國立交通大學

應用數學系

碩士論文

群試檢驗演算法的相關數學模型探討

Mathematical Models Related to Group Testing Algorithms

研究生:曾如汶　　Student:Ju-Wen Tseng
指導教授:翁志文　　Advisor:Chih-Wen Weng

中華民國一百零一年 六月

群試檢驗演算法的相關數學模型探討

Mathematical Models Related to Group Testing Algorithms

研究生:曾如汶　　Student:Ju-Wen Tseng

指導教授:翁志文　　Advisor:Chih-Wen Weng

碩士論文

A Thesis Submitted to Department of Applied Mathematics Collage of Science National Chiao Tung University In Partial Fulfillment of the Requirements For the Degree of Master In Applied Mathematics

Hsinchu, 30056, Taiwan

June 2011

中華民國一百零一年 六月

群試檢驗演算法的相關數學模型探討

學生: 曾如汶　　指導教授: 翁志文

國立交通大學

應用數學系

台灣,新竹

摘要

考慮一包含n個待測物，且最多有d個呈陽性的集合。我們的目的是藉由群試設計的概念找出所有呈陽性的待測物。一個群試設計含有多個測試，每個測試都包含兩個以上的待測物。我們探討群試設計的目的是去減少測試的個數和階段數。而在同一個階段裡的測試可同時執行。

我們修改並且分析了一個適用於在已知最多二個陽性物的情況下的二階段群試檢驗演算法。此演算法的測試次數是$O(3\log n)$。

Mathematical Models Related to Group Testing Algorithms

Student: Ju-Wen Tseng Advisor: Chih-Wen Weng

Department of Applied Mathematics

National Chiao Tung University

Hsinchu, Taiwan 30050

Abstract

Consider a set of n items which has at most d positive items. Our aim is to find all positive items by using the concept of group testing. A group testing consists of a few tests, each of them containing more than one item. The objective in the study of group testing is to reduce the number of test times and to reduce the number of stages which partition the tests into different time slots.

We modify and analysis a group testing algorithm, which has 2-stage for the case d=2 and the test number of this algorithm is $O(3 \log n)$.

致謝

碩士班的兩年雖然短暫，但卻豐富，有著許多的改變，許多不同的滋味。 謝謝交大的每一位師長們對我們的關懷，除了課業，也很關心我們的生活。謝謝翁志文老師在論文上面的建議、引導。老師您面對數學的耐心和對數學的要求也是我很好的榜樣。謝謝秋媛老師，您是一位處處爲學生著想的老師。常常在二樓看見秋媛老師您忙碌步伐的身影，但再怎麼忙碌，您對我們的關心卻從來都沒有減少。感謝黃大原老師，在我修課期間，給我課業上的鼓勵，讓我對數學比較有自信一點。

謝謝李光祥學長在論文上面給予我的協助、家安學長在數學上的教導。

謝謝爸爸媽媽一直以來對我的關心，給予我的自由。謝謝您們這兩年來開始願意給我很大的空間讓我自己來選擇，甚至願意放下您們自己原本的意見，和我溝通。您們的改變可能比我更多。

謝謝主耶穌，在這兩年當中，您是那位最清楚且參與我最多生活點滴的，雖然很多時候只能憑信心，但您眞的是那位信實的神。碩一暑假在您奇妙的引導下，不知不覺實現了小時候的夢想:我的房間眞的有一台琴了，並且再度開始學琴。

最後感謝陪伴我的朋友們，謝謝你們與我一同歡樂也一同分擔我的憂慮，讓我不至於一個人承擔。

Contents

Chapter 1

Introduction

1.1 The history of group testing

We first give the brief history of group testing and most of them is referred to [1].

Unlike many other mathematical fields which can track back to earlier centuries, group testing has developed only for about 70 years.

The idea of group testing origins from recent event World War II. We usually give credit for a single person-Robert Dorfman. During World War II, some economists is exhausted by examining blood samples from millions of draftees. Someone suggested that it is economical to pool the blood samples. We quote some paragraphs from Robert Dorfman's recollection in the following:

> "The drabness of life in those wings was relieves by occasional bull sessions. Group testing was first conceived in one of them, in which David Rosenblatt and I participated. Being economists, we were all struck by the wastefulness of subjecting blood samples from millions of draftees to identical analysis in order to detect a few thousand cases of syphilis. Someone suggest that it might be economical to pool the blood samples, and the idea was batted back and forth. There was lively give-and-take and some persiflage. I don't recall how explicitly the problem was formulated there. What is clear is that I took the idea seriously enough so that in the next few days I formulated the underlying probability problem and worked

though the algebra (which is pretty elementary). Shortly after, I wrote it up, presented it at a meeting of the Washington Statistical Association, and submitted the four-page note that was published in the Annals of Mathematical Statistics. By the time the note was published, Rosenblatt and I were both overseas and out of contact."

Robert Dorfman also applied group testing to examine syphilis, which is intended to be used by the United States Public Health Service and the Selective System to weep out all syphilitic men called up for induction. Although this group testing method for syphilis screening was not actually put to use, Dorfman's clear account of applying group testing to screen syphilitic individuals may have new impact to the medical world and the health service sector.

With the end of World War II and the release of millions of millions of inductees, the practical need of group testing disappeared, so the research related to group testing got fewer.

Two Bell Laboratories scientists, Sobel and Groll, again motivated by practical need, applied group testing on industrial sector, and they established many new grounds for future studies about group testing in their 74-page paper.

One of industrial application they apply is that testing condensers and resistors. This idea can be explained clearly by the Christmas tree lighting problem. A batch of light blubs is electrically arranged in series. If the lights are on, then whole tested subset of bulbs must be good; if the lights are off, then at least one bulb in the subset is defective.

Notice that Dorfman, as well as Sobel and Groll, studied group testing under probabilistic models. Katonia is the first people mentioned the combinatorial aspects of group testing. He give a more restrictive viewpoint on combinatorial group testing (CGT) is taken by completely deleting probability distributions on defectives. The assumption on the defective set is that it must be a member, called a sample point, of a given family called a sample space. For instance, the sample space can consist of all d-subsets of the n items, when the assumption is that there are exactly d defectives among the n items.

Recently, CGT has studied in many fields like complexity theory, graph theory, learning models, communication channels and fault tolerant computing.

1.2 Preliminaries

In this section, we will introduce the concept of group testing, then give the definition of d-pooling design which is the main theme of this paper.

Consider a set of n items, denoted by $1, 2, \ldots, n$ such that each of them can be either positive or negative. The concept of **group testing** is based on the following assumption:

The group testing assumption: Given any subset S of n items, if S has at least 1 positive item, then the group testing outcome of S is positive; otherwise (i.e. items in S are all negative), the group testing outcome of S is negative.

A subset S of $[n]$ will be called a **group test** or a **test** for short. A **group testing algorithm** is an organization of group tests such that from the outputs of these tests, one can identify which items are positive.

Definition 1.2.1. A **d-pooling design** is a group testing algorithm that can identify all positive items among items which have at most d positive items.

Generally, we can divide group testing algorithm into three types:

- sequential algorithm: The tests are conducted one by one, and the outcome of the previous tests are assumed known at the time of determining the current test.
- nonadaptive algorithm: All tests are conducted simultaneously.
- multistage algorithm: Tests are divided into several stages where the stages are considered sequential but all tests in the same stage are treated as nonadaptive.

1.3 Outline of each chapter

In this paper, we will first provide a 2-stage 2-pooling design in Chapter 2,and in this chapter we also give analysis of the test times of this 2-stage 2-pooling design. We will revisit this 2-stage 2-pooling design in after chapter.

In Chapter 3 , we will talk about d-separable and d-disjunct matrices and apply them to the 2-stage 2-pooling design in Chapter 2.

In Chapter 4 , we introduce $(d, s]$-disjunct matrices which can be seen as the generalization of d-disjunct matrix and discuss the lower bound of the number of rows of $(d, s]$-disjunct matrix.

Chapter 2

A 2-stage 2-pooling design

By modifying an idea in [4], we shall give a 2-stage 2-pooling design in this chapter. For convenience, we assume $n = c^k$, and represent each item as k-tuple $(c_1, c_2, \ldots, c_k)$, where $c_i \in \{1, 2, \ldots, c\}$. We assume that there are at most 2 positive items among these c^k items.

2.1 Algorithm

The algorithm for 2-stage 2-pooling design is described as the following:

Stage 1: Simultaneously apply group testing on each of the following ck subsets of items:

$$S_i(j) = \{(c_1, \ldots, c_{i-1}, j, c_{i+1}, \ldots, c_k) \mid c_\ell \in \{1, 2, \ldots, c\}\},$$

where $1 \le j \le c$, $1 \le i \le k$.

Analysis: Since there are at most 2 positive items, according to 0, 1 or 2 positive items in the beginning, there are the following (i)-(iii) cases for the outcomes of stage 1.

(i) For all $1 \le j \le c$, $1 \le i \le k$, the test on $S_i(j)$ is negative: This implies that there is no any positive item.

(ii) For all i, there exists a unique u_i such that the test on $S_i(u_i)$ is positive: This implies that the k-tuple $u = (u_1, u_2, \ldots, u_k)$ is the unique positive item.

(iii) There are p positions $d_1, d_2, \ldots, d_p \in \{1, 2, \ldots, k\}$ such that for each position $i \in \{d_1, d_2, \ldots, d_p\}$ among the c tests on $S_i(j)$ for $1 \le j \le c$, there are exactly two positive tests, say on $S_i(n_i)$ and on $S_i(m_i)$, and for each j of the remaining $(k-p)$ positions there is a unique positive test, say $S_j(n_j)$: This means that there are exactly two positive items u, v such that $u_i = v_i = n_i$ if $i \in [k] - \{d_1, d_2, \ldots, d_p\}$, and $u_j \neq v_j$ and $\{u_i, v_i\} = \{n_i, m_i\}$ for $i \in \{d_1, d_2, \ldots, d_p\}$.

We have to apply stage 2 if the case (iii) happens, otherwise we stop. Let $\{d_1, d_2, \ldots, d_p\} \subseteq \{1, 2, \ldots, k\}$ and $n_j, m_j \in \{1, 2, \ldots, c\}$ for $j \in \{d_1, d_2, \ldots, d_p\}$ be described in the case (iii) above.

Stage 2 : Do the $p-1$ group tests on $S_{d_1}(n_{d_1}) \cap S_i(n_i)$ for $i \in \{d_2, \ldots, d_p\}$.

Analysis: Let $D \subseteq \{d_2, \ldots, d_p\}$ such that the positive outputs in Stages 2 are $S_{d_1}(n_{d_1}) \cap S_i(n_i)$ for $i \in D$. Since the two positive items u, v take different values n_{d_1}, m_{d_1} on the coordinate d_1, we may assume $u_{d_1} = n_{d_1}$ and $v_{d_1} = m_{d_1}$. Then after Stage 2, one can identify the positive items u and v from the following descriptions:

(a) $u_i = v_i = n_i$, if $i \in [k] - \{d_1, d_2, \ldots, d_p\}$;

(b) $u_i = n_i$ and $v_i = m_i$, if $i \in D$;

(c) $u_i = m_i$ and $v_i = n_i$, if $i \in \{d_2, \ldots, d_p\} \setminus D$.

(a) is clear from the output of Stage 1. We shall prove (b)-(c). Since we assume $u_{d_1} = n_{d_1}$, the test on $S_{d_1}(n_{d_1}) \cap S_i(n_i)$ is positive iff $u_i = n_i$ (and hence $v_i = m_i$) for $i \in \{d_2, \ldots, d_p\}$. Then (b), (c) follow.

2.2 Analysis of the number of tests

In this section, we shall investigate the number of tests of the algorithm described in section 2.1. The number of tests at the first stage is apparently $c \cdot k$. The number of tests at the second stage is $p-1$, so the worse case is $k-1$. Hence the worst case of the number of tests of this algorithm of 2-stage 2-pooling design is $ck + (k-1) = (c+1)k - 1$.

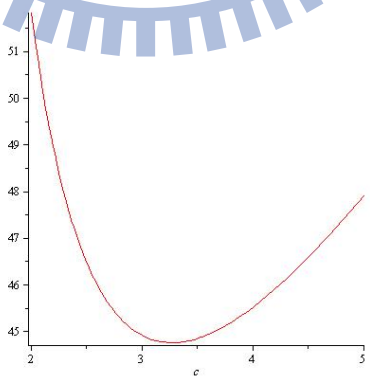
If n is given in general situation, we choose a constant number c and an integer k such that $c^{k-1} < n \leq c^k$. By adding more negative items if necessary, we can assume that there are c^k items and apply the 2-stage 2-pooling design. Then the lower bound and upper bound of the number t of tests in the expression of functions of n is

$$-1 + (c+1)\log_c n \leq t = (c+1)k - 1 < c + (c+1)\log_c n.$$

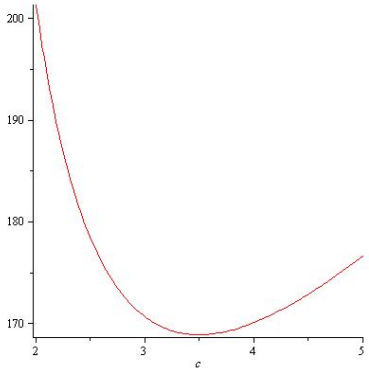
Note that if $c = 2$ then

$$-1 + 3\log_2 n \leq t < 2 + 3\log_2 n.$$

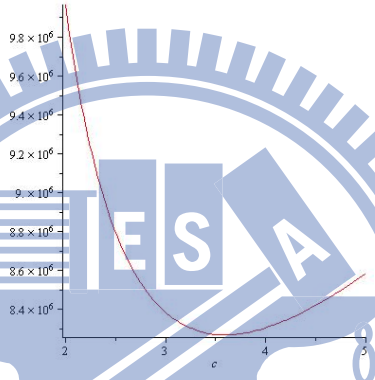
We shall indicate by three graphs that the values $c \in \{3, 4\}$ take smaller upper bounds for t.



The graph of $t = c + (c+1)\log_c n$ in ct-plane for $n = 10^5$ and $2 \leq c \leq 5$.



The graph of $t = c + (c+1)\log_c n$ in ct-plane for $n = 10^{20}$ and $2 \le c \le 5$.



The graph of $t = c + (c+1)\log_c n$ in ct-plane for $n = 10^{10^6}$ and $2 \le c \le 5$.

2.3 The minimum number of tests for 2-pooling design

Assume there are n items and among these items there are at most 2 positive items. Let $t(n, \overline{2})$ denote the minimum number of tests to identify all the positive items between n items in the worse case. We shall compare our 2-stage 2-pooling design with $t(n, \overline{2})$. Before that, we investigate the possible range of $t(n, \overline{2})$ first. The following lemma is about a lower bound of $t(n, \overline{2})$.

Lemma 2.3.1.

$$\frac{n^2 + n + 2}{2} \le 2^{t(n,\overline{2})}.$$

Proof. Given an algorithm of group testing, we can write down the outcomes of all tests as a

binary vector v of length $t(n,\overline{2})$ such that $v_i = 1$ iff the i-th test has been proceeded and has a positive outcome. If the algorithm works, then the $2^{t(n,\overline{2})}$ possible binary vectors v must distinguish those $\binom{n}{2} + \binom{n}{1} + \binom{n}{0} = (n^2+n+2)/2$ possible situations. Hence $(n^2+n+2)/2 \leq 2^{t(n,\overline{2})}$. □

The following lemma is about an upper bound of $t(n,\overline{2})$.

Lemma 2.3.2.

$$t(n,\overline{2}) \leq \lceil 2\log_2 n \rceil.$$

Proof. If we use the divide and conquer strategy, it takes at most $\lceil \log n \rceil$ steps to find 1 positive item. Since there are 2 positive items, it takes at most $\lceil 2\log n \rceil$ steps. □

From the above two lemmas, we have the following corollary.

Corollary 2.3.1. $-1 + \lceil 2\log_2 n \rceil \leq t(n,\overline{2}) \leq \lceil 2\log_2 n \rceil$. □

Proof. This follows from the above two lemmas and

$$-1 + \log_2(n^2+n+2) > -1 + 2\log_2 n.$$

□

From the previous Corollary, there are two possible values of $t(n,\overline{2})$, $t(n,\overline{2}) = -1 + \lceil 2\log_2 n \rceil$, or $t(n,\overline{2}) = \lceil 2\log_2 n \rceil$.

Comparing the number $t(n,\overline{2})$ with the test number t of our 2-stage 2-pooling design, our test number t is about $3t(n,\overline{2})/2$. Since our design has only 2 stages, it is suitable in biological experiments, which usually need long time to wait for a test result.

Chapter 3

Matrices related to pooling designs

Given a nonadaptive pooling design, one can construct a binary matrix $M = (m_{ij})$ whose rows are the t tests and columns are the n items such that

$$m_{ij} = 1 \text{ iff the } j\text{-th item is contained in the } i\text{-th test.}$$

On the other hand, given a binary matrix M, the above line also gives a nonadaptive group testing algorithm, possibly failing to work.

In this chapter, we will review d-separable matrix and d-disjunct matrix, which have good properties to ensure that their corresponding group testing algorithm works properly. We construct the matrices corresponding to the 2-stage 2-pooling designs given in section 2.1, and check how far for them to be d-separable or d-disjunct properties.

3.1 d-separable matrix and $\overline{d}$-separable matrix

We first give the definition of d-separable matrix and $\overline{d}$-separable matrix.

Definition 3.1.1. A binary matrix M is called **d-separable** if $\cup D \neq \cup D'$ for any two distinct d-sets D , D' of columns.

Definition 3.1.2. A binary matrix M is called **$\overline{d}$-separable** if $\cup D \neq \cup D'$ for any two distinct sets D , D' of columns of M with $|D|$, $|D'| \leq d$.

From the above two definitions, we have the following remark.

Remark 3.1.1. A $\overline{d}$-separable matrix is also a d-separable matrix.

Example 3.1. Consider the following two matrices

$$M = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix}, \qquad M' = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}.$$

One can easily check that M is a 2-separable matrix but not a $\overline{2}$-separable matrix, and M' is a $\overline{2}$-separable matrix. This example tells us that a $\overline{d}$-separable matrix can not contain an all zero columns.

In the following, we show that if there are exactly d unknown positive items we can find all these d positive items by using the outcome and a given d-separable matrix M.

Represent a nonadaptive group testing by using a $t \times n$ d-separable matrix M. There are t tests. The i-th test contains those items $j \in \{1, 2, \ldots, n\}$ with $m_{ij} = 1$ for $1 \leq i \leq t$.

Then the t outcomes can also be represented by a t-vector $v = (v_1, ..., v_t)^t$, where $v_i = 1$ iff the outcome of test i is positive; $v_i = 0$ otherwise. Let D be the set of positive items. Then v is the union of those columns indexed by D.

One of the methods is by comparing v with the union of any d columns of M to find the set D of positive items. The d-separable assumption on M ensures that D is unique, so this method works. However, in the worst case this method needs to $\binom{n}{d}$ times of comparison to find D. We give a method presented in [1] to reduce the number of comparisons.

Let M_D be a $t_D \times n_D$ matrix obtained from M by keeping only the rows with positive outcome and columns which represent items not appearing in any negative outcome. Let $T'_D := \{i \mid v_i = 1\}$ be the set of these t_D tests, and N_D be the set of these n_D items.

M :

		$\overbrace{\quad}^{N_D}$		v
T'_D {		M_D		1 ⋮ 1
	some 1	all 0	some 1	0 ⋮ 0

An illustration of the matrix M and its submatrix M_D.

Let T_D be the collection of subsets of items which are in each test in $T'(D)$. Then

$$T_D = \{\{j \mid m_{ij} = 1\} \mid v_i = 1\},$$

$$N_D = \bigcap_{i:v_i=0} \{j \mid m_{ij} = 0\}.$$

Lemma 3.1.1. $D \subseteq N_D$.

Proof. Let $j \in D$. Then $v_i = 1$ if $m_{ij} = 1$. Hence $j \notin \bigcup_{i:v_i=0} \{j \mid m_{ij} = 1\}$. Thus $j \in N_D$. □

Lemma 3.1.2. $D \cap T \neq \emptyset$ for all $T \in T_D$.

Proof. Fix $T \in T_D$. Then $T = \{j \mid m_{ij} = 1\}$ for some i with $v_i = 1$. Since $v_i = 1$, there exists $k \in D$ with $m_{ik} = 1$. Clearly $k \in D \cap T$. □

From the above two lemmas, the following hitting set problem is related to identifying the unknown subset D of items.

Reduced hitting set problem: Find a minimum-cardinality subset Y of N_D such that $Y \cap T \neq \emptyset$, for all $T \in T_D$.

Proposition 3.1. Suppose that there is a set of n items with exactly d unknown positive items. Let M be a $t \times n$ d-separable matrix. Given any test outcome vector, there exists a unique minimum solution Y for the reduced hitting set problem. Moreover, its size is d, except for $n_D = d$, the size can be $d - 1$.

Proof. First, note that $n_D \geq d$, since N_D contains all positive items.

Case 1 : If $n_D > d$.

We show that any minimum solution has size at least d.For contraction, if there exists a hitting set H of size $h < d$. Then putting other $d - h$ items from N_D into H would result in a hitting set of size d.

Since $n_D > d$, we can find two distinct hitting sets of size d. Note that the union of columns corresponding to any hitting set is the test outcome vector. Therefore, the two unions corresponding to two hitting sets of size d are equal, contradicting the definition of d-separability.

Moreover, all d positive items form a hitting set for positive pools. Therefore, the minimum hitting set has size exactly d. Furthermore, the hitting set of size d is unique since existence of two distinct hitting sets of size d yields the equality of two unions of d columns, contradicting the d-separability.

Case 2 : If $n_D = d$. The minimum hitting set K may have size k smaller than d, which would not result in any contradiction since N_D has the unique subset of size d. Since d-separable matrix is also $(d-1)$-separable. If $k < d-1$, then $K \cup \{x\} \neq K \cup \{y\}$ for $x \neq y$, $x, y \in N_D - K$, contradicting the $(d-1)$-separability. Hence $k = d-1$. The d-separability also assures the uniqueness of K. □

From Proposition 3.1, we know that if a d-pooling design whose corresponding matrix M is d-separable and we know that there are exactly d positive items additionally, then finding all positive items is equal to solve the hitting set problem. More precisely,

$$D = \begin{cases} Y & , n_D > d \\ N_d & , n_D = d \end{cases}.$$

Since for each $m \times n$ d-separable matrix M can correspond to a d-pooling design as follows: there are n items and m tests, and m_{ij}=1 iff test i contains item j. It means each d-separable

matrix M gives one d-pooling design which is non-adaptive.

Unfortunately, the hitting set problem is NP-hard, and it still has no good method except checking all possible d-subsets of all items to determine the hitting set.

Note that checking all possible d-subsets of all items spends $O(|N_d|^d)$.

Now we revisit 2-stage 2-pooling design given in section 2.1 . In the following we construct matrices corresponding to the first stage of the 2-stage 2-pooling design given in section 2.1. Recall that $n = c^k$ and the items are represented as k-tuple $(c_1, c_2, \ldots, c_k)$, where $c_i \in \{1, 2, \ldots, c\}$, and there are ck tests in the subsets

$$S_i(j) = \{(c_1, \ldots, c_{i-1}, j, c_{i+1}, \ldots, c_k) \mid c_\ell \in \{1, 2, \ldots, c\}\},$$

where $1 \leq j \leq c$, $1 \leq i \leq k$. So the corresponding matrix M has size ck by c^k. For convenience we use the k-tuple $c_1, c_2, \ldots, c_k$ over $\{1, 2, \ldots, c\}$ for the indices of columns, and 2-tuple (i, j) for the indices of rows, where $1 \leq i \leq k, 1 \leq j \leq c$. Then

$$m_{(i,j),(c_1,c_2,\ldots,c_k)} = \begin{cases} 1, & c_i = j; \\ 0, & \text{otherwise.} \end{cases}$$

Note that the matrix M has constant rowsum c^{k-1} and constant columnsum k.

We give an example of corresponding matrix to the first stage of the 2-stage 2-pooling design given in section 2.1 with $c = 2$ and $k = 3$ as following.

Example 3.2. Let M be the matrix corresponding to the first stage of the 2-stage 2-pooling design given in section 2.1 with $c = 2$ and $k = 3$. Then

$$M = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}$$

with columns indexed by $(1,1,1)$, $(2,1,1)$, $(1,2,1)$, $(2,2,1)$, $(1,1,2)$, $(2,1,2)$, $(1,2,2)$, $(2,2,2)$ in order, and rows indexed by $(1,1)$, $(1,2)$, $(2,1)$, $(2,2)$, $(3,1)$, $(3,2)$ in order.

Given any two items I_1 and I_2 whose representation are $(a_1, a_2, \ldots, a_k)$ and $(b_1, b_2, \cdots, b_k)$, where $a_i \neq b_i \ \forall i$ respectively. The union of two columns indexed by items I_1 and I_2 is the same as the union of columns indexed by items $(b_1, a_2, \cdots, a_k)$ and $(a_1, b_2, \cdots, b_k)$. This proves the following lemma.

Lemma 3.1.3. Let M be the matrix which represents the first stage of a 2-stage 2-pooling design given in section 2.1 . Then M is not a 2-separable matrix. □

Although M is not 2-separable matrix, M still have some good properties. For example, applying M in the first stage of a pooling design as described in section section 2.1 can conclude a smaller set including positive items, and reduce the number of tests in the next stage. Furthermore study of M is necessary.

3.2 s-disjunct matrix

In this section, we first give the definition of a binary matrix M to be d^e-disjunct and show that if there are at most d unknown positive items we can find all these d positive items by using the outcome and a given d-disjunct matrix M.

Definition 3.2.1. A binary matrix M is called **s^e-disjunct** if given any s+1 columns of M with one designated, there are e+1 rows with a 1 in the designated column and 0 in each of the other s columns.

An s^0-disjunct matrix is also called s-disjunct.

Example 3.3.

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

is a 2^1-disjunct matrix. Note that for a 2^1-disjunct matrix with 3 columns, the minimum number of row is 6.

In the following we talk about application of d-disjunct matrix on d-pooling design. Similarly to the section 3.1, represent d-pooling design as a binary matrix M, where the columns represent items, the rows represent tests, and $m_{ij} = 1$ iff item j is contained in the test i.

Suppose we have t tests. Then the t outcomes can also be represented by a t-vector $v = (v_1, ..., v_t)^t$, where v_i=1 iff the outcome of test i is positive; v_i=0 otherwise.

Proposition 3.2. Suppose M is a corresponding matrix of d-pooling design which is d-disjunct. An item is positive iff it (as a column) is contained by v.

Proof. Since a negative item (column) has at least one row not covered by the union of the up-to-d positive items ; such a row then has a negative outcome which identifies the item as negative. □

From this Proposition, we can conclude that if a pooling design whose corresponding matrix is d-disjunct, then it is simple to decode to find all positive items. Recall that in the section 3.1 we have mentioned if a d-pooling design whose corresponding matrix M is d-separable, then finding all positive items is equal to solve the hitting set problem. So finding all positive items with one d-pooling design corresponding matrix M which is d-separable is harder than with corresponding matrix M which is d-disjunct. This is reasonable. Since a d-disjunct matrix is also a d-separable matrix. If the reader want to know the proof, you may refer to [1].

We meet 2-stage 2-pooling design given in section 2.1 again.

Lemma 3.2.1. The matrix M corresponding to the first stage of the 2-stage 2-pooling design given in section 2.1 is not a 2-disjunct matrix.

Proof. Given any item I whose representation is $(a_1, a_2, \cdots, a_k)$. The union of columns indexed by items $(b_1, a_2, \cdots, a_k)$ and $(a_1, b_2, \cdots, b_k)$ covers the column indexed by I,where $b_i \neq a_i$.

□

From Example 4.2 and Example 4.4, we know that M is neither a 2-separable matrix nor a 2-disjunct matrix, and it seems to be impossible to modify this 2-stage algorithm to be 1-stage with small test times.

3.3 Method for constructing d^e-disjunct matrix

In [5], Macula gave a way of constructing disjunct matrices by the containment relation of subsets in a finite set. Now we introduce another construction mentioned in [3], which uses intersecting relation of subsets in a finite set. Before we give a construction, we define a notation which will be used to construct a d^e-disjunct matrix.

Definition 3.3.1. For positive integers $1 \leq d < k < n$, let $M(i; d, k, n)$ be the binary matrix with rows indexed with $\binom{[n]}{d}$ and columns indexed with $\binom{[n]}{k}$ such that $M(A, B) = 1$ iff $|A \cap B| = i$.

Let $B \in \binom{[n]}{k}$ and C=$[n]\backslash B$. Then, for any $D \in \binom{[n]}{d}$,$|D \cap B| = i$ iff $|D \cap C| = d - i$. Therefore,$M(i; d, k, n) = M(d - i; d, n - k, n)$ when $n > k + d - i$.
Since $i \leq \lfloor \frac{d}{2} \rfloor$ iff $d - i \geq \lfloor \frac{d+1}{2} \rfloor$, we always assume that $i \geq \lfloor \frac{d+1}{2} \rfloor$.

Theorem 3.3.1. Let $1 \leq s \leq i$, $\lfloor \frac{d+1}{2} \rfloor \leq i \leq d < k < n$ and $n - k - s(k + d - 2i) \geq d - i$. Then $M(i; d, k, n)$ is an s^{e_2}-disjunct matrix, where e_2=$\binom{k-s}{i-s}\binom{n-k-s(k+d-2i)}{d-i}$-1

Proof. Let $B_0, B_1, ..., B_s \in \binom{[n]}{k}$ be any $s + 1$ distinct columns of M. Let $x_i \in B_0\backslash B_i, i \in [s]$ and $X_0 = \{x_i | i \in [s]\}$. Let $\mathcal{A}_0 = \{A_0 \ | A_0 \in \binom{[n]}{i}$ such that $X_0 \subseteq A_0 \subseteq B_0\}$. Note that $|\mathcal{A}_0|$ =$\binom{k-|X_0|}{i-|X_0|}$= $k - |X_0|choosek - i$ and $|\mathcal{A}_0| \geq \binom{k-s}{k-i}$= $\binom{k-s}{i-s}$. Given any $A_0 \in \mathcal{A}_0$. We have $|A_0 \cap B_0| = i$, $|A_0 \cap B_j| < i$, $\forall j \in [s]$. Also notice that for B_j, $j \in [s]$. If $\exists$ $D \in \binom{[n]}{d}$ such that

$|D \cap B_0| = |D \cap B_j| = i$. Then $|B_0 \cap B_j| \geq 2i - d$. Let $Y = \{j | 1 \leq j \leq s, |D \cap B_j| \neq i, \forall D \in \binom{[n]}{d}$ such that $|D \cap B_0| = i\}$. We have $|\bigcup_{0 \leq j \leq s, j \notin Y} B_j| \leq k + (s - |Y|)[k - (2i - d)]$. Let $\mathcal{D}' = \{D' | D' \in \binom{[n]}{d}, |D' \cap B_0| = i, |D' \cap B_j| \neq i, \forall j \in [s]\}$. Then $|\mathcal{D}'| \geq \binom{k-s}{i-s}\binom{n-k-(s-|Y|)(k-2i+d)}{d-i}$ $\geq \binom{k-s}{i-s}\binom{n-k-s(k-2i+d)}{d-i}$.

□

We give some examples by using this construction.

Example 3.4. 1. $M(3; 3, 4, 8)$ is a 3-disjunct matrix, which has $\binom{8}{3}$ rows and $\binom{8}{4}$ columns.

2. $M(2; 2, 3, 6)$ is a 2-disjunct matrix, which has $\binom{6}{2}$ rows and $\binom{6}{3}$ columns.

3. $M(1; 1, 2, 4)$ is a 1-disjunct matrix, which has $\binom{4}{1}$ rows and $\binom{4}{2}$ columns.

   Note that the number of rows satisfies the equality of Sperner Theorem, which we will mention in the next Chapter.

Chapter 4

$(d, s]$-disjunct matrix and $(d, s]$-cover

In this chapter, we generalize the concept of d-disjunct matrix to $(d, s]$-disjunct matrix and its relative $(d, s]$-cover. The study of $(d, s]$-disjunct matrix may inspire us to realize d-disjunct matrix better.

4.1 $(d, s]$-disjunct matrix

Definition 4.1.1. A binary matrix M is called $(d, s]$-**disjunct** matrix if given any $d + s$ columns of M with d columns designated, there is at least 1 row with 1 in these s designated columns and 0 in each of the other d columns.

Remark 4.1.1. From the above definition, we can observe that $(d, 1]$-disjunct matrix is also a d-disjunct matrix.

Example 4.1.

$$\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

is a $(1, 2]$-disjunct matrix.

In the following, we provide a trivial construction for $(d, s]$-disjunct matrix. First we give an example to illustrate this construction.

Example 4.2. In this example, we construct $(d, s]$-disjunct matrix by trivial construction for the case $d = s = 2$. Fix w, $s \leq w \leq n - d$. Let M be the $\binom{n}{w} \times n$ binary matrix with each row consisting of the characteristic vector of each w-subset of $[n]$.

For the case $s = d = 2$ and $n = 4$. Fix w=4. We have

$$M = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}.$$

Note that in this construction we have the number of colimns=$n \leq \binom{n}{w}$ = the number of rows.

Proposition 4.1. Let M be as in Example 4.2. Then M is $(d, s]$-disjunct.

Proof. We shall prove the incidence matrix between $\binom{n}{w}$ and $[n]$ with row indexed by w-subset of [n] and column indexed by 1-subset of $[n]$ is $(d, s]$-disjunct. Pick $P, Q \subseteq [n]$ with $|P| = s$, $|Q| = d$, and $P \cap Q = \emptyset$. Then there exists a subset in $\binom{n}{w}$ and in the interval $[P, [n] - Q]$ since $s \leq w \leq n - d$. □

Definition 4.1.2. Let M be a binary matrix. The **complement matrix** $\overline{M}$ is defined by

$$\begin{cases} \overline{M}_{ij} = 1 & \text{if } m_{ij} = 0 \\ \overline{M}_{ij} = 0 & \text{if } m_{ij} = 1 \end{cases}.$$

Corollary 4.1.1. If M is a $(d, s]$-disjunct matrix, then $\overline{M}$ is a $(s, d]$-disjunct matrix.

Proof. Given any $d + s$ columns with d designed, we want to find a row in $\overline{M}$ with 1 in these d designated columns and 0 in the other s columns. Since M is $(d, s]$-disjunct. So we can find a row i with 1 in these s designated columns and 0 in the other d columns. By the definition

of complement matrix, for the matrix $\overline{M}$ row i is with 1 in these d designated columns and 0 in the other s columns.

□

From the Proposition 4.1 and Corollary 4.1.1, give any s, d, and n, with $n \geq s + d$. If $d < s$ then we find integer w in $[d, n - d]$ to obtain the minimum of $\binom{n}{w}$. Otherwise, we find integer w in $[s, n - s]$ to obtain the minimum of $\binom{n}{w}$.

4.2 $(d, s]$-cover and relation between $(d, s]$-disjunct matrix and $(d, s]$-cover

$(d, s]$-cover has a close relation with $(d, s]$-disjunct matrix. Roughly speaking, $(d, s]$-cover and $(d, s]$-disjunct matrix have the same meaning, but one uses the notation of subset the other uses the notation of matrix.

Definition 4.2.1. For $P, Q \subseteq [n]$, define $[P, \overline{Q}]$=$\{X : X$ is a set, $P \subseteq X \subseteq \overline{Q}\}$

Definition 4.2.2. Suppose $s + d \leq n$. An $(d, s]$**-cover** of $[n]$ is a family X of subsets of $[n]$ such that $[P, \overline{Q}] \cap X \neq \emptyset$ for each pair $P, Q \subseteq [n]$ with $|P| = s$, $|Q| = d$, and $P \cap Q = \emptyset$.

Proposition 4.2. Let X be a $(d, s]$-cover of $[n]$. Then the incidence matrix of X and $[n]$ is $(d, s]$-disjunct.

Note that the incidence matrix of X is a binary matrix with columns indexed with 1,2,...,n and rows indexed with subsets in family X.

Proof. Given any $d+s$ columns of M with s columns designed, namely $a_1, a_2, \ldots, a_s$, we want to find a row with 1 in these $a_1, a_2, \ldots, a_s$ columns and 0 in each of the other d columns, namely $b_1, b_2, \ldots, b_d$. Let Q=$\{x :$ column a_i is indexed by x, $1 \leq i \leq d\}$ and P=$\{x :$ column b_i is indexed by x, $1 \leq i \leq s\}$. Since X is a $(d, s]$-cover of $[n]$. So there exists D s.t. $D \in [P, \overline{Q}] \cap X$. Then the row indexed by D is the row we want to find. □

In fact the converse is also true.

Proposition 4.3. If an $t \times n$ binary M matrix is $(d, s]$-disjunct, and for $1 \leq i \leq t$ set $T_i = \{j \mid m_{ij} = 1\}$. Then $\{T_1, T_2, \ldots, T_t\}$ is an $(d, s]$-cover of $[n]$.

Proof. Pick any $P, Q \subseteq [n]$ with $|P| = s$ and $|Q| = d$ such that $P \cap Q = \emptyset$. Pick $i \in [t]$ such that $m_{ij} = 1$ for all $j \in P$ and $M_{ik} = 0$ for all $k \in Q$. Hence $P \subseteq T_i \subseteq \overline{Q}$. □

Combine Proposition 4.2 and Proposition 4.3, $(d, s]$-disjunct matrix and $(d, s]$-cover they have one-to-one correspondence.

Now we go back to the 2-stage 2-pooling design given in section 2.1 again, and investigate whether it is a (2.1]-disjunct matrix.

Consider P=$\{(1, 1, \ldots, 1)\}$ and Q=$\{(2, 1, 1, \ldots, 1), (1, 2, 1, \ldots, 1)\}$. Suppose $P \subseteq S_i(j) \subseteq \overline{Q}$ for some $(i, j) \in \{1, 2, \ldots, k\}$×$\{1, 2, \ldots, c\}$. Then j=1 since $P \subseteq S_i(j)$. Since $(2, 1, 1, \ldots, 1) \notin S_i(1), i = 1$. Since $(1, 2, \ldots, 1) \notin S_i(1), i = 2$, a contradiction. Also notice that $m_{(i,j)(c_1,c_2,\ldots,c_k)} = 1$ iff $(c_1, c_2, \ldots, c_k) \in S_i(j)$. Hence we have the following remark.

Remark 4.2.1. Let M be the matrix corresponding to the first stage of 2-stage 2-pooling design given in section 2.1 with size $ck \times c^k$, and for $(i, j) \in \{1, 2, \ldots, k\} \times \{1, 2, \ldots, c\}$, $T_{(i,j)}$=$\{(c_1, c_2, \ldots, c_k) : m_{(i,j)(c_1,c_2,\ldots,c_k)} = 1\}$. Then $T_{(i,j)}$ is not an $(2, 1]$-cover of all items $\{(q_1, q_2, \ldots, q_k) : 1 \leq q_i \leq c, \forall 1 \leq i \leq k\}$.

4.3 the minimum number of rows of $(d, s]$-disjunct matrix

Before we discuss the minimum number of rows with fixed number of columns in a $(d, s]$-disjunct matrix, we mention Sperner Theorem first, which gives the upper bound of the number of columns with the fixed number of rows and provide a lower bound of a $(1, 1]$-disjunct matrix with the fixed number of columns.

The following theorem is referred to [6].

Theorem 4.3.1. (Sperner 1928) If M is $(1,1]$-disjunct matrix then the number of columns of $M \leq \binom{t}{\lfloor t/2 \rfloor}$.

Note that for the number of columns of $M = \binom{t}{\lfloor t/2 \rfloor}$ we have the equality.

For $t = 4$ we have the number of columns of $M = \binom{4}{2} = 6$ here, but for the columns of $M = 6$ in the trivial construction we have $t = \binom{6}{w} \geq 6$.

For the rest of this section, we will discuss the minimum number of rows of $(d, s]$-disjunct matrix.

Definition 4.3.1.

1. Let $t(d, s, n)$ denote the minimum number of t such that a $(d, s]$-disjunct matrix of size $t \times n$ exists.
2. Let $t(d, s, n, w)$ denote the minimum number of t such that a $(d, s]$-disjunct matrix with constant row sum w of size $t \times n$ exists.
3. Let $t_c(d, s, n) = \min_{w \in [n]} t(d, s, n, w)$.

Remark 4.3.1.

1. $t(d, s, n) \leq t_c(d, s, n) \leq t(d, s, n, w)$ for $w \in [n]$.
2. $t(d, s, n)$ is the minimal size of a $(d, s]$-cover of $[n]$.
3. Sperner's Theorem says that if n is an integer of the form $\binom{t}{\lfloor t/2 \rfloor}$ then $\binom{t(1,1,n)}{\lfloor t(1,1,n)/2 \rfloor} = n$

Sperner Theorem implies $t(1, 1, \binom{t}{\lfloor t/2 \rfloor}) = t$. EFF proves $t(1, 1, n) = n$ if $n \leq 4$, and then $t(1, 1, 5) = 4$.

Proposition 4.4. $t(1, 1, n)$ is the least integer such that $\binom{t}{\lfloor t/2 \rfloor} \geq n$.

Proof. First we show that $t(1, 1, n) \leq t$. Let t is the least integer such that $\binom{t}{\lfloor t/2 \rfloor} \geq n$. Given any n $\in \mathbb{N}$. Since t is the least integer such that $\binom{t}{\lfloor t/2 \rfloor} \geq n$

By Sperner Theorem, we have a $t \times \binom{t}{\lfloor t/2 \rfloor}$ $(1, 1]$-disjunct matrix M. Delete any $\binom{t}{\lfloor t/2 \rfloor} - n$ columns from M, and let it be M'. Then M' is a $t \times n$ $(1, 1]$-disjunct matrix . Hence $t(1, 1, n) \leq t$.

Now, we show that $t(1,1,n) = t$, By Sperner Theorem, $n \leq \binom{t(1,1,n)}{\lfloor t(1,1,n)/2 \rfloor}$
Since t is "the least" integer such that $n \leq \binom{t}{\lfloor t/2 \rfloor}$. So $t(1,1,n) = t$.

Now we derive one lower bound for $t(d,s,n.w)$.

Lemma 4.3.1. Suppose $s \leq w \leq n-d$. Then

$$\begin{pmatrix} n \\ s \end{pmatrix} \times \begin{pmatrix} n-s \\ n-s-d \end{pmatrix} \leq t(d,s,n,w) \times \begin{pmatrix} n-w \\ n-w-d \end{pmatrix} \times \begin{pmatrix} w \\ s \end{pmatrix}.$$

Proof. Let $Y = \{(P,Q) : \text{Both } P \text{ and } Q \text{ are subsets of } [n], |P| = s, |Q| = d \text{ and } P \cap Q = \emptyset\}$. Let X be a $(d,s]$-cover of $[n]$ with size $t(d,s,n,w)$. Then given any pair (u,v) in Y, there exists subset $T \in X$ s.t. $U \subseteq T \subseteq [n] \setminus V$.
Let $T' = \{(A,B,T) : T \in X, A \subseteq T, B \subseteq [\text{n}] \setminus T, |A| = s, |B| = d\}$. Then $|T'| = t(d,s,n,w)\binom{w}{s}\binom{n-w}{n-w-d}$. Since for each pair (U,V) in Y, we can find a pair $(U,V,T) \in T'$. Hence $|Y| = \binom{w}{s}\binom{n-w}{n-w-d} \leq |T'| = t(d,s,n,w)\binom{w}{s}\binom{n-w}{n-w-d}$.

□

Lemma 4.3.1 gives us a lower bound for $t(d,s,n,w)$, we write it down as Corollary 4.3.1.

Corollary 4.3.1.

$$t(d,s,n,w) \geq \lceil u(w) \rceil,$$

where

$$u(w) = \frac{\begin{pmatrix} n \\ s \end{pmatrix}\begin{pmatrix} n-s \\ d \end{pmatrix}}{\begin{pmatrix} n-w \\ d \end{pmatrix}\begin{pmatrix} w \\ s \end{pmatrix}}.$$

We want to know when $u(w)$ obtains minimum. □

$$\frac{u(w)}{u(w+1)} = \frac{(w+1)(n-w-d)}{(w+1-s)(n-w)} \begin{cases} \geq 1, & \text{if } w \leq \lfloor \frac{ns-d}{d+s} \rfloor; \\ < 1, & \text{else.} \end{cases}$$

Corollary 4.3.2.

$$t_c(d,s,n) \geq u(\left\lfloor \frac{ns-d}{d+s} \right\rfloor).$$

In fact $u(\lfloor \frac{ns-d}{d+s} \rfloor)$ tends to a constant as $n \to \infty$. But we still can guess that the row weight of a good $(d,s]$-disjunct matrix is closed to $sn/(d+s)$.

Bibliography

[1] D.Z.Du and F.K.Hwang, *Pooling designs and nonadaptive group testing: important tools for DNA sequencing*, World Scientific Publishong Company, 2006.

[2] D.Z.Du and F.K.Hwang, *Combinatorial group testing and its applications*, World Scientific Publishong Company, 2nd Edition.

[3] Jun Guo and Kaishun Wang, "A construction of pooling designs with surprisingly high degree of error correction" *Journal of Combinatorial Theory, Series A, Vol. 118, Issue 7, pp.2056-2058, Oct., 2011.*

[4] 林呈翰兩階段群試設計, 台中一中科學班專題研究報告

[5] A.J. Macula, A simple construction of d-disjunct matrices with certain constant weights, Discrete Math. 162 (1996) 311- 312.

[6] J.H van Lint,and R.M Wilson, *A course in combinatorics,* Cambridge University Press, 2nd Edition.