On the detection of low rank matrices in the high-dimensional regime

Antoine Chevreuil, Philippe Loubaton

On the detection of low rank matrices in the high-dimensional regime.

Antoine Chevreuil and Philippe Loubaton
Gaspard Monge Computer Science Laboratory (LIGM) - UMR 8049 CNRS
Université de Paris-Est/Marne-la-Vallée
5 Bd. Descartes 77454 Marne-la-Vallée (France)

***Abstract*—We address the detection of a low rank $n \times n$ matrix $\mathbf{X}_0$ from the noisy observation $\mathbf{X}_0+\mathbf{Z}$ when $n \to \infty$, where $\mathbf{Z}$ is a complex Gaussian random matrix with independent identically distributed $\mathcal{N}_c(0, \frac{1}{n})$ entries. Thanks to large random matrix theory results, it is now well-known that if the largest singular value $\lambda_1(\mathbf{X}_0)$ of $\mathbf{X}_0$ verifies $\lambda_1(\mathbf{X}_0) > 1$, then it is possible to exhibit consistent tests. In this contribution, we prove *a contrario* that under the condition $\lambda_1(\mathbf{X}_0) < 1$, there are no consistent tests. Our proof is inspired by previous works devoted to the case of rank 1 matrices $\mathbf{X}_0$.**

***Index Terms*—statistical detection tests, large random matrices, large deviation principle.**

I. Introduction

The problem of testing whether an observed $n_1 \times n_2$ matrix $\mathbf{Y}$ is either a zero-mean independent identically distributed Gaussian random matrix $\mathbf{Z}$ with variance $\frac{1}{n_2}$, or $\mathbf{X}_0 + \mathbf{Z}$ for some low rank deterministic matrix $\mathbf{X}_0$, called also a *spike*, is a fundamental problem arising in numerous applications such as the detection of low-rank multivariate signals or the Gaussian hidden clique problem. When the two dimensions n_1, n_2 converge towards ∞ in such a way that $n_1/n_2 \to c > 0$ (the rank of $\mathbf{X}_0$ remaining fixed), known results on the so-called additive spiked large random matrix models have enabled to re-consider this fundamental detection problem (see *e.g.* [11], [4], [3]). It was established a long time ago (see *e.g.* [1] and the references therein) that in the above asymptotic regime, the largest singular value $\lambda_1(\mathbf{Z})$ of $\mathbf{Z}$ converges almost surely towards $1 + \sqrt{c}$. More recently, under mild technical extra assumptions, [3] proved that $\lambda_1(\mathbf{X}_0 + \mathbf{Z})$ still converges towards $1 + \sqrt{c}$ if $\lambda_1(\mathbf{X}_0)$ converges towards a limit strictly less than $c^{1/4}$. On the contrary, if the limit of $\lambda_1(\mathbf{X}_0)$ is strictly greater than $c^{1/4}$, then $\lambda_1(\mathbf{X}_0 + \mathbf{Z})$ converges towards a limit strictly greater than $1 + \sqrt{c}$. This result implies that the Generalized Likelihood Ratio Test (GLRT) is consistent (*i.e.* both the probability of false alarm and the probability of missed detection converge towards 0 in the above asymptotic regime) if and only if $\lambda_1(\mathbf{X}_0)$ is above the threshold $c^{1/4}$. In order to simplify the exposition, we assume from now on that $n_1 = n_2 = n$, so that ratio c reduces to 1.

While the detection problem was extensively addressed in the zone $\lambda_1(\mathbf{X}_0) > 1$, the case where $\lambda_1(\mathbf{X}_0) < 1$ was much less studied. Montanari *et al.* [10] consider the zone $\lambda_1(\mathbf{X}_0) < 1$ when $\mathbf{X}_0$ is a rank 1 matrix. Thanks to information geometry tools, [10] prove that, in this region, it is impossible to find a consistent test for the detection of the spike. Irrespective of the standard random matrix tools, this approach is extended in [10] to the more general case when $\mathbf{X}_0$ and $\mathbf{Z}$ are tensors of order $d \geq 3$; namely, if the Frobenius norm of the tensor $\mathbf{X}_0$ is stricly less than a threshold depending in d, then the probability distributions of the observation under the two hypotheses are asymptotically undistinguishable, so that any detection test cannot behave better than a random guess. This property, which is stronger than the non-existence of a consistent test, does not hold in the matrix case $d = 2$: see for instance [12] where a non-consistent test is exhibited that has a better performance than a random guess.

In this paper, we focus on the case where $\mathbf{X}_0$ has general rank r. Our contribution is to prove that under $\lambda_1(\mathbf{X}_0) < 1$, the consistent detection is impossible. While this theoretical result is not unexpected, we believe that it provides a better understanding of the above fundamental detection problem in large dimensions without resorting to the machinery of large random matrices.

II. Model, notation, asumption

The set of complex-valued matrices $\mathbb{C}^{n \times n}$ is a complex vector-space endowed with the standard scalar product $\langle \mathbf{X}, \mathbf{Y} \rangle = \mathrm{Tr}(\mathbf{X}\mathbf{Y}^*)$ and the Frobenius norm $\|\mathbf{X}\|_F = \sqrt{\langle \mathbf{X}, \mathbf{X} \rangle}$. The spectral norm of a matrix $\mathbf{X}$ is denoted by $\|\mathbf{X}\|_2$. The spike ("the signal") is assumed to be a matrix of fixed rank r and hence admits a SVD such as

$$\mathbf{X}_0 = \sum_{j=1}^{r} \lambda_j \mathbf{u}_j \mathbf{v}_j^* = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^* \tag{1}$$

where $\lambda_i = \lambda_i(\mathbf{X}_0)$ are the singular values of $\mathbf{X}_0$ sorted in descending order and where $\mathbf{\Lambda}$ is the diagonal matrix gathering the $(\lambda_j)_{j=1,\ldots,r}$ in the descending order. As $\mathbf{X}_0$ has to be defined for any n, we impose a non-erratic behavior of $\mathbf{X}_0$, namely that all its singular values $(\lambda_j)_{j=1,\ldots,r}$ do not depend on n for n large enough. This hypothesis could be replaced by the condition that $(\lambda_j)_{j=1,\ldots,r}$ all converge towards a finite limit at an *ad'hoc* rate. However, this would introduce purely technical difficulties.

The noise matrix $\mathbf{Z}$ is assumed to have i.i.d. entries distributed as $\mathcal{N}_c(0, 1/n)$. We consider the alternative $\mathcal{H}_0 : \mathbf{Y} = \mathbf{Z}$ versus $\mathcal{H}_1 : \mathbf{Y} = \mathbf{X}_0 + \mathbf{Z}$. We denote by $p_{1,n}(\mathbf{y})$ the probability probability density of $\mathbf{Y}$ under $\mathcal{H}_1$ and $p_{0,n}(\mathbf{y})$ the density of $\mathbf{Y}$ under $\mathcal{H}_0$. $\mathcal{L}(\mathbf{Y}) = \frac{p_{1,n}(\mathbf{Y})}{p_{0,n}(\mathbf{Y})}$ is the likelihood

ratio and we denote by $\mathbb{E}_0$ the expectation under $\mathcal{H}_0$. We now recall the fundamental information geometry results used in [10] in order to address the detection problem.The following properties are well known (see also [2] section 3):

- (i) If $\mathbb{E}_0\left[\mathcal{L}(\mathbf{Y})^2\right]$ is bounded, then no consistent detection test exists.
- (ii) If moroever $\mathbb{E}_0\left[\mathcal{L}(\mathbf{Y})^2\right] = 1+o(1)$, the total variation distance between $p_{0,n}$ and $p_{1,n}$ converges towards 0, and no test performs better than a decision at random.

We however mention that (i) and (ii) are only sufficient conditions. In particular, $\mathbb{E}_0\left[\mathcal{L}(\mathbf{Y})^2\right]$ unbounded does not imply the existence of consistent tests.

III. Prior on the spike. Expression of the second-order moment.

The density of $\mathbf{Z}$, seen as a collection of n^2 complex-valued random variables, is obviously $p_{0,n}(\mathbf{z}) = \kappa_n \exp\left(-n\left\|\mathbf{z}\right\|_F^2\right)$ where $\kappa_n = \left(\frac{n}{\pi}\right)^{n^2}$. On the one hand, we notice that the study of the second-order moment of the likelihood ratio is not suited to the deterministic model of the spike as presented previously. Indeed, in this case $\mathbb{E}_0\left[\mathcal{L}(\mathbf{Y})^2\right]$has the simple expression $\exp\left(2n\left\|\mathbf{X}_0\right\|_F^2\right)$ and always diverges. On the other hand, the noise matrix shows an invariance property: if $\mathbf{\Theta}_1, \mathbf{\Theta}_2$ are unitary $n\times n$ matrices , then the density of $\mathbf{\Theta}_1\mathbf{Z}\mathbf{\Theta}_2$ equals this of $\mathbf{Z}$. We hence modify the data according to the procedure: we pick two independent unitary $\mathbf{\Theta}_1, \mathbf{\Theta}_2$ according to the Haar measure (which corresponds to the uniform distribution on the set of all unitary $n \times n$ matrices), and change the data tensor $\mathbf{Y}$ according to $\mathbf{\Theta}_1\mathbf{Y}\mathbf{\Theta}_2$. As said above, this does not affect the distribution of the noise, but this amounts to assume a certain prior on the spike. Indeed, this amounts to replace $\mathbf{u}_i$ by $\mathbf{\Theta}_1\mathbf{u}_i$ and $\mathbf{v}_i$ by $\mathbf{\Theta}_2^*\mathbf{v}_i$. In the following, the data and the noise tensors after this procedure are still denoted respectively by $\mathbf{Y}$ and $\mathbf{Z}$.

We are now in position to give a closed-form expression of the second-order moment of $\mathcal{L}(\mathbf{Y})$. We have $p_{1,n}(\mathbf{Y}) = \mathbb{E}_X\left[p_{0,n}(\mathbf{Y}-\mathbf{X})\right]$ where $\mathbb{E}_X$ is the mathematical expectation over the prior distribution of the spike, or equivalently over the Haar matrices $\mathbf{\Theta}_1, \mathbf{\Theta}_2$. It holds that $\mathbb{E}_0\left[\mathcal{L}(\mathbf{Y})^2\right] = \mathbb{E}\left[\exp\left(2n\Re\left\langle\mathbf{X},\mathbf{X}'\right\rangle\right)\right]$ where the expectation is over independent copies $\mathbf{X}, \mathbf{X}'$ of the spike ($\Re$ stands for the real part); $\mathbf{X}$ and $\mathbf{X}'$ being respectively associated with $(\mathbf{\Theta}_1, \mathbf{\Theta}_2)$ and $(\mathbf{\Theta}_1', \mathbf{\Theta}_2')$, $\mathbb{E}_0\left[\mathcal{L}(\mathbf{Y})^2\right]$ has the expression

$$\mathbb{E}\left[\exp\left(2n\Re\mathrm{Tr}\left(\mathbf{\Theta}_1\mathbf{X}_0\mathbf{\Theta}_2\left(\mathbf{\Theta}_2'\right)^*\mathbf{X}_0^*\left(\mathbf{\Theta}_1'\right)^*\right)\right)\right].$$

As $\mathbf{\Theta}_k$ and $\mathbf{\Theta}_k'$ are Haar and independent, then $\left(\mathbf{\Theta}_1'\right)^*\mathbf{\Theta}_1$ and $\mathbf{\Theta}_2\left(\mathbf{\Theta}_2'\right)^*$ are also independent, Haar distributed and it holds

$$\mathbb{E}_0\left[\mathcal{L}(\mathbf{Y})^2\right] = \mathbb{E}\left[\exp\left(2n\eta\right)\right], \tag{2}$$

where the expectation is over the independent Haar matrices $\mathbf{\Theta}_1, \mathbf{\Theta}_2$ and $\eta = \Re\mathrm{Tr}\left(\mathbf{\Theta}_1\mathbf{X}_0\mathbf{\Theta}_2\mathbf{X}_0^*\right)$. The ultimate simplification comes from the decomposition (1) which implies that

$$\eta = \Re\mathrm{Tr}\left(\mathbf{\Lambda}\mathbf{\Psi}_1\mathbf{\Lambda}\mathbf{\Psi}_2\right) \tag{3}$$

where $\mathbf{\Psi}_1 = \mathbf{U}^*\mathbf{\Theta}_1\mathbf{U}$ and $\mathbf{\Psi}_2 = \mathbf{V}^*\mathbf{\Theta}_2\mathbf{V}$. It is clear that $\mathbf{\Psi}_1$ and $\mathbf{\Psi}_2$ are independent matrices that are both distributed as the upper $r \times r$ diagonal block of a Haar unitary matrix.

IV. Result

The main result of our contribution is the following

Theorem 1. *If $\lambda_1(\mathbf{X}_0) < 1$ then*

$$\limsup \mathbb{E}_0\left[\mathcal{L}(\mathbf{Y})^2\right] \leq \left(\frac{1}{1-\lambda_1(\mathbf{X}_0)^4}\right)^{r^2}$$

and it is not possible to find a consistent test.

We remind that we are looking for a condition on $\mathbf{X}_0$ (due to (2,3), this is a condition on $\mathbf{\Lambda}$) under which $\mathbb{E}\left[\exp\left(2n\eta\right)\right]$ is bounded. Evidently, the divergence may occur only when $\eta > 0$. We hence consider $E_1 = \mathbb{E}\left[\exp\left(2n\eta\right)\mathbb{I}_{\eta>\epsilon}\right]$ and $E_2 = \mathbb{E}\left[\exp\left(2n\eta\right)\mathbb{I}_{\eta\leq\epsilon}\right]$, and prove that, for a certain small enough $\epsilon > 0$ to be specified later, $E_1 = o(1)$ and that E_2 is bounded.

V. The E_1 term: computation of the GRF of η.

It is clear that the boundedness of the integral E_1 is achieved when η rarely deviates from 0. As remarked in [10], the natural machinery to consider is this of the Large Deviation Principle (LDP). In essence, if η follows the LDP with rate n, there can be found a certain non-negative function called Good Rate Function (GRF) I_η such that for any Borel set A of $\mathbb{R}$, $\frac{1}{n}\log\mathbb{P}\left(\eta \in A\right)$ converges towards $\sup_{x\in A} -I_\eta(x)$. The existence of a GRF allows one to analyze the asymptotic behaviour of the integral E_1. In the next section, we thus justify that η follows a Large Deviation Principle with rate n, and we compute the associated GRF.

A. Computation of the GRF of η

Eq. (3) and the Cauchy-Schwarz inequality imply that the random variable η is bounded: $|\eta| \leq \eta_{\max}$ with $\eta_{\max} = \sum_{j=1}^r \lambda_j^2$.

We first recall that for $i = 1, 2$, the random matrix $\mathbf{\Psi}_i$ follows a LDP with rate n and that its GRF at the parameter $\boldsymbol{\psi} \in \mathbb{C}^{r\times r}$, $\left\|\boldsymbol{\psi}\right\|_2 \leq 1$, is $\log\det\left(\mathbf{I}_r - \boldsymbol{\psi}^*\boldsymbol{\psi}\right)$ (see Theorem 3-6 in [8]). Besides, η is a function of the i.i.d. matrices $(\mathbf{\Psi}_i)_{i=1,2}$ and therefore, the contraction principle applies to η (see Theorem 4.2.1 in [7]): it ensures that η follows a LDP with rate n and its GRF is such that, for each real $|x| \leq \eta_{\max}$, $-I_\eta(x)$ is the solution of the following optimization problem:

Problem 2. Maximize in $\mathbb{C}^{r\times r}$

$$\log\det\left(\mathbf{I} - \boldsymbol{\psi}_1^*\boldsymbol{\psi}_1\right) + \log\det\left(\mathbf{I} - \boldsymbol{\psi}_2^*\boldsymbol{\psi}_2\right). \tag{4}$$

under the constraints

$$\Re\mathrm{Tr}\left(\mathbf{\Lambda}\boldsymbol{\psi}_1\mathbf{\Lambda}\boldsymbol{\psi}_2\right) = x \tag{5}$$

$$\left\|\boldsymbol{\psi}_i\right\|_2 \leq 1,\ i = 1, 2 \tag{6}$$

We provide a closed-form solution of Problem 2. In this respect, we define for each $k = 1, \ldots, r$ the interval $\mathcal{I}_k$ defined by

$$\forall k = 1, ..., r-1: \quad \mathcal{I}_k =]\sum_{i=1}^{k} \left(\lambda_i^2 - \lambda_k^2\right), \sum_{i=1}^{k+1} \left(\lambda_i^2 - \lambda_{k+1}^2\right)] \tag{7}$$

and $\mathcal{I}_r =]\sum_{i=1}^{r} \left(\lambda_i^2 - \lambda_k^2\right), \eta_{\max}]$. It is easy to check that $(\mathcal{I}_k)_{k=1,...,r}$ are disjoint and that $\cup_{k=1}^{r} \mathcal{I}_k =]0, \eta_{\max}]$. The following result holds:

Theorem 3. *The maximum of Problem 2 is given by*

$$-I_\eta(x) = 2 \sum_{k=1}^{r} \log \left(\left[\frac{\sum_{i=1}^{k} \lambda_i^2 - |x|}{k} \right]^k \frac{1}{\Pi_{i=1}^{k} \lambda_i^2} \right) \mathbb{I}_{\mathcal{I}_k}(|x|) \tag{8}$$

It is easy to check that the function $x \mapsto -I_\eta(x)$ is continuous on $]0, \eta_{\max}[$. The proof of Theorem 3 is provided in the Appendix.

We illustrate Theorem 3 through the following experiment. The rank of the spike is fixed to $r = 3$ and the singular values have been set to $(\lambda_1, \lambda_2, \lambda_3) = (1\ ,\ 0.7\ ,\ 0.2)$. We have computed millions of random samples of the matrices $(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2)$. Each pair is associated with a point (x, y) defined as $x = \Re\mathrm{Tr}\left(\boldsymbol{\Lambda}\boldsymbol{\psi}_1\boldsymbol{\Lambda}\boldsymbol{\psi}_2\right)$ and $y = \sum_{i=1}^{2} \log\det\left(\mathbf{I} - \boldsymbol{\psi}_i^*\boldsymbol{\psi}_i\right)$. We obtain a cloud of points, the upper envelope of which is expected to be $-I_\eta(x)$. We have also plotted the graph of the function $y = -I_\eta(x)$. In addition, we mention that, in the more general context of tensors of order d, the second-order moment of $\mathcal{L}(\mathbf{Y})$ is still given by (2) but the random variable - call it η_d - has a more complicated form than (3), see [5]; the asymptotics of the term E_1 can still be studied by evaluating the GRF of η_d. This GRF is the solution of an optimization problem that, apparently, cannot be solved in closed form for $d \geq 3$. In [5], an upper bound of the opposite of the true GRF is computed; this upper bound, valid for any d is given for $d = 2$ by $\log\left(1 - \frac{|x|}{\eta_{\max}}\right)$. We thus also represent in Figure V-A this upper bound; clearly, it is not tight.

B. Computation of E_1

The Varadhan lemma (see Theorem 4.3.1 in [7]) states that $\frac{1}{n}\log\mathbb{E}\left[\exp\left(2n\eta\right)\mathbb{I}_{\eta>\epsilon}\right] \to \sup_{x>\epsilon}\left(2x - I_\eta(x)\right)$ and hence the E_1 term converges towards 0 when $\sup_{x>\epsilon}\left(2x - I_\eta(x)\right) < 0$. Consider any of the intervals $\mathcal{I}_k$ defined in (7). The derivative of $2x - I_\eta(x)$ for any $x \in \mathcal{I}_k$ is $2 - 2k/(\lambda_1^2 + ... + \lambda_k^2 - x)$: it is decreasing on $\mathcal{I}_k$ and the limit in the left extremity of $\mathcal{I}_k$, i.e. $(\sum_{j=1}^{k-1} \lambda_j^2) - (k-1)\lambda_k^2$, is simply $2\left(1 - \frac{1}{\lambda_k^2}\right)$. If $\lambda_1\left(\mathbf{X}_0\right) < 1$, then for all the indices k, $1 - \frac{1}{\lambda_k^2} < 0$. This shows that $2x + I_\eta(x)$ is strictly decreasing on every $\mathcal{I}_k$. Hence, for every $x \in]0, \eta_{\max}]$, we have $2x - I_\eta(x) < 0 - I_\eta(0) = 0$. We have proved that $E_1 = o(1)$.

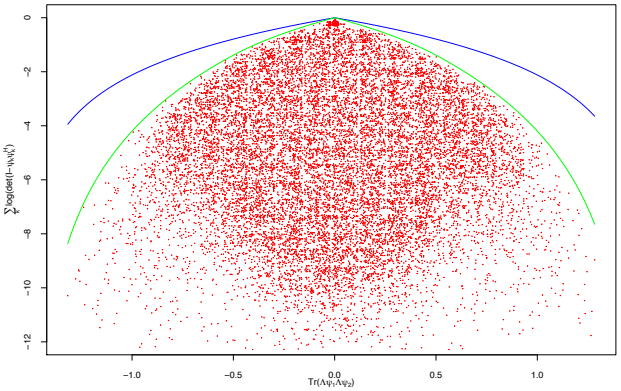

Fig. 1. graph of $-I_\eta(x)$ seen as an upper envelope. Upper curve: the upper bound computed in [5]

VI. The E_2 term: concentration of η.

Notice that the upper block $r \times r$ $\boldsymbol{\Psi}$ of a unitary Haar matrix $\boldsymbol{\Theta}$ has the same distribution as

$$\mathbf{G}\left(\tilde{\mathbf{G}}^*\tilde{\mathbf{G}}\right)^{-1/2}$$

where the $n \times r$ matrix $\tilde{\mathbf{G}}$ has i.i.d. entries distributed as $\mathcal{N}_{\mathbb{C}}(0,1)$ and $\mathbf{G}$ is the top $r \times r$ block of $\tilde{\mathbf{G}}$. Obviously, $\mathbb{E}[\tilde{\mathbf{G}}^*\tilde{\mathbf{G}}] = n\mathbf{I}$. It is a standard result that a random variable distributed as a $\chi^2(n)$ is concentrated around its mean. This can be easily extended to the matrix $\tilde{\mathbf{G}}^*\tilde{\mathbf{G}}$:

Lemma 4. *For any $0 < \delta < 1$, there exists a constant c such that*

$$\mathbb{P}\left(\left\|\frac{1}{n}\tilde{\mathbf{G}}^*\tilde{\mathbf{G}} - \mathbf{I}\right\|_2 > \delta\right) \leq c\exp\left(-n\frac{\delta^2}{2}\right).$$

We take $\tilde{\mathbf{G}}_1$ and $\tilde{\mathbf{G}}_2$ independent, distributed as $\tilde{\mathbf{G}}$ and consider the upper $r \times r$ blocks $\mathbf{G}_1$ and $\mathbf{G}_2$ of $\tilde{\mathbf{G}}_1$ and $\tilde{\mathbf{G}}_2$. It follows that η has the same distribution as $2\Re\mathrm{Tr}\left(\boldsymbol{\Lambda}\mathbf{G}_1\left(\tilde{\mathbf{G}}_1^*\tilde{\mathbf{G}}_1\right)^{-1/2}\boldsymbol{\Lambda}\mathbf{G}_2\left(\tilde{\mathbf{G}}_2^*\tilde{\mathbf{G}}_2\right)^{-1/2}\right)$. Take now any $\delta < 1$. We may split the integral E_2 in two parts:

$$\underbrace{\mathbb{E}\left[\exp\left(2n\eta\right)\mathbb{I}_{\{\eta\leq\epsilon\}\cap\mathcal{B}_1^c\cap\mathcal{B}_2^c}\right]}_{E_2'} + \underbrace{\mathbb{E}\left[\exp\left(2n\eta\right)\mathbb{I}_{\{\eta\leq\epsilon\}\cap(\mathcal{B}_1\cup\mathcal{B}_2)}\right]}_{E_2''}.$$

where we have defined the events $\mathcal{B}_i = \left\{\left\|\frac{1}{n}\tilde{\mathbf{G}}_i^*\tilde{\mathbf{G}}_i - I\right\|_2 > \delta\right\}$. Thanks to the above concentration result, we have

$$\begin{aligned} E_2'' &\leq \exp(2n\epsilon)\left(\mathbb{P}(\mathcal{B}_1) + \mathbb{P}(\mathcal{B}_2)\right) \\ &\leq \exp(2n\epsilon)2c\exp\left(-n\delta^2/2\right) \end{aligned}$$

As it is always possible to choose δ and ϵ such that $\delta^2 - 4\epsilon > 0$ and $\delta < 1$ it follows that $E_2'' = o(1)$.

Let us now inspect the term E_2'. Since we have, for $i = 1, 2$, $\left\|\frac{1}{n}\tilde{\mathbf{G}}_i^*\tilde{\mathbf{G}}_i - I\right\|_2 \leq \delta$, then there exist $\boldsymbol{\Delta}_i$ for $i = 1, 2$ such that

$\left(\tilde{\mathbf{G}}_i^*\tilde{\mathbf{G}}_i\right)^{-1/2} = \frac{1}{\sqrt{n}}(\mathbf{I}+\mathbf{\Delta}_i)$with $\|\mathbf{\Delta}_i\|_2 \leq \delta/2$. We hence have

$$E_2' \leq \mathbb{E}\left[\exp\left(2\Re\mathrm{Tr}\left(\mathbf{\Lambda}\mathbf{G}_1\left(\mathbf{I}+\mathbf{\Delta}_1\right)\mathbf{\Lambda}\mathbf{G}_2(\mathbf{I}+\mathbf{\Delta}_2)\right)\right)\right].$$

We expand $2\Re\mathrm{Tr}\left(\mathbf{\Lambda}\mathbf{G}_1\left(\mathbf{I}+\mathbf{\Delta}_1\right)\mathbf{\Lambda}\mathbf{G}_2\left(\mathbf{I}+\mathbf{\Delta}_2\right)\right)$ as the sum of four terms. Take for instance

$$T_2 = 2\Re\mathrm{Tr}\left(\mathbf{\Lambda}\mathbf{G}_1\mathbf{\Delta}_1\mathbf{\Lambda}\mathbf{G}_2\right)$$

Thanks to von Neumann's lemma [9], we have

$$\begin{aligned} T_2 &\leq 2\sum_{k=1}^{r}\lambda_k(\mathbf{\Delta}_1)\lambda_k\left(\mathbf{\Lambda}\mathbf{G}_2\mathbf{\Lambda}\mathbf{G}_1\right) \\ &\leq 2\left\|\mathbf{\Delta}_1\right\|\sum_{k=1}^{r}\lambda_k\left(\mathbf{\Lambda}\mathbf{G}_2\mathbf{\Lambda}\mathbf{G}_1\right)\end{aligned}$$

As $\sum_{k=1}^{r}\lambda_k\left(\mathbf{\Lambda}\mathbf{G}_2\mathbf{\Lambda}\mathbf{G}_1\right) \leq \sqrt{r}\sqrt{\sum_{k=1}^{r}\lambda_k^2\left(\mathbf{\Lambda}\mathbf{G}_2\mathbf{\Lambda}\mathbf{G}_1\right)}$, it yields

$$T_2 \leq 2\left\|\mathbf{\Delta}_1\right\|\sqrt{r}\sqrt{\mathrm{Tr}\left(\mathbf{\Lambda}\mathbf{G}_2\mathbf{\Lambda}\mathbf{G}_1\mathbf{G}_1^*\mathbf{\Lambda}\mathbf{G}_2^*\mathbf{\Lambda}\right)}.$$

Invoking the von Neumann's lemma three times, it holds that

$$\begin{aligned} T_2 &\leq 2\left\|\mathbf{\Delta}_1\right\|\sqrt{r}\left\|\mathbf{\Lambda}^2\right\|\sqrt{\mathrm{Tr}\ \left(\mathbf{G}_1\mathbf{G}_1^*\right)\mathrm{Tr}\ \left(\mathbf{G}_2\mathbf{G}_2^*\right)} \\ &\leq \sqrt{r}\left\|\mathbf{\Delta}_1\right\|\left\|\mathbf{\Lambda}^2\right\|\left(\mathrm{Tr}\ \left(\mathbf{G}_1\mathbf{G}_1^*\right)+\mathrm{Tr}\ \left(\mathbf{G}_2\mathbf{G}_2^*\right)\right)\end{aligned}$$

Similar manipulations can be done on the other terms of the expansion. so that E_2' is less than

$$\mathbb{E}\left[\exp\left(2\Re\mathrm{Tr}\left(\mathbf{\Lambda}\mathbf{G}_1\mathbf{\Lambda}\mathbf{G}_2\right)+\beta\mathrm{Tr}\left(\left(\mathbf{G}_1\mathbf{G}_1^*\right)+\mathrm{Tr}\left(\mathbf{G}_2\mathbf{G}_2^*\right)\right)\right)\right]$$

with $\beta = \frac{\sqrt{r}}{2}\delta(2+\delta)\left\|\mathbf{\Lambda}\right\|^2$. The above expectation is to be understood as the expectation over $(\mathbf{G}_1,\mathbf{G}_2)$. As $\mathbf{G}_1$ and $\mathbf{G}_2$ are independent, we consider first the expectation over $\mathbf{G}_1$. This gives, up to the factor $\exp\left(\beta\mathrm{Tr}\ \left(\mathbf{G}_2\mathbf{G}_2^*\right)\right)$

$$\pi^{-r^2}\int\exp\left(2\Re\mathrm{Tr}\left(\mathbf{g}_1\mathbf{E}\right)+(\beta-1)\,\mathrm{Tr}\ \left(\mathbf{g}_1^*\mathbf{g}_1\right)\right)d\mathbf{g}_1$$

with $\mathbf{E} = \mathbf{\Lambda}\mathbf{G}_2\mathbf{\Lambda}$. It is always possible to choose δ such that $\beta < 1$. With such a β, the above integral is

$$(1-\beta)^{-r^2}\exp\left(\frac{1}{4}\left(\frac{2}{\sqrt{1-\beta}}\right)^2\mathrm{Tr}\left(\mathbf{E}\mathbf{E}^*\right)\right)$$

As $\mathrm{Tr}\left(\mathbf{E}\mathbf{E}^*\right) \leq \left\|\mathbf{\Lambda}\right\|_2^4\mathrm{Tr}\left(\mathbf{G}_2\mathbf{G}_2^*\right)$ we finally obtain after multiplying by $\exp\left(\beta\mathrm{Tr}\ \left(\mathbf{G}_2\mathbf{G}_2^*\right)\right)$ and taking the expectation over $\mathbf{G}_2$:

$$E_2' \leq \frac{(1-\beta)^{-r^2}}{\pi^{r^2}}\int\exp\left(-\frac{(1-\beta)^2-\left\|\Lambda\right\|_2^4}{1-\beta}\mathrm{Tr}\left(\mathbf{g}_2^*\mathbf{g}_2\right)\right)d\mathbf{g}_2.$$

If $\left\|\mathbf{\Lambda}\right\|_2^2 < 1$, it is always possible to adjust δ such that the above integral converges. In this condition, we have

$$E_2' \leq \left(\frac{1}{(1-\beta)^2-\left\|\mathbf{\Lambda}\right\|_2^4}\right)^{r^2}.$$

This must be true for all β arbitrarily small, hence the result.

APPENDIX

We prove Theorem 3 when $x > 0$. As the function to be maximized converges towards $-\infty$ if $\|\boldsymbol{\psi}_1\| \to 1$ or $\|\boldsymbol{\psi}_2\| \to 1$, any argument $(\boldsymbol{\psi}_1,\boldsymbol{\psi}_2)$ of the maximization problem satisfies $\|\boldsymbol{\psi}_i\| < 1,\ i = 1,2$. Therefore, the Karush-Kuhn-Tucker (KKT) conditions imply the existence of a scalar Lagrange multiplier $\mu \geq 0$ such that $(\boldsymbol{\psi}_1,\boldsymbol{\psi}_2)$ is a stationary point of the Lagrangian $\ell(\boldsymbol{\psi}_1,\boldsymbol{\psi}_2,\mu)$ defined by $\sum_{i=1}^{2}\log\det\left(\mathbf{I}_r-\boldsymbol{\psi}_i^*\boldsymbol{\psi}_i\right)+\mu\ \Re\mathrm{Tr}\left(\mathbf{\Lambda}\boldsymbol{\psi}_1\mathbf{\Lambda}\boldsymbol{\psi}_2\right)$. As ℓ is a real valued function, a stationary point is computedwhen setting the differential w.r.t. the entries of $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$ to zero. It can be checked that $(\boldsymbol{\psi}_1,\boldsymbol{\psi}_2)$ is a stationary point of ℓ when

$$\begin{aligned}\mu\,\mathbf{\Lambda}\boldsymbol{\psi}_2\mathbf{\Lambda} &= \boldsymbol{\psi}_1^*(\mathbf{I}-\boldsymbol{\psi}_1\boldsymbol{\psi}_1^*)^{-1} \\ \mu\,\mathbf{\Lambda}\boldsymbol{\psi}_1\mathbf{\Lambda} &= \boldsymbol{\psi}_2^*(\mathbf{I}-\boldsymbol{\psi}_2\boldsymbol{\psi}_2^*)^{-1}\end{aligned}$$

In a first step, these equations can be shown to be satisfied only if $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$ are diagonal up to permutations of the columns. Then, is can be deduced that there exists a diagonal matrix $0 \leq \mathbf{P} \leq \mathbf{I}$ and a matrix of permutation $\mathbf{\Pi}$ such that $\log\det\left(\mathbf{I}_r-\boldsymbol{\psi}_1^*\boldsymbol{\psi}_1\right)+\log\det\left(\mathbf{I}_r-\boldsymbol{\psi}_2^*\boldsymbol{\psi}_2\right) = 2\log\det(\mathbf{I}-\mathbf{P})$and $\Re\mathrm{Tr}\left(\mathbf{\Lambda}\boldsymbol{\psi}_1\mathbf{\Lambda}\boldsymbol{\psi}_2\right) = \mathrm{Tr}(\mathbf{\Lambda}\mathbf{\Pi}^*\mathbf{\Lambda}\mathbf{\Pi}\,\mathbf{P})$. This invites us to consider the following

Problem 5. Maximize

$$\log\det(\mathbf{I}-\mathbf{P}) \tag{9}$$

jointly over all the $r!$ permutations $\mathbf{\Pi}$ and over diagonal matrices $\mathbf{P}$ verifying $0 \leq \mathbf{P} \leq \mathbf{I}$ and the constraint

$$\mathrm{Tr}(\mathbf{\Lambda}\mathbf{\Pi}^*\mathbf{\Lambda}\mathbf{\Pi}\,\mathbf{P}) = x. \tag{10}$$

In a first step, we set $\mathbf{\Pi} = \mathbf{I}$ in the above problem and consider the

Problem 6. Maximize

$$\sum_{i=1}^{r}\log(1-p_i) \tag{11}$$

under the constraints that $0 \leq p_i \leq 1$ for each $i = 1,\ldots,r$ and

$$\sum_{i=1}^{r}\lambda_i^2\,p_i = x. \tag{12}$$

The maximum is denoted by $J_\Lambda(x)$.

This is a variant of the celebrated water-filling problem (see *e.g.* [13] and Chap. 9 of [6]) that was solved to evaluate the capacity of a frequency selective Gaussian channel, the difference being that in the latter problem, $\log(1-p_i)$ is replaced by $\log(1+p_i)$.In order to solve Problem 6, we assume that the non zero singular values $(\lambda_i)_{i=1,\ldots,r}$ are distinct. If this is not the case, a standard perturbation argument can be used in order to address the general case. As the function to be maximized is strictly concave on the set defined by the constraints, the maximum is reached at a unique point $\mathbf{p}_*$ verifying $p_{i,*} < 1$ for each i. We consider the Lagrangian corresponding to Problem (6) given by $\sum_{i=1}^{r}\log(1-p_i)+\mu\left(\sum_{i=1}^{r}\lambda_i^2\,p_i\right)+\sum_{i=1}^{r}\delta_i p_i$

where $\mu \geq 0$ and $\delta_i \geq 0$ for $i = 1, \ldots, r$. The partial derivatives w.r.t. parameters $(p_i)_{i=1,\ldots,r}$ are zero at $\mathbf{p}_*$. This leads to

$$\text{for } i = 1, \ldots, r: \quad \frac{1}{1 - p_{i,*}} = \mu_* \lambda_i^2 + \delta_{i,*} \tag{13}$$

The first remark is that necessarily, these equations imply that the numbers $p_{i,*}$ are sorted in decreasing order. To verify this claim, we assume that $i < j$ and that $p_{i,*} = 0$ and $p_{j,*} > 0$. Then, it holds that $\mu_* \lambda_i^2 + \delta_{i,*} = 1$ and that $\mu_* \lambda_j^2 = \frac{1}{1-p_{j,*}} > 1$ because $p_{j,*} > 0$ implies $\delta_{j,*} = 0$. Therefore, $\lambda_i^2 \leq \frac{1}{\mu_*} < \lambda_j^2$, a contradiction because $\lambda_i^2 \geq \lambda_j^2$. We denote by $s(x)$ the number of non-zero entries of $\mathbf{p}_*$. Hence, the first $s(x)$ entries of $\mathbf{p}_*$ are non zero. Morever, the equations $\mu_* \lambda_i^2 = \frac{1}{1-p_{i,*}}$ for $i = 1, \ldots, s(x)$ imply that $p_{1,*} \geq \ldots \geq p_{s(x),*} > 0 = p_{s(x)+1,*} = \ldots = p_{r,*}$.

We now analytically characterize $s(x)$. On the one hand, (13) computed at for $i = s(x)$ and for $i = s(x) + 1$ both imply

$$\lambda_{s(x)+1}^2 \leq \frac{1}{\mu_*} < \lambda_{s(x)}^2 \tag{14}$$

On the other hand, the constraint (12) imposes that $1/\mu_*$ verifies

$$\frac{1}{\mu_*} = \frac{\sum_{i=1}^{s(x)} \lambda_i^2 - x}{s(x)}.$$

Therefore, it holds that

$$(\sum_{i=1}^{s(x)} \lambda_i^2) - s(x)\lambda_{s(x)}^2 < x \leq (\sum_{i=1}^{s(x)} \lambda_i^2) - s(x)\lambda_{s(x)+1}^2 \tag{15}$$

such that $s(x)$ coincides with the integer k for which $x \in \mathcal{I}_k$ (see (7) for the definition of these intervals). The maximum $\sum_{i=1}^{s(x)} \log(1 - p_{i,*})$ is direcly computed as

$$J_\Lambda(x) = \log \left(\left[\frac{\sum_{i=1}^{s(x)} \lambda_i^2 - x}{s(x)} \right]^{s(x)} \frac{1}{\Pi_{i=1}^{s(x)} \lambda_i^2} \right) \tag{16}$$

In order to show that the GRF of η is $I_\eta(x) = -2J_\Lambda(x)$, it remains to show that the solution of Problem 5 is reached when the permutation matrix $\mathbf{\Pi}$ is the identity. In this respect, we introduce a nested problem motivated by the following observation. We denote by $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ the r–dimensional vectors whose components are respectively the diagonal entries of $\mathbf{\Lambda}^2$ and of $\mathbf{\Lambda \Pi}^* \mathbf{\Lambda \Pi}$ arranged in the decreasing order. Evidently, $\boldsymbol{\alpha}$ majorizes $\boldsymbol{\beta}$ in the sense that

$$\text{for } k = 1, \ldots, r: \quad \sum_{i=1}^{k} \alpha_i \geq \sum_{i=1}^{k} \beta_i \tag{17}$$

We thus consider the relaxed problem

Problem 7. Maximize $\log\det(\mathbf{I} - \mathbf{P})$ over the diagonal matrices $0 \leq \mathbf{P} \leq \mathbf{I}$ and over vectors $\boldsymbol{\beta} = (\beta_1, ..., \beta_r)$ satisfying $\beta_1 \geq \beta_2 \geq \ldots \beta_r \geq 0$, the majorization constraint (17), and the equality constraint

$$\sum_{i=1}^{r} \beta_i \, p_i = x \tag{18}$$

The maximum of Problem 7 is above the maximum of Problem 5 which is itself above the maximum $J_\Lambda(x)$ of Problem 6. We actually show that the maximum of Problem 7 is less than $J_\Lambda(x)$, and that it is reached for a vector $\boldsymbol{\beta}$ that coincides with $\boldsymbol{\alpha}$. This will imply that the optimal permutation $\mathbf{\Pi}$ in Problem 5 is $\mathbf{I}$ and $I_\eta(x) = -2J_\Lambda(x)$.

We give some elements for solving Problem 7. We consider a stationary point $(\mathbf{p}_*, \boldsymbol{\beta}_*)$ of the associated Lagrangian and compute the KKT conditions. We suppose that this stationary point attains the maximum. If s denotes the number of non-zero components in $\mathbf{p}_*$, we prove that, necessarily, $p_{1,*} \geq p_{2,*} \geq ... \geq p_{s,*} > 0$ and $\beta_{1,*} \geq \beta_{2,*} \geq ... \geq \beta_{s,*}$. We let j_1 be the first index such that $\sum_{i=1}^{j_1} \alpha_i > \sum_{i=1}^{j_1} \beta_i$ (this index exists otherwise $\boldsymbol{\beta}_* = \boldsymbol{\alpha}$ and the problem is solved). This implies that $\beta_{i,*} = \alpha_i$ for all indices $i = 1, ..., j_1 - 1$. Notice this fact: if we suppose that the condition $\sum_{i=1}^{j_1+k} \alpha_i > \sum_{i=1}^{j+k} \beta_i$ is true whatever k, then it is possible to add a small $\epsilon > 0$, and update $\beta_{j_{1,*}}$ as $\beta_{j_1,*} + \epsilon$ in such a way that the majorization constraints still hold, the constraint (18) holds and the updated $\mathbf{p}_*$ increases the function to maximize. This is in contradiction with the definition of $(\mathbf{p}_*, \boldsymbol{\beta}_*)$. This means that there exists an index j_2 (we choose the smallest) such that $\sum_{i=1}^{j_1+j_2} \alpha_i = \sum_{i=1}^{j_1+j_2} \beta_{i,*}$. It can be shown that it is necessary that all the $\beta_{i,*}$ are equal for $i = j_1, ..., j_1 + j_2$. After some algebraic gymnastics, it can be shown that it in this case, all the inequalities (17) at $\boldsymbol{\beta}_*$ are saturated hence implying that $\boldsymbol{\beta}_* = \boldsymbol{\alpha}$. The value of $\sum_i \log(1 - p_{i,*})$ equals $J_\Lambda(x)$.

REFERENCES

[1] Z. Bai and J.W. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*. Springer-Verlag Series in Statistics, 2010.

[2] J. Banks, C. Moore, R. Vershynin, N. Verzelen, and J. Xu. Information-theoretic bounds and phase transitions in clustering, sparse PCA, and submatrix localization. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 1137–1141, June 2017.

[3] Florent Benaych-Georges and Raj Rao Nadakuditi. The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis*, 111:120–135, 2012.

[4] P. Bianchi, M. Debbah, M. Maïda, and M. Najim. Performance of statistical tests for single source detection using random matrix theory. *IEEE Transactions on Information Theory*, 57(4):2400–2419, 2011.

[5] A. Chevreuil and Ph. Loubaton. On the non-detectability of spiked large random tensors. *Arxiv, 1802.07093*, 2018.

[6] T.M. Cover and J.A. Thomas. *Elements of Information Theory, 2nd Edition*. Wiley Interscience, 2006.

[7] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Springer-Verlag Berlin Heidelberg, 2009.

[8] Fabrice Gamboa and Alain Rouault. Operator-valued spectral measures and large deviations. *J. of Stat. PLanning and Inference*, 154(3):72–86, 2014.

[9] L. Mirsky. A trace inequality of John von Neumann. *Monatshefte für Mathematik*, 79(4):303–306, Dec 1975.

[10] Andrea Montanari, Daniel Reichman, and Ofer Zeitouni. On the limitation of spectral methods: from the gaussian hidden clique problem to rank one perturbations of gaussian tensors. *IEEE Trans. Inf. Theor.*, 63(3):1572–1579, March 2017.

[11] R.R Nadakuditi and A. Edelman. Sample eigenvalue based detection of high-dimensional signals in white noise using relatively few samples. *IEEE Transactions on Signal Processing*, 56(7):2625–2637, 2008.

[12] A. Onatski, M.J. Moreira, and M. Hallin. Asymptotic power of sphericity tests for high-dimensional data. *Ann. Statistics*, 41(3):1204–1231, 2013.

[13] H.S. Witsenhausen. A determinant maximization problem occuring in the theory of data communications. *SIAM J. Appl. Math*, 29(3):515–522, 1975.