Significance and Recovery of Blocks Structures in Binary and Real-Valued Matrices with Noise

Xing Sun

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics and Operations Research (Statistics).

Chapel Hill
2007

Approved by

Advisor: Professor Andrew Nobel

Reader: Professor Amarjit Budhiraja

Reader: Professor Edward Carlstein

Reader: Professor Douglas Kelly

Reader: Professor Mark Huber

ABSTRACT

XING SUN: Significance and Recovery of Blocks Structures in Binary and Real-Valued Matrices with Noise

(Under the direction of Professor Andrew Nobel)

Biclustering algorithms have been of recent interest in the field of Data Mining, particularly in the analysis of high dimensional data. Most biclustering problems can be stated in the following form: given a rectangular data matrix with real or categorical entries, find every submatrix satisfying a given criterion. In this dissertation, we study the statistical properties of several commonly used biclustering algorithms under appropriate random matrix models. For binary data, we establish a three-point concentration result, and several related probability bounds, for the size of the largest square submatrix of 1s in a square Bernoulli matrix, and extend these results to non-square matrices and submatrices with fixed aspect ratios. We then consider the noise sensitivity of frequent itemset mining under a simple binary additive noise model, and show that, even at small noise levels, large blocks of 1s leave behind fragments of only logarithmic size. As a result, standard FIM algorithms that search only for submatrices of 1s cannot directly recover such blocks when noise is present. On the positive side, we show that an error-tolerant frequent itemset criterion can recover a submatrix of 1s against a background of 0s plus noise, even when the size of the submatrix of 1s is very small.

For data matrices with real-valued entries, we establish a concentration result for the size of the largest square submatrix with high average in a square Gaussian matrix. Probability upper bounds on the size of the largest non-square high average submatrix with a fixed row/column aspect ratio in a non-square real-valued matrix with fixed row/column aspect

ratio are also established when the entries of the matrix follow appropriate distributions. For biclustering algorithms targeting submatrices with low ANOVA residuals, we show how to assess the significance of the resulting submatrices. Lastly, we study the recoverability of submatrices with high average under an additive Gaussian noise model.

ACKNOWLEDGEMENTS

I owe my debt to my advisor Professor Andrew Nobel, who spent tremendous time guiding me. Without his valuable ideas, great support and continuous encouragement, I could not imagine completing this dissertation. I also would like to thank my committee members, Professor Amarjit Budhiraja, Professor Edward Carlstein, Professor Douglas Kelly, and Professor Mark Huber, who spent lots of their time reviewing my dissertation and made a lot of precious suggestions.

Many thanks to Professor Yufeng Liu, Professor Steve Marron, Professor Gabor Pataki and Professor Haipeng Shen, who spent their time discussing other interesting statistics and optimization problems with me; to all faculties and graduate fellows in UNC statistics and operation research department, who gave me kind helps during the past years; to Dr. Xiaoni Liu, who devoted a lot time to reading my dissertation and discussing with me. Finally, I would like to thank my parents for their spiritual supports.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

High-throughput technologies are widely used in scientific research, where large data sets are collected automatically with relatively low costs. These large data sets often contain a large number of variables and samples. Common examples of these large data sets can be transaction data which contains tens of thousands of different items and tens of thousands of transaction records (c.f. (12)), or drug activity data which contains hundreds of different compounds and less than a hundred atom types (c.f. (38)), or DNA Microarray data which contains from a thousand to twenty thousand human genes and less than a hundred samples (c.f. (28)).

Exploratory data analysis is often used when studying these large data sets. Exploratory data analysis employs techniques to better understand the data, specifically, to unveil the data structure, to build models, to identify important variables, and to test underlying assumptions. Data mining is often the first step in this exploratory analysis. It includes techniques such as supervised learning (classification, regression), unsupervised learning (clustering, biclustering, principle component analysis, singular value decomposition), and graphical visualization. Given response information, supervised learning tries to build models to connect the values of variables with the values of the responses, and further to predict the values of the responses based on the values of the variables. Unsupervised learning tries to explore the structure of data and build models by searching for consistent patterns and relationships between variables and samples, and this is done in the absence of a response.

In general, there are many different ways to define consistent patterns or relationships. In this dissertaion, a consistent pattern or relationship can be several variables taking a same value across different samples in binary or categorical data such as frequent itemsets

in the frequent itemset mining problem; it can also be strong correlations between variables across different samples, such as order preserving clusters; or it can be a partition of samples found by some clustering techniques such as hierarchical clustering, k-mean clustering etc.

By arranging the data into a matrix with rows representing variables and columns representing samples, these consistent patterns or relationships usually correspond to submatrices such as biclusters, or partitions of columns of the data matrix such as cluster structures. A real world example of these consistent patterns can be found in the gene expression analysis (54), where biologists apply clustering techniques to explore the DNA Microarray data from cancer patients. The resulting clustering structure suggests the biologists to further classify the patients into subgroups according to their different gene expression levels. An example of another type of consistent patterns can be found in drug activity analysis (38). Each chemical compound there is represented by a connection table and then coded to a canonical string with some characters representing atoms and some characters representing bonds. The patterns which are of interest to chemists are those substructures (substrings) frequently appearing across different compounds. If each compound is regarded as a sample, and each character in the canonical representation string as a variable, then a substructure corresponds to a subset of characters. A substructure is considered frequent if it appears in more than k compounds (samples), where the threshold k is predetermined by users.

1.1 Co-clustering

In the previous section, we briefly introduced an example of clustering analysis, which try to assign samples into groups such that within group distances are smaller than between group distances. However, a problem with this standard clustering technique is that the results do not directly reflect the variable structures, which are also of interest in research. To overcome this drawback, independent row and column clustering is proposed by Eisen et al. in (21), where they simply cluster columns and rows independently. Another concern on standard clustering algorithms and even the independent row and column clustering algorithm is that the distance they used has equal weights on all dimensions of variables. This may cause potential problems when the dimension is very high, where many variables are actually irrelevant noises. Some refined clustering techniques are proposed to deal with

this high dimensional problem such as Coupled Two Way Clustering (26), Iterative Two-way Clustering (64) and COSA (23). They are more flexible than the standard clustering methods, and they can reveal the associations between samples and variables. However, they are still based on iterative or weighted standard clustering. Thus, the association discovering is not done directly. They will not be studied in this dissertation.

1.2 Frequent Itemset Mining

Instead of discovering associations between samples and variables by indirect methods such as co-clustering, we are considering the problem in a more direct way. One of the simplest associations or patterns when the data only contain binary or categorical values is that a subset of variables take a same value across a subset of samples. This subset of variables are called frequent itemset in the data mining literature. For example, items which were purchased together in different transactions are called frequent items in market basket analysis. The technique to find these sets of items is called frequent itemset mining. The computational aspect of frequent itemset mining has been widely studied in (3; 2; 1) and the resulting methods have been applied in many different research areas. One example of the application of frequent itemset mining in drug discovery can be found in (38), where a modified frequent itemset mining algorithm is used to search for frequent substructures of chemical compounds. These frequent substructures can then be used as input variables in the drug activity classification model. It is shown that using frequent substructures can improve the accuracy of the prediction. Another example of applying frequent itemset mining to explore transaction data can be found in (12), where the real data from a Belgian retail store is analyzed by a frequent itemset mining technique. More examples of applications of frequent itemset mining can be found in data mining literature such as (3; 2; 1; 61; 27).

In general, the frequent itemset mining problem can be described as follows. The available data consists of n different items $S = \{s_1, ..., s_n\}$ and m transactions $T = \{t_1, \ldots, t_m\}$. Each transaction t_j is an index set. If item j appears in transaction i then the index $t_{ij} = 1$ otherwise $t_{ij} = 0$. Given such a binary data set, the objective of frequent itemset mining is to find all sets of items that appear in more than k transactions, where $k > 1$ is a user

determined parameter.

1.3 Biclustering

A number of common data mining techniques are similar to frequent itemset mining. They also try to identify distinguished associations between subsets of variables and subsets of samples. Again, if we represent each variable by a row in the data matrix and each sample by a column, these techniques are equivalent to identifying distinguished submatrices in the data matrix meeting different criteria. Techniques for doing this are called biclustering or biclustering techniques in the data mining literature. According to the type of data under study, biclustering techniques can be further classified into two types. In the first type, the data matrix consists of discrete entries; in the second type, the data matrix consists of continuous entries. Many different distinguishing criteria have been proposed to accommodate different types of data. Madeira and Oliveira in (44) summarize the biclustering criteria commonly used in gene expression data analysis into four types: biclusters with constant entries such as methods in (30; 13) , biclusters with constant rows/columns such as methods in (26; 14; 56; 57; 59), biclusters with coherent values such as methods in (15; 68; 69; 64; 40), and biclusters with coherent evolutions such as methods in (8; 43; 62). In this dissertation, distinguishing criteria studied for binary data include frequent itemset criterion described above, and error-tolerant frequent itemset criterion. When dealing with real-valued matrices, we study the high average criterion and ANOVA criterion (52; 67; 42; 41; 15; 40). For a general survey of biclustering algorithms, please refer to (63) and (44).

1.4 Statistical Significance Analysis

Numerous biclustering algorithms have been proposed and their implementations have been studied in the literature. However, little guidance can be found on how to evaluate the statistical significance of patterns identified by these algorithms. In this dissertation, we will address this problem. We will provide rigorous analysis on the statistical significance of output patterns by some commonly used biclustering algorithms. Some of our results are motivated by the existing works on clique numbers in graphic theory. Some of the results are new. They have never been studied before.

There are two primary reasons to consider the statistical significance of biclusters. The first reason is that biclustering algorithms usually produce a large number of output patterns. Among them, we want to know which are likely to be the results of noise and which are potentially interesting. By evaluating the statistical significance of patterns, we can identify out those potentially interesting ones. The second reason is from the computational consideration. Many biclustering methods are computationally intensive. A possible improvement can rely on only searching for those significant patterns rather than all patterns. For example, Koyutürk *et al.* in (37) propose a relatively efficient biclustering algorithm by only searching for those potentially significant patterns.

Before actually giving our analysis, we first briefly review the statistical significance analysis on standard clustering methods. Standard clustering refers to techniques which try to assign samples into subgroups such that the (average of) pairwise distance (based on all variables) between two samples within a same group is smaller than the distance between two samples from different groups. Validating and interpreting the clusters identified by standard clustering methods is not a thoroughly treated problem in statistics. One may ask whether the clusters have arisen by chance. Are there p-values associated with particular clusters? How many clusters should be there? Answers to these questions involve the multiple comparison problems. In general, multiple comparison problems are difficult. Some works on multiple testing can be found in studies of gene expression data, such as permutation based correction and false discovery rate (FDR) studied in (20). However, no existing work can successfully solve the multiple comparison problems in clustering.

The general validating methods for standard clustering can be divided into two main categories: external criterion methods and internal criterion methods. External criterion methods compare the clusters with prior information such as the method in (46). However, in most cases, prior information is not available. The internal criterion methods use information within the given data set and evaluate the goodness of fit. For example, Kaufman and Rousseeuw in (34) introduce the Silhouette statistic to assess clusters and estimate the optimal number of clusters. The gap statistic by Tibshirani *et al.* in (65) attempts to estimate the number of clusters by comparing within cluster dispersion to that of the reference null distribution. Simulation and resampling methods studied in (19; 39; 70) are

another two popular and widespread ways to assess the statistical significance or to validate the clustering result. However, they are computationally intensive. In standard clustering applications, such as spatial data analysis and epidemiology data analysis, probabilistic analyses on clustering results are available (*c.f.* (25)). Some of these probabilistic analyses (*c.f.* (31)) assume a spatial uniform distribution and the p-values are assigned based on the probability of finding a large number of observations in a small (usually spherical) area under Poisson or some other spatial distribution assumptions.

In conclusion, all the analyses mentioned above either can only play the role of validating clustering results rather than giving a rigorous statistical significant analysis such as a p-value; or by resampling method, they can give an estimated p-value but have the problem of being computationally intensive; or they only work for some particular clustering problems.

In contrast to the significance analysis of clusters, the analysis on biclusters is much easier. The reason is that when studying the significance of biclusters, one can directly treat the submatrices as a set of individual random variables, while in standard clustering, one needs to consider all the columns as a whole, which is more difficult to study. To analyze the significance of biclusters, Koyutürk *et al*. in (37) assume that the entries of data matrix follow a uniform memoryless distribution and by large deviation principle, they give p-values to submatrices. Note that the p-value given there is only the p-value for significance of an individual submatrix. If the null hypothesis is about significance among all submatrices, multiple test procedure is needed. Tanay *et al*. in (62) propose a p-value to evaluate the significance of a single bicluster by Central Limit Theorem. They also apply Bonferroni correction to achieve the overall p-value. However, this p-value is suboptimal due to the normal approximation when the data are actually non-Gaussian. In this dissertation, for a number of commonly used biclustering method, we will give the analyses directly.

1.5 Overview

This dissertation is organized as follows. We focus on binary-valued data in the first part and real-valued data in the second part. The last part discusses some future works. In Chapter 2, we give results on the statistical significance of frequent itemsets identified by standard frequent itemset mining. In Chapter 3, we propose a binary additive noise model

and then study the noise sensitivity of standard frequent itemset mining under this model. Due to the poor performance of standard frequent itemset mining in noisy environments, we begin to consider the error-tolerant frequent itemset mining methods which are natural relaxations of standard frequent itemset mining. We study the statistical significance of the patterns identified by some popular error-tolerant frequent itemset mining techniques. We then study the recoverability of the approximate frequent itemset proposed and studied in (42; 41) in a simple recovery problem. In Chapter 4, we switch our attentions to real-valued data. Under the assumption of i.i.d. Gaussian entries/i.i.d bounded entries, we study the statistical significance of submatrices identified by biclustering techniques based on high average criterion and ANOVA criterion. In Chapter 5, we study the consistency of submatrices with high average in a block recovery problem with Gaussian noise.

CHAPTER 2

Significance Analysis of Frequent Itemsets in Binary Matrix

2.1 Matrix Expression of Frequent Itemset Mining

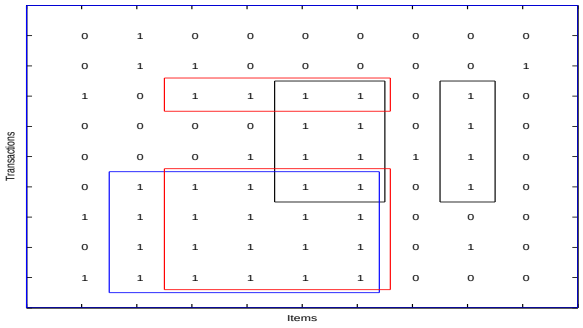
Recall that by definition, frequent itemset mining tries to find all sets of items that appear frequently in a given binary data set. It is easy to see that the frequent itemset mining problem can be expressed equivalently in a matrix form. In particular, one can express the data from a frequent itemset mining problem with m transactions and n items as an $m \times n$ binary matrix $\mathbf{X}$, where each row of $\mathbf{X}$ represents a transaction, and each column of $\mathbf{X}$ represents an item available to purchase. The entry $x_{ij} = 1$ if the j'th item is purchased in the i'th transaction, otherwise $x_{ij} = 0$.

Let A be a subset of rows and let B be a subset of columns. The index set $C = A \times B$ is called a submatrix of $\mathbf{X}$. Clearly, in the frequent itemset mining problem, a submatrix C contains information about whether items in A are purchased in those transactions in B.

Given $\mathbf{X}$, the objective of frequent itemset mining can then be translated to discovering all maximal submatrices of 1's with the number of rows greater than k. Here, a submatrix C is called a maximal submatrix of 1's if there does not exist another submatrix C' of 1's such that $C \subset C'$.

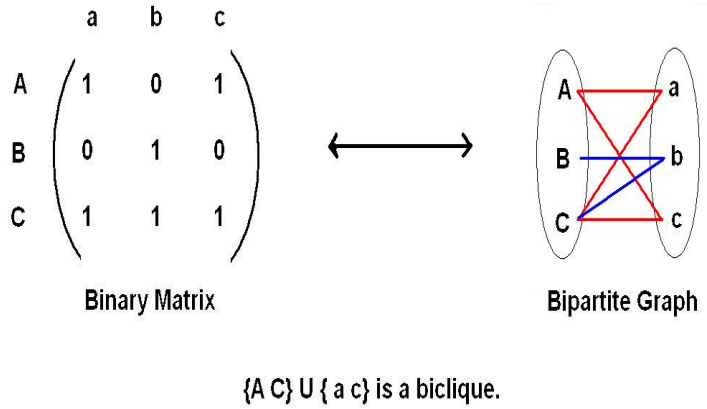
Figure 2.1 below is an example which illustrates the matrix form of the frequent itemset mining problem. Clearly, when $k \leq 4$, all three submatrices are considered frequent, and when $k = 5$, only the submatrix in red satisfies the requirement of being frequent. Note that the rows and the columns in the submatrices are not necessarily to be contiguous in the frequent itemset mining problem. In fact, this matrix form of the frequent itemset

Figure 2.1: Matrix Form Expression of Frequent Itemset Mining



mining problem also reveals an one to one correspondence with bipartite graphs. This correspondence has been used as the basis for biclustering algorithms such as (47; 62). To be specific, an $m \times n$ binary data matrix X can be represented as a graph $G = (V, E)$, whose vertex set V can be expressed as the union of two disjoint sets V_1 and V_2. The first set V_1 represents the set of transactions (rows), and the second set V_2 represents the set of items (columns). There is an edge $(v_1, v_2) \in E$ between vertices $v_1 \in V_1$ and $v_2 \in V_2$ iff $x_{v_1,v_2} = 1$. There are no edges connecting vertices within V_1 or vertices within V_2. Given any subset $V_1' \subset V_1$ and subset $V_2' \subset V_2$, the associated subgraph is defined as $H = (V', E')$ of G where $V' = V_1' \cup V_2'$ and E' is the induced set of edges. H is called a complete bipartite graph of G if there exists an edge between any pair of vertices in V_1' and V_2'. Moreover, if this bipartite graph is maximal, which means there does not exist another complete bipartite graph containing it, it is called a *biclique* of G. Clearly, a frequent itemset corresponds to a biclique with at least k vertices in V_1'. Figure 2.2 is an example demonstrating the connection between frequent itemsets and bicliques. It is known from (c.f. (24; 32; 51)) that the problem of finding the largest biclique in a given bipartite graph

Figure 2.2: Example of Biclique



G is NP-complete, and also the problem of finding the largest biclique with roughly equal vertex set sizes (c.f. (35; 22)). Some heuristic methods, such as those in (35; 22), and several approximate methods (c.f. (32; 47)) have been proposed to find the largest biclique, or the largest frequent itemset in polynomial time. Mirisha et al. also show in (47) that the results provided by their randomized algorithm can overlap a large proportion of the largest bicliques with high probability.

2.2 Significance Analysis of frequent itemset mining

In this dissertaion, we will not study how to search for frequent itemsets. Instead, we will focus on how to assess the statistical significance of frequent itemsets identified in frequent itemset mining problems. For this purpose, we consider the sizes of maximal submatrices of 1's in a binary random matrix. For simplicity, we will start by restricting ourselves to the case of square target submatrices in a square data matrix in this section. We will extend the results here to non-square target submatrices and matrices in later sections. In this section, we will also only focus on binary matrices. Most of our results obtained in

the binary case can be extended with little difficulty to the case of categorical data. This extension is trivial. Therefore it is omitted. Extensions to the case of real-valued matrices will be discussed in Chapter 4.

To begin, we first define a random matrix model and a random variable that will be studied throughout this chapter.

Definition: Let $\mathbf{Z} = \{z_{i,j} : i, j \geq 1\}$ be an infinite array of independent binary random variables with $P(z_{i,j} = 1) = p = 1 - P(z_{i,j} = 0)$, where the probability $p \in (0, 1)$ is fixed. For $n \geq 1$, let $\mathbf{Z}_n = \{z_{i,j} : 1 \leq i, j \leq n\}$.

Thus $\mathbf{Z}_n$ is an $n \times n$ binary random matrix comprising the "upper left corner" of the collection $\{z_{i,j}\}$. This definition allows us to make almost-sure type statements concerning the asymptotic behavior of functions of $\mathbf{Z}_n$.

Note that since entries in the submatrices we studied are all 1, and that the submatrix structures are invariant to row and column permutations, one of the only few quantities we can use to study the statistical significance of the submatrices of 1's is their size.

Definition: Given a binary matrix $\mathbf{X}$, let $M(\mathbf{X})$ be the largest integer k such that there exists a $k \times k$ submatrix of 1's in $\mathbf{X}$.

When assessing the statistical significance of the identified submatrices of 1's, the above binary random matrix model can be viewed as a null model and $M(\cdot)$ can be viewed as a natural test statistic . To be more specific, a $k \times k$ submatrix of 1's in $\mathbf{Z}_n$ has an associated significance value equal to $P\left(M(\mathbf{Z}_n) \geq k\right)$. To obtain this probability, one can follow the standard first moment method. Let $U_k(n)$ be the number of $k \times k$ submatrices of ones in $\mathbf{Z}_n$. It is easy to see that

$$P\left(M(\mathbf{Z}_n) \geq k\right) = P(U_k \geq 1) \leq EU_k = \binom{n}{k}^2 p^{k^2}.$$

Clearly, we need a bound on EU_k. Note that $EU_1 = n^2 p > 1$, $EU_n = p^{n^2} < 1$ and that EU_k is decreasing in k when $k > \log_{\frac{1}{p}} n$. Therefore, we wish to identify an integer $k(n)$ such that $EU_{k(n)} \approx 1$. The simple idea behind $k(n)$ is that when a submatrix of 1's with size $k > k(n)$ is observed, $EU_k < 1$, which suggests this submatrix might be significant; when a

submatrix of 1's with size $k < k(n)$ is observed, $EU_k > 1$, which suggests that it might be common. In order to obtain $k(n)$, we first consider the Stirling approximation of EU_k. Let

$$\phi(n,k) = (2\pi)^{-\frac{1}{2}}\, n^{n+\frac{1}{2}}\, k^{-k-\frac{1}{2}}\, (n-k)^{-(n-k)-\frac{1}{2}}\, p^{\frac{k^2}{2}} \approx (EU_k)^{1/2}. \tag{2.1}$$

Let $s(n)$ be any real-valued root of equation

$$1 = \phi(n,s). \tag{2.2}$$

The following lemma asserts that $s(n)$ uniquely exists and so does $k(n)$.

Lemma 2.2.1. *When n is sufficiently large, the $s(n)$ in (2.2) is unique, and satisfies $\log_b n < s(n) < 2\log_b n$, where $b = p^{-1}$.*

Let $k(n) = \lceil s(n) \rceil$ be the least integer larger than $s(n)$. Based on Lemma 2.2.1 and some technical but straightforward calculations, one can obtain the asymptotic expression of $s(n)$ as a deterministic function of n and p.

Lemma 2.2.2.

$$s(n) \;=\; 2\,\log_b n \;-\; 2\,\log_b \log_b n + C + o(1), \tag{2.3}$$

where $b = p^{-1}$ and $C = 2\log_b e - 2\log_b 2$.

The proofs of Lemma 2.2.1 and Lemma 2.8.1 can be found in Section 2.6.

Given the fact that $k(n)$ exists and is unique, one can establish the following proposition which provides an upper bound on $P\left(M(\mathbf{Z}_n) \geq k\right)$ for $k > k(n)$.

Proposition 2.2.3. *Fix $0 < \gamma < 1$. When n is sufficiently large, $P\left(M(\mathbf{Z}_n) \geq k(n) + r\right) \leq 2\, n^{-2\,r}\, (\log_b n)^{3r}$ for each $1 \leq r \leq \gamma\, n$.*

Remark: (i) Note that a crude upper bound on $M(\mathbf{Z}_n)$ may be obtained easily by verifying the following simple algebra. Indeed,

$$EU_k = \binom{n}{k}^2 p^{k^2} \leq \frac{n^{2k}}{k!^2} e^{-k^2 \ln b} \leq \frac{e^{2k\ln n - k^2 \ln b}}{k^2} \leq n^{-2}, \tag{2.4}$$

when $k \geq 2\log_b n + 1$. More generally, if p is replaced by $p(k,n)$, the above inequality still holds for any $k \geq -2\ln n/\ln p(k,n) + 1$. This trivial extension will be used in the later chapters. However, the proof of Proposition 2.2.3 in Section 2.7 can provide a more precise upper bound as shown in the next theorem.

(ii)This explicit bound on $M(\mathbf{Z}_n)$ can also be viewed as an extension of the result on clique numbers in Bollobás and Erdős (11) to the case of bicliques.

(iii) Note that the probability bound in Proposition 2.2.3 can be considered as a Bonferroni correction over all $k \times k$ submatrices in $\mathbf{Z}_n$. Usually the Bonferroni correction is conservative, but the next theorem shows that, in term of the critical threshold $k(n)$, it is tight.

Proposition 2.2.3 gives an upper bound on $P\left(M(\mathbf{Z}_n) > k(n)\right)$. One may also ask the question of how likely the size of the identified submatrix will be less than $k(n)$. This question corrsponds to the probability bound on $\{M(\mathbf{Z}_n) < k(n)\}$. The following theorem gives an answer to this question. It eventually implies that if the largest size of the observed square submatrix of 1's is much smaller than $k(n)$, then either this submatrix is still not the largest square submatrix of 1's in the whole matrix or one should suspect the i.i.d. Bernoulli random matrix model assumption. Note that as we have mentioned before, the concept of eventually almost sure convergency in the next theorem follows the convention in Bollobás and Erdős (11). The proof of the next theorem follows the general outlines by Bollobás and Erdős (11). The detailed proof can be found in Section 2.8.

Theorem 2.2.4. *When n is sufficiently large, $|M(\mathbf{Z}_n) - s(n)| < \frac{3}{2}$ eventually almost surely.*

It follows from Theorem 2.2.4 that $M(\mathbf{Z}_n)$ can take one of at most three (integer) values. This is similar to the result of clique numbers obtained by Bollobás and Erdős (11) and Matula (45), where they study the size of the largest clique, $cl(G_n)$, in a random graph G_n with n vertices and each edge being included independently with probability p. They show that when the size of graph n is sufficiently large, there exists a deterministic function $c(n)$, same as $s(n)$ up to a constant, such that $|cl(G_n) - c(n)| < 3/2$ eventually almost surely. Bollobás and Erdős in (10) give a good account of these results. In fact, the proof of Theorem 2.2.4 follows the basic outlines by Bollobás and Erdős in (11). However, due

to the difference between these two problems, we still need to handle some technical details carefully in the proof.

Dawande *et al.* in (16) use first and second moment arguments to show (in our terminology) that $P\left(\log_b n \leq M(\mathbf{Z}_n) \leq 2\log_b n\right) \to 1$ as n tends to infinity. Park and Szpankowshi in (50) improve the result of Dawande *et al.*. They show that

$$P\left((1+\epsilon)\log_b n \leq M(\mathbf{Z}_n) \leq (2-\epsilon)\log_b n\right) \to 1$$

as n tends to infinity for any fixed $0 < \epsilon < 1$. These are weaker versions of Theorem 2.2.4. Koyutürk *et al.* study the problem of finding dense patterns in binary data matrices in (37). They use a Chernoff type bound for the binomial distribution to assess whether an individual submatrix has an enriched fraction of ones, and employ the resulting test as the basis for a heuristic search for significant bi-clusters. However, the effects of multiple testing are not considered in their assessments of significance. Tanay *et al.* (62) assess the significance of bi-clusters in a real-valued matrix using likelihood-based weights, a normal approximation and a standard Bonferroni correction to account for the multiplicity of submatrices. Use of the normal approximation for individual submatrices leads to subtoptimal bounds in non-Gaussian setting.

Theorem 2.2.4 bounds $M(\mathbf{Z}_n)$, the size of the largest maximal square submatrices of 1's, from both above and below almost surely. It gives the range of values in which we expect to find the size of the largest square submatrices of 1's in an i.i.d. Bernoulli random matrix. However, in practice, one may not be able to find the largest square submatrix of 1's in a data matrix due to its computational complexity. Thus, it is also useful to ask what is the size of the smallest maximal square submatrices of 1's in $\mathbf{Z}_n$. When a maximal square submatrix of 1's is observed and its size is much smaller than what we expected, we should suspect the i.i.d. Bern(p) random matrix model assumption.

Definition: Let $L(\mathbf{Z}_n)$ be the smallest k such that there exists at least one $k \times k$ maximal submatrix of 1's in $\mathbf{Z}_n$ and this submatrix is not contained by any other square submatrix of 1's.

An analysis by adopting similar second moment arguments as those in the proof of Theorem 2.2.4 yields the following result for $L(\mathbf{Z}_n)$. The detailed proof can be found in Section 2.9.

Theorem 2.2.5. *With probability one,*

$$\lim_{n\to\infty} \frac{L(\mathbf{Z}_n)}{\log_b n} = 1.$$

Remark: Bollobás and Erdős (11) establish a related result on the size of the smallest cliques in a random graph. However, their proof can not be directly extended here to obtain the theorem above. What they actually consider, in our terminology, corresponds to the lower bound on the size of the smallest square submatrices of 1's, which is not contained by any rectangular submatrix of 1's. Obviously, this lower bound is always larger than $L(\mathbf{Z}_n)$, since the event that a square submatrix of 1's is not contained by a larger rectangular submatrix of 1's in $\mathbf{Z}_n$ implies it is also not contained by any larger square submatrix of 1's in $\mathbf{Z}_n$.

2.3 Non-Square Matrices

The results in the previous section apply to the special case of square matrices $\mathbf{Z}_n$ and square submatrices. This restriction can be readily relaxed, yielding bounds that are better suited to the data sets in recent scientific research with large numbers of variables and relatively few samples. In this section, we consider the case that $\mathbf{Z}_{m,n} \sim \text{Bern}(p)$ is an $m \times n$ random matrix with a fixed row/column aspect ratio $\alpha = \frac{m}{n}$ for some $\alpha > 0$ as n or m growing. We also allow the target submatrices to be rectangular with a fixed row/column aspect ratio β. Analogous to that of Proposition 2.2.3 and Theorem 2.2.4, we defined $M(\cdot)$ as follows.

Definition: Fix $\beta \geq 1$. Given an $m \times n$ i.i.d. random Bern(p) matrix $\mathbf{Z}_{m,n}$ with $\frac{m}{n} = \alpha > 0$, let $M(\mathbf{Z}, m, n, \beta)$ be the largest k such that $\mathbf{Z}_{mn}$ contains a $\lceil \beta k \rceil \times k$ submatrix of 1's.

The asymptotic behavior of $M(\mathbf{Z}, m, n, \beta)$ is the same as $M(\mathbf{Z}, n, m, \beta^{-1})$, so we only consider $\beta \geq 1$ here. Following similar steps as those in analyzing $M(Z_n)$, we first investigate

the value of k for which the expected number of $\lceil \beta k \rceil \times k$ submatrices of 1's in $\mathbf{Z}(\lceil \alpha n \rceil, n)$ approximately equal to one. Formally, for each k, let $U_k(m, n, \beta)$ be the number of $\lceil \beta k \rceil \times k$ submatrices of 1's in $\mathbf{Z}_{mn}$. Then

$$EU_k(m, n, \beta) = \binom{m}{\lceil \beta k \rceil} \binom{n}{k} p^{\lceil \beta k \rceil k}.$$

Define $s(m, n, \beta)$ to be the root of the equation

$$\begin{aligned} 1 &= \phi(s, m, n, \beta) \\ &= (2\pi)^{-1} n^{n+\frac{1}{2}} m^{m+\frac{1}{2}} s^{-s-\frac{1}{2}} (\beta s)^{-\beta s-\frac{1}{2}} (n-s)^{-(n-s)-\frac{1}{2}} (m-\beta s)^{-(m-\beta s)-\frac{1}{2}} p^{\beta s^2} \end{aligned} \tag{2.5}$$

over $s \in \mathbb{R}^+$. The uniqueness and existence of $s(m, n, \rho)$ is guaranteed by Lemma 2.11.1 in Section 2.11, and an asymptotic expression of $s(m, n, \beta)$ for large n and $m = \lceil \alpha n \rceil$ is given by Lemma 2.11.2 in Section 2.11.

$$s(m, n, \beta) = \frac{1+\beta}{\beta} \log_b n - \frac{1+\beta}{\beta} \log_b \left(\frac{1+\beta}{\beta} \log_b n \right) + \log_b \alpha + C(\beta) + o(1), \quad b = p^{-1} \tag{2.6}$$

for some constant $C(\beta) \geq 0$ depending only on β. Note that the aspect ratio α of the primary matrix appears only in the constant term, and therefore plays an insignificant role in the threshold value for k. The proofs of the following result are similar to that in the square case with some additional notation and work to handle the two aspect ratio. They can be found in Section 2.10 and Section 2.11 respectively.

Proposition 2.3.1. *Fix $0 < \gamma < 1$ and $\alpha > 0$. When n is sufficiently large, for each $1 \leq r \leq \gamma n$,*

$$P\{M(\mathbf{Z}, \lceil \alpha n \rceil, n, \beta) \geq k(\lceil \alpha n \rceil, n, \beta) + r\} \leq n^{-(\beta+1) r} 2(\log_b n)^{(\beta+2) r}, \tag{2.7}$$

where $k(\lceil \alpha n \rceil, n, \beta) = \frac{\beta+1}{\beta} \log_b n + \log_b \frac{\alpha}{\beta}$.

Since a fixed aspect ratio α of the primary matrix does not play an essential role in the

asymptotic behavior of $M(\mathbf{Z}, \lceil \alpha n \rceil, n, \beta)$, it is natural to consider a situation in which α can increase with n. This might model, for example, the scaling and cost structure of a given high-throughput technology. In the case where $\alpha(n) = \lceil n^{\gamma} \rceil$ for some $\gamma > 0$ the proof of Proposition 2.3.1 can be modified to show that

$$P\left(M(\mathbf{Z}, \lceil \alpha(n) \rceil, n, \beta) \geq \left(\gamma + \frac{\beta+1}{\beta}\right) \log_b n\right) \leq 2n^{-(\beta+1)\,r} \left(\log_b n\right)^{(\beta+2)r}.$$

This implies that large submatrices of 1's with aspect ratio β might still be significant.

On the other hand, one can easily show that when $\beta \geq 1$ is fixed and m grows exponentially in n, Z_{mn} can contain a $\lceil \beta n \rceil \times n$ submatrix of 1's with a positive probability. For example, suppose $m = \lceil e^n \rceil$. Let $k = \min\{n, \frac{n}{2 \ln b}\}$. We can show below that with high probability there exists a $\lceil \beta k \rceil \times k$ submatrix of 1's in the binary random matrix $\mathbf{Z}_{mn}$, given z_{ij} follows i.i.d. Bern(p). Indeed, let $\mathbf{Z}^*_{mk}$ be the binary matrix formed by the first k columns of $\mathbf{Z}_{mn}$. Clearly, for any $1 \leq i \leq m$, $P(z^*_{i,1} = ... = z^*_{ik} = 1) = p^k$, and the number of rows with all one entries in $\mathbf{Z}^*_{mk}$ follows a Binomial($\lceil e^n \rceil, p^k$) distribution. By the assumption that $k = \min\{n, \frac{n}{2 \ln b}\}$, the mean of this binomial distribution equals to $\lceil e^n \rceil \times p^k \geq \lceil e^n \rceil \times e^{-\frac{n}{2}} > \beta k$ for any constant β. Thus, with high probability, there exists a $\lceil \beta k \rceil \times k$ submatrix of 1's in $\mathbf{Z}_{mn}$.

Theorem 2.3.2. *Fix any $\alpha > 0$ and $\beta > 1$. Eventually almost surely, $|M(\mathbf{Z}, \lceil \alpha n \rceil, n, \beta) - s(\lceil \alpha n \rceil, n, \beta)| \leq \frac{5}{2}$.*

Remark: (i) When α and β are fixed, $s(\lceil \alpha n \rceil, n, \beta)$ is a deterministic function depending only on n.

(ii) Theorem 2.3.2 implies that $M(\mathbf{Z}, \lceil \alpha n \rceil, n, \beta)$ contains a submatrix of 1s having aspect ratio β and area $(\beta + 1) \log_b^2 n$, the latter increasing with β. Park and Szpankowski (50) establish a related result, showing that if we do not restrict β, the aspect ratio of the submatrices, then with high probability the submatrix of 1s in $\mathbf{Z}(m, n)$ with the largest area is of size $O(\alpha n) \times \ln b$ or $\ln b \times O(n)$.

For any discovered submatrix of 1's, we can use Propositions 2.3.1 to evaluate its statistical significance. We show a sample calculation explicitly in the following example.

Table 2.1: Simulation Results on $\hat{M}(\mathbf{Z}_n)$ Based on 400 Replications for Each n.

n	$s(n)$	k	Proportion of $\hat{M} = k$
40	3.553	3	85.75%
		4	14.25%
80	4.582	4	97%
		5	3%

Example. A frequent itemset mining algorithm is applied to a $4,000 \times 100$ binary matrix $\mathbf{Y}$, 65% of whose entries are equal to 1. Suppose that the algorithm finds a 44×25 submatrix $\mathbf{U}$ of ones in $\mathbf{Y}$. Applying Proposition 2.3.1 with $p = 0.65$, $\alpha = 40$ and $\beta = 1.76$ we find that $k(\lceil \alpha n \rceil, n, \beta) = 24$ and that the probability of finding such a matrix $\mathbf{U}$ in a purely random matrix is at most

$$2\, n^{-(1.76+1)\times(25-24)}\, (\log_b n)^{(1.76+2)\times(25-24)} \approx 0.04467.$$

Thus, a significant value $p(\mathbf{U}) \leq 0.04467$ may be assigned to $\mathbf{U}$.

2.4 Simulation

The results in the previous sections hold when n is sufficiently large. To check their validity for moderate values of n, we carried out a simple simulation study on $\mathbf{Z}_n$ with $n = 40$ and 80, and $p = .2$. In each case we generated 400 random matrices and applied the FP-growth algorithm (29) to identify all maximal submatrices of ones. For each maximal submatrix of ones we recorded the length of its shorter side, and let $\hat{M}$ be the maximum among these lengths. Thus $\hat{M}$ is equivalent to the side length of the largest square submatrix of 1's in the generated random matrix. We recorded the values of $\hat{M}$ over all 400 simulations and compared these values to the corresponding bounds $s(40) \approx 3.553$ and $s(80) \approx 4.582$ with respect to $p = 0.2$. Table 2.1 summarizes the results. Note that no value $\hat{M} \geq s(n)+1$ and no value $\hat{M} \leq s(n) - 1$.

In order to check our theoretical results on $M(\mathbf{Z}, n, n, \beta)$ with $\beta > 1$, we ran 100 simulations of 80×80 matrix with Bernoulli entries (p=0.1). By applying FP-growth algorithm,

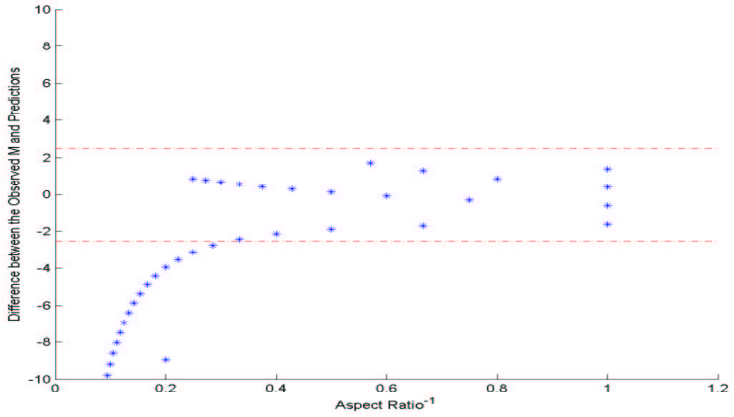
we found all rectangular maximal submatrices of 1's, and recorded the length of both their longer and shorter sides. For each appropriate aspect ratio $\beta \geq 1$, let $\hat{M}(\mathbf{Z}, n, n, \beta)$ be the largest k such that at least one $\beta k \times k$ or $k \times \beta k$ submatrix of 1's is observed. Note that since the length of the longer side and the length of the shorter side of the submatrices in $\mathbf{Z}_{80,80}$ take values among finite integers, β can only take a finite number of values. The difference between $\hat{M}(\mathbf{Z}, n, n, \beta)$ and $k(n, n, \beta)$ is calculated and displayed in Figure 2.3 and Figure 2.4. The value of X-axis in both of the following plots are $1/\beta$. The Y-axis in Figure 2.3 is the difference between $\hat{M}(\mathbf{Z}, n, n, \beta)$ and $k(n, n, \beta)$, and the Y-axis in Figure 2.4 is the proportion of simulations which are inconsistent with the theoretical predictions summarized by bins of β^{-1} with length 0.1. Note that even for moderate matrix size $n = 80$, the theoretical prediction is very accurate when the aspect ratio β is less than 2. In these cases, all observed size lengths are within the predicted value ranges. When the aspect ratio is large, corresponding to $\beta > 2.5$, the deviations from the predicted value are obvious. This reflects the fact that in Proposition 2.3.1, n needs to be sufficiently large for any fixed β, but this threshold of n depends on β. In current simulation setting, 80 is obviously not large enough for the larger aspect ratios.

2.5 Significance Analysis of Frequent Itemsets When Items are Dependent

In previous sections, we studied the statistical significance of frequent itemsets under the i.i.d. Bern(p) random matrix model. However, in some applications, the entries of the data matrix are known to be dependent. For example, in gene expression data, correlations exist between the expression levels of genes. In this scenario, to evaluate the statistical significance, we need incorporate dependence structure into the previous model. For binary matrix, a natural extension of the previous i.i.d binary random matrix model is to assume a Markov chain type dependence structure. In fact, one may assume the following model.

Alternative binary random matrix model: Let $c_1, ..., c_n$ be the columns of $\mathbf{Z}_n$, where $\mathbf{Z}_n$ is an $n \times n$ binary random matrix after suitable row-wise permutations. $c_1, ..., c_n$ are i.i.d. following a two state Markov chain with transition probability $P(z_{i+1,j} = 1 | z_{i,j} = 0) = p_0$

Figure 2.3: Difference between Prediction and Observed $\hat{M}(\cdot)$



and $P(z_{i+1,j} = 1 | z_{i,j} = 1) = p_1$.

Recall that by definition, $M(\mathbf{Z}_n)$ is the size of the largest square submatrix of 1's in $\mathbf{Z}_n$. By some simple arguments, one can still establish the following probability upper bound on $M(\mathbf{Z}_n)$.

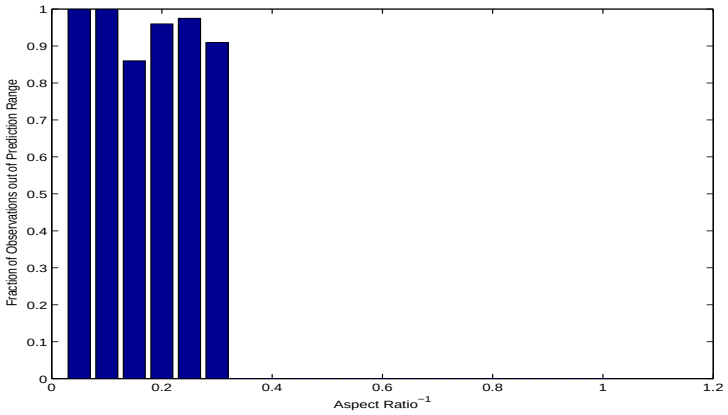
Proposition 2.5.1. *Fix any $0 < \epsilon < 1$, when n is sufficiently large,*

$$P(M(Z) \geq 2\log_b n + r) \leq n^{(2-\epsilon)r}, \tag{2.8}$$

where $b = \min\{p_0^{-1}, p_1^{-1}\}$.

Proof: Fix any column c_j. By the assumption, c_j is a Markov chain. Therefore, for any

Figure 2.4: Fraction of Observed $\hat{M}(\cdot)$ out of Predicted Range



row index $i_1 < ... < i_k$, it follows that

$$P(z_{i_1,j} = 1, ..., z_{i_k,j} = 1) = \Pi_{r=1}^{k} P(z_{i_r,j} = 1 | z_{i_{r-1},j} = 1).$$

When $i_r > i_{r-1} + 1$, one can verify that

$$\begin{aligned} P(z_{i_r,j} = 1 | z_{i_{r-1},j} = 1) &= \frac{P(z_{i_r,j} = 1, z_{i_{r-1},j} = 1)}{P(z_{i_{r-1},j} = 1)} \\ &= \sum_{u=0,1} \frac{P(z_{i_r,j} = 1, z_{i_r-1,j} = u, z_{i_{r-1},j} = 1)}{P(z_{i_{r-1},j} = 1, z_{i_{r-1},j} = 1)} \cdot \frac{P(z_{i_r-1,j} = u, z_{i_{r-1},j} = 1)}{P(z_{i_{r-1},j} = 1)} \\ &= \sum_{u=0,1} P(z_{i_r,j} = 1 | z_{i_r-1,j} = u) P(z_{i_r-1,j} = u | z_{i_{r-1},j} = 1) \\ &\leq \max\{p_0, p_1\} \sum_{u=0,1} P(z_{i_r-1,j} = u | z_{i_{r-1},j} = 1) = \max\{p_0, p_1\}. \end{aligned}$$

When $i_r = i_{r-1} + 1$, from the condition that $P(z_{i_r,j} = 1 | z_{i_{r-1},j} = 1) = p_1$, inequality

$P(z_{i_r,j} = 1 | z_{i_{r-1},j} = 1) \leq \max\{p_0, p_1\}$ holds immediately. By putting the above two cases together, one can conclude that $P(z_{i_1,j} = 1, ..., z_{i_k,j} = 1) \leq \max\{p_0, p_1\}^k$. Thus for any $k \times k$ submatrix V, it follows that $P(F(V) = 1) \leq (\max\{p_0, p_1\})^{k^2}$. Moreover, by following steps similar to those in the proof of Proposition 2.2.3, one can get inequality (2.8).

2.6 Proof of Lemma 2.2.1 and Lemma 2.2.2

Proof of Lemma 2.2.1: Differentiating $\log_b(\phi(n, s))$ yields

$$\frac{\partial \log(\phi(n,s))}{\partial s} = \frac{1}{2(n-s)} + \log_b(n-s) - s - \log_b s - \frac{1}{2s},$$

which is negative when $\log_b n < s < 2\log_b n$. A routine calculation shows that for $0 < s \leq \log_b n$,

$$\begin{aligned} \log_b \phi(n,s) &= (n+\frac{1}{2})\log_b n - (s+\frac{1}{2})\log_b s - (n-s+\frac{1}{2})\log_b(n-s) - \frac{s^2}{2} - \frac{1}{2}\log_b 2\pi \\ &\geq s\left(\log_b(n-\log_b n) - \frac{s}{2} - \log_b \log_b n\right) - \frac{1}{2}\log_b s - \frac{1}{2}\log_b 2\pi \ > 0 \end{aligned}$$

when n is sufficiently large. Similarly, for $2\log_b n \leq s < n$,

$$\begin{aligned} \log_b \phi(n,s) &\leq s\left(\log_b(n-s) - \frac{s}{2} - \log_b s\right) - \frac{1}{2}\log_b s - \frac{1}{2}\log_b 2\pi + 2s + \frac{s\log_b s}{2} \\ &\leq s\left(2 - \frac{\log_b s}{2}\right) - \frac{1}{2}\log_b s - \frac{1}{2}\log_b 2\pi \ < \ 0 \end{aligned}$$

when n is sufficiently large. Thus, when n is sufficiently large, there exists a unique solution $s(n)$ of the equation $\phi(n, s) = 1$ and $s(n) \in (\log_b n, 2\log_b n)$.

Proof of Lemma 2.2.2: Lemma 2.2.2 is a special case of Lemma 2.11.2 of Section 2.11. It is omitted here.

2.7 Proof of Proposition 2.2.3

To establish the bound with n independent of r, it suffices to consider a sequence $r = r_n$ that changes with n in such a way that $1 \leq r_n \leq \gamma n$. Fix n for the moment, let $l = k(n) + r_n$,

and let $U_l(n)$ be the number of $l \times l$ submatrices of 1's in $\mathbf{Z}_n$. Then by Markov's inequality and Stirling's approximation,

$$P(M(\mathbf{Z}_n) \geq r) \;=\; P(U_l \geq 1) \;\leq\; E(U_l) \;=\; \binom{n}{l}^2 p^{l^2} \;\leq\; 2\phi(n,l)^2. \tag{2.9}$$

A straightforward calculation using the definition of $\phi(n,\cdot)$ shows that

$$2\phi(n,l)^2 \;=\; 2\,\phi^2(n,k(n))\,p^{r\cdot k(n)}\,[\,A_n(r)\,B_n(r)\,C_n(r)\,D_n(r)\,]^2,$$

where

$$A_n(r) \;=\; \left(\frac{n-r-k(n)}{n-k(n)}\right)^{-n+r+k(n)+\frac{1}{2}} \qquad B_n(r) \;=\; \left(\frac{r+k(n)}{k(n)}\right)^{-k(n)-\frac{1}{2}}$$

$$C_n(r) \;=\; \left(\frac{n-k(n)}{r+k(n)}\,p^{\frac{k(n)}{2}}\right)^{r} \qquad D_n(r) \;=\; p^{\frac{r^2}{2}}$$

Note that $p^{r\cdot k(n)} = o(n^{-2r}(\log_b n)^{3r})$, and that $\phi^2(n,k(n)) \leq 1$ by the monotonicity of $\phi(n,\cdot)$ and the definition of $k(n)$. Thus it suffices to show that $A_n(r)\cdot B_n(r)\cdot C_n(r)\cdot D_n(r) \leq 1$ when n is sufficiently large. To begin, note that for any fixed $\delta \in (0,1/2)$, when n is sufficiently large,

$$C_n(r)^{\frac{1}{r}} \;=\; \frac{n-k(n)}{r+k(n)}\,p^{\frac{k(n)}{2}} \;\leq\; \frac{n}{k(n)}\,p^{\frac{k(n)}{2}} \;\leq\; \frac{n}{(2-\delta)\log_b n}\,\frac{\frac{2+\delta}{2}\log_b n}{n}$$

which is less than one. In order to show $A_n(r)\cdot B_n(r)\cdot D_n(r) \leq 1$, we consider two possibilities for the asymptotic behavior of $r = r_n$.

Case 1: Suppose $r/k(n) \to 0$ as $n \to \infty$. In this case, $B_n(r)^{\frac{1}{r}} = (1+o(1))\,e^{-1}$. Moreover, $r/n \to 0$, which implies that $A_n(r)^{\frac{1}{r}} = (1+o(1))\,e$. Thus

$$A_n(r)\cdot B_n(r)\cdot D_n(r) \;=\; ((1+o(1))^2\,p^{\frac{r}{2}})^r \;\leq\; 1$$

when n is sufficiently large.

Case 2: Suppose $\liminf_n r/k(n) > 0$. In this case a routine calculation shows that

$B_n(r) \leq 1$ for any $r \geq 1$, so it suffices to show that

$$A_n(r) \cdot D_n(r) \ \leq \ 1. \tag{2.10}$$

Note that $D_n(r) = (p^{\frac{r}{2}})^r$ and $A_n(r)^{\frac{1}{r}} = (1 + o(1))\, e$ when $r = o(n - k(n))$. Thus (2.10) holds when $r = o(n - k(n))$.

It remains to consider the case $o(n - k(n)) < r < \gamma\, n$. As $\sqrt{(2 + \frac{2}{1-\gamma})\, n/\log b} = o(n - k(n))$, it suffices to assume that $\sqrt{(2 + \frac{2}{1-\gamma})\, n/\log b} < r < \gamma\, n$. In this case,

$$\begin{aligned} \log_b A_n(r) \cdot D_n(r) &= \log_b \left[\left(1 + \frac{r}{n - k(n) - r} \right)^{n-r-k(n)-\frac{1}{2}} p^{\frac{r^2}{2}} \right] \\ &\leq n \log_b \left(1 + \frac{r}{n - r - k(n)} \right) - \frac{(2 + \frac{2}{1-\gamma})\, n}{2 \log b} \ \leq \ 0, \end{aligned}$$

where the last inequality comes from the fact that $\log_b(1 + x) \leq x/\log b$ for $x \geq 0$.

2.8 Proof of Theorem 2.2.4

The proof of Theorem 2.2.4 shares ideas similar to those in the proof of Theorem 2.3.2 in the later section. However, since in Theorem 2.2.4, the range in which $M(\mathbf{Z}_n)$ possibly takes value can be further improved from $s(n) \pm \frac{5}{2}$ to $s(n) \pm \frac{3}{2}$, we list both proofs in this dissertation.

To show Theorem 2.2.4, we need the following definitions.

Definition: Fix an $0 < \epsilon < \frac{1}{2}$. For any $k \geq 1$, let n'_k be the least integer n satisfying

$$EU_k(n) \geq k^{3+\epsilon}, \tag{2.11}$$

and let n_k be the largest integer n satisfying

$$EU_k(n) \leq k^{-3-\epsilon}. \tag{2.12}$$

Note that n_k and n'_k always exist since for any fixed k, $EU_k(n)$ is monotone increasing with n, $EU_k(k) = p^{k^2} \leq k^{-3-\epsilon}$, and $EU_k(n) \to k^{3+\epsilon}$ as $n \to \infty$.

Lemma 2.8.1. *When k is sufficiently large,*

1. $n'_k < n_{k+1}$.
2. $n'_k - n_k < \frac{C_1 \ln k}{k} n_k$ *for some constant* $C_1 > 0$.
3. $\lim_{k\to\infty} \frac{n_{k+2}-n_{k+1}}{n_{k+1}-n_k} = b^{\frac{1}{2}}$.

Proof of 1 : By the definition of n_k, it follows that

$$\binom{n_k}{k} p^{\frac{k^2}{2}} \leq k^{-\frac{(3+\epsilon)}{2}},$$

which yields

$$\frac{k^{\frac{(3+\epsilon)}{2}}}{k!\; b^{\frac{k^2}{2}}} \leq \frac{1}{(n_k-k)^k},$$

since $\frac{(n_k-k)!}{n_k!} \leq \frac{1}{(n_k-k)^k}$. Thus,

$$n_k \leq b^{\frac{k}{2}} \left(\frac{k!}{k^{\frac{(3+\epsilon)}{2}}}\right)^{\frac{1}{k}} + k.$$

On the other hand, by the definition of n_k, it also follows that

$$\binom{n_k+1}{k} p^{\frac{k^2}{2}} \geq k^{-\frac{(3+\epsilon)}{2}}.$$

Thus,

$$k^{\frac{(3+\epsilon)}{2}} \geq b^{\frac{k^2}{2}} \frac{k!}{(n_k+1)^k},$$

which leads to

$$n_k \geq b^{\frac{k}{2}} \left(\frac{k!}{k^{\frac{3+\epsilon}{2}}}\right)^{\frac{1}{k}} - 1.$$

Putting the above two bounds on n_k together, we have

$$b^{\frac{k}{2}} \left(\frac{k!}{k^{\frac{3+\epsilon}{2}}}\right)^{\frac{1}{k}} - 1 \leq n_k \leq b^{\frac{k}{2}} \left(\frac{k!}{k^{\frac{(3+\epsilon)}{2}}}\right)^{\frac{1}{k}} + k. \tag{2.13}$$

Consequently, we have

$$n_k = b^{\frac{k}{2}} (k!)^{\frac{1}{k}} + o(k\, b^{\frac{k}{2}}) \tag{2.14}$$

since $\frac{\ln(k^{-\frac{3+\epsilon}{2}})}{k} \to 0$ as $k \to \infty$.

Similarly, one can verify that

$$b^{\frac{k}{2}} \left(k!\ k^{\frac{3+\epsilon}{2}}\right)^{\frac{1}{k}} - 1 \leq n'_k \leq b^{\frac{k}{2}} \left(k!\ k^{\frac{(3+\epsilon)}{2}}\right)^{\frac{1}{k}} + k. \tag{2.15}$$

and therefore

$$n'_k = b^{\frac{k}{2}} (k!)^{\frac{1}{k}} + o(k\, b^{\frac{k}{2}}). \tag{2.16}$$

The fact that when k is sufficiently large, $n'_k < n_{k+1}$ is apparent from the above approximations on n_k and n'_k.

Proof of 2 : From inequalities (2.13) and (2.15), it follows that

$$\begin{aligned} n'_k - n_k &\leq b^{\frac{k}{2}} \left(k!\ k^{\frac{(3+\epsilon)}{2}}\right)^{\frac{1}{k}} + k - [b^{\frac{k}{2}} \left(\frac{k!}{k^{\frac{3+\epsilon}{2}}}\right)^{\frac{1}{k}} - 1] \\ &\leq b^{\frac{k}{2}} \left(\frac{k!}{k^{\frac{3+\epsilon}{2}}}\right)^{\frac{1}{k}} (k^{\frac{3+\epsilon}{k}} - 1) + k + 1 \\ &< (n_k + 1)(k^{\frac{3+\epsilon}{k}} - 1) + k + 1 \\ &< n_k\, C_1\, \frac{\ln k}{k}, \end{aligned} \tag{2.17}$$

where the third inequality comes from (2.13) and the last inequality comes by the fact that when $x \to 1$, there exists a constant C_1 such that $C_1 \ln x > x - 1$ and by letting $x = k^{\frac{3+\epsilon}{k}}$.

Proof of 3 : From equations (2.14) and (2.16), one can verify that

$$\frac{n_{k+1}}{n_k} = b^{\frac{1}{2}} + o(1),$$

and

$$\frac{n_{k+2}}{n_{k+1}} = b^{\frac{1}{2}} + o(1).$$

Therefore,

$$\frac{n_{k+2} - n_{k+1}}{n_{k+1} - n_k} = \frac{\frac{n_{k+2}}{n_{k+1}} - 1}{1 - \frac{n_k}{n_{k+1}}} \to b^{\frac{1}{2}}. \tag{2.18}$$

We will use U_k instead of $U_k(n)$ in the context below for simplicity when it does not cause confusion.

Lemma 2.8.2. *Following the definition of $U_k(n)$, we immediately have*

$$EU_k(n) \ = \binom{n}{k}^2 p^{k^2}$$

and

$$EU_k(n)^2 \ = \sum_{l=1}^{k} \binom{n}{k}\binom{k}{l}\binom{n-k}{k-l} \sum_{r=1}^{k} \binom{n}{k}\binom{k}{r}\binom{n-k}{k-r} \cdot p^{2\,k^2-lr}.$$

By Lemma 2.8.2, one has

$$g(U_k(n)) := \frac{Var\, U_k(n)}{(EU_k(n))^2} = \sum_{l=0}^{k}\sum_{r=0}^{k} \frac{\binom{k}{l}\binom{n-k}{k-l}}{\binom{n}{k}} \frac{\binom{k}{r}\binom{n-k}{k-r}}{\binom{n}{k}} b^{lr} - 1,$$

where $b = p^{-1}$. Now, we want to bound $g(U_k(n))$ from above by the following lemma. First, fix any $0 < \epsilon < \frac{1}{2}$.

Lemma 2.8.3. *When k is sufficiently large, for every $n'_k < n < n_{k+1}$,*

$$g(U_k(n)) \le C_0 k^{-1-\epsilon}. \tag{2.19}$$

Proof of Lemma 2.8.3: Note that $\frac{\binom{k}{l}\binom{n-k}{k-l}}{\binom{n}{k}}$ is the probability mass function of a hypergeometric distribution. Thus,

$$\begin{aligned} g(U_k) &= \sum_{l=0}^{k}\sum_{r=0}^{k} \frac{\binom{k}{l}\binom{n-k}{k-l}}{\binom{n}{k}} \frac{\binom{k}{r}\binom{n-k}{k-r}}{\binom{n}{k}} (b^{lr} - 1) \\ &= \sum_{l=1}^{k}\sum_{r=1}^{k} \frac{\binom{k}{l}\binom{n-k}{k-l}}{\binom{n}{k}} \frac{\binom{k}{r}\binom{n-k}{k-r}}{\binom{n}{k}} (b^{lr} - 1) \\ &< \sum_{l=1}^{k}\sum_{r=1}^{k} \frac{\binom{k}{l}\binom{n-k}{k-l}}{\binom{n}{k}} \frac{\binom{k}{r}\binom{n-k}{k-r}}{\binom{n}{k}} b^{lr} \le \left(\sum_{r=1}^{k} \frac{\binom{k}{r}\binom{n-k}{k-r}}{\binom{n}{k}} (b^{r^2/2}) \right)^2, \end{aligned}$$

where the last inequality is obtained by $b^{lr} \le b^{\frac{l^2+r^2}{2}}$. Thus, in order to show Lemma 2.8.3,

it suffices to show

$$\sum_{r=1}^{k} h(r) = O(k^{-1/2-\epsilon/2}) \quad \text{where} \quad h(r) := \frac{\binom{k}{r}\binom{n-k}{k-r}}{\binom{n}{k}}\, b^{r^2/2}. \tag{2.20}$$

When $n \geq n'_k$, by inequality (2.15), it follows that $k \leq 2\log_b n$. Similarly, inequality (2.13) implies that if $n \leq n_{k+1}$ then $k \geq (2-\eta)\log_b n$ for some fixed $0 < \eta < 1/2$. Moreover, by the definition of n'_k, $n > n'_k$ implies that $\binom{n}{k} p^{\frac{k^2}{2}} = \sqrt{EU_k(n)} \geq \sqrt{EU_k(n'_k)} \geq k^{3/2+\epsilon/2}$. Using these inequalities, one can bound $h(1)$, $h(k-1)$ and $h(k)$ from above as follows.

Note that when $n > n'_k$, we have shown that $k \leq 2\log_b n$. Thus, a routine calculation shows that

$$h(1) = \frac{\binom{k}{1}\binom{n-k}{k-1}}{\binom{n}{k}} b^{1/2} = \frac{b^{1/2}k^2(n-k)!(n-k)!}{(n-2k+1)!n!} < \frac{b^{1/2}k^2}{n-k} = O(k^2 b^{-k/2}),$$

$$h(k-1) = \frac{k(n-k)}{\binom{n}{k}}\, b^{\frac{k^2}{2}-k+\frac{1}{2}} \leq \frac{knb^{\frac{1}{2}-k}}{\sqrt{EU_k(n)}} = O(k^{-1/2-\epsilon/2}\, b^{-k(1-\eta)/(2-\eta)})$$

$$h(k) = \frac{b^{\frac{k^2}{2}}}{\binom{n}{k}} = \frac{1}{\sqrt{EU_k(n)}} \leq k^{-3/2-\epsilon/2}.$$

In order to show inequality (2.20), it now suffices to verify that

$$h(r) \leq h(1) + h(k-1)$$

when n is sufficiently large and $k > r$ and $(2-\eta)\log_b n < k < 2\log_b n$.

By the definition of $h(\cdot)$, one has

$$\frac{h(r+1)}{h(r)} = \frac{(k-r)^2 b^{r+\frac{1}{2}}}{(r+1)\,(n-2k+r+1)}.$$

When $r \leq \frac{1}{3}k$, the inequality $k \leq 2\log_b n$ implies that

$$\frac{h(r+1)}{h(r)} \leq \frac{bk^2 b^{\frac{k}{3}}}{n-2k+r+1} \leq \frac{bk^2 n^{\frac{2}{3}}}{n-2k+r+1} < 1.$$

When $\frac{2}{3}k \leq r < k-1$ the inequality $k \geq (2-\eta)\log_b n$ with $0 < \eta < 1/2$ implies that

$$\frac{h(r+1)}{h(r)} \geq \frac{3b^{\frac{2k}{3}}}{2k(n+r+1)} \geq \frac{3n^{\frac{2(2-\eta)}{3}}}{2k(n+r+1)} > 1.$$

Now, we have shown that when $r \leq \frac{1}{3}k$

$$\frac{h(r+1)}{h(r)} \leq 1,$$

and when $r \geq \frac{2}{3}k$,

$$\frac{h(r+1)}{h(r)} > 1.$$

Note that for $r \in [\lceil \frac{k}{3} \rceil - 1, \lceil \frac{2k}{3} \rceil]$, it follows that

$$h(r) = h(\lceil \frac{2k}{3} \rceil) \times \frac{h(\lceil \frac{2k}{3} \rceil - 1)}{h(\lceil \frac{2k}{3} \rceil)} \times ... \times \frac{h(r)}{h(r+1)}$$

and

$$h(r) = h(\lceil \frac{k}{3} \rceil - 1) \times \frac{h(\lceil \frac{k}{3} \rceil)}{h(\lceil \frac{k}{3} \rceil)} \times ... \times \frac{h(r)}{h(r-1)}.$$

Thus, if $\frac{h(r+1)}{h(r)}$ is monotone increasing on $[\lceil \frac{k}{3} \rceil - 1, \lceil \frac{2k}{3} \rceil]$, then

$$h(r) \leq \max\{h(\lceil \frac{k}{3} \rceil - 1), h(\lceil \frac{2k}{3} \rceil)\} \leq h(1) + h(k-1).$$

To verify the monotonicity, note that the derivative of $\frac{h(r+1)}{h(r)}$ is given by

$$\begin{aligned} & b^{\frac{2r+1}{2}} \left[\frac{-2(k-r)(r+1)(n-2k+r+1) - (k-r)^2(2r+n-2k+2)}{(r+1)^2(n-2k+r+1)^2} + \frac{(k-r)^2 \ln b}{(r+1)(n-2k+r+1)}\right] \\ = \; & \frac{b^{\frac{2r+1}{2}}(k-r)}{(r+1)(n-2k+r+1)}\left[\frac{-2(r+1)(n-2k+r+1) - (k-r)(2r+n-2k+2)}{(r+1)(n-2k+1)} + (k-r)\ln b\right]. \end{aligned} \tag{2.21}$$

When k is sufficiently large and $n \gg k > r$, the sum of those leading terms in the right hand side of (2.21) is

$$-2n(r+1) - (k-r)n + (k-r)(r+1)n\ln b = n(-r^2 \ln b + kr \ln b - k - r + (k-r)\ln b - 2).$$

By plugging $r = \frac{k}{3}$ and $r = \frac{2k}{3}$ in, it is not hard to check that when k is sufficiently large, the above quadratic form is nonnegative for any $r \in [\lceil \frac{k}{3} \rceil - 1, \lceil \frac{2k}{3} \rceil]$. Thus, one can further conclude that the ratio $\frac{h(r+1)}{h(r)}$ is monotone increasing on $[\lceil \frac{k}{3} \rceil - 1, \lceil \frac{2k}{3} \rceil]$.

Lemma 2.8.4. *Fix any sufficiently large k. $M(\mathbf{Z}_n) = k$ with probability one if $n'_k \leq n \leq n_{k+1}$.*

Proof of Lemma 2.8.4: By Lemma 2.8.2 and Markov's inequality , we have

$$P(M(\mathbf{Z}_n) > k) = P(U_{k+1}(n) > 0) \ \leq \ E(U_{k+1}(n)) \leq \ \frac{1}{k^{3+\epsilon}}, \tag{2.22}$$

when $n \leq n_{k+1}$.

By Lemma 2.8.2 and Chebyshev inequality, we have

$$P(M(\mathbf{Z}_n) < k) = P(\mathbf{U}_k(n) = 0) \ \leq \ \frac{Var(U_k(n))}{E^2(U_k(n))} = \sum_{l=0}^{r} \sum_{k=0}^{r} \frac{\binom{r}{l}\binom{n-r}{r-l}}{\binom{n}{r}} \frac{\binom{r}{k}\binom{n-r}{r-k}}{\binom{n}{r}} b^{-lk} - 1 \tag{2.23}$$

When $n'_k \leq n \leq n_{k+1}$, it immediately follows from Lemma 2.8.3 that

$$P(M(\mathbf{Z}_n) < k) \ \leq \ O(k^{-1-\epsilon}). \tag{2.24}$$

Note that $M(\mathbf{Z}_n)$ is monotone increasing with n. Therefore,

$$\begin{aligned} & \sum_k P(\exists n \text{ s.t. } n'_k \leq n \leq n_{k+1} \text{ and } M(\mathbf{Z}_n) \neq k) \\ \leq \ & \sum_k P(M(\mathbf{Z}_{n'_k}) < k) + \sum_k P(M(\mathbf{Z}_{n_{k+1}}) > k) \\ \leq \ & \sum_k O(k^{-1-\epsilon}) < \infty. \end{aligned}$$

By Borel-Cantelli lemma, one can conclude that when k is sufficiently large, $M(\mathbf{Z}_n) = k$ with probability 1 if $n'_k \leq n \leq n_{k+1}$.

Proof of Theorem 2.2.4: For each $m \geq 1$, let $A_m = \cup_{n \geq m} B_n$ with $B_n = \{|M(\mathbf{Z}_n) -$

$s(n)| \geq \frac{3}{2}\}$, and define index sets

$$\begin{aligned} I_m^1 &= \{n \geq m \text{ such that } n'_k \leq n \leq n_{k+1} \text{ for some } k \notin (s(n) - 3/2, s(n) + 3/2)\}, \\ I_m^2 &= \{n \geq m \text{ such that } n'_k \leq n \leq n_{k+1} \text{ for some } k \in (s(n) - 3/2, s(n) + 3/2)\}, \\ I_m^3 &= \{n \geq m \text{ such that } n \text{ belongs to no interval } [n'_k, n_{k+1}] \text{ for } k \geq 1\}. \end{aligned}$$

Then $A_m = A_m^1 \cup A_m^2 \cup A_m^3$, where $A_m^j = \cup_{n \in I_m^j} B_n$. It suffices to show that $P(A_m^j) \to 0$ as m tends to infinity for $j = 1, 2, 3$.

First, we want to show that when m is sufficiently large A_m^1 is an empty set. It is easy to verify from the proof of Lemma 2.2.1 that $k \in (\log_b n, 2 \log_b n)$. Suppose that there exists a $\tilde{k}$ such that $n'_{\tilde{k}} \leq n \leq n_{\tilde{k}+1}$. Then, one can verify the following inequalities,

$$\begin{aligned} (1 + o(1))\phi^2(n, \tilde{k} + 1) &= EU_{\tilde{k}+1}(n) \leq EU_{\tilde{k}+1}(n_{\tilde{k}+1}) \leq \tilde{k}^{-3-\epsilon} \\ &<< 1 = \phi^2(n, s(n)), \end{aligned}$$

where the first inequality follows from the monotonicity of $EU_{\tilde{k}}(\cdot)$, the second inequality follows from the definition of $n_{\tilde{k}}$ and the last equality follows from definition of $s(n)$. Note that $\phi(n, s)$ is monotone decreasing when $s \in (\log_b n, 2 \log_b n)$. Thus, it yields that $s(n) \leq \tilde{k} + 1$. Similarly,

$$\begin{aligned} (1 + o(1))\phi^2(n, \tilde{k}) &= EU_{\tilde{k}}(n) \geq EU_{\tilde{k}}(n'_{\tilde{k}}) \geq \tilde{k}^{3+\epsilon} \\ &> (1 + o(1)) = (1 + o(1))\phi^2(n, s(n)), \end{aligned}$$

which implies $s(n) \geq \tilde{k}$. Putting two bounds on $\tilde{k}$ together, one has that under the condition of index I_m^1, $|\tilde{k} - s(n)| < \frac{3}{2}$, which implies A_m^1 is empty.

Consider the index set I_m^2. When $|k - s(n)| < 3/2$ the event B_n implies that $M(\mathbf{Z}_n) \neq k$. Since $n_k, n'_k \to \infty$ as $k \to \infty$, it follows from Lemma 2.14 that

$$A_m^2 \subseteq \bigcup_{k \geq \kappa(m)} \bigcup_{n'_k \leq n \leq n_{k+1}} \{M(\mathbf{Z}_n) \neq k\}$$

for some function $\kappa(\cdot)$ such that $\kappa(m) \to \infty$ as $m \to \infty$. Lemma 2.8.4 implies that the probability of the latter set tends to zero as $\kappa(m) \to \infty$.

It remains to show that $\lim_m P(A_m^3) = 0$. By the definition of n_l, n_l' and the monotonicity of $\binom{n}{l}$, one can verify that $n_l < n_l'$. Moreover, by Lemma 2.8.1, one can conclude that

$$n_l \to \infty \text{ and } n_l' \to \infty \text{ as } l \to \infty\,,$$

and

$$... < n_l < n_l' < n_{l+1} < n_{l+1}' < ...$$

Thus, for any sufficiently large n, if there does not exist any l s.t. $n_l' \le n \le n_{l+1}$, then there must exist a l such that $n_l < n < n_l'$. Note that $M(\mathbf{Z}_n)$ is monotone increasing when n is increasing. Therefore, if $n_l < n < n_l'$ holds for any certain l, then $M(\mathbf{Z}_{n_l}) \le M(\mathbf{Z}_n) \le M(\mathbf{Z}_{n_l'})$. Moreover, from Lemma 2.8.4, we have with probability one,

$$l - 1 = M(\mathbf{Z}_{n_l}) \le M(\mathbf{Z}_n) \le M(\mathbf{Z}_{n_l'}) = l.$$

In order to show $\lim_m P(A_m^3) = 0$, it now suffices to show that $|l - s(n)| < \frac{3}{2}$ and $|l - 1 - s(n)| < \frac{3}{2}$. Note that the argument above for A_m^2 implies $|l - s(n_l')| < \frac{3}{2}$. Thus, it suffices to show $|s(n_l') - s(n)| \le |s(n_l') - s(n_l)| = o(1)$. Note that

$$s(n_l') - s(n_l) = 2\log_b \frac{n_l'}{n_l} - 2\log_b \frac{\log_b n_l'}{\log_b n_l} > 0.$$

When n_l' and n_l are sufficiently large, it is easy to check that $\frac{n_l'}{n_l} \ge \frac{\log_b n_l'}{\log_b n_l} > 1$ using the fact that function $f(x) = \frac{x}{\log_b x}$ is increasing for all sufficiently large x. Thus, it suffices to show that $\log_b \frac{n_l'}{n_l} = o(1)$. By 2 of Lemma 2.8.1, $n_l' = n_l(1 + C_1 \frac{\log l}{l})$, it follows immediately that $|s(n_l') - s(n_l)| = o(1)$. Thus $|l - s(n)| < \frac{3}{2}$ holds. Similarly, one can show $|l - 1 - s(n)| < \frac{3}{2}$.

2.9 Proof of Theorem 2.2.5

In order to establish Theorem 2.2.5, we begin with a definition and a lemma below.

Definition: Fix $0 < \epsilon < \frac{1}{2}$. For any k, define $n_k^* = \lceil b^{\frac{k}{1+\epsilon}} \rceil$.

Lemma 2.9.1. *Fix any* $0 < \epsilon < \frac{1}{2}$. *Eventually almost surely,* $\frac{1-\epsilon}{1+\epsilon}k < L(\mathbf{Z}_{n_k^*}) \leq k$.

Proof of Lemma 2.9.1: To establish the result in the above lemma, we will first show that eventually almost surely, $L(\mathbf{Z}_{n_k^*}) \leq k$, which is equivalent to

$$\lim_K P\left(\bigcup_{k\geq K} L(\mathbf{Z}_{n_k^*}) > k\right) = 0. \tag{2.25}$$

Let $\tilde{U}_l(n_k^*)$ be the number of $l \times l$ maximal submatrices of 1's in $\mathbf{Z}_{n_k^*}$ and not contained by any other square submatrices of 1's. It is clear that $\{L(\mathbf{Z}_{n_k^*}) > k\} \subset \{\tilde{U}_k(n_k^*) = 0\}$. Thus, to show (2.25), it suffices to show that

$$\lim_K P\left(\bigcup_{k\geq K} \{\tilde{U}_k(n_k^*) = 0\}\right) = 0. \tag{2.26}$$

Note that it can be verified that for any $0 \leq l \leq n$,

$$E(\tilde{U}_l(n)) = \binom{n}{l}^2 [2(1-p^l)^{(n-l)} - (1-p^l)^{2\,(n-l)}]\, p^{l^2}. \tag{2.27}$$

Moreover, by definition, it follows that $\tilde{U}_l(n) \leq U_l(n)$, where $U_l(n)$ is the number of $l \times l$ submatrices of 1's in $\mathbf{Z}_n$ with no other restriction. Thus,

$$E(\tilde{U}_l^2) \leq E(U_l^2) = \sum_{s=1}^{l} \binom{n}{l}\binom{l}{s}\binom{n-l}{l-s} \sum_{r=1}^{l} \binom{n}{l}\binom{l}{r}\binom{n-l}{l-r} p^{l^2-sr}. \tag{2.28}$$

Consequently,

$$\frac{E(\tilde{U}_l^2)}{E(\tilde{U}_l)^2} \leq [2(1-p^l)^{(n-l)} - (1-p^l)^{2\,(n-l)}]^{-2} \sum_{s=0}^{l}\sum_{r=0}^{l} \frac{\binom{l}{s}\binom{n-l}{l-s}}{\binom{n}{l}} \frac{\binom{l}{r}\binom{n-l}{l-r}}{\binom{n}{l}} b^{sr}. \tag{2.29}$$

Now, by a standard second moment argument and Borel-Cantelli lemma, in order to establish (2.26), it suffices to show that

$$\sum_k P\left(\tilde{U}_k(n_k^*) = 0\right) \leq \sum_k \frac{Var(\tilde{U}_k(n_k^*)^2)}{E(\tilde{U}_k(n_k^*))^2} < \infty. \tag{2.30}$$

Note that it is not hard to show that when k is sufficiently large,

$$(n_k^* - k)\ln(1-p^k) = (n_k^* - k)\ln(1 - {n_k^*}^{-1-\epsilon}) = -C_2 {n_k^*}^{-\epsilon},$$

where $C_2 > 0$ is a constant. When $l = k$ and $n = n_k^*$, since $(1-p^k)^{(n_k^*-k)} \to 1$ as $n_k^* >> k \to \infty$,

$$(2(1-p^l)^{(n-l)} - (1-p^l)^{2\,(n-l)})^{-2} - 1 \leq 4[2(1-p^l)^{(n-l)} - (1-p^l)^{2\,(n-l)} - 1] = O({n_k^*}^{-\epsilon})$$

for sufficiently large k. Consequently, the right hand side of inequality (2.29) is equal to

$$\sum_{s=1}^{k}\sum_{r=1}^{k} \frac{\binom{k}{s}\binom{n_k^*-k}{k-s}}{\binom{n_k^*}{k}} \frac{\binom{k}{r}\binom{n_k^*-k}{k-r}}{\binom{n_k^*}{k}} b^{sr} \cdot C_0,$$

for some constant $C_0 > 0$. Note that for any fixed $0 < \epsilon < \frac{1}{2}$, if $k = (1+\epsilon)\log_b n_k^*$ then $n_k' \leq n_k^* \leq n_{k+1}$, where n_k and n_k' follow the same definitions as those in Lemma 2.8.3. Therefore, it is clear from the proof of Lemma 2.8.3 and the definition of n_k^* that when k is sufficiently large,

$$\sum_{k=1}^{\infty}\left(\sum_{s=1}^{k}\sum_{r=1}^{k} \frac{\binom{k}{s}\binom{n_k^*-k}{k-s}}{\binom{n_k^*}{k}} \frac{\binom{k}{r}\binom{n_k^*-k}{k-r}}{\binom{n_k^*}{k}} b^{sr} \cdot C_0\right) < \infty,$$

which implies inequality (2.30).

Now, we wish to show $L(\mathbf{Z}_{n_k^*}) > (1-\epsilon)\log_b n_k^*$ eventually almost surely. This is equivalent to

$$P\left(\bigcup_{k=1}^{\infty}\bigcup_{l\leq(1-\epsilon)\log_b n_k^*}\{\tilde{U}_l(n_k^*) > 0\}\right) \leq \sum_{k=1}^{\infty}\sum_{l=1}^{(1-\epsilon)\log_b n_k^*} E(\tilde{U}_l(n_k^*)) < \infty, \tag{2.31}$$

where the first inequality follows from a standard first moment argument. It is easy to check that

$$E(\tilde{U}_l(n_k^*)) \leq \binom{n_k^*}{l}^2 (1-p^l)^{(n_k^*-l)} p^{l^2} =: E^*(l).$$

By Stirling approximation, one can verify the inequality below for any sufficiently large

k and $l = (1-\epsilon)\log_b n_k^*$.

$$\begin{aligned}\frac{\log_b E^*(l)^{\frac{1}{2}}}{l} &= (\frac{n_k^*}{l}+\frac{1}{2l})\log_b\frac{n_k^*}{n_k^*-l}+\log_b\frac{n_k^*-l}{l}-\frac{l}{2}-\frac{n_k^*-l}{2l}\log_b(1-p^l)+O(1)\\ &\leq 2\log_b n_k^* - O(n_k^{*\epsilon})+O(1)\leq -\gamma, \text{ where } \gamma>2. \end{aligned} \tag{2.32}$$

It can also be shown that when n_k^* is sufficiently large and $l \leq (1-\epsilon)\log_b n_k^*$,

$$\begin{aligned}\sqrt{\frac{E^*(l)}{E^*(l+1)}} &= (\frac{l+1}{n_k^*-l})(\frac{(1-p^{l+1})^{\frac{1}{2}}}{p^{l+\frac{1}{2}}})(\frac{1-p^l}{1-p^{l+1}})^{\frac{n-l}{2}}\\ &< (\frac{l+1}{n_k^*-l})(\frac{(1-p^{l+1})^{\frac{1}{2}}}{p^{l+\frac{1}{2}}})\\ &< 1 \end{aligned} \tag{2.33}$$

Therefore,by putting (2.32) and (2.33) together, one can obtain inequality (2.31).

Proof of Theorem 2.2.5: For any fixed $0<\epsilon<\frac{1}{2}$. Let $0<\epsilon'<\epsilon<\frac{1}{2}$. In this proof, n_k^* will be defined based on constant ϵ'. Note that by the definition of n_k^*, it is clear that as $k\to\infty$, $n_k^*\to\infty$, and that $n_k^*\leq n_{k+1}^*$ for any $k\geq 1$. Moreover, by the definition of $L(\mathbf{Z}_n)$, eventually almost surely, $L(\mathbf{Z}_n)\geq L(\mathbf{Z}_{n'})$ for any sufficiently large pair $n>n'$. Therefore,

$$\begin{aligned}&\lim_m P\left(\bigcup_{n\geq m}\{L(\mathbf{Z}_n)>(1+\epsilon)\log_b n\}\right)+\lim_m P\left(\bigcup_{n\geq m}\{L(\mathbf{Z}_n)\leq(1-\epsilon)\log_b n\}\right)\\ \leq\ &\lim_K P\left(\bigcup_{k\geq K}\{(1+\epsilon)\log_b n<L(\mathbf{Z}_n)\leq L(\mathbf{Z}_{n_{k+1}^*})\text{ when } n_k^*\leq n<n_{k+1}^*\}\right)+\\ &\lim_K P\left(\bigcup_{k\geq K}\{L(\mathbf{Z}_n)\leq L(\mathbf{Z}_{n_k^*})\leq(1-\epsilon)\log_b n\text{ when } n_k^*\leq n<n_{k+1}^*\}\right). \end{aligned} \tag{2.34}$$

Note that $\log_b n-\log_b n_k^*\leq \log_b n_{k+1}^*-\log_b n_k^*=1$ and $\log_b n_{k+1}^*-\log_b n\leq \log_b n_{k+1}^*-\log_b n_k^*=1$. Moreover, by definition, $\epsilon'<\epsilon$, it is then easy to check that when k is sufficiently large,

$$(1+\epsilon)\log_b n\geq(1+\epsilon)\log_b n_k^*>(1+\epsilon')\log_b n_{k+1}^*$$

and

$$(1-\epsilon)\log_b n\leq(1-\epsilon)\log_b n_{k+1}^*<(1-\epsilon')\log_b n_k^*,$$

when n is sufficiently large. Note that Lemma 2.9.1 implies

$$\begin{aligned}
&\lim_K P\left(\bigcup_{k\geq K}\{(1+\epsilon')\log_b n_{k+1}^* < L(\mathbf{Z}_n) \leq L(\mathbf{Z}_{n_{k+1}^*}) \text{ when } n_k^* \leq n < n_{k+1}^*\}\right)\\
= &\lim_K P\left(\bigcup_{k\geq K}\{L(\mathbf{Z}_n) \leq L(\mathbf{Z}_{n_k^*}) \leq (1-\epsilon')\log_b n_k^* \text{ when } n_k^* \leq n < n_{k+1}^*\}\right) = 0.
\end{aligned}$$

Therefore, $(1-\epsilon)\log_b n < L(\mathbf{Z}_n) \leq (1+\epsilon)\log_b n$ eventually almost surely.

2.10 Proof of Proposition 2.3.1

Fix n, let $r = k_1(n) + k$ and $U_r(\rho)$ be the number of $r\,\rho \times r$ submatrices of 1's in Y_{mn}. Note that $U_r(\rho)$ is an integer, so $E(U_r(\rho)) \geq 1 \cdot \sum_{i\geq 1} P(U_r = i)$. Then it follows that

$$P\left(M_\rho(\mathbf{Z}_{mn}) \geq r\right) = P\left(U_r > 0\right) \leq E\left(U_r(\rho)\right) = \binom{n}{r} \cdot \binom{\alpha n}{\rho r} \cdot p^{\rho\, r^2}. \tag{2.35}$$

When n is sufficiently large, $1 \leq k \leq \gamma\, n$, and $\log_{p^{-1}} n < k_1(n) < \frac{\rho+1}{\rho}\log_{p^{-1}} n$, one can apply Stirling approximation to show that

$$\begin{aligned}
E(U_r) &\leq 2\,[(2\pi)^{-\frac{1}{2}}\, n^{n+\frac{1}{2}}\, (n-r)^{-n+r+\frac{1}{2}}\, (r)^{-r-\frac{1}{2}}\, p^{\frac{\rho\,(r)^2}{2}}] \cdot\\
&\quad\ [(2\pi)^{-\frac{1}{2}}\, \alpha n^{\alpha n+\frac{1}{2}}\, (\alpha n-\rho r)^{-\alpha n+\rho r+\frac{1}{2}}\, (\rho r)^{-\rho r-\frac{1}{2}}\, p^{\frac{\rho\,(r)^2}{2}}]\\
&= 2\,E(U_{k_1(n)})\, p^{\rho\, k\, k_1(n)}\, [A(k)\, B(k)\, C(k)\, D(\rho\, k)]\\
&\quad\ \times [A_{\alpha n}(\rho\, k)\, B_{\alpha n}(\rho\, k)\, C_{\alpha n}(\rho\, k)\, D_{\alpha n}(\rho\, k)],
\end{aligned}$$

where $A(k) = (\frac{n-r}{n-k_1(n)})^{-n+r-\frac{1}{2}}$, $\quad B(k) = (\frac{r}{k_1(n)})^{-k_1(n)-\frac{1}{2}}$,
$C'(k) = (\frac{n-k_1(n)}{r} p^{\frac{\rho\, k_1(n)}{2}})^k$ and $D(\rho\, k) = p^{\frac{\rho\, k^2}{2}}$;
$A_{\alpha n}(\rho\, k) = (\frac{\alpha n-\rho r}{\alpha n-\rho k_1(n)})^{-\alpha n+\rho r-\frac{1}{2}}$, $B_{\alpha n}(\rho\, k) = (\frac{r}{k_1(n)})^{-\rho k_1(n)-\frac{1}{2}}$,
$C_{\alpha n}(\rho\, k) = (\frac{\alpha n-\rho k_1(n)}{\rho r} p^{\frac{k_1(n)}{2}})^{\rho k}$ and $D_{\alpha n}(\rho\, k) = p^{\frac{\rho\, k^2}{2}}$.

Since by definition, $k_1(n)$ is the solution to $E(U_{k_1(n)}(\rho)) = 1$, and when n is sufficiently large, $p^{\frac{\rho\, k k_1}{2}} \leq \frac{(\rho+2)k\, \log_{p^{-1}} n}{n^{2\rho\, k}}$, we only need to show that

$$A(k) \cdot B(k) \cdot C(k) \cdot D(\rho\, k) \leq 1$$

and

$$A_{\alpha n}(\rho k) \cdot B_{\alpha n}(\rho k) \cdot C_{\alpha n}(\rho k) \cdot D_{\alpha n}(\rho k) \leq 1,$$

when n is large.

Recall $C(k)$ and $D(k)$. Since $\rho \geq 1$, it follows that

$$D(\rho\, k) \leq D(k) \text{ and } C'(k) \leq C(k).$$

Then from the argument in the proof of Proposition 2.2.3, it concludes that when n is sufficiently large and $1 \leq k \leq \gamma\, n$,

$$A(k) \cdot B(k) \cdot C'(k) \cdot D(\rho\, k) \leq 1.$$

To show $A_{\alpha n}(\rho k) \cdot B_{\alpha n}(\rho k) \cdot C_{\alpha n}(\rho k) \cdot D_{\alpha n}(\rho k) \leq 1$, we have the following arguments.

First, fix any $0 < \delta < 1$ and for any sufficiently large n, we have the following inequalities,

$$\begin{aligned} C_{\alpha n}(\rho k)^{\frac{1}{\rho k}} &= \frac{\alpha n - \rho k_1(n)}{\rho k + \rho k_1(n)}\, p^{\frac{k_1(n)}{2}} \leq \frac{\alpha n}{\rho k_1(n)}\, p^{\frac{k_1(n)}{2}} \\ &\leq \frac{\alpha n}{(\rho+1)\log_{p^{-1}} n - (\rho+1+\delta)\log_{p^{-1}} \log_b n}\, \frac{\frac{(\rho+1+\delta)}{2\rho}\log_{p^{-1}} n}{\frac{\rho+1}{2\rho} n} \\ &\leq (1+o(1))\frac{\alpha}{1+\rho}, \end{aligned}$$

where the second inequality holds by the fact that

$$k_1(n) \geq \frac{\rho+1}{\rho}\log_{p^{-1}} n - (\frac{\rho+1}{\rho} + \delta)\log_{p^{-1}} \log_{p^{-1}} n.$$

In order to show

$$A_{\alpha n}(\rho k) \cdot B_{\alpha n}(\rho k) \cdot D_{\alpha n}(\rho k) \leq (\frac{2\,\alpha}{1+\rho})^{-\rho k}$$

when $\rho < 2\,\alpha - 1$, and

$$A_{\alpha n}(\rho k) \cdot B_{\alpha n}(\rho k) \cdot C_{\alpha n}(\rho k) \cdot D_{\alpha n}(\rho k) \leq 1$$

when $\rho \geq 2\,\alpha - 1$, we consider the following three cases.

Case 1: When $\frac{k}{k_1(n)} \to 0$ as $n \to \infty$, it follows that when n is sufficiently large

$$A_{\alpha n}(\rho k)^{\frac{1}{\rho k}} = (1 + o(1))\, e \text{ and } B_{\alpha n}(\rho k)^{\frac{1}{\rho k}} = [(1 + o(1))\, e]^{-1}.$$

Therefore

$$A_{\alpha n}(\rho k) \cdot B_{\alpha n}(\rho k) \cdot D_{\alpha n}(\rho k) \leq ((1 + o(1))\, p^{\frac{k}{2}})^{\rho k} \leq \min\left\{(\frac{2\,\alpha}{1+\rho})^{-\rho k}, 1\right\},$$

when n is sufficiently large.

Case 2: When $\sqrt{(2\alpha + \delta)\, n} \leq k \leq \gamma\, n$ and n is sufficiently large, it follows

$$\begin{aligned}
\log_{p^{-1}} A_{\alpha n}(\rho k) \cdot D_{\alpha n}(\rho k) &= \log_{p^{-1}} \left((1 + \frac{\rho k}{\alpha n - \rho k_1(n) - \rho k})^{\alpha n - \rho k_1 - \rho k + \frac{1}{2}} \cdot p^{\frac{\rho k^2}{2}} \right) \\
&\leq \alpha n \log_{p^{-1}} \left(1 + \frac{\rho k}{\alpha n - \rho k - \rho k_1(n)} \right) - \frac{(2\alpha + \delta)\, n}{2} \\
&\leq 0.
\end{aligned}$$

When $\sqrt{(2\alpha + \delta)\, n} \leq k \leq \gamma\, n$ and n is sufficiently large, we have

$$B_{\alpha n}(\rho k) \leq \left[\frac{\sqrt{(2\alpha + \delta)\, n}}{\frac{\rho+2}{\rho} \log_{p^{-1}} n} \right]^{-\rho k} \leq \min\left\{(\frac{2\,\alpha}{1+\rho})^{-\rho k}, 1\right\}.$$

Therefore, it follows that $A_\alpha(\rho k) \cdot B_\alpha(\rho k) \cdot D_\alpha(\rho k) \leq \min\left\{(\frac{2\,\alpha}{1+\rho})^{-\rho k}, 1\right\}$.

Case 3: When $\liminf \frac{k}{k_1(n)} > 0$ as $n \to \infty$ and $k < \sqrt{(2\alpha + \delta)\, n}$, it follows that $A_{\alpha n}(\rho k)^{\frac{1}{\rho k}} = (1 + o(1))\, e$ when n is sufficiently large. Therefore

$$A_{\alpha n}(\rho k) \cdot D_{\alpha n}(\rho k) \leq ((1 + o(1))e \cdot p^{\frac{k}{2}})^{\rho k} \leq (\frac{2\,\alpha}{1+\rho})^{-\rho k}, \tag{2.36}$$

when n is sufficiently large.

Note that when $k > 2\log_{p^{-1}}\frac{2\alpha}{1+\rho}$, (2.36) implies $((1+o(1))e\cdot p^{\frac{k}{2}})^{\rho k} \le \min\left\{(\frac{2\alpha}{1+\rho})^{-\rho k}, 1\right\}$. When $k \le 2\log_{p^{-1}}\frac{2\alpha}{1+\rho}$, (2.36) implies $A_{\alpha n}(\rho k)\cdot B_{\alpha n}(\rho k)\cdot D_{\alpha n}(\rho k) \le 1$ and $C_{\alpha n}(\rho k) \le (\frac{2\alpha}{1+\rho})^{2\log_{p^{-1}}\frac{2\alpha}{1+\rho}} = \Delta(\alpha,\rho,p)$. Therefore, the probability bound holds.

2.11 Proof of Theorem 2.3.2

Lemma 2.11.1. *When both n and m are sufficiently large, equation (2.5) has a unique root $s(m,n,\beta)$. Moreover, $s(m,n,\beta) \in (\log_b n + \frac{\beta}{\beta+1}\log_b\frac{m}{n}, \frac{\beta+1}{\beta}\log_b n + \log_b\frac{m}{n})$.*

Proof of Lemma 2.11.1: It is trivial to verify that

$$\frac{\partial \log_b(\phi(s,m,n,\beta))}{\partial s} = \log_b(n-s) + \log_b(m-\beta s) - 2\beta s - \log_b s - \log_b \beta s + O(1),$$

which is negative and bounded away from zero when $\frac{\log_b mn}{2\beta} < s(m,n,\beta) < \frac{\beta+1}{\beta}\log_b n + \log_b\frac{m}{n}$ and m, n are sufficiently large. Moreover, it is clear from the definition of $\phi(\cdot)$ that

$$\begin{aligned}\log_b \phi(s,m,n,\beta) &= (n+\frac{1}{2})\log_b n + (m+\frac{1}{2})\log_b m - (s+\frac{1}{2})\log_b s - (\beta s+\frac{1}{2})\log_b \beta s\\ &\quad -(n-s+\frac{1}{2})\log_b(n-s) - (m-\beta s+\frac{1}{2})\log_b(m-\beta s) - \beta s^2 - \frac{1}{2}\log_b 2\pi\\ &= s\log_b(n-s) + \beta s\log_b(m-\beta s) - \beta s^2 - (\beta+1)s\log_b s + O(s).\end{aligned}$$

It is easy to check that

$$s\log_b(n-s) + \beta s\log_b(m-\beta s) - \beta s^2 < s((\beta+1)\log_b n + \beta\log_b\frac{m}{n} - \beta s), \tag{2.37}$$

which is negative when $s \ge \frac{\beta+1}{\beta}\log_b n + \log_b\frac{m}{n}$, and that

$$s\log_b(n-s) + \beta s\log_b(m-\beta s) - \beta s^2 = s((\beta+1)\log_b n + \beta\log_b\frac{m}{n} - \beta s) + o(1), \tag{2.38}$$

which is positive when $s \le \log_b n + \frac{\beta}{\beta+1}\log_b\frac{m}{n}$. Now, to show the uniqueness and existence of the root, it suffices to check whether the lower bound in monotone interval of $\phi(\cdot)$,$\frac{\log_b mn}{2\beta}$, is less than $\log_b n + \frac{\beta}{\beta+1}\log_b\frac{m}{n}$ in above, which is obvious.

Lemma 2.11.2. *When n is sufficiently large, a routine analysis shows that*

$$s(m,n,\beta) \ = \frac{\beta+1}{\beta}\log_b n - \frac{\beta+1}{\beta}\log_b\log_b n + \log_b\frac{m}{n} + C(\beta) + o(1), \tag{2.39}$$

where $C(\beta)$ is some constant depending only on β and $b = p^{-1}$.

Proof: Recall that by definition,

$$\phi(s,m,n,\beta) = 2\pi n^{n+\frac{1}{2}} m^{m+\frac{1}{2}} s^{-s-\frac{1}{2}} (\beta s)^{-\beta s-\frac{1}{2}} (n-s)^{-(n-s)-\frac{1}{2}} (m-\beta s)^{-(m-\beta s)-\frac{1}{2}} p^{\beta s^2}.$$

Taking logarithm on both sides and after simple algebra, one can obtain that

$$\begin{aligned} & \frac{1}{2}\log_b\frac{n}{n-s} + \frac{1}{2}\log_b\frac{m}{m-\beta s} + n\log_b\frac{n}{n-s} + m\log_b\frac{m}{m-\beta s} \\ & -(s+\frac{1}{2})\log_b s - (\beta s+\frac{1}{2})\log_b \beta s + s\log_b(n-s) + \beta s\log_b(m-\beta s) - \beta s^2 \\ = \ & -\log_b 2\pi. \end{aligned}$$

Note that Lemma 1 implies that $s(m,n,\beta)$ belongs to interval

$$(\log_b n + \frac{\beta}{\beta+1}\log_b\alpha, \frac{\beta+1}{\beta}\log_b n + \log_b\alpha),$$

where $\alpha = \frac{m}{n}$. Thus, we will only consider the above equation for those $s << n$. Divide both sides of the above equation by s. It follows that

$$\beta\log_b(m-\beta s) + \log_b(n-s) - \beta s - (1+\beta)\log_b s = -(1+\beta)\log_b e + \beta\log_b\beta + O(\frac{\log_b s}{s}),$$

which can be rewritten as

$$\begin{aligned} & \beta\log_b m + \log_b n - \beta s - (1+\beta)\log_b\frac{1+\beta}{\beta}\log_b n \\ = \ & (1+\beta)\log_b\frac{s}{\frac{1+\beta}{\beta}\log_b n} - \beta\log_b\frac{m-\beta s}{m} - \log_b\frac{n-s}{n} \\ & -(1+\beta)\log_b e + \beta\log_b\beta + O(\frac{\log_b s}{s}). \end{aligned} \tag{2.40}$$

For any m, n and β, define $R(m,n,\beta)$ as the function of (m,n,β) satisfying

$$s(m,n,\beta) = \frac{\beta+1}{\beta}\log_b n - \frac{\beta+1}{\beta}\log_b\left(\frac{\beta+1}{\beta}\log_b n\right) + \log_b \alpha + R(m,n,\beta).$$

Since we have shown in Lemma 1 that $s(m,n,\beta)$ uniquely exists, to obtain the correct value of $R(m,n,\beta)$, one can directly plug the expression of $s(m,n,\beta)$ into (2.40). It is clear that $R(m,n,\beta)$ must be independent of m and n, and $R(m,n,\beta) = \frac{(1+\beta)\log_b e - \beta\log_b \beta}{\beta} + o(1)$. Therefore, we have shown that

$$s(m,n,\beta) = \frac{\beta+1}{\beta}\log_b n - \frac{\beta+1}{\beta}\log_b\left(\frac{\beta+1}{\beta}\log_b n\right) + \log_b \alpha + R(\beta) + o(1).$$

To study $M(\mathbf{Z},\beta)$, we define $n_k(\alpha,\beta)$ and $n'_k(\alpha,\beta)$ in a fashion analogous to n_k and n'_k in Section 2.8 respectively. In the rest of this section, we will use $EU_k(n)$ or EU_k instead of $EU_k(m,n,\beta)$ when they do not cause confusion.

Definition: Fix $\alpha > 0$ and $\beta > 1$. For any $k \geq 1$, let $n'_k(\alpha,\beta)$ be the least integer n s.t.

$$EU_k(\lceil \alpha n \rceil, n, \beta) \geq k^4. \tag{2.41}$$

Let $n_k(\alpha,\beta)$ be the largest integer n s.t.

$$EU_k(\lceil \alpha n \rceil, n, \beta) \leq k^{-4}. \tag{2.42}$$

The existence of $n_k(\alpha,\beta)$ and $n'_k(\alpha,\beta)$ is easy to check for any fixed α, β and $k > 0$ by arguments similar to those in Section 2.8. In fact, when $\alpha(n)$ is a function of n, as long as $m = \lceil \alpha n \rceil$ is non-decreasing in n, n_k and n'_k remain well defined.

Lemma 2.11.3. *Let $b = p^{-1}$. When k is sufficiently large,*

1. $n'_k(\alpha,\beta) < n_{k+1}(\alpha,\beta)$.
2. $\lim_{k\to\infty} \frac{n_{k+2}(\alpha,\beta) - n_{k+1}(\alpha,\beta)}{n_{k+1}(\alpha,\beta) - n_k(\alpha,\beta)} = b^{\frac{\beta}{\beta+1}}$.

Proof of 1: To begin, we find upper bounds on n_k. By definition of n_k, it follows that

$$EU_k(\lceil \alpha n_k \rceil, n, \beta) = \binom{\lceil \alpha n_k \rceil}{\lceil \beta k \rceil} \binom{n_k}{k} p^{\lceil \beta k \rceil k} \leq k^{-4}.$$

Simply using $\frac{(n_k-k)!}{n_k!} \leq \frac{1}{(n_k-k)^k}$ and $\frac{(\lceil \alpha n_k \rceil - \lceil \beta k \rceil)!}{\lceil \alpha n_k \rceil !} \leq \frac{1}{(\lceil \alpha n_k \rceil - \lceil \beta k \rceil)^{\lceil \beta k \rceil}}$, the above inequality yields

$$\frac{k^4}{\lceil \beta k \rceil ! k! \; b^{\lceil \beta k \rceil k}} \leq \frac{1}{(n_k-k)^k (\lceil \alpha n_k \rceil - \lceil \beta k \rceil)^{\lceil \beta k \rceil}} \leq \frac{1}{(n_k-k)^k (\alpha n_k - \beta k - 1)^{\lceil \beta k \rceil}}.$$

Rearranging the above inequality, one has $\beta k + 1 \leq \alpha k$,

$$n_k \leq \alpha^{-\frac{\beta}{\beta+1}} b^{\frac{\beta k+1}{\beta+1}} \left(\frac{k! \lceil \beta k \rceil !}{k^4} \right)^{\frac{1}{(\beta+1)k}} + k, \tag{2.43}$$

and when $\beta k + 1 > \alpha k$,

$$n_k \leq \alpha^{-\frac{\beta}{\beta+1}} b^{\frac{\beta k+1}{\beta+1}} \left(\frac{k! \lceil \beta k \rceil !}{k^4} \right)^{\frac{1}{(\beta+1)k}} + \frac{\beta}{\alpha} k - \alpha^{-1}. \tag{2.44}$$

Now, we look for the lower bounds on n_k. By the definition of n_k,

$$EU_k(\lceil \alpha(n_k+1) \rceil, n_{k+1}, \beta) = \binom{\lceil \alpha(n_k+1) \rceil}{\lceil \beta k \rceil} \binom{n_k+1}{k} p^{\lceil \beta k \rceil k} \geq k^{-4},$$

which implies

$$k^4 \geq b^{\beta k^2} \frac{k! \lceil \beta k \rceil !}{(n_k+1)^k (\alpha(n_k+1)+1)^{\beta k+1}} \geq b^{\beta k^2} \frac{k! \lceil \beta k \rceil !}{(n_k+1+\alpha^{-1})^k (\alpha(n_k+1)+1)^{\beta k+1}}.$$

The above inequality leads to

$$(n_k+1+\alpha^{-1})(\alpha(n_k+1)+1)^{\beta+\frac{1}{k}} \geq b^{\beta k} \left(\frac{k! \lceil \beta k \rceil !}{k^4} \right)^{\frac{1}{k}}.$$

Thus,

$$n_k \geq \alpha^{-\frac{\beta+k^{-1}}{\beta+1+k^{-1}}} b^{\frac{\beta k}{\beta+1+k^{-1}}} \left(\frac{k! \lceil \beta k \rceil !}{k^4} \right)^{\frac{1}{k(\beta+1+k^{-1})}} - 1 - \frac{1}{\alpha}. \tag{2.45}$$

Next, we turn our attention to n'_k. Similarly, one can verify that

$$n'_k \geq \alpha^{-\frac{\beta+k^{-1}}{\beta+1+k^{-1}}} b^{\frac{\beta k}{\beta+1+k^{-1}}} \left(k^4\, k!\lceil \beta k\rceil !\right)^{\frac{1}{k(\beta+1+k^{-1})}} - \frac{1}{\alpha}. \tag{2.46}$$

and when $\beta k + 1 \leq \alpha k$,

$$n'_k \leq \alpha^{-\frac{\beta}{\beta+1}} b^{\frac{\beta k+1}{\beta+1}} \left(k^4\, k!\lceil \beta k\rceil !\right)^{\frac{1}{(\beta+1)k}} + k - 1, \tag{2.47}$$

and when $\beta k + 1 > \alpha k$,

$$n'_k \leq \alpha^{-\frac{\beta}{\beta+1}} b^{\frac{\beta k+1}{\beta+1}} \left(k^4\, k!\lceil \beta k\rceil !\right)^{\frac{1}{(\beta+1)k}} + \frac{\beta}{\alpha}k - \alpha^{-1} - 1. \tag{2.48}$$

Obviously, when k is sufficiently large, $n'_k < n_{k+1}$, since

$$n'_k = O(b^{\frac{\beta k}{\beta+1}} k^{\frac{1}{\beta+1}}) \text{ and } n_{k+1} = O(b^{\frac{\beta(k+1)}{\beta+1}} (k+1)^{\frac{1}{\beta+1}}).$$

Proof of 2: From inequalities (2.43) - (2.45), one can verify that

$$\frac{n_{k+1}}{n_k} = b^{\frac{\beta}{\beta+1}} + o(1),$$

and

$$\frac{n_{k+2}}{n_{k+1}} = b^{\frac{\beta}{\beta+1}} + o(1) + o(1).$$

Therefore,

$$\frac{n_{k+2} - n_{k+1}}{n_{k+1} - n_k} = \frac{\frac{n_{k+2}}{n_{k+1}} - 1}{1 - \frac{n_k}{n_{k+1}}} \to b^{\frac{\beta}{\beta+1}}. \tag{2.49}$$

Lemma 2.11.4. *Following the definition of $U_k(m, n, \beta)$, we immediately have*

$$EU_k = \binom{m}{\lceil \beta k\rceil} \binom{n}{k} p^{\lceil \beta k\rceil k}$$

and

$$EU_k^2 = \sum_{l=0}^{k} \binom{n}{k}\binom{k}{l}\binom{n-k}{k-l} \sum_{r=0}^{\lceil \beta k \rceil} \binom{m}{\lceil \beta k \rceil}\binom{\lceil \beta k \rceil}{r}\binom{m-\lceil \beta k \rceil}{\lceil \beta k \rceil - r} \cdot p^{2\lceil \beta k \rceil k - lr}.$$

By Lemma 2.11.4, one has

$$g(U_k) := \frac{Var\, U_k}{(EU_k)^2} = \sum_{l=0}^{k} \sum_{r=0}^{\lceil \beta k \rceil} \frac{\binom{k}{l}\binom{n-k}{k-l}}{\binom{n}{k}} \frac{\binom{\lceil \beta k \rceil}{r}\binom{m-\lceil \beta k \rceil}{\lceil \beta k \rceil - r}}{\binom{m}{\lceil \beta k \rceil}} b^{lr} - 1,$$

where $b = p^{-1}$.

Lemma 2.11.5. *Fix $\alpha > 0$ and $\beta \geq 1$. When k is sufficiently large, for every $n'_k(\alpha, \beta) < n < n_{k+1}(\alpha, \beta)$ and $m = \lceil \alpha n \rceil$, one can show that*

$$g(U_k(m, n, \beta)) = O(k^{-2}). \tag{2.50}$$

Proof of Lemma 2.11.5: By reasons similar to those in the proof of Lemma 2.8.3 in Section 2.8, it suffices to upper bound the following quantity.

$$g^*(U_k) := \sum_{l=1}^{k} \sum_{r=1}^{\lceil \beta k \rceil} \frac{\binom{k}{l}\binom{n-k}{k-l}}{\binom{n}{k}} \frac{\binom{\lceil \beta k \rceil}{r}\binom{m-\lceil \beta k \rceil}{\lceil \beta k \rceil - r}}{\binom{m}{\lceil \beta k \rceil}} b^{lr}.$$

Let

$$h(l, r) := \frac{\binom{k}{l}\binom{n-k}{k-l}}{\binom{n}{k}} \frac{\binom{\lceil \beta k \rceil}{r}\binom{m-\lceil \beta k \rceil}{\lceil \beta k \rceil - r}}{\binom{m}{\lceil \beta k \rceil}} b^{lr}.$$

It is easy to verify that

$$h(1, 1) = \frac{k\lceil \beta k \rceil (n-k)(m - \lceil \beta k \rceil) b}{(k-1)(\lceil \beta k \rceil - 1)\binom{n}{k}\binom{m}{\lceil \beta k \rceil}} \leq k^{-4}$$

when k is sufficiently large and $n > n'_k(\alpha, \beta)$ (by inequalities (2.47) and (2.48)). Moreover, by definition of $n'_k(\alpha, \beta)$,

$$h(k, \lceil \beta k \rceil) = EU_k(m, n, \beta)^{-1} \leq k^{-4}$$

when $n > n'_k(\alpha, \beta)$.

In order to bound other terms in $g'(U_k)$, one can verify that for any fixed $l < \log_b \frac{m-2\beta k-2}{(\beta k+1)^2}$,

$$\frac{h(l, r+1)}{h(l, r)} = \frac{(\lceil \beta k \rceil - r)^2 b^l}{(r+1)\,(m - 2\lceil \beta k \rceil + r + 1)} \leq \frac{(\beta k+1)^2 b^l}{m - 2\beta k - 2} < 1,$$

and that for any fixed $l > \log_b(\beta k + 2)m$,

$$\frac{h(l, r+1)}{h(l, r)} = \frac{(\lceil \beta k \rceil - r)^2 b^l}{(r+1)\,(m - 2\lceil \beta k \rceil + r + 1)} \geq \frac{b^l}{(\beta k+2)m} > 1.$$

Thus,

$$h(l, r) \leq h(l, 1) + h(l, \lceil \beta k \rceil), \text{ when } l \in [\log_b(m - 2\beta k) - \log_b \beta k,\ \log_b m + \log_b(\beta k + 2)]^c.$$

Similarly,

$$h(l, r) \leq h(1, r) + h(k, r), \text{ when } r \in [\log_b(n - 2k) - \log_b k,\ \log_b n + \log_b(k + 2)]^c.$$

It remains to consider $h(l, r)$ when $l = \log_b m + o(\log_b m)$ and $r = \log_b n + o(\log_b n)$. A straightforward calculation yields that

$$\begin{aligned} \log_b h(r, l) &\leq 2l \log_b k + 2r \log_b \lceil \beta k \rceil - l \log_b(n - k) - r \log_b(m - \lceil \beta k \rceil) \\ &\quad + k \log_b k + \lceil \beta k \rceil \log_b \lceil \beta k \rceil + (1 + o(1)) \log_b m \log_b n \\ &\leq O(k) \log_b k - (1 + o(1)) \log_b m \log_b n \leq -4, \end{aligned}$$

when $l = \log_b m + o(\log_b m)$, $r = \log_b n + o(\log_b n)$, k is sufficiently large and $n'_k(m, \beta) < n < n_{k+1}(m, \beta)$ (by inequalities (2.43) - (2.48)).

Thus

$$\begin{aligned} g^*(U_k) &\leq \beta k^2 \cdot k^{-4} + \sum_{l=1}^{k} (h(l, 1) + h(l, \lceil \beta k \rceil)) + \sum_{r=1}^{\lceil \beta k \rceil} (h(1, r) + h(k, r)) \\ &\leq O(k^{-2}) + \sum_{l=1}^{k} (h(1, 1) + h(k, \lceil \beta k \rceil)) + \sum_{r=1}^{\lceil \beta k \rceil} (h(1, 1) + h(k, \lceil \beta k \rceil)) \end{aligned}$$

$$= \quad O(k^{-2}),$$

where the second inequality comes from the monotonicity of $h(\cdot, 1)$, $h(\cdot, \lceil \beta k \rceil)$, $h(1, \cdot)$ and $h(k, \cdot)$ and the last inequality comes from the upper bounds on $h(1, 1)$ and $h(k, \lceil \beta k \rceil)$.

Lemma 2.11.6. *Fix $\alpha > 0$, $\beta > 1$. For any sufficiently large k, when $n'_k(\alpha, \beta) \leq n \leq n_{k+1}(\alpha, \beta)$, $M(\mathbf{Z}, m, n, \beta) = k$ eventually almost surely.*

Proof of Lemma 2.11.6: By Lemma 2.11.4 and Markov's inequality , we have

$$P(M(\mathbf{Z}, m, n, \beta) > k) = P(U_{k+1}(m, n, \beta) > 0) \ \leq \ E(U_{k+1}(m, n, \beta)) \leq \ \frac{1}{k^4}, \tag{2.51}$$

when $n \leq n_{k+1}(\alpha, \beta)$.

By Lemma 2.11.4 and Chebyshev inequality, we have

$$P(M(\mathbf{Z}, m, n, \beta) < k) = P(U_k(m, n, \beta) = 0) \ \leq \ \frac{Var(U_k(m, n, \beta))}{E^2(U_k(m, n, \beta))} \tag{2.52}$$

When $n'_k(\alpha, \beta) \leq n \leq n_{k+1}(\alpha, \beta)$, it immediately follows from Lemma 2.11.5 that

$$P(M(\mathbf{Z}, m, n, \beta) < k) \ \leq \ O(k^{-2}). \tag{2.53}$$

Note that $M(\mathbf{Z}, m, n, \beta)$ is monotone increasing with n for any given $\alpha > 0$ and $m = \lceil \alpha n \rceil$. Therefore,

$$\begin{aligned} & \sum_k P(\exists n \text{ s.t. } n'_k \leq n \leq n_{k+1} \text{ and } M(\mathbf{Z}, \lceil \alpha n \rceil, , n, \beta) \neq k) \\ \leq \quad & \sum_k P(M(\mathbf{Z}, \lceil \alpha n'_k \rceil, n'_k, \beta) < k) + \sum_k P(M(\mathbf{Z}, \lceil \alpha n_{k+1} \rceil, n_{k+1}, \beta) > k) \\ < \quad & \infty. \end{aligned}$$

By Borel-Cantelli lemma, one can conclude that when k is sufficiently large, $M(\mathbf{Z}, \lceil \alpha n \rceil, n, \beta) = k$ with probability 1 if $n'_k \leq n \leq n_{k+1}$.

Proof of Theorem 2.3.2: Since the basic idea here is the same as that in the proof of Theorem 2.2.4, we will only address the difference.

For each $m \geq 1$, let $A_m = \cup_{n \geq m} B_n$ with $B_n = \{|M(\mathbf{Z}, \lceil \alpha n \rceil, n, \beta) - s(\lceil \alpha n \rceil, n, \beta)| \geq \frac{5}{2}\}$, and define index sets

$$\begin{aligned}
I_m^1 &= \{n \geq m \text{ such that } n'_k(\lceil \alpha n \rceil, n, \beta) \leq n \leq n_{k+1}(\lceil \alpha n \rceil, n, \beta) \text{ for some} \\
&\quad k \notin (s(\lceil \alpha n \rceil, n, \beta) - \frac{5}{2}, s(\lceil \alpha n \rceil, n, \beta) + \frac{5}{2})\}, \\
I_m^2 &= \{n \geq m \text{ such that } n'_k(\lceil \alpha n \rceil, n, \beta) \leq n \leq n_{k+1}(\lceil \alpha n \rceil, n, \beta) \text{ for some} \\
&\quad k \in (s(\lceil \alpha n \rceil, n, \beta) - \frac{5}{2}, s(\lceil \alpha n \rceil, n, \beta) + \frac{5}{2})\}, \\
I_m^3 &= \{n \geq m \text{ such that } n \text{ belongs to no interval} \\
&\quad [n'_k(\lceil \alpha n \rceil, n, \beta), n_{k+1}(\lceil \alpha n \rceil, n, \beta)] \text{ for } k \geq 1\}.
\end{aligned}$$

Again, we want to show $P(A_m^j) \to 0$ as m tends to infinity for $j = 1, 2, 3$.

First, we still want to show A_m^1 is empty. From inequalities (2.43) and (2.44), one can verify that if there exists a k such that $n'_k \leq n \leq n_{k+1}$, then $k \in (\log_b n + \frac{\beta}{\beta+1} \log_b \frac{m}{n}, \frac{\beta+1}{\beta} \log_b n + \log_b \frac{m}{n})$. By arguments to those in the proof of Theorem 2.2.4, it follows that

$$\begin{aligned}
(1 + o(1))\phi(k + 2, \alpha n, n, \beta) &\leq EU_{k+1}(n) \leq EU_{k+1}(n_{k+1}) \\
&\leq k^{-3-\epsilon} < 1 = (1 + o(1))\phi(s(\alpha n, n, \beta), \alpha n, n, \beta),
\end{aligned}$$

and

$$\begin{aligned}
(1 + o(1))\phi(k, \alpha n, n, \beta) &\geq EU_k(n) \geq EU_k(n'_k) \\
&\geq k^{3+\epsilon} > (1 + o(1)) = (1 + o(1))\phi(s(\alpha n, n, \beta), \alpha n, n, \beta).
\end{aligned}$$

By the monotonicity of $\phi(\cdot)$, it is clear that $k \leq s(\alpha n, n, \beta) \leq k + 2$, which implies A_m^1 is empty.

Since the arguments on $\lim_{m \to \infty} P(A_m^2) = 0$ are almost identical to those in the proof of Theorem 2.2.4, it will be omitted here.

It remains to show that $\lim_m P(A_m^3) = 0$. By a similar argument as that in the proof of Theorem 2.2.4, one can verify that for any sufficiently large n, if there does not exist any k s.t. $n'_k(\lceil \alpha n \rceil, n, \beta) \leq n \leq n_{k+1}(\lceil \alpha n \rceil, n, \beta)$, then there must exist a k such that

$n_k(\lceil \alpha n \rceil, n, \beta) < n < n'_k(\lceil \alpha n \rceil, n, \beta)$. Again, if we regard $\mathbf{Z}_{\lceil \alpha n \rceil, n}$ as the left upper corner of an infinite dimensional binary matrix, then $M(\mathbf{Z}, \lceil \alpha n \rceil, n, \beta)$ is monotone increasing when n is increasing. Therefore if $n_k(\lceil \alpha n \rceil, n, \beta) < n < n'_k(\lceil \alpha n \rceil, n, \beta)$ for any certain k, then $M(\mathbf{Z}, \lceil \alpha n_k \rceil, n_k, \beta) \leq M(\mathbf{Z}, \lceil \alpha n \rceil, n, \beta) \leq M(\mathbf{Z}, \lceil \alpha n'_k \rceil, n'_k, \beta)$. Moreover, from Lemma 2.11.6, we have $k - 1 \leq M(\mathbf{Z}, \alpha, n, \beta) \leq k$. By following the arguments in the proof of Theorem 2.2.4, it now suffices to show

$$s(\alpha n'_k, n'_k, \beta) - s(\alpha n_k, n_k, \beta) = o(1),$$

which is obvious from (2.6) and the bounds on $n_k(\lceil \alpha n \rceil, n, \beta)$ and $n'_k(\lceil \alpha n \rceil, n, \beta)$. Thus, $\lim_m P(A^3_m) = 0$.

CHAPTER 3

Noise Sensitivity of Frequent Itemset Mining and Recoverability of Error-Tolerant Frequent Itemset Mining in Binary Matrices with Noise

3.1 Noise Sensitivity Analysis

3.1.1 Noise

The data to which data mining methods are applied are typically obtained by high-throughput technologies. Data of this sort is subject to varying levels of error and noise effects. Systematic errors are often identified and removed in preprocessing before data mining. For example, in DNA microarray analysis, the biases such as those caused by the efficiency of dye incorporation, DNA concentration on arrays and batch variation are usually removed by normalization. Noise that remains after preprocessing is usually considered unavoidable and as randomness of the model. For example, in DNA Microarray analysis such as (66; 48), biologists carry out multiple replicate experiments on the aliquots that come from a same sample, and study the variations of the gene expression values among different aliquots. The result there shows that moderate noise exists for all genes in their experiment. Another example can be found in transaction data. Errors such as missing values, incorrect inputs caused by machine malfunctions are common and can be viewed as random errors.

Some data mining methods commonly used by computer scientists, such as standard frequent itemset mining algorithms, do not account for the effect of noise and errors in their search for distinguished submatrices. In the next section, we will show how the noise can severely affect the performance of standard frequent itemset mining.

3.1.2 Binary Statistical Additive Noise Model

In order to account for the potential effects of noise on data mining tasks such as frequent itemset mining, we study under a simple statistical model where the observed data is equal to the (modulo 2) sum of a "true" unobserved data matrix plus random noise. Formally,

$$\mathbf{Y} \;=\; \mathbf{X} \oplus \mathbf{Z}. \tag{3.1}$$

We define and interpret each matrix in turn. Each of the matrices $\mathbf{Y}$, $\mathbf{X}$ and $\mathbf{Z}$ has n columns and m rows. Each column corresponds to a sample or an experimental condition. Each row corresponds to a binary variable or a feature measured on each sample. The matrix $\mathbf{X} = \{x_{i,j}\}$ is a deterministic binary matrix that consists of the true data values in the absence of noise. $\mathbf{Z} \sim \text{Bern}(p)$ is a random matrix, whose entries $z_{i,j}$ are independent Bernoulli random variables (coin tosses) with $P(z_{i,j} = 1) = p = 1 - P(z_{i,j} = 0)$ for some $p \in (0, 1)$. The matrix $\mathbf{Y} = \{y_{i,j}\}$ represents the observed binary data.

The operation $\oplus$ is the standard exclusive-or: $0 \oplus 0 = 1 \oplus 1 = 0$ and $0 \oplus 1 = 1 \oplus 0 = 1$. The model (3.1) states that $y_{i,j} = x_{i,j} \oplus z_{i,j}$ for each $1 \le i \le m$ and $1 \le j \le n$. Thus the noise effect is equivalent to randomly flipping some of the values of $\mathbf{X}$ in $\mathbf{Y}$.

Here is a simple example illustrating the additive noise model described above.

$$\begin{pmatrix} \cdots & \cdots & \cdots \\ & 1 \;\; 1 & \\ \cdots & & \cdots \\ & 0 \;\; 1 & \\ \cdots & \cdots & \cdots \end{pmatrix} = \begin{pmatrix} \cdots & \cdots & \cdots \\ & 1 \;\; 1 & \\ \cdots & & \cdots \\ & 1 \;\; 1 & \\ \cdots & \cdots & \cdots \end{pmatrix} \oplus \begin{pmatrix} \cdots & \cdots & \cdots \\ & 0 \;\; 0 & \\ \cdots & & \cdots \\ & 1 \;\; 0 & \\ \cdots & \cdots & \cdots \end{pmatrix}$$

Observed Matrix $\mathbf{Y}$ Pattern $\mathbf{X}$ Noise $\mathbf{Z}$

The statistical model (3.1) is the binary version of the standard additive noise model in statistical inference. It is also equivalent to a standard communication model, where the values of $\mathbf{X}$ are observed after being passed through a binary symmetric channel. This model is motivated by statistical practice, and is intended to capture the effects of random errors on the search for structures in noisy environments.

Suppose that the pattern matrix $\mathbf{X}$ contains some sort of strong signal structure, for example a large submatrix of ones. If the noise is small (i.e. the error probability p is close to zero), we hope that this structure would be readily reflected in the observed matrix $\mathbf{Y}$ and could be approximately recovered by standard methods without too much additional effort. In the following section, we will study the recoverability of standard frequent itemset mining under this proposed binary additive noise model.

3.2 Noise Sensitivity of Frequent Itemset Mining

Frequent itemset mining is widely used in real world applications. However, it does not account the effects of noise (errors). The following discussion indicates that frequent itemset mining is very sensitive to noise. Indeed, this negative conclusion is already apparent from Theorem 2.2.4 and Proposition 2.2.3. Suppose as above that $\mathbf{Z} \sim \text{Bern}(p)$, and assume that $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ are of dimension $n \times n$. If each entry of $\mathbf{X}$ is zero, then $\mathbf{Y} = \mathbf{Z}$ and the largest $k \times k$ submatrix of ones in $\mathbf{Y}$ has k roughly equal to $2 \log_b n$, where $b = p^{-1}$. On the other hand, at the other extreme where each entry of $\mathbf{X}$ is equal to one, it is easy to see that the entries of $\mathbf{Y}$ are simply independent Bernoulli$(1-p)$ random variables. In this case the largest $k \times k$ submatrix of ones in $\mathbf{Y}$ has k roughly equal to $2 \log_{b'} n$, where $b' = (1-p)^{-1}$. Proposition 2.2.3 tells us that it is very unlikely to find a block with a larger size. In the extreme cases $\mathbf{X} = 0$ and $\mathbf{X} = 1$, the largest block structure in $\mathbf{Y}$ is of logarithmic size; the only change is in the base of the logarithm. The next result extends this conclusion to any underlying pattern matrix $\mathbf{X}$.

Proposition 3.2.1. *Fix any $\epsilon > 0$. Let $\mathbf{X}_n$ be a non-random $n \times n$ square binary matrix and let $\mathbf{Y}_n = \mathbf{X}_n \oplus \mathbf{Z}_n$, where $\mathbf{Z}_n \sim$ Bern(p). Eventually almost surely, $(2-\epsilon) \log_b n < M(\mathbf{Y}_n) \leq 2 \log_{b'} n$, where $b = p^{-1}$ and $b' = (1-p)^{-1}$.*

Proof of Proposition 3.2.1: Fix n and let $W_n = \{w_{i,j}\}$ be an $n \times n$ binary matrix with

independent entries, defined on the same probability space as $\{z_{i,j}\}$, such that

$$w_{i,j} = \begin{cases} \text{Bern}\left(\frac{1-2p}{1-p}\right) & \text{if } x_{ij} = y_{ij} = 0 \\ 1 & \text{if } x_{ij} = 0, y_{ij} = 1 \\ y_{i,j} & \text{if } x_{ij} = 1 \end{cases} \tag{3.2}$$

Note that the above definition is valid since we assume noise level $p < \frac{1}{2}$ here. Define $\tilde{\mathbf{Y}}_n = \mathbf{Y}_n \vee W_n$ to be the entry-wise maximum of $\mathbf{Y}_n$ and W_n. Clearly $M(\mathbf{Y}_n) \leq M(\tilde{\mathbf{Y}}_n)$, as any submatrix of ones in $\mathbf{Y}_n$ must also be present in $\tilde{\mathbf{Y}}_n$. Moreover, it is easy to check that $\tilde{y}_{i,j}$'s are i.i.d. with $P(\tilde{y}_{i,j} = 1) = 1 - p$ for every $1 \leq i, j \leq n$. Therefore, $\tilde{\mathbf{Y}}_n \sim \text{Bern}(1-p)$. Now it follows from Theorem 2.2.4 that $M(\mathbf{Y}_n) \leq 2 \log_{b'} n$ eventually almost surely.

To obtain the inequality of the other direction, let

$$\tilde{w}_{i,j} = \begin{cases} \text{Bern}\left(\frac{1-p}{p}\right) & \text{if } x_{ij} = y_{ij} = 1 \\ 0 & \text{if } x_{ij} = 1, y_{ij} = 0 \\ y_{i,j} & \text{if } x_{ij} = 0 \end{cases} \tag{3.3}$$

Define $\underline{\mathbf{Y}}_n = \mathbf{Y}_n \wedge \tilde{W}_n$ to be the entry-wise minimum of $\mathbf{Y}_n$ and $\tilde{W}_n$. By an argument similar to above for $\tilde{\mathbf{Y}}_n$, it follows that $M(\mathbf{Y}_n) \geq M(\underline{\mathbf{Y}}_n)$ and the entries in $\underline{\mathbf{Y}}_n$ are i.i.d. Bern(p). Thus, by Theorem 2.2.4, $M(\mathbf{Y}_n) \geq (2 - \epsilon) \log_b n$ eventually almost surely.

Proposition 3.2.1 can be interpreted as follows. No matter what type of block structures might exist in $\mathbf{X}$, in the presence of random noise these structures leave behind only logarithmic fragments in the observed data. In particular, under the additive noise model (3.1) block structures existing in the pattern matrix cannot be recovered, even approximately, by methods such as frequent itemset mining that look for maximal submatrices of 1's without errors.

3.3 Error-Tolerant Frequent Itemsets

As we argued in the previous sections, transaction related data is often contaminated with noise, and we also showed that the standard frequent itemset mining algorithm is

very sensitive to noise. In particular, the output submatrices of standard frequent itemset mining do not directly recover the block structures prior to the noise contamination. To overcome this potential drawback of standard frequent itemset mining, error-tolerant frequent itemset mining algorithms are proposed. To be specific, a modified frequent itemset mining algorithm is called error-tolerant if it allows some fraction of zeros existing in the resulting submatrices. There are a number of different error-tolerant frequent itemset mining algorithms (53; 52; 58; 42; 41). Most of them still require the average of the identified submatrices to be greater than a user specified threshold τ. We can use this common property to assess the statistical significance of the identified error-tolerant submatrices.

Definition: Given a binary matrix $\mathbf{U}$ with an index set C, let

$$F(\mathbf{U}) \;=\; \frac{\sum_{(i,j)\in C} u_{i,j}}{|\mathbf{U}|}$$

be the fraction of ones in $\mathbf{U}$, equivalently the average of the entries of $\mathbf{U}$.

Definition: Given $\tau > 0$, define $M_\tau(\mathbf{Z}_n)$ to be the largest k such that there exists at least one $k \times k$ submatrix $\mathbf{U}$ in $\mathbf{Z}_n$ satisfying $F(\mathbf{U}) > \tau$.

Under the same binary random matrix model in Chapter 2, and by applying a first moment argument analogous to that in Proposition 2.2.3 and a probability upper bound on the tails of the binomial distribution, one can easily establish the following proposition.

Proposition 3.3.1. *Fix* $0 < \gamma < 1$ *and suppose that* $0 < p < \tau < 1$. *When* n *is sufficiently large,* $P(M_\tau(\mathbf{Z}_n) \geq 2\log_{b^*} n + r) \leq 2n^{-2r}(\log_{b^*} n)^{3r}$ *for each* $1 \leq r \leq \gamma n$. *Here* $b^* = \exp\{3(\tau - p)^2/8p\}$.

Proof: For $l \geq 1$ let $V_l(n)$ be the number of $l \times l$ submatrices $\mathbf{U}$ of $\mathbf{Z}_n$ with $F(\mathbf{U}) \geq \tau$. Note that $E(V_l(n)) = \binom{n}{l}^2 P\left(F(\mathbf{Z}_l) \geq \tau\right)$. The random variable $l^2 \cdot F(\mathbf{Z}_l)$ has a Binomial(l^2, p) distribution. Using a standard inequality for the tails of the binomial distribution, (*c.f.* Problem 8.3 of (17)), we find that $P\left(F(\mathbf{Z}_l) \geq \tau\right) \leq q^{l^2}$ where $q = 1/b^*$. It then follows from Stirling's approximation that $EV_l(n) \leq 2$ when $l = l(n) = 2\log_{b^*} n$. For $l = r + l(n)$, $P(M_\tau(\mathbf{Z}_n) \geq l) \leq E(V_l(n))$ and the stated inequality then follows from arguments analogous

to those in the proof of Proposition 2.2.3.

Remark: It can be seen from the above proof that the base $b^* = \exp\{3(\tau - p)^2/8p\}$ of the logarithm is derived from an upper bound on the tails of the binomial distribution (See Problem 8.2 in (17)). This upper bound may not always be the sharpest one. For example, when $\tau \to 1$, $b^* = \exp\{3(\tau - p)^2/8p\}$ fails to converge to p^{-1}. Thus, the probability bound provided in Proposition 2.3.1 does not agree with that of Proposition 2.2.3. In this case, we need a sharper upper bound on the tails of the binomial distribution to get a base of logarithm larger than the above b^*. In fact, the probability bound in (33) can provide such an alternative base of the logarithm, namely $b^* = \left(\left(\frac{\tau}{p}\right)^{\tau} \left(\frac{1-\tau}{1-p}\right)^{1-\tau} \right)$, which tends to p^{-1} as $\tau \to 1$. When τ is not close to 1 and $p \geq \frac{1}{2}$, the upper bound on tails of the binomial distribution in (49) provides $b^* = \exp\{(\tau - p)^2/2p(1-p)\}$, which may be better than the b^* derived from the two probability upper bound mentioned above. In general, in order to get the best base of logarithm, one need calculate all three b^* described above or even more, and choose the largest one.

3.4 Non-Square Matrices

In this section, we extend the significance analysis results of square submatrices with a large fraction of ones in square matrices to the case of non-square submatrices in non-square matrices. We use the same notation and setting for row/column aspect ratios as those in Section 2.3, except that we consider the following quantity instead.

Definition: Fix $\alpha > 0, \tau > 0$ and $\beta \geq 1$. Given an $m \times n$ binary matrix $\mathbf{Z}_{mn}$ with $m = \lceil \alpha n \rceil$, let $M_\tau(\mathbf{Z}, m, n, \beta)$ be the largest k such that there exists a $\lceil \beta k \rceil \times k$ submatrix $\mathbf{U}$ with $F(\mathbf{U}) > \tau$ in $\mathbf{Z}_{mn}$.

By adopting similar steps as those in the proof of Proposition 2.3.1, one can easily establish the following result. Since the proof is trivial, it is omitted.

Proposition 3.4.1. *Fix $0 < \gamma < 1$ and $\alpha > 0$. For each $1 \leq r \leq \gamma\, n$, when n is sufficiently large,*

$$P\left(M_\tau(\mathbf{Z}, \lceil \alpha n \rceil, n, \beta) \;\geq\; k(n, \alpha, \beta, \tau) + r\right) \;\leq\; n^{-(\beta+1)\, r}\, 2(\log_{b^*} n)^{(\beta+2)r}, \qquad (3.4)$$

where $k(n, \alpha, \beta, \tau) = \frac{\beta+1}{\beta} \log_{b^*} n + \log_{b^*} \alpha$, *and* $b^* = \exp\{3(\tau - p)^2/8p\}$.

Remark: See the discussion following Proposition 3.3.1. The base b^* in the above proposition can also be replaced by values derived from other probability upper bounds on the tail of binomial distribution.

For discovered submatrices having a large fraction of ones, we can use Proposition 3.4.1 to evaluate their statistical significance. In the following example, we demonstrate how to do the calculation explicitly.

Example. Proposition 3.4.1 can be applied to find an approximate significance values for submatrices having a larger fraction of ones than the background level. Suppose that an error-tolerant frequent itemset mining algorithm is applied to a $4,000 \times 100$ binary matrix $\mathbf{Y}$, 65% of whose entries are equal to 1. This error tolerant frequent itemset mining algorithm finds a 73×25 submatrix $\mathbf{U}'$ in Y with 95% 1s. Since in this case $p > \frac{1}{2}$, the discussion immediately after Proposition 3.3.1 suggests using $b^* = \exp\{(0.95-p)^2/2p(1-p)\} = 1.2187$. By plugging each corresponding term into (3.4), one obtains a significance value $p(\mathbf{U}') \leq 0.04802$.

3.5 Simple Recovery Problem

In the previous sections, we showed that frequent itemset mining can not directly recover underlying block structures if noise is present. This motivates us to consider whether the algorithms other than those directly searching for submatrices of 1's can recover underlying block structures. We show below that block structures can, in principle, be recovered by some algorithms that search for submatrices having a large fraction of ones.

We referred several error-tolerant frequent itemset mining criteria in the previous sections. In this section, we will study the recoverability of a particular error-tolerant frequent itemset mining, approximate frequent itemset mining proposed in (42). The following definition of error-tolerant itemsets is introduced in (42). An algorithm for finding such itemsets is given in (41).

Definition: Given any binary matrix $\mathbf{Y}$, a $k \times l$ submatrix $\mathbf{U}$ of $\mathbf{Y}$ is a τ-*approximate*

frequent itemset ($\tau-$AFI) if each of its rows satisfies $F(u_{i*}) \geq \tau$ and each of its columns satisfies $F(u_{*j}) \geq \tau$. Let $\mathbf{AFI}_\tau(\mathbf{Y})$ be the collection of all τ-AFIs in $\mathbf{Y}$.

The recovery problem we are going to study in this dissertation is a simple recovery problem as follows.

A simple recovery problem: Let $\mathbf{X}$ be an $n \times n$ binary matrix that consists of an $l \times l$ submatrix of 1's, with an index set C^*, and all other entries equal to 0. (The rows and columns of C^* need not be contiguous.) Given an observation $\mathbf{Y} = \mathbf{X} \oplus \mathbf{Z}$ with $\mathbf{Z} \sim \text{Bern}(p)$ and $0 < p < 1/2$, we wish to recover the submatrix C^*.

To recover C^*, let p_0 be any number such that $p < p_0 < 1/2$, and let $\tau = 1 - p_0$ be the associated error threshold. We estimate C^* by the index set of the largest square $\tau-$AFI in the observed matrix $\mathbf{Y}$. More precisely, let $\hat{\mathcal{C}}$ be the family of index sets of square submatrices $C \in \mathbf{AFI}_\tau(\mathbf{Y})$, and define

$$\hat{C} \;=\; \text{argmax}_{C \in \hat{\mathcal{C}}} |C|$$

to be any maximal sized submatrix in $\hat{\mathcal{C}}$. Note that $\hat{\mathcal{C}}$ and $\hat{C}$ depend only on the observed matrix $\mathbf{Y}$. Let the ratio

$$\Lambda(\hat{C}) \;=\; |\hat{C} \cap C^*| / |\hat{C} \cup C^*|$$

measure the overlap between the estimated index set $\hat{C}$ and the true index set C^*. Thus $0 \leq \Lambda \leq 1$, and values of Λ close to one indicate better overlaps.

The following theorem is about the recoverability of AFI estimator $\hat{C}$.

Theorem 3.5.1. *When n is sufficiently large, for any $0 < \alpha < 1$ such that $8\alpha^{-1}(\log_b n + 2) \leq l$ we have*

$$P\left(\Lambda(\hat{C}) \leq \frac{1-\alpha}{1+\alpha}\right) \;\leq\; \Delta_1(l) + \Delta_2(\alpha, l). \tag{3.5}$$

Here $\Delta_1(l) = 2e^{-\frac{l(p-p_0)^2}{3p}}$, $\Delta_2(\alpha, l) = 2n^{-\frac{1}{4}\alpha l + 2\log_b n}$, and $b = \exp\{3(1-2\,p_0)^2/8p\}$.

Remarks: (i) Note that among the two terms on RHS of the above inequality, the second term is less than $2n^{-4/\alpha}$ and it is the dominant term in the probability upper bound when $l >> \ln n$ and $l >> \frac{(p-p_0)^2}{p}$.

(ii) Note that b in the above proposition is larger than $\tilde{b} = \exp\{3(1-2\,p_0)^2/8p_0\}$ and that when l is sufficiently large, crudely, $\Delta_1(l) \leq \tilde{\Delta}_1(l) := e^{-\sqrt{l}}$. Thus, by replacing b with $\tilde{b}$ and $\Delta_1(l)$ with $\tilde{\Delta}_1(l)$ in (3.5), one can obtain a probability bound which does not depend on the unknown parameter p.

Example: The following is an example illustrating Theorem 3.5.1. Let X be an $n \times n$ binary matrix with n = 800 and let C^* be an $l \times l$ submatrix of X with $l = 300$. Suppose the noise level $p = 0.1$ and suppose the user specified noise level $p_0 = 0.17$. When $\alpha = 1/4$, since $l > 8\alpha^{-1}(\log_b n + 2) = 156.8989$, it follows from Theorem 1 that $P(\Lambda(\hat{C}) \leq \frac{3}{5}) \leq 2(e^{-4.9}+800^{-12.944}) = 0.015$, i.e. the probability that the overlap between the AFI estimator and C* is less than 0.6 is small (less than 2%).

Theorem 3.5.1 can readily be applied to the asymptotic recovery of structure in a sequential framework. Suppose that $\{\mathbf{X}_n : n \geq 1\}$ is a sequence of square binary matrices, where $\mathbf{X}_n$ is $n \times n$ and consists of an $l_n \times l_n$ submatrix C_n^* of 1s with all other entries equal to 0. For each n we observe $\mathbf{Y}_n = \mathbf{X}_n \oplus \mathbf{Z}_n$, where $\mathbf{Z}_n \sim$Bern(p). Let $\Lambda(\hat{C}_n)$ measure the overlap between C_n^* and the estimate $\hat{C}_n$ produced by the AFI recovery method above. The result below follows from Theorem 3.5.1 and the Borel-Cantelli lemma.

Corollary 1. *If $l_n \geq 8\psi(n)(\log_b n + 2)$ where $\psi(n) \to \infty$ as $n \to \infty$, then eventually almost surely*

$$\Lambda(\hat{C}_n) \;\geq\; \frac{1-\psi(n)^{-1}}{1+\psi(n)^{-1}} \to 1.$$

Proof of Corollary 1: Theorem 3.5.1 implies that if we can bound both $\Delta_1(l_n)$ and $\Delta_2(\psi(n)^{-1}, l_n)$ by $2n^{-2}$ for any sufficiently large n, then Borel -Cantelli Lemma can be applied to establish the almost sure convergency.

When n is sufficiently large, the condition $l_n > 8\psi(n)(\log_b n - \log_b \log_b n + 2)$ and $\psi(n) \to \infty$, implies $l_n > \frac{6p}{(p-p_0)^2} \ln n$. By plugging this lower bound on l_n into $\Delta_1(l_n)$, one can get $\Delta_1(l_n) < 2n^{-2}$. Meanwhile, by plugging the condition that $l_n > 8\psi(n)(\log_b n - \log_b \log_b n + 2)$ into $\Delta_2(\psi(n)^{-1}, l_n)$, one can get $\Delta_2(\psi^{-1}(n), l_n) < 2n^{-2}$.

3.6 Discussion

Reuning-Scherer studies several recovery problems in (55). In the case considered above, he calculates the fraction of ones in every row and every column of $\mathbf{Y}$, and then selects those rows and columns with a large fraction of ones. His algorithm is consistent when $l \geq n^{\alpha}$ for $\alpha > 1/2$. However, it is easy to show that individual row and column sums alone are not sufficient to recover C^* in the above recovery problem when $l \leq n^{\alpha}$ for $\alpha < 1/2$. To be more concrete, suppose C^* has rows $c_1^*, ..., c_l^*$. By central limit theorem, one can show that for any row $c_i \in \{c_1^*, ..., c_l^*\}^c$ and $c_j^* \in \{c_1^*, ..., c_l^*\}$, $P\left(F(c_{i*}) > F(c_{j*}^*)\right) \geq \gamma > 0$ if $l \leq n^{\alpha}$ for $\alpha < 1/2$. In this case, one gains considerable power by directly considering submatrices, and as the result above demonstrates, one can consistently recover C_n^* if $l_n / \ln n \to \infty$. By Theorem 2.2.4, this ratio requirement on l_n is almost weakest if one wants to recover C^*.

3.7 Proof of Theorem 3.5.1

The following lemmas are used in the proof of Theorem 3.5.1. Among these, Lemma 3.7.1 implies that $|\hat{C}|$ is greater than or equal to $|C^*|$ with high probability, and Lemma 3.7.4 shows that $\hat{C}$ can only contain a small proportion of entries from outside C^*. Lemma 3.7.2 and Lemma 3.7.3 are used in the proof of Lemma 3.7.4.

Lemma 3.7.1. *Under the conditions of Theorem 3.5.1,* $P\left(|\hat{C}| < l^2\right) \leq \Delta_1(l)$.

Proof of Lemma 3.7.1: Let $u_{1*}, ..., u_{l*}$ be corresponding rows of C^* in $\mathbf{Y}$ and let V be the number of rows satisfying $F(u_{i*}) < 1 - p_0$, where $F(\cdot)$ is the function measuring the fraction of ones. By Markov's inequality,

$$P(V \geq 1) \ \leq \ E(V) \ = \ \sum_{i=1}^{l} P(F(u_{i*}) < 1 - p_0). \tag{3.6}$$

Using standard bounds on the tails of the binomial distribution, when l_n is sufficiently large,

$$P(V \geq 1) \leq l \cdot e^{-\frac{3l(p-p_0)^2}{8p}} \leq e^{-\frac{1}{3p}l(p-p_0)^2},$$

when l is sufficiently large.

Let $u_{*1}, ..., u_{*l}$ be corresponding columns of C^* in $\mathbf{Y}$ and let V' be the number of columns satisfying $F(u_{*i}) < 1 - p_0$. A similar calculation as above shows that

$$\begin{aligned} P(V' \geq 1) \leq E(V') &\leq l \cdot e^{-3\frac{l(p-p_0)^2}{8p}} \\ &\leq e^{-\frac{1}{3p}l(p-p_0)^2}. \end{aligned}$$

Since $\{|\hat{C}| < l^2 = |C^*|\} \subset \{C^* \notin AFI_\tau(\mathbf{Y})\} \subset \{V \geq 1\} \cup \{V' \geq 1\}$,

$$\begin{aligned} P\{|\hat{C}| < l^2\} &\leq P(V \geq 1) + P(V' \geq 1) \\ &\leq 2e^{-\frac{1}{3p}l_n(p-p_0)^2} = \Delta_1(l). \end{aligned}$$

Lemma 3.7.2. *Given $0 < \tau_0 < 1$, if there exists a $k \times r$ binary matrix M satisfying $F(M) \geq \tau_0$, then for $v = \min\{k, r\}$, there exists a $v \times v$ submatrix D of M such that $F(D) \geq \tau_0$.*

Proof of Lemma 3.7.2: Without loss of generality, we assume $v = k \leq r$. Then we rank each column according to its fraction of ones, and reorder the columns in descending order. Let the reordered matrix be M^1. Let $D = M^1[(1, ..., v) \times (1, ..., v)]$. One can verify that $F(D) \geq \tau_0$.

Lemma 3.7.3. *Let $1 < \gamma < 2$ be a constant, and let W be a $n \times n$ binary matrix. Let R_1 and R_2 be two square submatrices of W satisfying (i) $|R_2| = k^2$ with $k < n$, (ii) $|R_1 \backslash R_2| > k^\gamma$ and (iii) $R_1 \in \mathbf{AFI}_\tau(W)$. Then there exists a square submatrix $D \subset R_1 \backslash R_2$ such that $|D| \geq k^{2\gamma-2}/16$ and $F(D) \geq \tau$.*

Proof of Lemma 3.7.3: For any $R_1 \backslash R_2$, after suitable row and column permutations, it can be verified that $R_1 \backslash R_2$ can be expressed as a single maximal rectangular submatrix W_1 or can be expressed as the union of two overlapping maximal rectangular $W_1 \cup W_2$. Here we say W_i is a maximal rectangular submatrix of $R_1 \backslash R_2$, if there does not exist any other rectangular submatrix of $R_1 \backslash R_2$ that contains W_i.

Case 1: Suppose $R_1 \backslash R_2 = W_1$. Let l_1 and l_2 be the side length of W_1. Notice that $R_1 \backslash R_2 = W_1$ and $|R_2| = k^2$ imply the side length of square submatrix R_1 must be less than

k, which yields $\max(l_1, l_2) \leq k$. Since $|R_1 \backslash R_2| \geq k^\gamma$, it follows that $\min(l_1, l_2) \geq k^{\gamma-1}$. By the condition $AFI(R_1) \geq p$ and $AFI(R_2) \geq p$, it is trivial to conclude that $F(W_1) \geq p$. Then by Lemma 3.7.2, there exists a $v \times v$ submatrix D of W_1 such that $F(D) \geq p$ and $v \geq \min(l_1, l_2) \geq k^{\gamma-1} > k^{\gamma-1}/4$.

Case 2: Suppose $R_1 \backslash R_2 = W_1 \cup W_2$. It follows immediately that $\max(|W_1|, |W_2|) \geq \frac{|R_1 \backslash R_2|}{2}$. Without loss of generality, we assume $|W_1| \geq |W_2|$. By the definition of AFI and the condition that $AFI(R_1) \geq p$, it also follows that $F(W_1) \geq p$. Therefore, if we can show the length of the shorter side of W_1 is greater than $k^{\gamma-1}/4$, then there must exist a square submatrix $V \subset W_1$ such that $|V| \geq k^{2\gamma-2}/16$ and $F(V) \geq p$ by Lemma 3.7.2.

Let the side length of W_1 be l_1 and l_2. To show $\min(l_1, l_2) \geq k^{\gamma-1}/4$, we will instead show that $\min(l_1, l_2) < k^{\gamma-1}/4$ will lead to a contradiction.

Notice that when $\min(l_1, l_2) < k^{\gamma-1}/4$, it follows that $\max(l_1, l_2) > \frac{|R_1 \backslash R_2|}{2k^{\gamma-1}/4}$. Since square submatrix $|R_1|$ satisfies $|R_1| = \max(l_1, l_2)^2$, immediately we have $|R_1| > \frac{|R_1 \backslash R_2|^2}{k^{2\gamma-2}/4}$. Therefore

$$\begin{aligned} |R_1 \backslash R_2| &\geq |R_1| - |R_2| \\ &> \frac{|R_1 \backslash R_2|^2}{k^{2\gamma-2}/4} - k^2. \end{aligned} \tag{3.7}$$

Dividing both sides of inequality (3.7) by $|R_1 \backslash R_2|$, it yields

$$1 > \frac{|R_1 \backslash R_2|}{k^{2\gamma-2}/4} - \frac{k^2}{|R_1 \backslash R_2|}. \tag{3.8}$$

Since by the condition, $|R_1 \backslash R_2| \geq k^\gamma$, the right side of inequality (3.8) is greater than $4k^{(2-\gamma)} - k^{(2-\gamma)} > 1$ when $k > 0$. This leads to a contradiction with inequality (3.8). Therefore, $\min(l_1, l_2) \geq k^{\gamma-1}/4$.

Lemma 3.7.4. *Let $\mathcal{A}$ be the collection of $C \in \hat{\mathcal{C}}$ such that $|C| > \frac{l^2}{2}$ and $\frac{|C \cap C^{*c}|}{|C|} \geq \alpha$. Let A be the event that $\mathcal{A} \neq \emptyset$. If n is sufficiently large, then $l \geq 8\alpha^{-1}(\log_b n + 2)$ implies that*

$$P(A) \leq \Delta_2(\alpha, l).$$

Proof of Lemma 3.7.4: If $C \in \mathcal{A}$ then

(i) $|C^*| = l^2$,

(ii) $|C \backslash C^*| \ = \ |C| \ \cdot \ \frac{|C \cap C^{*c}|}{|C|} \ \geq \ \frac{l^2 \ \cdot \ \alpha}{2} \ = \ l^{\gamma}$, where $\gamma \ = \ 2 + \log_l \frac{\alpha}{2}$,

(iii) $C \in \mathbf{AFI}_{1-p_0}(\mathbf{Y})$.

Thus, by Lemma 3.7.3, there exists a $v \times v$ submatrix D of $C \backslash C^*$ such that $F(D) \ \geq \ 1 - p_0$ and $v \ \geq \ \frac{\alpha l}{4}$, which implies that

$$\max_{c \in \hat{\mathcal{C}}} M_{\tau}(C \cap C^{*c}) \ \geq \ v \ \geq \ \frac{\alpha l}{4},$$

where $\tau = 1 - p_0$.

Let $\mathbf{W}(\mathbf{Y}, C^*)$ be an $n \times n$ binary random matrix, where $w_{ij} = y_{ij}$ if $(i,j) \notin C^*$, and $w_{ij} \sim Bern(p)$ otherwise. It is clear that

$$M_{\tau}(\mathbf{W}) \ \geq \ \max_{c \in \hat{\mathcal{C}}} M_{\tau}(C \cap C^{*c}) \ \geq \ \frac{\alpha l}{4}.$$

By Proposition 3.3.1, when n is sufficiently large and $l \geq 8\alpha^{-1}(\log_b n + 2)$, we can bound $P(A)$ with

$$\begin{aligned} P(A) \quad &\leq \quad P(\max_{c \in \hat{\mathcal{C}}} M_{\tau}(C \cap C^{*c}) \ \geq \ \frac{\alpha l}{4}) \\ &\leq \quad P(M_{\tau}(\mathbf{W}) \geq \frac{\alpha l}{4}) \ \leq \ 2n^{-(\alpha l/4 - 2\log_{b'} n)}, \end{aligned} \tag{3.9}$$

where $b' = e^{\frac{3(1-p_0-p)^2}{8p}}$. As $p_0 > p$, it is trivial to verify that $b < b'$. Consequently, one can bound the RHS of inequality (3.9) by $\Delta_2(\alpha, l)$.

Proof of Theorem 3.5.1: Let E be the event that $\{\Lambda(\hat{C}) \leq \frac{1-\alpha}{1+\alpha}\}$. It is clear that E can be expressed as the union of two disjoint events E_1 and E_2, where

$$E_1 = \{|\hat{C}| < |C^*|\} \cap E$$

and

$$E_2 = \{|\hat{C}| \geq |C^*|\} \cap E$$

One can bound $P(E_1)$ by $\Delta_1(l)$ via Lemma 3.7.1.

It remains to bound $P(E_2)$. By the definition of $\Lambda(\cdot)$, the inequality $\Lambda(\hat{C}) \leq \frac{1-\alpha}{1+\alpha}$ can be rewritten equivalently as

$$1 + \frac{|\hat{C} \cap C^{*c}|}{|\hat{C} \cap C^*|} + \frac{|\hat{C}^c \cap C^*|}{|\hat{C} \cap C^*|} \geq \frac{1+\alpha}{1-\alpha}.$$

When $|\hat{C}| \geq |C^*|$, one can verify that $|\hat{C} \cap C^{*c}| \geq |\hat{C}^c \cap C^*|$, which implies that

$$1 + \frac{|\hat{C} \cap C^{*c}|}{|\hat{C} \cap C^*|} + \frac{|\hat{C}^c \cap C^*|}{|\hat{C} \cap C^*|} \leq 1 + 2\frac{|\hat{C} \cap C^{*c}|}{|\hat{C} \cap C^*|}.$$

Therefore, $E_2 \subset E_2^*$, where

$$\begin{aligned} E_2^* &= \{|\hat{C}| \geq |C^*|\} \cap \left\{1 + 2\frac{|\hat{C} \cap C^{*c}|}{|\hat{C} \cap C^*|} \geq \frac{1+\alpha}{1-\alpha}\right\} \\ &\subset \{|\hat{C}| > \frac{l^2}{2}\} \cap \left\{1 + 2\frac{|\hat{C} \cap C^{*c}|}{|\hat{C} \cap C^*|} \geq \frac{1+\alpha}{1-\alpha}\right\}. \end{aligned}$$

Notice that $1 + 2\frac{|\hat{C} \cap C^{*c}|}{|\hat{C} \cap C^*|} \geq \frac{1+\alpha}{1-\alpha}$ implies $\frac{|\hat{C} \cap C^{*c}|}{|\hat{C}|} \geq \alpha$. Therefore, by Lemma 3.7.4, $P(E_2^*) \leq \Delta_2(\alpha, l)$.

CHAPTER 4

Significance Analysis of Biclusters in a Real-Valued Matrix

In the previous chapters, we gave a detailed analysis of the statistical significance of submatrices of 1's and submatrices with a large fraction of ones in binary matrices. Many of these results can be easily extended into the case of categorical data. However, if the data is continuous, such extensions are less obvious, especially when we do not want to discretize the data. In this part of the dissertaion, we consider the problem of assessing the significance of submatrices with high average or low ANOVA residuals in real-valued matrices. By assuming the entries of the matrices follow i.i.d. Gaussian or other appropriate distributions, we obtains several probability bounds for the limiting distribution on the size of the largest submatrices with high average or low ANOVA residuals. These bounds are similar in form to those in Chapter 2.

Algorithms that search for submatrices with high average or low ANOVA residuals belong to the category of biclustering or subspace clustering in the data mining literature. In general, given an $m \times n$ data matrix, where entries are real values, a bicluster corresponds to a submatrix satisfying certain criterion. Some popular biclustering criteria are summarized by Madeira and Oliveira in (44) into four types: biclusters with constant entries, biclusters with constant rows/columns, biclusters with coherent values, biclusters with coherent evolutions. Applications of different biclustering algorithms in gene expression data analysis are also discussed in (44). In this dissertaion, we will only focus on biclustering algorithms based on average criterion and ANOVA criterion.

4.1 Average Criterion

One motivation to consider the average criterion comes from the gene expression analysis. There, a heat map is used to represent a data matrix by assigning a color to each entry in the data matrix according to its value. In DNA microarray analysis, biologists often reorder the locations of rows and columns in the heat map so that the resulting map displays large blocks of high red or high green areas. Such a visual display often reveals interesting information. For example, Perou *et al.* study the breast cancer subtypes in (54). They classify the patients into subtypes according to different expression levels based on several important genes, where the selection of important genes is motivated by the block patterns in the reordered heat map. Recalling the connection between heat maps and data matrices, we see that these red or green blocks correspond to submatrices with high absolute averages. For a simple illustration, we only consider submatrices with high positive averages in this dissertation. Formally speaking, the average criterion can be described as follows.

Average Criterion: Recall that for any given submatrix $\mathbf{U}$, $F(\mathbf{U}) = |\mathbf{U}|^{-1} \sum_{ij \in \mathbf{U}} u_{ij}$. Given a threshold $\tau > 0$, $\mathbf{U}$ is said having a high average if $F(\mathbf{U})$ is greater than τ.

Note that the criterion itself is the same as the average criterion in the case of binary matrices. For simplicity, for any given matrix $\mathbf{X}$ we will use a similar notation, $K_\tau(\mathbf{X})$, to denote the size of largest square submatrix $\mathbf{U} \subset \mathbf{X}$ with $F(\mathbf{U}) > \tau$.

In Section 2.1, we introduced a correspondence between binary data matrices and bipartite graphs. In fact, for real-valued data matrices, such a correspondence also exists, except that the regular bipartite graphs are replaced by edge-weighted bipartite graphs and a submatrix with high average corresponds to a high edge-weighted subgraph. The problem of finding the largest high average submatrices is still NP-complete, since its equivalent problem, finding the maximum edge-weighted subgraphs in bipartite graphs, is NP-complete. A slight variation to the problem of finding the largest high average submatrix is the problem of finding the CUT NORM of a matrix. Given an $m \times n$ data matrix $\mathbf{X}$, the CUT NORM $\|\mathbf{X}\|_c$ of $\mathbf{X}$ is $\max |\sum_{i \in A, j \in B} z_{ij}|$, where $A \subset \{1, ..., m\}$ and $B \subset \{1, ..., n\}$. Finding the CUT NORM of a data matrix is known to be NP-complete. Alon and Naor in (4) study

the CUT NORM problem for real-valued matrices. They propose a method that can find a ρ-approximation of the CUT NORM in polynomial time. Alon and Naor also show that with high probability, the absolute value of the sum of the entries in their output submatrix is greater than $0.56 \cdot \|\mathbf{X}\|_c$. Another variation related to the problem of finding the largest high average submatrices in a data matrix is studied by Dhillon *et al* in (18), where Dhillon *et al* propose a heuristic method using SVD to find a partition of the rows and columns of the data matrix. After reordering rows and columns according to this partition, the sum of the entries in the diagonal submatrices is largest. An extension of this algorithm with applications in DNA Microarray analysis can also be found in (36).

4.2 Significance Analysis under Average Criterion

We first propose a real-valued random matrix model for assessing significance of biclusters with high average. The most natural model is to assume that the entries in the data matrix follow i.i.d. standard Gaussian distribution. For simplicity, we first consider the case of square primary matrices and square submatrices.

Gaussian random matrix model: Let $\mathbf{W} = \{w_{i,j} : i, j \geq 1\}$ be an infinite array of independent N(0,1) random variables. For $n \geq 1$, let $\mathbf{W}_n = \{w_{i,j} : 1 \leq i, j \leq n\}$.

Thus $\mathbf{W}_n$ is an $n \times n$ random matrix with Gaussian entries comprising the "upper left corner" of the collection $\{w_{i,j}\}$.

To study the statistical significance of submatrices with high average in $\mathbf{W}_n$, we again use a first moment argument. Let $U_k(n, \tau)$ be the number of $k \times k$ submatrices in $\mathbf{W}_n$ with average greater than τ. Note that for any fixed $\tau > 0$, sufficiently large k and any $k \times k$ submatrix V, one has $P\left(F(V) \geq \tau\right) \leq e^{-\frac{\tau^2 k^2}{2}}$. Thus, $EU_k(n, \tau) \leq \binom{n}{k}^2 e^{-\frac{\tau^2 k^2}{2}}$. Using the Stirling approximation of $\binom{n}{k}$, one can then define

$$\tilde{\phi}(n, k) = (2\pi)^{-\frac{1}{2}} \, n^{n+\frac{1}{2}} \, k^{-k-\frac{1}{2}} \, (n-s)^{-(n-k)-\frac{1}{2}} \, e^{-\frac{\tau^2 k^2}{4}} \approx \left(EU_k(n, \tau)\right)^{1/2}. \qquad (4.1)$$

Clearly, $\tilde{\phi}(n, k)$ is an approximation of $EU_k(n, \tau)^{\frac{1}{2}}$. Let $\tilde{s}(n)$ be any real root of equation $\tilde{\phi}(n, s) = 1$, where $s \in \mathbb{R}^+$. By an argument similar to that in Section 2.2 and Lemma 2.2.1, it is not hard to verify that $\tilde{s}(n)$ always uniquely exists for any sufficiently large n.

Moreover, when n is sufficiently large, one can show that (c.f. Lemma 2.11.2)

$$\tilde{s}(n) \ = \frac{4}{\tau^2} \ln n \ - \ \frac{4}{\tau^2} \ln(\frac{4}{\tau^2} \ln n) + \tilde{C} + o(1), \tag{4.2}$$

where $\tilde{C} = \frac{4}{\tau^2} \ln e - \frac{4}{\tau^2} \ln 2$.

To assess the statistical significance of square submatrices with high averages under this Gaussian random matrix model, we establish the following proposition by bounding the probability $P\left(K_\tau(\mathbf{W}_n) > k\right)$ for $k > \tilde{s}(n) + 1$ from above.

Proposition 4.2.1. *Fix $0 < \gamma < 1$ and $\tau > 0$. When n is sufficiently large, for every integer $1 \le r \le \gamma n$ we have $P\left(K_\tau(\mathbf{W}_n) \ge \tilde{k}(n) + r\right) \le 4\tau^{-2} n^{-2r} (\frac{2 \ln n}{\tau^2})^{3r}$, where $\tilde{k}(n) = \lceil \tilde{s}(n) \rceil$.*

Proof of Proposition 4.2.1: The proposition above can be easily established by following a first moment argument similar to that in Proposition 2.2.3, in conjunction with the fact that $EU_k(n, \tau) \le \binom{n}{k}^2 e^{-\frac{\tau^2 k^2}{2}}$ when n and k are sufficiently large.

Remark: Hartigan (private communication) has pointed out that a similar result can also be obtained by applying the comparison principle for Gaussian sequences (cf. (60)). To be specific, for any individual submatrix V of size k, it is clear that $F(V)$ follows a Gaussian distribution. Thus, the set $\{F(V) : V = A \times B, A, B \subset \{1, ..., n\} \text{ and } |A| = |B| = k\}$ can be viewed as a positively correlated Gaussian sequence with $\binom{n}{k}^2$ different elements. Suppose $\{B_1^*, ..., B_m^*\}$, where $m = \binom{n}{k}^2$, is a sequence of i.i.d. Gaussian random variables with the same marginal means and variances as those of $\{F(V)\}$. According to the comparison principle for Gaussian sequences (60), for any given k,

$$P\left(K_\tau(\mathbf{W}_n) > k\right) = P\left(\max\{F(V)\} \ge \tau\right) \le P\left(\max\{B_1^*, ..., B_m^*\} \ge \tau\right).$$

Thus it suffices to bound $P\left(\max\{B_1^*, ..., B_m^*\} \ge \tau\right)$. Arratia *et al.* give a probability upper bound and lower bound on the extreme value of an independent Gaussian sequence via Poisson approximation in Section 4.4 of (7). Note that this upper bound is of the same magnitude as the probability upper bound in Proposition 4.2.1.

When the data matrix follows the i.i.d. Bernoulli random matrix model, we showed in

Theorem 2.2.4 that $M(\mathbf{Z}_n)$ takes at most three integer values around $s(n)$ eventually almost surely. When the matrix has i.i.d. Gaussian entries, one can establish a slightly weaker result for $K_\tau(\mathbf{W}_n)$, though more works on calculating and bounding the second moment are required.

Theorem 4.2.2. *Under the above model, eventually almost surely,*

$$\lfloor \tilde{s}(n) - \tilde{C} - 12\tau^{-2} \ln 2 \rfloor - 2 \leq K_\tau(\mathbf{W}_n) \leq \lceil \tilde{s}(n) \rceil + 1,$$

where $\tilde{C}$ is the same constant defined in (4.2).

Remarks: (i) Theorem 4.2.2 suggests that asymptotically, $K_\tau(\mathbf{W}_n)$ can only take values in a constant range around $\tilde{s}(n)$. This constant range is independent of n, but it varies for different threshold τ.

(ii) In the proof of Theorem 4.2.2, the assumption of normality is important. Without such an assumption, one may not be able to easily obtain a result as tight (up to a constant) as that above for $K_\tau(\mathbf{W}_n)$. In particular, we make critical use of the fact that under i.i.d. Gaussian assumption, for any individual $k \times k$ submatrix V of $\mathbf{W}_n$, there exist a simple upper bound and a simple lower bound on $P(F(V) \geq \tau)$ and that the ratio of this upper bound and this lower bound is less than $\tau^\gamma k^\gamma$ for some constant $\gamma \geq 0$. Note that, without this fact, it is still possible to show that there exists a positive constant $\rho < 1$ such that $K_\tau(\mathbf{W}_n) \geq \rho \cdot k_\tau(n)$ eventually almost surely. For example, given any threshold $\tau > 0$, suppose that $P(F(V) \geq \tau) \leq p_1$. A crude lower bound follows immediately from Theorem 2.2.4 in Chapter 2 that $K_\tau(\mathbf{W}_n) \geq 2 \log_{p_1^{-1}} n$ eventually almost surely.

4.3 ANOVA Criterion

The high average criterion is simple and intuitive. As we discussed in the previous sections, there are many other biclustering criteria. Among them, an important family of criteria try to identify coherent values in the submatrices. The ANOVA criterion is one of them.

The ANOVA criterion can be described as follows.

Definition: Suppose $\mathbf{X}$ is an $m \times n$ data matrix. A submatrix $V = A \times B$ of $\mathbf{X}$ satisfies

an ANOVA criterion with threshold $\tau > 0$, if there exist real constants a_i, b_j and c, where $i \in A$ and $j \in B$, s.t.

$$\frac{1}{|A||B|} \sum_{i \in A, j \in B} (w_{ij} - a_i - b_j - c)^2 \leq \tau. \tag{4.3}$$

Biclustering algorithms based on ANOVA type of criterion have been proposed and studied in (15) and in the PLAID model (40). Cheng and Church in (15) give a heuristic algorithm which finds submatrices with small sum of squares of ANOVA residuals. The PLAID model in (40) is based on a modified ANOVA type criterion, where the data matrix is modeled as a sum of layer submatrices and each layer submatrix has low ANOVA residuals. To implement their algorithm, Lazzeroni and Owen rewrote the original problem as an optimization problem, which can be relaxed and then solved iteratively. Some examples of applications of biclustering techniques based on ANOVA criteria can also be found in these two papers.

4.4 Significance Analysis under ANOVA Criterion

To assess the statistical significance of a submatrix with low ANOVA residuals, we make the following definitions.

Let

$$g(V) = k^{-2} \sum_{i \in A, j \in B} (w_{ij} - \overline{w}_{i.} - \overline{w}_{.j} + \overline{w}_{..})^2,$$

where $\overline{w}_{i.}$, $\overline{w}_{.j}$, and $\overline{w}_{..}$ correspond to the row, column, and the whole submatrix entry averages respectively. Note that by standard facts in ANOVA analysis, it is clear that

$$g(V) = \min_{a_i, b_j, c \in \mathbb{R}} k^{-2} \sum_{i \in A, j \in B} (w_{ij} - a_i - b_j - c)^2.$$

Definition: Given an $n \times n$ matrix $\mathbf{W}_n$ and $0 < \tau < 1$, let $L_\tau(\mathbf{W}_n)$ be the size of the largest square submatrix $V = A \times B$ in $\mathbf{W}_n$ such that $g(V) \leq \tau$.

By following similar steps as those in the proof of Proposition 4.2.1 and in conjunction with a simple probability upper bound on left tail of χ^2 distribution, one can establish the proposition below for $L_\tau^A(\mathbf{W}_n)$.

Proposition 4.4.1. *Fix $0 < \gamma, \tau < 1$ and fix $0 < \epsilon < 1$. When n is sufficiently large, for every integer $1 \leq r \leq \gamma\, n$ we have*

$$P\left(L_\tau^A(\mathbf{W}_n) \geq k_A(n) + r\right) \leq 2n^{-2\,r},$$

where $k_A(n) = \lceil \frac{2\ln n}{h(\tau)} \rceil$ and

$$h(\tau) = (1-\epsilon)\left(\frac{1-\tau}{2} - \frac{1}{2}\ln(2-\tau)\right).$$

Remark: Note that Proposition 4.4.1 immediately implies $L_\tau^A(\mathbf{W}_n) \leq k_A(n)$ eventually almost surely.

Proof of Proposition 4.4.1: It is clear from the proof in Proposition 2.2.3 that to obtain the right hand side of probability bound above, it suffices to obtain an exponential upper bound on $P\left(g(V) \leq \tau\right)$ for any individual square submatrix $V \in \mathbf{W}$ with size k. As entries of $\mathbf{W}_n$ are i.i.d. Gaussian. $g(V)$ has a χ^2 distribution with $k^2 - 2k + 1$ degrees of freedom. Using a Chernoff type argument, it is easy to check that for any $\delta > 0$,

$$\begin{aligned} P(g(V) \geq \delta \cdot (k-1)^{-2}) &\leq \min_{s \leq \frac{1}{2}} (1-2s)^{\frac{(k-1)^2}{2}} e^{-s\delta} \\ &= \left(\left(\frac{(k-1)^2}{\delta} \right)^{-\frac{(k-1)^2}{2\delta}} \exp\{-\frac{1-(k-1)^2/\delta}{2}\} \right)^{\delta}. \end{aligned}$$

In conjunction with the following lemma, it is easy to show that

$$P\left(g(V) \leq \tau\right) \leq \left(\frac{(k-1)^2}{(2-\tau)(k-1)^2 - 4} \right)^{-\frac{(k-1)^2}{2}} \times \exp\{-\frac{1}{2}((1-\tau)(k-1)^2 - 4)\},$$

which implies Proposition 4.4.1.

Lemma 4.4.2. *Suppose $X \sim \chi_k^2$ for any $k \geq 3$. If $0 < t < k-2$ is any constant, then*

$$P(X \leq t) \leq P(X \geq 2k - 4 - t).$$

Proof of Lemma 4.4.2: Let f be the density function of X. Since

$$P(X \leq t) = \int_0^t f(s)ds$$

and

$$P(X \geq 2k-4-t) \geq \int_{2k-4-t}^{2k-4} f(s)ds,$$

it suffices to show that the ratio

$$\frac{f(s)}{f(2k-4-s)} \leq 1, \tag{4.4}$$

for any $0 < s < t$. Eventually we show below that the above ratio is less than 1 for any $0 < s < k-2$.

Note that the ratio (4.4) can be rewritten as

$$\begin{aligned} \frac{f(s)}{f(2k-4-s)} &= \frac{s^{(k-2)/2}e^{-s/2}}{(2k-4-s)^{(k-2)/2}e^{-(2k-4-s)/2}} \\ &= \left[(1-\frac{2k-4-2s}{2k-4-s})e^{4(k-2-s)/(k-2)}\right]^{(k-2)/2} \end{aligned} \tag{4.5}$$

Let $u = \frac{2k-4-s}{2k-4-2s}$. The ratio (4.5) becomes

$$\frac{f(s)}{f(2k-4-s)} = \left[(1-\frac{1}{u})e^{\frac{2}{2u-1}}\right]^{(k-2)/2}$$

A routine calculation shows that the derivative of above over u is

$$\begin{aligned} \left[(1-\frac{1}{u})e^{\frac{2}{2u-1}}\right]' &= e^{\frac{2}{2u-1}} \cdot \left(u^{-2} + (1-\frac{1}{u})(\frac{2}{2u-1})'\right) \\ &= e^{\frac{2}{2u-1}} \cdot \frac{(2u-1)^2 - 4(u-1)u}{u^2(2u-1)^2} \geq 0; \end{aligned}$$

and that when $u \to \infty$, which is equivalent to $s \to k-2$,

$$\lim_{u\to\infty}(1-\frac{1}{u})e^{\frac{2}{2u-1}} = 1.$$

Therefore, $\frac{f(s)}{f(2k-4-s)} \leq 1$ for any $0 < s < k-2$.

4.5 Significant Analysis of Non-square Biclusters in Real Matrices

Here we extend the results obtained in Sections 5.2 and 5.4 to the case of non-square matrices $\mathbf{W}_{mn}$ and non-square target submatrices. Suppose the row/column aspect ratio of $\mathbf{W}_{mn}$, $\frac{m}{n} = \alpha$, is fixed for some $\alpha > 0$. For any $\beta \geq 1$, let $K_\tau(\mathbf{W}, \lceil \alpha n \rceil, n, \beta)$ be the largest k such that there exists at least one $\lceil \beta k \rceil \times k$ submatrix in $\mathbf{W}_n$ with its average greater than τ. One can generalize Proposition 4.2.1 and obtain the proposition below by following similar steps as those in the proof of Proposition 3.4.1.

Proposition 4.5.1. *Fix $0 < \gamma < 1$. When n is sufficiently large,*

$$P\left(K_\tau(\mathbf{W}, \lceil \alpha n \rceil, n, \beta) \;\geq\; k(\lceil \alpha n \rceil, n, \beta, \tau) + r\right) \;\leq\; 2n^{-(\beta+1)\,r} \tag{4.6}$$

for each $1 \leq r \leq \gamma\, n$. Here $k(\lceil \alpha n \rceil, n, \beta, \tau) = \frac{2\beta+2}{\tau^2 \beta} \ln n + \frac{2}{\tau^2} \ln \alpha$.

Definition: Given $0 < \tau < 1$, let $L^A_\tau(\mathbf{W}, \lceil \alpha n \rceil, n, \beta)$ be the largest k such that there exists at least one $\lceil \beta k \rceil \times k$ submatrix V in $\mathbf{W}_n$ with $g(V) \leq \tau$.

Similarly, the non-square result below can be generalized from Proposition 4.4.1 in the same fashion as above.

Proposition 4.5.2. *Fix $0 < \gamma < 1$ and $0 < \tau < 1$. When n is sufficiently large,*

$$P\{L^A_\tau(\mathbf{W}, \lceil \alpha n \rceil, n, \beta) \;\geq\; k^A(\lceil \alpha n \rceil, n, \beta, \tau) + r\} \;\leq 2\; n^{-(\beta+1)\,r} \tag{4.7}$$

for each $1 \leq r \leq \gamma\, n$, where

$$k^A(\lceil \alpha n \rceil, n, \beta, \tau) = \frac{\beta+1}{h(\tau)\beta} \ln n + h(\tau)^{-1} \ln \alpha,$$

and $h(\tau)$ is defined same as that in Proposition 4.4.1.

Note that by Borel-Cantelli lemma, it follows immediately that $K_\tau(\mathbf{W}, \lceil \alpha n \rceil, n, \beta) \leq$

$k(\lceil \alpha n \rceil, n, \beta, \tau)$ and $L^A(\mathbf{W}, \lceil \alpha n \rceil, n, \beta) \leq k^A(\lceil \alpha n \rceil, n, \beta, \tau)$ eventually almost surely. Since the proofs of the above two propositions are straightforward, they are omitted.

4.6 Significance Analysis Under Non-Gaussian Assumption

In the previous sections, we assume entries are i.i.d. Gaussian. In the following analysis, instead, we assume the entries of matrix $\Theta = \{\theta_{ij}\}$ are i.i.d. and follow any bounded distribution with $E\theta = 0$ and $Var(\theta) = 1$. Namely, we have the alternative random matrix model below. Note that we will only consider square primary matrices and square submatrices here. The generalization to non-square cases is apparent from Section 4.5.

Alternative real-valued random matrix model: Let Θ_n be an $n \times n$ real-valued matrix with i.i.d. entries such that $|\theta_{ij}| < \kappa < \infty$ with probability 1.

Define $K_\tau(\Theta_n)$ same as that in the previous section. Since the entries in Θ_n are bounded, one can easily establish the following proposition on $K_\tau(\theta_n)$ by applying Hoeffding's inequality and by following similar steps as those in the proof of Proposition 2.2.3.

Proposition 4.6.1. *Fix $0 < \gamma < 1$ and $0 < \epsilon < 1$. Let $k_B(n) = \frac{4\kappa^2 \ln n}{\tau^2}$. When n is sufficiently large, for every integer $1 \leq r \leq \gamma n$, we have $P\left(K_\tau(\Theta_n) \geq k_B(n) + r\right) \leq 2n^{-(2-\epsilon)r}$.*

Remark: In fact, to obtain a probability upper bound like above, the bounded assumption can be further relaxed. For example, one can assume that the following alternative random matrix model where the entries are i.i.d. sub-Gaussian distributed.

Alternative real-valued random matrix model II: Let $\tilde{\Theta}_n$ be an $n \times n$ real-valued matrix with i.i.d. entries $\tilde{\theta}_{ij}$ having distribution P. Let f be any real-valued differentiable function. Suppose $\tilde{\theta}_{ij}$ further satisfies

$$\frac{\int f^2 \ln f^2 dP}{\int f^2 dP} \leq 2c \int f'^2 dP,$$

where c is a constant.

Note that the above inequality is known as log-Sobolev inequality. This condition is also known to be equivalent to $\tilde{\theta}_{ij}$ following an absolutely continuous distribution with a

sub-Gaussian tail. To generalize Proposition 4.6.1 under this alternative random matrix model, one may also need the following well known result.

Lemma 4.6.2. *(Herbst) Suppose random variable $X \sim P$ with log-Sobolev inequality constant c. If G is a Lipschitz function on R^d with Lipschitz constant $|G|_L$, then for any $\delta > 0$,*

$$P(|G(x) - E_P(G(x))| \geq \delta) \leq 2e^{-\delta^2/2c|G|_l^2}.$$

Now, we are ready to establish the following result.

Proposition 4.6.3. *Suppose $\tilde{\theta}_{ij}$ in the above alternative real-valued random matrix model follows a distribution satisfying log-Sobolev inequality with constant c. Then, for any fixed τ, $0 < \epsilon < 1$, and sufficiently large n, it follows that*

$$P\left(K_\tau(\tilde{\Theta}_n) \geq k_L(n) + r\right) \leq 2\, n^{-(2-\epsilon)\, r},$$

where $k_L(n) = \frac{4c^2}{\tau^2} \ln n$.

Proof of Proposition 4.6.3: Let V be any $k \times k$ submatrix in $\tilde{\Theta}$. By the proof of Proposition 2.2.3, it suffices to show that $P(F(V) \geq \tau) \leq e^{-\tau^2 k^2/2c^2}$. Note that $F(V) = \frac{\sum v_{ij}}{k^2}$ is a Lipschitz function with constant k^{-2}. Thus, by the assumption that $\tilde{\theta}_{ij}$ satisfies log-Sobolev inequality with constant c and by Herbst lemma, it is clear that

$$P(F(V) \geq \tau) \leq e^{-\tau^2 k^2/2c^2},$$

which completes the proof.

Now, we want to get a result similar to Proposition 4.6.1 for biclusters satisfying the ANOVA criterion. Note that for any square submatrix V of Θ_n, $g(V)$ can be expressed as a sum of dependent random variables. Moreover, from the assumption that the entries of Θ_n are bounded, it is easy to see that $g(V)$ satisfies the conditions of McDiarmid inequality (17). By applying McDiarmid inequality to $P\,(g(V) \leq \tau)$, it is readily to establish the following proposition. The detailed proof can be found in Section 4.8.

Proposition 4.6.4. *Fix $0 < \gamma < 1$ and $0 < \tau < 1$. When n is sufficiently large, for every integer $1 \le r \le \gamma\, n$, $P\left(M_A^{\tau}(\mathbf{W}) \ge s_A(n) + r\right) = 2\, n^{-(2-\epsilon)\, r}$, where $s_A(n) = \frac{c \ln n}{(1-\tau)}$ and $c = 3600\kappa^4$.*

Remark: The constant c can be further reduced in particular applications with known correlation structures.

4.7 Proof of Theorem 4.2.2

Lemma 4.7.1. *Fix $\tau > 0$. There exist n_0 and k_0 such that for any $n \ge n_0$ and $k_0 \le k \le \tilde{s}(n) - 2$, there exists a constant $C_1 > 0$ such that*

$$\frac{Var\, U_k(\tau, n)}{(EU_k(\tau, n))^2} \le C_1 \sum_{l=1}^{k} \sum_{r=1}^{k} \frac{\binom{k}{l}\binom{n-k}{k-l}}{\binom{n}{k}} \frac{\binom{k}{r}\binom{n-k}{k-r}}{\binom{n}{k}} k^4 \cdot \exp\left\{ \frac{rl\tau^2}{2} \left(1 + \frac{k^2 - rl}{k^2 + rl} \right) \right\}. \tag{4.8}$$

Proof: Let $\mathcal{S}_k$ be the collection of all index sets of $k \times k$ square submatrices in Z_n with $k \le \tilde{s}(n) - 2$. It is clear that

$$EU_k(n, \tau) = \sum_{V \in \mathcal{S}_k} EI\{F(V) > \tau\} = \binom{n}{k}^2 \left(1 - \Phi(k\tau)\right), \tag{4.9}$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution. In a similar fashion, we have

$$EU_k^2(n, \tau) \;=\; \left(\sum_{V \in \mathcal{S}_k} EI\{F(V) > \tau\} \right)^2 = \sum_{V_i, V_j \in \mathcal{S}_k} EI\{F(V_i) > \tau\} \cdot I\{F(V_j) > \tau\}.$$

One can decompose $EU_k^2(n, \tau)$ as $EU_k^2(n, \tau) = H_1 + H_2 + H_3$ according to different degrees of overlap between index sets (V_i, V_j). Here H_1 includes pairs of submatrices without common entries, H_2 includes pairs of submatrices with both common and non-common entries, and H_3 includes pairs of submatrices with all entries being common. To be more precise, for any pair (V_i, V_j), let r be the number of common rows between V_i and V_j and let l be the number of common columns. Then,

$$H_1 \;=\; \sum_{\{V_i, V_j \in \mathcal{S}_k\}} I\{V_i \cap V_j = \emptyset\} \cdot EI\{F(V_i) > \tau\} \cdot I\{F(V_j) > \tau\}$$

$$= \sum_{\substack{r,l \text{ s.t.} \\ \min(r,l)=0 \\ \max(r,l)\le k}} I\{V_i \cap V_j = \emptyset\} P(F(V_i) \ge \tau)^2,$$

$$\begin{aligned} H_2 &= \sum_{\{V_i,V_j \in \mathcal{S}_k\}} I\{V_i \cap V_j \neq \emptyset, V_i \cap V_j \neq V_i\} \cdot [EI\{F(V_i) > \tau\} \cdot I\{F(V_j) > \tau\}] \\ &= \sum_{l=1}^{k}\sum_{r=1}^{k} \binom{n}{k}\binom{k}{l}\binom{n-k}{k-l}\binom{n}{k}\binom{k}{r}\binom{n-k}{k-r} \\ &\quad \int_{-\infty}^{\infty} \frac{\sqrt{rl}}{\sqrt{2\pi}} e^{-\frac{rlt^2}{2}} \times \left(1 - \Phi(\frac{k^2\tau - rlt}{\sqrt{k^2 - rl}})\right)^2 \cdot I\{rl \neq k^2\} dt \end{aligned} \tag{4.10}$$

and

$$\begin{aligned} H_3 &= \sum_{\{V_i,V_j \in \mathcal{S}_k\}} I\{V_i \cap V_j = V_i\} \cdot E[I\{F(V_i) > \tau\} \cdot I\{F(V_j) > \tau\}] \\ &= \binom{n}{k}^2 (1 - \Phi(k\tau)). \end{aligned}$$

Note that, as argued in the proof of Theorem 2.2.4,

$$\frac{Var\, U_k(n,\tau)}{(EU_k(n,\tau))^2} = \sum_{l=0}^{k}\sum_{r=0}^{k} \frac{\binom{k}{l}\binom{n-k}{k-l}}{\binom{n}{k}} \frac{\binom{k}{r}\binom{n-k}{k-r}}{\binom{n}{k}} \left(\frac{P(F(V_i) > \tau, F(V_j) > \tau)}{P(F(V_i) > \tau)^2} - 1\right).$$

Moreover, $P(F(V_i) > \tau, F(V_j) > \tau)/P(F(V_i) > \tau)^2 - 1 = 0$ for any pair (V_i, V_j) without common entries. Thus,

$$\frac{Var\, U_k(n,\tau)}{(EU_k(n,\tau))^2} \le \frac{H_2 + H_3}{(EU_k(n,\tau))^2}.$$

It is also clear that

$$\frac{H_3}{(EU_k(n,\tau))^2} = \binom{n}{k}^{-2} (1 - \Phi(k\tau))^{-1} \le C_2 \binom{n}{k}^{-2} \exp\left\{\frac{rl\tau^2}{2}\right\},$$

where the last inequality follows from the fact that when u is sufficiently large and $rl = k^2$ for H_3, $1 - \Phi(u) \ge c\, u^{-1}\, e^{-\frac{u^2}{2}}$ for some constant $c > 0$. This corresponds to $r = l = k$ in

(4.8). It remains to bound $\frac{H_2}{(EU_k(n,\tau))^2}$. When k is sufficiently large,

$$EU_k(n,\tau) \geq c\binom{n}{k}^2 (k\tau)^{-1} e^{-\frac{k^2\tau^2}{2}}.$$

Thus, we only need to show that

$$\begin{aligned} H_2 &\leq C \cdot \sum_{l=1}^{k}\sum_{r=1}^{k} \binom{k}{l}\binom{n-k}{k-l}\binom{n}{k}\binom{k}{r}\binom{n-k}{k-r}\binom{n}{k} k^2 \\ &\quad \times \exp\left\{-k^2\tau^2 + \frac{rl\tau^2}{2}\left(1 + \frac{k^2-rl}{k^2+rl}\right)\right\} \cdot I\{rl \neq k^2\}. \end{aligned} \tag{4.11}$$

By considering the cases $|k^2\tau - rlt| \geq 1$ and $|k^2\tau - rlt| < 1$ separately, one can rewrite H_2 as

$$\begin{aligned} H_2 &= \sum_{l=1}^{k}\sum_{r=1}^{k} \binom{n}{k}\binom{k}{l}\binom{n-k}{k-l}\binom{n}{k}\binom{k}{r}\binom{n-k}{k-r} \\ &\quad I\{rl \neq k^2\} \int_{-\infty}^{\infty} \frac{\sqrt{rl}}{\sqrt{2\pi}} e^{-\frac{rlt^2}{2}} \cdot \left(1 - \Phi\left(\frac{k^2\tau - rlt}{\sqrt{k^2-rl}}\right)\right)^2 I\{|k^2\tau - rlt| \geq 1\} \\ &\quad + \frac{\sqrt{rl}}{\sqrt{2\pi}} e^{-\frac{rlt^2}{2}} \cdot \left(1 - \Phi\left(\frac{k^2\tau - rlt}{\sqrt{k^2-rl}}\right)\right)^2 I\{|k^2\tau - rlt| < 1\} dt. \end{aligned} \tag{4.12}$$

To begin, we will bound the integral of the first term in the brackets above. By the assumptions that $rl \neq k^2$ and $|k^2\tau - rlt| \geq 1$, it follows that

$$\begin{aligned} 1 - \Phi\left(\frac{k^2\tau - rlt}{\sqrt{k^2-rl}}\right) &\leq \frac{\sqrt{k^2-rl}}{\sqrt{2\pi}(k^2\tau - rlt)} \exp\left\{-\frac{(k^2\tau - rlt)^2}{2(k^2-rl)}\right\} \\ &\leq O(\sqrt{k^2-rl}) \exp\left\{-\frac{(k^2\tau - rlt)^2}{2(k^2-rl)}\right\} =: G. \end{aligned}$$

Thus, the first term in the brackets in (4.12) is bounded by $\frac{\sqrt{rl}}{\sqrt{2\pi}} e^{-\frac{rlt^2}{2}} \cdot G^2 \cdot I\{|k^2\tau - rlt| \geq 1\}$. Moreover, we observe that the exponential part of $e^{-\frac{rlt^2}{2}} \cdot G^2$ is

$$\begin{aligned} -\frac{(k^2\tau - rlt)^2}{(k^2-rl)} - \frac{rlt^2}{2} &= -\frac{(k^2-rl)^2\tau^2 + 2rl(\tau-t)(k^2-rl)\tau + r^2l^2(\tau-t)^2}{k^2-rl} - \frac{rlt^2}{2} \\ &= -(k^2-rl)\tau^2 + 2rl\tau(t-\tau) - \frac{r^2l^2(\tau-t)^2}{k^2-rl} - \frac{rlt^2}{2} \\ &= -(k^2 - \frac{rl}{2})\tau^2 - \frac{3}{2}\tau^2 rl + 2rl\tau t - \frac{r^2l^2(\tau-t)^2}{k^2-rl} - \frac{rlt^2}{2}. \end{aligned}$$

Comparing this with (4.11), we only need to verify,

$$\frac{3}{2}\tau^2 rl \;-\; 2\tau rlt \;+\; \frac{rlt^2}{2} \;+\; \frac{r^2l^2(\tau-t)^2}{k^2-rl} \;\;\geq\;\; -\frac{rl\tau^2(k^2-rl)}{2(k^2+rl)}. \tag{4.13}$$

This inequality is easy to verify as a quadratic function of τ. Thus, when $rl \neq k^2$,

$$\begin{aligned} &\int_{-\infty}^{\infty} \frac{\sqrt{rl}}{\sqrt{2\pi}}\; e^{-\frac{rlt^2}{2}} \cdot G^2 \;\cdot\; I\{|k^2\tau - rlt| \geq 1\}\, dt \\ \leq\;\; & O(k^2-rl) \;\; \exp\left\{-(k^2-\frac{rl}{2})\tau^2 \;+\; \frac{rl\tau^2(k^2-rl)}{2(k^2+rl)}\right\}. \end{aligned}$$

Next, we consider the integral of the second term in the bracket of (4.12). Note that $|k^2\tau - rlt| < 1$ is equivalent to $t \in (\frac{k^2\tau-1}{rl}, \frac{k^2\tau+1}{rl})$. Thus, it follows that

$$\begin{aligned} & I\{rl \neq k^2\} \int_{-\infty}^{\infty} \frac{\sqrt{rl}}{\sqrt{2\pi}} e^{-\frac{rlt^2}{2}} \times \left(1 - \Phi\left(\frac{k^2\tau - rlt}{\sqrt{k^2-rl}}\right)\right)^2 \;\cdot\; I\{|k^2\tau - rlt| < 1\} dt \\ \leq\;\; & \int_{\frac{k^2\tau-1}{rl}}^{\frac{k^2\tau+1}{rl}} \frac{\sqrt{rl}}{\sqrt{2\pi}} e^{-\frac{rlt^2}{2}} dt = \Phi\left(\frac{k^2\tau+1}{\sqrt{rl}}\right) - \Phi\left(\frac{k^2\tau-1}{\sqrt{rl}}\right) \\ \leq\;\; & 1 - \Phi\left(\frac{k^2\tau-1}{\sqrt{rl}}\right) \leq \frac{k\sqrt{rl}}{\sqrt{2\pi}(k^2\tau-1)} e^{-\frac{(k^2\tau-1)^2}{2rl} - \ln k}. \end{aligned} \tag{4.14}$$

Comparing the right hand side of inequality (4.14) with (4.11), we wish to show that, for any fixed $\tau > 0$, when k is sufficiently large,

$$\frac{(k^2\tau-1)^2}{2rl} + \ln k \geq (k^2 - \frac{rl}{2})\tau^2.$$

By elementary algebra, this is equivalent to

$$(k^2-rl)^2\tau^2 - 2k^2\tau + 1 + 2rl\ln k \geq 0. \tag{4.15}$$

Suppose first that $rl \geq k^2 - \frac{k}{\sqrt{\ln k}}$. In this case, the quantity above is at least

$$-2k^2\tau + 1 + 2rl\ln k \geq -2k^2\tau + 1 + 2k^2\ln k - 2k\sqrt{\ln k} > 0,$$

when k is sufficiently large. Suppose now that $rl < k^2 - \frac{k}{\sqrt{\ln k}}$, or equivalently $k^2 - rl > \frac{k}{\sqrt{\ln k}}$,

As a quadratic function of τ, (4.15) takes its minimum value at $\tau = \frac{k^2}{(k^2-rl)^2}$, and the corresponding minimum value is $rl[-2k^2 + rl + 2(k^2 - rl)^2 \ln k]/(k^2 - rl)^2$. In this case, our assumption yields that

$$-2k^2 + rl + 2(k^2 - rl)^2 \ln k > rl > 0.$$

This establish (4.11) and complete the proof for $I\{rl \neq k^2\}$

Lemma 4.7.2. *Fix $\tau > 0$. There exists $k_0 > 0$ such that for any $k > k_0$ and n satisfies $k \leq \frac{4}{\tau^2} \ln n - \frac{4}{\tau^2} \ln\left(\frac{4}{\tau^2} \ln n\right) - \frac{12 \ln 2}{\tau^2}$,*

$$\frac{Var\, U_k(\tau, n)}{(EU_k(\tau, n))^2} \leq k^{-2}. \tag{4.16}$$

Comments: In fact, one only needs to show that the sum of the left hand side above over k is finite. Here, k^{-2} is obviously enough, and showing the inequality above for k^{-2} can also avoid the complicated notations in the proof.

Proof: By Lemma 4.7.1, it suffices to show that

$$\sum_{l=1}^{k} \sum_{r=1}^{k} \frac{\binom{k}{l}\binom{n-k}{k-l}}{\binom{n}{k}} \frac{\binom{k}{r}\binom{n-k}{k-r}}{\binom{n}{k}} \cdot k^4 \cdot \exp\left\{\frac{rl\tau^2}{2}\left(1 + \frac{k^2 - rl}{k^2 + rl}\right)\right\} \leq k^{-2}. \tag{4.17}$$

To establish (4.17), we wish to show that each term in the sum is less than k^{-4}. To begin, note that

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

and that $(n-k)^{-l} = O(1)n^{-l}$ when $l \leq k = o(n^{1/2})$. Therefore,

$$\frac{\binom{k}{l}\binom{n-k}{k-l}}{\binom{n}{k}} \frac{\binom{k}{r}\binom{n-k}{k-r}}{\binom{n}{k}} k^4 \leq k^4 \cdot \binom{k}{r}\binom{k}{l} \cdot k^{r+l} \cdot n^{-r-l} \cdot O(1). \tag{4.18}$$

Rewriting the condition on k as $\ln n \geq \frac{\tau^2 k}{4} + \ln(\frac{4}{\tau^2} \ln n) + 3 \ln 2$, one has

$$k^{r+l} \cdot n^{-r-l} \cdot \exp\left\{\frac{rl\tau^2}{2}\left(1 + \frac{k^2 - rl}{k^2 + rl}\right)\right\}$$

$$\leq \quad k^{r+l} \ \cdot \ e^{-3(r+l)\ln 2} \ \cdot \ (\frac{4}{\tau^2}\ln n)^{-r-l} \ \cdot \ \exp\left\{\frac{\tau^2}{2}\left(rl\frac{2k^2}{k^2+rl} \ - \ \frac{k}{2}(r+l)\right)\right\}.$$

Moreover, the assumption that $k < \frac{4}{\tau^2}\ln n$ implies

$$k^{r+l}\cdot e^{-3(r+l)\ln 2}\cdot(\frac{4}{\tau^2}\ln n)^{-r-l} \leq e^{-3(r+l)\ln 2}. \tag{4.19}$$

Thus, to establish (4.18), it suffices to show that

$$\binom{k}{r}\binom{k}{l}e^{-3(r+l)\ln 2}\cdot\exp\left\{\frac{\tau^2}{2}\left(rl\frac{2k^2}{k^2+rl}-\frac{k}{2}(r+l)\right)\right\} \leq k^{-8}. \tag{4.20}$$

To this end, we first examine the exponential term above. Since $r+l \geq 2\sqrt{rl}$ and $k^2+rl \geq 2\sqrt{k^2rl}$, it follows that

$$rl\frac{2k^2}{k^2+rl}-\frac{k}{2}(r+l) \leq \frac{2rlk^2}{2\sqrt{k^2\cdot rl}}-k\sqrt{rl}=0. \tag{4.21}$$

Suppose now that $r+l > \frac{3k}{4}$ and k is sufficiently large. Then,

$$\binom{k}{r}\binom{k}{l}e^{-3(r+l)\ln 2} \leq \binom{2k}{r+l}\cdot e^{-3(r+l)\ln 2} \leq 2^{2k}\cdot 2^{-\frac{9k}{4}} \leq k^{-8}, \tag{4.22}$$

where the first inequality follows from the simple fact that $\binom{k}{r}\binom{k}{l}=\binom{2k}{r+l}$ and the second inequality follows from the fact that $\binom{2k}{r+l}\leq 2^{2k}$.

Now, it remains to establish (4.20) when k is sufficiently large and $r+l \leq \frac{3k}{4}$. To show this, one may verify that

$$\begin{aligned}
& k^8\cdot\binom{2k}{r+l}\cdot\exp\left\{\frac{\tau^2}{2}[rl\frac{2k^2}{k^2+rl}-\frac{k}{2}(r+l)]\right\} \\
< \quad & \exp\left(\frac{\tau^2}{2}[\frac{(r+l)^2}{2}-\frac{k}{2}(r+l)]+8\ln k+(r+l)\ln 2k\right) \\
= \quad & \exp\left(\frac{\tau^2(r+l)}{2}[\frac{(r+l)}{2}-\frac{k}{2}+\frac{16\ln k}{(r+l)\tau^2}+\frac{2\ln 2k}{\tau^2}]\right) \\
\leq \quad & \exp\left(\frac{\tau^2(r+l)}{2}[\frac{3k}{8}-\frac{k}{2}+\frac{16\ln k}{\tau^2}+\frac{2\ln 2k}{\tau^2}]\right)\leq 1
\end{aligned} \tag{4.23}$$

where the first inequality follows by $rl\frac{2k^2}{k^2+rl} \leq \frac{(r+l)^2}{4}\cdot\frac{2k^2}{k^2+rl} \leq \frac{(r+l)^2}{2}$ and the last inequality

holds when k is sufficiently large.

Putting (4.21), (4.22) and (4.23) together, we have completed the proof.

Proof of Theorem 4.2.2: By Proposition 4.2.1 and the Borel-Cantelli lemma, eventually almost surely, $K_\tau(\mathbf{W}_n) \le \lceil \tilde{s}(n) \rceil + 1$. Thus, we only need to establish a lower bound on $K(\mathbf{W}_n)$. To this end, let

$$f(n) = \frac{4}{\tau^2} \ln n - \frac{4}{\tau^2} \ln \left(\frac{4}{\tau^2} \ln n \right) - \frac{12 \ln 2}{\tau^2}$$

for any integer $n > 0$, and let $g(k) = \min\{r \ge 1, \lfloor f(r) \rfloor = k\}$ for any integer $k > 0$. It is not hard to verify that for sufficiently large k, $g(k)$ is strictly monotone increasing, and $g(k)$ tends to infinity as k tends to infinity. Thus there exists an integer $n_0 \ge 1$ such that for any $n \ge n_0$, there exists $k = k(n)$ such that

$$g(k) \le n < g(k+1). \tag{4.24}$$

Now, let

$$A_m = \bigcup_{n>m} \{K_\tau(\mathbf{W}_n) < \tilde{s}(n) - \frac{12 \ln 2}{\tau^2} - \tilde{C} - 2\}.$$

By the above argument, when m is sufficiently large,

$$A_m \subset \bigcup_{k \ge \lfloor f(m) \rfloor} \bigcup_{g(k) \le n < g(k+1)} \{K_\tau(\mathbf{W}_n) < \tilde{s}(n) - \frac{12 \ln 2}{\tau^2} - \tilde{C} - 2\}.$$

Note that when n satisfies (4.24), the definition of $g(\cdot)$ ensures that $k \le f(n) < k+1$ and

$$\begin{aligned} 1 &= k + 1 - k > f(n) - \lfloor f(g(k)) \rfloor \ge f(n) - f(g(k)) \\ &= \tilde{s}(n) - \tilde{C} - o(1) - \frac{12 \ln 2}{\tau^2} - [\tilde{s}(g(k)) - \tilde{C} - o(1) - \frac{12 \ln 2}{\tau^2}], \end{aligned}$$

which is equivalent to $\tilde{s}(n) < \tilde{s}(g(k)) + 1$. Thus,

$$A_m \subset \bigcup_{k \geq \lfloor f(m) \rfloor} \bigcup_{g(k) \leq n < g(k+1)} \{K_\tau(\mathbf{W}_n) < \tilde{s}(g(k)) - \frac{12 \ln 2}{\tau^2} - \tilde{C} - 1\}.$$

By the monotonicity of $K_\tau(\mathbf{W}_n)$ in n, it immediately follows that $K_\tau(\mathbf{W}_{g(k)}) \leq K_\tau(\mathbf{W}_n)$. Consequently,

$$A_m \subset \bigcup_{k \geq \lfloor f(m) \rfloor} \bigcup_{g(k) \leq n < g(k+1)} \{K_\tau(\mathbf{W}_{g(k)}) < \tilde{s}(g(k)) - \frac{12 \ln 2}{\tau^2} - \tilde{C} - 1\}.$$

Let $k^* = \lfloor \tilde{s}(g(k)) - \frac{12 \ln 2}{\tau^2} - \tilde{C} \rfloor$. Then, by Chebyshev's inequality, it follows that

$$\begin{aligned} \sum_{k=1}^{\infty} P\left(K_\tau(\mathbf{W}_{g(k)}) < k^* - 1\right) &= \sum_{k=1}^{\infty} P\left(U_{k^*-1}(\tau, g(k)) = 0\right) \\ &\leq \sum_{k=1}^{\infty} \frac{Var\, U_{k^*-1}(\tau, g(k))}{(E U_{k^*-1}(\tau, g(k)))^2}. \end{aligned}$$

Note that by definition, $k^*(n) - 1 < \lfloor f(g(k)) \rfloor$, which satisfies the condition of Lemma 4.7.2. Thus

$$\sum_{k=1}^{\infty} P\left(K_\tau(\mathbf{W}_{g(k)}) < k^* - 1\right) < \infty,$$

and the Borel-Cantelli lemma immediately implies that $P(A_m) \to 0$ as $m \to \infty$.

4.8 Proof of Proposition 4.6.4

Let V be any $l \times l$ submatrix of Θ_n satisfying $g(V) \leq \tau$ with $l = k(n) + r$. From the proof of Proposition 4.2.1, it is clear that one only needs to show that $P(g(V_l) \leq \tau) \leq e^{-l^2(1-\tau)^2/C}$ for some constant C, where $0 < \tau < 1$. Let $V^{(i'j')}$ be an identical replicate submatrix of V except replacing the (i', j') entry of V with another random replicate $v^*_{i'j'}$ with the same distribution. One can verify that

$$g(V) - g(V^{(i'j')}) = \frac{I + II + III}{l^2},$$

where

$$I = \sum_{i\neq i',j\neq j'} \frac{(v_{i'j'} - v^*_{i'j'})^2}{l^4} - 2\sum_{i\neq i',j\neq j'} \frac{(v_{ij} - \overline{v}_{i.} - \overline{v}_{.j} + \overline{v}_{..})(v_{i'j'} - v^*_{i'j'})}{l^2},$$

$$\begin{aligned} II \;\; = \;\; & \sum_{i=i',j\neq j'} (v_{i'j'} - v^*_{i'j'})^2(\frac{1}{l} - \frac{1}{l^2})^2 + 2\sum_{i\neq i',j=j'} (v_{ij} - \overline{v}_{i.} - \overline{v}_{.j} + \overline{v}_{..})(v_{i'j'} - v^*_{i'j'})(\frac{1}{l} - \frac{1}{l^2}) \\ & + \sum_{i\neq i',j=j'} (v_{i'j'} - v^*_{i'j'})^2(\frac{1}{l} - \frac{1}{l^2})^2 + 2\sum_{i\neq i',j=j'} (v_{ij} - \overline{v}_{i.} - \overline{v}_{.j} + \overline{v}_{..})(v_{i'j'} - v^*_{i'j'})(\frac{1}{l} - \frac{1}{l^2}), \end{aligned}$$

and

$$III = \sum_{i=i',j=j'} (v_{i'j'} - v^*_{i'j'})^2(1 + \frac{1}{l} - \frac{1}{l^2})^2 + 2\sum_{i=i',j=j'} (v_{ij} - \overline{v}_{i.} - \overline{v}_{.j} + \overline{v}_{..})(v_{i'j'} - v^*_{i'j'})(1 + \frac{2}{l} - \frac{1}{l^2}).$$

Since $|v_{ij}|$ and $|v^*_{ij}|$ are both bounded by κ, it is easy to check that $\sup|I + II + III| \leq 29\kappa(2\kappa + \kappa/l)$ for any (i', j'). Let $C = 3600\kappa^4$. By the concentration type inequality of McDiarmid ((17)) and the condition that $E(g(V)) = 1 - o(1)$, it follows directly that $P(g(V) \leq \tau) \leq e^{-2l^2(1-\tau)^2/C}$.

CHAPTER 5

Recoverability of High Average Submatrices in Real Matrices with Noise

Recall that in order to account for, and study the potential effects of, noise on frequent itemset mining, we studied a simple binary additive noise model (3.1) in Chapter 3:

$$\mathbf{Y} \;=\; \mathbf{X} \oplus \mathbf{Z}.$$

We then obtained results on the recoverability of AFI estimator and showed that block estimation by AFI estimator asymptotically converges to the true underlying block in simple recovery problems. In the case of real-valued matrices, we are still interested in studying the recoverability of biclusters satisfying certain biclustering criteria. Here, we only focus on the simplest criterion, the average criterion.

5.1 Additive Gaussian Noise Model

We consider the standard Gaussian additive noise model,

$$\mathbf{Y} \;=\; \mathbf{X} + \mathbf{Z}, \tag{5.1}$$

where the matrix $\mathbf{Y} = \{y_{i,j}\}$ is the observed data, $\mathbf{X} = \{x_{i,j}\}$ is a deterministic matrix, and $\mathbf{Z} = \{z_{i,j}\}$ is a random matrix with independent standard Gaussian entries. Model (5.1) is widely used in statistics and engineering. For example, Arias-Castro *et al.* in (6; 5) study how to detect smoothed geometric shapes in noisy images under model (5.1). Their method, like most of other methods in image de-noising, deals with contiguous data structures. Some problems involving recovery of non-contiguous data with particular structures have also

been studied. For example, Zhou and Woodruff in (71) study how to recover a standard clustering similarity matrix from noise by matrix powering. However, none of the existing studies is coincident with our interests here. We are interested in a more general family of target matrices $\mathbf{X}$ than those considered above. For example, we do not assume $\mathbf{X}$ to be symmetric or have contiguous (smooth) structure.

Note that this additive noise model can be viewed as a special case of a more general additive model. In the more general model, $\mathbf{X}$ is not deterministic but has a Gaussian mixture structure, which can be regarded as foreground signals. The objective of recovery then becomes detecting and recovering foreground signals from background noise. This is also the model proposed in PLAID(40). Our model is a special case of the Gaussian mixture model in which each entry has the same variance.

5.2 Recoverability of Submatrices with High Average

Let $\mathbf{X}$ be an $n \times n$ target matrix. Suppose that $\mathbf{X}$ contains an $l \times l$ submatrix $\mathbf{U}$ with an index set C^*. The entries belonging to C^* are greater than some constant $\mu > 0$, and the entries belonging to $\mathbf{X} \backslash C^*$ are zero.

Fix $0 < \hat{\mu} < \mu$ and let $\hat{\mathcal{C}}$ be the family of index sets of square submatrices $\mathbf{U} \subset \mathbf{Y}$ with $F(\mathbf{U}) > \hat{\mu}$. We estimate C^* by the index of the largest high average square submatrix in the observed matrix $\mathbf{Y}$. More precisely, define

$$\hat{C} \ = \ \text{argmax}_{C \in \hat{\mathcal{C}}} |C|$$

and let the ratio

$$\Gamma(\hat{C}) \ = \ |\hat{C} \cap C^*| / |C^*|$$

measure the proportion of entries recovered by $\hat{C}$ in the true index set C^*.

Note that by Proposition 4.2.1, we should not expect to find a large submatrix with its average greater than μ. Therefore, we are using a threshold $\hat{\mu}$ which is less than μ. The estimator $\hat{C}$ described above satisfies the following theorem.

Theorem 5.2.1. *Fix $\frac{3}{2} < \delta < 2$, $0 < \epsilon < 2 - \delta$ and $0 < \alpha < 2\delta - 3$. When n is sufficiently*

large, if l satisfies $(\ln n)^{1/\alpha} \leq l < n$, then

$$P\left(\frac{|\hat{C}|}{|C^*|} \geq \frac{\mu}{\hat{\mu}}\right) \leq n^{-2} \text{ and } P\left(\Gamma(\hat{C}) < \frac{l^2}{l^2 + l^\delta}\right) \leq 3\,n\,\exp\{-l^{2-2\epsilon}\} + 3\,n^{-2}, \qquad (5.2)$$

where $\Gamma(\hat{C}) = \frac{|\hat{C} \cap C^*|}{|C^*|}$.

Remark: (i) The probability bounds n^{-2} and $3\,n^{-2}$ above can be improved. The resulting proof does not substantially differ from the current proof of 5.2.1 and is omitted.

(ii) It follows from the proof of Theorem 5.2.1 that the i.i.d Gaussian assumption can be further relaxed to the i.i.d. sub-Gaussian condition used in Proposition 4.6.3.

(iii) Theorem 5.2.1 shows that, with high probability, $\hat{C}$ contains a large proportion of C^*, but allows $\hat{C}$ to be larger than C^*. By contrast, the AFI estimator $\hat{C}$ of Theorem 3.5.1 is an estimator such that $|\hat{C}|$ is not much larger than $|C^*|$. One reason for the better performance of the AFI estimator is that the AFI criterion requires both the averages of each row and each column in the submatrix be large, which is stronger than the average criterion in this section. In fact, if μ can be estimated accurately enough such that $\hat{\mu} = \mu - O(l^{-2})$, then the result in Theorem 5.2.1 can be improved to

$$P\left(\Lambda(\hat{C}) < \frac{l^2}{l^2 + l^\delta}\right) = O(n^{-2}),$$

where $\Lambda(\hat{C})$ is defined as in Chapter 3. However, in practice, estimation of $\hat{\mu}$ can not always be guaranteed. In order to better recover C^*, one may further explore the estimator $\hat{C}$ obtained from the procedure described above. To be specific, the independent row and column scanning of Reuning-Scherer in (55) can be applied to $\hat{C}$. Since by Theorem 5.2.1 $\frac{|\hat{C}|}{|C^*|} \leq \frac{\mu}{\hat{\mu}}$ with high probability, the row and column scanning will be effective and by choosing a threshold properly, one can further show this two-stage estimation can provide a consistent estimate of target submatrix C^*.

5.3 Proof of Theorem 5.2.1

The proof of Theorem 5.2.1 requires a simple preliminary lemma which has been stated before in Chapter 3 in the binary case. It is easy to verify that it is also true here.

Lemma 5.3.1. *If for some $0 < \eta < 1$, there exists a $k \times s$ submatrix $\mathbf{R}$ satisfying $F(\mathbf{R}) \geq \eta$, then for any $v \leq \min\{k, s\}$, there exists a $v \times v$ submatrix $\mathbf{R}'$ of $\mathbf{R}$ such that $F(\mathbf{R}') \geq \eta$.*

Proof of Theorem 5.2.1: Let $\beta > \frac{\mu}{\hat{\mu}}$ be any constant. First, we want to bound $P(|\hat{C}| \geq \beta\,|C^*|)$. We begin with the following simple argument. Let $m = \lceil \sqrt{\beta|C^*|} \rceil$. When $|\hat{C}| > m^2$, one can rank the columns in $\hat{C}$ according to their averages, and drop the $\sqrt{|\hat{C}|} - m$ columns with smallest averages. After performing a similar row operation on the new rectangular submatrix, one obtains a submatrix $\tilde{C}$ such that satisfies $|\tilde{C}| = m^2$ and $F(\tilde{C}) \geq \hat{\mu}$ by the definition of $\hat{C}$. By the entry-wise normality assumption, it is easy to see that for any $m \times m$ submatrix V in $\mathbf{Y}$ having r rows and k columns in common with C^*, $F(V) \sim N\left(\frac{rk\mu}{m^2}, m^{-2}\right)$. Therefore,

$$\begin{aligned} P\left(\frac{|\hat{C}|}{|C^*|} \geq \beta\right) &\leq P\left(\max_{V \subset \mathbf{Y},} F(V) \geq \hat{\mu}\right) \\ &\leq \sum_{r=1}^{l}\sum_{k=1}^{l} \binom{n-l}{m-r} \cdot \binom{l}{r} \cdot \binom{n-l}{m-k} \cdot \binom{l}{k} \\ &\quad \times \exp\{-m^2 \cdot (\hat{\mu} - \beta^{-1}\,\mu)\} \\ &\leq n^{-2}, \end{aligned} \tag{5.3}$$

where the max in the first inequality is taken over all $m \times m$ submatrices in $\mathbf{Y}$, and the second inequality follows from the fact that for any $m \times m$ submatrix V, $\mathrm{E}\,[F(V)] = \frac{rk\mu}{m^2} \leq \frac{|C^*|\mu}{m^2} \leq \mu\beta^{-1} \leq \hat{\mu}$ (by the definition of β). The last inequality follows from the fact that $\sum_{r=1}^{l} \binom{n-l}{m-r} \cdot \binom{l}{r} \leq \binom{n}{m}$, the assumption on α, and the proof of Proposition 1.

Next, we want to give a lower bound on $|\hat{C}|$. Actually, we wish to bound $P\left(\frac{\mu - l^{-\epsilon}}{\hat{\mu} + l^{-\epsilon}} > \frac{|\hat{C}|}{|C^*|}\right)$ from above. To this end, we will show that with high probability, one can always construct a sufficiently large square submatrix D containing C^* and satisfying $F(D) \geq \hat{\mu}$. To begin, let ϵ be any number between $2 - \delta$ and $\frac{1}{2}$. Suppose $C^* = A \times B$. Consider each column c_i, $i = 1, ..., n - l$, in $\mathbf{Y}\,[A \times B^c]$. It is easy to see that $F(c_i) \sim N(0, l^{-1})$. Thus, when l is sufficiently large,

$$P\left(\min_{i=1,\ldots,n-l} F(c_i) \geq -l^{-\epsilon}\right) \geq 1 - (n-l) \cdot \exp\{-l^{1-2\epsilon}\}. \tag{5.4}$$

Similarly, for each row r_i, $i = 1, ..., n-l$, in $\mathbf{Y}\,[A^c \times B]$,

$$P\left(\min_{i=1,...,n-l} F(r_i) \geq -l^{-\epsilon}\right) \geq 1 - (n-l)\cdot \exp\{-l^{1-2\epsilon}\}. \tag{5.5}$$

Moreover, since $F(C^*) \sim N(\mu, l^{-2})$,

$$P\left(F(C^*) < \mu - l^{-\epsilon}\right) \leq \exp\{-l^{2-2\epsilon}\}. \tag{5.6}$$

Let $\mathcal{B}$ be the event that there does not exist a square submatrix $R = A' \times B'$ in $\mathbf{Y}[A^c \times B^c]$ with average greater than $-l^{-\epsilon}$ and $|A'| = |B'| = \left\lceil l\,(\sqrt{\frac{\mu - l^{-\epsilon}}{\hat{\mu} + l^{-\epsilon}}} - 1)\right\rceil$. Clearly,

$$P(\mathcal{B}) \leq P\left(\min_{V \in \mathbf{Y}[A^c \times B^c]} F(V) < -l^{-\epsilon}\right) \leq l^{\epsilon}\binom{n-l}{m}^{l} \exp\{-l^{-2\epsilon} m^2\}, \tag{5.7}$$

where the minimum is taken over all $m \times m$ submatrices $V \in \mathbf{Y}[A^c \times B^c]$ with $m = |A'| = |B'|$. Note that the assumption $l^{\alpha} \geq \ln n$ implies l satisfies

$$m = \left\lceil l\,(\sqrt{\frac{\mu - l^{-\epsilon}}{\hat{\mu} + l^{-\epsilon}}} - 1)\right\rceil \geq \frac{2\ln n}{l^{-2\epsilon}} \to \infty,$$

as $n \to \infty$. It follows from (2.4) that the right hand side of (5.7) is bounded by n^{-2}.

By (5.4), (5.5), (5.6), and (5.7), we have shown that with probability at least $1 - 3\,n\,\exp\{-l^{2-2\epsilon}\} - n^{-2}$, the average of every column in $\mathbf{Y}\,[A \times B^c]$, the average of every row in $\mathbf{Y}\,[A^c \times B]$, and the average of C^* are greater than $-l^{-\epsilon}$, $-l^{-\epsilon}$ and $\mu - l^{-\epsilon}$ respectively, and $\mathcal{B} \neq \emptyset$. In order to give probability upper bound on the event $\left\{\frac{\mu - l^{-\epsilon}}{\hat{\mu} + l^{-\epsilon}}\,|C^*| > |\hat{C}|\right\}$, one only need to verify that submatrix $D = (A \cup A') \times (B \cup B')$ satisfies

$$\begin{aligned} F(D) &= \frac{F(C^*)|C^*| + F(D \backslash C^*)|D \backslash C^*|}{|D|} \\ &\geq \frac{(\mu - l^{-\epsilon})\cdot l^2 - l^{-\epsilon}\cdot(\frac{\mu - l^{-\epsilon}}{\hat{\mu} + l^{-\epsilon}} - 1)l^2}{\frac{\mu - l^{-\epsilon}}{\hat{\mu} + l^{-\epsilon}}\,l^2} = \hat{\mu} + \frac{\mu + l^{-\epsilon}}{\mu - l^{-\epsilon}}\,l^{-\epsilon} \geq \hat{\mu}. \end{aligned}$$

Now, we have shown that with probability at least $1 - 3\,n\,\exp\{-l^{2-2\epsilon}\} - 2\,n^{-2}$, there

exists a square submatrix $D \in \mathbf{Y}$ which satisfies that $F(D) \geq \hat{\mu}$ and

$$|D| = \frac{\mu - l^{-\epsilon}}{\hat{\mu} + l^{-\epsilon}} |C^*| \leq |\hat{C}| \leq \beta |C^*|. \tag{5.8}$$

Let $\mathcal{A}$ be the event that the above inequality holds. Now, we will bound

$$P\left(\left\{\Gamma(\hat{C}) < \frac{l^2}{l^2 + l^\delta}\right\} \cap \mathcal{A}\right).$$

To begin, consider any integer $k \in I_0 := [\sqrt{\frac{\mu - l^{-\epsilon}}{\hat{\mu} + l^{-\epsilon}}}\, l, \sqrt{\beta}]$. It is clear that any $j \times j$ submatrix V_j in $\mathbf{Y}$ having r rows and k columns in common with C^* satisfies $F(V_j) \sim N\left(\frac{rk\mu}{j^2}, j^{-2}\right)$. Moreover, $j \geq k$ implies that when $\Gamma(V_j) < \frac{l^2}{l^2 + l^\delta}$,

$$\mathrm{E}\left[F(V_j)\right] = \frac{rk\mu}{j^2} = \Gamma(V(m)) \frac{l^2 \mu}{j^2} \leq \frac{l^2 \mu}{l^2 + l^\delta} \cdot \frac{\hat{\mu} + l^{-\epsilon}}{\mu - l^{-\epsilon}}.$$

Thus,

$$\begin{aligned}
\mathcal{A} \cap \left\{\Gamma(\hat{C}) < \frac{l^2}{l^2 + l^\delta}\right\} &\subset \left\{\max_{j \in I_0} \max_{V_j} \left[F(V_j) - \mathrm{E}\left[F(V_j)\right]\right] \geq \hat{\mu} - \frac{rk\mu}{j^2}\right\} \\
&\subset \left\{\max_{j \in I_0} \max_{V_j} \left[F(V_j) - \mathrm{E}\left[F(V_j)\right]\right] \geq \hat{\mu}\left(1 - \frac{l^2}{l^2 + l^\delta} \cdot \frac{\mu + \frac{\mu}{\hat{\mu}} l^{-\epsilon}}{\mu - l^{-\epsilon}}\right)\right\}.
\end{aligned}$$

It is easy to check that when l is sufficiently large and $\epsilon < 2 - \delta$ (by definition),

$$\frac{l^2}{l^2 + l^\delta} \cdot \frac{\mu + \frac{\mu}{\hat{\mu}} l^{-\epsilon}}{\mu - l^{-\epsilon}} \leq 1.$$

Thus, probability

$$\begin{aligned}
& P\left(\Gamma(\hat{C}) < \frac{l^2}{l^2 + l^\delta} \cap \mathcal{A}\right) \\
\leq\ & \sum_{j \in I_0} P\left(\max_{\text{all } j \times j\ V_j} [F(V_j) - \mathrm{E}[F(V_j)]] \geq \hat{\mu}\left(1 - \frac{l^2}{l^2 + l^\delta} \cdot \frac{\mu + \frac{\mu}{\hat{\mu}} l^{-\epsilon}}{\mu - l^{-\epsilon}}\right)\right) \\
\leq\ & \sqrt{\beta}\, l \cdot \max_{j \in I_0} \left[\exp\left\{-j^2 \cdot \hat{\mu}^2 \left(1 - \frac{l^2}{l^2 + l^\delta} \cdot \frac{\mu + \frac{\mu}{\hat{\mu}} l^{-\epsilon}}{\mu - l^{-\epsilon}}\right)^2\right\}\right.
\end{aligned}$$

$$
\begin{aligned}
& \times \left(\sum_{r=1}^{l} \sum_{k=1}^{l} \binom{n-l}{j-r} \cdot \binom{l}{r} \cdot \binom{n-l}{j-k} \cdot \binom{l}{k} \right) \Bigg] \\
\leq \quad & \sqrt{\beta}\, l \cdot \max_{j \in I_0} \left[\exp\left\{ -j^2 \cdot \hat{\mu}^2 \left(1 - \frac{l^2}{l^2 + l^\delta} \cdot \frac{\mu + \frac{\mu}{\hat{\mu}} l^{-\epsilon}}{\mu - l^{-\epsilon}} \right)^2 \right\} \times \binom{n}{j}^2 \right] \\
\leq \quad & \sqrt{\beta}\, l n^{-3} \qquad (5.9) \\
\leq \quad & n^{-2}.
\end{aligned}
$$

where the inequality (5.9) comes from the fact that the assumption $l^\alpha \geq \ln n$ is a sufficient condition to ensure that every $j \in I_0$ satisfies that

$$
j \geq 2 \ln n \cdot (1 - \frac{l^2}{l^2 + l^\delta} \cdot \frac{\mu + \frac{\mu}{\hat{\mu}} l^{-\epsilon}}{\mu - l^{-\epsilon}})^{-2} \geq 2\, l^{4-2\delta} \ln n,
$$

the right hand side of which goes to infinity as $n \to \infty$. Consequently,

$$
P\left(\Gamma(\hat{C}) < \frac{l^2}{l^2 + l^\delta} \right) \leq P\left(\Gamma(\hat{C}) < \frac{l^2}{l^2 + l^\delta} \cap \mathcal{A} \right) + P(\mathcal{A}^c) \leq 3\, n \, \exp\{-l^{2-2\epsilon}\} + 3\, n^{-2}.
$$

CHAPTER 6

Conclusion and Future Work

6.1 Conclusion

In this dissertation, we have studied some statistical problems related to biclustering algorithms. Biclustering algorithms are a type of data mining techniques used in bioinformatics, drug activity analysis and market basket analysis. Our goal is to provide a rigorous statistical theory to guide the application of biclustering techniques.

In Chapter 2, we focused on frequent itemset mining. Frequent itemset mining has an equivalent matrix form, where frequent itemsets correspond to maximal submatrices of 1's in binary matrices. The objective of the research is to evaluate the statistical significance of the identified submatrices of 1's. For this purpose, an i.i.d. Bernoulli random matrix model was assumed. By extending existing results on clique number by Bollobás and Erdős in random graph theory (11; 10), we established a probability upper bound on the existence of large-sized submatrices of 1's in the Bernoulli random matrix model. Further, we showed that as the size of the data matrix goes to infinity, eventually almost surely, the size of the largest submatrix of 1's (with a fixed row/column aspect ratio) only takes values in a set of five consecutive integers, whose values only depend on the size of the data matrix and the Bernoulli distribution parameter. An upper bound and a lower bound on the sizes of the smallest square maximal submatrices of 1's in square Bernoulli random matrices are also given in Chapter 2.

In Chapter 3, the noise sensitivity of standard frequent itemset mining was studied. It was shown that standard frequent itemset mining is very sensitive to noise and it can not be directly used to recover the block structures when the data is contaminated with ran-

dom noise. Then, we considered the statistical properties of error-tolerant frequent itemset mining, which had been proposed as a generalization of standard frequent itemset mining. We first showed how to evaluate the statistical significance of submatrices identified by a general class of error-tolerant frequent itemset mining algorithms. Then, we showed that approximate frequent itemset mining (41), a particular error-tolerant frequent itemset mining algorithm, can asymptotically recover the underlying block structure in simple recovery problems, where standard frequent itemset mining fails.

In Chapter 4, we considered the biclustering algorithms for real-valued matrices. We established results parallel to those of frequent itemset mining in Chapter 2 for biclusters with high averages under i.i.d. Gaussian random matrix assumption. In Chapter 4, we also studied the statistical properties of biclusters with low ANOVA residues, where the biclustering methods based on ANOVA type criterion are introduced and studied in Cheng and Church (15), and in Lazzeroni and Owen (40). In order to evaluate the significance of submatrices with low ANOVA residuals, several probability bounds on the size of identified submatrices under appropriate random matrix assumptions were given in Chapter 4.

In Chapter 5, we showed that in a simple recovery problem with Gaussian noise, there exists a procedure that is able to recover the underlying single block structure with high probability.

6.2 Future Work

Many of the results described here are based on an i.i.d. random matrix model. However, certain dependence structures are known to exist in particular applications. The statistical significance problem for frequent itemset mining under a simple Markov chain type of dependence for binary random matrix was considered in Chapter 2. Future work includes the study of biclustering methods under other dependence structures.

The recovery problems addressed in Chapter 3 and Chapter 5 consider recovery of a single square block from a square matrix. The results there could be further extended to the case of non-square blocks and non-square matrices under certain row/column aspect ratio restrictions. A further step is to consider multiple blocks in the underlying pattern matrix.

BIBLIOGRAPHY

[1] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan (1998). Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. *Proc. ACM SIGMOD*, 94-105.

[2] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. Verkamo(1996). Fast discovery of association rules. In U. M. Fayyad, U. and et. al, editors, *Advances in Knowledge Discove and Data Mining*, chapter 12, 307-328. AAAI Press.

[3] R. Agrawal, T. Imielinski, and A. Swami (1993). Mining association rules between sets of items in large databases. *Proc. ACM SIGMOD*, 207-216.

[4] N. Alon and A. Naor (2004). Approximating the cut-norm via Grothendieck's inequality. *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, 72 - 80.

[5] E. Arias-Castro, D. L. Donoho and X. Huo (2005). Near-optimal detection of geometric objects by fast multiscale methods. *IEEE Trans. Information Theory*.

[6] E. Arias-Castro, D. L. Donoho, and X. Huo (2006). Adaptive multiscale detection of filamentary structures embedded in a background of Uniform random points. *Annals of Statistics*, 34, 326-349.

[7] R. Arratia, L. Goldstein, and L. Gordon (1990). Poisson approximation and the chen-stein method. *Statistical Science*, 5, 403–434.

[8] A. Ben-Dor, B. Chor, R. Karp, and Z. Yakhini (2003). Discovering local structure in gene expression data: The order-preserving submatrix problem. *Journal of Computational Biology*, 10 (3-4), 373-384.

[9] S. Bergmann, J. Ihmels,and N. Barkai (2003). Iterative signature algorithm for the analysis of large-scale gene expression data. *Physical Review* E 67, 031902.

[10] B. Bollobás (2001). *Random Graphs*, 2nd ed., Cambridge University Press.

[11] B. Bollobás, P. Erdős (1976). Cliques in Random Graphs. *Mathematical Proceedings of the Cambridge Philosophical Society*, 80, 419-427.

[12] T. Brijs, G. Swinnen, K. Vanhoof and G. Wets (1999). Using Association Rules for Product Assortment Decisions: A Case Study. *Knowledge Discovery and Data Mining*, 254-260.

[13] S. Busygin, G. Jacobsen, and E. Kramer (2002). Double conjugated clustering applied o leukemia microarray data. *Proceedings of the 2nd SIAM International Conference on Data Mining*, Workshop on Clustering High Dimensional Data.

[14] A. Califano, G. Stolovitzky, and Y. Tu (2000). Analysis of gene expression microarays for phenotype classification. *Proceedings of the International Conference on Computacional Molecular Biology*, 75-85.

[15] Y. Cheng and G. M. Church (2000). Biclustering of expression data. *Proc. 8th Int'l Conference on Intelligent Systems for Molecular Biology*, 93-103.

[16] M. Dawande, P. Keskinocak, J. Swaminathan, and S. Tayur (2001). On bipartite and multipartite clique problems. *J. Algorithms*, 41, 388-403.

[17] L. Devroye, L. Gyorfi, G, Lugosi (1996). *A Probabilistic Theory of Pattern Recognition.* Springer, New York.

[18] I. S. Dhillon (2001). Co-clustering documents and words using bipartite spectral graph partitioning. *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD01)*, 269-274.

[19] S. Dudoit and J. Fridlyand (2003). Bagging to improve the accuracy of a clustering procedure. *Bioinformatics*, 19(9), 1090-1099.

[20] S. Dudoit, J. P. Shaffer, and J. C. Boldrick (2003). Multiple Hypothesis Testing in Microarray Experiments. *Statistical Science*, 18, 71.

[21] MB. Eisen, PT. Spellman , PO. Brown and D. Botstein (1998). Cluster Analysis and Display of Genome-Wide Expression Patterns. *Proc Natl Acad Sci U S A* 95, 14863-8.

[22] C. M. Fiduccia and R. M. Mattheyses (1982). A linear time heuristic for improving network partitions. Technical Report 82CRD130, GE Corporate Research.

[23] J. H. Friedman1 and J. J. Meulman (2004). Clustering objects on subsets of attributes. *Journal of the Royal Statistical Society Series B*, 66, 815.

[24] M. R. Garey and D. S. Johnson (1979). *Computers and Intractability, A guide to the theory of NP-completeness.* Freeman, San Francisco.

[25] F. Gebhardt. Survey on cluster tests for spatial area data. *http://home.vr-web.de/friedrich.gebhardt/Survey.pdf*

[26] G. Getz, E. Levine, and E. Domany (2000). Coupled two-way clustering analysis of gene microarray data. *Proceedings of the Natural Academy of Sciences USA*, 97, 12079-12084.

[27] B. Goethals. Survey on Frequent Pattern Mining. *h t t p: / / w w w.a d r e m.u a.a c.b e / g o e t h a l s / s o f t w a r e / s u r v e y . p d f*

[28] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, and C. D. Bloomfield (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*. 286(5439), 531-537.

[29] J. Han, J. Pei, and Y. Yin (2000). Mining Frequent Patterns without Candidate Generation. *Proc. ACM SIGMOD*, 1-12.

[30] J. A. Hartigan (1972). Direct clustering of a data matrix. *Journal of the American Statistical Association*, 67(337), 123-129.

[31] E. G. Hill, L. Ding and L. A. Waller (2000). A comparison of three tests to detect general clustering of a rare disease in Santa Clara County, California. *Statistics in Medicine*, 19(10), 1363 - 1378.

[32] D. S. Hochbaum (1998). Approximating clique and biclique problems. *Journal of Algorithms*, 29(1), 174-200.

[33] R. Karp (1988). *Probabilistic Analysis of Algorithms*. Class Notes, UC-Berkeley.

[34] L. Kaufman and P. J. Rousseeum (1990). *Finding Groups in Data*. John Wiley & Sons, Inc., New York, NY.

[35] B. Kernighan and S. Lin (1970). An efficient heuristic procedure for partitioning graphs. *The Bell System Technical Journal*, 29(2), 291-307.

[36] Y. Kluger, R. Basri, J. T. Chang and M. Gerstein (2003). Spectral Biclustering of Microarray Data: Coclustering Genes and Conditions. *Genome Res.* 13, 703-716.

[37] M. Koyutürk, W. Szpankowski and A. Grama (2004). Biclustering Gene-Feature Matrices for Statistically Significant Dense Patterns. *Proceedings of the 8th Annual International Conference on Research in Computational Molecular Biology*, 480-484.

[38] M. Kuramochi and G. Karypis (2001). Frequent Subgraph Discovery. *Proc. 2001 IEEE Int'l Conf. Data Mining (ICDM)*, 313-320.

[39] T. Lange, V. Roth, M. Braun, and J. Buhmann (2004). Stability-Based Validation of Clustering Solution. *Neural Computation*, 16(6), 1299-1323.

[40] L. Lazzeroni and A. Owen (2002). Plaid models for gene expression data. *Statistica Sinica*, 12, 61-86.

[41] J. Liu, S. Paulsen, X. Sun, W. Wang, A.B. Nobel, and J. Prins (2006). Mining approximate frequent itemsets in the presence of noise: algorithm and analysis. To appear in *Proceedings of SDM*.

[42] J. Liu, S. Paulsen, W. Wang, A. B. Nobel, and J. Prins (2005). Mining Approximate Frequent Itemsets from Noisy Data. *Proceedings of ICDM'05*, 721-724.

[43] J. Liu and W. Wang (2003). Op-cluster: Clustering by tendency in high dimensional space. *Proceedings of the 3rd IEEE International Conference on Data Mining*, 187-194.

[44] S. Madeira and A. Oliveira (2004). Biclustering Algorithms for Biological Data Analysis: A Survey. *IEEE Transactions on Computational Biology and Bioinformatics*, 1(1), 24-45.

[45] D. Matula (1976). The largest clique size in a random graph. Southern Methodist University, Tech. Report, CS 7608..

[46] G. W. Milligan (1983). Characteristics of four external criterion measures. In J. Felsenstein, (Ed.), *Proceedings of the 1982 NATO Advanced Studies Institute on Numerical Taxonomy*. New York: Springer-Verlag.

[47] N. Mishra, D. Ron and R. Swaminathan (2004). A New Conceptual Clustering Framework. *Machine Learning*. 56(1-3), 115-151.

[48] J. P. Novak, R. Sladek and T. J. Hudson (2002). *Genomics*, 79, 104-113.

[49] M. Okamoto (1958). Some inequalities relating to the partial sum of binomial probabilities. *Annals of the Institute of Statistical Mathematics*, 10, 29 - 35.

[50] G. Park and W. Szpankowski (2005). Analysis of Biclusters with Applications to gene Expression Data. *Conference on Analysis of Algorithms*, 267-274, Barcelona.

[51] R. Peeters (2003). The maximum edge biclique problem is np-complete. *Discrete Applied Mathematics*, 131(3), 651-654.

[52] J. Pei, G. Dong, W. Zou, J. Han (2002). Mining Condensed Frequent-Pattern Bases. *Knowledge and Information Systems*, 6(5).

[53] J. Pei, A. K. Tung, and J. Han (2001). Fault-tolerant frequent pattern mining: Problems and challenges. *Proceedings of DMKD'01*.

[54] C. M. Perou, T. Srlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, . Fluge, A. Pergamenschikov, C. Williams, S. X. Zhu, P. E. Lnning, A.-L. Brresen-Dale, Patrick O. Brown, and David Botstein (2000). Molecular Portraits of Human Breast Tumors. *Nature*, 406, 747-52.

[55] J. D. Reuning-Scherer (1997). Mixture Models for Block Clustering. Phd Thesis, Yale university.

[56] E. Segal, B. Taskar, A. Gasch, N. Friedman, and D. Koller (2001). Rich probabilistic models for gene expression. *Bioinformatics*, 17 (Suppl. 1), 243-252.

[57] E. Segal, B. Taskar, A. Gasch, N. Friedman, and D. Koller (2003). Decomposing gene expression into cellular processes. *Proceedings of the Pacific Symposium on Biocomputing*, 8, 89-100.

[58] J. K. Seppänen, and H. Mannila (2004). Dense Itemsets. *Proceedings of ACM SIGKDD'04*, 683-688.

[59] Q. Sheng, Y. Moreau, and B. D. Moor (2003). Biclustering micrarray data by gibbs sampling. *Bioinformatics*, 19 (Suppl. 2), 196- 205.

[60] D. Slepian(1962). The one-sided barrier problem for Gaussian noise. *Bell System Technical Journal*, 41, 463-501.

[61] M. Steinbach, P. Tan, V. Kumar (2004). Support envelopes: a technique for exploring the structure of association patterns. SIGKDD.

[62] A. Tanay, R. Sharan, R. Shamir (2002). Dicovering statistically significant biclusters in gene expression data. *Bioinformatics*, 18 (Suppl. 1), 136-144.

[63] A. Tanay, R. Sharan and R. Shamir (2005). Biclustering Algorithms: A Survey. *Handbook of Computational Molecular Biology*, Chapman & Hall/CRC, Computer and Information Science Series. In press.

[64] C. Tang, L. Zhang, I. Zhang, and M. Ramanathan (2001). Interrelated two-way clustering: an unsupervised approach for gene expression data analysis. *Proceedings of the 2nd IEEE International Symposium on Bioinformatics and Bioengineering*, 41-48.

[65] R. Tibshirani, G. Walther and T. Hastie. Estimating the number of clusters in a dataset via gap statistic. Technical Report 208, Dept of Statistics, Stanford University.

[66] Y. Tu , G. Stolovitzky , and U. Klein (2002). Quantitative noise analysis for gene expression microarray experiments. *PNAS*, 99:22, 14031-14036.

[67] C. Yang, U. Fayyad, P. S. Bradley (2001).Efficient discovery of error-tolerant frequent itemsets in high dimensions. *Conference on Knowledge discovery and data mining.*

[68] J. Yang, W. Wang, H. Wang, and P. Yu (2002). δ-clusters: Capturing subspace correlation in a large data set. *Proceedings of the 18th IEEE International Conference on Data Engineering*, 517-28.

[69] J. Yang, W. Wang, H. Wang, and P. Yu (2003). Enhanced biclustering on expression data. *Proceedings of the 3rd IEEE Conference on Bioinformatics and Bioengineering*, 321-327.

[70] K. Y. Yeung, D. R. Haynor, and W. L. Ruzzo (2000). Validating clustering for gene expression data. *Bioinformatics*, 17(4), 309-318.

[71] H. Zhou and D. Woodruff (2004). Clustering via matrix powering. *Proceedings of the twenty-third POD*, 136 - 142.