On the Maximal Size of Large-Average Submatrices in a Gaussian Random Matrix: Theoretical and Numerical Approaches

Jingxuan Bao
71671086
Supervisor: Prof. Robin Pemantle

University of Pennsylvania
Applied Math and Computational Science
School of Arts and Sciences
November, 2019

Abstract

Based on the article "On the Maximal Size of Large-Average and ANOVA-Fit Submatirces in a Gaussian Random Matrix" by Sun and Nobel, we investigate the maximal size of submatrices with average of the values of such submatrices more than a fixed positive number in the Gaussian random matrix. We identify the limit behavior of the threshold of the size of submatrices theoretically and numerically. Our principal result is an inconsistency between the results of two approaches and we propose our own analysis from the theoretical and numerical perspectives.

keywords: Gaussian random matrix; large-average submatrix; sandwiching method; large-average submatrix threshold searcher algorithm

Contents

1 Introduction

DNA microarrays (also commonly known as DNA chip or biochip) allow scientists to measure the coexpression levels of large numbers of genes simultaneously or to genotype multiple regions of a genome. With the development of the DNA microarrays, many technologies and applications in biostatistics or bioinformatics thrive in modern society, such as gene expression profiling (Adomas et al., 2008), comparative genomic hybridization (Pollack et al., 1999; Moran et al., 2004), and SNP detection (Hacia et al., 1999).

In particular, one important aspect of DNA microarray study is assessments of differential expression in a group of genes (functional category). More specifically, a subarray of a microarray is obtained by restricting to a subset of subjects and a subset of genes. We are supposed to examine when it has a surprisingly high degree of similarity when studying some subarrays of microarray population. A method in statistics which is wildly used to solve this kind of problems is called Statistical Hypothesis Testing. In this report, we provide a toy model for the null hypothesis on microarrays, in which we substitude high average value of independent and identically distributed (i.i.d.) standard normal random variable to the high genetic similarity of microarrays, examing whether such kind of high similarity results from the genetic data, or it is a common property in a large Gaussian random matrix. In summary, instead of examine the microarray dataset, we evaluate the maximum size of large-average submatrices in a Gaussian random matrix.

A Gaussian random matrix is a matrix whose elements are identical independent standard normal random variables. In our report, based on the analysis from Sun and Nobel (2013), " On the Maximal Size of Large-Average and ANOVA-Fit Submatircs in a Gaussian Random Matrix", we introduce the method of finding the maximum size of submatrices with average greater than a certain number in the Gaussian random matrix, specifically finding an expression of maximum size of submatrices of Gaussian random matrix in terms of a certain number, from both theoretical and numerical approaches.

In this report, literature review including previous work is introduced in Section 2. The theoretical approach to find the probability bounds for the size of large-average submatrices in the square matrix is presented in Section 3. The numerical approach to find the threshold for large-average submatrices is given in Section 4. Section 5 contains our integration of theoretical result and numerical result.

2 Literature Review

Our report is mainly based on the paper published by *Bernoulli* " On the Maximal Size of Large-Average and ANOVA-Fit Submatircs in a Gaussian Random Matrix" (Sun and Nobel, 2013).

In the original article, the authors introduce main three theoretical methods to analyze the thresholds for the large-average submatrices:

- **Method 1: Bipartite Graphs**

  In this method, the authors expressed the $m \times n$ matrix X using a bipartite graph $G = (V, E)$, where V represents the vertex set of G containing two disjoint sets V_1 and V_2 with $|V_1| = m$, $|V_2| = n$, representing the rows and columns of matrix X respectively; and E represents the set of the edges connecting row $i \in V_1$ and column $j \in V_2$ with weight $x_{i,j}$. Vertices in the same vertex set are not allowed to have edge. In such scenario, the large-average submatrices of X are one-to-one correspondence with subgraphs of G, showing in the Figure 1.

  However, according to Dawande et al. (2001), to find the edge with maximum weight subgraph in a general bipartite graph, a slightly variation of this problem, is NP-complete. In other words, there is no fast algorithm to solve this kind of problem.

- **Method 2: Random Matrix Theory**

  In this method, the authors define the notations in Table 1.

  We summarize the logic of Mehod 2 in Figure 2.

  With the assumption that m and n grow with $\frac{m}{n} \to \alpha$ for $\alpha \geq 1$, and the dimensions k and l grow with $\frac{k}{\ln n} \to \infty$ and $0 < \frac{k}{l} < \infty$, we have

$$\mathbb{P}(\exists\, k \times l \, submatrix \, in \, W \, whose \, average \, > \tau) \to 0$$

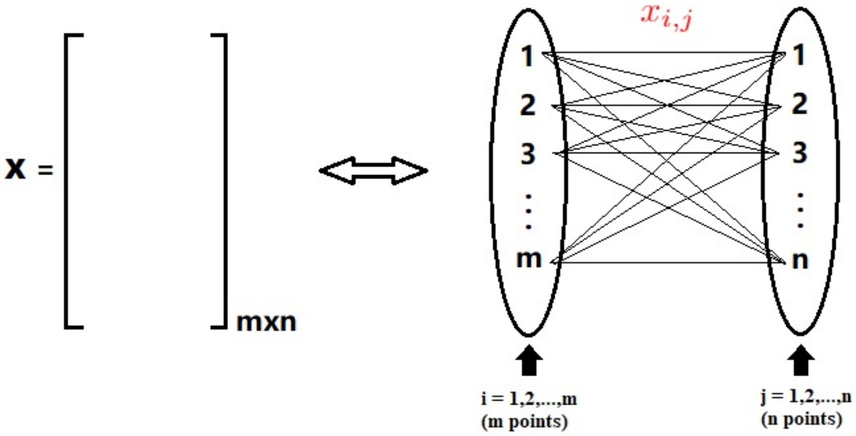

Figure 1: Matrix Bipartite Graph Correspondence

Name	Definition
W	$m \times n$ Gaussian random matrix
τ	Fixed number that is greater than 0
a	Indicator vectors having k non-zero components, $a \in \{0,1\}^m$
b	Indicator vectors having l non-zero components, $b \in \{0,1\}^n$
U	Submatrix whose rows and columns are indexed by a and b, $U = ab^T$
S	Rank-one matrix, $S = (1+\delta)\tau ab^T$, where $\delta > 0$
Y	Sum of W and S (a perturbed version of S), $Y = W + S$

Table 1: Definition of Notations

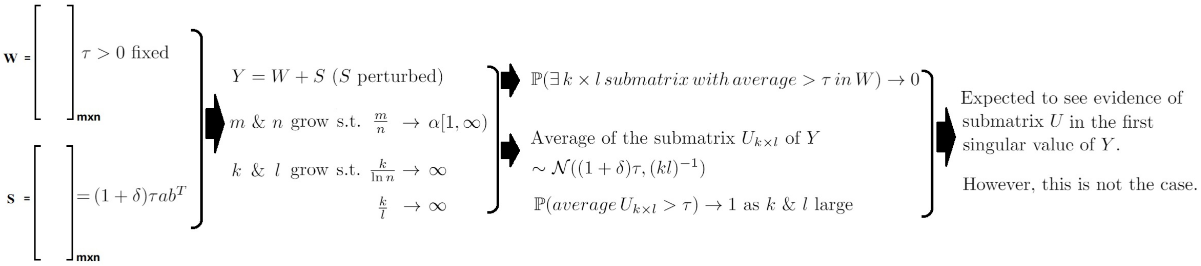
W =
$\tau > 0$ fixed
mxn
S =
$= (1+\delta)\tau ab^T$
mxn
$Y = W + S$ (S perturbed)
m & n grow s.t. $\frac{m}{n} \to \alpha[1,\infty)$
k & l grow s.t. $\frac{k}{\ln n} \to \infty$
$\frac{k}{l} \to \infty$
$\mathbb{P}(\exists\, k \times l\, submatrix\, with\, average > \tau\, in\, W) \to 0$
Average of the submatrix $U_{k\times l}$ of Y
$\sim \mathcal{N}((1+\delta)\tau, (kl)^{-1})$
$\mathbb{P}(average\, U_{k\times l} > \tau) \to 1$ as k & l large
Expected to see evidence of submatrix U in the first singular value of Y.
However, this is not the case.

Figure 2: Logic Graph of Method 2

On the other hand, we also have the average of the $k \times l$ submatrix U of Y has distribution $\mathcal{N}((1+\delta)\tau, (kl)^{-1})$, which means when k and l are large, we have

$$\mathbb{P}(average\, of\, U > \tau) \to 1$$

We expect to see evidence of submatrix U in the first singular value of Y; however this is not the case.

- **Method 3: Probability Bounds and A Finite Interval Concentration Result**

  Method 3 is the main method the authors introduce in the original article. This method can be divided into two parts: the first part concludes the expected number of $k \times k$ submatrices U of W_n with $F(U) \geq \tau$ is less than one, where $F(\cdot)$ denotes calculating the average of all the elements; and the second part deduces when the size of Gaussian random matrix n is sufficient large, the largest value k such that $n \times n$ Gaussian matrix W_n contains a $k \times k$ submatrix U with average greater or equal to τ for some fixed $\tau > 0$ approaches to $\frac{4}{\tau^2} \ln n$ almost surely when n goes to infinity.

We will study this method in detail in the next section.

3 Methodology

In this section, we review the paper written by Sun and Nobel (2013) and theoretically formalize the model to find the thresholds and bounds for large average submatrices following the same process of derivation of the original article.

3.1 Definition of Notations with Example Explanations

In this part of the section, we define the notions we will use later when deriving the theoretical result with example explanations.

- $W = \{w_{i,j} : i, j \geq 1\}$: Infinite array of independent $N(0, 1)$ random variables.

  Explanation: W has the form shown below

$$\begin{bmatrix} w_{1,1} & w_{1,2} & w_{1,3} & w_{1,4} & w_{1,5} & \cdots \\ w_{2,1} & w_{2,2} & w_{2,3} & w_{2,4} & w_{2,5} & \cdots \\ w_{3,1} & w_{3,2} & w_{3,3} & w_{3,4} & w_{3,5} & \cdots \\ w_{4,1} & w_{4,2} & w_{4,3} & w_{4,4} & w_{4,5} & \cdots \\ w_{5,1} & w_{5,2} & w_{5,3} & w_{5,4} & w_{5,5} & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

  where every $w_{i,j}$ denotes i.i.d. random variable with distribution $\mathcal{N}(0, 1)$.

- $W_n = \{w_{i,j} : 1 \leq i, j \leq n\}$: $n \times n$ Gaussian random matrix equal to upper left-hand corner of W.

  Example: W_5 has the form shown below:

$$\begin{bmatrix} w_{1,1} & w_{1,2} & w_{1,3} & w_{1,4} & w_{1,5} \\ w_{2,1} & w_{2,2} & w_{2,3} & w_{2,4} & w_{2,5} \\ w_{3,1} & w_{3,2} & w_{3,3} & w_{3,4} & w_{3,5} \\ w_{4,1} & w_{4,2} & w_{4,3} & w_{4,4} & w_{4,5} \\ w_{5,1} & w_{5,2} & w_{5,3} & w_{5,4} & w_{5,5} \end{bmatrix}$$

- $U = \{w_{i,j} : i \in A, j \in B\}$: submatrix of W_n, where $A, B \subseteq \{1, 2, .., n\}$; we write Cartesian product $C = A \times B$, and we can write $U = W_n[C]$.

  Example: if $A = \{1, 3\}$, $B = \{2, 4\}$, then $U = W_n[\{1, 3\} \times \{2, 4\}]$ is

$$\begin{bmatrix} w_{1,2} & w_{1,4} \\ w_{3,2} & w_{3,4} \end{bmatrix}$$

- $F(U)$: the average of submatrix U,

$$F(U) = \frac{1}{|C|} \sum_{(i,\, j) \in C} w_{i,\, j} = \frac{1}{|A||B|} \sum_{i \in A,\, j \in B} w_{i,\, j}$$

- $K_\tau(W_n)$: the largest $k \geq 0$ such that W_n contains a $k \times k$ submatrix U with $F(U) \geq \tau$ for fixed $\tau > 0$ and $n \geq 1$.

- $\Gamma_k(n, \tau)$, the number of $k \times k$ submatrices in W_n with average $\geq \tau$:

$$\Gamma_k(n, \tau) = \sum_{U \in S_k} \mathbf{1}_{F(U) \geq \tau}$$

  where S_k denotes all the $k \times k$ submatrices of W_n.

3.2 Mathematical Derivation

In this derivation part of the section, we follow exactly the steps of the article written by Sun and Nobel (2013), but we include more detail to let the original paper reading-friendly to those who do not have a very concrete math background. The main idea of the deduction in this report is to find an upper and a lower bound of $K_\tau(W_n)$, and then after some manipulation, we obtain an almost sure limit behavior of $K_\tau(W_n)$ by sending n, the dimension of Gaussian random matrix, to infinity.

First, we start from deducing there exists a special positive real value that is unique and when k is greater than such value, the expected number of $k \times k$ submatrices U of W_n with $F(U) \geq \tau$ is less than one.

Note that we have

$$F(W_k \geq \tau) = 1 - \Phi(\tau k),$$

where $\Phi(\cdot)$ represents the normal cumulative distribution function (CDF).

Proof. Since for any $i = 1, 2, \ldots, k^2$, we have $X_i \sim \mathcal{N}(0,1)$, and that

$$\frac{\sum_{i=1}^{k^2}}{\sqrt{k^2}} \sim \mathcal{N}(0,1).$$

We have

$$F(W_k) = \frac{\sum_{i=1}^{k^2}}{k^2} \sim \mathcal{N}(0, \frac{1}{k^2}).$$

As a result, we have

$$\begin{aligned} F(W_k \geq \tau) &= \int_\tau^\infty \frac{1}{\sqrt{2\pi k^2}} e^{-\frac{x^2}{2\sigma^2}} dx \\ &= \int_{\tau k}^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \\ &= 1 - \int_{-\infty}^{\tau k} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \\ &= 1 - \Phi(\tau k). \end{aligned}$$

□

Now, we are able to deduce an expression of the expected number of $k \times k$ submatrices in W_n with average $\geq \tau$

$$\begin{aligned} E\Gamma_k(n,\tau) =& (total\ number\ of\ k \times k\ submatrices\ of\ W_n) \times \\ & (probability\ of\ the\ average\ of\ a\ k \times k\ submatrix \geq \tau) \\ =& |S_k|\, \mathbb{P}(F(W_k) \geq \tau) \\ =& \begin{pmatrix} n \\ k \end{pmatrix}^2 (1 - \Phi(\tau k)). \end{aligned}$$

Using a standard bound on $(1 - \Phi(\cdot))$, i.e., $(1 - \Phi(\tau k)) \leq e^{-\frac{\tau^2 k^2}{2}}$ to get an upper bound,

$$E\Gamma_k(n,\tau) \leq \begin{pmatrix} n \\ k \end{pmatrix}^2 e^{-\frac{\tau^2 k^2}{2}}. \tag{1}$$

By application a slight variation version of Stirling approximation, $\sqrt{2\pi n}(\frac{n}{e})^n e^{\frac{1}{12n+1}} \leq n! \leq \sqrt{2\pi n}(\frac{n}{e})^n e^{\frac{1}{12n}}$ (Maria, 1965), to represent the combination; and rewrite the upper bound. Equation (1) becomes

$$\begin{aligned}
E\Gamma_k(n,\tau) \leq & \left(\begin{array}{c} n \\ k \end{array} \right)^2 e^{-\frac{\tau^2 k^2}{2}} \\
\leq & \Big(\frac{\sqrt{2\pi n}(\frac{n}{e})^n e^{\frac{1}{12n}}}{\sqrt{2\pi k}(\frac{k}{e})^k e^{\frac{1}{12k+1}} \sqrt{2\pi(n-k)}(\frac{n-k}{e})^{n-k} e^{\frac{1}{12(n-k)+1}}} \Big)^2 e^{-\frac{\tau^2 k^2}{2}} \\
= & \Big(\frac{1}{\sqrt{2\pi}} \sqrt{\frac{n}{k(n-k)}} \frac{n^n}{k^k (n-k)^{n-k}} e^{(1-144k^2-144n^2+144nk-12n)} \Big)^2 e^{-\frac{\tau^2 k^2}{2}} \\
\leq & \Big(\frac{1}{\sqrt{2\pi}} n^{n+\frac{1}{2}} k^{-k-\frac{1}{2}} (n-k)^{-(n-k)-\frac{1}{2}} \Big)^2 e^{-\frac{\tau^2 k^2}{2}}.
\end{aligned}$$

where the last inequality is obtained from the fact that $144nk - 144n^2 \leq 0$ as $n \geq k$; and $1 - 144k^2 - 12n \leq 0$ as $n \geq 1$, $1 \leq k \leq n$. So we have $e^{(1-144k-144n^2+144nk-12n)} \leq 1$ for any n and k positive.

We now define a new function named $\phi_{n,\tau}(s)$. And for $s \in (0,n)$, we have

$$\phi_{n,\tau}(s) = \frac{1}{\sqrt{2\pi}} n^{n+\frac{1}{2}} s^{-s-\frac{1}{2}} (n-s)^{-(n-s)-\frac{1}{2}} e^{-\frac{\tau^2 s^2}{4}}.$$

Then we have

$$\begin{aligned}
E\Gamma_k(n,\tau) \leq & \Big(\frac{1}{\sqrt{2\pi}} n^{n+\frac{1}{2}} k^{-k-\frac{1}{2}} (n-k)^{-(n-k)-\frac{1}{2}} \Big)^2 e^{-\frac{\tau^2 k^2}{2}} \\
= & \Big(\frac{1}{\sqrt{2\pi}} n^{n+\frac{1}{2}} k^{-k-\frac{1}{2}} (n-k)^{-(n-k)-\frac{1}{2}} e^{-\frac{\tau^2 k^2}{4}} \Big)^2 \\
= & 2\, \phi_{n,\tau}(k)^2. \qquad (2)
\end{aligned}$$

We consider the positive real root, $s(n,\tau)$, of the equation

$$\phi_{n,\tau}(s) = \frac{1}{\sqrt{2}}.$$

Now we introduce an lemma saying that the root of equation $\phi_{n,\tau}(s) = \frac{1}{\sqrt{2}}$ exists and is unique.

Lemma 1. *Let $\tau > 0$ be fixed. When n is sufficiently large, the equation $\phi_{n,\tau}(s) = 1$ has a unique positive real root $s(n,\tau)$, and*

$$s(n,\tau) = \frac{4}{\tau^2} \ln n - \frac{4}{\tau^2} \ln \Big(\frac{4}{\tau^2} \ln n \Big) + \frac{4}{\tau^2} + o(1)$$

where $o(1) \to 0$ as $n \to \infty$.

Proof. (*Proof of Lemma 1*)

Let $\tau > 0$ be fixed, and currently we have

$$\phi_{n,\tau}(s) = \frac{1}{\sqrt{2\pi}} n^{n+\frac{1}{2}} s^{-s-\frac{1}{2}} (n-s)^{-(n-s)-\frac{1}{2}} e^{-\frac{\tau^2 s^2}{4}}.$$

We multiply $\sqrt{2}$ from both sides and then taking logarithms to both sides of the equation, we have

$$\ln \big(\sqrt{2} \phi_{n,\tau}(s) \big) = -\frac{1}{2} \ln \pi + (n + \frac{1}{2}) \ln n - (s + \frac{1}{2}) \ln s - (n - s + \frac{1}{2}) \ln n - s - \frac{\tau^2 s^2}{4}.$$

Differentiating $\ln \big(\sqrt{2} \phi_{n,\tau}(s) \big)$ with respect to s yields

$$\frac{\partial \ln \big(\sqrt{2} \phi_{n,\tau}(s) \big)}{\partial s} = \frac{1}{2(n-s)} + \ln (n-s) = \frac{1}{2s} - \ln s - \frac{s\tau^2}{2}. \qquad (3)$$

From the expression of Equation (3), we have it is negative when $\frac{2 \ln n}{\tau^2} < s < \frac{4 \ln n}{\tau^2}$. Now, we calculate the value of $\ln \phi_{n,\tau}(s)$ for s outside such interval, and when $0 < s \leq \frac{2 \ln n}{\tau^2}$, we have

$$\ln \big(\sqrt{2} \phi_{n,\tau}(s) \big) \geq s \Big(\ln (n - \frac{2 \ln n}{\tau^2}) - \frac{s\tau^2}{4} - \ln \ln n - \ln \frac{2}{\tau^2} \Big) - \frac{\ln s}{2} - \frac{\ln \pi}{2},$$

which is positive when n is large enough.

We apply the expression of $\ln\big(\sqrt{2}\phi_{n,\tau}(s)\big)$ to find an upper bound

$$\begin{aligned}\ln\big(\sqrt{2}\phi_{n,\tau}(s)\big) &= -\frac{1}{2}\ln\pi + (n+\frac{1}{2})\ln n - (s+\frac{1}{2})\ln s - (n-s+\frac{1}{2})\ln n - s - \frac{\tau^2 s^2}{4} \\ &\le (n+\frac{1}{2})\ln n - (s+\frac{1}{2})\ln s - (n-s+\frac{1}{2})\ln n - s - \frac{\tau^2 s^2}{4} \\ &= s\Big(\ln(n-s) - \frac{s\tau^2}{4} - \ln s\Big) - \frac{\ln s}{2} + (n+\frac{1}{2})\ln\frac{n}{n-s}.\end{aligned}$$

Then, we have the right-hand side of the inequality is negative when $s > n-2$. We consider the cases $s+2 < n < \frac{s\ln s}{2\ln 2}$ and $n \ge \frac{s\ln s}{2\ln 2}$, and we can bound the $(n+\frac{1}{2})\ln\frac{n}{n-s}$ by $\frac{s\ln s}{2} + \frac{\ln 2}{2}$ and $2s + \frac{\ln 2}{2}$ respectively. Thus, we have for $s < n-2$

$$\ln\big(\sqrt{2}\phi_{n,\tau}(s)\big) \le s\Big(\ln(n-s) - \frac{s\tau^2}{4} - \ln s\Big) - \frac{\ln s}{2} + 2s + \frac{s\ln s}{2} + \frac{\ln 2}{2}.$$

Moreover, when $\frac{4\ln n}{\tau^2} \le s < n-2$, we have

$$\ln\big(\sqrt{2}\phi_{n,\tau}(s)\big) \le s\Big(2 - \frac{\ln s}{2}\Big) - \frac{\ln s}{2} + \frac{\ln 2}{2} < 0,$$

when n and s are large enough. Therefore, for large n, there exists a unique solution $s(n,\tau)$ of the equation $\sqrt{2}\phi_{n,\tau}(s) = 1$, i.e., $\phi_{n,\tau}(s) = \frac{1}{\sqrt{2}}$ with $\phi_{n,\tau}(s) \in (\frac{2\ln n}{\tau^2}, \frac{4\ln n}{\tau^2})$.

Taking the logarithms of both sides of the equation $\sqrt{2}\phi_{n,\tau}(s) = 1$ and rearranging terms yields

$$(\frac{1}{2}+n)\ln\frac{n}{n-s} - (s+\frac{1}{2})\ln s + s\ln(n-s) - \frac{\tau^2 s^2}{4} = \frac{\ln\pi}{2}.$$

Now we consider the case where s and $\frac{n}{s}$ tend to infinity with n. Dividing both sides of last expression by s, we have

$$\ln(n-s) - \frac{s\tau^2}{4} - \ln s = -1 + O(\frac{\ln s}{s}),$$

which equals to

$$\ln(n-s) - \ln n + \ln n - \frac{s\tau^2}{4} - \ln s + \ln\ln n - \ln\ln n = -1 + O(\frac{\ln s}{s}).$$

After simplifying the equation, we have

$$\ln n - \frac{s\tau^2}{4} - \ln\ln n = \ln\frac{s}{\ln n} - \ln\Big(\frac{n-s}{n}\Big) - 1 + O(\frac{\ln s}{s}) \tag{4}$$

For each $n \ge 1$, we define $R(n)$ via the equation

$$s(n,\tau) = \frac{4\ln n}{\tau^2} - \frac{4\ln\ln n}{\tau^2} + R(n)$$

Plugging the expression into Equation (4), we have

$$R(n) = \frac{4}{\tau^2}(1 - \ln\frac{4}{\tau^2}) + o(1).$$

Thus, we have

$$s(n,\tau) = \frac{4}{\tau^2}\ln n - \frac{4}{\tau^2}\ln\Big(\frac{4}{\tau^2}\ln n\Big) + \frac{4}{\tau^2} + o(1)$$

□

We have $s(n,\tau)$ exists and is unique, combining that if a value $k > s(n,\tau)$, then $\phi_{n,\tau}(k) < \frac{1}{\sqrt{2}}$; and combining the upper bound (2), we have deduced

$$E\Gamma_k(n,\tau) \le 2\,\phi_{n,\tau}(k)^2$$

We are able to conclude that

$$E\Gamma_k(n,\tau) \le 1$$

which means the expected number of $k \times k$ submatrices U of W_n with $F(U) \geq \tau$ is less than one.

Second, we start to deduce when n is sufficient large, the largest value k such that $n \times n$ Gaussian matrix W_n contains a $k \times k$ submatrix U with average greater or equal to τ for some fixed $\tau > 0$ has almost surely upper and lower bound. We derive the conclusion by the following 2 steps:

- **Step 1: Derive the upper bound**
  We introduce two necessary results (Sun and Nobel, 2013) to derive the upper bound:

  Proposition 1. *Let $\tau > 0$ be fixed. When n is sufficiently large, we have*

  $$\mathbb{P}(K_\tau(W_n) \geq s(n,\tau) + r) \leq 2e^{\frac{2}{\tau^2}} n^{-2r} \Big(\frac{4 \ln n}{\tau^2}\Big)^2$$

  for every $r = 1, 2, \ldots, n$.

  Notice Proposition 1 shows that the probability of seeing large-average submatrices is small.

  Lemma 2. *(Borel-Cantelli I) If $\sum_n \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(A_n \, infinitely \, often) = 0$.*

  An upper bound for $K_\tau(W_n)$ containing $s(n,\tau)$ can be obtained by applying Proposition 1 and Lemma 2 directly that

  $$K_\tau(W_n) \leq \lceil s(n,\tau) \rceil + 1 \leq s(n,\tau) + 2$$

  almost surely.

  Proof. (*Proof of upper bound*)

  Let $\tau > 0$ be fixed and for every $r = 1, 2, \ldots, n$, we denote a set A_r by

  $$A_r = \{K_\tau(W_n) \geq s(n,\tau) + r, K_\tau(W_n) \leq s(n,\tau) + (r+1)\}$$

  for n large enough.

  Then, we have
  $$\cup_{r=1}^{\infty} A_r = \{K_\tau(W_n) \geq s(n,\tau) + 1\}.$$

  Since A_r is almost disjoint for $r = 1, 2, \ldots, n$, we have when n approaches to infinity, by Proposition 1
  $$\sum_{r=1}^{\infty} \mathbb{P}(A_r) = \mathbb{P}(\cup_{r=1}^{\infty} A_r) = \lim_{n \to \infty} 2e^{\frac{2}{\tau^2}} n^{-2r} \Big(\frac{4 \ln n}{\tau^2}\Big)^2 = 0.$$

  Then we have $\mathbb{P}(A_r \, infinitely \, often) = 0$, and by Lemma 2, we have

  $$K_\tau(W_n) \leq \lceil s(n,\tau) \rceil + 1 \leq s(n,\tau) + 2$$

  almost surely. □

- **Step 2: Derive the lower bound**
  We first introduce a result derived by Sun and Nobel (2013) which we will use when derive the lower bound of $K_\tau(W_n)$:

  Lemma 3. *Let $\tau > 0$ be fixed. When k is sufficiently large, for every integer n satisfying the condition*
  $$k \leq \frac{4}{\tau^2} \ln n - \frac{4}{\tau^2} \ln \Big(\frac{4}{\tau^2} \ln n\Big) - \frac{12 \ln 2}{\tau^2} \tag{5}$$
  we have the bound
  $$\frac{Var \, \Gamma_k(n,\tau)}{(E \, \Gamma_k(n,\tau))^2} \leq k^{-2}$$

  Proof. (*Sketch proof of Lemma 3*)

  First we find an upper bound for $\frac{Var \, \Gamma_k(\tau,n)}{(E\Gamma_k(\tau,n))^2}$.

Lemma 4. *Fix $\tau > 0$. There exist integers $n_0, k_0 \geq 1$ and a positive constant C depending on τ but independent of k and n, such that for any $n \geq n_0$ and any $k \geq k_0$, we have*

$$\frac{Var\,\Gamma_k(\tau,n)}{(E\Gamma_k(\tau,n))^2} \leq Ck^4 \sum_{l=1}^{k}\sum_{r=1}^{k} \frac{\binom{k}{l}\binom{n-k}{k-l}}{\binom{n}{k}} \frac{\binom{k}{r}\binom{n-k}{k-r}}{\binom{n}{k}} \exp\{\frac{rl\tau^2}{2}(1+\frac{k^2-rl}{k^2+rl})\}$$

Proof. (*Sketch proof of Lemma 4*)

Note that we have

$$Var\,\Gamma_k(n,\tau) = E\Gamma_k^2(n,\tau) - (E\Gamma_k(n,\tau))^2$$

where we are able to derive that

$$E\Gamma_k(n,\tau) = \sum_{U\in S_k} \mathbb{P}(F(U) > \tau) = \binom{n}{k}^2 (1-\Phi(k\tau)).$$

For $1 \leq r, l \leq k$, we define

$$G(r,l) = \mathbb{P}(F(U) > \tau\, and\, F(V) > \tau)$$

where U and V are two fixed $k \times k$ submatrices of W having r rows and l solumns in commom, then we have

$$E\Gamma_k^2(n,\tau) = \sum_{r=0}^{k}\sum_{l=0}^{k} \binom{n}{k}^2 \binom{k}{r}\binom{n-k}{k-r}\binom{k}{l}\binom{n-k}{k-l} G(r,l)$$

Thus, we have an upper bound of $\frac{Var\,\Gamma_k(\tau,n)}{(E\Gamma_k(\tau,n))^2}$, which is

$$\frac{Var\,\Gamma_k(\tau,n)}{(E\Gamma_k(\tau,n))^2} \leq \sum_{r=0}^{k}\sum_{l=0}^{k} \frac{\binom{k}{l}\binom{n-k}{k-l}}{\binom{n}{k}} \frac{\binom{k}{r}\binom{n-k}{k-r}}{\binom{n}{k}} \Big[\frac{G(r,l)}{(1-\Phi(k\tau))^2} - 1\Big]$$

Then we can find an upper bound of $\frac{G(r,l)}{(1-\Phi(k\tau))^2} - 1$, i.e., there exists a positive constant C depends on τ but not on k and n such that

$$\frac{G(r,l)}{(1-\Phi(k\tau))^2} - 1 \leq Ck^4 \exp\{\frac{rl\tau^2}{2}(1+\frac{k^2-rl}{k^2+rl})\}$$

□

Then, after having derived Lemma 4, to prove Lemma 3, it suffices for us to show when n satisfies the condition inequality (5), we have

$$k^4 \sum_{l=1}^{k}\sum_{r=1}^{k} \frac{\binom{k}{l}\binom{n-k}{k-l}}{\binom{n}{k}} \frac{\binom{k}{r}\binom{n-k}{k-r}}{\binom{n}{k}} \exp\{\frac{rl\tau^2}{2}(1+\frac{k^2-rl}{k^2+rl})\} \leq k^{-2} \tag{6}$$

Therefore, to prove Inequality (6), we can show that each term in the sum is less than k^{-8}. We derive the following four inequalities, which are

- Inequality 1:

$$\frac{\binom{k}{l}\binom{n-k}{k-l}}{\binom{n}{k}} \leq \frac{\binom{k}{l} k^l (n-k)^{k-l}}{(n-k)^k} = \binom{k}{l} k^l (n-k)^{-l}$$

– Inequality 2:

$$\frac{\begin{pmatrix} k \\ l \end{pmatrix}\begin{pmatrix} n-k \\ k-l \end{pmatrix}}{\begin{pmatrix} n \\ k \end{pmatrix}}\frac{\begin{pmatrix} k \\ r \end{pmatrix}\begin{pmatrix} n-k \\ k-r \end{pmatrix}}{\begin{pmatrix} n \\ k \end{pmatrix}} \leq C\begin{pmatrix} k \\ r \end{pmatrix}\begin{pmatrix} k \\ l \end{pmatrix} k^{r+l} n^{-(r+l)}$$

– Inequality 3:

$$\begin{aligned} & n^{-(r+l)} \exp\{\frac{rl\tau^2}{2}(1+\frac{k^2-rl}{k^2+rl})\} \\ & \qquad \leq (\frac{4\ln n}{\tau^2})^{-(r+l)} \exp\{-3(r+l)\ln 2\} \exp\{\frac{\tau^2}{2}(\frac{2k^2rl}{k^2+rl}-\frac{k(r+l)}{2})\} \end{aligned}$$

– Inequality 4:

$$\begin{pmatrix} k \\ r \end{pmatrix}\begin{pmatrix} k \\ l \end{pmatrix} e^{-3(r+l)\ln 2} \exp\{\frac{\tau^2}{2}(\frac{2k^2rl}{k^2+rl}-\frac{k(r+l)}{2})\} \leq k^{-8}.$$

Finally, combining these four inequalities together, we have each term in the sum is less than k^{-8}, which completes the proof of Lemma 3.

□

By Lemma 3, with the help of Borel-Cantelli Lemma and Chebyshev's inequality, we obtain an almost sure lower bound of $K_\tau(W_n)$, which is

$$K_\tau(W_n) \geq s(n,\tau) - \frac{4}{\tau^2} - \frac{12\ln 2}{\tau^2} - 4.$$

Third, we derive the almost sure limit behavior of $K_\tau(W_n)$. By the previous two steps, we have our almost sure asymptotic upper and lower bound of the random variable $K_\tau(W_n)$ illustrated in the following theorem.

Theorem 1. *Let $W_n, n \geq 1$, be Gaussian random matrices derived from an infinite array W, and let $\tau > 0$ be fixed. When n is sufficiently large,*

$$s(n,\tau) - \frac{4}{\tau^2} - \frac{12\ln 2}{\tau^2} - 4 \leq K_\tau(W_n) \leq s(n,\tau) + 2$$

almost surely.

Now, we are able to derive the conclusion easily by sandwiching shown below.

Divide the conclusion in Theorem 1 by $\frac{4}{\tau^2}\ln n$, we obtain

$$\frac{s(n,\tau) - \frac{4}{\tau^2} - \frac{12\ln 2}{\tau^2} - 4}{\frac{4}{\tau^2}\ln n} \leq \frac{K_\tau(W_n)}{\frac{4}{\tau^2}\ln n} \leq \frac{s(n,\tau)+2}{\frac{4}{\tau^2}\ln n}$$

Substitute the expression of $s(n,\tau)$ that we derived in Lemma 1, the almost sure lower bound is

$$\begin{aligned} \textit{lower bound} &= \frac{\frac{4}{\tau^2}\ln n - \frac{4}{\tau^2}\ln\left(\frac{4}{\tau^2}\ln n\right) - \frac{12\ln 2}{\tau^2} - 4 + o(1)}{\frac{4}{\tau^2}\ln n} \\ &= 1 - \frac{\ln\left(\frac{4}{\tau^2}\ln n\right)}{\ln n} - \frac{3\ln 2}{\ln n} - \frac{\tau^2}{\ln n} + \frac{\tau^2 o(1)}{4\ln n} \\ &\to 1 \end{aligned}$$

as $n \to \infty$.

Similarly, the almost sure upper bound is

$$\begin{aligned}
\textit{upper bound} &= \frac{\frac{4}{\tau^2}\ln n - \frac{4}{\tau^2}\ln\left(\frac{4}{\tau^2}\ln n\right) + \frac{4}{\tau^2} + 2 + o(1)}{\frac{4}{\tau^2}\ln n} \\
&= 1 - \frac{\ln\left(\frac{4}{\tau^2}\ln n\right)}{\ln n} + \frac{1}{\ln n} + \frac{\tau^2}{2\ln n} + \frac{\tau^2 o(1)}{4\ln n} \\
&\to 1
\end{aligned}$$

when $n \to \infty$.

By sandwiching, we have

$$\frac{K_\tau(W_n)}{\frac{4}{\tau^2}\ln n} \to 1$$

almost surely as $n \to \infty$.

So when n is sufficient large, the largest value k such that $n \times n$ Gaussian matrix W_n contains a $k \times k$ submatrix U with average greater or equal to τ for some fixed $\tau > 0$ approaches to $\frac{4}{\tau^2}\ln n$ almost surely.

4 Numerical Evaluation

In this section, we introduce our original algorithm to find the largest size of square submatrices with average greater than a positive fixed number τ in a Gaussian random matrix. The algorithm is described in the first part of this section; and the numerical results is described in the second part of this section.

4.1 Algorithm

Algorithem 1. ***Large-Average Submatrix Threshold Searcher***

1. *Input an $n \times n$ dimensional Gaussian random matrix named W_n, and a fixed positive value named τ;*
2. *Rearrange columns of matrix W_n such that the column with largest sum is placed on the first column, till the column with smallest sum becomes the last column of the matrix, and do the same for the rows of matrix W_n. We denote the new matrix as W_n^{new};*
3. *Initialize $i = 0$, i denotes the current dimension of submatrix;*
4. *Start to check submatrix with dimension $i = i + 1$:*

   We evaluate whether the average of i dimensional principal minor of W_n^{new} is greater than τ:

   - *If True, go to Step 4;*
   - *If False, go to Step 5;*
5. *Find the $i \times i$ dimensional submatrix which contains the most largest i elements of W_n^{new}, and evaluate whether the average of such submatrix is greater than τ:*
   - *If True, go to Step 4;*
   - *If False, go to Step 6;*
6. *Enumerate all the $i \times i$ submatrices of W_n^{new}, and evaluate whether the average of each submatrix is greater than τ:*
   - *If one submatrix has average greater than τ, then stop enumerating and go to Step 4;*
   - *If no submatrix has average greater than τ, then output $i - 1$ to be the largest dimension of submatrices that have average greater than τ.*

4.2 Numerical Results

To examine the relation between largest size of submatrices with average greater than a certain fixed positive number in an Gaussian random matrix and the dimension of the Gaussian random matrix, we evaluate the largest size of the submatrices with average greater than fixed value τ being $0.5, 1, 1.5, 2, 2.5, 3$, when dimension n varies from 3 to 15. For each τ and each n, we calculate the threshold for 100 times and evaluate the mean of the 100 values of the thresholds. The results are presented in the following table:

Fixed value τ	Dimension and Corresponding Result						
Dimension	3	4	5	6	7	8	9
$\tau = 0.5$	1.5500	2.2400	2.9900	3.3500	3.9600	4.3200	4.9400
Dimension	10	11	12	13	14	15	
$\tau = 0.5$	5.2400	5.8100	6.1900	6.5200	7.0300	7.4500	
Dimension	3	4	5	6	7	8	9
$\tau = 1.0$	0.9000	1.4300	1.6300	1.9500	2.1200	2.4300	2.6400
Dimension	10	11	12	13	14	15	
$\tau = 1.0$	2.8200	3.1200	3.3000	3.5800	3.7300	3.8300	
Dimension	3	4	5	6	7	8	9
$\tau = 1.5$	0.4300	0.6700	1.0000	1.0000	1.2100	1.3300	1.5200
Dimension	10	11	12	13	14	15	
$\tau = 1.5$	1.6400	1.8400	1.9400	2.0400	2.0500	2.1800	
Dimension	3	4	5	6	7	8	9
$\tau = 2.0$	0.1400	0.3300	0.4900	0.6100	0.6900	0.8300	0.9300
Dimension	10	11	12	13	14	15	
$\tau = 2.0$	0.9300	1.0200	1.0500	1.1500	1.2100	1.2000	
Dimension	3	4	5	6	7	8	9
$\tau = 2.5$	0.0300	0.0900	0.1100	0.2600	0.2400	0.2900	0.4400
Dimension	10	11	12	13	14	15	
$\tau = 2.5$	0.4400	0.4600	0.6600	0.6400	0.7700	0.7300	
Dimension	3	4	5	6	7	8	9
$\tau = 3.0$	0.0300	0.0200	0.1000	0.0300	0.0700	0.1000	0.0900
Dimension	10	11	12	13	14	15	
$\tau = 3.0$	0.1100	0.1600	0.2200	0.1600	0.2100	0.3300	

Table 2: Results of $\tau = 0.5, 1, 1.5, 2, 2.5$ in Different Dimensions

5 Discussion

In this section, we integrate the theoretical result and numerical result and examine their consistency. We calculate the numerical result and theoretical result respectively when τ equals to $0.5, 1.0, 1.5, 2.0, 2.5, 3.0$ and n varies from 3 to 15 with the application of the algorithm presented in Section 4 part 4.1 and the theoretical we derived before in Section 3, that when n is sufficient large, the largest value k such that $n \times n$ Gaussian matrix W_n contains a $k \times k$ submatrix U with average greater or equal to τ for some fixed $\tau > 0$ approaches to $\frac{4}{\tau^2} \ln n$ almost surely. The results are shown in Figure 3 to Figure 8

As we may observed from the graphs, the theoretical results are observed much larger than the numerical results; and for each graph, when τ taking different values, the difference between theoretical results and numerical results become increasingly large. This phenomena may undermine our conclusion deduced in Section 3 that $\frac{K_\tau(W_n)}{\frac{4}{\tau^2}\ln n} \to 1$ almost surely; and we post our analysis to this inconsistency in the following two points:

- **Impact from The Mistake Made in the Original Article**
  In the original article written by Sun and Nobel (2013), the authors derived the upper bound for the expected value of $\Gamma_k(n, \tau)$, the number of $k \times k$ submatrices in W_n having an average

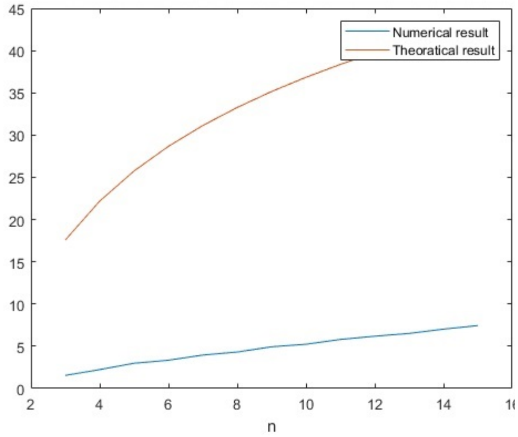

Figure 3: Results of $\tau = 0.5$

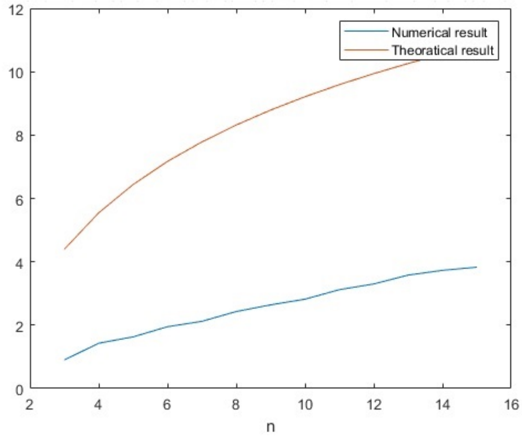

Figure 4: Results of $\tau = 1$

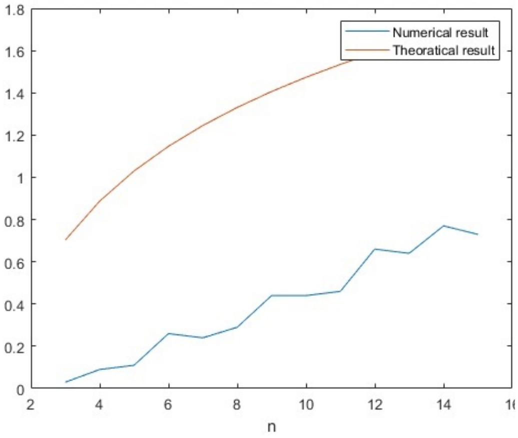

Figure 5: Results of $\tau = 1.5$

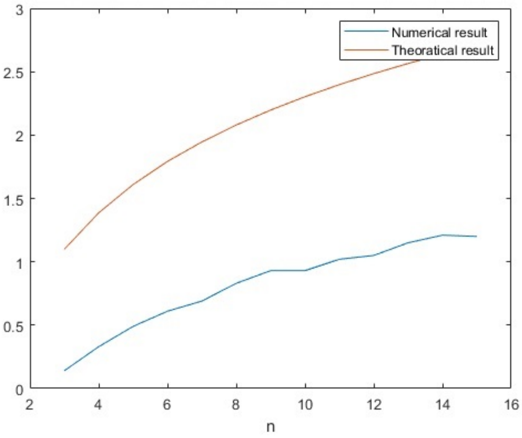

Figure 6: Results of $\tau = 2.0$

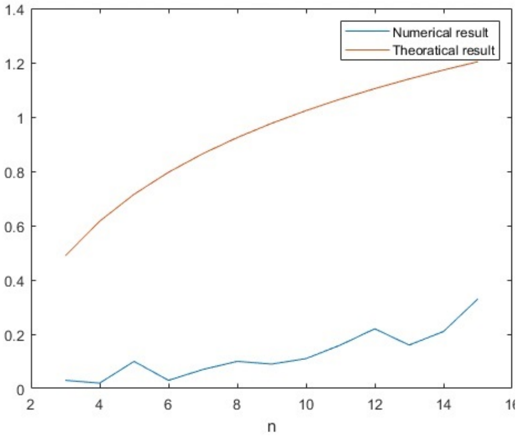

Figure 7: Results of $\tau = 2.5$

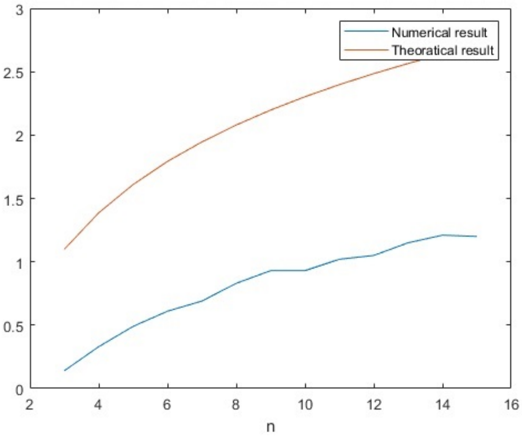

Figure 8: Results of $\tau = 3.0$

greater than or equal to τ, to be

$$E\Gamma_k(n,\tau) \leq 2\,\phi_{n,\tau}(k)^2$$

After the derivation of the upper bound for $E\Gamma_k(n,\tau)$, the authors present a lemma showing the existence and uniqueness of the positive real of the equation $\phi_{n,\tau}(s) = 1$.

The problem occurs when the authors conclude that "for values of k greater than $s(s,\tau)$, the expected number of $k \times k$ submatrices U of W_n with $F(U) \geq \tau$ is less than one". Actually, when k greater than $s(s,\tau)$, we have $\phi_{n,\tau}(s) \leq 1$; and hence we obtain

$$E\Gamma_k(n,\tau) \leq 2\,\phi_{n,\tau}(k)^2 \leq 2$$

Therefore, the correct conclusion is that for values of k greater than $s(s, \tau)$, the expected number of $k \times k$ submatrices U of W_n with $F(U) \geq \tau$ is less than two instead of one. In other words, we should find the positive real root of equation $\phi_{n,\tau}(s) = \frac{1}{\sqrt{2}}$ to ensure $E\Gamma_k(n, \tau) \leq 1$.

Although we have corrected the lemma in the original article with our refined proof (see Lemma 1), we still cited the Proposition 1 and Lemma 3 which are based on the wrong version of Lemma 1 in the original article. Therefore, we are still not clear whether the wrong version of Lemma 1 will affect our almost sure upper and lower bound, and furthermore, affect the limit behavior of our size threshold.

- **The Small Value of Dimensions in Numerical Approach**
  Due to the computational restriction, our algorithm is only efficient when the dimensions of Gaussian random matrix are relatively small, i.e., $n \leq 15$. Moreover, since the question of finding the largest threshold of the size of submatices with average greater than a certain positive fixed number is itself a NP-complete question, we are only able to use enumeration or enumeration-like algorithm such as our algorithm raised in Section 4, which is a slightly optimized enumeration algorithm. This kind of question is computation expensive and we will never obtain the results when n is large, especially when τ is also relative small. However, our conclusion from theoretical deduction says that when the dimension n approaching to infinity, the $K_\tau(W_n)$ shows the pattern approaching to $\frac{4}{\tau^2} \ln n$ almost surely. Our numerical result only evaluates $3 \leq n \leq 15$, which is far from infinity, and maybe when n is at such range, the value of threshold does not show any convergence.

6 Conclusion

In conclusion, we summarize our original work for this project in the first part; and we conclude our main results and propose some future work to this project in the second part.

6.1 Original Work

In this project, we analyze the maximal threshold of large-average submatices in a Gaussian random matrix mainly based on the article written by Sun and Nobel (2013) with three points of our original work listing in the following:

- We add more detail to some part of the proof and deduction when deriving the limit behavior theoretically so that the article is more reading-friendly. Moreover, we refine the proof of Lemma 1, which has a mistake that may undermine the theoretical analysis in the original article.
- We design the optimized enumeration algorithm "Large-Average Submatrix Threshold Searcher" to search the largest size of submatrices from Gaussian random matrix.
- We compare the numerical results by applying our original algorithm and theoretical results from the article written by Sun and Nobel (2013); and we present two explanations for the inconsistency of the results from different perspectives.

6.2 Conclusion and Future Work

In this report, based on the article "On the Maximal Size of Large-Average and ANOVA-Fit Submatirces in a Gaussian Random Matrix" by Sun and Nobel, we reformulate the model of finding the largest size of submatrices with average more than a fixed positive number in the Gaussian random matrix following the same process as the original article with more details to make it more smooth to read in terms of those without a very solid mathematical background. Moreover, we also propose our self-designed optimized enumeration algorithm to search the largest thresholds. We compare our numerical results and the theoretical results presented on the original article and find appearance of inconsistency of two different approaches; and we propose our own analysis for the inconsistency from both the theoretical perspective that the typo made by authors of original article has a relatively huge impact on the process of derivation of the later parts and eventually affect the conclusion, and the numerical perspective that the computational restriction limits the dimension of Gaussian random matrix we are able to examine, resulting in the dimensions not large enough to reach the critical value of convergence.

In the future, due to the nature of a NP-complete question, we are not able to improve our algorithm saving more time and hence examine a larger value of the dimensions of the submatrices in the Gaussian random matrix. However, we are able to look into the proof of Proposition 1 and Lemma 3 written by Sun and Nobel and try to refine the proof such that we are able to derive the accurate expression for the almost sure upper and lower bound, and refine the part of deriving the limit behavior of $K_\tau(W_n)$.

References

Aleksandra Adomas, Gregory Heller, Åke Olson, Jason Osborne, Magnus Karlsson, Jarmila Nahalkova, Len Van Zyl, Ron Sederoff, Jan Stenlid, Roger Finlay, and Frederick O. Asiegbu. Comparative analysis of transcript abundance in Pinus sylvestris after challenge with a saprotrophic, pathogenic or mutualistic fungus. *Tree Physiology*, 28(6):885–897, 06 2008. ISSN 0829-318X. doi: 10.1093/treephys/28.6.885. URL `https://doi.org/10.1093/treephys/28.6.885`.

Milind Dawande, Pinar Keskinocak, Jayashankar M Swaminathan, and Sridhar Tayur. On bipartite and multipartite clique problems. *Journal of Algorithms*, 41(2):388 – 403, 2001. ISSN 0196-6774. doi: https://doi.org/10.1006/jagm.2001.1199. URL `http://www.sciencedirect.com/science/article/pii/S019667740191199X`.

Joseph Hacia, J.B. Fan, Oliver Ryder, Li Jin, Keith Edgemon, Ghassan Ghandour, R. Mayer, Bryan Sun, Linda Hsie, Christiane Robbins, Lawrence Brody, David Wang, Eric Lander, Robert Lipshutz, Stephen Fodor, and Francis Collins. Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. *Nature genetics*, 22: 164–7, 07 1999. doi: 10.1038/9674.

A. J. Maria. A remark on stirling's formula. *The American Mathematical Monthly*, 72(10):1096–1098, 1965. ISSN 00029890, 19300972. URL `http://www.jstor.org/stable/2315957`.

Gary Moran, Cheryl Stokes, Sascha Thewes, Bernhard Hube, David C. Coleman, and Derek Sullivan. Comparative genomics using candida albicans dna microarrays reveals absence and divergence of virulence-associated genes in candida dubliniensis. *Microbiology*, 150 (10):3363–3382, 2004. ISSN 1350-0872. doi: https://doi.org/10.1099/mic.0.27221-0. URL `https://www.microbiologyresearch.org/content/journal/micro/10.1099/mic.0.27221-0`.

Jonathan Pollack, Charles Perou, Ash Alizadeh, Michael Eisen, Alexander Pergamenschikov, Cheryl Williams, Stefanie Jeffrey, David Botstein, and Patrick Brown. Genome-wide anaiysis of dna copy-number changes using cdna microarrays. *Nature genetics*, 23, 01 1999.

Xing Sun and Andrew Nobel. On the maximal size of large-average and anova-fit submatrices in a gaussian random matrix. *Bernoulli : official journal of the Bernoulli Society for Mathematical Statistics and Probability*, 19:275–294, 02 2013. doi: 10.3150/11-BEJ394.

Appendix

matRearrange

```
function y = matRearrange(mat)
% This function input a matrix and return a new matrix with
% rearrange the order of the row sum and column sum from
% maximum to minimum.

% Calculate the row sum and column sum
colsum = sum(mat);
rowsum = sum(mat');

% Rearrange the matrix from largest to smallest
[~, colsort]=sort(-colsum);
mat = mat(:,colsort);
[~,rowsort]=sort(-rowsum);
y = mat(rowsort,:);
end
```

evaluateLargestMat

```
function y = evaluateLargestMat(mat,dim)
% This function returns a submatrix which contains the largest
% (dim) number of matrix (mat)

% Unique sorted values
sortedValues = unique(mat(:));
% Get the dim largest values
maxValues = sortedValues(end-dim+1:end);
% Get a logical index of all values
maxIndex = ismember(mat,maxValues);
% Locate these values
[m,n] = find(maxIndex==1);
% Find the maximum matrix
diffLength = 0;
diffLength_previous = 0;

% If some of the largest elements are in the same row or
% column, then we are able to find an extra largest number
% of element and hopefully this submatrix is the largest
% submatrix we can find without using enumeration.
while length(unique(m))~=dim || length(unique(n))~=dim

    % If some of the largest (dim) elements share the same column
    if  length(unique(m)) > length(unique(n))
        sortedValues = unique(mat(unique(m),:));
        diffLength_previous =diffLength;
        diffLength = diffLength + length(unique(m))-length(unique(n));
        maxValues = sortedValues(end-(dim+diffLength)+1:end-(dim+diffLength_previous));
        maxIndex = ismember(mat,maxValues);
        [temp_m,temp_n] = find(maxIndex==1);
        m = [m;temp_m];
        n = [n;temp_n];

        % If some of the largest (dim) elements share the same row
    elseif length(unique(m)) < length(unique(n))
        sortedValues = unique(mat(:,unique(n)));
        diffLength_previous =diffLength;
        diffLength = diffLength + length(unique(n))-length(unique(m));
        maxValues = sortedValues(end-(dim+diffLength)+1:end-(dim+diffLength_previous));
        maxIndex = ismember(mat,maxValues);
        [temp_m,temp_n] = find(maxIndex==1);
        m = [m;temp_m];
        n = [n;temp_n];

        % If the largest (dim) elements are in a small submatrix of the
        % submatrix
    elseif length(unique(m)) == length(unique(n))
        mat_temp=mat;
        mat_temp(m,n) = -100;
        sortedValues = unique(mat_temp(:));
        maxValues = sortedValues(end-(dim-length(unique(m)))+1:end);
        maxIndex = ismember(mat_temp,maxValues);
        [temp_m,temp_n] = find(maxIndex==1);
        m = [m;temp_m];
        n = [n;temp_n];
        diffLength = 0;
        diffLength_previous = 0;
    end
end
y = mat(unique(m),unique(n));
end
```

model

```
function result = model(mat,tau)
% This function input the matrix with mat and average value
% tau and returns the value of largest size of matrix with
% average greater than tau.

fprintf("Start to search the largest submatrix with average greater than tau \n")
dim = 1:length(mat(:,1));
i = 1;
succ = 0;
mat = matRearrange(mat);

% Check whether size i submatrix is the largest size of
% submatrix with average greater than tau.
while i <= length(dim) && succ == 0
    % Apply the calculation-friendly way to check whether some of the
    % average of current dimension matrix is greater than tau
    fprintf('Checking the submatrix with size %24.16f; \n',i)
    method1_result = sum(sum(mat(1:i,1:i)));
    method2_result = sum(sum(evaluateLargestMat(mat,i)));
    if method1_result > (tau*i^2)
        succ_sub = 1;
        result = i;
        fprintf('Method 1 (size %24.16f submatrix) succeess; \n',i)
    elseif method2_result > (tau*i^2)
        fprintf('Method 2 (size %24.16f submatrix) success; \n',i)
        succ_sub = 1;
        result = i;
    else
        succ_sub = 0;
        fprintf('Method 1 and 2 (size %24.16f submatrix) fail; \n',i)
        fprintf('Start enumeration')
    end
    % If either of the method works, then we will go directly to i = i + 1;
    % if not, we have to use enumeration in below
    % Find all the combination of row and column
    x = nchoosek(1:length(dim),i);
    y = x;
    % Initialization
    temp_mat = eye(i);
    j = 1;
    while j <= length(x(:,1)) && succ_sub == 0
        fprintf(’Enumerate size %24.16f submatrix %24.16f/%24.16f time; \n’,i,j,length(x(:,1)))
        k = 1;
        while k <= length(y(:,1)) && succ_sub == 0
            temp_mat = mat(x(j,:),y(k,:));
            % Check whether the current matrix has sum
            % greater than tau
            if sum(sum(temp_mat)) > (tau*i^2)
                % If it does, stop checking the size i
                % by setting succ_sub = 1; and continue
                % to check size i + 1 by setting succ = 0
                succ_sub = 1;
                succ = 0;
                % The largest size of submatrix with
                % average value greater than tau is i
                result = i;
            else
                % If not, check next size i submatrix by
                % setting succ_sub = 0; and if we check
                % all the size i submatrix and no one is
                % greater than tau, then stop by setting
                % succ = 1
                k = k + 1;
                succ_sub = 0;
                succ = 1;
                % The largest size of submatrix with
                % average value greater than tau is i - 1
                result = i-1;
            end
        end
        j = j + 1;
    end
    i = i + 1;
end
end
```

outputResult

```
function result = outputResult(tau,maxDimension,iterationTimes)
% This function input the fixed value tau, maximum dimension you want to
% check, and number of iterations for one dimension;
% The function returns a matrix with the first row being number of
% dimension, the second row being the theoretical value of corresponding
% dimension, and the third row being the real value of the corresponding
% dimension.
if maxDimension <= 3
    fprintf(’Please input a number with maximum size of the Gaussian matrix greater than 3’)
    result = 0;
else
   result_temp = zeros(iterationTimes,maxDimension-2);
   for i = 3:maxDimension
       for j = 1:iterationTimes
           result_temp(j,i-2)=model(randn(i),tau);
       end
   end
   n = 3:maxDimension;
   % Calculate the theoretical value and the real value
```

```
    meanLargestSize = mean(result_temp);
    theoreticalValue = 4/(tauˆ2)*log(n);
    result = [n;theoreticalValue;meanLargestSize];
    plot(n,meanLargestSize);
    hold on;
    plot(n,theoreticalValue);
    legend(’Numerical result’,’Theoratical result’)
    xlabel(’n’)
    title(’Plot of numerical and theoratical result of the maximum size of submatrix’)
    hold off;
end
```

Published with MATLAB® R2019b