta)$. Given the sample covariance matrix K, we replace Σ in the expression of $\mathcal{G}_\Sigma(\theta)$ with K and choose the parameter

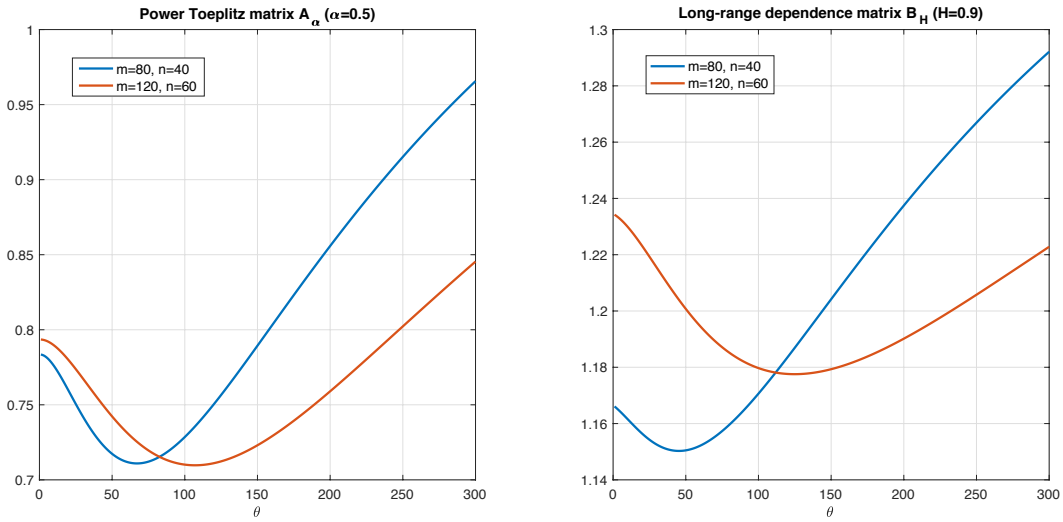
$$\hat{\theta} = \text{argmin}_{\theta>0}\mathcal{G}_K(\theta). \tag{25}$$

We estimate the true covariance matrix Σ using the Ewens estimator $K_{\hat{\theta}}$.

In Figure 3, we plot the graphs of $\sqrt{\mathcal{G}_\Sigma(\theta)}$ as a function of $\theta > 0$ for given pairs of m, n, for the power Toeplitz matrix and long-range dependence matrix respectively. In all plots, we can see that $\sqrt{\mathcal{G}_\Sigma(\theta)}$ achieves the unique minimum at an oracle value $\theta_0 > 0$.

In Table I and Table II, we numerically compute the oracle parameter θ_0 and its corresponding loss value $(\mathbb{E}\|K_{\theta_0} - \Sigma\|_{NF}^2)^{1/2} = \sqrt{\mathcal{G}_\Sigma(\theta_0)}$. We also find the estimated $\hat{\theta}$ and its loss value $\|K_{\hat{\theta}} - \Sigma\|_{NF}$, as well as the loss value $\|K - \Sigma\|_{NF}$

Fig. 3. Plots of $\sqrt{\mathcal{G}_\Sigma(\theta)}$ for $\Sigma = A_\alpha$ and $\Sigma = B_H$.



of using the sample covariance matrix K directly. These three quantities are averaged over 50 repetitions. In both tables, we note that both θ_0 and $\hat{\theta}$ increase with the matrix size m and decrease with the ratio n/m. However, our suggested $\hat{\theta}$ is quite far from the oracle θ_0. This happens possibly because the coefficients a_i's are perturbed by a large value when we replace Σ with K. It is not clear to us yet how to select a better parameter θ. Comparing Table I with Table II, we see that for the long-range dependence matrix, $\|K_{\hat{\theta}} - \Sigma\|_{NF}$ differs very little from $\sqrt{\mathcal{G}_\Sigma(\theta_0)}$, even though $\hat{\theta}$ is not a good approximation of θ_0. In all cases, directly using the sample covariance matrix K provides the worst performance.

Performance comparision. We compare the performance of the Ewens estimator, LW estimator, the Invcov_p estimator and the sample covariance matrix, for both models: power Toeplitz matrix A_α $(\alpha = 0.5)$ and long-range dependence matrix B_H $(H = 0.8)$.

For the Invcov_p estimator, we approximate the true covariance matrix Σ by $(p/m)\mathrm{invcov}_p(K)^{-1}$ and consider the loss function

$$\|(p/m)\mathrm{invcov}_p(K)^{-1} - \Sigma\|_{NF}.$$

Due to the complicated expression of the Invcov_p operator, it is hard to suggest how to turn the parameter p. In Figure 4, we plot the graphs of $\|(p/m)\mathrm{invcov}_p(K)^{-1} - \Sigma\|_{NF}$ for all values of $5 \le p \le n$ for given pairs of m, n. For the power Toeplitz matrix, the optimum values of p are approximately $p = 8$ for $m = 40, n = 20$, $p = 13$ for $m = 80, n = 40$, $p = 18$ for $m = 120, n = 60$ and $p = 26$ for $m = 160, n = 80$. For the long-range dependence matrix, the optimum values of p happen at its largest possible value n. We take these optimum values p in later comparison. Although it does not seem a fair game for other estimators, we will see that the Invcov_p estimator is never the best estimator, even with the optimum parameter p.

In Figure 5, we compare the performance of the estimators. We plot the loss function values

$$\|\mathrm{Estimator} - \Sigma\|_{NF}$$

for $m = 40, 80, 120, 160$ and $n = m/2$, averaged over 50 repetitions, for Σ the power Toeplitz matrix and the long-range dependence matrix.

For the power Toeplitz matrix (left figure in Figure 5), we observe that the LW estimator (yellow line) has the best performance and for the oracle θ_0 (red dashed line), the Ewens estimator has almost the identical performance. This is in accordance with Theorem V-B (see also Remark V-E), that is, the Ewens estimator is asymptotically equivalent to the linear shrinkage estimator $\rho I_m + (1-\rho)K$. In our finite sample study, we further observe that the Ewens estimator with oracle θ_0 performs roughly the same as the linear shrinkage estimator with the best ρ which is provided in the LW estimator. However, our suggested parameter $\hat{\theta}$ does not seem a good approximation. The invcov_p (purple line) with optimum p outperforms the Ewens estimator with $\hat{\theta}$, but is not comparable with the LW estimator. Directly using the sample covariance matrix K (green dotted line) provides the worst approximation. Nevertheless, when Σ is the power Toeplitz matrix and possesses some level of sparsity, the LW estimator is the best choice. By providing a better parameter $\hat{\theta}$, the Ewens estimator might be comparable with the LW estimator.

For the long-range dependence matrix (right figure in Figure 5), we see that the Ewens estimator (for both oracle θ_0 and estimated $\hat{\theta}$) outperforms the other estimators. Actually, the Ewens estimator $K_{\hat{\theta}}$ performs almost as good as the oracle K_{θ_0}. The LW estimator is only slightly better than using the sample covariance matrix directly. The invcov_p estimator (even with optimum p) always gives the largest errors and is not a good estimator for the long-range dependence matrix.

G. Comments

The simulations suggest that for the true covariance matrix with power decay Toeplitz structure, the Ewens estimator with the oracle parameter is asymptotically as good as the LW estimator. At present, we do not have a satisfying algorithm for choosing the parameter θ very close to the oracle value. For

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2888734, IEEE Transactions on Information Theory

TABLE I

POWER TOEPLTIZ MATRIX $\Sigma = A_\alpha$ ($\alpha = 0.5$): ORACLE AND ESTIMATED θ AND THEIR CORRESPONDING LOSS VALUES AND LOSS OF THE SAMPLE COVARIANCE MATRIX.

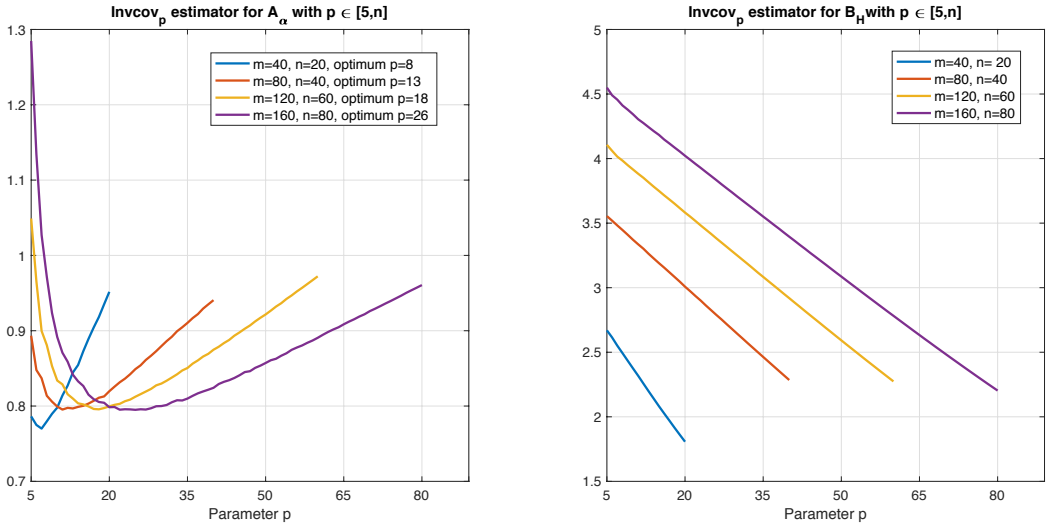
$n = m/2$	θ_0	$\sqrt{\mathcal{G}_\Sigma(\theta_0)}$	$\hat{\theta}$	$\|K_{\hat{\theta}} - \Sigma\|_{NF}$	$\|K - \Sigma\|_{NF}$
$m = 40, n = 20$	27.47	0.7145	106.01	0.8929	1.4344
$m = 80, n = 40$	67.11	0.7109	226.27	0.8908	1.4296
$m = 120, n = 60$	106.99	0.7097	350.02	0.8857	1.4240
$m = 160, n = 80$	146.93	0.7091	472.59	0.8836	1.4206
$n = m/4$	θ_0	$\sqrt{\mathcal{G}_\Sigma(\theta_0)}$	$\hat{\theta}$	$\|K_{\hat{\theta}} - \Sigma\|_{NF}$	$\|K - \Sigma\|_{NF}$
$m = 40, n = 10$	12.36	0.7661	88.95	1.1517	2.0448
$m = 80, n = 20$	36.52	0.7602	199.56	1.1473	2.0235
$m = 120, n = 30$	60.78	0.7586	308.10	1.1409	2.0081
$m = 160, n = 40$	85.06	0.7579	418.75	1.1416	2.0098

TABLE II

LONG-RANGE DEPENDENCE MATRIX $\Sigma = B_H$ ($H = 0.9$): ORACLE AND ESTIMATED θ AND THEIR CORRESPONDING LOSS VALUES AND LOSS OF THE SAMPLE COVARIANCE MATRIX.

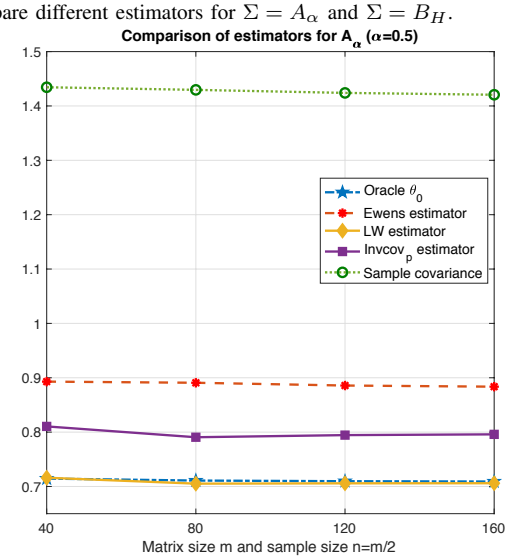
$n = m/2$	θ_0	$\sqrt{\mathcal{G}_\Sigma(\theta_0)}$	$\hat{\theta}$	$\|K_{\hat{\theta}} - \Sigma\|_{NF}$	$\|K - \Sigma\|_{NF}$
$m = 40, n = 20$	4.30	1.1263	73.83	1.1696	1.6254
$m = 80, n = 40$	45.30	1.1503	195.66	1.1829	1.5107
$m = 120, n = 60$	124.86	1.1776	325.09	1.1814	1.5194
$m = 160, n = 80$	228.00	1.1978	512.75	1.2074	1.4825
$n = m/4$	θ_0	$\sqrt{\mathcal{G}_\Sigma(\theta_0)}$	$\hat{\theta}$	$\|K_{\hat{\theta}} - \Sigma\|_{NF}$	$\|K - \Sigma\|_{NF}$
$m = 40, n = 10$	1.88	1.4787	80.60	1.5858	2.1031
$m = 80, n = 20$	5.51	1.4322	152.59	1.5186	2.1461
$m = 120, n = 30$	23.49	1.4504	261.23	1.5407	2.0868
$m = 160, n = 40$	69.52	1.4782	367.96	1.5396	2.0972

Fig. 4. Plots of $\|(p/m)\text{invcov}_p(K)^{-1} - \Sigma\|_{NF}$ for $\Sigma = A_\alpha$ and $\Sigma = B_H$.



the current suggested parameter $\hat{\theta}$, the LW estimator outperforms the Ewens estimator. However, for the true covariance matrix that has long-range dependence structure, the Ewens estimator always performs better than all other estimators considered. Even our suggested parameter $\hat{\theta}$ is not an accurate approximation to the oracle parameter, it has little influence on the performance. Provided a more accurate algorithm for choosing the parameter θ, the Ewens estimator seems a better choice than the LW estimator since it is less sensitive to the sparsity of the true covariance matrix. There are still many questions to be answered: How does the operator K_θ change the eigenvalues and eigenvectors of the original matrix K? Is there a better way to select the parameter for the Ewens estimator, using the samples? Is it possible to analyze the performance of the Ewens estimator under other loss functions? A more comprehensive understanding on the Ewens estimator K_θ will shed lights on analyzing the performance of the hybrid operators $K_{\theta,m,p}$ and $\tilde{K}_{\theta,m,p}$ defined in Section IV. We did not include simulations on the performance of these hybrid operators in this paper. However, it is an intriguing future research question to explore how the parameters p and θ affect the estimations.

Fig. 5. Compare different estimators for $\Sigma = A_\alpha$ and $\Sigma = B_H$.



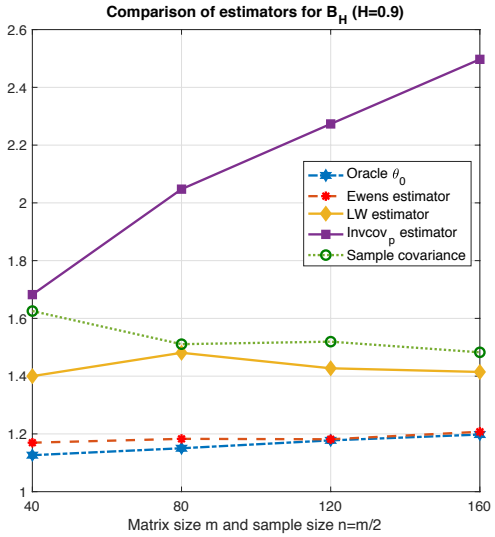


TABLE III
BORDER–STRIP TABLEAUX OF SHAPE λ_j AND TYPE ρ

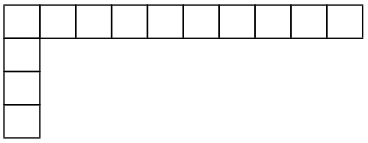
	$\rho=(1,1,1)$	$\rho=(1,2)$	$\rho=(3)$
$\lambda_0=(3)$	1 2 3	1 2 2	1 1 1
$\lambda_1=(2,1)$	1 2 / 3 & 1 3 / 2	Does not exist	1 1 / 1
$\lambda_2=(1,1,1)$	1 / 2 / 3	1 / 2 / 2	1 / 1 / 1

$\chi^{\lambda_j}(\rho)$	$\rho=(1,1,1)$	$\rho=(1,2)$	$\rho=(3)$
$\lambda_0=(3)$	1	1	1
$\lambda_1=(2,1)$	2	0	-1
$\lambda_2=(1,1,1)$	1	-1	1

APPENDIX A
SMALL DIMENSIONAL EXAMPLES FOR COMPUTING $\mathbb{E}(\Phi^*(\Phi D_n \Phi^*)^l \Phi)$

In this appendix, we provide small dimensional examples for computing $\mathbb{E}(\Phi^*(\Phi D_n \Phi^*)^l \Phi)$ using formulas derived in Section II-E.

Let $\lambda_j = (N-j, 1^j)$ be the partition of N with j ones. This one has a hook shape with $N-j$ blocks in the row and $j+1$ blocks in the column.



For $l=1$, it was shown in [20] that

$$\mathbb{E}(\Phi^*(\Phi D_n \Phi^*)^l \Phi) = \frac{p(np-1)}{n(n^2-1)} D_n + \frac{p(n-p)}{n(n^2-1)} \mathrm{Tr}(D_n) I_n.$$

For $l=2$ and $\rho=(1,1,1),(1,2),(3) \vdash 3$, we list all border–strip tableaux of shape λ_j and type ρ in the Table III.

Thus,

$$s_{\lambda_0}(D) = \frac{\mathrm{Tr}(D)^3}{3!} + \frac{\mathrm{Tr}(D)\mathrm{Tr}(D^2)}{2} + \frac{\mathrm{Tr}(D^3)}{3},$$
$$s_{\lambda_1}(D) = 2\frac{\mathrm{Tr}(D)^3}{3!} - \frac{\mathrm{Tr}(D^3)}{3},$$
$$s_{\lambda_2}(D) = \frac{\mathrm{Tr}(D)^3}{3!} - \frac{\mathrm{Tr}(D)\mathrm{Tr}(D^2)}{2} + \frac{\mathrm{Tr}(D^3)}{3}.$$

and

$$\frac{\partial s_{\lambda_0}}{\partial d_i} = d_i^2 + \mathrm{Tr}(D) d_i + \frac{\mathrm{Tr}(D)^2 + \mathrm{Tr}(D^2)}{2},$$
$$\frac{\partial s_{\lambda_1}}{\partial d_i} = -d_i^2 + \mathrm{Tr}(D)^2,$$
$$\frac{\partial s_{\lambda_2}}{\partial d_i} = d_i^2 - \mathrm{Tr}(D) d_i + \frac{\mathrm{Tr}(D)^2 - \mathrm{Tr}(D^2)}{2}.$$

Furthermore,

$$\begin{aligned}
&\left(\mathbb{E}(\Phi^*(\Phi D \Phi^*)^2 \Phi)\right)_{ii} \\
&= \frac{1}{3}\sum_{j=0}^{2} (-1)^j \frac{(2+p-j)!(n-j-1)!}{(2+n-j)!(p-j-1)!} \frac{\partial s_{\lambda_j}(D)}{\partial d_i} \\
&= (c_0+c_1+c_2) d_i^2 + (c_0-c_2)\mathrm{Tr}(D) d_i \\
&\quad + c_0 \frac{\mathrm{Tr}(D)^2 + \mathrm{Tr}(D^2)}{2} - c_1 + c_2 \frac{\mathrm{Tr}(D)^2 - \mathrm{Tr}(D^2)}{2},
\end{aligned}$$

where

$$c_0 = \frac{1}{3}\frac{(2+p)!(n-1)!}{(2+n)!(p-1)!}, \quad c_1 = \frac{1}{3}\frac{(1+p)!(n-2)!}{(1+n)!(p-2)!},$$

and

$$c_2 = \frac{1}{3}\frac{p!(n-3)!}{n!(p-3)!}.$$

Finally,

$$\begin{aligned}
&\mathbb{E}\Big(\Phi^*(\Phi D \Phi^*)^2 \Phi\Big) \\
&= (c_0+c_1+c_2) D^2 + (c_0-c_2)\mathrm{Tr}(D) D \\
&\quad + \Big(c_0 \frac{\mathrm{Tr}(D)^2 + \mathrm{Tr}(D^2)}{2} - c_1 \mathrm{Tr}(D)^2 \\
&\quad + c_2 \frac{\mathrm{Tr}(D)^2 - \mathrm{Tr}(D^2)}{2}\Big) I_n.
\end{aligned}$$

APPENDIX B
COMPUTING $\mathbb{E}\|K_\theta - \Sigma\|_{NF}^2$

In this section, we compute the explicit formula for $\mathbb{E}\|K_\theta - \Sigma\|_{NF}^2 = \frac{1}{m}\mathbb{E}\|K_\theta - \Sigma\|_F^2$ and express the formula in terms of Σ. We assume the m-dimensional random vector X has the normal distribution $N(0,\Sigma)$. Let $x_1,\dots,x_n$ be n independent copies of X. Recall $M = (x_1,\dots,x_n)$ and $K = MM^T/n = (a_{ij})$. Then

$$\mathbb{E}\|K_\theta - \Sigma\|_F^2 = \sum_{i=1}^m \mathbb{E}(K_\theta - \Sigma)_{ii}^2 + \sum_{i\neq j}\mathbb{E}(K_\theta - \Sigma)_{ij}^2.$$

By Theorem III-A, we first have

$$\begin{aligned}(K_\theta - \Sigma)_{ii}^2 &= \Big(\frac{\theta-1}{\theta+m-1}a_{ii} + \frac{1}{\theta+m-1}\mathrm{Tr}K - \Sigma_{ii}\Big)^2\\ &= \frac{(\theta-1)^2}{(\theta+m-1)^2}a_{ii}^2 + \frac{1}{(\theta+m-1)^2}(\mathrm{Tr}K)^2\\ &\quad + \Sigma_{ii}^2 + \frac{2(\theta-1)}{(\theta+m-1)^2}a_{ii}\mathrm{Tr}K\\ &\quad - \frac{2(\theta-1)}{\theta+m-1}a_{ii}\Sigma_{ii} - \frac{2}{\theta+m-1}\Sigma_{ii}\mathrm{Tr}K.\end{aligned}$$

Note that $\Sigma_{ii} = \mathbb{E}a_{ii}$ and $\mathbb{E}\mathrm{Tr}K = \sum_{i=1}^m \Sigma_{ii}$. Thus

$$\begin{aligned}&\sum_{i=1}^m \mathbb{E}(K_\theta - \Sigma)_{ii}^2\\ &= \frac{(\theta-1)^2}{(\theta+m-1)^2}\Big(\sum_{i=1}^m \mathbb{E}a_{ii}^2\Big) + \frac{m\mathbb{E}(\mathrm{Tr}K)^2}{(\theta+m-1)^2}\\ &\quad + \sum_{i=1}^m \Sigma_{ii}^2 + \frac{2(\theta-1)}{(\theta+m-1)^2}\mathbb{E}(\mathrm{Tr}K)^2\\ &\quad - \frac{2(\theta-1)}{\theta+m-1}\sum_{i=1}^m \Sigma_{ii}^2 - \frac{2}{\theta+m-1}(\mathbb{E}\mathrm{Tr}K)^2.\end{aligned}$$

Plugging in

$$\mathbb{E}(\mathrm{Tr}K)^2 = \sum_{i=1}^m \mathbb{E}a_{ii}^2 + \sum_{i\neq j}\mathbb{E}a_{ii}a_{jj},$$

we get

$$\begin{aligned}&\sum_{i=1}^m \mathbb{E}(K_\theta - \Sigma)_{ii}^2\\ &= \frac{\theta^2+m-1}{(\theta+m-1)^2}\Big(\sum_{i=1}^m \mathbb{E}a_{ii}^2\Big) - \frac{\theta-m-1}{\theta+m-1}\sum_{i=1}^m \Sigma_{ii}^2\\ &\quad + \frac{2\theta+m-2}{(\theta+m-1)^2}\sum_{i\neq j}\mathbb{E}a_{ii}a_{jj} - \frac{2}{\theta+m-1}\Big(\sum_{i=1}^m \Sigma_{ii}\Big)^2.\end{aligned}$$

For brevity, denote $\beta = (\theta+m-1)(\theta+m-2)$. Next, by the formula obtained in Theorem III-A, we get for $i\neq j$

$$\begin{aligned}(K_\theta - \Sigma)_{ij}^2 &= \frac{1}{\beta^2}\big((\theta^2-1)a_{ij} + (\theta-1)a_{ji}\\ &\quad + (\theta-1)\sum_{k\neq i,j}(a_{ik}+a_{kj}) + \sum_{l\neq k}a_{lk} - \beta\Sigma_{ij}\big)^2\\ &= \frac{1}{\beta^2}\big(\theta(\theta-1)a_{ij} + (\theta-1)\sum_{k\neq i}a_{ik}\\ &\quad + (\theta-1)\sum_{k\neq j}a_{jk} + \sum_{l\neq k}a_{lk} - \beta\Sigma_{ij}\big)^2.\end{aligned}$$

Expanding the square above and taking the expectation over the sum of all $i\neq j$, one obtains

$$\begin{aligned}&\sum_{i\neq j}\mathbb{E}(K_\theta - \Sigma)_{ij}^2\\ &= \frac{1}{\beta^2}\Big[\theta^2(\theta-1)^2\sum_{i\neq j}\mathbb{E}a_{ij}^2 + 2(\theta-1)^2\sum_{i\neq j}\mathbb{E}(\sum_{k\neq i}a_{ik})^2\\ &\quad + m(m-1)\mathbb{E}(\sum_{i\neq j}a_{ij})^2 + \beta^2\sum_{i\neq j}\Sigma_{ij}^2\\ &\quad + 4\theta(\theta-1)^2\sum_{i\neq j}\sum_{k\neq i}\mathbb{E}a_{ij}a_{ik} + 2\theta(\theta-1)\mathbb{E}(\sum_{i\neq j}a_{ij})^2\\ &\quad - 2\beta\theta(\theta-1)\sum_{i\neq j}\Sigma_{ij}^2 + 2(\theta-1)^2\sum_{i\neq j}\mathbb{E}(\sum_{k\neq i}a_{ik})(\sum_{k\neq j}a_{jk})\\ &\quad + 4(\theta-1)\mathbb{E}\big(\sum_{i\neq j}\sum_{k\neq i}a_{ik}\big)(\sum_{l\neq k}a_{lk})\\ &\quad - 4\beta(\theta-1)\sum_{i\neq j}\sum_{k\neq i}\Sigma_{ij}\Sigma_{ik} - 2\beta(\sum_{l\neq k}\Sigma_{lk})^2\Big].\end{aligned}$$

We observe in the above summation that

$$\sum_{i\neq j}\mathbb{E}(\sum_{k\neq i}a_{ik})^2 = (m-1)\sum_{i=1}^m\mathbb{E}(\sum_{k\neq i}a_{ik})^2,$$

$$\sum_{i\neq j}\sum_{k\neq i}\mathbb{E}a_{ij}a_{ik} = \sum_{i=1}^m\mathbb{E}(\sum_{k\neq i}a_{ik})^2,$$

$$\sum_{i\neq j}\mathbb{E}(\sum_{k\neq i}a_{ik})(\sum_{k\neq j}a_{jk}) = \mathbb{E}(\sum_{l\neq k}a_{lk})^2 - \sum_{i=1}^m\mathbb{E}(\sum_{k\neq i}a_{ik})^2$$

and

$$\begin{aligned}&\mathbb{E}\big(\sum_{i\neq j}\sum_{k\neq i}a_{ik}\big)(\sum_{l\neq k}a_{lk})\\ &= (m-1)\mathbb{E}\big(\sum_{i=1}^m\sum_{k\neq i}a_{ik}\big)(\sum_{l\neq k}a_{lk})\\ &= (m-1)\mathbb{E}(\sum_{l\neq k}a_{lk})^2.\end{aligned}$$

Thus, after simplification, we get

$$\begin{aligned}&\sum_{i\neq j}\mathbb{E}(K_\theta - \Sigma)_{ij}^2\\ &= \frac{1}{\beta^2}\Big[\theta^2(\theta-1)^2(\sum_{i\neq j}\mathbb{E}a_{ij}^2)\end{aligned}$$

$$+ 2(\theta-1)^2(2\theta+m-2)\sum_{i=1}^{m}\mathbb{E}(\sum_{j\neq i}a_{ij})^2$$
$$+ \big[m(m-1)+2(\theta-1)(2\theta+2m-3)\big]\mathbb{E}(\sum_{i\neq j}a_{ij})^2$$
$$- 4\beta(\theta-1)\sum_{i=1}^{m}(\sum_{i\neq j}\Sigma_{ij})^2$$
$$- 2\beta(\sum_{i\neq j}\Sigma_{ij})^2+\big(\beta^2-2\beta\theta(\theta-1)\big)\sum_{i\neq j}\Sigma_{ij}^2\Big].$$

Finally, we get the explicit formula

$$\mathbb{E}\|K_\theta-\Sigma\|_F^2 \tag{26}$$
$$= \frac{\theta^2+m-1}{(\theta+m-1)^2}\big(\sum_{i=1}^{m}\mathbb{E}a_{ii}^2\big)-\frac{\theta-m-1}{\theta+m-1}\sum_{i=1}^{m}\Sigma_{ii}^2$$
$$+\frac{2\theta+m-2}{(\theta+m-1)^2}\sum_{i\neq j}\mathbb{E}a_{ii}a_{jj}-\frac{2}{\theta+m-1}\big(\sum_{i=1}^{m}\Sigma_{ii}\big)^2$$
$$+\frac{\theta^2(\theta-1)^2}{(\theta+m-1)^2(\theta+m-2)^2}(\sum_{i\neq j}\mathbb{E}a_{ij}^2)$$
$$+\frac{2(\theta-1)^2(2\theta+m-2)}{(\theta+m-1)^2(\theta+m-2)^2}\sum_{i=1}^{m}\mathbb{E}(\sum_{j\neq i}a_{ij})^2$$
$$+\frac{2(\theta-1)(2\theta+2m-3)+m(m-1)}{(\theta+m-1)^2(\theta+m-2)^2}\mathbb{E}(\sum_{i\neq j}a_{ij})^2$$
$$-\frac{4(\theta-1)}{(\theta+m-1)(\theta+m-2)}\sum_{i=1}^{m}(\sum_{j\neq i}\Sigma_{ij})^2$$
$$-\frac{2}{(\theta+m-1)(\theta+m-2)}(\sum_{i\neq j}\Sigma_{ij})^2$$
$$+\Big[1-\frac{2\theta(\theta-1)}{(\theta+m-1)(\theta+m-2)}\Big](\sum_{i\neq j}\Sigma_{ij}^2). \tag{27}$$

Since we assume $X=(X^1,\dots,X^m)^T\sim N(0,\Sigma)$, we can further express (26) in terms of the entries of Σ. We use x_s^i to denote the ith entry of the vector x_s. Note that $a_{ij}=\frac{1}{n}\sum_{s=1}^{n}x_s^ix_s^j$ by our definition of K. Besides, $\mathbb{E}K=\Sigma$. We also use the following facts about multivariate normal distribution:

$$\mathbb{E}(X^i)^2=\Sigma_{ii},\quad \mathbb{E}(X^i)^4=3\Sigma_{ii}^2,\quad \mathbb{E}X^iX^j=\Sigma_{ij}$$

and

$$\mathbb{E}X^iX^{k_1}X^jX^{k_2}=\Sigma_{ik_1}\Sigma_{jk_2}+\Sigma_{ij}\Sigma_{k_1k_2}+\Sigma_{ik_2}\Sigma_{jk_1}$$

for arbitrary $1\le i,j,k_1,k_2\le m$.

It is elementary to verify the following calculation.

$$\sum_{i=1}^{m}\mathbb{E}a_{ii}^2=\frac{1}{n^2}\sum_{i=1}^{m}\big[\mathbb{E}\sum_{s=1}^{n}(x_s^i)^4+\sum_{s\neq t}\mathbb{E}(x_s^i)^2\mathbb{E}(x_t^i)^2\big]$$
$$=\frac{1}{n^2}\sum_{i=1}^{m}\big[3n\Sigma_{ii}^2+n(n-1)\Sigma_{ii}^2\big]=\frac{n+2}{n}\sum_{i=1}^{m}\Sigma_{ii}^2 \tag{28}$$

and

$$\sum_{i\neq j}\mathbb{E}a_{ii}a_{jj}=\frac{1}{n^2}\sum_{i\neq j}\sum_{s,t=1}^{n}\mathbb{E}(x_s^i)^2(x_t^j)^2$$
$$=\frac{1}{n^2}\sum_{i\neq j}\big[n\mathbb{E}(X^i)^2(X^j)^2+n(n-1)\Sigma_{ii}\Sigma_{jj}\big]$$
$$=\sum_{i\neq j}\Sigma_{ii}\Sigma_{jj}+\frac{2}{n}\sum_{i\neq j}\Sigma_{ij}^2 \tag{29}$$

and

$$\sum_{i\neq j}\mathbb{E}a_{ij}^2=\frac{1}{n^2}\sum_{i\neq j}\big[n\mathbb{E}(X^i)^2(X^j)^2+n(n-1)(\mathbb{E}X^iX^j)^2\big]$$
$$=\frac{1}{n}\sum_{i\neq j}\Sigma_{ii}\Sigma_{jj}+\frac{n+1}{n}\sum_{i\neq j}\Sigma_{ij}^2. \tag{30}$$

Similarly, we also obtain

$$\sum_{i=1}^{m}\mathbb{E}(\sum_{j\neq i}a_{ij})^2$$
$$=\frac{1}{n^2}\sum_{i=1}^{m}\sum_{j_1,j_2\neq i}(n\Sigma_{ii}\Sigma_{j_1j_2}+2n\Sigma_{ij_1}\Sigma_{ij_2}$$
$$+n(n-1)\Sigma_{ij_1}\Sigma_{ij_2})$$
$$=\frac{1}{n}\sum_{i=1}^{m}\sum_{j_1,j_2\neq i}\Sigma_{ii}\Sigma_{j_1j_2}+\frac{n+1}{n}\sum_{i=1}^{m}(\sum_{j\neq i}\Sigma_{ij})^2 \tag{31}$$

and

$$\mathbb{E}(\sum_{i\neq j}a_{ij})^2=\frac{1}{n^2}\sum_{i_1\neq j_1,i_2\neq j_2}(n\Sigma_{i_1j_1}\Sigma_{i_2j_2}+n\Sigma_{i_1i_2}\Sigma_{j_1j_2}$$
$$+n\Sigma_{i_1j_2}\Sigma_{i_2j_1}+n(n-1)\Sigma_{i_1j_1}\Sigma_{i_2j_2})$$
$$=(\sum_{i\neq j}\Sigma_{ij})^2+\frac{2}{n}\sum_{i_1\neq j_1,i_2\neq j_2}\Sigma_{i_1i_2}\Sigma_{j_1j_2}. \tag{32}$$

Also note that

$$\sum_{i\neq j}\Sigma_{ii}\Sigma_{jj}=(\sum_{i=1}^{m}\Sigma_{ii})^2-\sum_{i=1}^{m}\Sigma_{ii}^2.$$

Thus we obtain the following formula of $\mathbb{E}\|K_\theta-\Sigma\|_{NF}^2$ by plugging (28)-(32) to (26) and dividing m on both sides:

$$\frac{1}{m}\mathbb{E}\|K_\theta-\Sigma\|_F^2 \tag{33}$$
$$=\Big[\frac{(n+2)(\theta^2+m-1)}{n(\theta+m-1)^2}-\frac{\theta-m-1}{\theta+m-1}-\frac{2\theta+m-2}{(\theta+m-1)^2}$$
$$-\frac{\theta^2(\theta-1)^2}{n(\theta+m-1)^2(\theta+m-2)^2}\Big]\frac{1}{m}\sum_{i=1}^{m}\Sigma_{ii}^2$$
$$+\Big[\frac{(2\theta+m-2)}{(\theta+m-1)^2}+\frac{\theta^2(\theta-1)^2}{n(\theta+m-1)^2(\theta+m-2)^2}$$
$$-\frac{2}{\theta+m-1}\Big]\frac{1}{m}(\sum_{i=1}^{m}\Sigma_{ii})^2$$
$$+\Big[\frac{2(2\theta+m-2)}{n(\theta+m-1)^2}+\frac{(n+1)\theta^2(\theta-1)^2}{n(\theta+m-1)^2(\theta+m-2)^2}$$
$$+1-\frac{2\theta(\theta-1)}{(\theta+m-1)(\theta+m-2)}\Big]\frac{1}{m}\sum_{i\neq j}\Sigma_{ij}^2$$

$$
\begin{aligned}
&+\Big[\frac{2(n+1)(\theta-1)^2(2\theta+m-2)}{n(\theta+m-1)^2(\theta+m-2)^2} \\
&\quad-\frac{4(\theta-1)}{(\theta+m-1)(\theta+m-2)}\Big]\frac{1}{m}\sum_{i=1}^{m}(\sum_{j\neq i}\Sigma_{ij})^2 \\
&+\Big[\frac{2(\theta-1)(2\theta+2m-3)+m(m-1)}{(\theta+m-1)^2(\theta+m-2)^2} \\
&\quad-\frac{2}{(\theta+m-1)(\theta+m-2)}\Big]\frac{1}{m}(\sum_{i\neq j}\Sigma_{ij})^2 \\
&+\frac{2(\theta-1)^2(2\theta+m-2)}{n(\theta+m-1)^2(\theta+m-2)^2}(\frac{1}{m}\sum_{i=1}^{m}\sum_{j_1,j_2\neq i}\Sigma_{ii}\Sigma_{j_1j_2}) \\
&+\frac{2(\theta-1)(2\theta+2m-3)+m(m-1)}{n(\theta+m-1)^2(\theta+m-2)^2} \\
&\qquad\cdot(\frac{2}{m}\sum_{i_1\neq j_1,i_2\neq j_2}\Sigma_{i_1i_2}\Sigma_{j_1j_2}). \qquad (34)
\end{aligned}
$$

ACKNOWLEDGMENT

We would like to thank the anonymous referees for their careful reading and many insightful suggestions.

REFERENCES

[1] Z. D. Bai. Methodologies in spectral analysis of large-dimensional random matrices, a review. *Statist. Sinica*, vol. 9, no. 3, pp. 611–677, 1999.

[2] V. Betz, D, Ueltschi and Y. Velenik. Random permutations with cycle weights *Ann. Appl. Probab.*, vol. 21, no. 1, pp. 312331, 2011.

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

[4] P. J. Bickel and E. Levina. Covariance regularization by thresholding. *The Annals of Statistics*, vol. 36, no. 6, pp. 2577–2604, 2008.

[5] A. Böttcher and S. M. Grudsky. Spectral properties of banded Toeplitz matrices. Society for Industrial and Applied Mathematics, 2005.

[6] A. Böttcher and B. Silbermann. Introduction to large truncated Toeplitz matrices. Springer Verlag, 1999.

[7] 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.

[8] T. T. Cai and H. H. Zhou. Minimax estimation of large covariance matrices under l_1 norm, *Statistica Sinica*, 2011.

[9] N. R. Draper and H. Smith. Applied Regression Analysis (Wiley Series in Probability and Statistics). Wiley-Interscience, 1998.

[10] N. El Karoui. Operator norm consistent estimation of large-dimensional sparse covariance matrices. *The Annals of Statistics*, pp. 2717–2756, 2008.

[11] N. Ercolani and D. Ueltschi. Cycle structure of random permutations with cycle weights, 2011.

[12] Y. V. Fyodorov and B. A. Khoruzhenko. A few remarks on colour–flavour transformations, truncations of random unitary matrices, Berezin reproducing kernels and Selberg-type integrals. *Journal of Physics A: Mathematical and Theoretical*, 40(4):669, 2007.

[13] H. Fulton, Representation Theory, *Springer*, 1991.

[14] J. Galambos. *Advanced probability theory*, volume 10. CRC, 1995.

[15] R. M. Gray. *Toeplitz and circulant matrices: A review*. Information Systems Laboratory, Stanford University, 1971.

[16] T. Kurmayya and K. C. Sivakumar. Moore-penrose inverse of a gram matrix and its nonnegativity. *Journal of Optimization Theory and Applications*, vol. 139, no. 1, pp.201–207, 2008.

[17] O. Ledoit and M. Wolf. Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *Annals of statistics*, pp. 1081–1102, 2002.

[18] O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, vol. 88, no. 2, pp. 365–411, 2004.

[19] I. Macdonald, Symmetric functions and Hall Polynomials *Clarendon Press*, Oxford University Press, New York, 1995.

[20] T. Marzetta, G. Tucci, and S. Simon. A random matrix–theoretic approach to handling singular covariance estimates, IEEE Transactions on Information Theory, vol. 57, no. 9, pp. 6256–6271, 2011.

[21] X. Mestre. Improved estimation of eigenvalues and eigenvectors of covariance matrices using their sample estimates. *Information Theory, IEEE Transactions on*, vol. 54, no. 11, pp. 5113–5129, 2008.

[22] X. Mestre and M. A. Lagunas. Diagonal loading for finite sample size beamforming: an asymptotic approach. *Robust adaptive beamforming*, pp. 201–257, 2006.

[23] R. Muirhead. Aspects of Multivariate Statistical Theory. *John Wiley & Sons*, New York, 1982.

[24] C. D. Richmond, R. Rao Nadakuditi, and A. Edelman. Asymptotic mean squared error performance of diagonally loaded capon–mvdr processor. In *Signals, Systems and Computers, 2005. Conference Record of the Thirty-Ninth Asilomar Conference on*, pp. 1711–1716, 2005.

[25] A. J. Rothman, P.J. Bickel, E. Levina and J. Zhou. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, vol. 2, pp. 494–515, 2008.

[26] B. Sagan. The Symmetric Group: Representations. *Combinatorial Algorithms, and Symmetric Functions*, Springer, 2nd edition, 2010.

[27] J. SchŁfer and K. Strimmer A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4.1 (2005): 32.

[28] R. P. Stanley. Enumerative Combinatorics: Volume 2. *Cambridge university press Cambridge*, 1999.

[29] G. W. Stewart and J. G. Sun. Matrix perturbation theory. *Academic Press*, 1990.

[30] P. Stoica and A. Nehorai. MUSIC maximum likelihood and Cramr-Rao bound. *IEEE Trans. Acoust. Speech Signal Processing*, vol. 37, pp. 720-741, 1989.

[31] P. Stoica and A. Nehorai. Performance study of conditional and unconditional direction-of-arrival estimation. *IEEE Trans. Acoust. Speech Signal Processing*, vol. 38, pp. 1783-1795, 1990.

[32] M. A. G. Viana. The covariance structure of random permutation matrices. *Algebraic methods in statistics and probability: AMS Special Session on Algebraic Methods and Statistics, April 8–9, 2000, University of Notre Dame, Notre Dame, Indiana*, pp. 287–303, 2001.

[33] K. Wieand. Eigenvalue distributions of random permutation matrices. *The Annals of Probability*, 28.4 (2000): 1563-1587.

[34] K. Wieand. Eigenvalue distributions of random unitary matrices. *Probability Theory and Related Fields*, 123.2 (2002): 202-224.

[35] W. B. Wu and M. Pourahmadi. Banding sample autocovariance matrices of stationary processes *Statistica Sinica*, vol. 19, no. 4, pp. 1755, 2009.

Gabriel H. Tucci was born in Montevideo, Uruguay. He's a mathematician and electrical engineer with expertise in probability, statistics, random matrices, graph theory, machine learning, and applied mathematics. He received a Ph.D. in Mathematics from Texas A&M University in 2008 under the direction of Ken Dykema. He worked at Bell Labs (Murray Hill NJ), where he was a member of Technical Staff in the Industrial Mathematics and Operations Research department for more than 4 years. While at Bell Labs he worked on a variety of research projects including random matrix theory, probability, stochastic analysis/processes, machine learning, statistical analysis, information theory, complex networks, hyperbolic geometry/graphs/networks, traffic congestion, energy-related problems and electricity consumption/prediction/forecast. He was the recipient of grants from the AFOSR, NIST, and GERI in South Korea. In 2013, he joined Barclays as a quantitative researcher in the electronic trading group for the equities desk. He worked on problems related to portfolio trading strategies, implementation shortfall strategies, optimal order placement, dark pools, TCA, impact models, and other topics. In 2015, he joined Swiss Re as a quant/underwriter in the energy, weather and commodity market trading desk. He worked on the valuation of financial derivatives whose underlying are a combination of weather indices as well as energy and commodity prices. In 2016 he joined the Global Markets Equities team in Citi. More recently, in 2018, he was promoted to Quant Global Head of Cash Trading at Citi.

Ke Wang received the B.S. from University of Science and Technology of China in 2006 and the Ph.D. in Mathematics from Rutgers University in 2013. She was a postdoctoral researcher at Institute for Mathematics and its Applications (IMA) from 2013-2015. She later joined Hong Kong University of Science and Technology (HKUST) Jockey Club Institute for Advanced Study as a postdoctoral fellow from 2015-2016. She is now a research assistant professor in the department of Mathematics at HKUST. Her interests lie in the fields of probabilistic combinatorics, random structures, random matrices and their applications.