UNIVERSITY OF CALIFORNIA, SAN DIEGO

Detection and Localization of a Submatrix: Theory, Methods and Algorithms

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Mathematics (with a specialization in Statistics)

by

Yuchao Liu

Committee in charge:

Professor Ery Arias-Castro, Chair
Professor Jelena Bradic
Professor Kamalika Chaudhuri
Professor Fan Chung Graham
Professor Young-Han Kim

2018

The dissertation of Yuchao Liu is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2018

DEDICATION

To my family, who have my back all the time.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

First, I would like to express my sincere thanks for my advisor, Professor Arias-Castro, for his support and guidance throughout my Ph.D. study at UC San Diego. Professor Arias-Castro introduced me to this research field. His sparking wisdom and deep insight into the problem helped me tame the seemingly impossible research projects. And I can not owe him more for the constant and endless patience and encouragement. Without him, the probability of the existence of this thesis is simply zero.

I would like to thank Professor Bradic and Professor Chaudhuri for offering me excellent courses as well as guidance on research and professional development. Their courses provide me solid foundations of theoretical statistics and machine learning. Also, I would like to thank Professor Chung and Professor Kim for serving as my thesis committee members. Professor Eggers and Professor Schweinsberg, as well as all my students I had during my TA career, have sharpened my teaching skills and made me a better statistics educator. Outside UCSD, I would like to thank Dr. Xinqun Yang and Dr. Yuqing Yang from BioMarin Pharmaceutical, and Dr. Emil Laftchiev from Mitsubishi Electric Research Labs, for hosting me during my summer internships in their respective departments.

I am immensely grateful to all my friends at UCSD Math department. Shout out to Kuang, Stephan, Michelle, Pieter, Emma, for those trips, Wednesday lunches, and Super Bowls; to Alex, Jiaqi, Jingwen, Andrew, Selene, Shiyun, Liyu, for sharing food, research ideas, and gaming time. There have been a lot of things supporting me during depression and despair. Thank Lovelive project, Ayakura Mei and her group, Boston Celtics Team (R.I.P. SD Chargers), and every stunning scenery and gourmet chef I met in my several trips to Japan.

Last but not least, my sincere thanks go to my whole family. My parents brought me to the world, raised me up, and they are the role models teaching me to learn, be independent, be strong, and pursue the better. My grandparents have taught me to be a

person with love and kindness. Finally, my wife Chenchen has brought the most sunshine and laughter into my life. Everything I achieved was tiny compared to her boundless love and sacrifice.

Chapter 2, partially, is a version of the paper "Distribution-free Detection of a Submatrix", *Journal of Multivariate Analysis* 156 (2017): 29-38. The dissertation author is the corresponding author of this material.

Chapter 2, partially, is a version of the paper "Distribution-Free, Size Adaptive Submatrix Detection with Acceleration". The dissertation author is the principal investigator and corresponding author of this material. The manuscript is being prepared to be submitted to a major statistics journal.

Chapter 3, in full, is a version of the paper "A Multiscale Scan Statistic for Adaptive Submatrix Localization". The dissertation author is the principal investigator and corresponding author of this material. The manuscript is being prepared to be submitted to a major statistics journal.

VITA

2011	B. S. in Statistics, Fudan University, Shanghai China
2011-2013	Ph. D. Student in Probability and Mathematical Statistics, Fudan University, Shanghai China
2013-2018	Graduate Teaching Assistant, University of California, San Diego
2015	C. Phil. in Mathematics, University of California, San Diego
2018	Ph. D. in Mathematics, University of California, San Diego

PUBLICATIONS

Arias-Castro, Ery, and Liu, Yuchao. "Distribution-free Detection of a Submatrix", *Journal of Multivariate Analysis* 156 (2017): 29-38.

Liu, Yuchao, and Guo, Jiaqi. "Distribution-Free, Size Adaptive Submatrix Detection with Acceleration", *Manuscript in Preparation*, 2018.

Liu, Yuchao, and Arias-Castro, Ery. "A Multiscale Scan Statistic for Adaptive Submatrix Localization", *Manuscript in Preparation*, 2018.

Laftchiev, Emil, and Liu, Yuchao. "Finding Multidimensional Patterns in Multidimensional Time Series", *Submitted*, 2018.

ABSTRACT OF THE DISSERTATION

Detection and Localization of a Submatrix: Theory, Methods and Algorithms

by

Yuchao Liu

Doctor of Philosophy in Mathematics (with a specialization in Statistics)

University of California, San Diego, 2018

Professor Ery Arias-Castro, Chair

We consider the problem of detecting and localizing an submatrix with larger-than-usual entries inside a large, noisy matrix. This problem arises from analysis of data in genetics, bioinformatics, and social sciences. We consider that entries of the data matrix are independently following distributions from a natural exponential family, which generalizes the common Gaussian assumptions in the literature. In Chapter 2 a permutation test for testing the existence of the elevated submatrix is studied. The test's asymptotic power is illustrated, and its robust variation (rank method) is also studied. In The latter part of Chapter 2 and Chapter 3 we remove the prior knowledge of the submatrix size, aiming to develop adaptive methods for detection and localization. Latter part of Chapter 2

proposes a Bonferroni testing framework based on the permutation scan test, to solve the detection problem. An accelerating framework is also developed without sacrificing asymptotic power. In Chapter 3, a new size-adaptive estimator is proposed to solve the localization problem. Its asymptotic performance is studied, and two fast algorithms to approximate the estimator are developed.

Chapter 1

Introduction

This chapter introduces the background of the submatrix detection and localization problems as well as their applications in different areas. Following that we describe the problem setup, and build a statistical model to transfer the problem to a statistical hypothesis testing / parameter estimation problem. A natural exponential family of distributions is also introduced, as an important parameterization tool in the evaluation of the methods proposed in the thesis. Finally, the thesis structure is described in the last section.

1.1 An Example

Matrix type data appears in different research fields, such as genetics, social sciences, and many other fields such as ecology and bioinformatics. In many cases, the row indexes and column indexes are individuals or units that can interact with each other, and the entry values inside the data matrix represents the strength of the interaction. We use the data from a study in biomedical sciences [STQ08] as an example. In Figure 1.1(a), the row indexes represent genes, and the column indexes represent different conditions. Each element inside the data matrix, denoted a_{ij}, indicates the transcript abundance of gene i

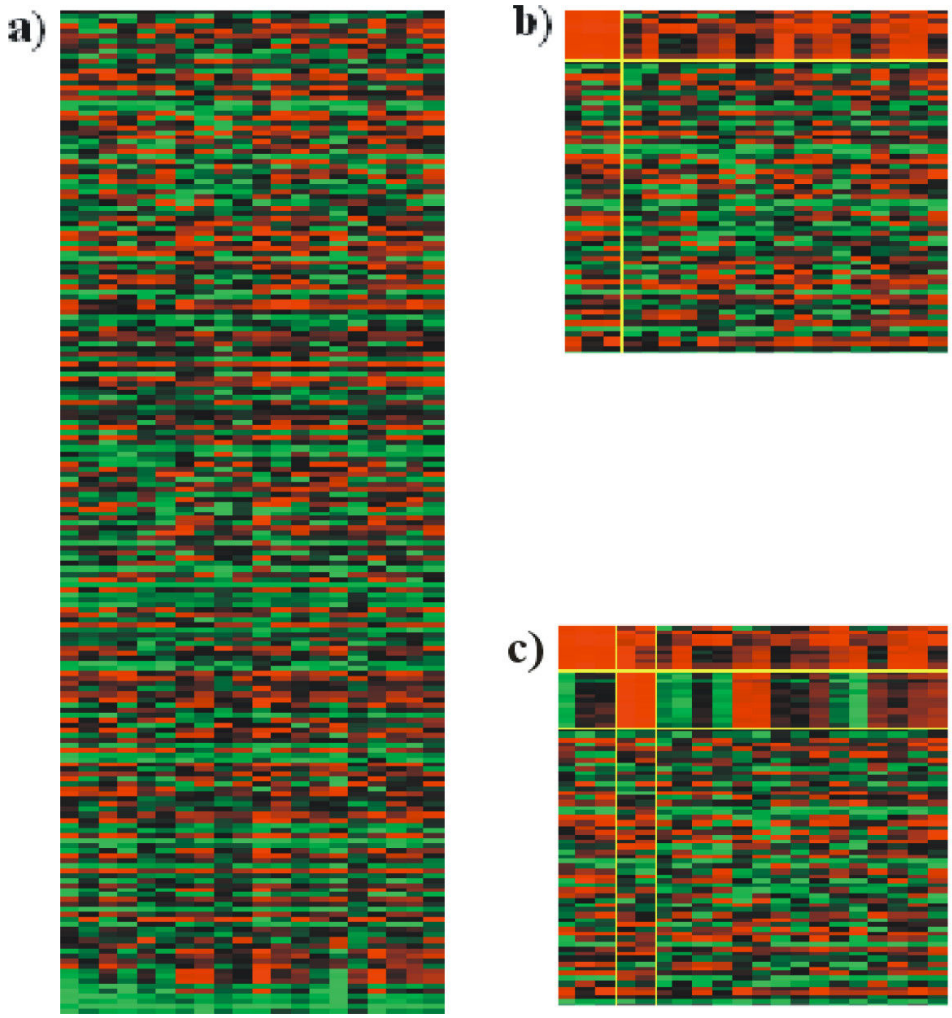

Figure 1.1: [STQ08] Transcript abundance heatmap

under condition j. This type of data appears in the applications of microarray technology - a technique exploring the relationship of gene expressions and different situations. See [SD01] for a general description of microarray technology.

Usually researchers are interested in finding useful information from the data in order to make inferences or help making scientific discovery. As shown in Figure 1.1(b), if the rows and columns are permuted in a proper way, a submatrix containing comparatively high values will appear in the upper left corner. This indicates that the genes, marked by the submatrix's rows, have relatively high transcript abundance under some conditions,

which are also marked by the submatrix's columns. Figure 1.1(c) shows the finding of another such submatrix, which is not overlapping with the one found in Figure 1.1(b).

Therefore, a submatrix containing larger-than-usual data in the data matrix are usually of special interest to people wanting to make use of the data. The process of finding such submatrices is usually referred as bi-clustering, or co-clustering. See [CC00] for analysis of gene expression data and [MO04, TSS05, KKZ09, CA11] for a survey and [PBZ+06] for a comparison of the methods. For the applications in social science, we refer the reader to [For10]. In this thesis, we mainly solve the following two problems:

- Does the data matrix contain a submatrix, such that entries inside the submatrix is larger-than-usual? (Detection problem)
- Given the data matrix, how to find a submatrix containing larger-than-usual entries? (Localization problem)

For each problem we propose methods, analyze the methods under specific parameterized statistical models, and give corresponding algorithms.

1.2 Problem Setup

We now restate the problem in a general setup. Denote $\mathbf{X}$ as the data matrix observed, and (M, N) as the number of rows and columns of $\mathbf{X}$. The detection problem in the previous section can be formalized into the a hypothesis testing problem, distinguishing the situations when the data matrix $\mathbf{X}$ contains only IID entries, against the situation where there are submatrices containing larger-than-usual entries. Usually the former situation is referred as the null hypothesis, or zero hypothesis, and the latter one as alternative hypothesis.

A hypothesis testing problem requires us to find a proper measurable function ϕ that maps from the sample space to $[0, 1]$. If $\mathbf{X}$ is observed, one reject H_0 with probability

$\phi(\mathbf{X})$. In other words, one would like to have $\phi(\mathbf{X})$ close to 1 when $\mathbf{X}$ is generated under H_1. In practice, p-value is usually used to perform the hypothesis testing. The p-value $\mathfrak{P}(\mathbf{X})$ is valued in $[0,1]$, and satisfies

$$\mathbb{P}(\mathfrak{P}(\mathbf{X}) \leq \alpha) \leq \alpha \tag{1.1}$$

with H_0 true and $\alpha \in [0,1]$. The function ϕ based on the p-values take the form $\phi(\mathbf{X}) = \mathbb{1}(\mathfrak{P}(\mathbf{X}) \leq \alpha_0)$, with some predetermined $\alpha_0 \in [0,1]$ controlling type-I error rate.

For the localization problem, the main task is to output the row and column indexes of the submatrix containing larger-than-usual values. Denote $[M] = \{1, \ldots, M\}$. We would like to find a function Φ mapping from the sample space to the collection of subsets of $[M] \times [N]$, such that the submatrix indexed by $\Phi(\mathbf{X})$ contains larger-than-usual values. We need the function to be accurate, that is, when there is an elevated submatrix, with high probability the function will find its row and column indexes. And the function should enjoy computational tractability. This means when a data matrix is inputted, the time and computational power consumed in calculating Φ, or approximating Φ, should be reasonable.

1.2.1 Related Works

There is an active research line focusing on different aspects of submatrix detection and localization. [ACL17, BI13] concentrate on detection of the existence of such an elevated submatrix, and asymptotic behaviors of tests. [BIS15, KBRS11] displayed the minimal signal strength needed for the existence of successful submatrix localizer without the introduction of sparsity and proved some finite sample property. [HWX15a] also considered minimal signal strength issues under symmetric data structure, and [HWX15b] developed upper and lower signal bounds for weak consistent estimators (w.h.p. estimators with error rate converging to 0). A convex optimization framework for biclustering is proposed

and associated algorithms are developed in [CAB17]. A selective inference framework for quantifying the information contained within a selected submatrix is proposed by [LST15].

The calculation issue is attracting increasing attentions from the researchers in the field, and some papers are addressing with the tradeoff between statistical power and computational tractability [MW15, CX16, CLR+17]. In the meantime, theoretically computationally tractable (a.k.a. running in polynomial time with respect to sample size) algorithms are developed. See [KBRS11] for a survey. Other computationally tractable methods are analyzed such as semidefinite relaxation [CX16], spectral method [CLR+17], message passing [HWX15b], and largest marginal gaps [BC16]. Iterative algorithm with Lasso type optimization target is developed in [TW14].

Another active research line worth mentioning here is on the stochastic block model (SBM). In the setup of SBM the observation is a graph, with edges independently connected. See [HLL83] for a detailed introduction. The detection problem is to distinguish between an Erdős-Rényi graph and a graph with group of nodes usually closely connected. The localization problem is to cluster the nodes by the closeness of their connection.

If the adjacency matrix of the graph is considered, the problem shares many properties with submatrix detection and localization (the adjacency matrix is symmetric with independent upper triangle nodes, compared with submatrix localization problem). There are works considering the existence of consistency detectors [ZZ+16], the existence of consistency clustering methods [MNS15], semi-definite programming [CX16, ABH16], and spectral methods [CCT12, McS01].

1.3 Natural Exponential Family

We introduce a natural exponential family of distributions, that is useful in the construction of statistical model in the thesis. We consider a one parameter exponential

family to build up the parametric framework, as studied in papers such as [BI13, ACTW17, ACL17]. We define such a distribution family by selecting a distribution ν with mean zero and variance 1. Denote $\varphi(\theta)$ as the moment generating function of ν, say $\varphi(\theta) = \int e^{\theta t}\nu(dt)$, and furthermore we require $\varphi(\theta) < \infty$ for all θ in $[0, \theta_\star)$. Here $\theta_\star$ is defined as $\sup\{\theta : \varphi(\theta) < \infty\}$ and could be equal to infinity. The parametrization is finished by defining $f_\theta, \theta \in [0, \theta_\star)$ as the density indexed with θ with respect to ν as

$$f_\theta(x) = \exp\{\theta x - \log \varphi(\theta)\}. \tag{1.2}$$

By selecting different ν, the distribution family covers Normal family ($\nu = \mathcal{N}(0,1)$), Poisson family ($\nu = \mathcal{P}(1) - 1$), and Rademacher family ($\nu = \text{Rade}(0.5)$), etc. Note that with ν fixed, the distribution family is stochastically monotone in θ ([LR05], Lemma 3.4.2), which is to say, for $X_1 \sim f_{\theta_1}$ and $X_2 \sim f_{\theta_2}$ with $0 \le \theta_2 \le \theta_1 \le \theta_\star$ and fixed ν,

$$\mathbb{P}_0(X_1 \ge t) \ge \mathbb{P}(X_2 \ge t), \forall t \in \mathbb{R}. \tag{1.3}$$

This fact enables us to model the submatrix localization problem with θ controlling the signal strength. If $X \sim f_\theta$, with θ increasing, X is more 'elevated', or is more larger than usual, if we denote ν or f_0 as the distribution of usual entries.

With this parametrization tool we may model the microarray data, which usually is described as Gaussian, and integer data such as $0-1$ data (directed graph or presence-absence matrices in Ecology. See [Got00] for examples).

1.4 Thesis Structure

The rest of the thesis will go as follows. In Chapter 2 we will introduce a permutation test framework dealing with the detection problem. The asymptotic power of the test

under the parametric framework based on the natural exponential family will be analyzed, and is compared with some state-of-the-art parametric methods. A robust variation of the permutation test using ranks is also analyzed and the power loss is exactly illustrated. Then, the testing procedure is used to construct a Bonferroni framework dealing with the case when the submatrix's size is unknown. The Bonferroni style permutation test's asymptotic power is shown. An accelerated version of the Bonferroni type permutation test is also introduced, and we show that the acceleration will not sacrifice the first order asymptotic power of the test.

In Chapter 3, an estimator of the submatrix indexes is introduced to tackle the localization problem, with submatrix size unknown. We illustrate the conditions under which the estimator is accurate with high probability, and two associated algorithms are introduced to approximate the estimator.

Chapter 2

Detection of a Submatrix via Permutation Test

2.1 Introduction

In this chapter we introduce a permutation test framework to discover submatrix type anomalies lying inside large data matrix. The problem is inspired by the work from bi-clustering, which has emerged as an important set of tools in bioinformatics, in particular, in the analysis of gene expression data [CC00]. It comes in different forms, and in fact the various methods proposed under that umbrella may target different goals. Here we follow [SWPN09], where the problem is posed as that of discovering a submatrix of unusually large values in a (large) data matrix. For example, in the context of a microarray dataset, the data matrix is organized by genes (rows) and samples (columns). As described before, we let $\mathbf{X} = (X_{ij})$ denote the matrix, M denote the number of rows and N denote the number of columns, so the data matrix $\mathbf{X}$ is M-by-N.

2.1.1 Submatrix detection

We consider the detection problem in its simplest form, that of merely detecting of the presence of an anomalous (or unusual) submatrix, which leads to a hypothesis testing problem. This was considered by Butucea and Ingster [BI13] from a minimax perspective. Their work relies on parametric assumptions. For example, in the normal model, they assume that the X_{ij}'s are independent and normal, with mean θ_{ij} and unit variance. Under the null hypothesis $\theta_{ij} = 0$ for all $i \in [M] := \{1, \dots, M\}$ and all $j \in [N]$. Under the (composite) alternative there is a m-by-n submatrix indexed by $\mathcal{I}_{\text{true}} \subset [M]$ and $\mathcal{J}_{\text{true}} \subset [N]$ such that

$$\theta_{ij} \geq \theta_{\ddagger}, \quad \forall (i,j) \in \mathcal{I}_{\text{true}} \times \mathcal{J}_{\text{true}}, \tag{2.1}$$

while $\theta_{ij} = 0$ otherwise. Here $\theta_{\ddagger} > 0$ controls the signal-to-noise ratio. In that paper, Butucea and Ingster precisely establish how large $\theta_{\ddagger}$ needs to be as a function of (M, N, m, n) in order for there to exist a procedure that has (worst-case) risk tending to zero in the large-sample limit (i.e., as the size of the matrix grows). They consider two tests which together are shown to be minimax optimal. One is the 'sum test' based on

$$\text{SUM}(\mathbf{X}) = \sum_{i \in [M]} \sum_{j \in [N]} X_{ij}. \tag{2.2}$$

It is most useful when the submatrix is large relative to the data matrix. The other one is the 'scan test' which, when the submatrix size is known (meaning m and n are known), is based on

$$\text{SCAN}(\mathbf{X}) = \max_{\mathcal{I} \subset [M], |\mathcal{I}| = m} \quad \max_{\mathcal{J} \subset [N], |\mathcal{J}| = n} \quad \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} X_{ij}. \tag{2.3}$$

To avoid making parametric assumptions, some works such as [BNW05, HTE+00] have suggested a calibration by permutation. We consider two somewhat stylized permutation approaches:

- *Unidimensional permutation.* The entries are permuted within their row. (One could permute within columns, which is the same after transposition.)
- *Bidimensional permutation.* The matrix is vectorized, the entries are permuted uniformly at random as one would in a vector, and the vector is reshaped into a matrix of same dimensions.

The first method is most relevant when one is not willing to assume that the entries in different rows are comparable. It is appealing in the context of microarray data and was suggested, for example, in [HTE+00]. The second method is most relevant in a setting where all the variables are comparable. In the parlance of hypothesis testing, the first method derives from a model where the entries within each row are exchangeable under the null, while the second method arises when assuming that all the entries are exchangeable under the null.

Contribution 1 (Calibration by permutation)**.** We analyze the performance of the scan test when calibrated using one of these two permutation approaches. We show that, regardless of the variant, the resulting test is (first order) asymptotically as powerful as a calibration by Monte Carlo with full knowledge of the parametric model. We prove this under some standard parametric models.

Remark 1. We focus on the scan statistic (2.3) and abandon the sum statistic (2.2) for at least two reasons: 1) the sum statistic cannot be calibrated without knowledge of the null distribution; 2) the sum statistic is able to surpass the scan statistic when it is impossible to locate the submatrix with any reasonable accuracy, which is somewhat less interesting to the practitioner.

A calibration by permutation is computationally intensive in that it requires the repeated computation of the test statistic on permuted data. In practice, several hundred

permutations are used, which can cause the method to be rather time-consuming. A possible way to avoid this is to use ranks, which was traditionally important before the availability of computers with enough computational power. [Het84] is a classical reference. In line with the two permutation methods described above, we consider the corresponding methods for ranking the entries:

- *Unidimensional ranks.* The entries are ranked relative to the other entries in their row.
- *Bidimensional ranks.* The entries are ranked relative to the all other entries.

The use of ranks has the benefit of only requiring calibration (typically done on a computer nowadays) once for each matrix size $M \times N$. It has the added benefit of yielding a method that is much more robust to outliers.

Contribution 2 (Rank-based method)**.** We analyze the performance of the scan test when the entries are replaced by their ranks following one of the two methods just described. We show that, regardless of the variant, there is a mild loss of asymptotic power, which we precisely quantify. We do this under some standard parametric models.

In reality, the elevated submatrix's size, namely (m, n) in (2.3), is rarely known. When m and n are unknown, one can perform a scan test for each (m, n) in some range of interest and control for multiple testing using the Bonferroni method. From [BI13], and also from other prior work (e.g., [ACV14]), we know that the resulting procedure achieves the same first-order asymptotic performance. In this chapter, we manage to show that out permutation test framework also enjoys such property.

Contribution 3 (First-order asymptotic performance of Bonferroni permutation test)**.** We develop a Bonferroni testing procedure based on the permutation test. We show that the testing procedure (at the first order) is asymptotically as powerful as the permutation

test and the test proposed by [BI13] when (m, n) is known. This result is analyzed and proved under some standard exponential family parameter assumptions.

As permutation test is computationally hard to calibrate (usually the total number of permutations are increasing exponentially with the sample size), in practice a Monte Carlo calibration method is executed. This requires independent sampling from the group of permutation patterns for a decent number of times. However still, due to the computational complexity of (2.3), performing permutation tests on all combinations of (m, n) could be extremely consuming in time and computational power. We construct a subset of $[M]\times[N]$, such that by performing permutation test on all (m, n) inside this subset, we can still detect the existence of the elevated submatrix with no sacrifice of the power on the first order.

Contribution 4 (Bonferroni procedure on an approximate net)**.** We propose an power-preserving fast test based on the Bonferroni permutation test, with the Bonferroni procedures working on a proper approximate net of $[M] \times [N]$. We show that this test is as powerful as the Bonferroni permutation test on all pairs of (m, n) in $[M] \times [N]$. This is also analyzed and proved under some standard exponential family parameter assumptions.

2.1.2 More related work

The scan statistic (2.3) is computationally intractable and there has been efforts to offer alternative approaches. We already mentioned [SWPN09], which proposes an alternate optimization strategy: given a set of rows, optimize over the set of columns, and vice versa, alternating in this fashion until convergence to a local maximum. This is the algorithm we use in our simulations. It does not come with theoretical guarantees (other than converging to a local maximum, a recently work [GL16] proves that the algorithm gives "promising" result with high probability under IID Gaussian entries in square matrices, but still it is

not guaranteed to be exact) but performs well numerically.

Our work here is not on the computational complexity of the problem. Rather we assume that we can compute the scan statistic and proceed to study it. In effect, we contribute here to a long line of work that studies permutation and rank-based methods for nonparametric inference. Most notably, we continue [ACTW17] where it studies the detection problem under a similar premise but under much more stringent structural assumptions. The setting there would correspond to an instance where the submatrix is in fact a block, meaning, that $\mathcal{I}_{\text{true}}$ and $\mathcal{J}_{\text{true}}$ are of the form $\mathcal{I}_{\text{true}} = \{i+1, \ldots, i+k\}$ and $\mathcal{J}_{\text{true}} = \{j+1, \ldots, j+l\}$. The present setting assumes much less structure. The related applications are very different in the end. Nevertheless, the technical arguments developed in [ACTW17] apply here with only minor adaptation. The main differences are that we consider two types of permutation and ranking protocols.

2.1.3 Content

The rest of the chapter is organized as follows. In Section 2.2 we describe a parametric setting where likelihood methods have been shown to perform well. This parametric setting will serve as benchmark for the nonparametric methods that ensue. In Section 2.3 we study the scan statistic with each of the two types of calibration by permutation. In Section 2.4 we study the rank-based scan statistic using each of the two types of ranking. In Section 2.5 we present some numerical experiments on simulated data. All the proofs are in Section 2.7.

2.2 The parametric scan

Following the classical line in the literature on nonparametric tests, we will evaluate the nonparametric methods introduced later on a family on parametric models. As in [BI13], and in [ACTW17], we consider a one-parameter exponential family in Section 1.3.

Such a parametric model is attractive as a benchmark because it includes these popular models and also because likelihood methods are known to be asymptotically optimal under such a model. [BI13] shows this is the case for the problem of detection, where the generalized likelihood ratio test is based on the scan statistic (2.3).

Under such a parametric model, the detection problem is formalized as a hypothesis testing problem where ν plays the role of null distribution. In detail, suppose that the submatrix is known to be $m \times n$. The search space is therefore

$$\mathbb{S}_{m,n} := \big\{\mathcal{S} = \mathcal{I} \times \mathcal{J} : \mathcal{I} \subset [M], |\mathcal{I}| = m \text{ and } \mathcal{J} \subset [N], |\mathcal{J}| = n\big\}.$$

We assume that the X_{ij}'s are independent with $X_{ij} \sim f_{\theta_{ij}}$, and the testing problem is

$$H_0 : \theta_{ij} = 0, \quad \forall (i,j) \in [M] \times [N],$$

versus

$$H_1 : \exists \mathcal{S}_{\text{true}} \in \mathbb{S}_{m,n} \text{ such that } \begin{cases} \theta_{ij} \geq \theta_\ddagger, & \forall (i,j) \in \mathcal{S}_{\text{true}}, \\ \theta_{ij} = 0, & \text{otherwise.} \end{cases}$$

Here $\theta_\ddagger$ controls the signal-to-noise ratio and is assumed to be known in this formulation. (The sum test and the scan test do not need to know this parameter.)

In this context, we have the following.

Theorem 1 ([BI13])**.** *Consider an exponential model as described above, with ν having finite fourth moment. Assume that*

$$M, N, m, n \to \infty, \quad \frac{m}{M}, \frac{n}{N} \to 0, \quad \frac{\log(M \vee N)}{m \wedge n} \to 0. \tag{2.4}$$

Then the sum test based on (2.2)*, at any fixed level* $\alpha > 0$*, has limiting power 1 when*

$$\theta_\sharp \frac{mn}{\sqrt{MN}} \to \infty.$$

Then the scan test based on (2.3)*, at any fixed level* $\alpha > 0$*, has limiting power 1 when*

$$\liminf \frac{\theta_\sharp \sqrt{mn}}{\sqrt{2(m \log \frac{M}{m} + n \log \frac{N}{n})}} > 1. \tag{2.5}$$

Conversely, the following matching lower bound holds. Assume in addition that $\log M \asymp \log N$ *and* $m \asymp n$*. Then any test at any fixed level* $\alpha > 0$ *has limiting power at most* α *when*

$$\theta_\sharp \frac{mn}{\sqrt{MN}} \to 0 \quad \textit{and} \quad \liminf \frac{\theta_\sharp \sqrt{mn}}{\sqrt{2(m \log \frac{M}{m} + n \log \frac{N}{n})}} < 1.$$

We note that [BI13] derived their lower bound under slightly weaker assumptions on M, N, m, n.

Remark 2. Proper calibration in this context is based on knowledge of the null distribution ν. In more detail, consider a test that rejects for large values of a statistic $T(\mathbf{X})$. Assuming a desired level of $\alpha > 0$ and that ν is either diffuse or discrete (for simplicity), the critical value for T is set at t_α, where $t_\alpha = \inf\{t : \nu(T(\mathbf{X}) \geq t) \leq \alpha\}$. The test is then $\mathbb{I}\{T(\mathbf{X}) \geq t_\alpha\}$. In practice, t_α may be approximated by Monte Carlo sampling.

2.3 Permutation scan tests

In the previous section we described the work of Butucea and Ingster [BI13], who in certain parametric models show that the sum test (2.2) and scan test (2.3) are jointly optimal for the problem of detecting a submatrix. This is so if they are both calibrated with full knowledge of the null distribution (denoted ν earlier).

What if the null distribution is unknown? A proven approach is via permutation. This is shown to be optimal in some classical settings [LR05] and was recently shown to also be optimal in more structured detection settings [ACTW17]. We prove that this is also the case in the present setting of detecting a submatrix. We consider the two types of permutation, unidimensional and bidimensional, described in Section 2.1.1. More elaborate permutation schemes have been suggested, e.g., in [BNW05], but these are not considered here, in part to keep the exposition simple. Indeed, we simply aim at showing that a calibration by permutation performs very well in the present context.

Let Π be a subgroup of permutations of $[M] \times [N]$, identified with $[MN]$. Then a calibration by permutation of the scan statistic (or any other statistic) yields the P-value

$$\mathfrak{P}(\mathbf{X}) = \frac{\#\{\pi \in \Pi : \textsc{scan}(\mathbf{X}_\pi) \geq \textsc{scan}(\mathbf{X})\}}{|\Pi|}, \tag{2.6}$$

where $\mathbf{X}_\pi = (X_{\pi(i,j)})$ is the matrix permuted by π. The permutation scan test at level α is the test $\mathbb{I}\{\mathfrak{P}(\mathbf{X}) \leq \alpha\}$. It is well-known that $\mathfrak{P}(\mathbf{X})$ is a valid P-value in the sense that it dominates the uniform distribution on $[0, 1]$ under the null [LR05]. (This remains true of a Monte Carlo approximation.)

The set of unidimensional permutations, denoted Π_1, is that of all permutations that permute within each row, while the set of bidimensional permutations, denoted Π_2, is simply the set of all permutations. Obviously, $\Pi_1 \subset \Pi_2$ with $|\Pi_1| = (N!)^M$ and $|\Pi_2| = (MN)!$, and they are both groups.[1]

Theorem 2. *Consider an exponential model as described in Section 2.2. In addition to* (2.4)*, assume*

$$\log^3(M \vee N)/(m \wedge n) \to 0, \tag{2.7}$$

and that either (i) ν *has support bounded from above, or* (ii) $\max_{i,j} \theta_{ij} \leq \bar{\theta}$ *for some* $\bar{\theta} < \theta_\star$

[1] The group structure is important. See the detailed discussion in [HG14].

fixed. Let the group of permutations Π *be either* Π_1 *or* Π_2*; if* $\Pi = \Pi_1$*, we require that* $\varphi(\theta) < \infty$ *for some* $\theta < 0$*. Then the permutation scan test based on* (2.6)*, at any fixed level* $\alpha > 0$*, has limiting power 1 when* (2.5) *holds.*

The additional condition (on ν or the nonzero θ_{ij}'s) seems artificial, but just as in [ACTW17], we are not able to eliminate it. Other than that, in view of Theorem 1 we see that the permutation scan test — just like the parametric scan test — is optimal to first-order under a general one-parameter exponential model.

2.4 Rank-based scan tests

Rank tests are classical special cases of permutation tests [Het84]. Traditionally, when computers were not as readily available and not as powerful, permutation tests were not practical, but rank tests could still be, as long as calibration had been done once for the same (or a comparable) problem size. Another well-known advantage of rank tests is their robustness to outliers.

We consider the two ranking protocols described in Section 2.1.1. After the observations are ranked, the distribution under the null is the permutation distribution, either uni- or bi-dimensional depending on the ranking protocol. This is strictly true under an appropriate exchangeability condition, which holds in the null model we consider here where all observations are IID. In fact, the unidimensional rank scan test is a form of unidimensional permutation test, and the bidimensional rank scan is a form of bidimensional permutation test, each time, the statistic being the rank scan

$$\text{SCAN}(\mathbf{R}) = \max_{\mathcal{I} \subset [M], |\mathcal{I}|=m} \quad \max_{\mathcal{J} \subset [N], |\mathcal{J}|=n} \quad \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} R_{ij}, \tag{2.8}$$

where $\mathbf{R} = (R_{ij})$ is the matrix of ranks.

Rank tests have been studied in minute detail in the classical setting [Het84, HS67]. Typically, this is done, again, by comparing their performance with the likelihood ratio test in the context of some parametric model. There is usually some loss in efficiency, unless one tailors the procedure to a particular parametric family.[2] Such a performance analysis was recently carried out for the rank scan in more structured settings [ACTW17]. We again extend this work here and obtain the following.

Define

$$\Upsilon = \mathbb{E}(Z\mathbb{1}_{(Z>Y)}) + \frac{1}{2}\mathbb{E}(Z\mathbb{1}_{(Z=Y)})$$

where Y, Z are IID with distribution ν. (This is the same constant introduced by [ACTW17].)

Theorem 3. *Consider an exponential model as described in Section 2.2. Assume that* (2.4) *holds. Let the group of permutations* Π *be either* Π_1 *or* Π_2*. The rank scan test at any fixed level* $\alpha > 0$ *has limiting power 1 when*

$$\liminf \frac{\theta_\ddagger \sqrt{mn}}{\sqrt{2(m \log \frac{M}{m} + n \log \frac{N}{n})}} > \frac{1}{2\sqrt{3}\Upsilon}. \tag{2.9}$$

Compared to the (optimal) performance of the parametric and permutation scan tests in the same setting (Theorem 1 and Theorem 2), we see that there is a loss in power. However, the loss can be quite small. For example, as argued in [ACTW17], in the normal model $1/(2\sqrt{3}\Upsilon) = \sqrt{\pi/3} \approx 1.023$.

[2] Actually, [Haj62] proposes a more complex method that avoids the need for knowing the parametric family.

2.5 Numerical experiments

We performed some numerical experiments to assess the accuracy of our asymptotic theory. To do so, we had to deal with two major issues in terms of computational complexity. The first issue is the computation of the scan statistic defined in (2.3). There are no known computationally tractable method for doing so. As [BI13] did, we opted instead for an approximation in the form of the alternate optimization (or hill-climbing) algorithm of [SWPN09]. Since in principle this algorithm only converges to a local maximum, we run the algorithm on several random initializations and take the largest output. The second issue is that of computing the permutation P-value defined in (2.6). (This is true for the permutation test and also for the special case of the rank test.) Indeed, examining all possible permutations in Π (either Π_1 or Π_2) is only feasible for very small matrices. As usual, we opted for Monte Carlo sampling. Specifically, we picked $\pi_1, \dots \pi_B$ IID uniform from Π with $B = 500$ in our setup. We then estimate the permutation P-value by

$$\hat{\mathfrak{P}}(\mathbf{X}) = \frac{\#\{b \in [B] : \mathrm{SCAN}(\mathbf{X}_{\pi_b}) \geq \mathrm{SCAN}(\mathbf{X})\} + 1}{B + 1}.$$

We mention that when rank methods are applied, the ties in the data are broken randomly.

Simulation setup Our simulation strategy is as follows. A data matrix $\mathbf{X}$ of size $M \times N$ is generated with the anomaly as $[m] \times [n]$. All the entries of $\mathbf{X}$ are independent with distribution f_0 (same as ν) except for the anomalous ones which have distribution $f_{\theta_\ddagger}$ for some $\theta_\ddagger > 0$. We compare the permutation tests and rank tests (unidimensional and bidimensional) with the scan test calibrated by Monte Carlo (using 500 samples), which serves as an oracle benchmark as it has full knowledge of the null distribution f_0. By construction, all tests have the prescribed level. As we increase $\theta_\ddagger$, the P-values of the

different tests are recorded. Each setting is repeated 200 times.

As one of the main purposes of our simulations is to confirm our theory, we zoom in on the region near the critical value

$$\theta_{\text{crit}} = \sqrt{\frac{2(m \log \frac{M}{m} + n \log \frac{N}{n})}{mn}},$$

which comes from (2.5). Specifically, we increase $\theta_\ddagger$ from $0.5 \times \theta_{\text{crit}}$ to $1.5 \times \theta_{\text{crit}}$ with step size $0.125 \times \theta_{\text{crit}}$ to explore the behavior of P-values around the critical value.

The Normal Case Here we generate data from normal family, where f_θ corresponds to $\mathcal{N}(\theta, 1)$. We used two setups, $(M, N, m, n) = (200, 100, 10, 15)$ and $(M, N, m, n) = (200, 100, 30, 10)$, to assess the performance of the tests under different anomaly sizes. The resulting boxplots of the averaged P-values are shown in Figure 2.1.

From the plots we see that the P-values are generally very close to 0 when $\theta_\ddagger$ exceeds θ_{crit}. When $(m, n) = (10, 15)$ the convergence towards 0 is slower, which may be due to the small size of the anomalous submatrix. As expected, the (oracle) Monte Carlo test is best, followed by the bidimensional permutation test, followed by the unidimensional permutation test. That said, the differences appear to be minor, which confirms our theoretical findings.

For the rank tests, we observe a similar behavior of the P-values, with the bidimensional showing superiority over the unidimensional rank test, but the loss of power with respect to the oracle test is a bit more substantial, as predicted by the theory. As shown before, $1/(2\sqrt{3}\Upsilon) \approx 1.03$ for the standard normal, so that we should place the critical threshold approximately at $1.03 \times \theta_{\text{crit}}$. This appears to be confirmed in the setting where $(m, n) = (30, 10)$. While the P-values for the rank tests converge relatively slowly when $(m, n) = (10, 15)$ (for unidimensional rank test the P-value is close to 0 at $\theta = 1.5 \times \theta_{\text{crit}}$), this may be due to the relatively small size of the anomaly.

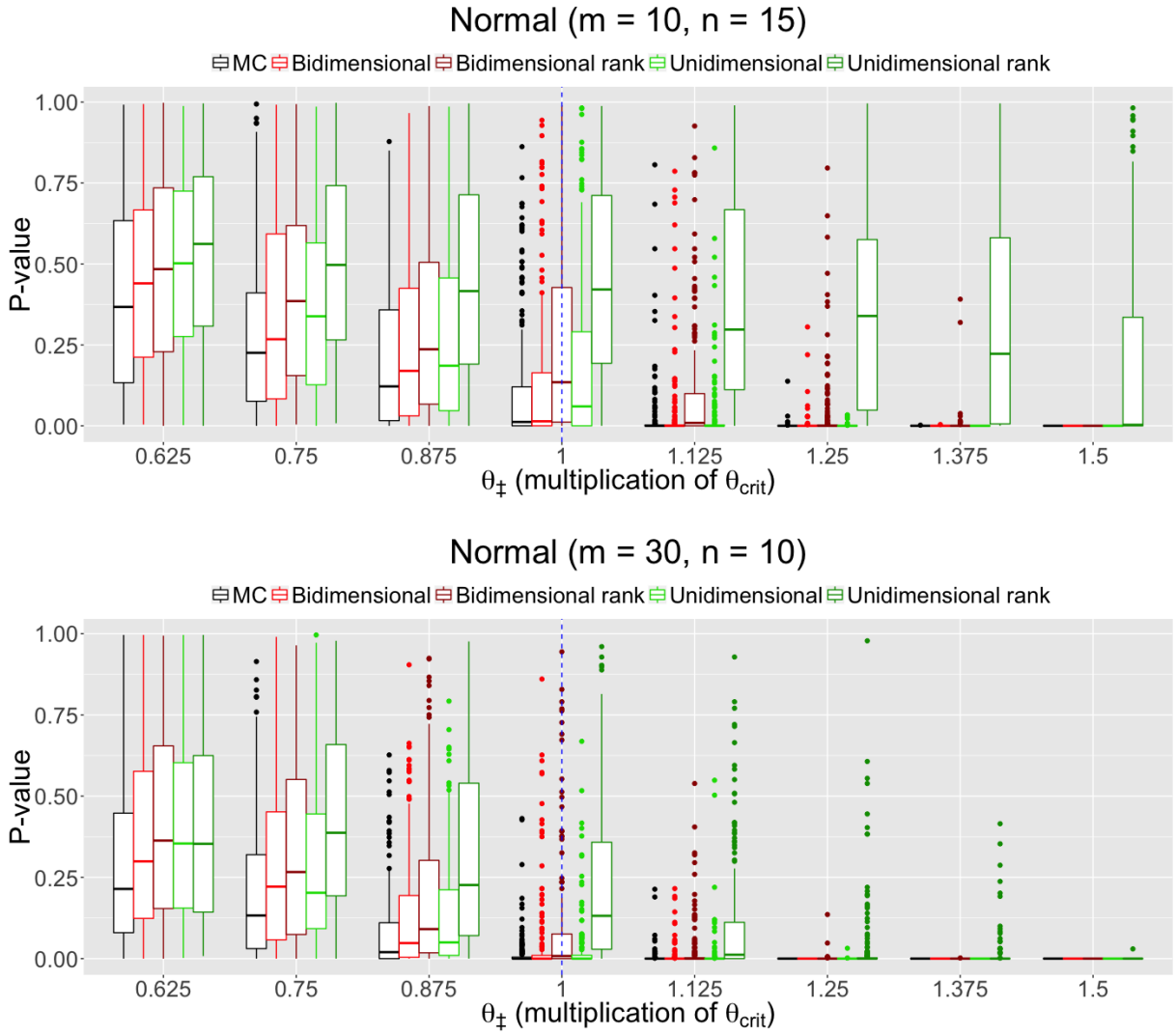

Figure 2.1: P-values of various forms of scan tests in the normal model

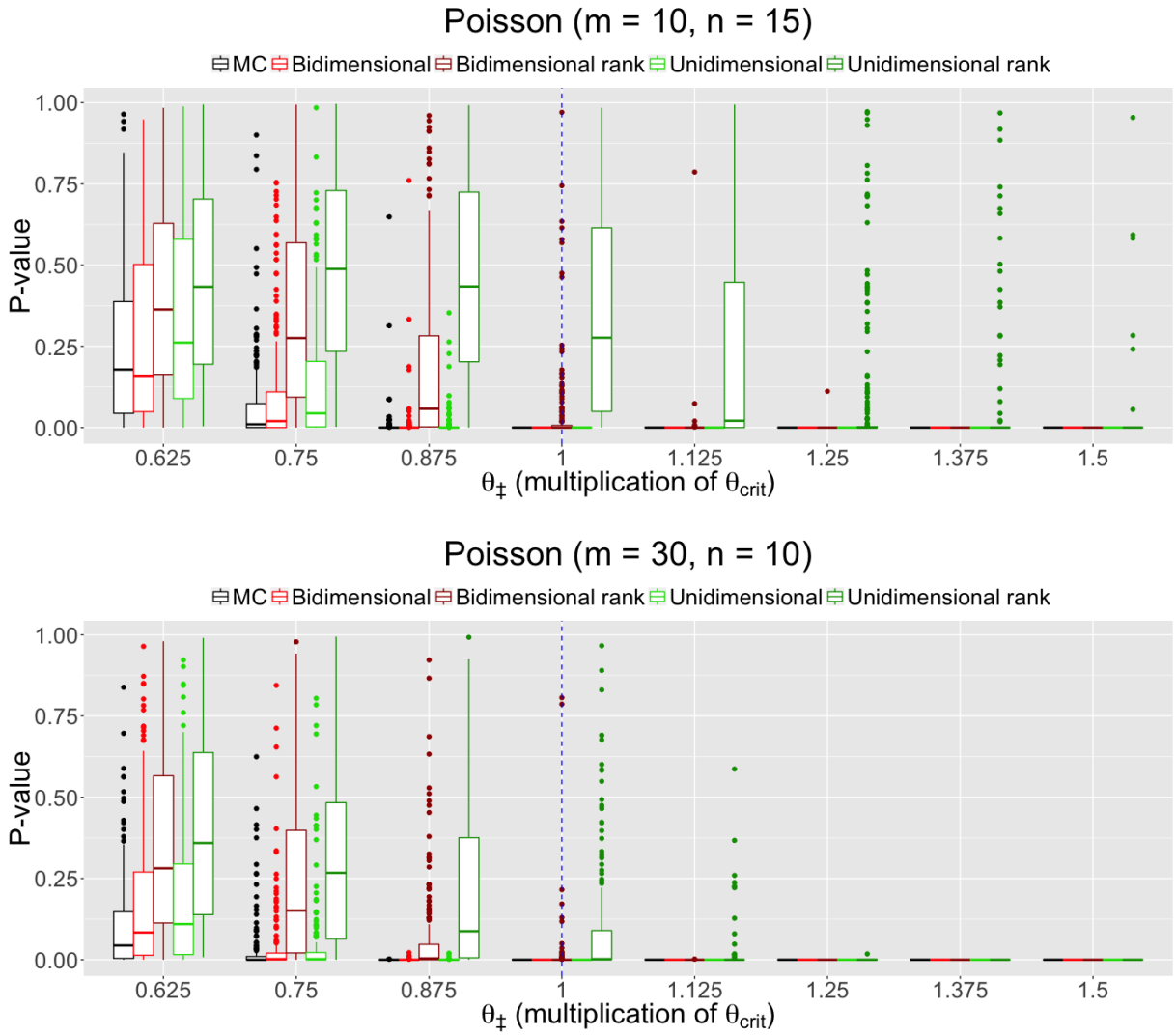

Figure 2.2: P-values of various forms of scan tests in the Poisson model

The Poisson Case As another example, we consider the Poisson family, where f_θ corresponds to $\mathcal{P}(e^\theta) - 1$. The data matrix and anomaly sizes are the same as they are in the normal case. The resulting boxplots of the P-values are shown in Figure 2.2. Overall, we observe a similar behavior of the P-values.

2.6 Bonferroni Permutation Test

2.6.1 Ordinary Bonferroni Procedure

In this section we construct the Bonferroni permutation test procedure under the situation where (m, n) is unknown. While the null hypothesis remains unchanged, the information of (m, n), contained inside the alternative hypothesis as mentioned in Section 2.2, will be removed. Therefore the hypotheses will be changed as

$$H_0 : \theta_{ij} = 0, \quad \forall (i, j) \in [M] \times [N],$$

versus

$$H_1 : \exists \mathcal{S}_{\text{true}} \text{ such that } \begin{cases} \theta_{ij} \geq \theta_\ddagger, & \forall (i, j) \in \mathcal{S}_{\text{true}}, \\ \theta_{ij} = 0, & \text{otherwise.} \end{cases}$$

Here $\mathcal{S}_{\text{true}}$ is a subset of $[M] \times [N]$ and do not have size regulations. Correspondingly, we define $H_1(m, n)$ as the alternative hypothesis defined in Section 2.2. So under $H_1(m, n)$, $\mathcal{S}_{\text{true}}$ has size (m, n) and has entries equipped with θ bounded from below. By realizing the fact that H_1 is true if and only if there exists some (m, n) such that $H_1(m, n)$ is true, we can perform test on $H_1(m, n)$ for all pairs of (m, n), and use Bonferroni correction in order to control the type I error. We adapt the permutation test from Section 2.3 here.

We start by defining some notations which has similar meanings in the previous sections, but have footnotes with (m, n) since (m, n) is allowed to change in this case. We start with the scan statistics:

$$\text{SCAN}_{m,n}(\mathbf{X}) = \max_{\mathcal{I} \subset [M], |\mathcal{I}| = m} \quad \max_{\mathcal{J} \subset [N], |\mathcal{J}| = n} \quad \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} X_{ij}. \tag{2.10}$$

and the permutation test distinguishing H_0 and $H_1(m,n)$ will be of the same from of (2.6), but with a bit of different notation:

$$\mathfrak{P}_{m,n}(\mathbf{X}) = \frac{|\{\pi \in \Pi : \mathrm{SCAN}_{m,n}(\mathbf{X}_\pi) \geq \mathrm{SCAN}_{m,n}(\mathbf{X})\}|}{|\Pi|}. \tag{2.11}$$

For each pair of (m,n), calculate $\mathfrak{P}_{m,n}(\mathbf{X})$, and calculate the final Bonferroni corrected p-value as

$$\mathfrak{P}_B(\mathbf{X}) = \min(MN \min_{m,n} \mathfrak{P}_{m,n}(\mathbf{X}), 1). \tag{2.12}$$

One rejects H_0 when $\mathfrak{P}_B(\mathbf{X})$ is less than some pre-determined level α. Due to the property of Bonferroni type of tests, this test has level α if each test concerning H_0 and $H_1(m,n)$ has level α, which is a proved fact in [LR05], regardless of the dependencies between tests.

Being a conservative method in multiple testing, Bonferroni method usually loses statistical power in exchange for controlling the family wise error rate. However indicated by some previous work (as mentioned before, [ACV14, BI13]), the Bonferroni procedure achieves the same first-order asymptotic power as the scan test with oracle knowledge of the submatrix size.

Theorem 4. *Consider an exponential model as described in* (1.2)*. Assume that there exists a pair of* (m,n) *such that* $H_1(m,n)$ *is true, and all the assumptions in Theorem 2 are satisfied. Then in probability,*

$$\mathfrak{P}_B(\mathbf{X}) \to 0.$$

2.6.2 A Power-preserving Fast Test

We build our Bonferroni testing framework on the permutation test, which is by nature a computationally intensive method. Calculation of scan statistic (2.10) is NP-hard, and the total number of permutation $|\Pi|$ will skyrocket when the number of data increases.

In practice the scan statistic is calculated by LAS algorithm proposed by [SWPN09], as in [BI13], and the permutation test is done by Monte-Carlo sampling, as in Section 2.5. In detail, a large number B is fixed, and permutations $\{\pi_1, \ldots, \pi_B\}$ is the IID sample from uniform distribution on $|\Pi|$. Then $\mathfrak{P}_{m,n}$ is approximated by

$$\hat{\mathfrak{P}}_{m,n}(\mathbf{X}) = \frac{|\{i : \mathrm{SCAN}_{m,n}(\mathbf{X}_{\pi_i}) \geq \mathrm{SCAN}_{m,n}(\mathbf{X})\}| + 1}{B+1}. \tag{2.13}$$

As illustrated in Section 2.5, the calculation of the scan statistic (2.10) is already difficult, even if the size (m, n) is known. Now with the submatrix size unknown, we have added the difficulty since the scan statistic under all possible combinations of (m, n), will be calculated during the Bonferroni process. In principle, the Bonferroni method requires going over all submatrix sizes, but we only scan a carefully chosen subset to lighten up the computational burden. Inspired by [ACDH05], we illustrate the construction of such subset on $[M] \times [N]$, and show that the first-order statistical power is preserved.

The subset of $[M] \times [N]$ we are going to construct in order to approximate the elements in $[M] \times [N]$ is called an approximate net. We first construct one-dimensional approximate net on $[M]$. We start by the following definition.

Definition 1. The binary expansion of an integer c is the sequence $\{a_i(c)\}$ with $a_i(c) \in \{0, 1\}$ and $i \leq \lfloor \log_2 c \rfloor$, such that

$$c = \sum_{i=0}^{\lfloor \log_2 c \rfloor} a_i(c) 2^i.$$

After representing an integer in the binary numeral system, one may approximate this integer by keeping the first k digits of its binary expansion. Denote c' as the integer satisfying

$$a_i(c') = a_i(c) \mathbb{1}\{i \geq \lfloor \log_2 c \rfloor - k + 1\}. \tag{2.14}$$

To find such an approximation of c, represent c in its binary expansion, keep the first k digits, and shrink the rest to zero. Finally calculate c' by the formula in Definition 1. It captures the main part of the integer and the difference could be controlled by k. The following lemma gives an upper bound of the difference rate.

Lemma 1. *If c' defined as in* (2.14)*,*

$$0 \leq \frac{c - c'}{c} \leq 2^{1-k}.$$

Note that the difference rate is only associated with k, and not associated with the value of c. Therefore if we apply the approximation to a collection of integers, the difference rate will be controlled *uniformly* among all the integers in the collection by the choice of k.

Now we construct the approximation net of $[M]$ based on the approximation.

Definition 2. An approximation net $S_k(M)$, of set $[M]$, is defined as

$$S_k(M) = \{c' : \text{There exists } c \in [M] \text{ such that (2.14) holds}\}.$$

The cardinality of $S_k(M)$ can be much less than M as k is chosen properly. For example, set $k = \log_2 \log_2 M + 1$ and it can be shown that $|S_k(M)| = O((\log_2 M)^2)$. Note that in this case $k \to \infty$ when $M \to \infty$, and by Lemma 1 we know that for every $c \in [M]$, there exists some $c' \in S_k(M)$ such that $c' = (1 + o(1))c$, and $o(1)$ is uniform among $[M]$.

Based on the one-dimensional approximation net defined in Definition 2, we can similarly extend the idea to sets of two-dimensional integer pairs. We perform the Bonferroni-type testing procedure on $S_k(M) \times S_l(N)$, instead of $[M] \times [N]$. In detail, we use the following Bonferroni corrected p-value:

$$\mathfrak{P}^A_{k,l}(\mathbf{X}) = \min\left(|S_k(M)||S_l(N)| \min_{(s,t)\in S_k(M)\times S_l(N)} \mathfrak{P}_{s,t}(\mathbf{X}), 1\right). \tag{2.15}$$

The idea is to use the property of approximation net to eliminate a significant portion of calculation by reducing the scanning region of the Bonferroni process, while keeping the accuracy through choosing a proper pair of (k, l). Assume we are under $H_1(m, n)$. When setting $k, l \to \infty$, there is a pair of $(m', n') \in S_k(M) \times S_l(N)$ that is close enough to (m, n), and $\mathfrak{P}_{m',n'}(\mathbf{X})$ will converge to zero fast enough such that brings the Bonferroni corrected p-value to zero as well.

It is clear that (2.15) defines a proper p-value. The following theorem describes the asymptotic power of the Bonferroni test on the approximate net.

Theorem 5. *Assume all the assumptions in Theorem 4 hold. Further set $k, l \to \infty$. Then in probability*

$$\mathfrak{P}^A_{k.l}(\mathbf{X}) \to 0.$$

2.7 Proofs

2.7.1 Preliminaries

We start with some preliminary results. First, for any one-parameter exponential family $(f_\theta : \theta \in \Theta)$ described in Section 1.3 with a standardized base distribution ν, as we consider to be here,

$$\mathbb{E}_\theta(X) \geq \theta, \quad \forall \theta \in \Theta. \tag{2.16}$$

Next, in the same context, if $\sup \Theta > 0$ (which we assume throughout), then f_θ has a sub-exponential right tail, which is uniform in $\theta \in \bar{\theta}$ if $\bar{\theta} \in \Theta$. In particular, there is $\bar{\gamma}$ that depends on $\bar{\theta} > 0$ such that, if $X_1, \ldots, X_k$ are independent, with $X_j \sim f_{\theta_j}$ and $\theta_j \leq \bar{\theta}$, then

$$\max_{j \in [k]} X_j \leq \bar{\gamma} \log k, \quad \text{with probability tending 1 as } k \to \infty. \tag{2.17}$$

By symmetry, if $\inf \Theta < 0$ (which we assume in the case of unidimensional permutations), the same is true on the left. In particular, ν itself (corresponding to $\theta = 0$) has a sub-exponential left tail in this case, meaning that there is a constant $\gamma_0 > 0$ such that, if $X_1, \ldots, X_k$ are IID ν, then

$$\min_{j \in [k]} X_j \geq -\gamma_0 \log k, \quad \text{with probability tending 1 as } k \to \infty. \tag{2.18}$$

2.7.2 Proof of Theorem 2

Unidimensional Case

In what follows, we take $\Pi = \Pi_1$. Recall that in this case we assume in addition that $\varphi(\theta) < \infty$ for some $\theta < 0$. This implies that ν as sub-exponential tails.

Case *(i)* We first focus on the condition where ν has support bounded from above and let b_0 denote such an upper bound. (Necessarily, $b_0 > 0$ since ν has zero mean.) Thus, regardless of the θ_{ij}'s,

$$\mathbb{P}(\max_{i,j} X_{ij} \leq b_0) = 1. \tag{2.19}$$

The permutation scan test has limiting power 1 if and only if $\mathbb{P}(\mathfrak{P}(\mathbf{X}) \leq \alpha) \to 1$ under the alternative. We show that by proving the stronger claim that $\mathfrak{P}(\mathbf{X}) \to 0$ in probability under the alternative.

We first work conditional on $\mathbf{X} = \mathbf{x}$, where $\mathbf{x} = (x_{ij})$ denotes a realization of $\mathbf{X} = (X_{ij})$ where $x_{ij} \leq b_0$ for all (i, j), which is typical due to (2.19). We may equivalently center the rows of $\mathbf{X}$ before scanning, and the resulting test remains unchanged. Therefore, we may assume that all the rows of $\mathbf{x}$ sum to 0. Let $\zeta = \textsc{scan}(\mathbf{x})$ for short. We have

$$\mathfrak{P}(\mathbf{x}) = \mathbb{P}(\textsc{scan}(\mathbf{x}_\pi) \geq \zeta),$$

where the randomness comes solely from π, uniformly drawn from Π. Using the union bound, we get

$$\mathfrak{P}(\mathbf{x}) \le |\mathbb{S}_{m,n}| \max_{\mathcal{S}\in\mathbb{S}_{m,n}} \mathbb{P}\Big(\textstyle\sum_{(i,j)\in\mathcal{S}} x_{\pi(i,j)} \ge \zeta\Big).$$

For each $i \in [N]$, let $(A_{ij} : j \in [n])$ be a sample from $(x_{ij} : j \in [N])$ *without* replacement and let $A_i = \sum_{j\in[n]} A_{ij}$. Note that $A_1, \dots, A_M$ are independent and, for $\mathcal{S} = \mathcal{I} \times \mathcal{J}$, we have

$$\sum_{(i,j)\in\mathcal{S}} x_{\pi(i,j)} \sim \sum_{i\in\mathcal{I}} A_i.$$

Fix $\mathcal{I} \subset [M]$ of size m. Using Markov's inequality and the independence of the A_i's, we get

$$\mathbb{P}\Big(\textstyle\sum_{i\in\mathcal{I}} A_i \ge \zeta\Big) \le e^{-c\zeta} \prod_{i\in\mathcal{I}} \phi_i(c),$$

where ϕ_i is the moment generating function of A_i. The key is [Hoe63, Theorem 4], which implies that $\phi_i \le \psi_i$, where ψ_i is the moment generating function of B_i, where $B_i = \sum_{j\in[n]} B_{ij}$ and $(B_{ij} : j \in [n])$ is a sample from $(x_{ij} : j \in [N])$ *with* replacement, meaning that these are IID random variables uniformly distributed in $(x_{ij} : j \in [N])$. We have $B_{ij} \le b_0$, and the classical arguments leading to the (one-sided) Bernstein inequality yield the usual bound

$$\psi_i(c) \le \exp\left(\frac{nc^2\sigma_i^2}{2} \frac{e^{cb_0} - 1 - cb_0}{c^2 b_0^2/2}\right),$$

where σ_i^2 is the variance of B_{i1}, meaning, $\sigma_i^2 := \frac{1}{N}\sum_{j\in[N]} (x_{ij} - \bar{x}_i)^2$, with $\bar{x}_i := \frac{1}{N}\sum_{j\in[N]} x_{ij}$ being the mean. Letting $\sigma^2 = \max_{i\in[M]} \sigma_i^2$, we derive

$$e^{-c\zeta} \prod_{i\in\mathcal{I}} \phi_i(c) \le e^{-c\zeta} \prod_{i\in\mathcal{I}} \exp\left(\frac{nc^2\sigma_i^2}{2} \frac{e^{cb_0} - 1 - cb_0}{c^2 b_0^2/2}\right) \le e^{-c\zeta} \exp\left(\frac{mnc^2\sigma^2}{2} \frac{e^{cb_0} - 1 - cb_0}{c^2 b_0^2/2}\right),$$

the latter being the usual bound that leads to Bernstein's inequality, and the same

optimization over c then yields

$$\mathbb{P}\left(\sum_{i\in\mathcal{I}} A_i \geq \zeta\right) \leq \exp\left(-\frac{\zeta^2}{2mn\sigma^2 + \frac{2}{3}b_0\zeta}\right).$$

We now emphasize the dependency of ζ and σ^2 on $\mathbf{x}$ by adding $\mathbf{x}$ as a subscript. Noting that this bound is independent of $\mathcal{I}$ (of size m), we get

$$\mathfrak{P}(\mathbf{x}) \leq |\mathbb{S}_{m,n}| \exp\left(-\frac{\zeta_{\mathbf{x}}^2}{2mn\sigma_{\mathbf{x}}^2 + \frac{2}{3}b_0\zeta_{\mathbf{x}}}\right). \tag{2.20}$$

We now free $\mathbf{X}$ and bound $\zeta_{\mathbf{X}}$ from below, and $\sigma_{\mathbf{X}}^2$ from above. When doing so, we need to take into account that we assumed the rows summed to 0. When this is no longer the case, $\zeta_{\mathbf{X}}$ denotes the scan of $\mathbf{X}$ after centering all the rows. Let $\bar{X}_i$ denote the mean of row i. By definition of the scan in (2.3),

$$\zeta_{\mathbf{X}} \geq \zeta_{\text{true}} \coloneqq \sum_{i\in\mathcal{I}_{\text{true}}} \sum_{j\in\mathcal{J}_{\text{true}}} (X_{ij} - \bar{X}_i) = (1 - \tfrac{n}{N}) \sum_{i\in\mathcal{I}_{\text{true}}} \sum_{j\in\mathcal{J}_{\text{true}}} X_{ij} - \tfrac{n}{N} \sum_{i\in\mathcal{I}_{\text{true}}} \sum_{j\notin\mathcal{J}_{\text{true}}} X_{ij}.$$

For the expectation, by (2.4) and (2.16), we have

$$\mathbb{E}(\zeta_{\text{true}}) \geq (1 - \tfrac{n}{N}) \sum_{i\in\mathcal{I}_{\text{true}}} \sum_{j\in\mathcal{J}_{\text{true}}} \theta_{ij} \geq (1 - o(1))mn\theta_{\ddagger}.$$

For the variance, we have $\text{Var}(X_{ij}) = 1$ when $(i,j) \notin \mathcal{S}_{\text{true}}$ (since ν has variance 1) and $\text{Var}(X_{ij}) \leq \mathbb{E}(X_{ij}^2) \leq b_0^2$ always. Using this, we derive

$$\text{Var}(\zeta_{\text{true}}) \leq mnb_0^2 + (\tfrac{n}{N})^2 mN = mn(b_0^2 + \tfrac{n}{N}) = O(mn).$$

Because of (2.4) and (2.5), $\mathbb{E}(\zeta_{\text{true}}) \gg \sqrt{\text{Var}(\zeta_{\text{true}})}$, and thus by Chebyshev's inequality,

$$\zeta_{\text{true}} = (1 + o_P(1))\,\mathbb{E}(\zeta_{\text{true}}) \geq (1 + o_P(1))mn\theta_{\ddagger}. \tag{2.21}$$

We now bound $\sigma_{\mathbf{x}}^2$. For $i \in \mathcal{I}_{\text{true}}$, we have

$$\sigma_i^2(\mathbf{X}) \le \frac{1}{N}\sum_{j\in[N]} X_{ij}^2 = \frac{1}{N}\sum_{j\in\mathcal{J}_{\text{true}}} X_{ij}^2 + \frac{1}{N}\sum_{j\notin\mathcal{J}_{\text{true}}} X_{ij}^2 \le \frac{nb_0^2}{N} + \frac{1}{N}\sum_{j\notin\mathcal{J}_{\text{true}}} X_{ij}^2.$$

For $i \notin \mathcal{I}_{\text{true}}$,

$$\sigma_i^2(\mathbf{X}) \le \frac{1}{N}\sum_{j\in[N]} X_{ij}^2.$$

Therefore

$$\sigma_{\mathbf{X}}^2 \overset{\text{sto}}{\le} 1 + o(1) + \max_{i\in[M]} \frac{1}{N}\sum_{j\in[N]} T_{ij}, \tag{2.22}$$

where $(T_{ij} : (i,j) \in [M]\times[N])$ are IID with distribution that of $X^2 - 1$ when $X \sim \nu$. Note that $\mathbb{E}(T_{ij}) = 0$ since ν has variance 1 and

$$\max_{i,j} T_{ij} \le \bar{t} := b_0^2 \vee (\gamma_0 \log(MN))^2,$$

by (2.19) and when the following event holds

$$\mathcal{A} := \Big\{ \min_{i,j} X_{ij} \ge -\gamma_0 \log(MN) \Big\},$$

which by (2.18) happens with probability tending to 1. Let $\mathbb{P}_{\mathcal{A}}$ be the probability conditional on $\mathcal{A}$ and $\mathbb{E}_{\mathcal{A}}$ the corresponding expectation. Let $\mu_A = \mathbb{E}_{\mathcal{A}}(T_{ij})$ and $\tau_{\mathcal{A}}^2 = \text{Var}_{\mathcal{A}}(T_{ij})$. Note that $\tau_{\mathcal{A}}^2 < \infty$ because ν has finite fourth moment. By Bernstein's inequality, for any $c > \mu_{\mathcal{A}}$,

$$\mathbb{P}_{\mathcal{A}}\left(\frac{1}{N}\sum_{j\in[N]} T_{ij} > c \right) \le \exp\left(-\frac{N(c-\mu_{\mathcal{A}})^2}{2\tau_{\mathcal{A}}^2 + \frac{2}{3}\bar{t}c} \right).$$

Then, using a union bound,

$$\mathbb{P}_{\mathcal{A}}\left(\max_{i\in[M]}\frac{1}{N}\sum_{j\in[N]}T_{ij}>c\right)\leq M\exp\left(-\frac{N(c-\mu_{\mathcal{A}})^2}{2\tau_{\mathcal{A}}^2+\frac{2}{3}\bar{t}c}\right). \tag{2.23}$$

Taking natural logs, noting that $\mu_{\mathcal{A}}\to 0$ and $\tau_{\mathcal{A}}^2\to\tau^2:=\mathrm{Var}(T_{ij})$, as well as $\bar{t}=O(\log(MN))$, and using (2.4) and (2.7), we see that the RHS tends to 0 for any $c>0$ fixed. Therefore $\max_{i\in[M]}\frac{1}{N}\sum_{j\in[N]}T_{ij}=o_P(1)$ conditional on $\mathcal{A}$, and since $\mathbb{P}(\mathcal{A})\to 1$, also unconditionally. Coming back to (2.22), we conclude that

$$\sigma_{\mathbf{X}}^2=1+o_P(1). \tag{2.24}$$

The upper bound on $\zeta_{\mathbf{X}}$ and the lower bound on $\sigma_{\mathbf{X}}^2$, combined, imply by monotonicity that

$$\frac{\zeta_{\mathbf{X}}^2}{2mn\sigma_{\mathbf{X}}^2+\frac{2}{3}b_0\zeta_{\mathbf{X}}}\geq(1+o_P(1))\frac{mn\theta_\ddagger^2}{2+\frac{2}{3}b_0\theta_\ddagger}.$$

We also have $|\mathbb{S}_{m,n}|=\binom{M}{m}\binom{N}{n}$, so that

$$\log|\mathbb{S}_{m,n}|=\log\binom{M}{m}+\log\binom{N}{n}\leq(1+o(1))\Lambda,$$

with

$$\Lambda:=\left(m\log\tfrac{M}{m}+n\log\tfrac{N}{n}\right),$$

where in the last inequality we used (2.4) and the fact that $\log\binom{K}{k}\leq k\log(K/k)+k$ for all integers $1\leq k\leq K$.

Coming back to (2.20) and collecting all the bounds in between, we find that

$$\log \mathfrak{P}(\mathbf{X}) \le (1+o(1))\Lambda - (1+o_P(1))\frac{mn\theta_\ddagger^2}{2+\frac{2}{3}b_0\theta_\ddagger}.$$

Under (2.5), there is $\varepsilon > 0$ such that, eventually,

$$\theta_\ddagger \ge (1+\varepsilon)\sqrt{2\Lambda/(mn)}. \tag{2.25}$$

When that's the case, we get

$$\log \mathfrak{P}(\mathbf{X}) \le (1+o(1))\Lambda - (1+o_P(1))\frac{(1+2\varepsilon)\Lambda}{1+\frac{1}{3}b_0(1+\varepsilon)\sqrt{2\Lambda/(mn)}}. \tag{2.26}$$

Noting that $\Lambda/(mn) = o(1)$ and $\Lambda \to \infty$ under (2.4), we get

$$\log \mathfrak{P}(\mathbf{X}) \le -(1+o_P(1))2\varepsilon\Lambda \to -\infty,$$

which is what we needed to prove.

Case *(ii)* We now consider the case where $\theta_{ij} \le \bar{\theta}$ for all $(i,j) \in [M]\times[N]$ for some $\bar{\theta} < \theta_*$. Although (2.19) may not hold for any b_0, we redefine $b_0 = \bar{\gamma}\log(MN)$, where $\bar{\gamma}$ depends on $\bar{\theta}$, and condition on the event

$$\mathcal{B} := \Big\{\max_{i,j} X_{ij} \le b_0\Big\},$$

which holds with probability tending to 1 by (2.17). The bound (2.20) holds unchanged (assuming that $\max_{i,j} x_{ij} \le b_0$). What is different is how $\zeta_{\mathbf{X}}$ and $\sigma^2_{\mathbf{X}}$ are handled, now that we conditioned on $\mathcal{B}$. Let $\mathbb{P}_{\mathcal{B}}$ and $\mathbb{E}_{\mathcal{B}}$ denote the probability and expectation conditional on $\mathcal{B}$.

We have

$$\begin{aligned}\mathbb{E}_{\mathcal{B}}(\zeta_{\text{true}}) &\geq (1-\tfrac{n}{N}) \sum_{i\in\mathcal{I}_{\text{true}}} \sum_{j\in\mathcal{J}_{\text{true}}} \mathbb{E}_{\mathcal{B}}(X_{ij}) - \tfrac{n}{N} \sum_{i\in\mathcal{I}_{\text{true}}} \sum_{j\notin\mathcal{J}_{\text{true}}} \mathbb{E}_{\mathcal{B}}(X_{ij}) \\ &= (1-\tfrac{n}{N}) \sum_{i\in\mathcal{I}_{\text{true}}} \sum_{j\in\mathcal{J}_{\text{true}}} \mathbb{E}(X_{ij}|X_{ij}\leq b_0) - \tfrac{n}{N} \sum_{i\in\mathcal{I}_{\text{true}}} \sum_{j\notin\mathcal{J}_{\text{true}}} \mathbb{E}(X_{ij}|X_{ij}\leq b_0) \\ &\geq (1+o(1))mn\theta_{\ddagger}.\end{aligned}$$

In the last inequality, for $j \notin \mathcal{J}_{\text{true}}$ we used the fact that $\mathbb{E}(X_{ij}) = 0$, which implies that $\mathbb{E}_{\mathcal{B}}(X_{ij}) \leq 0$ in that case. And for $j \in \mathcal{J}_{\text{true}}$ we used the fact that $\mathbb{E}(X_{ij}|X_{ij} \leq b_0) \to \theta_{ij} \geq \theta_{\ddagger}$ combined with a Cèsaro-type argument. On the other hand, in a similar way, we also have

$$\text{Var}_{\mathcal{B}}(\zeta_{\text{true}}) = O(mnb_0^2) = O(mn\log^2(MN)).$$

So we still have $\mathbb{E}_{\mathcal{B}}(\zeta_{\text{true}}) \gg \sqrt{\text{Var}_{\mathcal{B}}(\zeta_{\text{true}})}$, by (2.4) and (2.5), and in addition (2.7). In particular, (2.21) holds under $\mathcal{B}$. In very much the same way, one can verify that the same is true of (2.24).

From there we get to (2.26) in exactly the same way, conditionally on $\mathcal{B}$, and then unconditionally since $\mathbb{P}(\mathcal{B}) \to 1$. Then, to conclude, we only need to check that $b_0\sqrt{\Lambda/(mn)} = o(1)$, which is the case by (2.7).

Bidimensional Case

The following lemma is at the center of our argument.

Lemma 2 ([ACTW17] Lemma 2, Bernstein's inequality for sampling without replacement)**.** *Let $(Z_1, \ldots, Z_m)$ be obtained by sampling without replacement from a given a set of real numbers $\{z_1, \ldots, z_J\} \subset \mathbb{R}$. Define $z_{\max} = \max_j z_j$, $\bar{z} = \frac{1}{J}\sum_j z_j$, and $\sigma_z^2 = \frac{1}{J}\sum_j (z_j - \bar{z})^2$. Then*

the sample mean $\bar{Z} = \frac{1}{m}\sum_i Z_i$ *satisfies*

$$\mathbb{P}\left(\bar{Z} \geq \bar{z} + t\right) \leq \exp\left[- \frac{mt^2}{2\sigma_z^2 + \frac{2}{3}(z_{\max} - \bar{z})t} \right], \quad \forall t \geq 0.$$

The lemma is a result from [Ser74]. We also refer the reader to [BM+15, BLM13] for further details of the lemma.

Starting from here we denote $S^* = \mathcal{I}_{\text{true}} \times \mathcal{J}_{\text{true}}$ as the true anomaly submatrix, and recall that (m, n) are its row and column sizes. For a submatrix index set S, denote $Y_S(X) = \sum_{(i,j)\in S} X_{ij}$. Now fix some $S \in \mathbb{S}_{m,n}$ and uniformly sample a permutation $\pi \in \Pi_2$. Also, if we conditional a realization of $\mathbf{X}$, say $\mathbf{x}$, by Lemma 2, we have

$$\mathbb{P}_\pi \left(\frac{Y_S(\mathbf{x}_\pi)}{mn} - \bar{\mathbf{x}} > t \right) \leq \exp\left(- \frac{t^2 mn}{2\sigma_{\mathbf{x}}^2 + \frac{2}{3}(\mathbf{x}_{\max} - \bar{\mathbf{x}})t} \right),$$

where $\mathbf{x}_{\max}$ is the largest value in the data matrix $\mathbf{x}$, and $\sigma_{\mathbf{x}}^2$ here, is redefined as the sample variance of the data if $\mathbf{x}$ is treated as a one dimensional data vector. Notice that the probability is on the permutation process. Therefore we rewrite the above inequality as

$$\sum_{\pi\in\Pi_2} \frac{\mathbb{1}\left(\frac{Y_S(\mathbf{x}_\pi)}{mn} - \bar{\mathbf{x}} > t\right)}{|\Pi_2|} \leq \exp\left(- \frac{t^2 mn}{2\sigma_{\mathbf{x}}^2 + \frac{2}{3}(\mathbf{x}_{\max} - \bar{\mathbf{x}})t} \right).$$

Notice that with π fixed,

$$\mathbb{1}\left(\frac{\text{SCAN}(\mathbf{x})}{mn} - \bar{\mathbf{x}} > t \right) \leq \sum_{S\in\mathbb{S}_{m,n}} \mathbb{1}\left(\frac{Y_S(\mathbf{x}_\pi)}{mn} - \bar{\mathbf{x}} > t \right).$$

Therefore, if set $t = \textsc{scan}(\mathbf{x})/mn - \bar{\mathbf{x}}$, we have

$$\begin{aligned}\mathfrak{P}(\mathbf{x}) &= \sum_{\pi\in\Pi_2} \frac{\mathbb{1}\left(\frac{\textsc{scan}(\mathbf{x})}{mn} - \bar{\mathbf{x}} > t\right)}{|\Pi_2|} \\ &\leq |\mathbb{S}_{m,n}| \exp\left(-\frac{mn(\textsc{scan}_{m,n}(\mathbf{x})/mn - \bar{\mathbf{x}})^2}{2\sigma_{\mathbf{x}}^2 + \frac{2}{3}(\mathbf{x}_{\max} - \bar{\mathbf{x}})(\textsc{scan}_{m,n}(\mathbf{x})/mn - \bar{\mathbf{x}})}\right)\end{aligned}$$

Note that this inequality holds for all the realizations of $\mathbf{X}$, thus we allow $\mathbf{X}$ change and focusing on the quantity

$$|\mathbb{S}_{m,n}| \exp\left(-\frac{mn(\textsc{scan}_{m,n}(\mathbf{X})/mn - \bar{\mathbf{X}})^2}{2\sigma_{\mathbf{X}}^2 + \frac{2}{3}(\mathbf{X}_{\max} - \bar{\mathbf{X}})(\textsc{scan}_{m,n}(\mathbf{X})/mn - \bar{\mathbf{X}})}\right). \tag{2.27}$$

We show that this quantity is $o_P(1)$.

We start with bounding $\bar{\mathbf{X}}$ by rewriting $\bar{\mathbf{X}}$ as

$$\bar{\mathbf{X}} = \frac{\sum_{i,j} X_{ij}}{MN} = \frac{mn}{MN} \cdot \frac{\sum_{(i,j)\in S^*} X_{ij}}{mn} + \frac{MN - mn}{MN} \cdot \frac{\sum_{(i,j)\notin S^*} X_{ij}}{MN - mn}.$$

With θ in the anomalous submatrix bounded from above, or the support of f_θ being bounded, the first term is $o_P(1)$ since $mn = o(MN)$. By Law of Large Numbers, the second term is $o_P(1)$ given that distribution f_0 has mean zero. So

$$\bar{\mathbf{X}} = o_P(1).$$

Follow (2.17), we can bound $\mathbf{X}_{\max}$ with

$$\mathbb{P}(\mathbf{X}_{\max} < \bar{\gamma}\log(MN)) \to 1. \tag{2.28}$$

Denote the event $\mathcal{A} = \{\mathbf{X}_{\max} < \bar{\gamma}\log(MN)\}$. All the following arguments are conditional on

$\mathcal{A}$. But since $\mathcal{A}$ happens with probability tending to 1, all the conditional high probability events will happen unconditionally with high probability as well.

We do similar operations to bound $\sigma_{\mathbf{X}}^2$ as follows.

$$\begin{aligned}\sigma_{\mathbf{X}}^2 &= \frac{1}{MN}\sum_{i,j}(X_{ij}-\bar{\mathbf{X}})^2 \le \frac{1}{MN}\sum_{i,j}X_{ij}^2 \\ &= \frac{mn}{MN}\cdot\frac{\sum_{X_{ij}\in S^*}X_{ij}^2}{mn} + \frac{MN-mn}{MN}\cdot\frac{\sum_{X_{ij}\notin S^*}X_{ij}^2}{MN-mn}.\end{aligned}$$

The first term is $o_P(1)$ if distribution f_θ has finite second moment, which can be derived from the assumption. The second term is $1+o_P(1)$ by Law of Large Numbers and Slutsky's Lemma. Therefore

$$\sigma_{\mathbf{X}}^2 = 1 + o_P(1) \tag{2.29}$$

Finally we bound $\textsc{scan}_{m,n}(\mathbf{X})/mn - \bar{\mathbf{X}}$ as a whole. By the definition of the scan statistic,

$$\frac{\textsc{scan}_{m,n}(\mathbf{X})}{mn} - \bar{\mathbf{X}} \ge \frac{Y_{S^*}(\mathbf{X})}{mn} - \bar{\mathbf{X}} = \bar{\mathbf{X}}_{S^*} - \bar{\mathbf{X}},$$

here $\bar{\mathbf{X}}_{S^*} = \sum_{(i,j)\in S^*}X_{ij}/mn$ represents the average in the submatrix indexed by S^*. Rewrite $X_{ij} = E(X_{ij}) + Z_{ij}$ for $X_{ij}\in S^*$, where Z_{ij} has mean zero and bounded second moment. By Law of Large numbers, as well as (2.16),

$$\bar{\mathbf{X}}_{\mathbf{S}^*} = \frac{1}{mn}\sum_{X_{ij}\in S^*}\mathbb{E}\,X_{ij} + O_P(\frac{1}{\sqrt{mn}}) \ge \theta_\ddagger + O_P(\frac{1}{\sqrt{mn}}).$$

Note that $\bar{\mathbf{X}} = O_P(1/\sqrt{MN})$ and $mn = o(MN)$,

$$\frac{\textsc{scan}_{m,n}(\mathbf{X})}{mn} - \bar{\mathbf{X}} \ge \theta_\ddagger + O_P(\frac{1}{\sqrt{MN}})$$

By (2.5) we know that $\sqrt{mn}\theta_\ddagger \to \infty$, or $\theta_\ddagger \gg 1/\sqrt{MN}$, so we can rewrite the equation above as

$$\frac{\text{SCAN}_{m,n}(\mathbf{X})}{mn} - \bar{\mathbf{X}} \geq \theta_\ddagger(1 + o_P(1)). \tag{2.30}$$

Plug in (2.28), (2.29) and (2.30) into (2.27), we have

$$\mathfrak{P}(\mathbf{X}) \leq |\mathbb{S}_{m,n}| \exp\left(- \frac{(1 + o_P(1))\theta_\ddagger^2 mn}{2(1 + o_P(1)) + \frac{2}{c}(\log MN)\theta_\ddagger(1 + o_P(1))} \right).$$

From (2.7) we get the second term in the denominator in the exponent component is $o_P(1)$. Assume in (2.5),

$$\liminf \frac{\theta_\ddagger^2 mn}{2(m \log \frac{M}{m} + n \log \frac{N}{n})} \geq 1 + \epsilon$$

with some constant $\epsilon > 0$, then eventually with high probability we well have

$$\begin{aligned} \log(\mathfrak{P}(\mathbf{X})) &\leq \log(|\mathbb{S}_{m,n}|) - \frac{(1 - \epsilon/2)\theta_\ddagger^2 mn}{2(1 + \epsilon/8)} \\ &\leq \log(|\mathbb{S}_{m,n}|) - (1 + \frac{\epsilon}{4 + \epsilon/2})\Lambda. \end{aligned}$$

By the fact that $|\mathbb{S}_{m,n}| \leq (1 + o(1))\Lambda$, we have eventually

$$\log(|\mathbb{S}_{m,n}|) \leq ((1 + \frac{\epsilon}{8 + \epsilon}))\Lambda.$$

So the with high probability the permutation p-value is bounded from above by

$$\log(\mathfrak{P}(\mathbf{X})) \leq -(\frac{\epsilon}{8 + \epsilon})\Lambda \tag{2.31}$$

which finishes the proof.

2.7.3 Proof of Theorem 3

Unidimensional Case

As rank tests are permutation tests, it suffices to specialize and detail the arguments given in Section 2.7.2. We focus again on the unidimensional case. In that case, in each row i, the ranks $(R_{ij} : j \in [N])$ are a permutation of $\{1, \dots, N\}$. Therefore the row mean is $(N+1)/2$ and the variance is $(N^2-1)/12$. Hence, with $\mathbf{R} = \mathbf{r}$ denoting the ranks of $\mathbf{X} = \mathbf{x}$, (2.20) becomes

$$\mathfrak{P}(\mathbf{r}) \le |\mathbb{S}_{m,n}| \exp\left[-\frac{\zeta_{\mathbf{r}}^2}{2mn((N^2-1)/12) + \frac{2}{3}N\zeta_{\mathbf{r}}} \right], \tag{2.32}$$

where, as before, $\zeta_{\mathbf{r}}$ is the rank scan of $\mathbf{r}$ with the row means removed.

As in [ACTW17], by a monotonicity argument, we may reduce the situation to the last favorable case where $\theta_{ij} = \theta_\ddagger$ when $(i,j) \in \mathcal{S}_{\text{true}}$ and assume an equality in (2.25). Following the same arguments used in [ACTW17] for bounding $\zeta_{\mathbf{R}}$ from below, we arrive at

$$\zeta_{\mathbf{R}} \ge mn(N-n)\big(\theta_\ddagger \Upsilon - O(\theta_\ddagger^2)\big) + O_p(N\sqrt{mn}) = (1 + o_P(1))mnN\theta_\ddagger\Upsilon.$$

$$\frac{\zeta_{\mathbf{r}}^2}{2mn((N^2-1)/12) + \frac{2}{3}N\zeta_{\mathbf{r}}} \ge (1 + o_P(1))\frac{mn\theta_\ddagger^2\Upsilon^2}{2/12 + \frac{2}{3}\theta_\ddagger\Upsilon}. \tag{2.33}$$

From (2.9), there exists some $\varepsilon > 0$ such that eventually,

$$\Upsilon\theta_\ddagger \ge (1+\varepsilon)\sqrt{\frac{2\Lambda}{12mn}}.$$

Along with (2.33) and (2.32), we have the bound

$$\log \mathfrak{P}(\mathbf{R}) \le (1 + o(1))\Lambda - (1 + o_P(1))\frac{(1+2\varepsilon)\frac{2\Lambda}{12}}{2/12 + \frac{2}{3}(1+\varepsilon)\sqrt{\frac{2\Lambda}{12mn}}}.$$

With $\Lambda/mn \to 0$ and $\Lambda \to \infty$,

$$\log \mathfrak{P}(\mathbf{R}) \leq -(1 + o(1))2\varepsilon\Lambda \to -\infty,$$

which concludes the proof.

When $\theta_{ij} \geq \theta_\ddagger$ with $(i, j) \in \mathcal{S}_{\text{true}}$, note that p_θ is increasing with θ, therefore the P-value will be stochastically dominated by the one during proof.

Bidimensional Case

According to the analysis before, all that remains to be done is to study the performance of the rank scan test under the alternative.

We may directly go through the same procedure as in the data case, to obtain

$$\mathfrak{P}(\mathbf{r}) \leq |\mathbb{S}_{m,n}| \exp\left(- \frac{\Gamma(\mathbf{r})^2}{\frac{(MN)^2}{6} + \frac{MN}{2\sqrt{|S^*|}}\Gamma(\mathbf{r})} \right),$$

where we used $\sigma_{\mathbf{r}}^2 = ((MN)^2 - 1)/12 < (MN)^2/12$, $\mathbf{r}_{\max} = MN$ and $\bar{\mathbf{r}} = (MN + 1)/2$, so that $\mathbf{r}_{\max} - \bar{\mathbf{r}} < (MN)/2$. Here we use $\Gamma(\mathbf{r}) = \textsc{scan}(\mathbf{r})/\sqrt{|S^*|} - \sqrt{|S^*|}(MN + 1)/2$, where $|S^*| = mn$. The previous bounds can be directly computed when there are no ties in the ranks, and it is easy to verify that they also hold if ties are dealt with in any of the classical ways (assigning the average rank, randomly breaking ties, etc). As before, this is a result conditional on the observations $\mathbf{X} = \mathbf{x}$ and hence the ranks $\mathbf{R} = \mathbf{r}$. The next step is to remove this conditioning, which now amounts to controlling the term $\Gamma(\mathbf{R})$.

Let S^* denote the anomalous submatrix under the alternative and first assume that $\theta = \theta_\ddagger$ for all entries inside S^*. By our assumptions on the size of the anomalous submatrix, we assume $\theta_\ddagger \to 0$ as $M, N \to \infty$.

Since

$$\Gamma(\mathbf{R}) \geq Y_{S^*}(\mathbf{R})/\sqrt{|S^*|} - \sqrt{|S^*|}\frac{MN+1}{2},$$

we focus on obtaining a lower bound on $\tilde{Y}(\mathbf{R}) \coloneqq Y_{S^*}(\mathbf{R})/\sqrt{|S^*|}$ that applies with high probability.

Take any $v, w \in S^*$ distinct and define

$$\zeta_\theta = \mathbb{E}_\theta(R_v), \quad \sigma_\theta^2 = \mathrm{Var}_\theta(R_v), \quad \xi_\theta = \mathrm{Cov}_\theta(R_v, R_w).$$

Note that

$$\mathbb{E}_\theta[\tilde{Y}_{S^*}(\mathbf{R})] = \sqrt{|S|}\zeta_{\theta_\ddagger}, \quad \mathrm{Var}_\theta[\tilde{Y}_{S^*}(\mathbf{R})] = \sigma_{\theta_\ddagger}^2 + (|S^*|-1)\xi_{\theta_\ddagger},$$

and so, by Chebyshev's inequality,

$$\tilde{Y}_{S^*}(\mathbf{R}) = \sqrt{|S^*|}\zeta_{\theta_\ddagger} + O_P\left(\sqrt{\sigma_{\theta_\ddagger}^2 + |S^*|\xi_{\theta_\ddagger}}\right).$$

With the lemma 3 in [ACTW17], we can obtain

$$\zeta_{\theta_\ddagger} = \tfrac{MN+1}{2} + (MN - |S^*|)(\zeta_{\theta_\ddagger} - \tfrac{1}{2}), \quad \sigma_{\theta_\ddagger}^2 \leq 4(MN)^2, \quad \xi_{\theta_\ddagger} \leq 2MN.$$

which gives

$$\tilde{Y}_{S^*}(\mathbf{R}) - \sqrt{|S^*|}\tfrac{MN+1}{2} = \sqrt{|S^*|}(MN - |S^*|)(\zeta_{\theta_\ddagger} - \tfrac{1}{2}) + O_P(MN). \tag{2.34}$$

Use Lemma 3 from [ACTW17] again, we have

$$\zeta_{\theta_\ddagger} \geq \tfrac{1}{2} + \theta_\ddagger \Upsilon - O(\theta_\ddagger^2).$$

Plug in (2.34), noting that $\sqrt{|S|}\theta_\ddagger \to \infty$ in (2.9) and $|S^*| = mn = o(MN)$,

$$\tilde{Y}_{S^*}(R) - \sqrt{|S^*|}\tfrac{MN+1}{2} \geq (1+o(1))MN\sqrt{|S^*|}\theta_\ddagger \Upsilon \tag{2.35}$$

By the same argument of the previous proof,

$$\begin{aligned}\log(\mathfrak{P}(\mathbf{R})) &\leq \log(|\mathbb{S}_{m,n}|) - \frac{(1+o(1))(MN)^2|S^*|\theta_\ddagger^2\Upsilon^2}{\frac{(MN)^2}{6} + \frac{MN}{2\sqrt{|S^*|}}(1+o(1))MN\sqrt{|S^*|}\theta_\ddagger\Upsilon} \\ &= \log(|\mathbb{S}_{m,n}|) - \frac{(1+o(1))|S^*|\theta_\ddagger^2\Upsilon^2}{\frac{1}{6} + \frac{1}{2}(1+o(1))\theta_\ddagger\Upsilon}\end{aligned}$$

Assuming $\theta_\ddagger \to 0$ ensures the second term of the denominator is $o(1)$. Assume in (2.9),

$$\liminf \frac{mn\theta_\ddagger^2}{2(m\log(M/m) + n\log(N/n))} = \frac{1}{12\Upsilon^2}(1+\epsilon)$$

with some constant $\epsilon > 0$, eventually we will have

$$\log(\mathfrak{P}(\mathbf{R})) \leq \log(|\mathbb{S}|) - (1+\tfrac{2\epsilon}{3})\Lambda.$$

By the same argument on $|\mathbb{S}_{m,n}|$ such that $\log|\mathbb{S}_{m,n}| \leq (1+o(1))\Lambda$, eventually

$$\log|\mathbb{S}| \leq (1+\tfrac{\epsilon}{3})\Lambda,$$

so we can derive

$$\log(\mathfrak{P}(r)) \leq -\frac{\epsilon}{3}(m\log(\frac{M}{m}) + n\log(\frac{N}{n})).$$

We reach the conclusion that the rank test calibrated by permutation has power going to 1 as $M, N \to \infty$ when $\theta = \theta_\ddagger$ for all the entries inside the elevated submatrix.

Finally, to establish the result when $\theta_{ij} \geq \theta_\ddagger$, we use the fact that the permutation p-value under this case is stochastically dominated by the p-value used in the proof.

2.7.4 Proof of Theorem 4

We first show that the test has level α. To show this, we fix a pair of (m, n) and show that $\mathbb{P}(\mathfrak{P}_{m,n}(X) \geq \alpha) \leq \alpha$. By the standard argument on the level of Bonferroni test, we will finish the proof on the level. Assuming the null is true, $\mathbf{scan}_{m,n}(\pi(X))$ has the same distribution with $\mathbf{scan}_{m,n}(X)$ under either permutation methods. Therefore we define

$$T_k = \mathbf{scan}_{m,n}(\pi_k(X)), k \in [(MN)!],$$

and assume $T_{k_0} = \mathbf{scan}_{m,n}(X)$, then $\text{rank}(T_{k_0})$ is uniformly distributed on $[(MN)!]$ (if the ties are broken randomly). We have

$$\mathbb{P}(\mathfrak{P}_{m,n}(X) \leq \alpha) \leq \mathbb{P}(\text{rank}(T_{i_0}) \leq \lfloor \alpha(MN!) \rfloor) \leq \frac{\lfloor \alpha(MN!) \rfloor}{MN!} \leq \alpha.$$

Therefore all we need to show is that the p-value tends to zero under the alternative.

Bidimensional permutation

We refer to the proof in Theorem 2 and directly use the results there. We only consider the permutation p-value calibrated under the true anomaly size (m, n), which

denoted as $\mathfrak{P}_{m,n}(\mathbf{X})$. By the definition of $\mathfrak{P}(\mathbf{X})$, we have

$$\mathfrak{P}(\mathbf{X}) \leq MN\mathfrak{P}_{m,n}(\mathbf{X}).$$

By the inequality controlling $\mathfrak{P}_{m,n}(\mathbf{X})$, namely (2.31), we have

$$\log[MN\mathfrak{P}_{m,n}(\mathbf{X})] \leq \log(MN) - (\frac{\epsilon}{8+\epsilon})\Lambda.$$

Recall that $\Lambda = \left(m \log \frac{M}{m} + n \log \frac{N}{n}\right)$, it is easily to see that $\log(MN) = o(1)\Lambda$ with the help of (2.7). Therefore eventually

$$\log(MN) \leq \epsilon\Lambda/16$$

which implies

$$\log[MN\mathfrak{P}_{m,n}(\mathbf{X})] \leq \log(MN) - (\frac{\epsilon}{16+\epsilon})\Lambda.$$

which concludes the proof.

Unidimensional permutation

We also refer to the proof in Section 2.3. Under the unidimensional permutation, for the permutation p-value with (m, n) known, we have

$$\log \mathfrak{P}_{m,n}(\mathbf{X}) \leq -(1 + o_P(1))\delta \log(|\mathbb{S}_{mn}|).$$

Here δ is a positive constant related to the difference between both sides of (2.5). Note that $\log(MN) = o(1)\log(|\mathbb{S}_{mn}|)$ by (2.7), therefore directly,

$$\begin{aligned}\log \mathfrak{P}(\mathbf{X}) &\leq \log(MN) + \log \mathfrak{P}_{m,n}(\mathbf{X}) \\ &\leq -(1+o_P(1))\delta \log|\mathbb{S}_{mn}| + \log(MN) = -(1+o_P(1))\delta \log(|\mathbb{S}_{mn}|).\end{aligned}$$

2.7.5 Proof of Lemma 1

From (2.14),

$$c' - c = \sum_{i=0}^{\lfloor \log_2 c \rfloor - k} a_i(c) 2^i \leq \sum_{i=0}^{\lfloor \log_2 c \rfloor - k} 2^i \leq 2^{\lfloor \log_2 c \rfloor - k + 1}.$$

Observing that $c = 2^{\log_2 c}$, we have

$$\frac{c - c'}{c} \leq 2^i \leq 2^{\lfloor \log_2 c \rfloor - k + 1 - \log_2 c} \leq 2^{1-k}.$$

2.7.6 Proof of Theorem 5

We illustrate the bidimensional case here, since the unidimensional case is following the same proof strategy and basically a rework of the existing proof of Theorem 2 on the unidimensional case.

From Lemma 1, we may find $(m', n') \in S_k(M) \times S_l(N)$ such that $m' \leq m, n' \leq n$ and $m' = (1+o(1))m, n' = (1+o(1))n$. Now we consider performing permutation test on (m', n') and bound $\mathfrak{P}_{m',n'}$ by the same way in the previous proof of Theorem 4.

We rewrite (2.27) as follows, as well as multiplied by the Bonferroni correction

factor. The $\mathfrak{P}^A_{k,l}(\mathbf{X})$ is bounded by the following,

$$|S_k(M)||S_l(N)||\mathbb{S}_{m',n'}| \exp\left(- \frac{m'n'(\mathbf{scan}_{m',n'}(\mathbf{X})/m'n' - \bar{\mathbf{X}})^2}{2\sigma^2_{\mathbf{X}} + \frac{2}{3}(\mathbf{X}_{\max} - \bar{\mathbf{X}})(\mathbf{scan}_{m',n'}(\mathbf{X})/m'n' - \bar{\mathbf{X}})} \right).$$

Note that $|S_k(M)||S_l(N)| \leq MN$. Combined with $m' = (1 + o(1))m, n' = (1 + o(1))n$, all we need to verify is the following,

$$\frac{\mathbf{scan}_{m',n'}(\mathbf{X})}{m'n'} - \bar{\mathbf{X}} \geq \theta_{\ddagger}(1 + o_P(1)). \tag{2.36}$$

This is done by realizing that $\mathbf{scan}_{m',n'}(\mathbf{X}) \geq Y_{S'}(\mathbf{X})$, where $S' \subset S^*$ and have m' rows and n' columns. Since $m', n' \to \infty$, the same argument yields

$$\frac{\mathbf{scan}_{m',n'}(\mathbf{X})}{m'n'} \geq \theta_{\ddagger} + O_P(\frac{1}{\sqrt{m'n'}}) = \theta_{\ddagger} + O_P(\frac{1}{\sqrt{mn}}). \tag{2.37}$$

Everything else follows the argument of the proof of Theorem 2, and (2.36) will be verified, which leads to the following,

$$|\mathbb{S}_{m',n'}| \exp\left(- \frac{m'n'(\mathbf{scan}_{m',n'}(\mathbf{X})/m'n' - \bar{\mathbf{X}})^2}{2\sigma^2_{\mathbf{X}} + \frac{2}{3}(\mathbf{X}_{\max} - \bar{\mathbf{X}})(\mathbf{scan}_{m',n'}(\mathbf{X})/m'n' - \bar{\mathbf{X}})} \right) \leq -\epsilon\Lambda$$

for some positive constant ϵ. Now realize that

$$|S_k(M)||S_l(N)| \leq MN,$$

we have $\log(|S_k(M)||S_l(N)|) = o(1)\Lambda$. Therefore eventually $\log(|S_k(M)||S_l(N)|) \leq 0.5\epsilon\Lambda$ and this will conclude the proof.

2.8 Acknowledgment

Chapter 2, partially, is a version of the paper "Distribution-free Detection of a Submatrix", *Journal of Multivariate Analysis* 156 (2017): 29-38. The dissertation author is the corresponding author of this material.

Chapter 2, partially, is a version of the paper "Distribution-Free, Size Adaptive Submatrix Detection with Acceleration". The dissertation author is the principal investigator and corresponding author of this material. The manuscript is being prepared to be submitted to a major statistics journal.

Chapter 3

Size-adaptive Submatrix Localization

3.1 Introduction

In this chapter we consider the problem of localizing the anomalous submatrix inside a large data matrix. Observing a data matrix $\mathbf{X}$, the problem of submatrix localization, or bi-clustering, is to locate some submatrix, whose entries are 'elevated', or 'significant', compared to the entries outside the submatrix. Usually the index sets are of special interest, since the elevated entries may stand for some potential association or relationship between the index sets. Notice that unlike in the previous chapter, where we are mainly interested in distinguishing data sets containing such information from pure noise, here we are interested in finding the submatrix that contains information, and separate it from the data matrix.

3.1.1 Submatrix localization

As in the previous chapters, we denote M and N as the number of rows and columns of the observed data $\mathbf{X}$. We further assume that there is only one submatrix, of size $m^* \times n^*$, to be localized (notice the notation change), as in the simplest case. Furthermore we introduce the Gaussian assumption to the distribution of the entries, as well as homogeneity

both inside and outside the anomaly, namely:

$$X_{ij} = \theta 1_{\{(i,j)\in(I^*\times J^*)\}} + \varepsilon_{ij} \tag{3.1}$$

where ε_{ij} are IID following standard normal distribution, and $I^* \subset [M], J^* \subset [N]$ are index sets of the elevated submatrix with size (m^*, n^*) respectively. The parameter θ, which is assumed to be strictly positive in this case, quantifies the signal contained in the submatrix.

This parametric setup is used in [KBRS11] and [BIS15], where they focus on locating the submatrix by the scan statistic [BI13], with full knowledge of the anomaly size (m^*, n^*). In detail, the scan statistic, with (m^*, n^*) *known*, is defined as

$$\text{SCAN}_{m^*,n^*}(\mathbf{X}) = \max_{I\subset[M],|I|=m^*,J\subset[N],|J|=n^*} \sum_{(i,j)\in I\times J} X_{ij}. \tag{3.2}$$

Here we use the symbol $[M]$ to represent integer set $\{1, \ldots, M\}$, and function $|\cdot|$ represents the cardinality of a set. Localizing the submatrix by the scan statistic is simply returning the row and column indexes reaching the maximal, namely:

$$\Phi_{\text{SCAN}}(\mathbf{X}) = \{(I, J) : \sum_{(i,j)\in I\times J} X_{ij} = \text{SCAN}_{m^*,n^*}(\mathbf{X})\}$$

It is shown that this estimator has the 'minimax' property under Gaussian assumption:

Theorem 6 ([BIS15] Theorem 2.1, 2.2)**.** *Under* (3.1)*, assume that*

$$M, N, m^*, n^* \to \infty, \quad , \frac{\max(m^*, n^*)}{\min(M, N)} \to 0. \tag{3.3}$$

Denote

$$\theta_0 = \max\left\{\frac{\sqrt{2\log n^*} + \sqrt{2\log(N-n^*)}}{\sqrt{m^*}}, \frac{\sqrt{2\log m^*} + \sqrt{2\log(M-m^*)}}{\sqrt{n^*}}, \frac{\sqrt{2n^*\log(N/n^*) + 2m^*\log(M/m^*)}}{\sqrt{m^*n^*}}\right\}.$$

Then if

$$\liminf \theta/\theta_0 > 1,$$

the estimator $\Phi_{\text{SCAN}}(\mathbf{X})$ *is consistent, meaning that* $\mathbb{P}(\Phi_{\text{SCAN}}(\mathbf{X}) \neq I^* \times J^*) \to 0$.

Moreover, if

$$\limsup \theta/\theta_0 < 1,$$

there does not exist a uniformly consistent estimator that recovers $I^* \times J^*$.

Theorem 6 provides lower bound of the signal for the existence of highly probably exact estimators, and proved that the estimator based on scan statistic will reach the lower bound. In order to make use of the scan statistic, *exact prior knowledge* of the anomaly size (m, n) is critical.

We consider the case when the submatrix size (m, n) is *not known* and as part of the parameters to be estimated. Therefore another form of scan statistic, different from proposed in (3.2), is in need in order to recover the submatrix.

Contribution 5 (Multiscale scan statistic and its property)**.** We propose a multiscale scan statistic as an estimator under the parametric model defined in Section 3.1.1. We prove that under some regularity conditions, this statistic exactly recovers the submatrix with high probability, when the signal strength is similar to order described in Theorem 6.

Additionally, the theory is extended to the case when entries follow distributions from a one-parameter exponential family.

Unfortunately, the finding of such a scan statistic is proved to be NP-hard, which is from the fact that calculation of the scan statistic (3.2) is NP-hard ([CC00] Theorem 1). To tackle the problem of calculating the scan statistic defined in (3.2), [SWPN09] develops an iterative algorithm (LAS) for a efficiently approximation. Based on the LAS algorithm, we develop two algorithms in order to approach our proposed statistic:

Contribution 6. We develop iterative algorithms to solve the problem of finding the proposed multiscale scan statistic. One algorithm adapts the idea of [SWPN09] and is a hill-climbing type optimization algorithm. The other is a golden section search type optimization procedure. Both algorithms are multiscale and apply to the case when (m, n) are unknown, while providing good performance in the simulation study.

3.1.2 Content

The following of this chapter are arranged as follows. In Section 3.2 we introduce the multiscale scan statistic, and reveal the connection between the statistic and the likelihood ratio test under the Gaussian parametrization model, with its theoretical property proved in Section 3.4. Section 3.3 introduces the main algorithms approaching the statistic. A hill-climbing search algorithm and a golden section search algorithm are introduced in details. The chapter is finished by a simulation study in Section 3.5. The technical proofs go into the appendix.

3.2 The multiscale scan statistic

3.2.1 Normalized scan statistic

When (m^*, n^*) is unknown, the original scan statistic from (3.2) will change with different inputs of predetermined submatrix sizes. We would like to continue using the scan statistic framework, but a criteria for carrying out comparison between different scan statistic results under different submatrix sizes inputs is in need. For example, if two potential submatrix sizes (m_1, n_1) and (m_2, n_2) are investigated, how to determine which one of $(\text{SCAN}_{m_1,n_1}(\mathbf{X}), \text{SCAN}_{m_2,n_2}(\mathbf{X}))$ is more 'significant'? It is clear that one can not compare them directly with the value since it heavily favors the larger submatrix sizes. Also, it is worth mentioning that comparing the average of entries inside the submatrix returned by the scan statistic, namely, $\text{SCAN}_{m,n}(\mathbf{X})/(mn)$, is not proper since this criteria will mark the 1×1 submatrix containing $\max_{i,j} X_{ij}$ as the most significant.

Our proposed multiscale scan statistic is inspired by [ACTW17] where the authors aims to detect an interval with unusual large mean inside a sequence of observations, with interval length unknown. In their work the authors introduce a scan statistic, which is searching for the interval containing the largest normalized mean. Following this idea, we focus the normalized scan statistic as $\text{SCAN}_{m,n}(\mathbf{X})/\sqrt{mn}$. This statistic is also used in the testing procedure for detecting the existence of elevated submatrix by [BI13].

Another reason to focus on such normalization approach is that, as discussed in [ACG13], the scan statistic as the test statistic in [ACTW17] is asymptotically equivalent to a form of generalized likelihood ratio test. We illustrate the same idea here with Gaussian assumption, as specified previously in (3.1). Consider the hypothesis testing problem regarding data $\mathbf{X}$, where

$$H_0 : X_{ij} \text{ are iid } \mathcal{N}(0, 1)$$

versus

$$H_1 : \text{There exists nonempty } I^* \text{ and } J^* \text{ such that (3.1) holds.}$$

The likelihood ratio test statistic takes the form of rejecting H_0 when

$$\Lambda = \frac{\max_{\theta\in\mathbb{R},(I,J)\subset[M]\times[N]} \prod_{(i,j)\in I\times J} \phi(X_{ij}-\theta) \prod_{(i,j)\notin I\times J} \phi(X_{ij})}{\prod_{(i,j)\in[M]\times[N]} \phi(X_{ij})},$$

is large. With some algebra, it can be shown that the test is equivalent to rejecting H_0 when the following quantity is large:

$$\max_{I\subset[M],J\subset[N]} \frac{\sum_{(i,j)\in I\times J} X_{ij}}{\sqrt{|I||J|}},$$

Therefore, normalized scan statistics captures the deviation of a model parametrized by (3.1) from pure Gaussian noise. And naturally the submatrix responsible for generating such a deviation is the estimator of indexes for the elevated submatrix.

3.2.2 Comparing normalized scan statistics

When comparing two normalized scan statistic, especially when the sizes of the scan statistics are differing significantly, we need to consider the effect of background noise into the comparison. For example, consider two sets of sizes (m_1, n_1) and (m_2, n_2), and denote $\mathcal{S}_{m,n}$ as the collection of submatrices with size (m, n). If $|\mathcal{S}_{m_1,n_1}|$ is larger than $|\mathcal{S}_{m_2,n_2}|$ by a very large margin, and further assume all X_{ij} are iid standard normal (which means that there is no signal contained), it will be likely that $\text{SCAN}_{m_1,n_1}(\mathbf{X})/\sqrt{m_1 n_1} > \text{SCAN}_{m_2,n_2}(\mathbf{X})/\sqrt{m_2 n_2}$ since the former is taking maximal over more elements than the latter. This tells us that we need more adjusting elements to construct our multiscale scan statistic.

Inspired by the multiscale testing procedure in [BI13], we denote the following

quantity:

$$\lambda_{m,n} = \sqrt{2\log\left[MN\binom{M}{m}\binom{N}{n}\right]}, \tag{3.4}$$

and our multiscale scan statistic is defined as

$$\text{MSCAN}(\mathbf{X}) = \max_{I\subset[M],J\subset[N]} \left\{ \frac{\sum_{(i,j)\in I\times J} X_{ij}}{\sqrt{|I||J|}} - \lambda_{|I|,|J|} \right\}. \tag{3.5}$$

Naturally, the corresponding estimator for localizing the elevated submatrix will be

$$\Phi_{\text{MSCAN}}(\mathbf{X}) = \{(I,J) : \sum_{(i,j)\in I\times J} X_{ij} = \sqrt{|I||J|}(\text{MSCAN}(\mathbf{X}) + \lambda_{|I|,|J|})\}. \tag{3.6}$$

In the following chapters we will answer the two major problems centered at this proposed estimator: how to calculate the estimator provided data, and when will this estimator be successful for finding the true elevated submatrix.

3.3 Two iterative searching algorithm

The calculation of (3.6) requires operations exponentially increasing with (M, N) in the worst case, because to obtain the statistic, one may screen over all the submatrices of $\mathbf{X}$ and the totally number of submatrices is equal to 2^{M+N}. This fact makes the direct calculation of the multiscale scan statistic computationally intractable. Another method to illustrate the computational intractability of (3.6) is to realize that the computation of (3.6) is harder than that of (3.2), and the latter is proved to be NP-hard by [CC00]. Thus an fast and scalable algorithm approaching the estimator is needed in order to put the methodology into practice.

When the true anomaly size (m^*, n^*) is known, [SWPN09] proposed an iterative hill-climbing algorithm (LAS) to approach the statistic in (3.5).

Algorithm 1: Shabalin's Hill-climbing Algorithm (LAS)

Input: m^*, n^*, initial row indexes $\hat{I}$ with $|\hat{I}| = m^*$
Output: $\hat{I}$, $\hat{J}$

1 Calculate the sum $S_j = \sum_{i\in\hat{I}} X_{ij}$, and let $\hat{J}$ be the column index set corresponding to the largest n^* items in $\{S_j\}_{j=1}^N$
2 Calculate the sum $T_i = \sum_{j\in\hat{J}} X_{ij}$, and renew $\hat{I}$ be the row index set corresponding to the largest m^* items in $\{T_i\}_{i=1}^M$
3 Repeat *Steps 1 and 2*, till converge.

As a hill-climbing algorithm, Algorithm 1 may be trapped into local maximums therefore in practice people initiate the algorithm several times with random $\hat{I}$, and take the output with the largest entry sum, as [BI13] and [ACL17] did in their simulations.

3.3.1 Adaptive hill-climbing

We follow the idea of Algorithm 1. Each round we fixing row or column indexes while optimizing the other one. In each round of the iteration, we fix either row or column indexes of the temporary estimator, and renew the other by maximizing the multiscale scan statistic according to (3.5). Therefore each round keeps improving the target function denoted by (3.5).

Algorithm 2: Adaptive Hill-climbing Algorithm

Input: m, n, initial row indexes $\hat{I}_0$
Output: $\hat{I}$, $\hat{J}$

1 Run Algorithm 1 based on the initial input, obtaining $(\hat{I}, \hat{J})$
2 Calculate the sum $S_j = \sum_{i\in\hat{I}} X_{ij}$. Let $S_j^\dagger = \sum_{k=1}^j S_{(k)}/\sqrt{j} - \lambda_{|\hat{I}|,j}$ where $\{S_{(k)}\}_{k=1}^N$ are order statistics for $\{S_k\}_{k=1}^N$;
3 Renew n by $S_n^\dagger = \max_j S_j^\dagger$, and renew $\hat{J}$ by $\sum_{j\in\hat{J}} S_j = \sum_{k=1}^n S_{(k)}$
4 Calculate the sum $T_i = \sum_{j\in\hat{J}} X_{ij}$. Let $T_i^\dagger = \sum_{l=1}^i T_{(l)}/\sqrt{i} - \lambda_{i,|\hat{J}|}$ where $\{T_{(l)}\}_{l=1}^M$ are order statistics for $\{T_l\}_{l=1}^M$;
5 Renew m by $T_m^\dagger = \max_i T_i^\dagger$, and renew $\hat{I}$ by $\sum_{i\in\hat{I}} T_j = \sum_{l=1}^m T_{(l)}$;
6 Repeat *Steps 2-5*, till converge.

Note that the initial input m and n in Algorithm 2 does not necessarily represent

any prior knowledge of the anomaly size. One can start with a moderately small guess (e.g. $m = n = 5$) and the algorithm will renew the estimated anomaly size automatically. Also, by randomly initiating the algorithm several times and picking up the result with largest corresponding normalized entry mean, one can prevent the algorithm from being stuck at local maximals.

3.3.2 Golden section search

An alternative algorithm was inspired by the 2 dimensional golden section search on finding the maximal of a unimodal target function (see [Cha09] for a general survey of the algorithm in N dimensions). The idea is to set the target function f as the function of the submatrix size (m, n):

$$f_{\mathbf{X}}(m, n) = \frac{\mathrm{SCAN}_{m,n}(\mathbf{X})}{\sqrt{mn}} - \lambda_{m,n}, \tag{3.7}$$

and the multiscale scan statistic is the maximal of f:

$$\mathrm{MSCAN}(\mathbf{X}) = \max_{m,n} f_{\mathbf{X}}(m, n)$$

Note that the scan statistic part of (3.7) could be computed, or approximated, by Algorithm 1, with reasonable fast speed and good accuracy (just modify the output from index sets to the corresponding sum-of-entries in the algorithm). Therefore we may apply 2 dimensional golden section search on f and eventually we will land on the pair (m, n) maximizing (3.7), and Algorithm 1 will return the solution to (3.6).

It is worth noticing that the algorithm needs to stop the loop when the searching frame is less or equal to 3×3, otherwise the loop would not break because of the discrete nature in the index sets (the searching frame will not shrink). Therefore the algorithm

Algorithm 3: Golden Section Search Algorithm

Input: $\bar{m}$, $\bar{n}$, initial row indexes $\hat{I}$
Output: $\hat{I}$, $\hat{J}$

1 Calculate $f(s,t)$ by getting the value of $\frac{\sum_{(i,j)\in\hat{I}\times\hat{J}} X_{ij}}{\sqrt{st}} - \lambda_{s,t}$, where $(\hat{I},\hat{J})$ are the output of Algorithm 1 when the input submatrix size is (s,t);
2 Denote $(m_{\min}, n_{\min}) = (1,1)$, $(m_{\max}, n_{\max}) = (\bar{m},\bar{n})$;
3 Calculate $m_1 = \lceil m_{\max} + (m_{\min} - m_{\max})\phi \rceil$, $m_2 = \lfloor m_{\min} + (m_{\max} - m_{\min})\phi \rfloor$, $n_1 = \lceil n_{\max} + (n_{\min} - n_{\max})\phi \rceil$, $n_2 = \lfloor m_{\min} + (n_{\max} - n_{\min})\phi \rfloor$;
4 Calculate $f(m_1,n_1), f(m_2,n_1), f(m_1,n_2), f(m_2,n_2)$, denote $(m^{\#},n^{\#}) = \arg\max_{(i,j)\in\{1,2\}^2} f(m_i,n_j)$;
5 Denote $(m_*,n_*) = \arg\max_{(i,j)\in\{1,2\}^2} \|(m^{\#},n^{\#}) - (m_i,n_j)\|$, $(m^*,n^*) = \arg\min_{(k,l)\in\{\min,\max\}^2} \|(m^{\#},n^{\#}) - (m_k,n_l)\|$;
6 Renew $(m_{\max}, n_{\max}) = (\max(m_*,m^*), \max(n_*,n^*))$, $(m_{\min}, n_{\min}) = (\min(m_*,m^*), \min(n_*,n^*))$;
7 Repeat *Steps 3-6*, till $\max(m_{\max} - m_{\min}, n_{\max} - n_{\min}) \le 3$;
8 Calculate $f(s,t)$ for all $(s,t) \in [m_{\min}, m_{\max}] \times [m_{\min}, m_{\max}] \cap \mathbb{N}^2$. Denote the maximal among the results as $f(\hat{m},\hat{n})$. Finalize $(\hat{I},\hat{J})$ as the output of Algorithm 1 with input $(\hat{m},\hat{n})$.

switches to exhaust search after the searching frame is less or equal to 3×3, during which there are less than 9 extra search operations, which does not contribute significantly to the total running time of the algorithm.

Different from Algorithm 2, the initial submatrix size pair guess $(\bar{m},\bar{n})$ is an potential upper bound of the anomaly size, since the algorithm is searching for the anomaly with size less or equal to this pair. Possibly some prior knowledge will apply in the initialization of the algorithm, but again, it does not require any exact prior knowledge of the anomaly size.

A rationale behind the golden section search algorithm is that, we consider function f to be unimodal on $[\bar{m}] \times [\bar{n}]$, as well as that $f(m^*,n^*)$ is the global maximum in this region. While we are unable to give a rigorous condition for the signal strength to ensure the former statement, the latter statement could be ensured by the theorems in Section 3.4. We perform a simulation on the unimodal issue in Section 3.5.4, to illustrate the unimodality of (3.7) when the signal strength is beyond a proper threshold.

3.4 Theoretical property

3.4.1 Gaussian entries

We establish the minimal information bound, over which the estimator based on multiscale scan statistic will be equal to the true anomaly with high probability, with a constraint on the parameter space. The information contained in the anomaly is quantified with the value of θ. Recall that the true anomaly $I^* \times J^*$ has size (m^*, n^*), as defined in Section 3.2. We allow θ to change with (M, N, m^*, n^*).

We have the following theorem describing the information needed for the proposed estimator to succeed:

Theorem 7 (Exact recovery)**.** *Under the assumption of* (3.1) *and suppose* (3.3) *holds. Set a pair of positive constants* (τ_1, τ_2) *such that* $\tau_2 = 1 + \tau_1^{-1}$*. And define*

$$\theta_1 = \left\{ (1+\tau_1) \frac{\lambda_{m^*,n^*}}{\sqrt{m^* n^*}} \right\},$$

and

$$\theta_2 = \left\{ 2(\tau_2^{1.5} + \tau_2^{1.25}) \max\left[\sqrt{\frac{\log(M-m^*) + \log m^*}{n^*}}, \sqrt{\frac{\log(N-n^*) + \log n^*}{m^*}} \right] \right\}.$$

If

$$\liminf \frac{\theta}{\max(\theta_1, \theta_2)} > 1, \tag{3.8}$$

the estimator defined by (3.6) *is equal to* (I^*, J^*) *with high probability as* $(M, N, m^*, n^*) \to \infty$.

Remark 3. The θ_1 corresponds to the last component of θ_0 defined in Theorem 6. And θ_2 is corresponding to the rest of θ_0, which represent different requirements to the signal from row and column structures. We can see that with sparsity assumption (3.3), these signal

bounds essentially represents the same order of signal.

3.4.2 An extension to an exponential family

In practice the Gaussian assumption is not satisfied in many situations. Therefore it is of importance to extend the result to a distribution family that covers several common data types. We consider a one parameter exponential family to build up the parametric framework, as mentioned in Section 1.3. The localization problem can now be formalized with ν as the role of noise, and entries in the anomaly are distributed as f_θ. Formally,

$$X_{ij} \sim \begin{cases} f_\theta & (i,j) \in I^* \times J^* \\ \nu & (i,j) \notin I^* \times J^* \end{cases} \tag{3.9}$$

Note that distribution $f_0 = \nu$.

We may model the submatrix localization problem with θ controlling the signal strength. With θ increasing, the anomaly is more 'elevated', making the localization problem relatively easier to solve.

Here we build the similar minimax theory under the assumption of exponential family. This provide theoretical foundation for applying our proposed method under much wider data types.

Theorem 8 (Lower bound under exponential family assumption)**.** *Under the assumption of* (1.2) *and suppose* (3.3) *holds. Additionally, we assume*

$$\frac{\max(\log M, \log N)}{\min(m^*, n^*)} \to 0. \tag{3.10}$$

Then if

$$\theta = \sqrt{\alpha} \max\left(\sqrt{\frac{\log M}{n^*}}, \sqrt{\frac{\log N}{m^*}} \right).$$

with $\alpha \in (0, 1/8)$*, every estimator of* (I^*, J^*) *with full knowledge of* (m^*, n^*) *will fail to exactly recover* (I^*, J^*) *with probability bounded away from zero.*

This theorem is an extension of Theorem 1 in [KBRS11] where the authors considered Gaussian assumption instead of the exponential family. This result also partially answers the open problem raised by [BIS15] asking for a lower bound of signal strength under distribution of the exponential family.

Before we head into evaluating the performance of our proposed estimator (3.6) applied under the exponential family assumption, we need to re-define the adjusting quantity in (3.4).

Fixing a constant $\delta > 0$, we re-define

$$\lambda_{m,n} = \sqrt{(2+\delta)\log\left[MN\binom{M}{m}\binom{N}{n}\right]}. \tag{3.11}$$

The multiscale scan statistic, and its associated estimator of the elevated submatrix, is defined again according to (3.5) and (3.6).

Now we illustrate the minimal information needed for the estimator defined in (3.6) to succeed under the exponential family case:

Theorem 9 (Exact recovery under exponential family assumption)**.** *Under the assumption of* (1.2) *and suppose* (3.3) *and* (3.10) *holds. Furthermore define a sequence of integer sets* $\mathbb{D}(M, N) \subset [M] \times [N]$ *such that*

$$\limsup_{(m,n)\in\mathbb{D}(M,N)} \frac{\lambda_{m,n}}{\sqrt{mn}} \xrightarrow{M,N\to\infty} 0. \tag{3.12}$$

Additionally, assume $(m^*, n^*) \in \mathbb{D}(M, N)$.

Denote positive constant (τ_1, τ_2) *such that* $\tau_2 = 1 + \tau_1^{-1}$, *and again,*

$$\theta_1 = \left\{(1+\tau_1)\frac{\lambda_{m^*,n^*}}{\sqrt{m^* n^*}}\right\},$$

and

$$\theta_2 = \left\{2(\tau_2^{1.5} + \tau_2^{1.25}) \max\left[\sqrt{\frac{\log(M-m^*) + \log m^*}{n^*}}, \sqrt{\frac{\log(N-n^*) + \log n^*}{m^*}}\right]\right\}.$$

If

$$\liminf \frac{\theta}{\max(\theta_1, \theta_2)} > 1,$$

the estimator defined by

$$\left\{(I,J) : (|I|,|J|) \in \mathbb{D}(M,N); \sum_{(i,j)\in I\times J} X_{ij} = \sqrt{|I||J|}(\text{MSCAN}(\mathbf{X}) + \lambda_{|I|,|J|})\right\}.$$

is equal to (I^*, J^*) *with high probability as* $(M, N, m, n) \to \infty$.

Remark 4. $\mathbb{D}(M,N)$ and (3.10), (3.12) are requirements on the anomaly size which refrains it from being too small. Then Central Limit Theorem drives the asymptotic behavior of the statistic close to the normal case, which makes these two theorems similar.

3.5 Numerical experiments

We perform some simulation experiments to check the asymptotic performance of our proposed estimator, as well as evaluate the performance of the proposed two algorithms. The framework is to generate matrix with random entries according to (3.1) as well as (3.9), and perform the algorithms on the generated data. The elevated submatrix is set to be $[m^*] \times [n^*]$, and the evaluation criteria for a result $(\hat{I}, \hat{J})$ from algorithms is the

logarithm of cardinality of disjunctive union between the result and the truth, namely

$$\mathrm{Err}(\hat{I}, \hat{J}) = \log(|\hat{I} \triangle [m^*]| + |\hat{J} \triangle [n^*]| + 1), \tag{3.13}$$

where the operator $\triangle$ stand for the symmetric difference between sets.

We compare the proposed algorithms with two computationally tractable algorithms with proved consistency. One is spectral method [CLR+17], which uses singular value decomposition along with k-means clustering on the singular vectors. The other is greatest marginal gap method [BC16], which splits the indexes at the place when the largest drop of marginal sums occurs.

We take a balanced setup where $(M, N, m^*, n^*) = (1000, 1200, 170, 140)$, and an unbalanced setup where $(M, N, m^*, n^*) = (4000, 500, 70, 250)$. The latter is aiming to mimic a common case-control setup during biomedical study. For each fixed signal strength, the data is independently generated 30 times and error counts defined by (3.13) for all the algorithms are saved. Three data types (normal, Poisson, Rademacher) are investigated to illustrate the performance of algorithms under different types of data.

3.5.1 Signal strength

The signal strength is quantified by the parameter θ of the entries inside the submatrix. As discussed before, it represents how significant the submatrix is elevated from the noisy background. In order to capture the change point of algorithm performances during the increasing of signal strength, we denote the following quantity θ_{crit}:

$$\theta_{\mathrm{crit}} = \max\left(\sqrt{\frac{\log M}{n^*}}, \sqrt{\frac{\log N}{m^*}}\right).$$

This is part of the quantity at the left side of (3.8), inside Theorem 7. We zoom in to the interval $[1.0 \times \theta_{\text{crit}}, 4.0 \times \theta_{\text{crit}}]$. The signal strength is increased in the process of simulation, each time by $0.1\theta_{\text{crit}}$ in order to accurately sketch the change of (3.13) when the signal strength goes up.

3.5.2 Simulation result

Figure 3.1 demonstrates the result for the balanced case, and shows that both algorithms have there error counts converging to 0 when the signal strength is larger than $2\theta_{\text{crit}}$. Under the unbalanced case, Figure 3.2 exhibits a phase transition phenomenon for both algorithms, where the error counts sharply drop to 0 after signal strength goes beyond a certain bound. We discover that there is a potential performance difference between these two algorithms, however we should see that both algorithms would provide perfect submatrix recovery before the signal strength reaches $4\theta_{\text{crit}}$, and the multiscale scan statistic is actually recovering the submatrix perfectly when θ is beyond $1.5\theta_{\text{crit}}$ (see the result from golden section search).

It is worth mentioning that in both cases, our proposed methods outperform the other two computationally tractable methods. In the balanced case, both hill-climbing and golden section search beats the spectral method by a small margin, while the error of greatest marginal gap method does not converge in the region. In the imbalanced case, though the hill-climbing falls short comparing to the spectral method, golden section search method beats the other three, proving that the criteria we proposed has better performance.

In other data types, similar results are illustrated, except for the Rademacher imbalanced design where the hill-climbing algorithm's error counts fails to converge to zero. However in that case the golden section search manages to beat the other three algorithms in performance. See Figure 3.3 to Figure 3.6 for the simulation results.

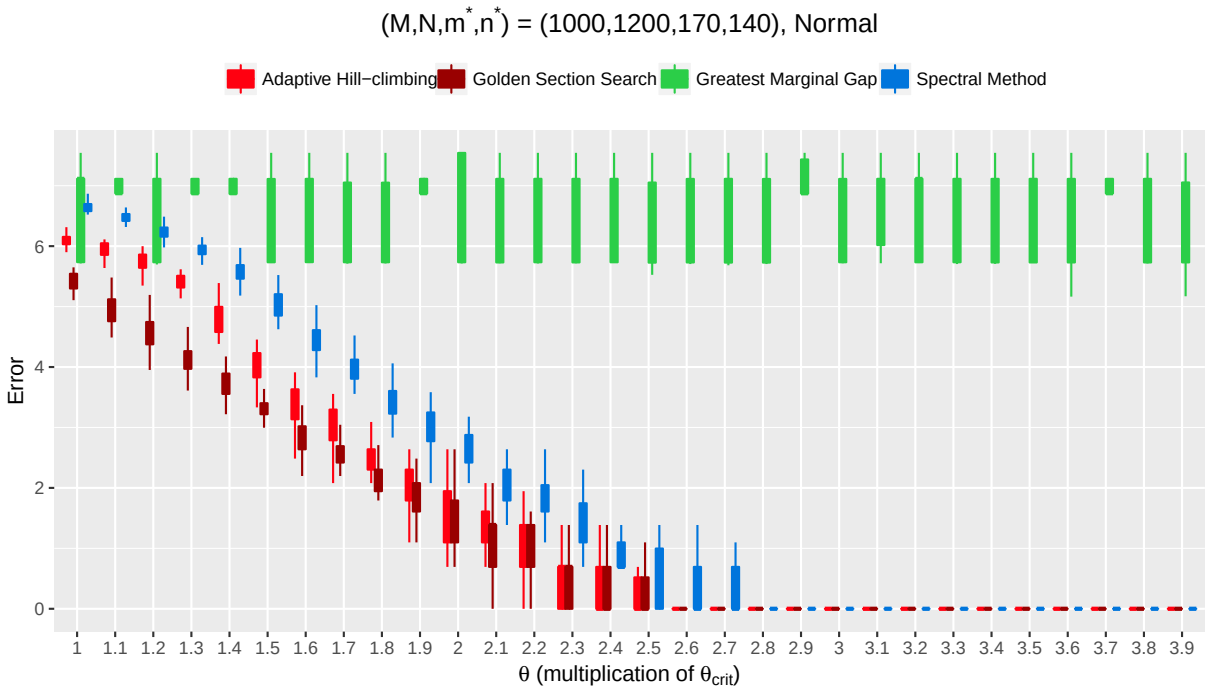

Figure 3.1: Error counts for a balanced design, Normal entries

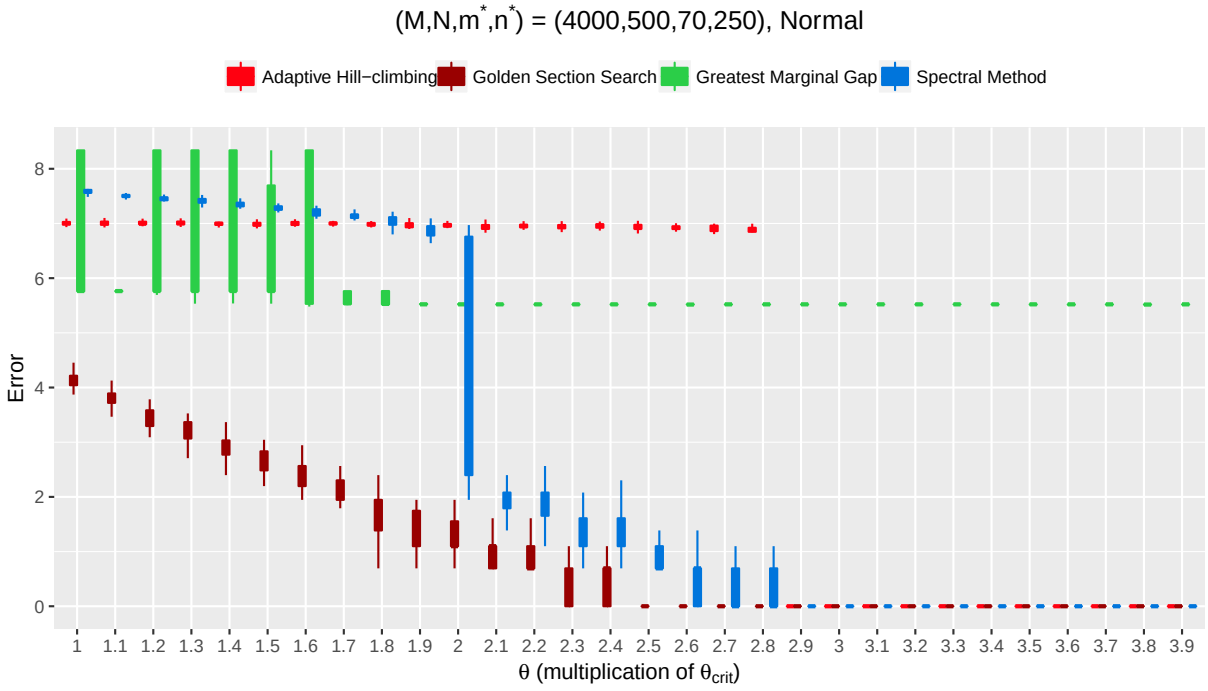

Figure 3.2: Error counts for an imbalanced design, Normal entries

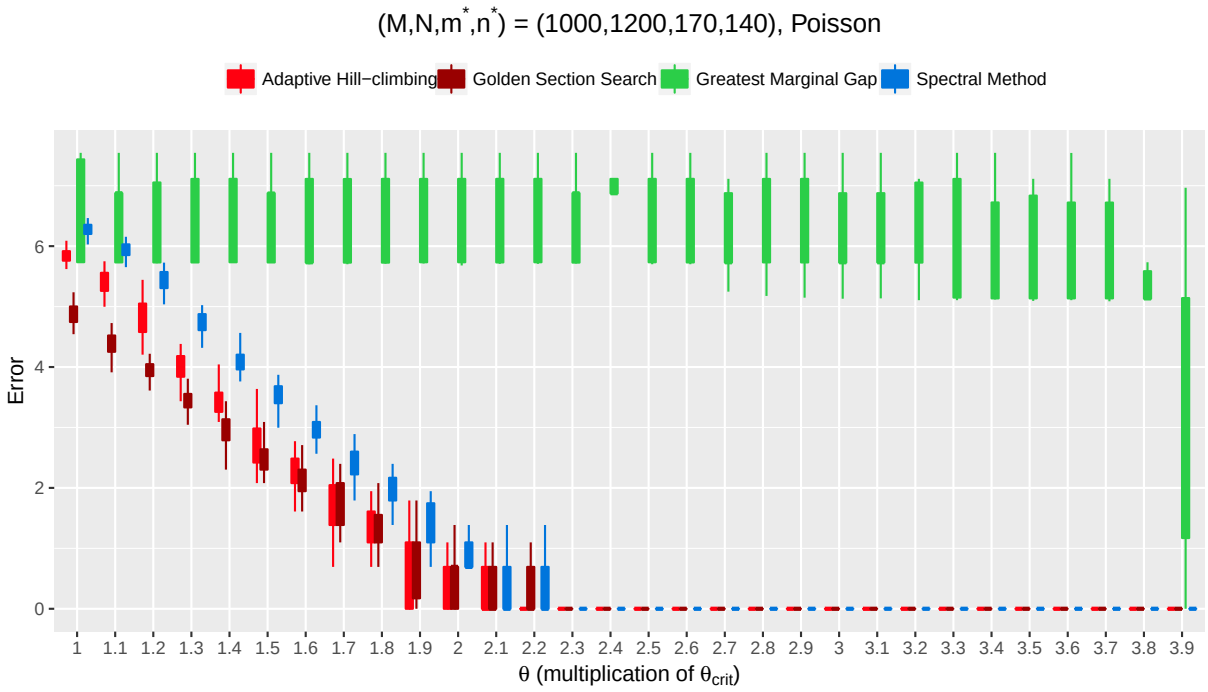

Figure 3.3: Error counts for a balanced design, Poisson entries

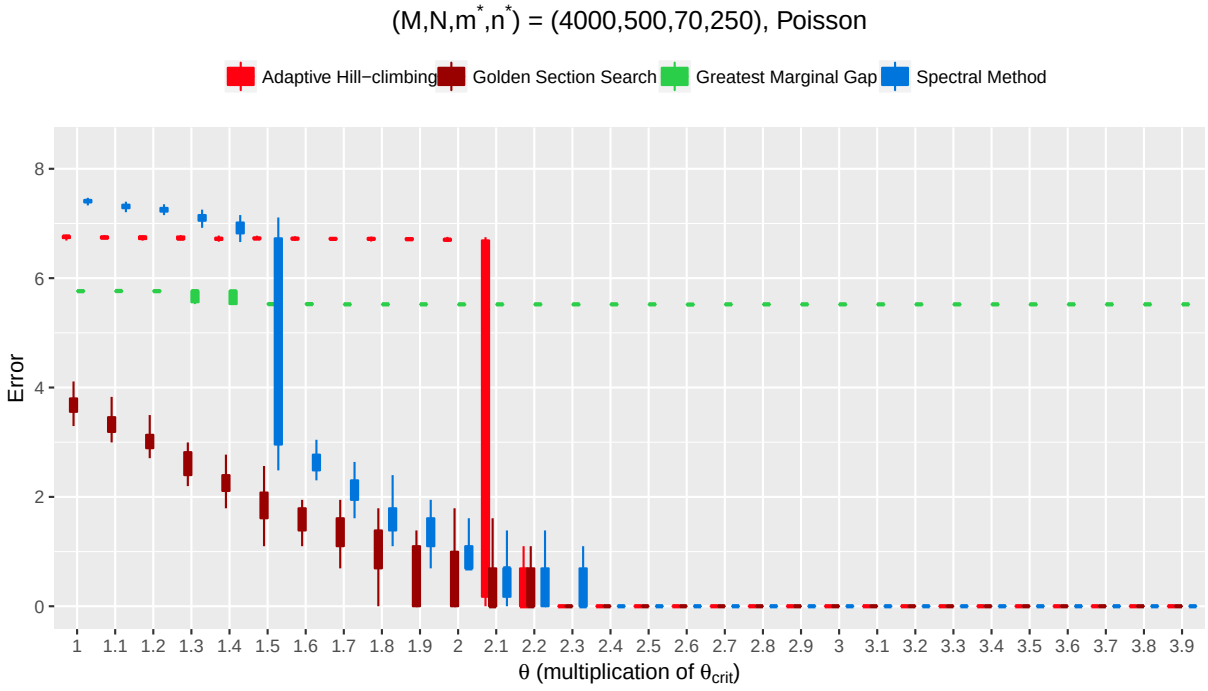

Figure 3.4: Error counts for an imbalanced design, Poisson entries

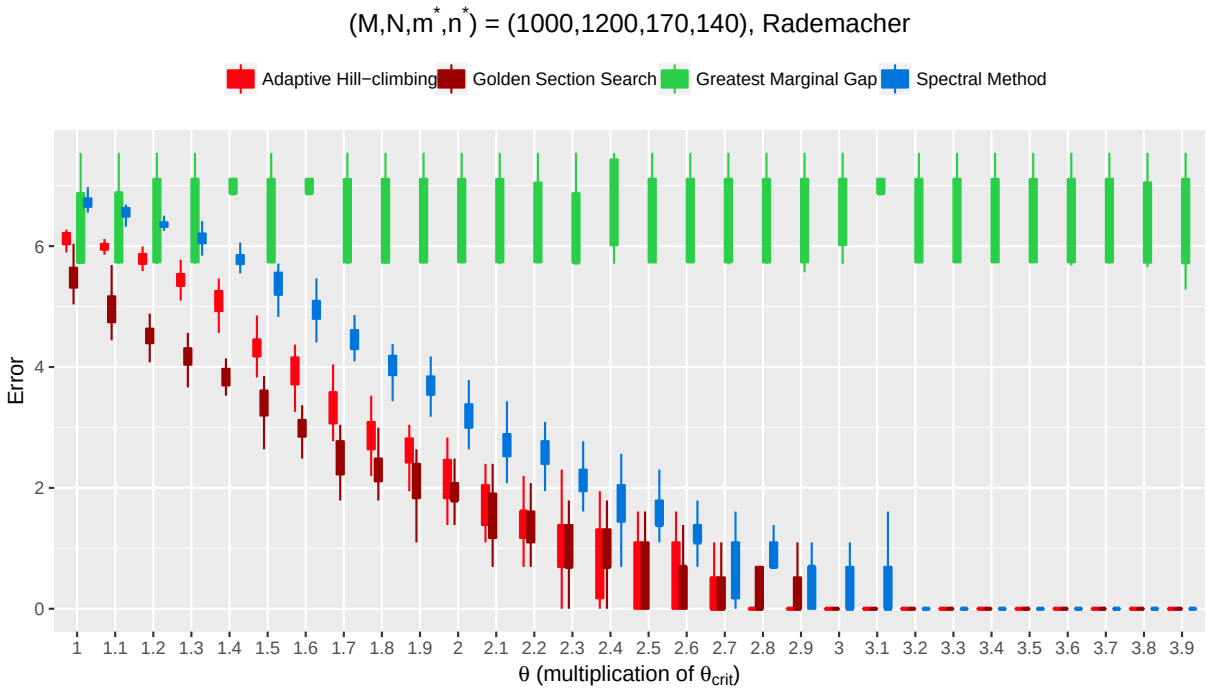

Figure 3.5: Error counts for a balanced design, Rademacher entries

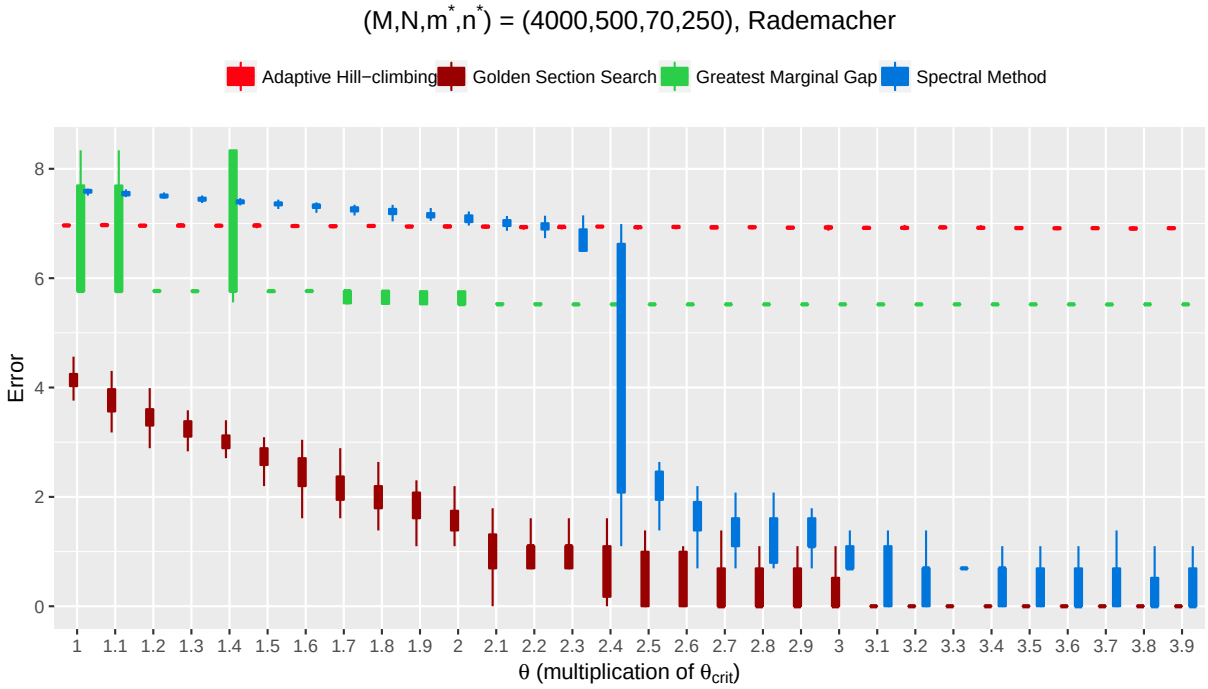

Figure 3.6: Error counts for an imbalanced design, Rademacher entries

3.5.3 Computing time

Here we present the computing time for executing the programs described in the previous section. The simulated data is from model (3.1), with corresponding sizes $(M, N, m^*, n^*) = (1000, 1000, 100, 100)$ and signal strength $\theta = 2.5\theta_{\text{crit}}$. Each data set generated goes through the four algorithms exhibited in the previous section, and the time consumed for executing the program is recorded. The procedure is repeated 100 times.

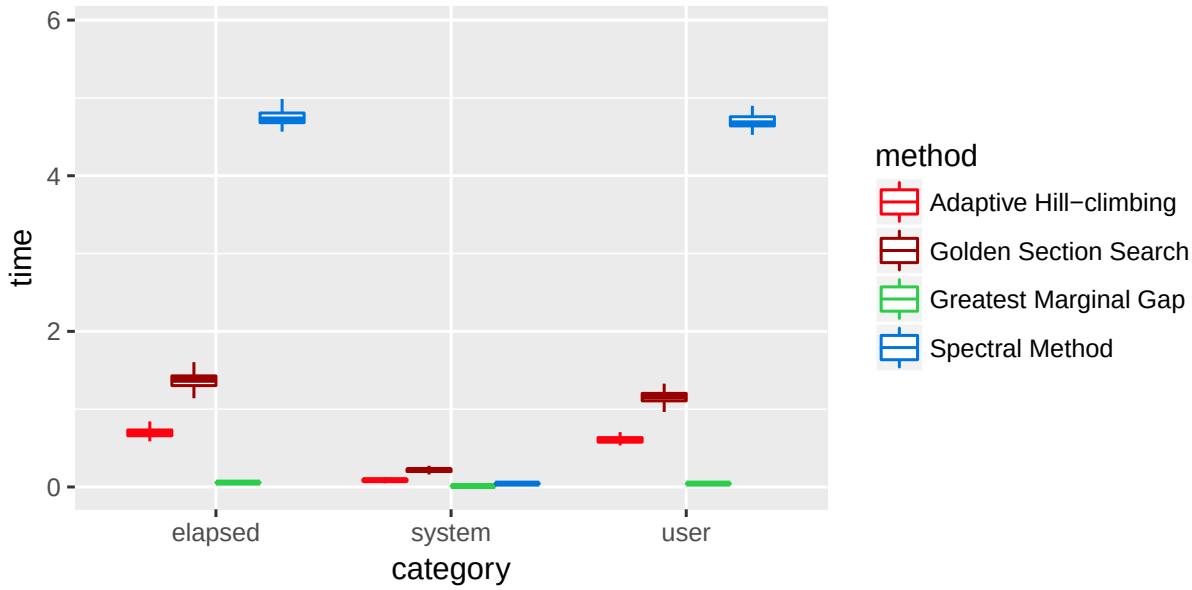

Figure 3.7: Computing times for different algorithms

The computing time is calculated by function 'proc.time' in programming language R. Category 'user' represents the time for executing the program codes, 'system' is the CPU time charged for execution by the system on behalf of the calling process, and 'elapsed' is the sum of the other two. As an algorithm linear in computing time [CLR+17], spectral method consumed significantly more time than the two iterative algorithms proposed. As expected, the clustering algorithm based on the largest marginal gap is the fastest, however it does not own promising clustering accuracy according to the previous simulation.

3.5.4 Unimodality issue in Algorithm 3

Golden section search presented in Algorithm 3 requires the function defined in (3.7) to be unimodal, in order to successfully discover the global maximal. In general, checking the unimodality of a function is often hard and here we present some numerical experiment for illustrating the unimodality of the target function defined in (3.7).

We use two simulated Gaussian data set to illustrate the result. For each combination of (M, N, m^*, n^*), set the level of signal as $\theta = 2\theta_{\text{crit}}$, which is just above the signal level such that the search algorithm makes few to zero mistakes. Then for every pair of (m, n) such that $m \le \bar{m}, n \le \bar{n}$, we calculate the function value according to (3.7). Due to the computing power constraint, the scan statistic $\text{SCAN}_{m,n}(\mathbf{X})$ is calculated by Algorithm 1, and $\lambda_{m,n}$ is approximated by

$$\hat{\lambda}_{m,n} = \sqrt{2(\log M + \log N + m\log(M/m) + n\log(N/n)}.$$

Here we exploit the fact that $\log\binom{K}{k} \approx k\log(K/k)$ when $k \ll K$. We follow the simulation ideas in the previous section, examining both balanced and imbalanced cases. In the balanced setup, $(M, N, m, n) = (300, 360, 40, 60)$ and $(\bar{m}, \bar{n}) = (100, 120)$. In the imbalanced setup, $(M, N, m, n) = (500, 50, 10, 25)$ and $(\bar{m}, \bar{n}) = (100, 50)$.

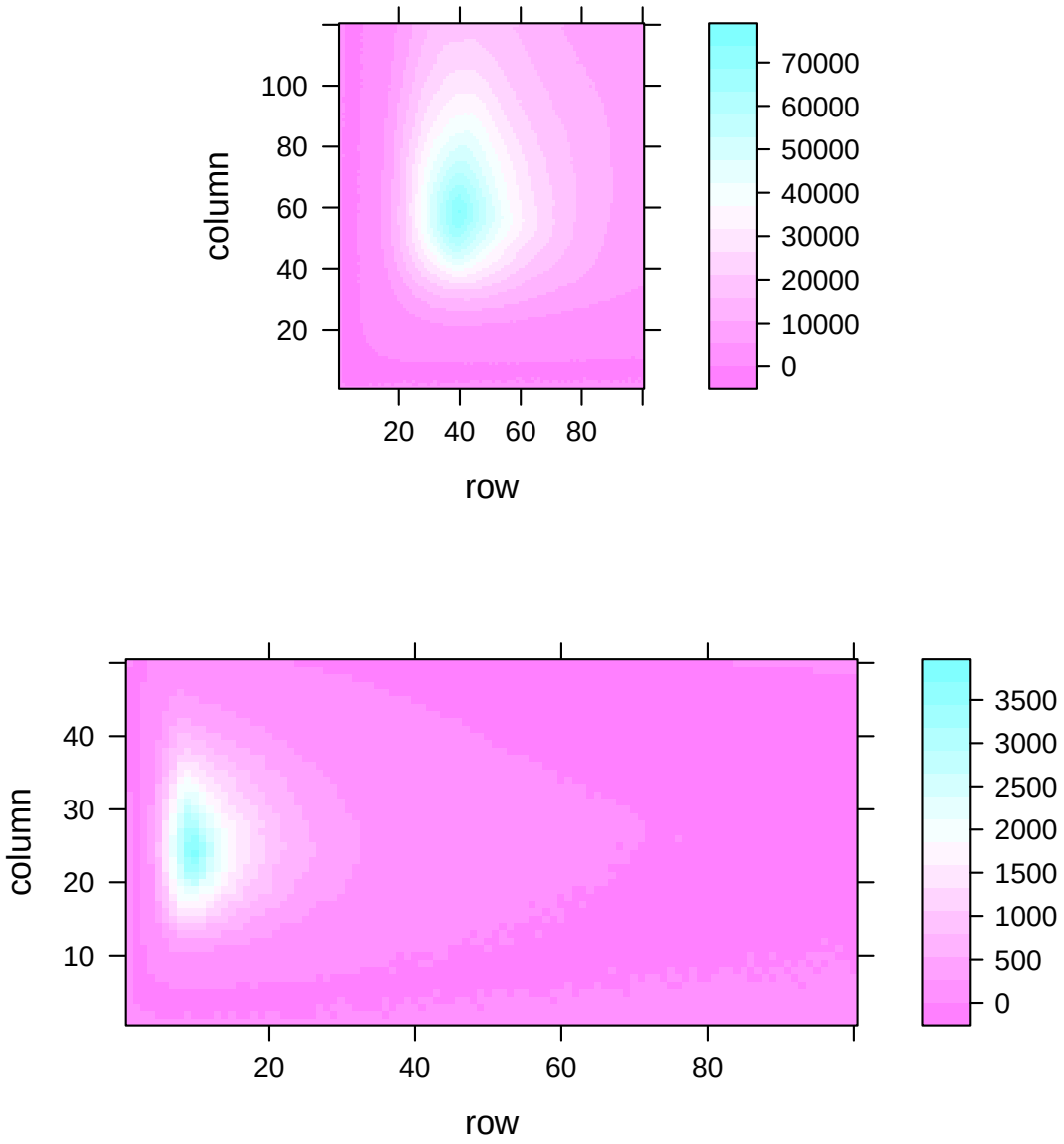

Figure 3.8: Levelplots for illustrating unimodality

To better illustrate the unimodality, the simulated data is raised to its fourth moment to enlarge the difference around the mode. We can see from Figure 3.8 that except a few points around the edge, the target function (3.7) has a clear unimodal structure on the majority of the search field $[\bar{m}] \times [\bar{n}]$.

3.6 Conclusion and discussion

In this chapter we propose a new multiscale scan statistic for localizing the elevated submatrix inside a large noisy matrix, which does not require prior knowledge on the elevated submatrix size. We show that our estimator successes will high probability with signal at the same order of the minimax bound, and design two algorithms with decent approximating accuracy and computing speed. There are, however, some problems for future work and discussions.

Minimaxity of the multiple scan statistic: [BIS15] showed a sharp minimax signal bound, as we described in Theorem 6. Can our estimator based on the multiscale scan statistic reach the minimax bound? While we are unable to reduce the constant in Theorem 7 to the bound, we conjecture that our estimator is essentially minimax.

Proof of unimodality of (3.7): What is the relationship of $(M, N, m^*, n^*, \bar{m}, \bar{n})$ and θ, such that (3.7) is unimodal on $[\bar{m}] \times [\bar{n}]$ with high probability? Solution to this problem will directly lead to the success guarantee of Algorithm 3.

Computationally tractable algorithms for the multiscale scan: There are literature about using computationally tractable methods such as semidefinite programming to solve NP-hard problems (for example, [CX16] on using SDP on finding the maximal likelihood estimator). Designing and analyzing such an algorithm for the multiscale scan statistic will be a interesting topic on the computational side.

Tight minimax bound for exponential family case: As [BIS15] mentioned, minimax theory on the case of exponential family is still open. We give a bound which is based on [KBRS11], which we believe is not tight enough. Also, although the scan statistic under the exponential family setup is proved to be minimax in testing by [BI13], the analysis of its localization performance is still open.

3.7 Proofs

3.7.1 Technical lemmas

In this section, we present several technical lemmas which are used in our proofs to the theorems.

Denote $\mathbf{E} = (E_{ij})$ as an $M \times N$ matrix with IID entries. For a submatrix S of E, define

$$Z_S = \frac{E_S}{\sqrt{|S|}}, \quad E_S = \sum_{(i,j)\in S} E_{ij}.$$

Let $\mathcal{S}_{m,n}$ be the class of subsets $S = I \times J \subset [M] \times [N]$ with $|I| = m$ and $|J| = n$. Consider the event

$$\mathcal{A}_{m,n} : \max_{S\in\mathcal{S}_{m,n}} Z_S \le \lambda_{m,n}, \tag{3.14}$$

where $\lambda_{m,n}$ are defined in (3.4).

Lemma 3. *If E_{ij} are IID standard normal, $\mathbb{P}(\bigcap_{m,n} \mathcal{A}_{m,n}) \to 1$ as $M \wedge N \to \infty$.*

Proof. It is not hard to see that for any S, Z_S has standard normal distribution. By a tail bound on the cumulative density function of standard normal distribution, namely

$$\Phi(-t) \le \frac{e^{-t^2/2}}{t}, \tag{3.15}$$

along with union bound, we have

$$\begin{aligned}
\mathbb{P}(\bigcup_{m,n} \mathcal{A}^c_{m,n}) &\le \sum_{m,n} \mathbb{P}(\mathcal{A}^c_{m,n}) \le \sum_{m,n} \sum_{S\in\mathcal{S}_{m,n}} \mathbb{P}(Z_S > \lambda_{m,n}) = \sum_{m,n} \sum_{S\in\mathcal{S}_{m,n}} \Phi(-\lambda_{m,n}) \\
&\le \sum_{m,n} \sum_{S\in\mathcal{S}_{m,n}} \frac{1}{\lambda_{m,n}} \exp\{-\lambda^2_{m,n}/2\} = \sum_{m,n} \sum_{S\in\mathcal{S}_{m,n}} \frac{1}{\lambda_{m,n}} \left(MN \binom{M}{m} \binom{N}{n} \right) \\
&= \frac{1}{\lambda_{m,n}} \to 0.
\end{aligned}$$

By taking complement we finish the proof of the lemma. □

Now we consider the situation where the entries of $\mathbf{E}$ are IID following ν defined in Section 1.3. Consider the event defined in (3.14) where $\lambda_{m,n}$ are defined in (3.11).

Lemma 4. *If E_{ij} are IID following ν, $\mathbb{P}(\bigcap_{(m,n)\in\mathbb{D}(M,N)} \mathcal{A}_{m,n}) \to 1$ as $M \wedge N \to \infty$, where $\mathbb{D}(M,N)$ is defined in* (3.12).

Proof. Consider a submatrix $S \in \mathcal{S}_{m,n}$, with $(m,n) \in \mathbb{D}(M,N)$. We first show that when E_{ij} are IID ν, $\mathbb{P}(Z_S > \lambda_{m,n}) \le \exp\{-(\lambda_{m,n})^2(1+o(1))/2\}$, and we can bring those $o(1)$ *uniformly* to 0 on all pairs of $(m,n) \in \mathbb{D}(M,N)$.

Note that

$$\mathbb{P}(Z_S > \lambda_{m,n}) = \mathbb{P}(E_S > \sqrt{mn}\lambda_{m,n}) = \mathbb{P}(e^{\zeta E_S} > e^{\zeta\sqrt{mn}\lambda_{m,n}}) \le \frac{(\mathbb{E}\, e^{\zeta E_{11}})^{mn}}{e^{\zeta\sqrt{mn}\lambda_{m,n}}},$$

where ζ is an arbitrary positive constant, and the last inequality is from Markov inequality.

We denote $\phi(\cdot)$ as the moment generating function of ν, as we did in Section 1.3. Taking natural log at both sides yields

$$\log \mathbb{P}(Z_S > \lambda_{m,n}) \le mn \log[\phi(\zeta)] - \zeta\sqrt{mn}\lambda_{m,n}.$$

Since ν has mean zero and variance 1, we do a Taylor expansion style decomposition, which yields $\log[\phi(x)] = x^2/2 + \psi(x^2)$, where when $x \to 0^+$, $\psi(x^2)/x^2 \to 0$.

Collecting all the components, and plugging in $\zeta = \lambda_{m,n}/\sqrt{mn}$, we have

$$\log \mathbb{P}(Z_S > \lambda_{m,n}) \le -\frac{(\lambda_{m,n})^2}{2}\left(1 + \frac{\psi((\lambda_{m,n})^2/(mn))}{(\lambda_{m,n})^2/(mn)}\right).$$

Now we see that from the definition of $\mathbb{D}(M,N)$, the latter part of the right hand side is uniformly $(1+o(1))$ for all pairs of (m,n) on $\mathbb{D}(M,N)$.

We apply the same argument in the proof of Lemma 3:

$$\begin{aligned}
\mathbb{P}\Big(\bigcap_{(m,n)\in\mathbb{D}(M,N)}(\mathcal{A}_{m,n})^c\Big) &\le \sum_{(m,n)\in\mathbb{D}(M,N)}\mathbb{P}((\mathcal{A}_{m,n})^c) \le \sum_{(m,n)\in\mathbb{D}(M,N)}\sum_{S\in\mathcal{S}_{m,n}}\mathbb{P}(Z_S > \lambda_{m,n}) \\
&\le \sum_{(m,n)\in\mathbb{D}(M,N)}\sum_{S\in\mathcal{S}_{m,n}}\exp\left\{-\frac{(\lambda_{m,n})^2}{2}(1+o(1))\right\} \\
&\le \sum_{(m,n)\in\mathbb{D}(M,N)}\sum_{S\in\mathcal{S}_{m,n}}\exp\left\{-(1+\delta/3)\log\left[MN\binom{M}{m}\binom{N}{n}\right]\right\} \\
&\le \frac{1}{(MN)^{\delta/3}} \to 0.
\end{aligned}$$

This finishes the proof. □

Lemma 5. *For any positive integer k and K such that $k < K$,*

$$\left(\frac{K}{k}\right)^k \le \binom{K}{k} \le \left(\frac{eK}{k}\right)^k$$

Some quick results from this Lemma are:

Corollary 1.

$$\log\binom{K}{k} \le \min(k, K-k)\log\left(\frac{K}{k}+1\right)$$

The proof is straight forward form the latter part of Lemma 5. And:

Corollary 2. *Fix K, $\binom{K}{k}/k$ is monotone decreasing in k.*

Proof. Pick an arbitrary k, the aim is to show

$$\frac{\log\binom{K}{k-1}}{k-1} \ge \frac{\log\binom{K}{k}}{k},$$

which is equivalent to

$$k \log \binom{K}{k} + k \log \left(\frac{K-k+1}{k} \right) \geq (k-1) \log \binom{K}{k}. \tag{3.16}$$

Here we used the fact $\binom{K}{k}(k/(K-k+1)) = \binom{K}{k-1}$. From Lemma 5,

$$\log \binom{K}{k} \geq k \log \left(\frac{K}{k} \right) \geq k \log \left(\frac{K-k+1}{k} \right),$$

which is enough to show (3.16). □

3.7.2 Proof of Theorem 7

Denote $S^* = I^* \times J^*$ as the index set of the true anomaly. For the submatrices in $\mathcal{S}_{m,n}$, define their affinity in size as

$$\rho_{m,n} = \max_{S \in \mathcal{S}_{m,n}} \frac{|S \cap S^*|}{\sqrt{|S||S^*|}} = \frac{(m \wedge m^*)}{\sqrt{mm^*}} \frac{(n \wedge n^*)}{\sqrt{nn^*}}.$$

For a specific submatrix S with size (m, n), such that $S \cap S^*$ is a $s \times t$ submatrix, define the affinity of coverage as

$$\upsilon_{m,n,s,t} = \frac{st}{\sqrt{mm^* nn^*}}.$$

The proof strategy is that, with high probability (converging to 1), we can sequentially rule out the submatrices with low affinity in size and coverage from the result of proposed estimator. Then we analyze those submatrices left behind, which have size and coverage comparable with the true anomaly.

Low affinity in size

Denote $\vartheta = \theta\sqrt{m^* n^*}$ and from (3.8) we know that $\vartheta \geq (1+\tau_1)\lambda_{m^*,n^*}$. Here we follow the style of notations in Section 3.7.1, for a submatrix $S \in \mathcal{S}_{m,n}$, defining

$$Y_S = \frac{X_S}{\sqrt{|S|}}, \quad X_S = \sum_{(i,j)\in S} X_{ij}.$$

From Lemma 3, assuming that $\bigcap_{m,n} \mathcal{A}_{m,n}$ holds, we have

$$\max_{S\in\mathcal{S}_{m,n}} Y_S - \lambda_{m,n} \leq \vartheta\rho_{m,n}.$$

On the other hand,

$$Y_{S^*} - \lambda_{S^*} = \vartheta - \lambda_{m^*,n^*} + O_p(1).$$

The latter is strictly larger than the former when

$$\vartheta - \lambda_{m^*,n^*} + O_p(1) - \vartheta\rho_{m,n} \geq [\tau_1 - (1+\tau_1)\rho_{m,n}]\lambda_{m^*,n^*} + O(1) > 0. \tag{3.17}$$

If $\rho_{m,n} < \tau_2^{-1}$, or $\tau_1 - (1+\tau_1)\rho_{m,n} > 0$, (3.17) will eventually hold with high probability (since $\lambda_{m^*,n^*} \to \infty$). Therefore we rule out all the submatrices belonging to $\mathcal{S}_{m,n}$ such that $\rho_{m,n} < \tau_2^{-1}$.

Low affinity in coverage

With a bit abuse of notation, we define

$$\mathcal{S}_{m,n,s,t} = \left\{S = I \times J \in \mathcal{S}_{m,n} : |I \cap I^*| = s \text{ and } |J \cap J^*| = t\right\},$$

and it is trivial to see that all $S \in \mathcal{S}_{m,n,s,t}$ have the same affinity in coverage. We aim to exclude those $\mathcal{S}_{m,n,s,t}$ such that $\upsilon_{m,n,s,t} < \tau_2^{-1}$.

In fact, this is done by the same argument of the previous section. From Lemma 3, assuming that $\bigcap_{m,n} \mathcal{A}_{m,n}$ holds, we have

$$\max_{S \in \mathcal{S}_{m,n,s,t}} Y_S - \lambda_{m,n} \leq \vartheta \upsilon_{m,n,s,t}.$$

and by replacing $\rho_{m,n}$ by $\upsilon_{m,n,s,t}$ in the previous section we can exclude submatrices in $\mathcal{S}_{m,n,s,t}$ such that $\upsilon_{m,n,s,t} < \tau_2^{-1}$ from the result of our proposed estimator with high probability.

One-point failure probability

Remember $S^* = I^* \times J^*$ as the true anomaly index set. For a candidate submatrix $S' = I' \times J'$, we make the following notation:

$$S_1 = S^* \cap S', \quad S_2 = S^* \setminus S', \quad S_3 = S' \setminus S^*,$$

and

$$m' = |I'|, \quad n' = |J'|, \quad s = |I^* \cap I'|, \quad t = |J^* \cap J'|.$$

Notice that, from the previous two sections, we have

$$\tau_2^{-1} m^* \leq m' \leq \tau_2 m^*, \quad \tau_2^{-1} m^* \leq m' \leq \tau_2 m^*, \quad \tau_2^{-1} n^* \leq n' \leq \tau_2 n^*, \quad \tau_2^{-1} n^* \leq t \leq \tau_2 n^* \quad (3.18)$$

We quantify the probability that S', instead of S^*, is returned by the estimator in

(3.6), namely

$$\begin{aligned}
&\mathbb{P}\left(Y_{S^*} - \lambda_{m^*,n^*} \le Y_{S'} - \lambda_{m',n'}\right) \\
&= \mathbb{P}\left(\sqrt{m'n'}(X_{S^*}) \le (\sqrt{m^*n^*}X_{S'}) + \sqrt{m^*n^*m'n'}(\lambda_{m^*,n^*} - \lambda_{m',n'})\right) \\
&= \mathbb{P}\left(\sqrt{m'n'}\big(X_{S_1} + X_{S_2}\big) \le \sqrt{m^*n^*}\big(X_{S_1} + X_{S_3}\big) + \sqrt{m^*n^*m'n'}(\lambda_{m^*,n^*} - \lambda_{m',n'})\right) \\
&= \mathbb{P}\left(\sqrt{m'n'}X_{S_2} + (\sqrt{m'n'} - \sqrt{m^*n^*})X_{S_1} - \sqrt{m^*n^*}X_{S_3} \le \sqrt{m^*n^*m'n'}(\lambda_{m^*,n^*} - \lambda_{m',n'})\right)
\end{aligned} \tag{3.19}$$

Under the assumption of (3.1), the last line of (3.19) is equivalent to

$$\mathbb{P}(Z(m',n',s,t) \le \sqrt{m^*n^*m'n'}(\lambda_{m^*,n^*} - \lambda_{m',n'})),$$

where $Z(m',n',s.t)$ is following normal distribution with mean $(\sqrt{m'n'}m^*n^* - \sqrt{m^*n^*}st)\theta$ and variance $2(m^*n^*m'n' - \sqrt{m^*n^*m'n'}st)$.

We claim that, for all pairs of (m',n') such that $\lambda_{m',n'} \le \lambda_{m^*,n^*}$, we have

$$\frac{\sqrt{m^*n^*m'n'}(\lambda_{m^*,n^*} - \lambda_{m',n'})}{(\sqrt{m'n'}m^*n^* - \sqrt{m^*n^*}st)\theta} \le \frac{1}{1+\tau_1}. \tag{3.20}$$

To see this, note that $\theta\sqrt{m^*n^*} \ge (1+\tau_1)\lambda_{m^*,n^*}$ from (3.8), and $(s,t) \le (m^* \wedge m', n^* \wedge n')$ by definition. Therefore it suffices to show the following:

$$\frac{1 - \lambda_{m',n'}/\lambda_{m^*,n^*}}{(1 - (m^* \wedge m')(n^* \wedge n')/\sqrt{m^*n^*m'n'})} \le 1.$$

or equivalently,

$$\frac{\lambda_{m',n'}}{\lambda_{m^*,n^*}} \ge \frac{(m^* \wedge m')(n^* \wedge n')}{\sqrt{m^*n^*m'n'}}. \tag{3.21}$$

Without loss of generality we focus on the case when $(m',n') \le (m^*,n^*)$ (to see this, assume

$m' > m$ and $n' \le m$ and consider $m'' \approx (m^*)^2/m'$. It provides the same right hand side but smaller left hand side). Now we may transform (3.21) to

$$\sqrt{\frac{2(\log(MN) + \log\binom{M}{m'} + \log\binom{N}{n'}}{m'n'}} \ge \sqrt{\frac{2(\log(MN) + \log\binom{M}{m^*} + \log\binom{N}{n^*}}{m^*n^*}}, \tag{3.22}$$

which holds due to Corollary 2. Now (3.20) directly leads to

$$\begin{aligned}\mathbb{P}(Z(m', n', s, t) &\le \sqrt{m^*n^*m'n'}(\lambda_{m^*,n^*} - \lambda_{m',n'})) \\ &\le \mathbb{P}\left(Z(m', n', s, t) \le \frac{1}{1+\tau_1}\,\mathbb{E}\,Z(m', n', s, t)\right).\end{aligned}$$

By a variation of (3.15) (namely: $\Phi(-t) \le e^{-t^2/2}$ with $t > 0$),

$$\mathbb{P}(Z(m', n', s, t) \le \sqrt{m^*n^*m'n'}(\lambda_{m^*,n^*} - \lambda_{m',n'})) \le \exp\left\{-\frac{m^*n^*\theta^2}{4\tau_2^2}\left(1 - \frac{st}{\sqrt{m^*n^*m'n'}}\right)\right\}. \tag{3.23}$$

A union bound approach

First from union bound, the probability of a submatrix in $\mathcal{S}_{m',n',s,t}$ being returned is bounded by

$$\binom{m^*}{s}\binom{M-m^*}{m'-s}\binom{n^*}{t}\binom{N-n^*}{n'-t}\exp\left\{-\frac{m^*n^*\theta^2}{4\tau_2^2}\left(1 - \frac{st}{\sqrt{m^*n^*m'n'}}\right)\right\},$$

which by Corollary 1, is further bounded from above by

$$\begin{aligned}q_{m'n'st} = \exp\Bigg\{&[(m^* - s) \wedge s]\log m^* + (m' - s)\log(M - m^*) + [(n^* - t) \wedge t]\log n^* \\ &+ (n' - t)\log(N - n^*) - \frac{m^*n^*\theta^2}{4\tau_2^2}\left(1 - \frac{st}{\sqrt{m^*n^*m'n'}}\right)\Bigg\}.\end{aligned}$$

We partition the parameter space by different values of (m', n', s, t) and apply a

union bound:

$$\mathbb{P}(S^* \text{ not returned}) \le C\Bigg[\underbrace{n^{*2}m^* \max_{(n',s,t)} q_{m^*n'st}}_{\text{(I)}} + \underbrace{m^{*2}n^* \max_{(m',s,t)} q_{m'n^*st}}_{\text{(II)}} + \underbrace{m^{*2}n^{*2} \max_{(m',n',s,t)} q_{m'n'st}}_{\text{(III)}}\Bigg].$$

Here C is a fixed constant associated with τ_2, but is not changing with M, N, m^*, n^*. And we have applied the result from (3.18) to obtain the coefficients. Part (I) and (II) are corresponding to the cases when $m' = m$ and $n' = n$, while (III) excludes the union of these two cases. The maximal is taken correspondingly. Note that we completely ignore the case when $m' = m$ and $n' = n$ since it was considered in [BIS15].

We show how to bound (I) and (III), and (II) could be bounded similarly with (I) by symmetry.

Part 1: In this part we show that (I) is bounded towards zero. First realize from (3.18) that

$$\frac{st}{\sqrt{m^{*2}n^*n'}} > \tau_2^{-1}. \tag{3.24}$$

Note that (3.24) infers $t/\sqrt{m^{*2}n^*n'} > 1/(\tau_2 s) \ge 1/(\tau_2 m^*)$, where the latter part of the inequalities is from the fact that $s \le m^*$ is always holding. Immediately we have

$$\frac{m^*n^*\theta^2}{4\tau_2^2}\frac{t}{\sqrt{m^{*2}n^*n'}} > \frac{n^*\theta^2}{4\tau_2^3} > \log m^* + \log(M - m^*).$$

The last part is from (3.8). This ensures $q_{m^*n'st}$ is monotone increasing with s therefore we set $s = m^*$ in the following argument.

We now take a closer look at

$$q_{m^*n'm^*t} = \exp\left\{[(n^* - t) \wedge t]\log n^* + (n' - t)\log(N - n^*) - \frac{m^*n^*\theta^2}{4\tau_2^2}\left(1 - \frac{t}{\sqrt{n^*n'}}\right)\right\}.$$

Since

$$\frac{m^*n^*\theta^2}{4\tau_2^2}\frac{1}{\sqrt{n^*n'}} = \frac{m^*\theta^2}{4\tau_2^2}\sqrt{\frac{n^*}{n'}} \geq \frac{m^*\theta^2}{4\tau_2^3} \geq \log n^* + \log(N-n^*),$$

the same argument on the monotonicity of s implies that $q_{m^*n'm^*t}$ is monotone increasing in t as well. Therefore we take $t = \min(n^*, n')$.

Case 1 ($n' < n^*, t = n'$): Under this, we may simplify

$$\begin{aligned} q_{m^*n'm^*n'} &= \exp\left\{[(n^*-n')\wedge n']\log n^* - \frac{mn\theta^2}{4\tau_2^2}\left(1-\sqrt{\frac{n'}{n}}\right)\right\} \\ &\leq \exp\left\{(n^*-n')\log n^* - \frac{mn\theta^2}{4\tau_2^2}\left(1-\sqrt{\frac{n'}{n}}\right)\right\}. \end{aligned}$$

Denote the right hand side as $Q_1(n')$. We compare $Q_1(n'+1)$ and $Q_1(n')$:

$$\begin{aligned} \frac{Q_1(n'+1)}{Q_1(n')} &= \exp\left\{-\log n^* + \frac{m^*n^*\theta^2}{\tau_2^2}\left(\sqrt{\frac{n'+1}{n^*}} - \sqrt{\frac{n'}{n^*}}\right)\right\} \\ &= \exp\left\{-\log n^* + \frac{m^*\theta^2}{4\tau_2^2}\cdot\sqrt{n^*}\left(\sqrt{n'+1}-\sqrt{n'}\right)\right\} \\ \text{(convexity of function } f(x)=\sqrt{x}) \qquad &\geq \exp\left\{-\log n^* + \frac{m^*\theta^2}{4\tau_2^2}\cdot\frac{\sqrt{n^*}}{2\sqrt{n'+1}}\right\} \\ \text{(from (3.8) and } n'+1\leq n) \qquad &\geq \exp\left\{-\log n^* + \frac{1}{2}(\log n^* + \log(N-n^*))\right\} \geq 1. \end{aligned} \tag{3.25}$$

Therefore $q_{m^*n'm^*n'}$ is monotone non-decreasing in n'. By setting $n' = n^* - 1$ we will bound

(I) with

$$
\begin{aligned}
n^{*2}m^*Q_1(n^*-1) &= \exp\left\{3\log n^* + \log m^* - \frac{m^*n^*\theta^2}{4\tau_2^2}\left(1-\sqrt{\frac{n^*-1}{n^*}}\right)\right\} \\
(n^* - \sqrt{n^*(n^*-1)} = 0.5 + o(1)) \qquad &= \exp\left\{3\log n^* + \log m^* - \frac{m^*\theta^2}{8\tau_2^2}(1+o(1))\right\} \\
(\text{from (3.8)}) \qquad &< \exp\left\{3\log n^* + \log m^* - 2\log n^* - 2\log(N-n^*)\right\} \to 0
\end{aligned}
$$

Note that for the last line we used the fact that (3.8) infers $m^*\theta^2 \geq 16\tau_2^2(\log n^* + \log(N-n^*))$ eventually. We will use this back and forth in the following text.

Case 2 ($n' > n^*, t = n^*$): The plugging-in procedure yields

$$
q_{m^*n'm^*n^*} = \exp\left\{(n'-n^*)\log(N-n^*) - \frac{m^*n^*\theta^2}{4\tau_2^2}\left(1-\sqrt{\frac{n^*}{n'}}\right)\right\}.
$$

Denote the right hand side as $Q_2(n')$. We see that the maximal of $Q_2(n')$ on the interval $[n^*+1, \tau_2 n^*]$ can only happen on the interval endpoints. First we set $n' = n^*+1$. Now look at

$$
\begin{aligned}
n^{*2}m^*Q(n^*+1) &= \exp\left\{2\log n^* + \log m^* + \log(N-n^*) - \frac{m^*n^*\theta^2}{4\tau_2^2}\left(1-\sqrt{\frac{n^*}{n^*+1}}\right)\right\} \\
&\leq \exp\left\{2\log n^* + \log m^* + \log(N-n^*) - \frac{m^*\theta^2}{8\tau_2^2}(1+o(1))\right\} \\
&< \exp\left\{3\log n^* + \log m^* - 2\log(N-n^*) - 2\log(n^*)\right\} \to 0 \qquad (3.26)
\end{aligned}
$$

The first inequality is from $n^* - \sqrt{n^*(n^*-1)} = 0.5 + o(1)$, and the second is from (3.8). Next we set $n = \tau_2 n^*$:

$$
n^{*2}m^*Q(\tau_2 n^*) = \exp\left\{(\tau_2-1)n^*\log(N-n^*)(1+o(1)) - \frac{m^*n^*\theta^2}{4\tau_2^2}\left(1-\sqrt{\frac{1}{\tau_2}}\right)\right\} \qquad (3.27)
$$

In order to send the right hand side to 0, we need

$$\liminf \frac{m^* n^* \theta^2 (1 - \sqrt{\tau_2^{-1}})/4\tau_2^2}{(\tau_2 - 1) n^* \log(N - n^*)} > 1,$$

or

$$\liminf \frac{m^* \theta^2 / 4\tau_2^2}{\log(N - n)} \geq \tau_2 + \sqrt{\tau_2}$$

which is ensured by (3.8).

By summing up all the results we finished bounding (I) towards 0, and by symmetry we can bound (II) as well.

Part 2: In the following part we bound (III) by the similar arguments as above. We start with

$$\frac{st}{\sqrt{m' m^* n' n^*}} > \tau_2^{-1}. \tag{3.28}$$

The first task is to show that $q_{m'n'st}$ is monotone non-decreasing for both s and t. This can be done by realizing that (3.28) implies

$$\begin{cases} \dfrac{m^* n^* \theta^2}{4\tau_2^2} \dfrac{t}{\sqrt{m^* n^* m' n'}} & \geq \dfrac{n^* \theta^2}{4\tau_2^3} \\ \dfrac{m^* n^* \theta^2}{4\tau_2^2} \dfrac{s}{\sqrt{m^* n^* m' n'}} & \geq \dfrac{m^* \theta^2}{4\tau_2^3}, \end{cases}$$

and with (3.8), eventually

$$\begin{cases} \dfrac{n^* \theta^2}{4\tau_2^3} & \geq \log m^* + \log(M - m^*) \\ \dfrac{m^* \theta^2}{4\tau_2^3} & \geq \log n^* + \log(N - n^*). \end{cases}$$

The monotonicity holds from the same argument in Case 2 of Part 1. Due to this, we set $(s, t) = (\min(m^*, m'), \min(n^*, n'))$. Naturally there are for cases inside the discussion,

namely:

$$\begin{cases} m' < m^*, & n' < n^*, & (s,t) = (m', n') \\ m' < m^*, & n' > n^*, & (s,t) = (m', n^*) \\ m' > m^*, & n' < n^*, & (s,t) = (m^*, n') \\ m' > m^*, & n' > n^*, & (s,t) = (m^*, n^*) \end{cases} \tag{3.29}$$

Note that we exclude all the cases with equal signs since they are included in the previous argument. We illustrate the proof under the first case here and the rest three are similar therefore omitted in the text. When $m' < m^*, n' < n^*$, we have $(s,t) = (m', n')$. And this leads our attention to

$$q_{m'n'm'n'} \le \exp\left\{(m^* - m')\log m^* + (n^* - n')\log n^* - \frac{\theta^2 m^* n^*}{4\tau_2^2}\left(1 - \sqrt{\frac{m'n'}{m^* n^*}}\right)\right\}.$$

Compare to the argument in (3.25), in order to show $q_{m'n'm'n'}$ is monotone non-decreasing in m', all we need to show is

$$\frac{m^* n^* \theta^2}{4\tau_2^2}\left(\sqrt{\frac{(m'+1)n'}{m^* n^*}} - \sqrt{\frac{m'n'}{m^* n^*}}\right) \ge \log m^*.$$

This can be achieved by

$$\frac{m^* n^* \theta^2}{4\tau_2^2}\left(\sqrt{\frac{(m'+1)n'}{m^* n^*}} - \sqrt{\frac{m'n'}{m^* n^*}}\right) \ge \frac{n^* \theta^2}{8\tau_2^2}\sqrt{\frac{n'}{n^*}\frac{m^*}{m'+1}} \ge \frac{n^* \theta^2}{8\tau_2^3} \ge \frac{1}{2}[\log m^* + \log(M - m^*)],$$

where the first inequality is from the convexity of function $f(x) = \sqrt{x}$, the second coming from the fact that $(s,t) = (m', n')$, therefore

$$\sqrt{\frac{n'}{n}\frac{m}{m'+1}} \ge \sqrt{\frac{n'}{n}\frac{m'}{m}} = \frac{st}{\sqrt{mm'nn'}} > \tau_2^{-1},$$

and the last coming form (3.8). Similarly, $q_{m'n'm'n'}$ is non-decreasing in n' as well. Therefore

in this case $(m', n') = (m-1, n-1)$. Similar arguments would shrink the search of maximal to the following four cases, according to (3.29), and we omit the process here:

$$\begin{cases} m' < m^*, \quad n' < n^* & \Rightarrow \quad (m', n') = (m^* - 1, n^* - 1) \\ m' < m^*, \quad n' > n^* & \Rightarrow \quad (m', n') = (m^* - 1, n^* + 1) \quad \text{or} \quad (m^* - 1, \tau_2 n^*) \\ m' > m^*, \quad n' < n^* & \Rightarrow \quad (m', n') = (m^* + 1, n^* - 1) \quad \text{or} \quad (\tau_2 m^*, n^* - 1) \\ m' > m^*, \quad n' > n^* & \Rightarrow \quad (m', n') = (m^* + 1 \text{ or } \tau_2 m^*, n^* + 1 \text{ or } \tau_2 n^*). \end{cases} \tag{3.30}$$

The left hand side is the regulations for (m', n') and the right hand side is the value(s) of (m', n') that achieves the maximal. For multiple possibilities of (m', n'), we evaluate all the choices and show that all of them will bring (III) to converge to zero.

Now, what we need to check is that $m^{*2} n^{*2} q_{m'n'st} \to 0$ under all the four cases with (3.29) and (3.30). Since the process of checking is similar, we display one under the last case here and the other three cases can be similarly extended. We display the last case here, which includes four further cases.

Under $(m', n') = (m^* + 1, n^* + 1)$, we have

$$m^{*2} n^{*2} q_{(m^*+1)(n^*+1)m^* n^*} = \exp\Bigg\{ 2(\log m^* + \log n^*) + \log(N - n^*) + \log(M - m^*) \\ - \frac{m^* n^* \theta^2}{4\tau_2^2} \left(1 - \sqrt{\frac{m^* n^*}{(m^* + 1)(n^* + 1)}} \right) \Bigg\}$$

whose right hand side is bounded by

$$\begin{aligned}
&\exp\Bigg\{2(\log m^* + \log n^*) + \log(N-n^*) + \log(M-m^*) \\
&\quad - \frac{m^*n^*\theta^2}{4\tau_2^2}\left(\frac{(m^*+n^*+1)/(m^*+1)(n^*+1)}{1+\sqrt{m^*n^*/(m^*+1)(n^*+1)}}\right)\Bigg\} \\
&\quad \le \exp\left\{2(\log m^* + \log n^*) + \log(N-n^*) + \log(M-m^*) - (1+o(1))\frac{(m^*+n^*)\theta^2}{8\tau_2^2}\right\} \\
&\quad \le \exp\left\{-(1+o(1))(\log(M-n^*) + \log(N-n^*))\right\} \to 0,
\end{aligned}$$

where the last line is from (3.8).

Under $(m', n') = (\tau_2 m^*, \tau_2 n^*)$, we have

$$\begin{aligned}
m^{*2}n^{*2}q_{(\tau_2 m^*)(\tau_2 n^*)m^*n^*} = \exp\Bigg\{&(1+o(1))(\tau_2-1)[n^*\log(N-n^*) + m^*\log(M-m^*)] \\
&- \frac{m^*n^*\theta^2}{4\tau_2^2}\left(1-\tau_2^{-1}\right)\Bigg\}. \qquad (3.31)
\end{aligned}$$

Note that by (3.8)

$$\frac{m^*n^*\theta^2}{4\tau_2^3} = \frac{m^*n^*\theta^2}{8\tau_2^3} + \frac{m^*n^*\theta^2}{8\tau_2^3} \ge 2[m^*\log(N-n^*) + n^*\log(M-m^*)],$$

the right hand side of (3.31) is bounded by

$$\exp\left\{-(1+o(1))(\tau_2-1)[m^*\log(N-n^*) + n^*\log(M-m^*)]\right\}.$$

which converges to 0.

Under $(m', n') = (\tau_2 m^*, n^* + 1)$, we have

$$m^{*2} n^{*2} q_{(\tau_2 m^*)(n^*+1)m^* n^*} = \exp\Bigg\{(1+o(1))(\tau_2 - 1)m^* \log(M - m^*) + 2\log n^* + \log(N - n^*) - \frac{m^* n^* \theta^2}{4\tau_2^2}\left(1 - \sqrt{\frac{n^*}{\tau_2(n^*+1)}}\right)\Bigg\}. \quad (3.32)$$

with a proper chosen ϵ, which converge to zero eventually, we may have

$$\left(1 - \sqrt{\frac{n^*}{\tau_2(n^*+1)}}\right)(1+\epsilon) \geq 1 - \sqrt{\frac{1}{\tau_2}} + 1 - \sqrt{\frac{n^*}{(n^*+1)}}.$$

So we can bound the right hand side of (3.32) with

$$\underbrace{\exp\left\{(1+o(1))\left[(\tau_2 - 1)m^* \log(M - m^*) + 2\log n^* - \frac{m^* n^* \theta^2}{4\tau_2^2}\left(1 - \sqrt{\frac{1}{\tau_2}}\right)\right]\right\}}_{\text{(A)}}$$
$$\times \underbrace{\exp\left\{(1+o(1))\left[\log(N - n^*) - \frac{m^* n^* \theta^2}{4\tau_2^2}\left(1 - \sqrt{\frac{n^*}{(n^*+1)}}\right)\right]\right\}}_{\text{(B)}}$$

Note that $\log(n^*) \leq \log(M - m^*)$ by (3.3), so $(\tau_2 - 1)m^* \log(M - m^*) + 2\log n^* = (1 + o(1))(\tau_2 - 1)m^* \log(M - m^*)$. Then we can bound (A) by the techniques in (3.27) and (B) by (3.26). This case is finished. The case left over is checked by symmetry.

By checking all four cases we manage to control (III) towards 0, which completes the proof.

3.7.3 Proof of Theorem 8

Without loss of generality, assume $S^* = [m^*] \times [n^*]$. Denote the joint distribution of the entries under this assumption as $\mathcal{P}_0$. Correspondingly denote $\mathcal{P}_j$ as the joint distribution of the entries when $S^* = [m^*] \times \{[n^* - 1] \cup \{j\}\}$, with $j \in \mathbb{N} \cap [n^* + 1, N]$. The proof is

complete by following the proof of Theorem 1 in [KBRS11], if the following holds,

$$\mathcal{D}(\mathcal{P}_0||\mathcal{P}_j) \geq m^*\theta^2(1+o(1)),$$

where $\mathcal{D}(\cdot||\cdot)$ is the Kullback Leibler divergence between two distributions. In fact, with some direct calculation,

$$\begin{aligned}
\mathcal{D}(\mathcal{P}_0||\mathcal{P}_j) &= \sum_{i\in[m^*]} \int \log(e^{\theta x_{in^*}-\log(\varphi(\theta))}) e^{\theta x_{in^*}-\log(\varphi(\theta))}\nu(dx_{in^*}) \\
&\quad + \sum_{i\in[m^*]} \int \log(e^{-\theta x_{ij}+\log(\varphi(\theta))})\nu(dx_{ij}) \\
&= \sum_{i\in[m^*]} \int (\theta x_{in^*} - \log(\varphi(\theta))) e^{\theta x_{in^*}-\log(\varphi(\theta))}\nu(dx_{in^*}) \\
&\quad + \sum_{i\in[m^*]} \int (-\theta x_{ij} + \log(\varphi(\theta)))\nu(dx_{ij}) \\
&\overset{\text{(i)}}{=} \sum_{i\in[m^*]} x_{in^*}\theta e^{\theta x_{in^*}-\log(\varphi(\theta))}\nu(dx_{in^*}) \\
&\overset{\text{(ii)}}{\geq} \sum_{i\in[m^*]} \theta^2(1+o(1)) = m^*\theta^2(1+o(1)).
\end{aligned}$$

where (i) uses the fact that ν has mean 0, and (ii) follows from the fact that $\mathbb{E}(X) \geq \theta(1+o(1))$ if $X \sim f_\theta$ with $\theta \to 0$ (see the arguments under (16) of [ACTW17] and (7) of [ACL17]). Here $\theta \to 0$ is ensured by (3.10). Replace (15) of the proof of Theorem 1 in [KBRS11], and the result follows.

3.7.4 Proof of Theorem 9

Proof. The proof is essentially following the proof for Theorem 7, except that we need to check (i) Lemma 3 still holds under the exponential family assumption; (ii) (3.23) holds eventually. (i) can be done by Lemma 4, therefore we focus on (ii) here.

We use the same notations and representations as the previous proof of Theorem 7,

evaluating the following quantity

$$\mathbb{P}\left(Y_{S^*} - \lambda_{m^*,n^*} \le Y_{S'} - \lambda_{m',n'}\right). \tag{3.33}$$

Note that $X_{ij} \sim f_\theta$ when $(i,j) \in (I^* \times J^*)$, and $X_{ij} \sim \nu$ otherwise. Furthermore we assume that

$$\theta = C_e \theta_0.$$

with a constant $C_e > 1$. For θ larger than quantities of this order, the stochastic monotone property from (1.3) ensures the recovery becoming easier. Note that in this case $\theta \to 0$ as $M, N \to \infty$. By the same decomposition of (3.19), we rewrite, and bound (3.33) as

$$\begin{aligned}
&\mathbb{P}\left(\sqrt{m'n'}(X_{S^*}) \le (\sqrt{m^*n^*}X_{S'}) + \sqrt{m^*n^*m'n'}(\lambda_{m^*,n^*} - \lambda_{m',n'})\right) \\
&= \mathbb{P}\left(\exp\left\{-\sqrt{m'n'}\zeta X_{S_2}\right\}\exp\left\{-(\sqrt{m'n'} - \sqrt{mn})\zeta X_{S_1}\right\}\exp\left\{\sqrt{mn}\zeta X_{S_3}\right\} \ge \right. \\
&\qquad\qquad \left. \exp\left\{-\tau\sqrt{m^*n^*m'n'}(\lambda_{m^*,n^*} - \lambda_{m',n'})\right\}\right) \\
&\overset{\text{(i)}}{\le} \mathbb{E}\left(\exp\left\{-\sqrt{m'n'}\zeta X_{S_2}\right\}\right)\mathbb{E}\left(\exp\left\{-(\sqrt{m'n'} - \sqrt{mn})\zeta X_{S_1}\right\}\right)\mathbb{E}\left(\exp\left\{\sqrt{mn}\zeta X_{S_3}\right\}\right) \\
&\qquad\qquad \exp\left\{\zeta\sqrt{m^*n^*m'n'}(\lambda_{m^*,n^*} - \lambda_{m',n'})\right\} \\
&\overset{\text{(ii)}}{\le} \left[\mathbb{E}\left(\exp\{-\sqrt{m'n'}\zeta X_\theta\}\right)\right]^{m^*n^* - st}\left[\mathbb{E}\left(\exp\{-(\sqrt{m'n'} - \sqrt{m^*n^*})\zeta X_\theta\}\right)\right]^{st} \\
&\qquad \left[\mathbb{E}\left(\exp\{\sqrt{m^*n^*}\zeta X_0\}\right)\right]^{m'n' - st}\exp\left\{\frac{\zeta(\sqrt{m'n'}m^*n^* - \sqrt{m^*n^*}st)\theta)}{1+\tau_1}\right\}.
\end{aligned} \tag{3.34}$$

Here $\zeta > 0$ is a constant, $X_\theta \sim f_\theta$ and $X_0 \sim \nu$. Meanwhile, (i) is from Markov inequality and (ii) is from independence among the entries, as well as (3.20).

The following fact enables us to further simplify the above expression, with a little abusing of notations:

$$\mathbb{E}\left(\exp\{-\zeta X_\theta\}\right) = \int e^{-\zeta x} e^{\theta x - \log\varphi(\theta)} \nu(dx) = \frac{\varphi(\theta-\zeta)}{\varphi(\theta)},$$

with $\zeta \in (0,\theta)$. Therefore, if $(\sqrt{m'n'}\zeta, (\sqrt{m'n'} - \sqrt{m^*n^*})\zeta) < \theta$ and $\sqrt{m^*n^*}\zeta < \theta_*$, (3.34) is equal to

$$\frac{\varphi(\theta - \sqrt{m'n'}\tau)^{m^*n^*-st}\varphi(\theta - (\sqrt{m'n'} - \sqrt{m^*n^*})\tau)^{st}\varphi(\sqrt{m^*n^*}\tau)^{m'n'-st}}{\varphi(\theta)^{m^*n^*}} \times \exp\left\{\frac{\zeta(\sqrt{m'n'}m^*n^* - \sqrt{m^*n^*}st)\theta)}{1+\tau_1}\right\} \tag{3.35}$$

By the fact that $\mathbb{E}(X_0) = 0$ and $\mathrm{Var}\,(X_0) = 1$, when $x > 0$ is close to zero, Taylor expansion on $\log\varphi(\cdot)$ shows

$$\log\varphi(x) = \frac{1}{2}x^2 + o(x^2).$$

We now set

$$\sqrt{m'n'}\tau = \frac{\theta}{2}\left[1 + \frac{1}{1+\tau_1}\right],$$

and from (3.10) we have $\sqrt{m'n'}\tau \to 0$. Meanwhile, (3.12) tells that

$$\sqrt{m^*n^*}\tau \propto \theta\sqrt{m^*n^*/m'n'} \to 0$$

as well. To see this, note that

$$\theta = O\left(\sqrt{\frac{\log M + \log N}{m^* + n^*}}\right),$$

therefore

$$\theta\sqrt{\frac{m^*n^*}{m'n'}} = O\left(\sqrt{\frac{(\log M + \log N)m^*n^*}{m'n'(m^* + n^*)}}\right) = O\left(\sqrt{\frac{\max(\log M, \log N)}{m^* + n^*}}\right),$$

and this is $o(1)$ by (3.10). The last part is from (3.18).

Take a natural log on (3.35), and by the expansion above and some algebra, we have

$$\log \mathbb{P}\left(Y_{S^*} - \lambda_{m^*,n^*} \le Y_{S'} - \lambda_{m',n'}\right) \le -\frac{m^*n^*\theta^2}{4}\left(1 - \frac{st}{\sqrt{m^*n^*m'n'}}\right)(1 + o(1))$$

and this finishes the proof.

3.8 Acknowledgment

Chapter 3, in full, is a version of the paper "A Multiscale Scan Statistic for Adaptive Submatrix Localization". The dissertation author is the principal investigator and corresponding author of this material. The manuscript is being prepared to be submitted to a major statistics journal.

Bibliography

[ABH16] Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62(1):471–487, 2016.

[ACDH05] Ery Arias-Castro, David L Donoho, and Xiaoming Huo. Near-optimal detection of geometric objects by fast multiscale methods. *IEEE Transactions on Information Theory*, 51(7):2402–2425, 2005.

[ACG13] Ery Arias-Castro and Geoffrey R Grimmett. Cluster detection in networks using percolation. *Bernoulli*, 19(2):676–719, 2013.

[ACL17] Ery Arias-Castro and Yuchao Liu. Distribution-free detection of a submatrix. *Journal of Multivariate Analysis*, 156:29–38, 2017.

[ACTW17] Ery Arias-Castro, Rui M Castro, Ervin Tánczos, and Meng Wang. Distribution-free detection of structured anomalies: permutation and rank-based scans. *Journal of the American Statistical Association*, 2017. To appear. Available online as an arXiv preprint arXiv:1508.03002.

[ACV14] Ery Arias-Castro and Nicolas Verzelen. Community detection in dense random networks. *The Annals of Statistics*, 42(3):940–969, 2014.

[BC16] Vincent Brault and Antoine Channarond. Fast and consistent algorithm for the latent block model. *arXiv preprint arXiv:1610.09005*, 2016.

[BI13] Cristina Butucea and Yuri I Ingster. Detection of a sparse submatrix of a high-dimensional noisy matrix. *Bernoulli*, 19(5B):2652–2688, 2013.

[BIS15] Cristina Butucea, Yuri I Ingster, and Irina A Suslina. Sharp variable selection of a sparse submatrix in a high-dimensional noisy matrix. *ESAIM: Probability and Statistics*, 19:115–134, 2015.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[BM+15] Rémi Bardenet, Odalric-Ambrym Maillard, et al. Concentration inequalities for sampling without replacement. *Bernoulli*, 21(3):1361–1385, 2015.

[BNW05] William T Barry, Andrew B Nobel, and Fred A Wright. Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, 21(9):1943–1949, 2005.

[CA11] Malika Charrad and Mohamed Ben Ahmed. Simultaneous clustering: A survey. In *International Conference on Pattern Recognition and Machine Intelligence*, pages 370–375. Springer, 2011.

[CAB17] Eric C Chi, Genevera I Allen, and Richard G Baraniuk. Convex biclustering. *Biometrics*, 73(1):10–19, 2017.

[CC00] Yizong Cheng and George M Church. Biclustering of expression data. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pages 93–103. AAAI Press, 2000.

[CCT12] Kamalika Chaudhuri, Fan Chung, and Alexander Tsiatas. Spectral clustering of graphs with general degrees in the extended planted partition model. In *Conference on Learning Theory*, pages 1–23, 2012.

[Cha09] Yen-Ching Chang. N-dimension golden section search: Its variants and limitations. In *2009 2nd International Conference on Biomedical Engineering and Informatics*, pages 1–6. IEEE, 2009.

[CLR+17] T Tony Cai, Tengyuan Liang, Alexander Rakhlin, et al. Computational and statistical boundaries for submatrix localization in a large noisy matrix. *The Annals of Statistics*, 45(4):1403–1430, 2017.

[CX16] Yudong Chen and Jiaming Xu. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *Journal of Machine Learning Research*, 17(27):1–57, 2016.

[For10] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3-5):75–174, 2010.

[GL16] David Gamarnik and Quan Li. Finding a large submatrix of a gaussian random matrix. *The Annals of Statistics, to appear. Manuscript available at arXiv preprint arXiv:1602.08529*, 2016.

[Got00] Nicholas J Gotelli. Null model analysis of species co-occurrence patterns. *Ecology*, 81(9):2606–2621, 2000.

[Haj62] Jaroslav Hajek. Asymptotically most powerful rank-order tests. *The Annals of Mathematical Statistics*, 33(3):1124–1147, 1962.

[Het84] Thomas P. Hettmansperger. *Statistical inference based on ranks*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, 1984.

[HG14] Jesse Hemerik and Jelle Goeman. Exact testing with random permutations. *arXiv preprint arXiv:1411.7565*, 2014.

[HLL83] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[Hoe63] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963.

[HS67] Jaroslav Hájek and Zbynek Sidak. *Theory of rank tests*. Academic Press, Academia Publishing House of the Czechoslovak Acad, 1967.

[HTE+00] Trevor Hastie, Robert Tibshirani, Michael B Eisen, Ash Alizadeh, Ronald Levy, Louis Staudt, Wing C Chan, David Botstein, and Patrick Brown. 'gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology*, 1(2):1–21, 2000.

[HWX15a] Bruce Hajek, Yihong Wu, and Jiaming Xu. Information limits for recovering a hidden community. *arXiv preprint arXiv:1509.07859*, 2015.

[HWX15b] Bruce Hajek, Yihong Wu, and Jiaming Xu. Submatrix localization via message passing. *arXiv preprint arXiv:1510.09219*, 2015.

[KBRS11] Mladen Kolar, Sivaraman Balakrishnan, Alessandro Rinaldo, and Aarti Singh. Minimax localization of structural information in large noisy matrices. In *Advances in Neural Information Processing Systems*, pages 909–917, 2011.

[KKZ09] Hans-Peter Kriegel, Peer Kröger, and Arthur Zimek. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(1):1, 2009.

[LR05] E. L. Lehmann and Joseph P. Romano. *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, New York, third edition, 2005.

[LST15] Jason D Lee, Yuekai Sun, and Jonathan E Taylor. Evaluating the statistical significance of biclusters. In *Advances in Neural Information Processing Systems*, pages 1324–1332, 2015.

[McS01] Frank McSherry. Spectral partitioning of random graphs. In *Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium on*, pages 529–537. IEEE, 2001.

[MNS15] Elchanan Mossel, Joe Neeman, and Allan Sly. Consistency thresholds for the planted bisection model. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 69–75. ACM, 2015.

[MO04] Sara C Madeira and Arlindo L Oliveira. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 1(1):24–45, 2004.

[MW15] Zongming Ma and Yihong Wu. Computational barriers in minimax submatrix detection. *The Annals of Statistics*, 43(3):1089–1116, 2015.

[PBZ+06] Amela Prelić, Stefan Bleuler, Philip Zimmermann, Anja Wille, Peter Bühlmann, Wilhelm Gruissem, Lars Hennig, Lothar Thiele, and Eckart Zitzler. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22(9):1122–1129, 2006.

[SD01] Almut Schulze and Julian Downward. Navigating gene expression using microarraysa technology review. *Nature cell biology*, 3(8):E190, 2001.

[Ser74] Robert J Serfling. Probability inequalities for the sum in sampling without replacement. *The Annals of Statistics*, pages 39–48, 1974.

[STQ08] Rodrigo Santamaría, Roberto Therón, and Luis Quintales. A visual analytics approach for understanding biclustering results from microarray data. *BMC bioinformatics*, 9(1):247, 2008.

[SWPN09] Andrey A Shabalin, Victor J Weigman, Charles M Perou, and Andrew B Nobel. Finding large average submatrices in high dimensional data. *The Annals of Applied Statistics*, 3(3):985–1012, 2009.

[TSS05] Amos Tanay, Roded Sharan, and Ron Shamir. Biclustering algorithms: A survey. *Handbook of computational molecular biology*, 9(1-20):122–124, 2005.

[TW14] Kean Ming Tan and Daniela M Witten. Sparse biclustering of transposable data. *Journal of Computational and Graphical Statistics*, 23(4):985–1008, 2014.

[ZZ+16] Anderson Y Zhang, Harrison H Zhou, et al. Minimax rates of community detection in stochastic block models. *The Annals of Statistics*, 44(5):2252–2280, 2016.