On the Optimal Stacking of Information-plus-Noise Matrices

Øyvind Ryan, *Member, IEEE*

***Abstract*— Observations of the form $\mathbf{D}+\mathbf{X}$, where $\mathbf{D}$ is a matrix representing information, and $\mathbf{X}$ is a random matrix representing noise, can be grouped into a compund observation matrix, on the same information + noise form. There are many ways the observations can be stacked into such a matrix, for instance vertically, horizontally, or quadratically. An unbiased estimator for the spectrum of $\mathbf{D}$ can be formulated for each stacking scenario in the case of Gaussian noise. We compare these spectrum estimators for the different stacking scenarios, and show that all kinds of stacking actually decrease the variance of the corresponding spectrum estimators when compared to just taking an average of the observations, and find which stacking is optimal in this sense. When the number of observations grow, however, it is shown that the difference between the estimators is marginal. with only the cases of vertical and horizontal stackings having a higher variance asymptotically.**

***Index Terms*— Gaussian matrices, Random Matrices, free convolution, deconvolution, spectrum estimation.**

I. Introduction

Random matrices find applications in many fields of research, such as digital communication [1], mathematical finance [2] and nuclear physics [3]. Free probability theory [4], [5], [6], [7], [8] has strong connections with random matrix theory, and can be used for high dimensional statistical inference by addressing the following questions:

Given $\mathbf{A}$, $\mathbf{B}$ two $n\times n$ independent square Hermitian (or symmetric) random matrices:
1) Can one derive the eigenvalue distribution of $\mathbf{A}$ from those of $\mathbf{A}+\mathbf{B}$ and $\mathbf{B}$?
2) Can one derive the eigenvalue distribution of $\mathbf{A}$ from those of $\mathbf{AB}$ and $\mathbf{B}$?

More generally, such questions can be asked starting with any functional of the involved random matrices. If 1) or 2) can be answered for given random matrices $\mathbf{A}$ and $\mathbf{B}$, the corresponding operation for finding the eigenvalue distribution is called deconvolution. Deconvolution can be easier to perform in the large n-limit, and the literature contains result in this respect both for Vandermonde matrices [9], [10], and Gaussian matrices. For Gaussian matrices in the large n-limit, there exist several results which can be stated in terms of the Stieltjes transform [11], [12], and several results on inference as in 1) and 2) [13], [14], [15], [16], [17], [18], [19].

This work was supported by Alcatel-Lucent within the Alcatel-Lucent Chair on flexible radio at SUPELEC
Øyvind Ryan is with the Centre of Mathematics for Applications, University of Oslo, P.O. Box 1053 Blindern, NO-0316 Oslo, NORWAY, and with SUPELEC, Gif-sur-Yvette, France. Email: oyvindry@ifi.uio.no, Phone: +47 22 85 59 25, Fax: +47 22 85 43 49

In this contribution we will focus on the method of moments [4], [20], which also applies in the large n-limit for Gaussian matrices, and has a more general role in the context of free probability [8]. The method of moments is based on the relations between the moments of the matrices involved. The p-th moment of an $n\times n$ random matrix $\mathbf{A}$ is defined as

$$t_{\mathbf{A}}^{n,p}=\mathbb{E}\left[\mathrm{tr}(\mathbf{A}^p)\right]=\int \lambda^p d\rho(\lambda) \tag{1}$$

where $\mathbb{E}[\cdot]$ is the expectation, tr the normalized trace, and $d\rho(\lambda)=\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\delta(\lambda-\lambda_i)\right)$ the associated empirical mean measure, with $\{\lambda_1,...,\lambda_n\}$ the eigenvalues of $\mathbf{A}$. The p-th moment is thus defined only if $\mathbb{E}\left[\mathrm{tr}(\mathbf{A}^p)\right]$ exists. Both the Stieltjes transform of the eigenvalue distribution of A, and the moments of A, can be used to retrieve the eigenvalues of A, and can therefore be used for spectrum estimation. For many types of random matrices, $t_{\mathbf{A}}^{n,p}$ converges almost surely when $n\to\infty$ to an analytical expression $t_{\mathbf{A}}^{p}$, depending only on some specific parameters, such as the distribution of the entries of $\mathbf{A}$. This enables us to reduce the dimensionality of the problem, which can be useful also for n of moderate size, due to the rate of convergence.

Deconvolution in terms of moments turns out to be quite simple if asymptotic freeness holds, and can be performed using the R- and S-transforms [8]. Although Gaussian matrices are asymptotically free, they are not free for any finite dimensions. In fact, the literature contains no non-trivial cases of finite matrices which display freeness. In [21], a moment-based framework which applies for Gaussian matrices of any dimensions is presented, and which has many similarities with the moment-based free probability framework. This contribution willl address the following questions on how this framework can be best adapted for spectrum estimation purposes:

1) Observations of the form $\mathbf{D}+\mathbf{X}$ allow for stacking into a compound observation matrix. Can the compound observation matrix be supported within the same framework?
2) Is one stacking of the observations better than another, for the purpose of spectrum estimation?

A popular way of combining observations is the so-called sample covariance matrix, which essentially results from stacking observations of a random vector horizontally into a compound matrix. Results in this paper will actually challenge this construction, at least for the simple case under consideration, showing that it is not always the best way of combining observations. Another facet of stacking is that it can make asymptotic results more applicable, and eliminate the need for results on finite-dimensional matrices. This can be useful,

since asymptotic results can be simpler to obtain, and have a nicer form. We will give a partial answer to the above questions in this paper, in the sense that we characterize the stacking of observations which is optimal in terms of the variance of the corresponding spectrum estimators, and we characterize what we gain in comparison with methods where observations are not stacked.

The paper is organized as follows. Section II provides background essentials on random matrix theory needed to state the main results, and define different ways of stacking observations. Section III defines the unbiased spectrum estimators we will analyse, states the form of this which is useful for implementation. Section IV states the main result, which concerns finding the stacking which gives a minimum variance estimator. In Section V we present some useful simulations verifying the results.

II. Random matrix Background Essentials

In the following, upper boldface symbols will be used for matrices, and lower symbols will represent scalar values. $(.)^T$ will denote the transpose operator, $(.)^\star$ conjugation, and $(.)^H = \left((.)^T\right)^\star$ Hermitian transpose. The $n \times n$ identity matrix will be written as $\mathbf{I}_n$. We let Tr be the (non-normalized) trace for square matrices, defined by

$$\mathrm{Tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii},$$

where a_{ij} is the (i,j) entry of the $n \times n$ matrix $\mathbf{A}$. We also let tr be the normalized trace, defined by $\mathrm{tr}(\mathbf{A}) = \frac{1}{n}\mathrm{Tr}(\mathbf{A})$. In addition to the moments, we define the mixed moments by

$$A_{p_1,\ldots,p_k} = \mathbb{E}\left[\mathrm{tr}\left(\mathbf{A}^{p_1}\right)\cdots\mathrm{tr}\left(\mathbf{A}^{p_k}\right)\right].$$

When $\mathbf{A}$ is non-random, the moments are simply $A_p = \mathrm{tr}(\mathbf{A}^p)$. $\mathbf{X}$ will denote a standard complex Gaussian matrix, meaning that it has i.i.d. complex Gaussian entries with zero mean and unit variance Moreover, the real and imaginary parts of the entries are independent, each with variance $\frac{1}{2}$.

From $L = L_1 L_2$ observations of an $n \times N$ random matrix $\mathbf{Y}$, we can form the $(nL_1) \times (NL_2)$ compound observation matrix, denoted $\mathbf{Y}_{L_1,L_2}$, by stacking the observations into a $L_1 \times L_2$ block matrix in a given order. Similarly, if $\mathbf{D}$ is non-random, we will denote by $\mathbf{D}_{L_1,L_2}$ the compound matrix formed in the same way from $\mathbf{D}$. We will be concerned with the following question:

Assume that

$$\mathbf{Y} = \mathbf{D} + \mathbf{X}, \tag{2}$$

with $\mathbf{D}$ non-random and $\mathbf{X}$ Gaussian, both of size $n \times N$. How can we infer the spectrum of $\mathbf{D}$ from independent observations $\mathbf{Y}_1, \ldots, \mathbf{Y}_L$ of $\mathbf{Y}$ in an unbiased way, and with minimum variance?

[21] states a moment-based method to infer the spectrum of $\frac{1}{N}\mathbf{D}\mathbf{D}^H$ from that of $\frac{1}{N}\mathbf{Y}\mathbf{Y}^H$. Since $\mathbf{X}_{L_1,L_2}$ is also Gaussian, and since the compound matrices satisfy $\mathbf{Y}_{L_1,L_2} = \mathbf{D}_{L_1,L_2} + \mathbf{X}_{L_1,L_2}$, the same method can be used to infer the spectrum of $\frac{1}{NL_2}\mathbf{D}_{L_1,L_2}\mathbf{D}_{L_1,L_2}^H$ from the compound observation matrix. But since

$$\mathrm{tr}\left(\left(\frac{1}{NL_2}\mathbf{D}_{L_1,L_2}\mathbf{D}_{L_1,L_2}^H\right)^p\right) = L_1^{p-1}\mathrm{tr}\left(\left(\frac{1}{N}\mathbf{D}\mathbf{D}^H\right)^p\right), \tag{3}$$

the method from [21] applied to the compound observation matrix help us infer the spectrum of $\frac{1}{N}\mathbf{D}\mathbf{D}^H$ also. We will state the corresponding estimator in Section III. To ease notation, we will let $D_p = \mathrm{tr}\left(\left(\frac{1}{N}\mathbf{D}\mathbf{D}^H\right)^p\right)$ in the following. We will see that different stackings L_1, L_2 give rise to different spectrum estimators, all of them unbiased, and we will compare their variances. Different stackings will be denoted as follows:

Definition 1: Assume that we are given a number of $L > 0$ observations We will call a stacking scenario

- horizontal if the observations are grouped into a $1 \times L$ block matrix,
- vertical if the observations are grouped into a $L \times 1$ block matrix,
- rectangular (of ratio c with $0 < c < \infty$) if the observations are grouped into a $L_1 \times L_2$ block matrix with $L = L_1 \times L_2$, and with $(L_1, L_2) = \arg\min \left| c - \frac{L_1}{L_2} \right|$.

These three types of stackings are also denoted by H, V, and R, respectively.

Horizontal stacking in particular has been considered previously [22].

III. Formulation of the estimator

To state our estimators, we need the following concepts, taken from [21]. These concepts are better motivated geometrically in terms of pairings of Gaussian elements. This is further explained in Appendix A:

Definition 2: Let p be a positive integer. By a partial permutation we mean a one-to-one mapping π between two subsets ρ_1, ρ_2 of $\{1, \ldots, p\}$. We denote by SP_p the set of partial permutations of p elements. When $\pi \in \mathrm{SP}_p$, we define $\hat{\pi} \in \mathrm{SP}_{2p}$ by

$$\hat{\pi}(2j-1) = 2\pi^{-1}(j), \quad j \in \rho_2$$
$$\hat{\pi}(2j) = 2\pi(j) - 1, \quad j \in \rho_1.$$

We associate to π an equivalence relation $\rho = \rho(\pi)$ on $\{1, \ldots, 2p\}$ generated by

$$j \sim_\rho \hat{\pi}(j) + 1,\ j+1 \sim_\rho \hat{\pi}(j), \text{ for } j \in \rho_1, \tag{4}$$

and we let $k(\rho)$ and $l(\rho)$ denote the number of blocks of ρ consisting of only even or odd numbers, respectively. With $\mathcal{D} = (\rho_1 \bigcup \rho_2)^c$ $\sigma = \sigma(\pi)$ is defined as the equivalence relation on $\mathcal{D}$ generated by the relations

$$k \sim_\sigma k+1 \quad \text{if} \quad k, k+1 \in \mathcal{D} \tag{5}$$
$$k \sim_\sigma l \quad \text{if} \quad k, l \in \mathcal{D}, k+1 \sim_\rho l, \tag{6}$$

and we let $kd(\rho)$ be the number of blocks of ρ contained within the even numbers which intersect $\mathcal{D} \cup (\mathcal{D}+1)$, and $ld(\rho)$ be the number of blocks of ρ contained within the odd numbers which intersect $\mathcal{D} \cup (\mathcal{D}+1)$.

Our estimators can now be stated as follows:

Lemma 1: Let $\mathbf{Y} = \mathbf{D} + \mathbf{X}$, and

$$\begin{aligned} D_{p_1,\ldots,p_k} &= \operatorname{tr}\left(\left(\frac{1}{N}\mathbf{D}\mathbf{D}^H\right)^{p_1}\right)\operatorname{tr}\left(\left(\frac{1}{N}\mathbf{D}\mathbf{D}^H\right)^{p_2}\right)\cdots \\ &\quad \times\operatorname{tr}\left(\left(\frac{1}{N}\mathbf{D}\mathbf{D}^H\right)^{p_k}\right) \qquad (7) \\ Y_{p_1,\ldots,p_k} &= \operatorname{tr}\left(\left(\frac{1}{N}\mathbf{Y}\mathbf{Y}^H\right)^{p_1}\right)\operatorname{tr}\left(\left(\frac{1}{N}\mathbf{Y}\mathbf{Y}^H\right)^{p_2}\right)\cdots \\ &\quad \times\operatorname{tr}\left(\left(\frac{1}{N}\mathbf{Y}\mathbf{Y}^H\right)^{p_k}\right), \qquad (8) \end{aligned}$$

and let $|\rho_1|, |\rho_2|$ be the number of elements in ρ_1, ρ_2. Then

$$\begin{aligned} \widehat{D_{p_1,\ldots,p_k}} = \sum_{\substack{\pi\in \mathrm{SP}_p \\ \pi=\pi(\rho_1,\rho_2,q)}} & (-1)^{|\rho_1|}\frac{n^{|\sigma(\pi)|-k}}{N^{|\rho_1|}} \\ & \times N^{k(\rho(\pi))-kd(\rho(\pi))}n^{l(\rho(\pi))-ld(\rho(\pi))} \\ & \times Y_{l_1,\ldots,l_r} \qquad (9) \end{aligned}$$

is an unbiased estimator for $D_{p_1,\ldots,p_k}$, i.e. $\mathbb{E}\left(\widehat{D_{p_1,\ldots,p_k}}\right) = D_{p_1,\ldots,p_k}$ for all p. In particular, $\widehat{D_p}$ is an unbiased estimator for D_p.

Similarly, given $L = L_1L_2$ observations $\mathbf{Y}_i = \mathbf{D} + \mathbf{X}_i$ $(1 \le i \le L)$, form the compound observation matrix $\mathbf{Y}_{L_1,L_2}$ and let instead Y_p be the moments

$$Y_p = \operatorname{tr}\left(\left(\frac{1}{NL_2}\mathbf{Y}_{L_1,L_2}\mathbf{Y}_{L_1,L_2}^H\right)^p\right). \qquad (10)$$

Then

$$\begin{aligned} &\widehat{D_{p_1,\ldots,p_k,L_1,L_2}} \\ &= L_1^{k-p_1-\cdots-p_k}\sum_{\substack{\pi\in \mathrm{SP}_p \\ \pi=\pi(\rho_1,\rho_2,q)}} (-1)^{|\rho_1|}\frac{(nL_1)^{|\sigma(\pi)|-k}}{(NL_2)^{|\rho_1|}} \\ &\quad \times(L_2N)^{k(\rho(\pi))-kd(\rho(\pi))}(L_1n)^{l(\rho(\pi))-ld(\rho(\pi))} \\ &\quad \times Y_{l_1,\ldots,l_r} \qquad (11) \end{aligned}$$

is also an unbiased estimator for $D_{p_1,\ldots,p_k}$ for any L_1, L_2. In particular $\widehat{D_{p,L_1,L_2}}$ is an unbiased estimator for D_p [1].

Note that there is a constant term in $\widehat{D_{p_1,\ldots,p_k}}$, coming from π where $\rho_1 = \rho_2 = \{1, ..., p\}$. The proof of Lemma 1 can be found in Appendix B. The appendix concentrates on establishing Equation (9), since Equation (11) is immediate: the term trailing $L_1^{k-p_1-\cdots-p_k}$ in Equation (11) is an unbiased estimator for the moments $F_p = \operatorname{tr}\left(\left(\frac{1}{NL_2}\mathbf{D}_{L_1,L_2}\mathbf{D}_{L_1,L_2}^H\right)^p\right)$, once Equation (9) is established, so that the entire right hand side of Equation (11) is an unbiased estimator for

$$\begin{aligned} & L_1^{k-p_1-\cdots-p_k}F_{p_1,\ldots,p_k} \\ &= L_1^{k-p_1-\cdots-p_k}F_{p_1}\cdots F_{p_k} \\ &= (L_1^{1-p_1}F_{p_1})\cdots(L_1^{1-p_k}F_{p_k}) \\ &= D_{p_1}\cdots D_{p_k}, \end{aligned}$$

where we have used Equation (3).

[1]There can also be a known noise variance σ_2 present, so that Model (2) takes the form $\mathbf{Y} = \mathbf{D} + \sigma_2\mathbf{X}$. Generalizations of the result to this case is obviuos, and the implementation used supports arbitrary noise variances.

Equation (9) is important: it reveals a similarity in the expressions for convolution and deconvolution, so that the implementation of Equation (13) from [21] for convolution can also be used for deconvolution. Secondly, Equation (9) can be used for obtaining an expression for the variances of $\widehat{D_p}$, as will be seen.

IV. Statement of the main result

The following result says that, among the estimators we have defined, any rectangular stacking asymptotically has the lowest variance, and that horizontal and vertical stackings, as well as averaging of observations, have a higher variance asymptotically. We will let $v_{p,\cdot,L}$ denote the variance of $\widehat{D_{p,L_1,L_2}}$, with $L = L_1L_2$ the number of observations, and $\cdot$ the stacking (H, V, or R). We will in addition let A denote taking the average of L applications of $\widehat{D_p}$, i.e. $\frac{1}{L}\sum_{i=1}^{L}\widehat{D_p}(\mathbf{Y_i})$ and denote the variance of the corresponding estimator by $v_{p,A,L}$. It is clear that this is also an unbiased estimator for D_p, with $v_{p,A,L} = \frac{1}{L}v_{p,R,1}$, since observations are assumed independent. When P is a polynomial in several variables, also denote by the degree of P, or $\deg(P)$, the highest sum of the exponents in any term therein. We will use the notation $O(L^k)$ to denote any polynomial in L where there are no terms of higher order than k.

Theorem 1: The variance $v_{p,\cdot,L}$ of $\widehat{D_{p,L_1,L_2}}$ is $O(L^{-1})$. Moreover,

$$\lim_{L\to\infty} Lv_{1,\cdot,L} = \frac{2}{nN}D_1 + \frac{1}{nN},$$

where $\cdot$ can be H, V, R, or A. For $p \ge 2$ we have that

$$\begin{aligned} \lim_{L\to\infty} Lv_{p,R,L} &= \frac{2p^2}{nN}D_{2p-1} \\ \lim_{L\to\infty} Lv_{p,V,L} &= \frac{2p^2}{nN}D_{2p-1} + \frac{p^2}{N^2}D_{2p-2} \\ \lim_{L\to\infty} Lv_{p,H,L} &= \frac{2p^2}{nN}D_{2p-1} + \frac{p^2}{nN}D_{2p-2} \\ \lim_{L\to\infty} Lv_{p,A,L} &= \frac{2p^2}{nN}D_{2p-1} + \left(\frac{p^2}{N^2} + \frac{p^2}{nN}\right)D_{2p-2} \\ &\quad + Q(D_{2p-3}, ..., D_1), \end{aligned}$$

where Q is a polynomial in $D_{2p-3}, D_{2p-4}, \ldots, D_1$ of degree $2p-2$, with only positive coefficients. In particular, all rectangular stackings asymptotically have the same variance, and

$$\begin{aligned} \lim_{L\to\infty} Lv_{p,R,L} &\le \lim_{L\to\infty} Lv_{p,V,L} \le \lim_{L\to\infty} Lv_{p,A,L} \\ \lim_{L\to\infty} Lv_{p,R,L} &\le \lim_{L\to\infty} Lv_{p,H,L} \le \lim_{L\to\infty} Lv_{p,A,L} \end{aligned}$$

(since $Q(D_{2p-3}, ..., D_1) \ge 0$ and all $D_p \ge 0$). Also, the variance decreases with L for a fixed stacking aspect ratio, and, for a given L and any rectangular stackings R_1, R_2 into $L = L_1^{(1)} \times L_2^{(1)}$ and $L = L_1^{(2)} \times L_2^{(2)}$ observations, respectively. $v_{p,R_1,L} < v_{p,R_2,L}$ if and only if

$$\max\left(\frac{nL_1^{(1)}}{NL_2^{(1)}}, \frac{nL_2^{(1)}}{NL_1^{(1)}}\right) < \max\left(\frac{nL_1^{(2)}}{NL_2^{(2)}}, \frac{nL_2^{(2)}}{NL_1^{(2)}}\right). \qquad (12)$$

Also, $v_{p,\cdot,L} < v_{p,A,L}$ for any stacking.

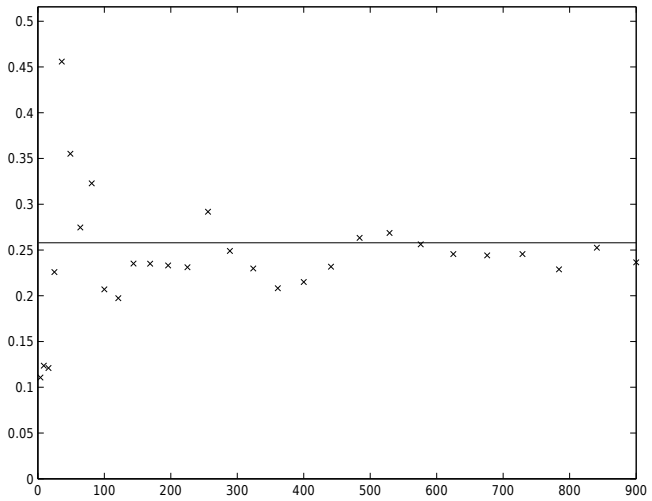


Fig. 1. The estimator expressed in (10) with quadratic stacking applied with different number of observations. $\mathbf{D}$ is a 4×4 matrix. The actual third moment of $\frac{1}{N}\mathbf{D}\mathbf{D}^H$ is also shown.

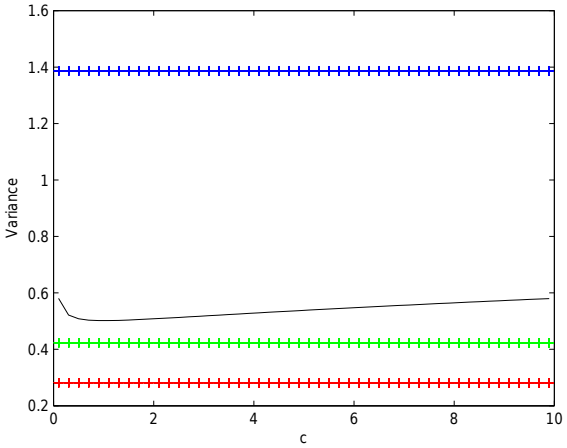
The proof of Theorem 1 can be found in Appendix C. Note that Equation (12) expresses that the first compound observation matrix is "more square" than the second, thereby providing an even stronger conclusion in the case of any given finite number of observations. Theorem 1 is a statement on the leading order term of the variances of moments of certain random matrices. Similarly, the recently developed theory of second order freeness [23], [24], [25] studies the leading order terms in covariance expressions for many types of random matrices.

V. Simulations

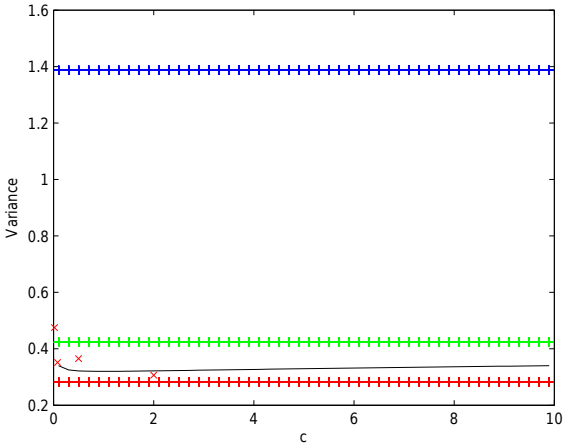
In [21], Proposition 1 was implemented. In the following, the estimators and the expression (16) for the variance are computed using this implementation [2].

Figure 1 shows results for the third moment estimator expressed in (11), applied to a diagonal matrix $\mathbf{D}$, with diagonal entries assumed to be $2, 1, 1, 0.5$ (i.e. $n = N = 4$). The estimator was applied to quadratic stackings of $L = 1, 4, 9, 16, ...$, all the way up to $L = 900$ observations. Although Theorem 1 says that the quadratic stacking is optimal, the difference between the different estimators may be hard to detect in practice, since differences may be small. Figure 2 gives a comparison for the actual variances for different number of observations and different stacking aspect ratios, verifying Theorem 1. The theoretical limits for rectangular and horizontal stacking and averaging are also shown. We have used the same 4×4 matrix, and computed an exact expression for the variance obtained in Appendix B. As predicted by Theorem 1, the variance tends towards the theoretical lower bounds for rectangular and horizontal stacking when the number of observations grow. For $L = 50$ observations, to verify the results, we have also plotted

[2] A guide to the Matlab source code running the following simulations can be found in [26].



(a) $L = 5$



(b) $L = 50$. Empirical variances are also shown for $c = 0.02, 2/25, 0.5, 2$.

Fig. 2. Figures displaying $Lv_{3,\cdot,L}$ for the different estimators for Model (2), for different number of observations L. A diagonal matrix $\mathbf{D}$ with entries $2, 1, 1, 0.5$ on the diagonal has been chosen. The three rectangular lines are the theoretical limits $\lim_{L\to\infty} Lv_{3,\cdot,L}$ for rectangular stacking, horizontal stacking, and averaging, as predicted by Theorem 1, in increasing order. It is seen that aspect ratio near 1 gives lowest variance, and that the variances decreases towards the theoretical limit predicted by Theorem 1 when L increases.

the empirical variances

$$\frac{1}{K-1}\sum_{i=1}^{K}(x_i - \bar{x})^2,$$

where $\{x_i\}_{i=1}^K$ are K outputs from the estimator (i.e. a number of KL observations is needed, since each run of the estimator requires L observations), and $\bar{x} = \frac{1}{K}\sum_{i=1}^{K} x_i$ is the mean. We have set $K = 1000$, and indicated the empirical variances for $L_1 = 1, 2, 5, 10$, which correspond to $c = 0.02, 0.08, 0.5, 2$.

VI. Conclusion and Further Work

We have analyzed an unbiased spectrum estimator from observations of the form $\mathbf{D} + \mathbf{X}$, where $\mathbf{X}$ is Gaussian, and shown that the way the observations are stacked can play a role. More specifically, it is desirable to stack observations so

that the compound observation matrix is square, as this will give rise to spectrum estimators with lowest possible variance. Asymptotically (i.e. when the number of observations grow to infinity), the variance of the estimators are the same, with only vertical and horizontal stacking and averaging displaying different asymptotic behaviour. All cases of stacking were shown to reduce the variance when compared to averaging.

This paper only considers estimators which perform averaging or stacking of observations. Future work could consider non-linear ways of combining observations, and compare results on these with the results obtained here. Theorem 1 should also have some significance when the noise is not Gaussian, since many random matrices with non-Gaussian, i.i.d. entries display the same asymptotic behaviour as Gaussian matrices. Future work could also consider this, and explore to which extent results generalize to the finite regime.

APPENDIX A
GEOMETRIC INTERPRETATION OF $\pi, \rho(\pi)$ AND $\sigma(\pi)$

The following geometric interpretations explain the concepts in Definition 2, and is a summary of [21]:

- We draw k disconnected circles with $2p_1, 2p_2, ..., 2p_k$ edges, respectively, and number the edges clockwise from 1 to $2p_1+\cdots+2p_k$. The set ρ_1 is visualized as a subset of the even edges $(2, 4, ..., 2p)$ under the mapping $i \to 2i$, ρ_2 is visualized as a subset of the odd edges $(1, 3, ..., 2p-1)$ under the mapping $i \to 2i-1$.
- $\pi(i) = j$ means that the corresponding even and odd edges $2i$ and $2j-1$ are identified, and with opposite orientation.
- The vertices on the circles are also labeled clockwise, so that edge i borders to vertices i and $i+1$. When edges are identified as above, we also get an identification between the vertices bordering to the edges. This gives rise to an equivalence relation on the vertices. The corresponding partition in $\mathcal{P}(2p_1+\cdots+2p_k)$ of the equivalence classes of vertices is written $\rho(\pi)$, where $\mathcal{P}(n)$ denotes the partition of n elements.
- It turns out that a block of ρ either consists of odd numbers only (odd vertices), or of even numbers only (even vertices). $k(\rho(\pi))$ is defined as the number of blocks consisting of even numbers only, $l(\rho(\pi))$ as the number of blocks consisting of odd numbers only.
- Edges from ρ_1 and ρ_2 are called random edges, other edges are called deterministic edges. $kd(\rho(\pi))$ is the number of even equivalence classes of vertices bordering to a deterministic edge, $ld(\rho(\pi))$ is defined similarly for odd equivalence classes of vertices.
- $\sigma = \sigma(\pi)$ is the partition where the blocks are the connected components of deterministic edges after identification of edges.
- By the graph of random edges we will mean the graph constructed when we, after the identification of edges, join vertices which are connected with a path of deterministic edges, and afterwards remove the set of deterministic edges,

The quantities $k(\rho(\pi)) - kd(\rho(\pi))$ and $l(\rho(\pi)) - ld(\rho(\pi))$ in Equation (13) thus describe the number of even and odd vertices, respectively, which do not border to deterministic edges in the graph after the identification of edges. Note that when $\rho_1 = \rho_2 = \{1, ..., p\}$, $\sigma(\pi)$ is a partition of zero elements. In this case we define $D_{l_1,...,l_r} = 1$.

APPENDIX B
THE PROOFS OF LEMMA 1

We will need the following result, taken from [21], where the general statement is for the case when $\mathbf{D}$ is random, independent from $\mathbf{X}$:

Proposition 1: Let $\mathbf{X}$ be an $n \times N$ standard, complex, Gaussian matrix and $\mathbf{D}$ be an $n \times N$ non-random matrix. Set

$$\begin{aligned} D_{p_1,...,p_k} &= \operatorname{tr}\left(\left(\frac{1}{N}\mathbf{D}\mathbf{D}^H\right)^{p_1}\right)\operatorname{tr}\left(\left(\frac{1}{N}\mathbf{D}\mathbf{D}^H\right)^{p_2}\right)\cdots \\ &\quad \times \operatorname{tr}\left(\left(\frac{1}{N}\mathbf{D}\mathbf{D}^H\right)^{p_k}\right) \\ M_{p_1,...,p_k} &= \mathbb{E}\left[\operatorname{tr}\left(\left(\frac{1}{N}(\mathbf{D}+\mathbf{X})(\mathbf{D}+\mathbf{X})^H\right)^{p_1}\right)\right. \\ &\quad \times \operatorname{tr}\left(\left(\frac{1}{N}(\mathbf{D}+\mathbf{X})(\mathbf{D}+\mathbf{X})^H\right)^{p_2}\right)\cdots \\ &\quad \left.\times \operatorname{tr}\left(\left(\frac{1}{N}(\mathbf{D}+\mathbf{X})(\mathbf{D}+\mathbf{X})^H\right)^{p_k}\right)\right], \end{aligned}$$

We have that

$$\begin{aligned} M_{p_1,...,p_k} = \sum_{\substack{\pi\in \mathrm{SP}_p \\ \pi=\pi(\rho_1,\rho_2,q)}} & \frac{n^{|\sigma(\pi)|-k}}{N^{|\rho_1|}} \\ & \times N^{k(\rho(\pi))-kd(\rho(\pi))} n^{l(\rho(\pi))-ld(\rho(\pi))} \\ & \times D_{l_1,...,l_r}, \end{aligned} \tag{13}$$

and where $l_1, \ldots, l_r$ are the cardinalities of the blocks of $\sigma(\pi)$, divided by 2.

Lemma 1 will be proved from Proposition 1. To ease the expressions in the following, we will set

$$P_\pi(n, N) = N^{k(\rho(\pi))-kd(\rho(\pi))} n^{l(\rho(\pi))-ld(\rho(\pi))}.$$

To prove that the estimators $\widehat{D_{p_1,...,p_k}}$ expressed in (9) are unbiased, we will first find alternative recursive expressions for them, and prove by induction that these are unbiased. Assume that we have found unbiased estimators $\widehat{D_{q_1,...,q_l}}$ building on Equation (13), whenever $q_1+\cdots+q_l < p_1+\cdots+p_k$. Define $\widehat{D_{p_1,...,p_k}}$ by reorganizing Equation (13) to

$$\begin{aligned} \widehat{D_{p_1,...,p_k}} &= Y_{p_1,...,p_k} - \textstyle\sum_{p\geq 1}\sum_{\substack{\pi\in \mathrm{SP}_p \\ \pi=\pi(\rho_1,\rho_2,q)}} \frac{n^{|\sigma(\pi)|-k}}{N^{|\rho_1|}} \\ &\quad \times P_\pi(n, N)\widehat{D_{l_1,...,l_r}}. \end{aligned} \tag{14}$$

Here the term for the empty partial permutation has been separated from the other terms, and, by convention, $\widehat{D_{l_1,...,l_r}} = 1$ whenever $\pi = \pi(\rho_1, \rho_2, q)$ with $\rho_1 = \rho_2 = \{1, ..., p\}$. Taking

expectations on both sides in Equation (14) we get

$$\begin{aligned}
&\mathbb{E}(\widehat{D_{p_1,\ldots,p_k}}) \\
&= \mathbb{E}(Y_{p_1,\ldots,p_k}) \\
&\quad - \sum_{p\geq 1} \sum_{\substack{\pi\in \mathrm{SP}_p \\ \pi=\pi(\rho_1,\rho_2,q)}} \frac{n^{|\sigma(\pi)|-k}}{N^{|\rho_1|}} P_\pi(n,N)\mathbb{E}(\widehat{D_{l_1,\ldots,l_r}}) \\
&= D_{p_1,\ldots,p_k} \\
&\quad + \sum_{p\geq 1} \sum_{\substack{\pi\in \mathrm{SP}_p \\ \pi=\pi(\rho_1,\rho_2,q)}} \frac{n^{|\sigma(\pi)|-k}}{N^{|\rho_1|}} P_\pi(n,N) D_{l_1,\ldots,l_r} \\
&\quad - \sum_{p\geq 1} \sum_{\substack{\pi\in \mathrm{SP}_p \\ \pi=\pi(\rho_1,\rho_2,q)}} \frac{n^{|\sigma(\pi)|-k}}{N^{|\rho_1|}} P_\pi(n,N) D_{l_1,\ldots,l_r} \\
&= D_{p_1,\ldots,p_k},
\end{aligned}$$

where we have again used Equation (13). This shows that $\widehat{D_{p_1,\ldots,p_k}}$ also is unbiased. We will now show that this recursive definition of $\widehat{D_{p_1,\ldots,p_k}}$ coincides with Equation (9), which will complete the proof of Lemma 1.

Recursively replacing the $\widehat{D_{l_1,\ldots,l_r}}$ in Equation (14) until there are only terms on the form $Y_{p_1,\ldots,p_k}$ left, we arrive at an expression on the form

$$\begin{aligned}
&\textstyle\sum_l \sum_{\pi_1,\ldots,\pi_l} (-1)^l \left(\prod_{i=1}^l \frac{n^{|\sigma(\pi_i)|-\sigma(\pi_{i-1})}}{N^{|\rho_{1i}|}} P_{\pi_i}(n,N)\right) Y_{l_1,\ldots,l_r} \\
&= \textstyle\sum_l \sum_{\pi_1,\ldots,\pi_l} (-1)^l \frac{n^{|\sigma(\pi_l)|-k}}{N^{\sum_{i=1}^l |\rho_{1i}|}} \left(\prod_{i=1}^l P_{\pi_i}(n,N)\right) Y_{l_1,\ldots,l_r},
\end{aligned} \tag{15}$$

where $\pi_1,\ldots,\pi_l$ are non-empty partial permutations, and where $l_1,\ldots,l_r$ are the cardinalities of the blocks after the identification of edges from all $\pi_1,\ldots,\pi_l$. We will call a $\pi=\pi_1,\ldots,\pi_l$ a *nested partial permutation*, since it corresponds to a nested application of partial permutations. The factor $(-1)^l$ comes from l applications of Equation (14), where each application contributes a factor (-1) therein. Due to this alternating sign, many terms in Equation (15) will cancel. The following class of permutations will be useful to see these cancellations:

Definition 3: Let $\Pi_{l,k}$ be the set of nested partial permutations on the form $\{\pi_1,\ldots,\pi_l\}$, where $|\rho_{\pi_1}|+\cdots+|\rho_{\pi_l}|=k$. Also, when $\pi=\{\pi_1,\ldots,\pi_l\}$ are nested partial permutations which do not contain any identifications involving edges i or j, let $\Pi_{\pi,i,j}\subset \Pi_{l,k+2}\bigcup\Pi_{l+1,k+2}$ be the set of nested partial permutations which equals π, with the exception that the identification (i,j) is added.

It is clear that any $\pi\in\Pi_{\pi,i,j}$ gives equal contribution in Equation (15) up to sign, since each such π embraces the same edges, and the order of the identification of edges does not matter for the final graph. It is also clear that

$$\begin{aligned}
|\Pi_{\pi,i,j}\cap\Pi_{l,k+2}| &= l, \\
|\Pi_{\pi,i,j}\cap\Pi_{l+1,k+2}| &= l+1,
\end{aligned}$$

and that the contributions from the two sets $\Pi_{\pi,i,j}\cap\Pi_{l,k+2}$ and $\Pi_{\pi,i,j}\cap\Pi_{l+1,k+2}$ have opposite signs, since the sign for any $\pi\in\Pi_{l,k}$ is $(-1)^l$. Adding the contributions, we get that the total contribution from $\Pi_{\pi,i,j}$ equals that from just one nested partial permutation in $\Pi_{l+1,k+2}$ where we set $\pi_{l+1}=(i,j)$. Summing over all π and l where $\pi=\{\pi_1,\ldots,\pi_l\}$ does not contain any identifications involving i or j, we get that the contribution from the set of π which contain (i,j) equals the sum over $\{\pi_1,\ldots,\pi_{l-1},\pi_l=(i,j)\}$. In the same way we can sum over π with (i,j) replaced by all other edge possibilities, to arrive at the sum over all $\pi=\{\pi_1,\ldots,\pi_l\}$, where all $|\rho_{\pi_i}|=1$, and where we need only to sum over sets (i.e. the order of the elements does not matter). In other words, and since there are $l=|\rho_1|$ partial permutations nested in this way, we can replace Equation (14) with

$$\widehat{D_p} = \sum_{\substack{\pi\in \mathrm{SP}_p \\ \pi=\pi(\rho_1,\rho_2,q)}} (-1)^{|\rho_1|} \frac{n^{|\sigma(\pi)|-1}}{N^{|\rho_1|}} P_\pi(n,N) Y_{l_1,\ldots,l_r}.$$

This coincides with Equation (9), and the proof of Lemma 1 is finished.

In a similar way to how Lemma 1 was proved, we can also obtain the following expression for the variance of $\widehat{D_p}$. We will only state it for the case of no stacking, and apply it for the stacked observation model in Appendix C:

Lemma 2: Let SPR_{2p} be the set of partial permutations of $\{1,\ldots,2p\}$ such that all identifications are from $\{1,\ldots,p\}$ to $\{p+1,\ldots,2p\}$, or vice versa. The variance

$$v_p = \mathbb{E}\left[\widehat{D_p}^2\right] - \mathbb{E}\left[\widehat{D_p}\right]^2$$

of $\widehat{D_p}$ equals

$$\sum_{\pi\in\mathrm{SPR}_{2p}} \frac{n^{|\sigma(\pi)|-2}}{N^{|\rho_1|}} P_\pi(n,N) D_{l_1,\ldots,l_r}. \tag{16}$$

Proof: Inserting the expression (9) twice we get

$$\begin{aligned}
v_p =& \mathbb{E}\left[\widehat{D_p}^2\right] - \mathbb{E}\left[\widehat{D_p}\right]^2 \\
=& \sum_{\substack{\pi_1\in\mathrm{SP}_p \\ \pi_1=\pi(\rho_1^{(1)},\rho_2^{(1)},q)}} \sum_{\substack{\pi_2\in\mathrm{SP}_p \\ \pi_2=\pi(\rho_1^{(2)},\rho_2^{(2)},q)}} \\
& (-1)^{|\rho_1^{(1)}|}(-1)^{|\rho_1^{(2)}|} \frac{n^{|\sigma(\pi_1)|-1}}{N^{|\rho_1^{(1)}|}} \frac{n^{|\sigma(\pi_2)|-1}}{N^{|\rho_1^{(2)}|}} \\
& \times P_{\pi_1}(n,N) P_{\pi_2}(n,N) \\
& \times \left(\mathbb{E}\left[Y_{l_1^{(1)},\ldots,l_{r_1}^{(1)} Y_{l_1^{(2)},\ldots,l_{r_2}^{(2)}}}\right] \right. \\
& \left. \quad - \mathbb{E}\left[Y_{l_1^{(1)},\ldots,l_{r_1}^{(1)}}\right] \mathbb{E}\left[Y_{l_1^{(2)},\ldots,l_{r_2}^{(2)}}\right]\right),
\end{aligned} \tag{17}$$

where $l_1^{(1)},\ldots,l_{r_1}^{(1)}$ are the cardinalities of the blocks of $\sigma(\pi_1)$ divided by 2, $l_1^{(2)},\ldots,l_{r_2}^{(2)}$ those of $\sigma(\pi_2)$ divided by 2. Using Equation (13) we can write

$$\begin{aligned}
&\mathbb{E}\left[Y_{l_1^{(1)},\ldots,l_{r_1}^{(1)} Y_{l_1^{(2)},\ldots,l_{r_2}^{(2)}}}\right] \\
&= \sum_{\substack{\pi\in\mathrm{SP}_{2p-|\rho_1^{(1)}|-|\rho_1^{(2)}|} \\ \pi=\pi(\rho_1,\rho_2,q)}} \frac{n^{|\sigma(\pi)|-r_1-r_2}}{N^{|\rho_1|}} P_\pi(n,N) D_{l_1,\ldots,l_r},
\end{aligned} \tag{18}$$

where $l_1, \ldots, l_r$ are the cardinalities of $\sigma(\pi)$ divided by 2, and

$$\begin{aligned}\mathbb{E}\left[Y_{l_1^{(1)},\ldots,l_{r_1}^{(1)}}\right]\mathbb{E}\left[Y_{l_1^{(2)},\ldots,l_{r_2}^{(2)}}\right] &= \sum_{\substack{\pi^{(1)}\in \mathrm{SP}_{p-|\rho_1^{(1)}|} \\ \pi^{(1)}=\pi(\rho_{11},\rho_{12},q)}} \sum_{\substack{\pi^{(2)}\in \mathrm{SP}_{p-|\rho_1^{(2)}|} \\ \pi^{(2)}=\pi(\rho_{21},\rho_{22},q)}} \frac{n^{|\sigma(\pi^{(1)})|-r_1}}{N^{|\rho_{11}|}}\frac{n^{|\sigma(\pi^{(2)})|-r_2}}{N^{|\rho_{21}|}} \\ &\quad \times P_{\pi^{(1)}}(n,N)P_{\pi^{(2)}}(n,N) \\ &\quad \times D_{l_1^{(1)},\ldots,l_{r_1}^{(1)}}D_{l_1^{(2)},\ldots,l_{r_2}^{(2)}}, \end{aligned} \tag{19}$$

where $l_1^{(1)}, \ldots, l_{r_1}^{(1)}$ are the cardinalities of $\sigma(\pi^{(1)})$ divided by 2, $l_1^{(2)}, \ldots, l_{r_2}^{(2)}$ those of $\sigma(\pi^{(2)})$ divided by 2. The powers of n and N in Equation (19) can be written as

$$\frac{n^{|\sigma(\pi^{(1)})|+|\sigma(\pi^{(2)})|-r_1-r_2}}{N^{|\rho_{11}|+|\rho_{21}|}} P_{\pi^{(1)}}(n,N)P_{\pi^{(2)}}(n,N),$$

which match the powers of n and N in Equation (18) when $\pi = \pi_1 \times \pi_2$ does not contain any identification of edges from different expectations. These terms thus cancel, and we are left with summing over π containing identification of edges between the two expectations.

To see that we need only to sum over π containing only identification of edges from one expectation to another, note that a π_1 containing (i,j) cancels the contribution from a π containing (i,j), since the former has an additional power of (-1). The same can be said for π_2. The only terms not canceling therefore occur when π_1 and π_2 are empty, and π only contains identifications between the two expectations. These correspond to SPR_{2p} by definition. All of them contribute with a positive sign, and all in all we get that v_p equals

$$\sum_{\pi\in\mathrm{SPR}_{2p}} \frac{n^{|\sigma(\pi)|-2}}{N^{|\rho_1|}} P_\pi(n,N) D_{l_1,\ldots,l_r}$$

(since $r_1 = r_2 = 1$), which is what we had to show. ■

APPENDIX C
THE PROOF OF THEOREM 1

The geometric interpretation of $\pi \in \mathrm{SPR}_{2p}$ is as an identification among some of $4p$ edges, where even edges are only identified with odd edges and vice versa, and where there are only identifications between $\{1, ..., 2p\}$ and $\{2p+1, ..., 4p\}$, and vice versa. It is clear that $\pi \in \mathrm{SPR}_{2p}$ is invariant under cyclic shifts of the form $\pi \to s_{1k}s_{2l}\pi(s_{1k}s_{2l})^{-1}$, where

$$\begin{aligned} s_{1k}(r) &= \begin{cases} r+k \text{ for } r \in \{1,...,2p\} \\ r \text{ for } r \in \{2p+1,...,4p\} \end{cases} \\ s_{2l}(r) &= \begin{cases} r \text{ for } r \in \{1,...,2p\} \\ r+l \text{ for } r \in \{2p+1,...,4p\} \end{cases} \end{aligned}$$

(addition performed so that result stays within the same interval, either $[1, ..., 2p]$ or $[2p+1, ..., 4p]$) as long as k and l either are both odd, or both even, in order for the identification to remain between even and odd elements and vice versa. The equivalence class of $\pi \in \mathrm{SPR}_{2p}$ under cyclic shifts is given by $\bigcup_{k,l} s_{1k}s_{2l}\pi(s_{1k}s_{2l})^{-1}$, where k and l either are both odd, or both even. We will denote by SPE_{2p} the set of such equivalence classes, and denote by $\bar{\pi} \in \mathrm{SPE}_{2p}$ the equivalence class of $\pi \in \mathrm{SPR}_{2p}$.

From the geometric interpretation of π it is clear that, when we instead of π use $s_{1k}s_{2l}\pi(s_{1k}s_{2l})^{-1}$,

1) $|\rho_1|$ and $|\rho_2|$ is the same for π and $s_{1k}s_{2l}\pi(s_{1k}s_{2l})^{-1}$,
2) $|\sigma(\pi)| = |\sigma(s_{1k}s_{2l}\pi(s_{1k}s_{2l})^{-1})|$. The block cardinalities $l_1, ..., l_r$ of $\sigma(\pi)$ and $\sigma(s_{1k}s_{2l}\pi(s_{1k}s_{2l})^{-1})$ are also equal,
3) when k and l are both even, k, kd, l, ld are the same for $\rho(\pi)$ and $\rho(s_{1k}s_{2l}\pi(s_{1k}s_{2l})^{-1})$,
4) when k and l are both odd,

$$\begin{aligned} k(\rho(\pi) &= l(\rho(s_{1k}s_{2l}\pi(s_{1k}s_{2l})^{-1})) \\ kd(\rho(\pi) &= ld(\rho(s_{1k}s_{2l}\pi(s_{1k}s_{2l})^{-1})) \\ l(\rho(\pi) &= k(\rho(s_{1k}s_{2l}\pi(s_{1k}s_{2l})^{-1})) \\ ld(\rho(\pi) &= kd(\rho(s_{1k}s_{2l}\pi(s_{1k}s_{2l})^{-1})). \end{aligned}$$

By definition of P_π, the last two statements say that

$$P_{s_{1k}s_{2l}\pi(s_{1k}s_{2l})^{-1}}(n,N) = P_\pi(n,N)$$

when k, l are both even, and

$$P_{s_{1k}s_{2l}\pi(s_{1k}s_{2l})^{-1}}(n,N) = P_\pi(N,n)$$

when k, l are both odd. Since there are equally many elements with k, l odd and k, l even under cyclic equivalence, we see that

$$Q_{\bar{\pi}}(n,N) = \sum_{\pi_1\sim\pi} P_{\pi_1}(n,N) \tag{20}$$

is a polynomial symmetric in n and N, where $\sim$ denotes equivalence under cyclic shifts. The first statements above say that the rest of the powers of n and N in Equation (16) are unchanged under cyclic equivalence. By summing over the cyclic equivalence classes in Equation (16), we see that it can be rewritten as

$$v_p = \sum_{\bar{\pi}\in\mathrm{SPE}_{2p}} \frac{n^{|\sigma(\pi)|-2}}{N^{|\rho_1|}} Q_{\bar{\pi}}(n,N) D_{l_1,\ldots,l_r}, \tag{21}$$

with $Q_{\bar{\pi}}$ symmetric in n and N. Moreover, $Q_{\bar{\pi}}$ has the form $Q_{\bar{\pi}}(n,N) = an^kN^l + bn^lN^k$, where $a+b$ is the number of elements in the cyclic equivalence class of π.

Since $\widehat{D_{p,L_1,L_2}}$ is L_1^{1-p} times the estimator for the p-th moment F_p of the compound matrix by the comments following the statement of Lemma 1, the variance $v_{p,\cdot,L}$ of $\widehat{D_{p,L_1,L_2}}$ in Equation (11) is, after replacing n with nL_1, and N with NL_2 in Equation (21),

$$\begin{aligned} v_{p,\cdot,L} &= L_1^{2-2p} \sum_{\pi\in\mathrm{SPE}_{2p}} \frac{n^{|\sigma(\pi)|-2}L_1^{|\sigma(\pi)|-2}}{N^{|\rho_1|}L_2^{|\rho_1|}} Q_{\bar{\pi}}(nL_1,NL_2) \\ &\quad \times F_{l_1,\ldots,l_r} \\ &= L_1^{2-2p} \sum_{\pi\in\mathrm{SPE}_{2p}} \frac{n^{|\sigma(\pi)|-2}L_1^{|\sigma(\pi)|-2}}{N^{|\rho_1|}L_2^{|\rho_1|}} Q_{\bar{\pi}}(nL_1,NL_2) \\ &\quad \times L_1^{2p-|\rho_1|-|\sigma(\pi)|} D_{l_1,\ldots,l_r} \\ &= \sum_{\pi\in\mathrm{SPE}_{2p}} \frac{n^{|\sigma(\pi)|-2}}{N^{|\rho_1|}L^{|\rho_1|}} Q_{\bar{\pi}}(nL_1,NL_2) D_{l_1,\ldots,l_r}, \end{aligned} \tag{22}$$

where we have used Equation (3), and set $L = L_1 L_2$.

$\deg(Q_{\bar{\pi}})$ describes the number of vertices in the graph of random edges not bordering to deterministic edges. Each vertex is associated with a value $\leq L\max(n, N)$, so that $Q_{\bar{\pi}}$ has order at most L to the power of the number of vertices not bordering to deterministic edges. We will use this in the following, and consider the following possibilities:

1) There are no deterministic edges: in this case, $p = |\rho_1|/2$. Since there are only crossidentifications between $\{1, ..., 2p\}$ and $\{2p+1, ..., 4p\}$ for $\pi \in \mathrm{SPR}_{2p}$, any vertex in $\{2p+1, ..., 4p\}$ is identified with a vertex in $\{1, ..., 2p\}$, so that $\{1, ..., 2p\}$ contains representatives for all equivalence classes of vertices. There are thus at most p even equivalence classes, and at most p odd equivalence classes. Thus

$$\begin{aligned} Q_{\bar{\pi}}(nL_1, NL_2) &\leq O\left((nL_1)^p (NL_2)^p\right) = O(L^p) \\ &= O(L^{|\rho_1|/2}). \end{aligned}$$

When $p = 1$, $|\rho_1| = 2$, and $|\rho_1|/2 = |\rho_1| - 1$, so that $Q_{\bar{\pi}}(nL_1, NL_2) \leq O\left(L^{|\rho_1|-1}\right)$, and it is easy to check that we have equality for the only partial permutation in SPR_2, and that $Q_{\bar{\pi}}(nL_1, NL_2) = nNL^{|\rho_1|-1}$ for this $\bar{\pi}$. When $p > 1$, $|\rho_1|/2 < |\rho_1| - 1$, so that $Q_{\bar{\pi}}(nL_1, NL_2) = O(L^{|\rho_1|-2})$ for such $\bar{\pi}$.

2) The graph of random edges is a tree, and there exist deterministic edges: since any crossidentification between $\{1, ..., 2p\}$ and $\{2p+1, ..., 4p\}$ does not give rise to a leaf node when all edges are considered, any leaf in the tree of random edges must be bordering to a deterministic edge. Since the tree contains $|\rho_1|+1$ vertices, and since there are at least two leafs in any tree, we have that $Q_{\bar{\pi}}(nL_1, NL_2)$ has order at most $O\left(L^{|\rho_1|-1}\right)$, with equality only if the graph of random edges borders to exactly two deterministic edges. It is easily seen that this occurs if and only if $|\rho_1|$ pairs of edges are identified in successive order.

3) The graph of random edges is not a tree, and there exist deterministic edges: if there are two cycles in the graph of random edges, $Q_{\bar{\pi}}(nL_1, NL_2)$ has order at most $O\left(L^{|\rho_1|-2}\right)$ (two subtracted for the cycles, one for the deterministic edge). Similarly, if there is one cycle, and more than one vertex bordering to a deterministic edge, $Q_{\bar{\pi}}(nL_1, NL_2)$ has order at most $O\left(L^{|\rho_1|-2}\right)$. Assume thus that there is only one vertex bordering to a deterministic edge, and only one cycle. It is easily checked that this vertex must be on the cycle, and that we must end up in the same situation as in 2) where edges are identified in successive order, for which we actually have a tree. Thus, there is nothing more to consider.

We see that $Q_{\bar{\pi}}(nL_1, NL_2)$ has order at most $O\left(L^{|\rho_1|-1}\right)$ in any case. Inserting into Equation (22), the first case above contributes with $L^{-1}\frac{1}{nN}$ for $p = 1$, for $p > 1$ we get only terms of order $O(L^{-2})$. The third case contributes only with terms of order $O(L^{-2})$. For the second case, contributions are of order $O(L^{-2})$ when $|\rho_1|$ pairs of edges are not identified in successive order. When they are identified in successive order, we consider the following different possibilities:

- When $|\rho_1|$ is odd we will have $k(\rho(\pi)) - kd(\rho(\pi)) = l(\rho(\pi)) - ld(\rho(\pi)) = \frac{|\rho_1|-1}{2}$, so that

$$\begin{aligned} &Q_{\bar{\pi}}(nL_1, NL_2) \\ &= (L_2 N)^{k(\rho(\pi))-kd(\rho(\pi))} (L_1 n)^{l(\rho(\pi))-ld(\rho(\pi))} \\ &= (L_2 N)^{\frac{|\rho_1|-1}{2}} (L_1 n)^{\frac{|\rho_1|-1}{2}} \\ &= (nN)^{\frac{|\rho_1|-1}{2}} L^{\frac{|\rho_1|-1}{2}}, \end{aligned}$$

  so that the term for $\bar{\pi}$ in Equation (22) is of order $L^{(|\rho_1|-1)/2-|\rho_1|} = L^{-|\rho_1|/2-1/2}$. When $|\rho_1| = 1$, this is $O\left(L^{-1}\right)$, and the contribution in this case is $\frac{1}{nNL}$ times the number of partitions in the equivalence class of $\bar{\pi}$ When $|\rho_1| > 1$, all terms are of order $O(L^{-2})$.
- When $|\rho_1|$ is even, either
  1) $k(\rho(\pi)) - kd(\rho(\pi)) = \frac{|\rho_1|}{2} - 1$, $l(\rho(\pi)) - ld(\rho(\pi)) = \frac{|\rho_1|}{2}$, for which

$$\begin{aligned} &Q_{\bar{\pi}}(nL_1, NL_2) \\ &= (L_2 N)^{\frac{|\rho_1|}{2}-1} (L_1 n)^{\frac{|\rho_1|}{2}} \\ &= N^{\frac{|\rho_1|}{2}-1} n^{\frac{|\rho_1|}{2}} L^{\frac{|\rho_1|}{2}-1} L_1, \end{aligned}$$

     so that the term for $\bar{\pi}$ in Equation (22) is of order

$$L^{|\rho_1|/2-1-|\rho_1|} L_1 = L^{-|\rho_1|/2-1} L_1.$$

     When the stacking is not vertical, we have that $L_1 \leq O(L^{1/2})$, so that the term for $\bar{\pi}$ is of order $\leq O(L^{-|\rho_1|/2-1} L^{1/2}) = O(L^{-|\rho_1|/2-1/2}) \leq O(L^{-3/2})$ When the stacking is vertical, the term is of order $L^{-|\rho_1|/2}$, which is $O(L^{-2})$ when $|\rho_1| > 2$. When $|\rho_1| = 2$, the contribution in Equation (22) is seen to be $\frac{1}{N^2 L}$ times the number of partitions in the equivalence class of $\bar{\pi}$.
  2) $k(\rho(\pi)) - kd(\rho(\pi)) = \frac{|\rho_1|}{2}$, $l(\rho(\pi)) - ld(\rho(\pi)) = \frac{|\rho_1|}{2} - 1$, for which the term for $\bar{\pi}$ in Equation (22) similarly is shown to be of order

$$L^{|\rho_1|/2-1-|\rho_1|} L_2 = L^{-|\rho_1|/2-1} L_2,$$

     and, similarly, only horizontal stacking with $|\rho_1| = 2$ gives contributions of order $O(L^{-1})$. The contribution in Equation (22) is seen to be $\frac{1}{nNL}$ times the number of partitions in the equivalence class of $\bar{\pi}$.

When it comes to the number of elements in the corresponding equivalence classes, it is easy to see that

- there are $2p^2$ elements for the class where $|\rho_1| = 1$, corresponding to any choice of the $2p$ edges $\{1, ..., 2p\}$, and any choice of the p even or odd edges in $\{2p+1, ..., 4p\}$.
- p^2 elements for each class where $|\rho_1| = 2$.

Summing up, we see that for $p = 1$, $v_{1,\cdot,L} = L^{-1}\frac{2}{nN} D_1 + L^{-1}\frac{1}{nN}$ for any type of stacking/averaging. For $p \geq 2$ we get that

$$\begin{aligned} v_{p,R,L} &= L^{-1}\frac{2p^2}{nN} D_{2p-1} + O\left(L^{-3/2}\right) \\ v_{p,V,L} &= L^{-1}\frac{2p^2}{nN} D_{2p-1} + L^{-1}\frac{p^2}{N^2} D_{2p-2} + O\left(L^{-3/2}\right) \\ v_{p,H,L} &= L^{-1}\frac{2p^2}{nN} D_{2p-1} + L^{-1}\frac{p^2}{nN} D_{2p-2} + O\left(L^{-3/2}\right), \end{aligned}$$

and the first formulas in Theorem 1 follows after multiplying both sides with L, and taking limits. The $O\left(L^{-3/2}\right)$-terms make up the polynomial Q in Theorem 1, which has only positive coefficients due to Expression (16). The case of averaging follows by noting that there are only positive coefficients in Equation (22) for the variance, and that the variance is divided by L when one takes L independent observations.

Finally, we prove why the least variance is obtained when the compound observation matrix is as square as possible. With $c_1 = nL_1, c_2 = NL_2, c = \frac{c_1}{c_2}$, we can write each $Q_{\bar{\pi}}(nL_1, NL_2)$ as a scalar multiple of

$$\begin{aligned} & c_1^k c_2^l + c_1^l c_2^k \\ &= (nN)^{\frac{k+l}{2}} L^{\frac{k+l}{2}} \left(c_1^{\frac{k-l}{2}} c_2^{\frac{l-k}{2}} + c_1^{\frac{l-k}{2}} c_2^{\frac{k-l}{2}} \right) \\ &= (nN)^{\frac{k+l}{2}} L^{\frac{k+l}{2}} \left(c^{\frac{k-l}{2}} + c^{\frac{l-k}{2}} \right), \end{aligned}$$

It is clear that $f(c) = c^{(k-l)/2} + c^{(l-k)/2}$ has a global minimum at $c = 1$ on $(0, \infty)$, and the result follows.

REFERENCES

[1] E. Telatar, "Capacity of multi-antenna gaussian channels," *Eur. Trans. Telecomm. ETT*, vol. 10, no. 6, pp. 585–596, Nov. 1999.

[2] J.-P. Bouchaud and M. Potters, *Theory of Financial Risk and Derivative Pricing - From Statistical Physics to Risk Management*. Cambridge: Cambridge University Press, 2000.

[3] T. Guhr, A. Müller-Groeling, and H. A. Weidenmüller, "Random matrix theories in quantum physics: Common concepts," *Phys.Rept. 299*, pp. 189–425, 1998.

[4] D. V. Voiculescu, "Addition of certain non-commuting random variables," *J. Funct. Anal.*, vol. 66, pp. 323–335, 1986.

[5] ——, "Multiplication of certain noncommuting random variables," *J. Operator Theory*, vol. 18, no. 2, pp. 223–235, 1987.

[6] D. Voiculescu, "Circular and semicircular systems and free product factors," *Operator algebras, unitary representations, enveloping algebras and invariant theory*, vol. 92, 1990.

[7] ——, "Limit laws for random matrices and free products," *Inv. Math.*, vol. 104, pp. 201–220, 1991.

[8] F. Hiai and D. Petz, *The Semicircle Law, Free Random Variables and Entropy*. American Mathematical Society, 2000.

[9] Ø. Ryan and M. Debbah, "Asymptotic behaviour of random Vandermonde matrices with entries on the unit circle," *IEEE Trans. on Information Theory*, vol. 55, no. 7, pp. 3115–3148, 2009.

[10] ——, "Convolution operations arising from Vandermonde matrices," *Submitted to IEEE Trans. on Information Theory*, 2009.

[11] J. W. Silverstein and P. L. Combettes, "Signal detection via spectral theory of large dimensional random matrices," *IEEE Trans. Signal Process.*, vol. 40, no. 8, pp. 2100–2105, 1992.

[12] B. Dozier and J. W. Silverstein, "On the empirical distribution of eigenvalues of large dimensional information-plus-noise type matrices," *J. Multivariate Anal.*, vol. 98, no. 4, pp. 678–694, 2007.

[13] N. E. Karoui, "Spectrum estimation for large dimensional covariance matrices using random matrix theory," *Ann. Statist.*, vol. 36, no. 6, pp. 2757–2790, 2008.

[14] X. Mestre, "Improved estimation of eigenvalues of covariance matrices and their associated subspaces using their sample estimates," *IEEE Trans. Signal Process.*, vol. 56, no. 11, pp. 5353–5368, 2008.

[15] R. Couillet, J. W. Silverstein, Z. Bai, and M. Debbah, "Eigen-inference for energy estimation of multiple sources," *submitted to IEEE Trans. on Information Theory*, pp. 1–17, 2010, http://arxiv.org/abs/1001.3934.

[16] P. Vallet, P. Loubaton, and X. Mestre, "Improved subspace estimation for multivariate observations of high dimension: the deterministic signals case," *Submitted*, pp. 1–68, 2010, http://arxiv.org/abs/1002.3234.

[17] N. R. Rao and A. Edelman, "The polynomial method for random matrices," *Foundations of Computational Mathematics*, vol. 8, no. 6, pp. 649–702, December 2008.

[18] "Sample eigenvalue based detection of high-dimensional signals in white noise using relatively few samples," *Submitted to IEEE Trans. Signal Process.*, vol. 7, pp. 1–21, 2008, http://arxiv.org/abs/0705.2605.

[19] N. R. Rao, J. Mingo, R. Speicher, and A. Edelman, "Statistical eigen-inference from large Wishart matrices," *Ann. Statist.*, vol. 36, no. 6, pp. 2850–2885, 2008.

[20] F. Benaych-Georges and M. Debbah, "Free deconvolution: from theory to practice," *submitted to IEEE Transactions on Information Theory*, 2008.

[21] Ø. Ryan, A. Masucci, S. Yang, and M. Debbah, "Finite dimensional statistical inference," *Submitted to IEEE Trans. on Information Theory*, 2009, http://arxiv.org/abs/0911.5515.

[22] Ø. Ryan and M. Debbah, "Channel capacity estimation using free probability theory," *IEEE Trans. Signal Process.*, vol. 56, no. 11, pp. 5654–5667, November 2008.

[23] J. A. Mingo and R. Speicher, "Second order freeness and fluctuations of random matrices: I. Gaussian and Wishart matrices and cyclic Fock spaces," *J. Funct. Anal.*, vol. 235, no. 1, pp. 226–270, 2006.

[24] J. A. Mingo, P. Śniady, and R. Speicher, "Second order freeness and fluctuations of random matrices: II. unitary random matrices," *Adv. in Math.*, vol. 209, pp. 212–240, 2007.

[25] B. Collins, J. A. Mingo, P. Śniady, and R. Speicher, "Second order freeness and fluctuations of random matrices: III. higher order freeness and free cumulants," *Documenta Math.*, vol. 12, pp. 1–70, 2007.

[26] Ø. Ryan, *Tools for the optimal stcaking of noisy observations*, 2010, http://folk.uio.no/oyvindry/findimstacking/.

Øyvind Ryan was born in Oslo, Norway. He studied mathematics at the University of Oslo, where he received the M.Sc and the Ph.D. degrees in 1993 and 1997, respectively.

From 1997 to 2004, he worked as a consultant and product developer in various information technology projects. From 2004 to 2007, he was a postdoctoral fellow at the Institute of Informatics at the University of Oslo. He is currently employed as a researcher at the Centre of Mathematics for Applications at the University of Oslo, and at Supelec, Gif-sur-Yvette, France. His research interests are applications of free probability theory and random matrices to the fields of wireless communication, finance, and information theory.