PREFACE

In a bid to standardise higher education in the country, the University Grants Commission (UGC) has introduced Choice Based Credit System (CBCS) based on five types of courses viz. *core, discipline specific, generic elective, ability* and *skill enhancement* for graduate students of all programmes at Honours level. This brings in the semester pattern, which finds efficacy in sync with credit system, credit transfer, comprehensive continuous assessments and a graded pattern of evaluation. The objective is to offer learners ample flexibility to choose from a wide gamut of courses, as also to provide them lateral mobility between various educational institutions in the country where they can carry acquired credits. I am happy to note that the University has been accredited by NAAC with grade 'A'.

UGC (Open and Distance Learning Programmes and Online Learning Programmes) Regulations, 2020 have mandated compliance with CBCS for U.G. programmes for all the HEIs in this mode. Welcoming this paradigm shift in higher education, Netaji Subhas Open University (NSOU) has resolved to adopt CBCS from the academic session 2021-22 at the Under Graduate Degree Programme level. The present syllabus, framed in the spirit of syllabi recommended by UGC, lays due stress on all aspects envisaged in the curricular framework of the apex body on higher education. It will be imparted to learners over the *six* semesters of the Programme.

Self Learning Materials (SLMs) are the mainstay of Student Support Services (SSS) of an Open University. From a logistic point of view, NSOU has embarked upon CBCS presently with SLMs in English / Bengali. Eventually, the English version SLMs will be translated into Bengali too, for the benefit of learners. As always, all of our teaching faculties contributed in this process. In addition to this we have also requisitioned the services of best academics in each domain in preparation of the new SLMs. I am sure they will be of commendable academic support. We look forward to proactive feedback from all stakeholders who will participate in the teaching-learning based on these study materials. It has been a very challenging task well executed, and I congratulate all concerned in the preparation of these SLMs.

I wish the venture a grand success.

Professor (Dr.) Subha Sankar Sarkar
Vice-Chancellor

Netaji Subhas Open University

Under Graduate Degree Programme
Choice Based Credit System (CBCS)
Subject: UG Mathematics (HMT)
Course : Applications of Algebra
Course Code - GE-MT-31

First Print : August, 2022

Printed in accordance with the regulations of the
Distance Education Bureau of the University Grants Commission.

Netaji Subhas Open University

Under Graduate Degree Programme
Choice Based Credit System (CBCS)
Subject: UG Mathematics (HMT)
Course : Applications of Algebra
Course Code - GE-MT-31

: Course Writer :

Chapter-1 & 2 **Mrinal Nath**
Assistant Professor of Computer Science, NSOU

Chapter-3, 4 & 5 **Dr. Satyabrota Kundu**
Assistant Professor of Mathematics, Loreto College

: Course Editor :
Dr. Sujit Kumar Sardar
Professor of Mathematics, Jadavpur University

: Format Editors :
Mr. Mrinal Nath
NSOU

Notification

Kishore Sengupta
Registrar

Netaji Subhas Open University

UG Mathematics
(HMT)

Course : Applications of Algebra

Course Code - GE-MT-31

Unit-1 ❒ Coding Theory 7–42

Unit-2 ❒ Block Design 43–67

Unit-3 ❒ Symmetry Groups and Color Pattern 68–97

Unit-4 ❒ Application of Linear Transformation 98–144

Unit-5 ❒ Matrix Theory 145–205

Unit 1 ❑ Coding Theory

Structure

1.0 Objectives

1.1 Introduction

1.2 Algebraic preliminaries for Coding Theory

1.3 Linear Block Codes

1.4 Cyclic Codes

1.5 Summary

1.6 Exercises

1.7 Reference and further reading

1.0 Objectives

After going through this unit the learner should be able to :

- define the Linear Block Code.
- understand the generator matrix of linear code.
- define parity check matrix.
- understand process to create parity check matrix for a linear code.
- define Hamming Distance and Hamming Weight.
- understand how Hamming Code can detect and correct the error in data transmission.
- define Cyclic code and their relation with polynomials of ring.
- use abstract and linear algebra as a tool for coding theory.

1.1 Introduction

In the recent years, a significant increase of interests has been noticed in the field of digital data transmissions and storage systems. With the advent of large-scale, high speed data network, efficiency and reliability becomes two most important parameters to measure the quality of digital data transmission. Error is inherent in any digital transmission. In particular, when the transmission media or the channel is noisy, it becomes a major concern for the designer to control the errors so that the reliable reproduction of data is obtained. In 1948, Shannon demonstrated in a

landmark paper that, by proper encoding of the information, errors induced by the noisy channel can be reduced to any desired level without sacrificing the rate of information transmission. Since Shannon's work a great deal of effort has been expended on the problem of devising effective encoding and decoding methods for error control in noisy environment.

A typical transmission system can be represented by the block diagram shown in **Figure 1.1.** The ***Information Source*** is a digital computer or similar machine. It sends the output to the Source Decoder either as a continuous waveform or a sequence of discrete symbols. The ***Source Encoder*** transforms this into a sequence of binary digits called the information sequence u. The ***Channel Encoder*** transforms the information sequence u into a discrete encoded sequence v (mostly binary) known as code word. Since the discrete symbols are not suitable for transmission in physical

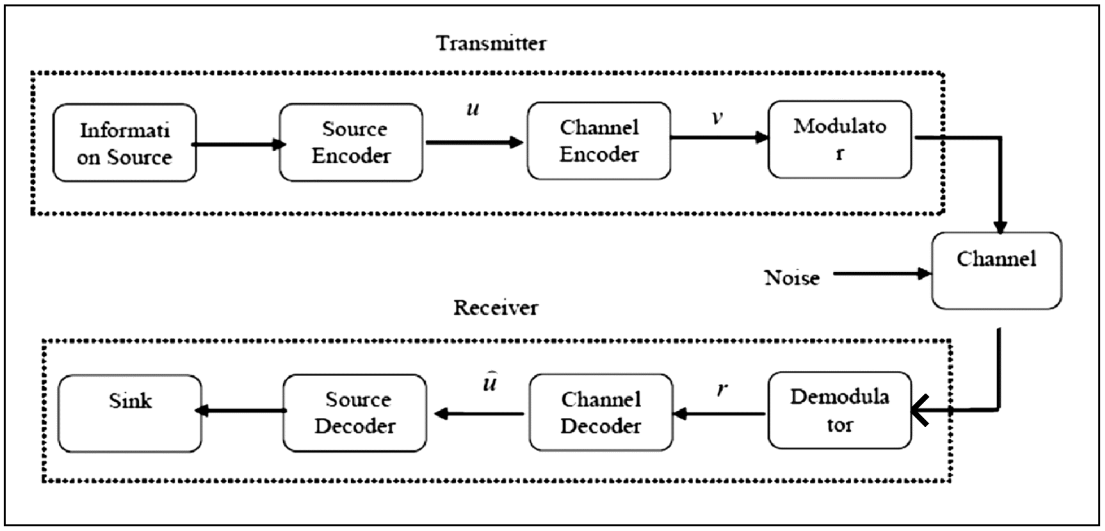


Figure 1.1

channel, the ***Modulator*** transforms output symbol into a waveform which is suitable for transmission. This waveform enters into the channel and is corrupted by noise. The output from channel is received by the ***Demodulator*** which produces a sequence r corresponding to the encoded sequence v. The Channel Decoder transforms the received sequence r into a binary sequence $\hat{u}$ known as estimated sequence. The ***Source Decoder*** transforms the estimated sequence $\hat{u}$ into an estimate of the source output and delivers it to the final destination. Now the primary focus of this unit is to design of Channel Encoders and Channel Decoders to combat the noisy environment. The strategy and principle to transform the sequence u into v or r into $\hat{u}$ will be discussed with the help of abstract and linear algebra.

1.2 Algebraic preliminaries for Coding Theory

1.2.1 Group

Definition 1.1

A set G on which a binary operation * is defined is called group if the following conditions are satisfied.

(1) The binary operation * is associative i.e. for any $a, b, c \in G$

$$a * (b * c) = (a * b) * c$$

(2) G contains an element e such that, for any $a \in G$

$$a * e = e * a = a.$$

This element e is called an identity element of G w.r.t. the binary operation *. (It can be proved that e is unique and once it is proved one can say the identity instead of an identity.)

(3) For any $a \in G$, there exists an element $a' \in G$ such that

$$a * a' = a' * a = e$$

This element a' is known as an inverse of a . (It can be proved that a' is unique and once it is proved one can call it the inverse.).

A group G is said to be commutative or Abelian if its binary operation satisfies the condition : For all $a, b \in G, a * b = b * a$

The number of elements in a group is called the order of the group. A group of finite order is called finite group.

Example 1.1

Consider the set of two integers, $G = \{0, 1\}$. Let us define a binary operation, denoted by, on G as follows :

$$0 \oplus 0 = 0, \qquad 0 \oplus 1 = 1, \qquad 1 \oplus 0 = 1, \qquad 1 \oplus 1 = 0.$$

This binary operation is called modulo-2 addition. To be more precise, $a \oplus b$ denotes the remainder after dividing the result of $(a + b)$ by 2. The operation can be thought of as clock operation which has 0 at top of the clock and 1 at the bottom. Let us consider we are at the top (0). Now adding 1 will send our position at the bottom (1) in the clockwise direction. Adding another 1 will again send us to the top (0) from where we started our journey. (Refer the diagram 1.2).

The set $G = \{0, 1\}$ is a group under modulo-2 addition. It follows from the definition of the modulo-2 addition $\oplus$ that G is closed under $\oplus$ and $\oplus$ is

commutative. It can easily be checked that $\oplus$ is associative. The element 0 is the identity element. The inverse of 0 is itself and the inverse of 1 is also itself. Thus G together with $\oplus$ is a commutative group and this group is usually denoted by $\mathbf{Z}_2$.

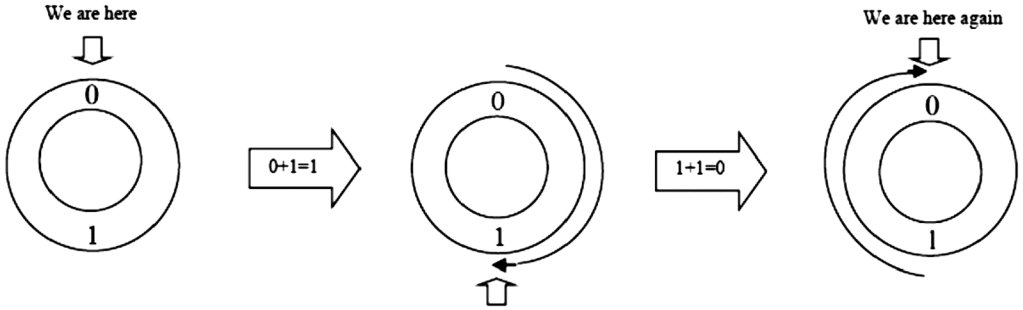


Figure 1.2

N.B. The set {0, 1, 2, 3,....., m – 1} is a group under modulo-m addition for any positive integer m and this group is usually denoted by $\mathbf{Z}_m$.

1.2.2 Ring

Definition 1.2

Let R be a set of elements on which two binary operations, called addition '+' and multiplication '.' are defined (for any a, b $\in$ R, a.b will be written as ab). The set R together with the two binary operations + and. is said to be a ring if the following conditions are satisfied.

(1) R is a commutative group under addition +. The identity element with respect to addition is called the zero element and denoted by 0 .

(2) Multiplication '.' associative: $a(bc) = (ab)c$ for all $a, b, c \in R$.

(3) Multiplication '.' is distributive over addition + from both left and right i.e., $a(b + c) = ab + ac$ and $(b + c)a = ba + ca$ for all $a, b, c \in R$.

A ring R is said to be commutative if multiplication is commutative i.e., if $ab=ba \forall a,b \in R$. A ring R is said to have identity (denoted by 1) if 1 is the identity with respect to multiplication i.e., if $a1 = 1a \forall a \in R$.

1.2.3 Field

Definition 1.3

Roughly speaking, a field is a set of elements in which addition, subtraction, multiplication, division can be done without leaving the set. A formal definition of the field is given below.

Let F be a set of elements on which two binary operations, called addition '+' and multiplication '.' are defined . The set F together with the two binary operations '+' and '.' is said to be a field if the following conditions are satisfied.

(4) F is commutative group under addition +. The identity element with respect to addition is called the zero element and denoted by 0 .

(5) The set of non-zero elements in F ($F \setminus \{0\}$) is commutative group under multiplication '.'.The identity element with respect to multiplication is known as the unit element and denoted by 1 .

(6) Multiplication is distributive over addition; that is for any three elements $a, b, c \in F$,

$$a(b + c) = ab + ac$$

In other words, a commutative ring with identity is called a field if every non zero element has multiplicative inverse.

The no of elements of a field is the order of the field. A field with finite number of elements is known as a finite field. The additive inverse of an element a is denoted by $-a$ and the multiplicative inverse is denoted by a^{-1}, provided that a $\neq 0$.

Example 1.2

Consider the set {0,1} together with modulo-2 addition and modulo-2 multiplication shown in table 1.1 and 1.2. In, Example 1 it has been shown that {0, 1} is a commutative group under modulo-2 addition. It can be easily checked that {1} is also a commutative group under modulo-2 multiplication. It is easy to verify that modulo-2 multiplication is distributive over modulo-2 addition by computing $a(b + c)$ and $ab + ac$ for eight (2^3) possible combinations of a, b and c. Therefore, the set {0, 1} is a field of two elements under modulo-2 addition and modulo-2 multiplication.

The field given by Example 1.2 is known as binary field which plays an important role in coding theory. Normally we denote binary field by B. It is also denoted by $\mathbf{Z}_2$. Finite fields are also called Galois Fields (GF), in honour of their discoverer. Therefore, the binary field is also represented by GF(2).

Modulo-2 Addition

+	0	1
0	0	1
1	1	0

Table 1.1

Modulo-2 Multiplication

×	0	1
0	0	0
1	0	1

Table 1.2

For any prime number p, $\mathbf{Z}_p$ is a field with addition as addition modulo p and multiplication as multiplication modulo p.

1.2.4 Vector Space

Definition 1.4

Let V be a set of elements on which the binary operation addition + is defined. Let F be a field with two operations '+' and '.' (+ of V and + of F will not create any confusion as context will make it clear and . of F will be written by juxtaposition i.e., for any a, b ε F, a.b will be written as ab). A multiplication operation also denoted by juxtaposition, between the elements in F and the elements in V is defined. Then V is said to be a vector space over the field F if it satisfies the following conditions.

(1) V is a commutative group under addition +.

(2) For any $a \in F$ and any $v \in V$, av is an element V.

(3) For any $u, v \in V$ and any $a, b \in F$, $a(u + v) = au + av$ and $(a + b)v = av+bv$.

(4) For any $v \in V$ and any $a, b \in F$, $(ab)v = a(bv)$.

(5) Let 1 be the unit element of F Then for any $v \in V$, $1v = v$.

The elements of V are called vectors and elements of the field F are called scalars. The addition + in V is called vector addition and the multiplication that combines the scalar in F and vector in V is referred to as multiplication of a vector by a scalar. The additive identity of V is denoted by 0 (called the null vector). The additive identity of F is also denoted by 0 (called the scalar 0 and the context will not create any confusion). The following two properties will be used several times in the sequel.

Property 1 : For any vector $v \in V$ and $0 \in F$, $0v = 0$

Property 2 : For any scalar $c \in F$ and $0 \in$ V, c0 = 0

Let us consider an ordered sequence of n components, $(a_0, a_1, a_2 \ldots\ldots\ldots a_{n-1})$ where each components a_i is an element from the binary field GF(2) (i.e. $a_i = 0$ or 1. This sequence is generally called an n-tuple over GF(2). Since there are two choices for each a_i, 2^n distinct n tuples can be constructed. Let us denote the set of this 2^n distinct n tuples by V_n. Now let us define addition + on V_n in the following way : For any $\boldsymbol{u} = (u_0, u_1, \ldots\ldots, u_{n-1})$ and $v = (v_0, v_1, \ldots\ldots\ldots, v_{n-1})$ in V_0,

$$u + v = (u_0 + v_0, u_i + v_i, \ldots\ldots u_{n-1} + V_{n-1}) \qquad (1)$$

here $u_i + v_i$ is carried out in modulo-2 addition. Since u_i and v_i both are either 0 or 1, $u_i + v_i$ will also be either 0 or 1(refer to Table 2.1).Clearly $\boldsymbol{u} + \boldsymbol{v}$ also in n-tuple over GF(2). Hence V_n is closed under addition defined in (1).

It can be seen that all 0 n-tuple $\mathbf{0} = (0, 0, \ldots\ldots, 0)$ is the additive identity. For any v in V_n , $\mathbf{v}+\mathbf{v} = (v_0 + v_0, v_i + v_i, \ldots\ldots v_{n-1} + v_{n-1}) = (0, 0, \ldots\ldots, 0) = 0$. Hence the additive inverse of each n-tuple in V_0 is itself. The addition defined in (1) is commutative and associative. Therefore, V_0 is commutative group under the addition.

Now let us also define scalar multiplication '.' of an n-tuple v in V_n by an element a from GF(2) as follows :

$$a(v_0, v_1, v_2, \ldots\ldots, v_{n-1}) = (av_0, av_1, av_2, \ldots\ldots, av_{n-1}) \qquad (2)$$

where av_i is carried out in modulo 2 multiplication (refer to Table 2.2). Then it can be easily shown that V_0 satisfies all the conditions outlined in the definition of vector space. Therefore, the set V_n of all n-tuples over GF(2) forms a vector space over GF(2).

Example 1.3

Let n = 4. The vector space 4 V of all 4-tuples over GF(2) consist of following 16 vectors.

(0 0 0 0), (0 0 0 1), (0 0 1 0), (0 0 1 1),
(0 1 0 0), (0 1 0 1), (0 1 1 0), (0 1 1 1),
(1 0 0 0), (1 0 0 1), (1 0 1 0), (1 0 1 1),
(1 1 0 0), (1 1 0 1), (1 1 1 0), (1 1 1 1),

The result of vector addition of (1 0 0 1) and (1 1 0 1) is :

(1 0 0 1) + (1 1 0 1) = (1 + 1, 0 + 1, 0 + 0, 1 + 1) = (0 1 0 1)

The result of scalar multiplication on some vector (1, 0, 1, 1) by the element of GF(2) is :

0(1 0 1 1) = (01, 00, 01, 01) = (0 0 0 0)
1(1 0 1 1) = (11, 10, 11, 11) = (1 0 1 1)

The vector space of all n tuples over GF(2) or over an extension field of GF(2) [e.g. GF (2^m), m is a positive integer] can be constructed similarly.

1.2.5 Subspace

Definition 1.5

A nonempty subset S of vector space V over a field F is said to be a subspace if S itself is a vector space over the field F .

Let S be a subset of vector space V over a field F . Then S be subspace of V if the following conditions are satisfied.

(1) The 0 vector is in S .

(2) S is closed under vector addition. That is, for any two vectors

$$u, v \in S \Rightarrow u + v \in S$$

(3) S is closed under scalar multiplication. That is, for any

$$u, \in S, a \in F \Rightarrow au \in S$$

Let $S = \{v_1, v_2, ..., v_k\}$ be a subset of a vector space V over a field F. Let $a_1, a_2, ..., ak$ F be k scalars. The sum $a_1v_1 + a_2v_2 + ... + a_kv_k$ is called linear combination of $v_1, v_2, ..., v_k$. It can be proved that the set of all linear combinations $v_1, v_2, ..., v_k$ forms a subspace of V and it is denoted by L(S) and called the linear span of S. It can also be checked that L(S) is the intersection of all subspaces of V containing S i.e., the smallest subspace of V containing S.

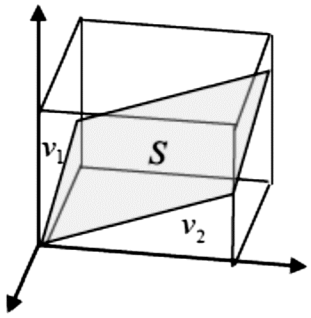


Figure 1.3

In Figure 1.3 the shaded area shows a subspace S formed by taking all linear combinations of two vector $v_1, v_2 \in V_3$ where vector space V_3 is comprised of 3-tuples over the field of real numbers **R**. The subspace S can be represented as $a_1v_1 + a_2v_2$ where $a_1, a_2 \in R$. Clearly the subspace is a plane in 3 dimensions.

Example 1.4

Consider the vector space V_4 of all 4-tuples over GF(2) given by example 4.

The linear combination of (0 0 1 1) and (1 0 1 0) are

$$0(0\ 0\ 1\ 1) + 0(1\ 0\ 1\ 0) = (0\ 0\ 0\ 0)$$
$$0(0\ 0\ 1\ 1) + 1(1\ 0\ 1\ 0) = (1\ 0\ 1\ 0)$$
$$1(0\ 0\ 1\ 1) + 0(1\ 0\ 1\ 0) = (0\ 0\ 1\ 1)$$
$$0(0\ 0\ 1\ 1) + 0(1\ 0\ 1\ 0) = (0\ 0\ 0\ 0)$$

These 4 vectors form a subspace of V_4.

A set $\{v_1, v_2, ..., v_k\}$ 1 of vectors in a vector space V over a field F is said to be linearly dependent if and only if there exist k scalars $a_1, a_2, ..., a_k \in F$, not all zero, such that

$$a_1v_1 + a_2v_2 + ... + a_kv_k = 0 \text{ (the null vector)} \qquad (3)$$

A set $(v_1, v_2, \ldots\ldots, v_k)$ of vectors in a vector space V is said to be linearly independent if it is not linearly dependent. That is,

$$a_1v_1 + a_2v_2 + \ldots\ldots + a_kv_k \neq 0 \qquad (4)$$

unless $a_1 = a_2 = \ldots\ldots\ldots\ldots = a_k = 0$

A set S of vectors is said to span a vector space V if L(S) = V i.e., every vector of V is a linear combination of the vectors in the set S. In any vector space or subspace there exists at least one set B of linearly independent vectors which span the space. This set is called a basis of the vector space. The no of vectors in the basis of a vector space is called the dimension of the vector space.

Example 1.5

Consider the vector space V_n of all n-tuples over GF(2). Let us form the following n n-tuples :

$$e_0 = (1\ \ 0\ \ 0\ \ldots\ldots\ 0)$$
$$e_1 = (0\ \ 1\ \ 0\ \ldots\ldots\ 0)$$
$$e_2 = (0\ \ 0\ \ 1\ \ldots\ldots\ 0)$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$e_{n-1} = (0\ \ 0\ \ 0\ \ldots\ldots\ 1)$$

where tuple e_i has only one non-zero component at i^{th} position. Then every n-tuple $(a_0, a_1, a_2, \ldots\ldots\ldots, a_{n-1})$ in V_n can be expressed as a linear combination of $(e_0, e_1, e_2, \ldots\ldots\ldots, e_{n-1})$ as follows:

$$(a_0, a_1, a_2, \ldots\ldots\ldots\ a_{n-1}) = a_0e_0 + a_1e_1 + a_2e_2 + \ldots\ldots + a_{n-1}e_{n-1}$$

Therefore, $(e_0, e_1, e_2, \ldots\ldots\ldots, e_{n-1})$ span the vector space V_n of n-tuples over GF(2). Also it can be seen that $(e_0, e_1, e_2, \ldots\ldots\ldots, e_{n-1})$ is linearly independent.

Hence they form the basis of V_n and dimension of V_n is n.

1.2.6 Inner Product

Definition 1.6

Let $u = (u_0, u_1, u_2, \ldots\ldots\ldots, u_{n-1})$ and $v = (v_0, v_1, v_2, \ldots\ldots\ldots, v_{n-1})$ be two n-tuples in v_n.

The inner product (dot product) between u and v is defined as

$$u.v = u_0 \cdot v_0 + u_1 \cdot v_1 + \ldots\ldots\ldots\ldots + u_{n-1} \cdot v_{n-1}$$

where $u_i \cdot v_i$ and $u_i \cdot v_i + u_{i+1} \cdot v_{i+1}$ are carried out in modulo-2 multiplication and addition. Hence the inner product is a scalar in GF(2). If $u \cdot v = 0$ then u and v are orthogonal to each other.

Let S be the k dimensional subspace of vector space V_n and let S_d be the set of vector in V_n such that for any $u \in S$ and $v \in S_d$, $u \cdot v = 0$. The set S_d contains at least the all-zero n-tuples (0 = (0, 0, 0, ..., 0), since for any $u \in S$, and $u.0 = 0$. Thus S_d is non-empty. For any element $a \in GF(2)$ and $v \in S_d$,

$$av \in \begin{cases} 0 & if \ a = 0 \\ v & if \ a = 1 \end{cases}$$

Therefore, av is also in S_d. Let v and w be any two vectors in S_d. For any vector u in S, $u.(v + w) = u \cdot v + u \cdot w = 0 + 0 = 0$. This says that $v + w$ is also orthogonal to u . Consequently, $v + w$ is a vector in S_d. This proves that S_d is a subspace of vector space V_n. This subspace S_d is called the null (or dual) space of S and vice-versa.

Theorem 1.1

Let S be k dimensional space of the vector space V_n of all n-tuples over GF(2). The dimension of its null space S_d is $n - k$, In other words dim(S) + dim(S_d) = n.

1.2.7 Matrices

A $k \times n$ matrix over GF(2) (or over any other field) is a rectangular array with k rows and n columns.

$$G = \begin{bmatrix} g_0 \\ g_1 \\ g_{2-} \\ \cdots\cdots \\ g_{k-1} \end{bmatrix} = \begin{bmatrix} g_{00} & g_{01} & g_{02} & \cdots\cdots & g_{0(n-1)} \\ g_{10} & g_{11} & g_{12} & \cdots\cdots & g_{1(n-1)} \\ g_{20} & g_{21} & g_{22} & \cdots\cdots & g_{2(n-1)} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ g_{(k-1)0} & g_{(k-1)1} & g_{(k-1)2} & \cdots\cdots & g_{(k-1)(n-1)} \end{bmatrix} \qquad (5)$$

where each g_{ij} with $0 \le i < k$ and $0 \le j < n$ is an element from the binary field GF(2). If the k ($k < n$) rows of G are linearly independent, 2^k linear combinations of these rows form a k dimensional subspace of vector space V_n of all n-tuples over GF(2). This subspace is called a row space of G.

Let S be the row space of a $k \times n$ matrix G over GF(2) whose k rows (g_0, g_1, g_2, ..., g_{k-1}) are linearly independent. Let S_d be the null space of S.

Then the dimension of S_d is $n - k$. Let (h_0, h_1, h_2, ..., h_{n-k-1}) be $n - k$ linearly independent vectors of S_d. Clearly, these vectors span S_d. An $(n - k) \times n$ matrix H may be formed using (h_0, h_1, h_2, ..., h_{n-k-1}) as rows :

$$H = \begin{bmatrix} h_0 \\ h_1 \\ h_2 \\ \cdots\cdots \\ h_{n-k-1} \end{bmatrix} = \begin{bmatrix} h_{00} & h_{01} & h_{02} & \cdots\cdots & h_{0(n-1)} \\ h_{10} & h_{11} & h_{12} & \cdots\cdots & h_{1(n-1)} \\ h_{20} & h_{21} & h_{22} & \cdots\cdots & h_{2(n-1)} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ h_{(n-k-1)0} & h_{(n-k-1)1} & h_{(n-k-1)2} & \cdots\cdots & h_{(n-k-1)(n-1)} \end{bmatrix}$$

The row space of H is S_d. For each row $g_i \in G$ and $h_j \in H$, the inner product between g_i and h_j must be zero (i.e. $g_i \cdot h_j = 0$). Since row space (S) of G is the null space of the row space (S_d) of H, we call S the null space (dual space) of H. The result could be expressed by the following theorem :

Theorem 1.2

For any $k \times n$ matrix over GF(2) with k linearly independent rows, there exists an (n – k) × n matrix H over GF(2) with $n - k$ linearly independent rows such that for any row $g \in G$ and any row $h_j = H$, $g_i \cdot h_j = 0$. The row space of G is the null space of H, and vice versa.

1.3 Linear Block Codes

In this unit, the output of information source is thought of a sequence of binary digits '0' and '1'. In block coding, this sequence is divided into message blocks of some fixed length (say k), denoted by u. Therefore, total no of possible distinct messages is 2^k. The encoder transforms each of these message u into a binary n-tuple v where $n > k$. This binary n -tuple v is referred to as the code word(code vector) of the message u. Since code word must be distinct, the total no of code word also should be 2^k. Therefore, a one-to-one correspondence should exists between message u and code word v. This set of 2^k code words is known as block code. To reduce the encoding complexity a desirable structure of block code is linearity.

Definition 1.7

A block code of length n and 2^k code wards is called linear (n, k) code if and only if 2^k code words is k dimensional subspace of the vector space of all n-tuples over the field GF(2).

Since (n, k) linear code C is a k dimensional subspace of the vector space V_n of all the binary n-tuples, there are k linearly independent code words in C . Table 2.3 shows linear code block with $n = 7$ and $k = 4$. It can be checked easily that sum of any two code words in this is also a code word.

Messages	Code words
(0 0 0 0)	(0 0 0 0 0 0 0)
(1 0 0 0)	(1 1 0 1 0 0 0)
(0 1 0 0)	(0 1 1 0 1 0 0)
(1 1 0 0)	(1 0 1 1 1 0 0)
(0 0 1 0)	(1 1 1 0 0 1 0)
(1 0 1 0)	(0 0 1 1 0 1 0)
(0 1 1 0)	(1 0 0 0 1 1 0)
(1 1 1 0)	(0 1 0 1 1 1 0)
(0 0 0 1)	(1 0 1 0 0 0 1)
(1 0 0 1)	(0 1 1 1 0 0 1)
(0 1 0 1)	(1 1 0 0 1 0 1)
(1 1 0 1)	(0 0 0 1 1 0 1)
(0 0 1 1)	(0 1 0 0 0 1 1)
(1 0 1 1)	(1 0 0 1 0 1 1)
(0 1 1 1)	(0 0 1 0 1 1 1)
(1 1 1 1)	(1 1 1 1 1 1 1)

Table 1.3

It is always possible to find the k linearly independent code words (g_0, g_1, g_2, ..., g_{k-1}) such that every v in C is a linear combination of these k code words, that is,

$$v = u_0 g_0 + u_1 g_1 + u_2 g_2 + ... + u_{k-1} g_{k-1} \qquad (5)$$

where u_i = 0 or 1 for $0 \le i < k$,

1.3.1 Generator Matrix

Now let us arrange this k linearly independent code words as the rows of a k × n matrix as follows.

$$G = \begin{bmatrix} g_0 \\ g_1 \\ g_2 \\ .... \\ g_{k-1} \end{bmatrix} = \begin{bmatrix} g_{00} & g_{01} & g_{02} & ..... & g_{0(n-1)} \\ g_{10} & g_{11} & g_{12} & ..... & g_{1(n-1)} \\ g_{20} & g_{21} & g_{22} & ..... & g_{2(n-1)} \\ .... & ..... & ..... & ..... & ..... \\ g_{(k-1)0} & g_{(k-1)1} & g_{(k-1)2} & ..... & g_{(k-1)(n-1)} \end{bmatrix} \qquad (6)$$

where $g_i = (g_{i0}, g_{i1}, g_{i2}, \ldots\ldots\ldots, g_{i(n-1)})$ for $0 \le i < k$, If u = (u_0, u_1, u_2, ……, u_{k-1}) is the message to be encoded, the corresponding code word can be given as :

$$v = u.G = (u_o, u_1, u_2, \ldots\ldots\ldots, u_{k-1}) \begin{bmatrix} g_0 \\ g_1 \\ g_2 \\ \ldots\ldots \\ g_{k-1} \end{bmatrix} = u_0 g_0 + u_1 g_1 + u_2 g_2 + \ldots\ldots\ldots + u_{k-1} g_{k-1} \quad (7)$$

It is clear that the rows of generator matrix G generates (or span) the (n, k) linear code C. For this reason G is known as generator matrix for C.

Example 1.6

The (7,4) linear code is given below as a generator matrix G. Let us consider a message u = (1 1 0 1) . The code word corresponding to the message u needs to be determined.

$$G = \begin{bmatrix} g_0 \\ g_1 \\ g_2 \\ g_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Using the equation (7), the code word would be $u.\ G$

= 1(1 1 0 1 0 0 0) + 1(0 1 1 0 1 0 0) + 0 (1 1 1 0 0 1 0)
+ 1(1 0 1 0 0 0 1)

= (1 1 0 1 0 0 0) + (0 1 1 0 1 0 0) + (0 0 0 0 0 0 0)
+ (1 0 1 0 0 0 1)

= (0 0 0 1 1 0 1)

1.3.2 Linear Systematic Block Code

In example 1.6, it can be observed that the last four digits of the code word v (0 0 0 1 1 0 1) is same as the message u = (1 1 0 1). The first three digits are redundant and could be written as linear sums of information digits (message) (refer example 1.7). These redundant digits are known as parity-check digits. Therefore, a code word of length n could be divided into following two parts.

1. Message part : Consist of k unaltered information digits
2. Checking part : Consist of k_{n-k} parity – check digits.

Redundant checking part $\leftarrow n-k$ digits $\rightarrow$	Message part $\leftarrow k$ digits $\rightarrow$

Figure 1.4

The linear block code with this structure (Figure 1.4) is referred to as linear systematic block code.

A linear systematic *(n, k)* code is completely specified by $k \times n$ matrix G of the following form :

$$G = \begin{bmatrix} g_0 \\ g_1 \\ g_2 \\ \cdots \\ g_{k-1} \end{bmatrix} = \left[\begin{array}{cccccc|cccccc} P_{00} & P_{01} & P_{02} & \cdots & P_{0(n-k-1)} & & 1 & 0 & 0 & \cdots & 0 \\ P_{10} & P_{11} & P_{12} & \cdots & P_{1(n-k-1)} & & 0 & 1 & 0 & \cdots & 0 \\ P_{20} & P_{21} & P_{22} & \cdots & P_{2(n-k-1)} & & 0 & 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & & \cdots & \cdots & \cdots & \cdots & \cdots \\ P_{(k-1)0} & P_{(k-1)1} & P_{(k-1)2} & \cdots & P_{(k-1)(n-k-1)} & & 0 & 0 & 0 & 0 & 0 \end{array}\right] \quad (8)$$

Where $P_{ij} = 0$ or 1. Let I_k denote the $k \times k$ identity matrix. Then $G = [P\ I_k]$. Let $\mathbf{u} = (u_0, u_1, u_2, \ldots\ldots, u_{k-1})$ be the message to be encoded. The corresponding code word is

$$v = (v_0, v_1, v_2, \ldots\ldots\ldots\ldots, v_{n-1})$$

$$= (u_0, u_1, u_2, \ldots\ldots, u_{k-1}).G \quad (9)$$

Using (8) and (9), the components of v could be written as following :

$$v_{n-k-i} = u_i \quad for\ 0 \le i < k \quad (10)$$

$$v_j = u_0 p_{oj} + u_1 p_{1j} + \ldots\ldots + u_{k-1} p_{(k-1)j} \quad for\ 0 \le j < n-k \quad (11)$$

Equation (10) shows that the rightmost k digits of code word is identical to information digits and Equation (11) leftmost $n - k$ redundant digits are linear sum of information digits. Equation (11) is known as parity-check equation of the code.

Example 1.7

The matrix given in example 1.6 is in the systematic form. The equation of the digits of the code word needs to be determined.

Let $u = (u_0, u_1, u_2, u_3)$ be the message and $v = (v_0, v_1, v_2, v_3, v_4, v_5, v_6)$ be the code word.

$$G = \begin{bmatrix} g_0 \\ g_1 \\ g_2 \\ g_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Then, we know that $v = u$. G

$$\Rightarrow (v_0, v_1, v_2, v_3, v_4, v_5, v_6) = (u_0, u_1, u_2, u_3).\begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Solving the above equation, we get,

$$v_6 = u_3,\ v_5 = u_2,\ v_4 = u_1,\ v_3 = u_0$$

$$v_2 = u_1 + u_2 + u_3,$$

$$v_1 = u_0 + u_1 + u_2$$

$$v_0 = u_0 + u_2 + u_3$$

We can check easily that if u = (1 0 1 1) then v can be derived using the above equations as (1 0 0 1 0 1 1)

1.3.3 Conversion of Generator Matrix into Systematic Form

Elementary Row Operation :

An elementary row operation on a binary matrix (elements are from GF(2)) of replacing a row of the matrix with the sum of that row and any other row.

If we have a generator matrix G for a linear code L , all other generator matrices for L can be obtained by applying a sequence of elementary row operations to G.

Example 1.8

Let $G = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}$ is a generator matrix of the

Figure 1.5

linear code {0000, 0011, 0110, 1100, 0101, 1111, 1010, 1001}.

Matrix G consists of four column (figure 1.5), out of these C_0 and C_3 are already the columns of identity matrix I_3. Therefore, we need to create a

$$G = \begin{matrix} C_0 & C_1 & C_2 & C_3 \end{matrix}\begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}$$

column of the form $(0\ 1\ 0)^T$. To do that, we begin by applying the elementary row operation of replacing the second row of G by the sum of the first and second rows

$(R_3 \rightarrow R_2 + R_3)$ to give

$$G_1 = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}$$

Now after rearranging the columns we can easily get the generator matrix G in systematic form.

$$G_1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

1.3.4 Parity-Check Matrix

There is another useful matrix associated with every linear block code. As per the Theorem 1.2 stated in section 1.2.7, For any $k \times n$ matrix G over GF(2) with k linearly independent rows, there exists an (n – k) × n matrix H over GF(2) with $n - k$ linearly independent rows such that any vector in the row space of G is orthogonal to rows of H and vice versa. Therefore, we can describe the (n, k) linear code generated by G in an alternate ways as follows: An n - tuple v is a code word if and only if v. $H^T = 0$. The matrix H is parity check matrix of the code. The 2^{n-k} linear combination of the rows of H form the $(n - k, n)$ linear code C_d . This code is the null space of the (n, k), linear code C generated by matrix G (Refer section 1.2.6). C_d is called dual code of C.

Example 1.9

Let us define the parity check matrix for the linear code (7,4) generated by G in example 1.7.

We have seen following seven check equation for the code in Example 1.7.

$v_0 = u_3$ *(12.a)*, $v_5 = u_2$ *(12.b)*, $v_4 = u_1$ *(12.c)*, $v_3 = u_0$ *(12.d)*

$v_2 = u_1 + u_2 + u_3$ *(12.e)*, $v_1 = u_0 + u_1 + u_2$ *(12.f)*, $v_0 = u_0 + u_2 + u_3$ *(12.g)*

Replacing the value of u_i 's in equation (12.e, f, g) from equation (12.a, b, c, d) we get new equations as follows :

$v_2 = u_1 + u_2 + u_3$
$\Rightarrow v_2 = v_4 + v_5 + v_6$ (13.a) | $v_1 = u_0 + u_1 + u_2$ $\Rightarrow v_1 = v_3 + v_4 + v_5$ (13.b) | $v_0 = u_0 + u_2 + u_3$ $\Rightarrow v_0 = v_3 + v_5 + v_6$ (13.c)

Now consider the equation (13.a)

$v_2 = v_4 + v_5 + v_6$.

$\Rightarrow v_2 + v_2 = v_4 + v_5 + v_6 + v_2$ (Adding v_2 in both the side)

$\Rightarrow v_2 + v_4 + v_5 + v_6 = 0$ ($\because v_2 + v_2 = 0$ in modulo - 2 addition where v_2 = GF(2))

Applying the above method for equation (13.b) and (13.c) also, we get new set of equation as follows.

$v_0 + v_3 + v_5 + v_6 = 0$

$v_1 + v_3 + v_4 + v_5 = 0$

$v_2 + v_4 + v_5 + v_6 = 0$

Now if we will write this equation in matrix form $v \cdot H^T = 0$ as follows :

$$(v_0, v_1, v_2, v_3, v_4, v_5, v_6).\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix} = 0$$

Therefore, the parity-check matrix

$$H = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}$$

This shows that parity-check matrix takes the form $H = I_{n-k} \; P^T$.

Example 1.10

The parity check matrix H can be generated from the generator matrix G (in systematic form) of Example 1.8.

$$G = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Here G is in the form of $[P\ I_3]$, where $P = (1\ 1\ 0)^T$. We know that, the matrix $H = [I_{n-k}\ PT]$ where $n = 4$, $k = 3$, Therefore, $H = [I_1, PT] = [1\ 1\ 1\ 0]$. It can also be checked that $G \cdot H^T = 0$.

1.3.5 Minimum Distance of a Block Code

The error correcting and error detecting capability of a code depends on the parameter known as minimum distance. To define minimum distance let us introduce the term Hamming Weight and Hamming Distance.

Hamming Weight

If $v = (v_0, v_1, \ldots\ldots, v_{n-1})$ be a binary n – tuple, then the Hamming Weight wt

(v) is defined as no of nonzero digits in *v*. For example the Hamming Weight of (1 1 0 0 1 0 1) is 4.

Hamming Distance

If *v* and *w* be two *n* – tuples, then Hamming Distance d *(v, w)* between *v* and *w* is defined as the no of places where they differ. For example, the Hamming Distance between *v* = (1 1 0 0 1 0 1) and *w* = (0 0 0 1 0 0 1) is 4. They differ in zeroth, first, third, fourth places. Hamming Distance is a metric on the vector space v_n over GF(2). Therefore, for any three vectors *v*, *w* and x ∈ v_n obeys the following conditions.

(1) $d(v, w) \geq 0$

(2) $d(v, w) = d(w, v)$

(3) $d(v, w) + d(w, x) \geq d(v, x)$

The important point to be notices that, Hamming weight of the modulo-2 sum of *v*, *w* actually gives the Hamming distance between *v* and *w*. For example,

v + *w* = (1 1 0 0 1 0 1) + (0 0 0 1 0 0 1) = (1 1 0 1 1 0 0)

Therefore,

$$d(v,w) = wt\ (v + w). \qquad (14)$$

Since for any linear code, the sum of two code words gives another code word, so equation (14) can be written as:

$d(v, w) = wt(x)\ :\ x \in C$ where *v*, *w* and *x* are code words of linear code *C* (15)

Minimum Distance

The minimum Hamming distance between two distinct code of any code *C* is known as minimum distance (d_{min}) of the code *C* .

$$d_{min} = \min\{d(v, w)\ :\ v, w \in C, v \neq w\} \qquad (16)$$

$$= \min\{wt(v + w)\ :\ v, w \in C, v \neq w\} \quad \text{(from equation) (14)}$$

$$= \min\{wt(x)\ :\ x \in C, x \neq 0\} \text{ (from equation} \qquad (15)$$

Let us introduce a parameter min W_{min}= $\min\{wt(x)\ :\ x \in C, x \neq 0\}$ which is called the minimum weight of linear code *C*. The above result is summarized in the following theorem.

Theorem 1.3

The minimum distance of a linear block code is equal to the minimum weights of its non-zero code words.

Therefore, the minimum distance of code (7, 4) given in table 1.3 is 3. There is another way to calculate the minimum distance of a linear block code from the parity check matrix using following theorem.

Theorem 1.4

Let C be an (n, k) linear code with parity check matrix H. For each code vector of Hamming Weight l, there exist l columns of H such that the vector sum of these l columns is equal to zero vector. Conversely, if there exist l columns of H whose vector sum is zero vector, there exist a code vector of Hamming Weight l in C .

Corollary 1.4.1

Let C be a linear block code with parity check matrix H. The minimum weight of C is equal to the smallest no of columns of H that sum to zero.

The parity check matrix H for (7,4) linear block code is following :

$$H = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}$$

Since no two columns are exactly same in H ,so sum of two columns cannot be zero. Therefore, the minimum distance of C is > 2. However, the zeroth, fourth and fifth column sum to zero. Thus the minimum distance of code C is 3.

1.3.6 Error Detecting Capability of Block Code

Theorem 1.5

A linear block code C of minimum distance d_{min}, detects up to t errors if and only if its minimum distance is greater than t i.e. $d_{min}(C) > t$.

Proof

Let v° be the code word transmitted over a noisy channel from the block code) $C(n,k)$ and r be the received word which differs from v° in at most t places. Therefore,

$$d(v^\circ, r) < t$$

$$\Rightarrow d(v^\circ, r) = t - k_1 \text{ where } k_1 \geq 0 \qquad (17)$$

To detect that r is in error it is sufficient to ensure that r does not correspond to any of the valid code words v_i, that is, $d(r, v_i) > 0$ for $0 \leq i < 2^k$.

Using the triangle inequality, we have that for any code word v_i :

$$d(v^\circ, r) + d(r, v_i) \geq d(v^\circ, v_i)$$

$$\Rightarrow d(r, v_i) \geq d(v^\circ, v_i) - d(v^\circ, r) \qquad (18)$$

Given that $d_{min}(C) > t \Rightarrow d(v^\circ, v_i) > t \quad \forall i$

$$\Rightarrow d(v^\circ, v_i) = t + k_2 \text{ where } k_2 > 0 \qquad (19)$$

Replacing the value of $d(v°, r)$ and $d(v°, v_i)$ from equation (17) and (19), in equation (18) we get,

$$d(r, v_i) \geq (t + k_2 - (t - k_1) \Rightarrow d(r, v_i) \geq k_1 + k_2$$

$$\Rightarrow d(r, v_i) > 0 \ (\text{since } k_1 + k_2 > 0) \qquad (20)$$

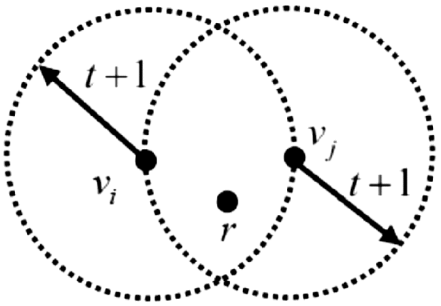


Figure 1.6

Figure 1.6 depicts the two closest code words, v_i and v_j, at a distance of $(t + 1)$ (i.e., $d_{\min}(C) = t + 1$). Here the value of is $k_2 = 1$. The hypersphere of distance $(t + 1)$ drawn around each code word may touch another code word but no code word falls within the hypersphere of another code word since this would violate $d_{\min}(C) = t + 1$. Clearly any received word r of distance t from any code word will fall within the hypersphere of radius $t + 1$ from one of more code words and hence be detectable since it can never be mistaken for a code word.

1.3.7 Error Correcting Capability of Block Code

Theorem 1.6

A linear block code C of minimum distance $d_{\min}$, corrects up to t errors if and only if its minimum distance is greater than t i.e. $d_{\min}(C) > 2t$.

Proof : Let $v°$ be the code word transmitted over a noisy channel from the block code $C(n, k)$ and r be the received word which differs from $v°$ in at most t places.

Therefore,

$$d(v°, r) < t$$

$$\Rightarrow d(v°, r) = t - k_1 \text{ where } k_1 \geq 0 \qquad (21)$$

To detect that r is in error and ensure that the Hamming distance decoding rule uniquely identify $v°$ as a corrected code, it is sufficient to ensure $d(r, v_i) > d(r, v°)$ *for* $0 \leq i < 2^k$. This means the closest correct code word of r is $v°$ which was originally transmitted. Using the triangle inequality, we have that for any code word v_i :

$$d(v^\circ, r) + d(r, v_i) \geq d(v^\circ, v_i)$$

$$\Rightarrow d(r, v_i) \geq d(v^\circ, v_i) - d(v^\circ, r) \qquad (22)$$

Given that, $d_{\min}(C) > 2t$

$$\Rightarrow d(v^\circ, v_i) > 2t \quad \forall i$$

$$\Rightarrow d(v^\circ, v_i) > 2t + k_2 \text{ where } k_2 > 0 \qquad (23)$$

Now, using equation (22), we can write,

$$d(r, v_i) - d(r, v^\circ) \geq d(v^\circ, v_i) - d(v^\circ, r) - d(r, v^\circ)$$

$$= d(v^\circ, v_i) - 2d(v^\circ, r) \quad (\text{since } d(v^\circ, r) = d(r, v^\circ)$$

$$= (2t + k_2) - 2(t - k_i) \quad (\text{using equation (21) and (23)}$$

$$= k_2 + 2k_1 > 0$$

Therefore, finally we get,

$$d(r, v_i) - d(r, v^\circ) > 0$$

$$\Rightarrow d(r, v_i) > d(r, v^\circ) \qquad (24)$$

Figure 1.7 depicts the two closest code words, v_i and v_j, at a distance of $(2t + 1)$ (i.e.. $d_{\min}(C) = 2t + 1$). Here the value of is $k_2 = 1$. The hyperspheres of distance t drawn around each code word do not touch each other. Clearly any received word r of distance $\leq t$ from any code word will only fall within the hypersphere of radius t from that code word and hence can be corrected.

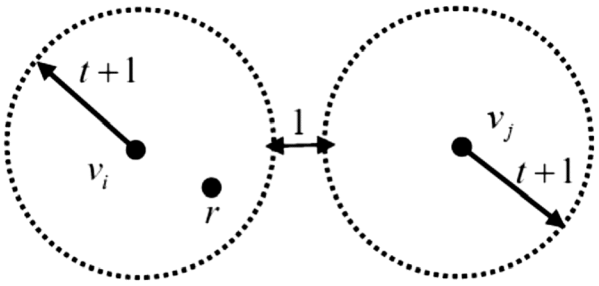


Figure 1.7

Example 1.11

Let us analyze the error detecting and correcting capability of (3, 2) linear block code given in the following table.

Messages (k = 2)	Code words (n = 3)
00	000
01	001
02	011
03	111

Table 1.4

For (3,2) code given in the table 1.4, d_{min} = 1.

As d_{min} –1 = 0, this code cannot detect any error.

Example 1.12

Let us analyse the error detecting and correcting capability of (6,3) linear block code given in the following table.

Messages (k = 3)	Code words (n = 6)
000	000000
001	001110
010	010101
011	011011
100	100011
101	101101
110	110110
111	111000

Table 1.5

For (3,2) code given in the table 1.5, d_{min} = 1.

As d_{min} – 1 = 0, this code cannot detect any error.

By computing the distance between all pairs of distinct code words, requiring $\binom{8}{2} = 28$ computations of the Hamming distance, we find that d_{min} = 3. This means the code can detect d_{min} – 1 = 2 bit error and can detect $\left[\frac{d_{min}-1}{2}\right] = 1$ bit error.

1.3.8 Hamming Code

Hamming codes are the linear code which have been used in error control in digital communication and data storage system. For any positive integer $m > 3$, there exists a Hamming code with following parameters mentioned in table 1.6 :

Parameter Name(Symbol)	Expression
Code Length (n)	$(2^m - 1)$
Number of information bit (k)	$(2^m - m - 1)$
Number of parity-check bit ($n - k$)	(m)
Minimum Hamming Distance (d_{min})	3
Error correcting capability(t)	$\left[\frac{d_{min} - 1}{2}\right] = 1$

The parity check matrix consists of all non-zero m – tuples.

Total no of m tuples formed by m bits = 2^m. Out of all these m – tuples, 1 tuple has all zero's. Therefore, no of non-zero m – tuples = $2^m - 1$.

The parity check matrix in systematic form :

$$H = \left[I_m : P^T \right] \tag{25}$$

where I_m is the identity matrix and contains m columns of weight 1, and P^T consists of remaining $(2^m - m - 1)$ columns of m – tuples of weight 2 or more. For $m = 3$, The Hamming Code has length $n = 2^3 - 1 = 7$.

Therefore, $k = n - m = 7 - 3 = 4$.

Now the parity check matrix can be constructed by all non-zero 3 tuples as columns in the form of equation (25).

$$H = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}$$

The columns of P^T can be arranged in any order without affecting the distance property and weight distribution of the code.

Now it is also easy to write the generator matrix for the (7, 4) code using following formula.

$$G = [P : I_k]$$

$$G = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

The minimum distance of (7,4) code can be easily determined form H. It can be checked that at least three (e.g. the zeroth, first and third) columns of H need to be added to get 0.Therefore, using corollary

1.4.1 It can be concluded that the $d_{mm} = 3$ and the code can correct $\left[\frac{d_{min}-1}{2}\right] = 1$ bit error.

Error correction by Hamming Code

The parity check matrix columns of Hamming code can be rearranged such that column in position i represents the integer i. Then the parity check matrix will be :

$$H = \begin{array}{c} \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{matrix} \\ \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix} \end{array} \qquad (26)$$

Here the column $T(x, y, z)$ represents the number $x(2^0) + y(2^1) + z(2^2)$. Let us also consider r be the received code word and v be the transmitted code word. Now we calculate the value of r. H^T which is known as syndrome. If syndrome (r . H^T) is 0, then $r = v$ since for any valid code word v, we can write v . $H^T = 0$ (refer section 1.3.5). Otherwise, for any non-zero syndrome it can be concluded a single bit error has occurred while passing through the noisy channel and received word r is no longer a valid code. Now in Hamming code the non-zero value of r . H^T gives the bit position of the single bit error in r. Once the error bit position is identified, it can be corrected.

Let (0 1 0 1 0 1 0) be a code word in (7, 4) Hamming code. Suppose a single bit error occurred in second bit and the received word r becomes (0 0 0 1 0 1 0). Using the parity check matrix H from equation (26) we calculate the syndrome (r . H^T) as follows :

$$r.H^T = [0\ 0\ 0\ 1\ 0\ 1\ 0] . \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} = [0\ 1\ 0]$$

The number represented by the syndrome is 2 hence the error is in second bit position. Therefore, the estimated code word is (0 1 0 1 0 1 0).

1.4 Cyclic Codes

Cyclic codes are important subclass of linear code. These codes are useful for following two reason.

(1) Encoding and syndrome (refer section 1.3.9) computations are easy to implement.
(2) It has considerable inherent algebraic structure which makes it possible to use it in various practical methods.

Before detailing the topic, it is important to mention few algebraic properties of polynomials.

1.4.1 Rings of Polynomial

Polynomial of degree n : A polynomial of degree n with coefficients in a ring R is an element of R^{n+1}. Therefore, a $n + 1$ tuple $(r_0, r_1, r_2, ..., r_n)$ can represent the polynomial $r_0 + r_1 x + r_2 x + ... + r_n x^n$.

Example 1.13

The following polynomials over the ring of integers, Z.

$p_1(X) = 1 + 2X + 3X^2 + 4X^3 + 5X^4$ and $p_2(X) = 1 + 2X - 4X^2 + 6X^3 - 2X^4$

Adding these polynomials means adding the coefficients and multiplying by an integer means multiplying the coefficient :

$$p_1(X) + p_2(X) = (1-1) + (2-2)X + (3+4)X^2 + (4-6)X^3 + (5-2)X^4$$

$$= 7X^2 - 2X^3 + 3X^4$$

$$6p_1(X) = 6 + 12X + 18X^2 + 24X^3 + 30X^4$$

Example 1.14

Polynomials also can be defined over the binary field GF(2). Following Two polynomials over GF(2) are added using modulo-2 addition.

$p_1(X) = 1 + X + X^2 + X^3 + X^4$ and $p_2(X) = 1 + X + X^5$

$$p_1(X) + p_2(X) = (1+1) + (1+1)X + X^2 + X^3 + X^4 + X^5$$

$$= X^2 + X^3 + X^4 + X^5$$

Multiplying polynomials by elements of GF(2) is trivial : we either multiply by 0 to get the zero polynomial, or multiply by 1, which gives the same polynomial.

The set of polynomials of degree n over GF(2) is a vector space V^{n+1}. The elements of V^{n+1} with polynomials can be identified by matching the components of the vector with coefficients of the polynomial, matching the leftmost bit with the constant term. So, for $n + 1 = 8$, 11100101 is matched with $p_1(X) = 1 + X + X^2 + X^5 + X^7$.

Table 1.7 shows the vector and their corresponding polynomial for different values of n.

Value of n	Vector	Polynomial	Vector	Polynomial
1	00	0	01	X
	10	1	11	$1 + X$
	000	0	001	X^2
2	100	1	101	$1 + X^2$
	010	X	011	$X + X^2$
	110	1 + X	111	$1 + X + X^2$

Table 1.7

Polynomial over GF(2) also can be multiplied in the usual way. At the time of adding the intermediate result modulo-2 addition should be applied.

$$(1+X+X^2)(1+X) = 1+X+X^2+X+X^2+X^3$$

$$= 1+(X+X)+(X^2+X^2)+X^3 = 1+X^3$$

In the vector form the above result can be written as $111 \times 11 = 1001$. If the ring of coefficient is a field, then the division operation for polynomial over that field can also be defined by synthetic division method. The following example.

Example 1.15

Consider $X + 4$ and $X^3 + 2X^2 - 5X + 15$ to be polynomials with coefficients in the field of real numbers. We start the synthetic division by setting them out as follows :

$$X+4 \overline{)\, X^3+2X^2-5X+15} \quad \Big| \; X^2-2X+3$$

$$\begin{array}{r} X^3+4X^2 \\ \hline -2X^2-5X \\ -2X^2-8X \\ \hline 3X+15 \\ 3X+12 \\ \hline 3 \end{array}$$

This shows that $X^3+2X^2-5X+15 = (X+4)(X^2-2X+3)+3$

Example 1.16

We can perform synthetic division in the same way when the coefficients of the polynomials come from the binary field GF(2). In this case, addition and subtraction are the same operation. To justify this argument let us recall the module -2 addition as a clock operation mentioned in the example 1 of section 1.2.1. It is seen that adding 1 means clockwise rotation. Since subtraction is negative addition, Subtracting 1 means anticlockwise rotation. Figure 1.8 clearly shows that addition and subtraction in modulo-2 arithmetic gives same result.

To divide $X^5 + 1$ by $X^2 + X$ we start by setting the polynomials out in the usual way:

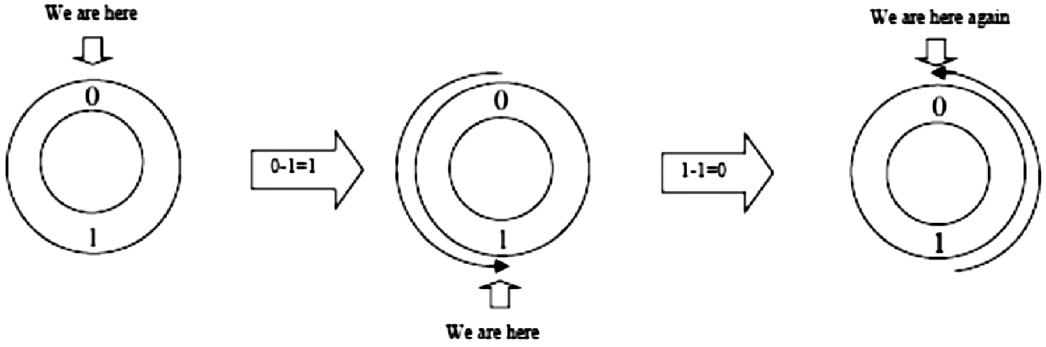


Figure 1.8

$$
\begin{array}{r}
X^3 + X^2 + X + 1 \\
X^2 + X \overline{)\, X^5 \qquad\quad + 1} \\
\underline{X^5 + X^4} \\
X^4 \\
\underline{X^4 + X^3} \\
X^3 \\
\underline{X^3 + X^2} \\
X^2 \\
\underline{X^2 + X} \\
X + 1
\end{array}
$$

This shows that $X^5 + 1 = (X^2 + X)(X^3 + X + 1) + (X + 1)$

If two polynomials of degree n are multiplied together, the result is usually a polynomial of degree $2n$. This means that the set of polynomials of degree n can't be a ring under the

operations of polynomial addition and multiplication, as multiplying two polynomials in this set will give a polynomial that is not in this set. To get rid of this problem the multiplication operation is defined differently to create a new ring structure.

The operation of multiplication modulo the polynomial $X^n + 1$ gives products which are polynomials of degree $n - 1$ or less. The set of polynomials of degree less than n forms a ring with respect to addition and multiplication modulo $X^n + 1$; this ring will be denoted $B_n(X) / (X^n + 1)$.

Theorem 1.7 (Remainder theorem for polynomial over GF(2)) :

The remainder that is obtained when a polynomial in $B_n(X)$ is divided by $X^n + 1$ can be calculated by replacing X^n with 1 in the polynomial. This operation "wraps" the powers of X around from X^n to $X^0 = 1$.

Example 1.17

The remainder of $X^3 + X^2 + X$ when divided by $X^3 + 1 = 1 + X^2 + X$.

The remainder of $X^4 + X^3 + X^2$ when divided by

$$X^3 + 1 = 1X + 1 + X^2 = 1 + X + X^2$$

The remainder of $X^7 + X^6 + X^5$ when divided by

$$X^4 + 1 = 1X^3 + 1X^2 + 1X = X + X^2 + X^3$$

We can now compute the multiplication tables of the rings $B_n(X) / (X^n + 1)$.

The elements of $B_2(X) / (X^2 + 1)$ are the polynomials 0, 1, X and $1 + X$.

Multiplication by 0 and 1 is trivial. The other products are,

$$(X)(X) \text{ modulo } (X^2 + 1) = 1$$

$(X)(1 + X)$ modulo $(X^2 + 1) = X + X^2$ modulo $(X^2 + 1) = X + 1$

$(1 + X)(1 + X)$ modulo $(X^2 + 1) = 1 + X + X + X^2$ modulo $(X^2 + 1)$

$= 1 + X^2$ modulo $(X^2 + 1) = 1 + 1 = 0$

The multiplication table for the polynomial of $B_2(X) / (X^2 + 1)$ is given in Table 1.8.

	0	1	X	1 + X
0	0	0	0	0
1	0	1	X	1 + X

X	0	X	1	1 + X
1 + X	0	1 + X	1 + X	0

Table 1.8

Identifying the polynomial 0 with 00, 1 with 10, X with 01 and $1 + X$ with 11, the multiplication table becomes :

	00	10	01	11
0	00	00	00	00
1	00	10	01	11
X	00	01	10	11
X + 1	00	11	11	00

Table 1.9

Theorem 1.8

Multiplication by X in $B_n(X) / (X^n + 1)$ is equivalent to shifting the components of the corresponding vector in vector space V^n over GF(2) cyclically one place to the right.

Proof :

Let $p(X)$ be a polynomial in $B_n(X) / (X^n + 1)$,

$p(X) = b_0 + b_1X + b_2X^2 + \ldots\ldots\ldots\ldots + b_{n-1}X^{n-1}$ and the corresponding vector in $V^n = b_0b_1b_2, \ldots\ldots\ldots\ldots b_{n-1}$ where $b_i \in \{0,1\}$.

$Xp(X) = b_0X + b_1X^2 + b_2X^3 + \ldots\ldots\ldots\ldots + b_{n-1}X^n$

To find out $Xp(X)$ modulo $(X^n + 1)$, we need to replace X^n by 1. Therefore, $Xp(X)$ modulo $(X^n + 1) = b_0X + b_1X^2 + b_2X^3, \ldots\ldots\ldots\ldots + b_{n-2}X^{n-1} + b_{n-1}$.

$= b_{n-1} + b_0X + b_1X^2 + b_2X^3, \ldots\ldots\ldots\ldots + b_{n-2}X^{n-1}$

The vector corresponding to the polynomial $Xp(X)$ modulo $(X^n + 1)$ is $b_{n-1}b_0b_1b_2, \ldots\ldots\ldots\ldots b_{n-2}$ which is the result of cyclic shifting $b_0b_1b_2, \ldots\ldots\ldots b_{n-1}$ one place to the right.

Definition 1.8

A cyclic code is a linear code with the property that any cyclic shift of a code word is also a code word.

Consider the (4,2) linear code whose generator matrix is

$$G = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}$$

The linear code generated by G is $\{0000,1010,0101,1111\}$. Note that shifting 0000 cyclically gives 0000, shifting 1010 one place cyclically gives 0101, shifting 0101 one place cyclically gives 1010 and shifting 1111 cyclically gives 1111. So this is a cyclic code.

In polynomial notation, the code is $\{0,1+X^2,X+X^3,1+X+X^2+X^3\}$. A cyclic shift to the right can be accomplished by multiplying by X modulo $X^4 + 1$. Multiplying 0 by X gives 0, multiplying $1 + X^2$ by X gives $X + X^3$, multiplying $X + X^3$ by X (modulo $X^4 + 1$) gives $1+X^2$ and multiplying $1 + X + X^2 + X^3$ by X (also modulo $X^4 + 1$) gives $1 + X + X^2 + X^3$.

Theorem 1.9

A cyclic code contains a unique non-zero polynomial of minimal degree.

1.4.2 Generator Matrix

Generator Polynomial

The unique non-zero polynomial of minimal degree in a cyclic code is the generator polynomial of the code.

Theorem 1.10

If $g \in B_n(X)/(X^n+1)$ is the generator polynomial for some cyclic code, then every polynomial in the code can be generated by multiplying g by some polynomial in $B_n(X)/(X^n+1)$.

Proof :

Since multiplication by X modulo $(X^n + 1)$ has the effect of shifting a code word cyclically one place to the right, multiplying by X^k modulo $(X^n + 1)$ has the effect of shifting a code word cyclically k places to the right. It follows that the product g by X^k modulo $(X^n + 1)$ will be a code word for any $k \geq 0$.

Multiplying g by any polynomial is equivalent to multiplying g by various powers of X modulo $(X^n + 1)$ and adding the products together. Since the products are all code words, so is their sum.

To show that every code word can be generated in this way, note that if the degree of g is r, then the polynomials g, gX, gX^2 ....., gX^{n-r-1} form a basis for the code and hence that every code word can be generated by taking a linear combination of these basis elements.

If the generator of a cyclic code is $g(X) = g_0 + g_1X + g_2X^2 + \ldots\ldots\ldots\ldots + g_rX^{r-1}$ the fact that the polynomials g, gX, gX^2 ....., gX^{n-r-1} form a basis for the code means that the generator matrix of the code can be written in the form :

$$G = \begin{bmatrix} g_0 & g_1 & \ldots. & g_r & 0 & 0 & \ldots. & 0 \\ 0 & g_0 & \ldots. & g_{r-1} & g_r & 0 & \ldots. & 0 \\ 0 & 0 & \ldots. & \ldots. & \ldots. & \ldots. & \ldots. & \ldots. \\ 0 & \ldots. & \ldots. & \ldots. & \ldots. & \ldots. & \ldots. & \ldots. \\ 0 & \ldots. & 0 & g_0 & g_1 & \ldots. & g_r & 0 \\ 0 & \ldots. & 0 & 0 & g_0 & \ldots. & g_{n-1} & g_r \end{bmatrix} \qquad (27)$$

G is a cyclic matrix (each row is obtained by shifting the previous row one column to the right).

Example 1.18

The following code is a cyclic code.

0000000	1011100	0010110	0010111
1001011	1100101	1110010	0111001

The code word 1011100 corresponds to the polynomial $1 + X^2 + X^3 + X^4$, which is the polynomial of minimal degree and hence the generator polynomial.

The generator Matrix

$$G = \begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}$$

The generator matrix in systematic form $G = \begin{bmatrix} P_{k\times(n-k)} & I_{k\times k} \end{bmatrix}$ (refer to section 1.3.3) can be derived by interchanging columns $(C_0 \leftrightarrow C_4, C_2 \leftrightarrow C_5)$:

$$G = \begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

The parity check matrix $H = [I_{n-k} \quad P^T]$ (refer to section 1.3.5) could be expressed as :

$$H = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix}$$

1.5 Summary

Several key aspects coding theory with help of abstract and linear algebra are discussed in this unit. Learners can now appreciate why the encoding and decoding is required in the context of digital transmission. They can now differentiate between linear and cyclic code. How these codes are generated along with the parity check matrix. Learners can now explain how the polynomials of ring and vector spaces are important for different coding technique. They may use the concepts of this unit while specializing in information and coding theory in future course.

1.6 Exercises

1. Consider a linear block code C, with parity check matrix given by

$$H = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 \end{bmatrix}$$

What is (n, k) of C ?

2. Consider following binary code block, C

$$C = \{00000, 110011, 011101, 11111\}$$

Show that C can't be a linear block code? What are the code words those needs to be added in C so that C becomes a linear code words.

3. Consider a systematic (8.4) code whose parity check equations are

$$v_0 = u_1 + u_2 + u_3$$

$$v_1 = u_0 + u_1 + u_2$$

$$v_2 = u_0 + u_1 + u_3$$

$$v_3 = u_0 + u_2 + u_3$$

where u_0, u_1, u_2, u_3 are message digits and v_0, v_1, v_2, v_3 are parity-check digits. Find the generator and parity-check matrix for this code. Show analytically that the minimum distance of this code is 4.

4. Let H_1 be the parity matrix of an (n,k) linear code C that has both odd-andeven weight code vectors. Construct a new linear code C_1 with following parity-check matrix:

$$H_1 = \begin{bmatrix} 0 & & \\ 0 & & \\ \vdots & H & \\ 0 & & \\ 1 & 1 \quad 1 \ \ldots \ 1 & \end{bmatrix}$$

a. Show that C_1 is $(n + 1, k)$ linear code, C_1 is called an extension of C.

b. Show that every code vector of C_1 has even weight.

c. Show that C_1 can be obtained C from by adding an extra-parity-check digit, denoted v_∞, to the left of each code vector v as follows :

1) if v has odd weight, then $v_\infty = 1$, and

2) if v has even weight, then $v_\infty = 0$

The parity-check digit v_∞ is called the overall parity-check-digit.

5. Write down the following generator matrix in systematic form. Then write down the parity check matrix H. Also verify that $GH^T = 0$.

$$G = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

6. If C be a linear code with both even and odd weight code words then show that the number of even weight code words is equal to the number of odd weight code words.

7. Write down the following generator matrix in systematic form. Then write down the parity check matrix H. Also verify that $GH^T = 0$.

$$G = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

8. Suppose that the code words of a k -dimensional linear code of length n are arranged as the rows of a matrix with 2^k rows and n columns, with no column consisting only of "0"s. Show that each column consists of 2^{k-1} "0"s and 2^{k-1} "1"s. Use this to show that the sum of the weights of the code words is $n2^{k-1}$.

9. Use the result of the previous Exercise to show that the minimum distance of a k-dimensional linear code of length n is no more than $\dfrac{n2^{k-1}}{(2^k - 1)}$

10. Write down the generator polynomials of the cyclic codes whose generator matrices are given below :

$$G = \begin{bmatrix} 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}$$

11. Write down the generator matrices of the cyclic codes whose generator polynomials and lengths are given below :

(a) $g(X) = (1 + X + X^4 + X^5), n = 8$

(b) $g(X) = (1 + X^3 + X^6), n = 9$

12. Construct the parity check matrix of the Hamming code for $m = 4$, and show that it is equivalent to the cyclic code with $n = 15$ and generator polynomial $(1 + X + X^4)$.

1.7 Solution & Hints

Exercise 1.1 $n = 7$, $k = 3$

Exercise 1.2 Find all the subsets of code words from the given set C. Total no of subset by choosing 2 and 3 code words from the set is ${}^4C_2 + {}^4C_3 = 12$. Check whether the linear sum of members of each of these subsets is a code word or not. If not, then add the new code words into the set C to make it linear block code.

Exercise 1.3 The parity matrix $H = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \end{bmatrix}$

Exercise 1.4

Part a : The H_1 is $(n - k + 1) \times (n + 1)$ matrix. Since H is a linear code with $n - k$ linearly independent rows, the first $n = k$ rows of H_1 are also linearly independent. The last row of H_1 has a "1" in the first position but the other rows of H_1 have "0" at their first position. Any linear combination including the last row of H_1 will never yield a zero vector. Hence all rows in H_1 are linearly independent. The dimension of row space of H_1 is $(n - k + 1)$. Since the dimension of the null space of H_1, C_1 we can write using theorem 1.1,

$$\dim (C_1) = (n + 1) - (n - k + 1) = k$$

Therefore, C_1 is $(n + 1, k)$ linear code.

Part b : We will prove it by contradiction. Assume there is a valid code vector v of odd weight. As the last row of H_1 is an all "1" vector the expression $v.H_1^T$ (Adding odd no of 1 will produce 1) which can't be true for ant valid code word. Therefore v can't be of odd weight.

Part c : Let v be a code word in C. Then $v \,.\, H^T = 0$. Extend v by adding a digit

v_∞ to its left. This results in a vector v' of $n + 1$ digits,

$$v' = (v_\infty, v) = (v_\infty, v_0, v_1, ..., v_{n-1})$$

Since v be valid code word in C_1, we can write $v'.H_1^T = 0$. Considering the last row of H_1, the last component of inner product becomes,

$$v_\infty + v_0 + v_1 + ... + v_{n-1} = 0$$

To satisfy the above equation v_∞ must be 1 if $v = (v_0, v_1, ..., v_{n-1})$ has odd weight and v_∞ must be 0 if $v = (v_0, v_1, ..., v_{n-1})$ has even weight.

Exercise 1.5

After following sequence of row operations,

1. $R_3 \rightarrow R_1 + R_3$
2. $R_2 \rightarrow R_2 + R_3$
3. $R_1 \rightarrow R_1 + R_2$

G becomes

$$G = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 \end{bmatrix}$$

Which is in $[I_k : P]$ form. We can also represent it in $[P : I_k]$ after interchanging the columns.

So the matrix H can be expressed in the form $[P^T : 1_{n-k}]$ as following.

$$H = \begin{bmatrix} 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

It can be easily checked that $GH^T = 0$.

Exercise 1.6

Let C_e = Set of all even weight code words in C

C_0 = Set of all odd weight code words in C.

x be an odd weight code word in C_0. Adding x to each vector in C_0, we obtain a set C'_e, of all even weight vector.

Therefore, $$|C| = |C'_0| \leq |C_e| \Rightarrow |C_0| \leq |C_e| \qquad \text{(E1)}$$

Adding x to each vector in C_e, we obtain a set C'_0 of all odd weight vector.

Therefore, $|C_e| = |C_0'| \le |C_0| \Rightarrow |C_e| \le |C_0|$ (E2)

From equation (E1) and (E2) we can conclude $|C_e| = |C_0|$

Exercise 1.8 (1st Part)

Every column of the matrix contains at least one non-zero entry. Now consider any arbitrary column i. Let S_0 be the set of code words which contain "0" in i^{th} column and Let S_1 be the set of code words which contain "1" in i^{th} column. Let x be a code word from S_1 . Adding x to each vector of S_0 gives a set S'_1 of code words with a "1" in i^{th} column, Therefore,

$$|S_0| = |S_1'| \le |S_1| \Rightarrow |S_0| \le |S_1| \quad \text{(E3)}$$

On the other hand adding x to each vector of S_1 gives a set S'_0 of code words with a "0" in i^{th} column, Therefore,

$$|S_1| = |S_0'| \le |S_0| \Rightarrow |S_1| \le |S_0| \quad \text{(E4)}$$

From equation (E3) and (E4), we can say that $|S_0| = |S_1|$. Since the i^{th} column contains 2^k entry, we can conclude $|S_0| = |S_1| = \frac{2^k}{2} = 2^{k-1}$.

Exercise 1.8

The total no of "1"s in the matrix is $n2^{k-1}$.

Using theorem 1.3 we can write, each non-zero code word has weight at least min d_{min}. Out of 2^k code words one code word has all "0"s. So total no nonzero code words is = $2^k - 1$. Therefore,

$$(2^k - 1)d_{\min} \le n2^{k-1} \Rightarrow d_{\min} \le \frac{n2^{k-1}}{(2^k - 1)}$$

1.8 Reference and Further reading

1. Fundamentals of Information Theory and Coding Design, Roberto Togneri, Christopher J.S. deSilva, Taylor & Francis Group.
2. Error Control Coding : Fundamentals and Applictions, Shu Lin, Daniel J. Costello, Jr. Prentice - Hall.
3. Error-Correcting Codes, Second Edition, W.W.Peterson and E.J. Weldon, Jr., MIT Press, Cambridge, Mass.

Unit 2 ❑ Block Design

Structure

2.0 Objectives
2.1 Introduction
2.2 Definition and Properties of Design
2.3 Incidence Matrix
2.4 New BIBDs from Old BIBD
2.5 Fisher's Inequalities
2.6 Symmetric BIBD
2.7 BIBD Construction
2.8 Summary
2.9 Exercises
2.10 Reference and Further reading

2.0 Objectives

After going through this unit the learner should be able to :

- define what is a block design?
- understand balanced incomplete block design (BIBD) and its parameters.
- represent the block design in algebraic structure matrix.
- design new blocks from old blocks.
- define symmetric BIBD and its parameters.
- construct symmetric BIBD using different algebraic techniques.
- use block design in practical problems.
- use abstract and linear algebra as a tool for combinational block design technique.

2.1 Introduction

Design theory, specifically the combinational design theory, concerns questions about whether it is possible to arrange elements of a finite set into subsets so that certain "balance" properties are satisfied. Let us start with following puzzle of seven golfers to understand the content of this unit.

Golfers's Puzzle

Seven golfers are to spend a week's holiday at a seaside town which has two splendid golf courses. They decide each should play a round of golf on each of the seven days. They also decide that on each day they should split into two groups, one of size 3 to play on one course, and the other of size 4 on the other course. Can the groups be arranged so that each pair of the golfers plays together in a group of 3 the same no of times, and each pair plays together in a group of 4 the same no of times?

We will see the solution of this problem later. But the important point is to be noted that, this kind of problem is more recreational than most of the other brunches of mathematics. Even though combinational design theory that are studied today were first considered in the context of mathematical puzzles or brain-teasers in the eighteenth and nineteenth centuries. The study of design theory as a mathematical discipline really began in the twentieth century due to applications in the design and analysis of statistical experiments. Designs have many other applications as well, such as tournament scheduling, lotteries, mathematical biology, algorithm design and analysis, networking, group testing, and cryptography.

Design theory makes use of tools from linear algebra, groups, rings and fields, and number theory, as well as combinatorics. In this unit, the mathematical theorems relevant to design will be discussed with the examples rather than rigorous proofs.

2.2 Definition and Properties of Design

Definition 2.1

A design is a pair (X, A) such that the following properties are satisfied :

(1) X is a set of elements called points.

(2) A is a collection (i.e., multiset) of nonempty subsets of X called blocks.

If two blocks in a design are identical, they are said to be repeated blocks. Because of this repetition A is referred as a multiset of blocks rather than a set which contains distinct element. A design without repeated blocks is known as simple design. The order of the elements in a multiset is irrelevant, as with a set. Balanced incomplete block designs are probably the moststudied type of design. In this unit, our main focus will be on balanced incomplete block designs which is known as BIBD, in short.

Definition 2.2

Let v, k, and λ be positive integers such that $v > k \geq 2$. A (v, k, λ)- balanced incomplete block design (which we abbreviate to (v, k, λ) - BIBD) is a design (X, A) such that the following properties are satisfied :

1) $|X| = v$.

2) each block contains exactly k points.

3) every pair of distinct points is contained in exactly λ blocks.

Property 3 in the definition above is the "balance" property. A BIBD is called an incomplete block design because $k < v$, and hence all its blocks are incomplete blocks. A BIBD may possibly contain repeated blocks if $\lambda > 1$.

Example 2.1 (Solution of the Golfers' Puzzle)

One solution of the Golfers' Puzzle mentioned in Section 2.0 is given in Table 2.1. The players are marked by the numbers 0,1,2,3,4,5,6.

Day of the week	Group A(3 Players)	Group B(4 players)
Day 1	{0, 1, 3}	{2, 4, 5, 6}
Day 2	{1, 2, 4}	{0, 3, 5, 6}
Day 3	{2, 3, 5}	{0, 1, 4, 6}
Day 4	{3, 4, 6}	{0, 1, 2, 5}
Day 5	{4, 5, 0}	{1, 2, 3, 6}
Day 6	{5, 6, 1}	{0, 2, 3, 4}
Day 7	{6, 0, 2}	{1, 3, 4, 5}

Table 2.1

It can be checked that each pair of group A plays once together and each pair of group B plays twice together. Therefore, group A is an example of (7, 3, 1)-BIBD and group B is an example of (7, 4, 2)- BIBD. For the (7, 3, 1)- BIBD we can see :

X = {0,1,2,3,4,5,6} and A = {013, 124, 235, 346, 450, 561, 602}

This BIBD has a nice diagrammatic representation; see Figure 2.1. The blocks of the BIBD are the six lines and the circle in this diagram. The diagram is known as Fano Plane which satisfies the axioms of Fano's (Gino Fano) Geometry.

Example 2.2

A (9, 3, 1) - BIBD, where

X = (1, 2, 3, 4, 5, 6, 7, 8, 9) and

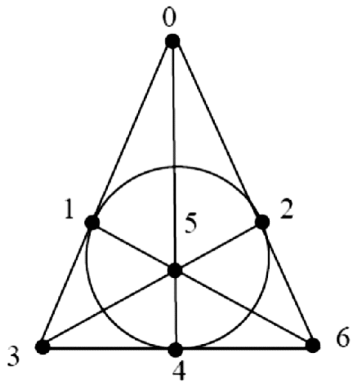


Figure 2.1

A = {123, 456, 789, 147, 258, 369, 159, 267, 348, 168, 249, 357}

This BIBD can also be presented diagrammatically; see Figure 2.2. The 12 blocks of the BIBD are depicted as eight lines and four triangles. Observe that the blocks can be separated into four sets of three, where each of these four sets cover every point in the BIBD.

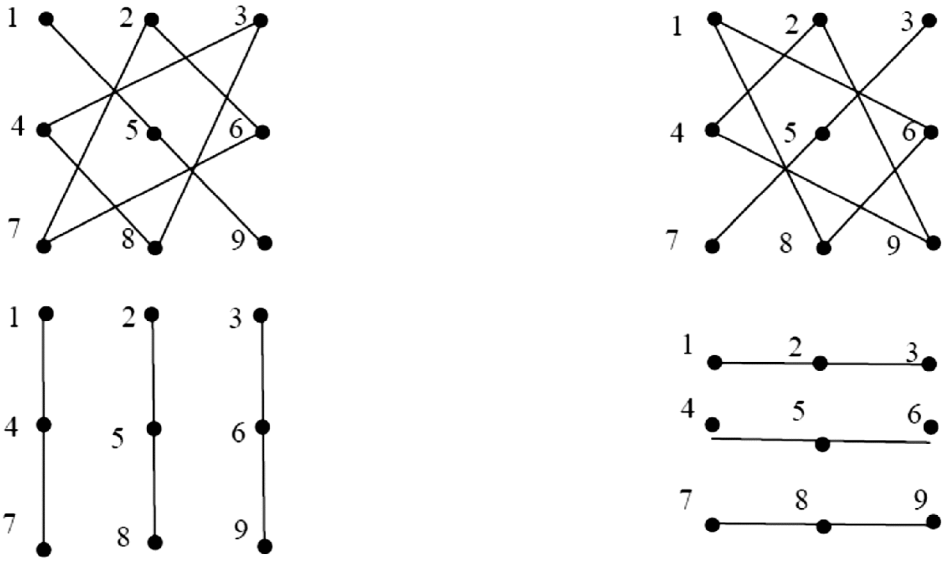


Figure 2.2

Theorem 2.1

In a *(v, k, λ)* - BIBD, every point occurs in exactly $r = \frac{\lambda(v-1)}{k-1}$ blocks.

Proof.

Let (X, A) - be a (v, k, λ) - BIBD. Suppose $x \in X$, and let r_x denote the number of blocks containing x. Define a set

$$I_x = \{(y, A_m) : y \in X, y \neq x, A_m \in A, \{x, y\} \subseteq A_m\}$$

We will compute $|I_x|$ in two different ways.

First, there are $v - 1$ ways to choose $y \in X$ such that $y \neq x$. For each such y, there are λ blocks A_m such that m $\{x, y\} \subseteq A_m$. Hence,

$$|I_x| = \lambda(v-1)$$

On the other hand, there are r_x ways to choose a block A_m such that $x \in A_m$. For each choice of A_m, there are $k - 1$ ways to choose $y \in A_m, y \neq x$. Hence,

$$|I_x| = r_x(k-1).$$

Combining these two equations, we see that

$$\lambda(v-1) = r_x(k-1).$$

Hence $r_x = \frac{\lambda(v-1)}{(k-1)}$ is independent of x and we denote $r_x = r$, Therefore,

$$r = \frac{\lambda(v-1)}{(k-1)}$$

The value r is often called the replication number of the BIBD.

Theorem 2.2

A *(v, k, λ)* - BIBD has exactly $b = \frac{vr}{k} = \frac{\lambda(v^2-v)}{k^2-k}$ blocks.

Proof.

Let (X, A) be a (v, k, λ) - BIBD, and let b = |A|. Define a set

$$I = \{(x, A_m) : x \in X, A_m \in A, x \in A_m\}.$$

We will compute $|I|$ in two different ways.

First, there are v ways to choose $x \in X$. For each such x, there are r blocks A_m such that $x \in A_m$. Hence,

$$I = vr.$$

On the other hand, there are b ways to choose a block $A_m \in$ A. For each choice of A_m, there are k ways to choose $x \in A_m$. Hence,

$$I = bk .$$

Combining these two equations, we see that

$$bk = vr$$

$$\Rightarrow b = \frac{vr}{k} = \frac{\lambda v(v-1)}{k(k-1)} = \frac{\lambda(v^2 - v)}{(k^2 - k)}$$

We can also use the notation (v, b, r, k, λ) -BIBD if we want to record the values of all five parameters.

Corollary 2.3

If a (v, k, λ)-BIBD exists, then λ (v – 1) ≡ 0 mod (k – 1) and λv (v–1) ≡ 0 mod $k(k–1)$. [Here $a \equiv b$ mod c means the remainder will be same when a and b both divided by c].

For example, an (8, 3, 1)- BIBD does not exist because $\lambda(v - 1) = 7 \neq 0$ (mod 2).

As another example, let us consider the parameter set (19, 4, 1). Here, we see that λ (v – 1) = 18 = 0 (mod 3) but v (v – 1) = 342 ≠ 0(mod 12). Hence a (19, 4, 1)-BIBD cannot exist.

Corollary 2.3 is necessary conditions for existence of a BIBD with fixed values of v, k and λ. This means if we find that the corollary 2.3 is not satisfied for the specific values of parameter v, k and λ, we can conclude construction of (v, k, λ)-BIBD is impossible. On the other hand, satisfying the corollary 2.3 for the specific values of parameter v,k and λ doesn't guarantee the existence of (v, k, λ)-BIBD. One of the main goals of combinatorial design theory is to determine necessary and sufficient conditions for the existence of a (v, k,λ)-BIBD. This is a very difficult problem in general, and there are many parameter sets where the answer is not yet known. For example, it is currently unknown if there exists a (22, 8, 4)-BIBD even though the necessary condition (corollary 2.3) are satisfied.

2.3 Incidence Matrix

It is often convenient to represent a BIBD by means of an incidence matrix.

Definition 2.3

Let (X, A) be a design where X = {x_1, x_2, ..., x_v} and {A_1, A_2, ..., A_b}. The incidence matrix of (X, A) is the $v \times b$ 0–1 matrix $M = (m_{ij})$ defined as

$$(m_{i,j}) = \begin{cases} 1 & if \quad x_i \in A_j \\ 0 & if \quad x_i \notin A_j \end{cases}$$

The incidence matrix, M, of a (v, b, r, k, λ)-BIBD satisfies the following properties :

1. every column of M contains exactly k "1"s ;
2. every row of M contains exactly r "1"s ;
3. two distinct rows of M both contain "1"s in exactly λ columns.

Example 2.3

Consider the (9, 3, 1)- BIBD of example 2.2. The parameter for the BIBD is given below.

$X = \{1,2,3,4,5,6,7,8,9\}$

$A = \{123, 456, 789, 147, 258, 369, 159, 267, 348, 168, 249, 357\}$

$$v = 9, k = 3, \lambda = 1, b = \frac{\lambda v(v-1)}{k(k-1)} = 12, r = \frac{\lambda(v-1)}{(k-1)} = 4$$

The incidence matrix of this design is the following 9 × 12 matrix (Figure 2.3).

$$M = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \\ x_9 \end{array} \begin{array}{c} \begin{array}{cccccccccccc} A_1 & A_2 & A_3 & A_4 & A_5 & A_6 & A_7 & A_8 & A_9 & A_{10} & A_{11} & A_{12} \end{array} \\ \left[\begin{array}{cccccccccccc} 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \end{array}\right] \end{array}$$

Figure 2.3

We can validate that every row sum is $r = 4$ and every column sum is $k = 3$. The corollary 2.3 can also be proved from the incidence matrix M.

If we add the "1"s for all rows, then total no "1"s in the matrix M is vr.

Again if we add the "1"s for all columns, then total no "1"s in M is bk.

Therefore, equating total no of "1" from the above two count we can conclude,

$$vr = bk$$

We can also observe that the intersection of any two rows has a common "1" in λ (in this case, 1) block positions. This implies every pair of distinct points of X are contained in will λ bocks.

We know that, there are $\binom{v}{2}$ ways to choose a pair of points from v distinct points of set X.

Next we consider the $\binom{v}{2} \times b$ pair-incidence matrix

$$M' = \begin{array}{c} \\ x_1x_2 \\ x_1x_3 \\ \vdots \\ \vdots \\ x_{v-1}x_v \end{array} \begin{array}{c} \begin{array}{ccccc} A_1 & A_2 & \cdots & A_{b-1} & A_b \end{array} \\ \left[\begin{array}{ccccc} e_{12,1} & e_{12,2} & \cdots & e_{12,(b-1)} & e_{12,b} \\ e_{13,1} & e_{13,2} & \cdots & e_{13,(b-1)} & e_{13,b} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \cdots & \ddots & \vdots \\ e_{(v-1)v,1} & e_{(v-1)v,2} & \cdots & e_{(v-1)v,(b-1)} & e_{(v-1)v,b} \end{array}\right] \end{array}$$

Now the every row sum of M is λ and there are $\binom{v}{2}$ rows in M.

Therefore, total no of "1"s in M for all rows is $\binom{v}{2}\lambda$.

On the other hand, since any block has k points, we can choose $\binom{k}{2}$ pair of points where each pair is associated with that block. Therefore, every column has $\binom{k}{2}$ no of "1"s.

The total no of "1" s in the matrix for all columns is $\binom{k}{2}b$.

We can write,

$$\binom{v}{2}\lambda = \binom{k}{2}b \Rightarrow \frac{v(v-1)}{2}\lambda = \frac{k(k-1)}{2}b$$

$\Rightarrow v(v-1)\lambda = vr(k-1)$ [$\because bk = br$]

$\Rightarrow r = \frac{\lambda(v-1)}{k-1}$

The same result was also obtained in theorem 2.1.

Corollary 2.4

For every non-empty BIBD, λ < r

Proof :

From definition of BIBD we know that, $k < v \Rightarrow k - 1 < v - 1 \Rightarrow \frac{k-1}{v-1} < 1$ (1)

from theorem 2.1 we know that,

$\lambda(v-1) = r(k-1) \Rightarrow \frac{\lambda}{r} = \frac{k-1}{v-1} \Rightarrow \frac{\lambda}{r} < 1$ [From equation (1)]

$\Rightarrow \lambda$ < r

2.4 New BIBDs from Old BIBD

A new BIBD can be constructed from old BIBD in two different methods. The methods are following.

2.4.1 Sum Construction :

Theorem 2.5

Suppose there exists a v, k, λ_1-BIBD (X, A_1) and a v, k, λ_2-BIBD (X, A_2). Then there exists a v, k, $\lambda_1 + \lambda_2$ -BIBD (X, A) on the set X.

Proof :

Let $A = A_1 \cup A_2$ be the multiset union of the multisets A_1 and A_2. Then A is a multiset of nonempty subsets of X. We have already seen in Section 2.2 that multiset may contain repeated block, therefore A can have same block repetitively.

- Clearly $|x| = v$
- Furthermore, since every block in A_1 contains k points and every block in A_2 contains the same k points, we can conclude that every block in A (union of two sets each having same k points) contains k points.
- Let us consider an arbitrary point $x, y \in X$ be such that $x \neq y$. Then the pair $\{x, y\}$ is contained in λ_1 blocks in A_1, and the pair $\{x, y\}$ is contained in λ_2 blocks in A_2, so the pair $\{x, y\}$ is contained in $\lambda_1 + \lambda_2$ blocks in A (can have repeated blocks as A is a multiset.

- Hence X, A is a v, k, $\lambda_1 + \lambda_2$-BIBD.

Corollary 2.6

Suppose there exists a (v, k, λ)-BIBD. Then there exists a (v, k, $s\lambda$)-BIBD for all integers $s \geq 1$.

Note that the BIBDs produced by Corollary 2.6 with $s \geq 2$ are not simple designs, even if the initial (v, k, λ)-BIBD is simple. For ($\lambda > 1$), construction of simple BIBDs is, in general, more difficult than construction of BIBDs with repeated blocks. The construction process used to prove theorem 2.5 can be used to create new BIBD from old BIBD.

2.4.2 Block Complementation:

Theorem 2.7

Suppose there exists a (v, b, r, k, λ)-BIBD, where $k \leq v - 2$. Then there also exists a (v,b,$b - r$,$v - k$, $b - 2r + \lambda$)-BIBD.

Proof:

Suppose (X, A) is a (v, b, r, k, λ)-BIBD. Then block complementation is done by replacing every block $A_m \in A$ by $X \setminus A_m$ (all other points of X not belonging to A_m. We will show that new design (X, $\{X \setminus A_m : A_m \in A\}$) is a BIBD.

- Clearly the new design has v points since no point has been deleted.
- The new design has b blocks since we have created one block corresponding to each block of the old design.
- Any block in the new design (X, $\{X \setminus A_m : A_m \in A\}$) has been created with the points which were not there in the corresponding block of old design (X, A).

  Suppose, $X = \{0,1,2,3,4,5,6\}$, $A = \{013, 124, 235, 346, 450, 561, 602\}$. Then the block in the new design corresponding to the block $A_1 = (013) \in A$ will be $A'_1 = 2456$. Therefore, the new design will have $k' = v - k$ points in every block.
- Any point $x \in X$ of design (X, A) with parameter (v, b, r, k, λ) occurs in r blocks (A_1, A_2, ..., A_r). It is clear that x can't be there in any of the remaining $b - r$ (A_{r+1}, A_{r+2}, ..., A_b) blocks of (X, A). Now the new design is constructed in such a way that point x will present in all $b - r$ ($A'_{r+1}, A'_{r+2}, \ldots\ldots, A'_b$) blocks corresponding to (A_{r+1}, A_{r+2}, ..., A_b) of old design (X, A). Therefore, any point $x \in X$ of new design will be there in $b - r$ blocks. Therefore, $r' = b - r$.

- We need to prove the last parameter λ' of new design is $b - 2r + \lambda$.

  Let $x, y \in X, x \neq y$, then considering design (X, A) we can write,

$$|A| = b \Rightarrow (|A_m \in A : x, y \in A_m| + |A_m \in A : x \in A_m, y \notin A_m|$$

$$+ |A_m \in A : x \notin A_m, y \in A_m| + |A_m \in A : x, y \notin A_m|) = b \qquad (2)$$

  Define, $b_1 = |A_m \in A : x, y \in A_m| = \lambda$

$$b_2 = |A_m \in A : x \in A_m, y \notin A_m| = |A_m \in A : x \in A_m| - |A_m \in A : x \in A_m, y \in A_m| = r - \lambda$$

$$b_3 = |A_m \in A : x \notin A_m, y \in A_m| = |A_m \in A : y \in A_m| - |A_m \in A : x \in A_m, y \in A_m| = r - \lambda$$

$$b_4 = |A_m \in A : x, y \notin A_m|) = \lambda'$$

  We can rewrite the equation (2) as

$$b_1 + b_2 + b_3 + b_4 = b \Rightarrow \lambda + (r - \lambda) + (r - \lambda) + \lambda' = b$$

$$\Rightarrow \lambda' = b - 2r + \lambda$$

Therefore, we can conclude new design $(X, \{X \setminus A_m : A_m \in A\})$ is a BIBD with parameters $(v, b, b - r, v - k, b - 2r + \lambda)$.

Now if we refer the solution of golfer's puzzle in table 2.1 of Section 2.2, it can be verified that BIBD for group B(4 players) with parameters) (7, 7, 4, 4, 2) can be constructed using block complementation method from the BIBD of group A(3 players) with parameters (7, 7, 3, 3, 1).

2.5 Fisher's Inequalities

We have already discussed two necessary conditions for the existence of a (v, k, λ)-BIBD, namely theorems 2.1 and 2.2. Another important necessary condition is known as "Fisher's Inequality". Before we elaborate this, it is important understand few basic properties of linear algebra.

Rank of a Matrix :

The dimensions of column space or row space (refer to section 1.2.7 of unit 01) of a matrix is known as rank of the matrix.

The following theorem can be proved easily using linear algebra.

Theorem 2.8

If M_1 M_2 be the product of matrix M_1 and M_2 then

$$rank(M_1M_2) \leq \min(rank(M_1), rank(M_2))$$

Theorem 2.9 (Fisher's Inequalities) :

In Any BIBD, $b \geq v$

Proof:

Let $M(v \times b)$ be an incident matrix of a BIBD with parameters (v,b, r, k, λ). $M = \left[m_{ij}\right]$ where row i = 1, 2, ..., v denotes the no of points and column j = 1, 2, ..., b denotes the no of blocks of BIBD. Let us also define a new $(v \times v)$ matrix $M' = MM^T = \left[m'_{ij}\right]$. From the definition of matrix multiplication, we can write,

$$m'_{ij} = \sum_{k=1}^{b} m_{ik} m_{jk} = \begin{cases} r \ when \ i = j \\ \lambda \ when \ i \neq j \end{cases}$$

The matrix,

$$M' = MM^T = \begin{bmatrix} r & \lambda & \lambda & \lambda & \ldots & \lambda \\ \lambda & r & \lambda & \lambda & \ldots & \lambda \\ \lambda & \lambda & r & \lambda & \ldots & \lambda \\ \lambda & \lambda & \lambda & r & \ldots & \lambda \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \lambda & \lambda & \lambda & \lambda & \ldots & r \end{bmatrix} = (r-\lambda)I_v + \lambda J_v$$

where I_n denotes an $n \times n$ identity matrix, J_n denotes the $n \times n$ matrix in which every entry is a "1". Subtracting the first column of a matrix from the other columns does not change the determinant. Hence,

$$\det(M') = \begin{bmatrix} r & \lambda - r & \lambda - r & \lambda - r & \ldots & \lambda - r \\ \lambda & r - \lambda & 0 & 0 & \ldots & 0 \\ \lambda & 0 & r - \lambda & 0 & \ldots & 0 \\ \lambda & 0 & 0 & r - \lambda & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \lambda & 0 & 0 & 0 & \ldots & r - \lambda \end{bmatrix}$$

Adding the other rows of a matrix to the first row does not change the determinant. Hence,

$$\det(M') = \begin{bmatrix} r + (v-1)\lambda & 0 & 0 & 0 & \ldots & 0 \\ \lambda & r - \lambda & 0 & 0 & \ldots & 0 \\ \lambda & 0 & r - \lambda & 0 & \ldots & 0 \\ \lambda & 0 & 0 & r - \lambda & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \lambda & 0 & 0 & 0 & \ldots & r - \lambda \end{bmatrix}$$

Since the upper triangle of this matrix is all zeroes, the determinant is the product of the diagonal entries. Thus,

$$\det(M') = [r + (v-1)\lambda](r-\lambda)^{v-1}$$

Now $(r-\lambda)>0$ (form corollary 2.4) and $[r + (v-1)\lambda] > 0$ ($\because$ v > 2), det (M') is non-zero.

Accordingly, the rank of the $(v \times v)$-matrix MM^T is v. Since the rank of the $(v \times b)$ incidence matrix M is at most b, and since the rank, v, of the product matrix MM^T cannot exceed the rank of the matrix M (theorem 2.8), it follows that $v \leq b$ or $b \geq v$.

2.6 Symmatric BIBD :

Definition 2.4

A BIBD in which $b = v$ (or, equivalently, $r = k$ or $\lambda(v-1) = k^2 - k$) is called a symmetric BIBD. The term symmetric is a poor choice inherited from the statistical history of the subject. The incidence matrices of these designs are not symmetric matrices.

Theorem 2.10

Suppose that (X, A) is a symmetric (v, k, λ)-BIBD and denote $A = \{A_1, A_2, ..., A_v\}$ Suppose that $1 \leq i, j \leq v, i \neq j$. Then $|A_i \cap A_j| = \lambda$

Proof :

To prove the above theorem first we state few important properties of determinant and matrices in table 2.2. The reader can refer the book "Linear Algebra and its Application" by David C. Lay mentioned in the Section 2.11 (Reference and Further Reading) to proof these theorems.

Property 1 :

If A is an $n \times n$ matrix, then det(A) = det(A^T)

Property 2 (Multiplicative property) :

If A and B are $n \times n$ matrices, then det (AB) = det(A) det(B)

Definition 2.5

An $n \times n$ matrix A is said to be invertible, if there is an $n \times n$ matrix C such that

$$CA = I \text{ and } AC = I$$

where $I = I_n$, the $n \times n$ identity matrix. In this case, C is inverse of A, which also is denoted by A^{-1}.

Property 3 :

If inverse of matrix A exists then $\det(A) \neq 0$

Table 2.2

In theorem 2.9, we have already proved that for incidence matrix

$\det(MM^T) > 0$

$\Rightarrow \det(M)\det(M^T) > 0$ [using property 2 of table 2.2] (3)

Also since the design is symmetric, it's incidence matrix is square using property 1 of table 2.2 we can write,

$$\det(M) = \det(M^T) \quad (4)$$

Now from (3) and (4), we can conclude $\det(M) \neq 0$, thus M^{-1} exists [using property 3 of table 2.2]. Now from theorem 2.9 we get,

$$MM^T = (r-\lambda)I_v + \lambda J_v \Rightarrow MM^T M = ((r-\lambda)I_v + \lambda J_v)M$$

$$\Rightarrow MM^T M = ((r-\lambda)I_v M + \lambda J_v M)$$

$$\Rightarrow MM^T M = ((r-\lambda)MI_v + \lambda MJ_v) \quad [\because I_v M = MI_v, J_v M = MJ_v]$$

$$\Rightarrow MM^T M = M((r-\lambda)I_v + \lambda J_v) \Rightarrow MM^T M = MMM^T$$

$$\Rightarrow M^{-1}MM^T M = M^{-1}MMM^T \Rightarrow M^T M = MM^T$$

The $(i, j)^{th}$, entry in the product on the right is λ if $i \neq$ j (refer theorem 2.9) however this entry in the product on the left is the inner product of columns i and j of A, i.e., it gives the number of elements in common in the two blocks which are represented by these columns.

2.7 BIBD Construction

We now discuss a method for constructing symmetric BIBDs that uses the arithmetic of the integers modulo n (Z_n).

2.7.1 Method using Difference Set

In this method, the points are the integers in Z_n, so, to make our notation consistent, we use v instead of n.

Let $v \geq 2$ be an integer, and consider the set of integers mod v :

$$Z_v = \{0, 1, 2, ..., v-1\}$$

Note that addition and multiplication in Z_v are denoted by the usual symbols + and ×. Let $B = \{i_1, i_2, ..., i_k\}$ be a subset of Z_v consisting of k integers. For each integer j in Z_v, we define

$$B + j = \{i_1 + j, i_2 + j, ..., i_k + j\}$$

to be the subset of Z_v obtained by adding j and v to each of the integers in B. The set $B + j$ also contains k integers. We can prove this by method of contradiction. So let us assume two integer $i_{p'}, i_{q'} \in Z_n$, are found to be the same after adding j mod v to two integer $i_{p'}, i_{q'} \in Z_n$, respectively. That is,

$$i_{p'} = i_{q'} \Rightarrow i_p + j = i_q + j$$

$$\Rightarrow i_p = i_q \text{ [By adding the additive inverse } j \text{ } f\{ \text{ into both side]}$$

The v sets can be formed by

$$\{B + 0, B + 1, ..., B + v - 1\}$$

are called the blocks developed from the block B, and B is called the starter.

Example 2.4

Let $v = 7$ and consider $Z_v = \{0,1,2,3,4,5,6\}$. The starting block B = {0,1,3}. Then we have,

$B + 0 = \{0,1,3\}$
$B + 1 = \{1,2,4\}$
$B + 2 = \{2,3,5\}$
$B + 3 = \{3,4,6\}$
$B + 4 = \{4,5,0\}$ [$\therefore$ 6 + 1 mod (7) = 0]
$B + 5 = \{5,6,1\}$
$B + 6 = \{6,0,2\}$

Each set in this list, other than the first, is obtained by adding 1 mod 7 to the previous set. In addition, the first set B on the list can be found from the last by adding 1 mod 7. This is a BIBD, indeed, the same one in the introductory example 2.1(golfers' puzzle) in Section 2.2. Since b = v, we have an symmetric BIBD with $b = v = 7$, $k = r = 3$ and $\lambda = 1$.

Example 2.5

Let $v = 7$ and $Z_v = \{0, 1, 2, 3, 4, 5, 6\}$ as previous example. The starting block $B = \{0, 1, 4\}$. Then we have,

$$B + 0 = \{0,1,4\}$$
$$B + 1 = \{1,2,5\}$$
$$B + 2 = \{2,3,6\}$$
$$B + 3 = \{3,4,0\}$$
$$B + 4 = \{4,5,1\}$$
$$B + 5 = \{5,6,2\}$$
$$B + 6 = \{6,0,3\}$$

In this case, we do not obtain a BIBD because, for instance, the pair of points (1, 2) occurs in one block, while the pair of points (1, 5) is in two blocks.

It follows from these two examples that sometimes, but not always, the blocks developed from a starter block are the blocks of a symmetric BIBD. Therefore, we must define additional property so that the development process ensures the design to be symmetric BIBD.

Difference Set

Definition 2.6

Suppose $(G, +)$ is a finite group of order v in which the identity element is denoted by "0". Unless explicitly stated, we will not require that G be an Abelian group. Let k and λ be positive integers such that $2 \leq k < v$. A (v, k, λ)-difference set in $(G, +)$ is a subset $B \subseteq G$ that satisfies the following properties :

(1) $|B| = k$

(2) The multiset $[x - y : x, y \in B, x \neq y]$ contains every element in $G \setminus \{0\}$ exactly λ times.

Example 2.6

A (7, 3, 1)- difference set in $(Z_7, +)$ is $B = \{0, 1, 3\}$.

Let us first understand the addition and subtraction operation in modulo 7 arithmetic using a 7-hour clock which is marked with 0,1, 2, 3, 4, 5, 6 along its perimeter (figure 2.3). Let us choose any arbitrary pair of points (x, y) from set B. Say $x = 1$ and $y = 3$. The

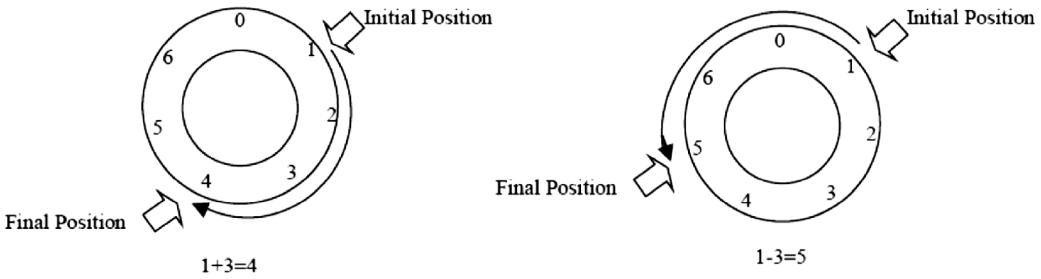


Figure 2.3

1+3 means, clockwise movement by 3 steps from initial position 1, therefore, the result of addition is 4. On the other hand, 1-3 means, anticlockwise movement by 3 steps from initial position 1, therefore, the result is 5.

Addition modulo (v)	Subtraction modulo (v)
$x+y\begin{cases} x+y & \text{if } x+y<v \\ (x+y)\bmod(v) & \text{if } x+y\geq v \end{cases}$	$x-y\begin{cases} x-y & \text{if } x\geq y \\ x+(v-y) & \text{if } x<y \end{cases}$

Table 2.3

In general, the modulo v operations can be written algebraically as mentioned in table 2.3. Now we compute $x - y$ for all pair x, y $\in B$, $x \neq y$ and note down the result in table 2.4.

-	0	1	3
0	*	6	4
1	1	*	5
3	3	2	*

Table 2.4

Examining this table, we see that the nonzero integers 1, 2, 3, 4, 5, 6 in Z_7, each occur exactly once in the off-diagonal positions and hence exactly once as a difference. Hence, B is a difference set mod 7. All the diagonal entries are marked as "*" to indicate them redundant as $x = y$ is not considered to find out the difference set.

Example 2.7

Let us create the subtraction table (table 2.5) for example 2.5.

-	0	1	4
0	*	6	3
1	1	*	4
4	4	3	*

Table 2.5

We see that 1 and 6 each occur once as a difference, 3 and 4 each occur twice, and 2 and 5 do not occur at all. Thus, B is not a difference set in this case.

Theorem 2.11 :

Let B be a subset of $k < v$ elements of Z_v, that forms a difference set mod v. Then the blocks developed from B as a starter block form a symmetric BIBD with

$$\lambda = \frac{k(k-1)}{v-1}$$

Proof.

Since $k < v$, the blocks are not complete. Each block contains k elements. In section 2.7.1, we have already seen that v sets can be developed by

$$\{B + 0, B + 1, ..., B + v - 1\}$$

Therefore, the number of blocks is the same as the number v of points.

Since B is a difference set, as per the definition each nonzero integer in Z_v occurs as a difference exactly same no of times. Let us assume this no is λ. Now total no of ordered pair from k elements of set B is $k(k-1)$. This means the subtraction table of B contain $k(k-1)$ no of differences.

Again total no of non-zero element in Z_v is $v-1$. Each of these elements is repeated λ times as a difference in the subtraction table of set B. Therefore, we can write,

$$\lambda(v-1) = k(k-1) \Rightarrow \lambda = \frac{k(k-1)}{(v-1)}$$

This proves that each nonzero integer in Z_v, occurs as a difference exactly $\frac{k(k-1)}{(v-1)}$ times in the subtraction table of difference set. Now we show that each pair of elements of Z_v is in $\lambda = \frac{k(k-1)}{(v-1)}$ blocks.

Let p and q be distinct integers in Z_v. Then,

$p - q = r \neq 0$, and $r \in Z_v$ (by property (2) of definition 2.6)

Now r must be present exactly λ times in the subtraction table of set B as a difference. Therefore, there must be λ distinct equation for, x_j, $y_j \in$ B, where $1 \leq j \leq \lambda$, such that

$$x_j - y_j = r \Rightarrow x_j - y_j = p - q \Rightarrow p - x_j = q - y_j$$

Let us suppose, $g_j = p - x_j \Rightarrow p = g_j + x_j$

Then, $q = p - x_j + y_j = g_j + y_j$

Now since $g_j \in Z_n$, $B + g_j$ is another block that can be developed from B. If the pair $(x, y) \in B$ then, $(x + g_j, y + g_j) \in B + g_j \Rightarrow (p,q) \in B + g_j$

Now since, $1 \leq j \leq \lambda$, p and q are together in exactly λ blocks and the blocks are denoted by $\{B + g_1, B + g_2, ..., B + g_\lambda\}$.

Thus each pair of elements of Z_v is in $\lambda = \frac{k(k-1)}{(v-1)}$ blocks.

2.7.2 Quadratic Residue Difference Sets

The difference set can also be formed using quadratic residue. Most of the topics of this section will use different results from number theory. We will discuss them without rigorous proofs. The reader can refer the books mentioned in the section 2.11 (Reference and Further Reading) to proof these results.

Quadratic Residue

Definition 2.7

Suppose that a, $m(\geq 2)$ are two integers and gcd(a, m) = 1 (a,m are coprime). Then we say that a is a quadratic residue modulo m if and only if the congruence

$$x^2 \equiv a(\text{mod } m)$$

has a solution $x \in Z_m \setminus \{0\}$, otherwise a is said to be non-quadratic residue.

Example 2.8

To find quadratic residues mod 11 we square all the numbers 1. 2, 3,....,10 except 0 of $x \in Z_{11} \setminus \{0\}$ and reduce to mod 11(table 2.6)

x	1	2	3	4	5	6	7	8	9	10
x^2	1	4	9	16	25	36	49	64	81	100
x^2(mod 11)	1	4	9	5	3	3	5	9	4	1

Table 2.6

Therefore, quadratic residues mod 11 denoted by $QR(11)$ is {1,3,4,5,9} and quadratic non-residues mod 11 denoted by $QNR(11)$ is {2,6,7,8,10}.

Let us now define quadratic residues in a finite field $F_q = GF(q)$, where q is an odd prime power. The quadratic residues of F_q are the elements in the set

$$QR(q) = \{x^2 : x \in F_q, x \neq 0\}$$

We will also define

$$QNR(q) = F_q \setminus (QR(q) \cup \{0\})$$

The elements of $QNR(q)$ are called the quadratic nonresidues of F_q.

We will now characterize the quadratic residues and nonresidues in a different way. To do that let us first define the cyclic group as follows.

Cyclic Group :

Definition 2.8

A group G is called cyclic if $\exists x \in G$ such that,

$$G = \langle x \rangle = \{x^n : n \in Z\}.$$

We say x is a generator of G. (A cyclic group may have many generators.) Although the list ..., x^{-2}, x^{-1}, x^0, x^1, x^2, ... has infinitely many entries, the set $\{x^n : n \in Z\}$ may have only finitely many elements. Cyclic groups are Abelian. And all groups of prime order are cyclic.

We make use of the important fact (which we do not prove) that the multiplicative group $(Fq \setminus \{0\}, *)$ is a cyclic group. A generator of this group, say ω, is called a primitive element of the field F_q. Clearly, an element $\omega \in F_q$ is a primitive element if and only if,

$$\{\omega^i : 0 \le i \le q-2\} = F_q \setminus \{0\}.$$

Observe that we have not considered $i = q - 1$. This is because according to Fermat's Little Theorem, if q be a prime number then we can write,

$$\omega^{q-1} = 1 \bmod q$$

Again $\omega^0 = 1 (\bmod q)$, so we can conclude $\omega^{q-1} = \omega^0 1 (\bmod q)$. Since we have considered the ω^0, there is no need to consider ω^{q-1} again in the set.

Example 2.9

Consider the field F_7. Let us find the primitive element of this field. We have determined all the powers of each element of F_7 except "0" in table 2.7. Suppose to find out 4^3 (mod 7) we do the following.

$4^3 = 64$, now the remainder when we divide 64 by 7 is 1. Therefore, $4^3 = 1 (\text{mod } 7)$.

x	x^0	x^1	x^2	x^3	x^4	x^5
1	1	1	1	1	1	1
2	1	2	4	1	2	4
3	1	3	2	6	4	5
4	1	4	2	1	4	2
5	1	5	4	6	2	3
6	1	6	1	6	1	6

Table 2.7

We can easily check that 3 is the primitive element as 3^i for $0 \le i \le q-2$ generates all the elements of $F_7 \setminus \{0\}$. We also find the quadratic residue of mod(7) in table 2.8.

x	1	2	3	4	5	6
x^2	1	4	9	16	25	36
x^2(mod 7)	1	4	2	2	4	1

Table 2.8

The $QR(7) = \{1,2,4\}$ and $QNR(7) = \{3,5,6\}$. Therefore, we can observe that in terms of a primitive element, ω of F^q, the quadratic residues are the even powers of ω, and the quadratic nonresidues are the odd powers. So, from the example 2.9, we can state the following result (without proof) :

$$QR(q) = \{\omega^{2i} : 0 \le i \le \frac{q-3}{2}\}$$

The cardinality or no of element (n) of this set can be determined by the help of arithmetic progression in the following manner.

$$t_0 + (n-1)d = t_n$$

where t_i is the i th element and d is the common difference. Replacing all the values we get,

$$0 + (n-1) = \frac{q-3}{2} \Rightarrow n = \frac{q-1}{2}$$

Finally, $\left|\{\omega^{2i} : 0 \le i \le \frac{q-3}{2}\}\right| = \frac{q-1}{2}|QR(q)|$

Therefore, The $QR(q)$ can be generated from primitive element of F_q.

Theorem 2.12

Suppose q = 3(mod 4) is a prime power, Then $QR(q)$ is a difference set in $(F_q, +)$.

Example 2.10

The QR difference set obtained when q = 11 is an (11, 5, 2)-difference set. We have already seen in example 2.8 that $QR(11) = \{1, 3, 4, 5, 9\}$. Then by theorem 2.12 we can choose that $B = \{1,3,4,5,9\}$ as a difference set. Then we have,

$B + 0 = \{1,3,4,5,9\}$

$B + 1 = \{2,4,5,6,10\}$

$B + 2 = \{3,5,6,7,0\}$

$B + 3 = \{4,6,7,8,1\}$

$B + 4 = \{5,7,8,9,2\}$

$B + 5 = \{6,8,9,10,3\}$

$$B + 6 = \{7,9,10,0,4\}$$
$$B + 7 = \{8,10,0,1,5\}$$
$$B + 8 = \{9,0,1,2,6\}$$
$$B + 9 = \{10,1,2,3,7\}$$
$$B + 10 = \{0,2,3,4,8\}$$

It can be checked that $k = \frac{q-1}{2} = 5$ $\lambda = \frac{q-3}{4} = 2$ for this symmetric BIBD.

2.8 Summary

The theory and principle of block design with help of abstract and linear algebra are discussed in this unit. Learners can now appreciate that the technique of block design is highly dependent on the principle of number theory, abstract and linear algebra. To realize the beauty of block design, someone should have basic knowledge of these fields. After studying this unit learners can opt more advanced courses in combinational design.

2.9 Exercises

Exercise 2.1

Find out value of b in a (46, 6, 1)-BIBD (if it exists) and the value of r in a (65, 5,1)-BIBD (if it exists) ?

Exercise 2.2

Does there exist a BIBD with following parameters

(1) $b = 10$, $k = 4$, $v = 8$ and $r = 5$.

(2) $b = 12$, $k = 4$, $v = 16$ and $r = 3$.

(3) $b = 20$, $k = 9$, $v = 18$ and $r = 10$.

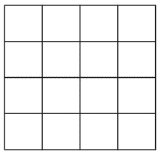
Exercise 2.3

Let M be the incidence matrix of a (v, b, r, k, 1)-BIBD and define a ($b \times b$) matrix $N = M^T M$. Denote $N = \left[n_{ij} \right]$. Prove that

$$n_{ij} = \begin{cases} k & \text{if } i = j \\ 0 \text{ or } 1 & \text{if } i \neq j \end{cases}$$

Exercise 2.4

Consider the squares of 4 × 4 board.



There are total 16 squares in the board. We define blocks as follows: For each given square, we take the 6 other squares that are either in its row or in its column (so not the given square itself). Prove that the design is a BIBD.

Exercise 2.5

Determine the complementary design of the BIBD with parameters

(1) $b = v = 7$, $k = r = 3$, $\lambda = 1$.

(2) $b = v = 16$, $k = r = 6$, $\lambda = 2$.

Exercise 2.6

How are the incidence matrices of a BIBD and its complement related ?

Exercise 2.7

Show that a BIBD, with v points whose block size v equals $v - 1$ does not have a complementary design.

Exercise 2.8

Show that (21, 5, 1)-difference set in $(Z_{21}, +)$ is B = {0, 1, 2, 4, 5, 8, 10}.

Exercise 2.9

Show that (15, 7, 3)- difference set in $(Z_{15}, +)$ is B = {0, 1, 2, 4, 5, 8, 10}.

Exercise 2.10

Develop all the blocks of the symmetric BIBD (11, 5, 2) using the starter block as B = {0, 2, 3, 4, 8} difference set in $(Z_{11}, +)$.

Exercise 2.11

Show that B = {0, 1, 3, 9} is a difference set in $(Z_{13}, +)$, and use this difference set as a starter block to construct a symmetric BIBD. Identify the parameters of the block design.

Exercise 2.12

Is B = {0, 2, 5, 11} a difference set in $(Z_{12}, +)$?

Exercise 2.13

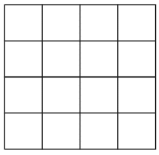
Using The QR(7) difference set, generate the (7, 3, 1) symmetric BIBD.

2.10 Solution & Hints

Exercise 2.4

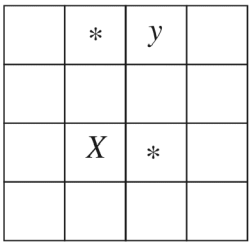
Solution :

The squares of 4 × 4 board has 16 squares. This means the no of points $v = 16$. Each square has a block defined, therefore, no of block $b = 16$. Each square



belongs to 6 blocks, since each square lies in a row with 3 other squares and in a column with 3 more squares. Thus, we also have $r = 6$. To determine λ let us choose a pair of squares x and y. There are three possibilities :

1. x and y are in the same row. Then x and y are together in the two blocks determined by the other two squares in their row.
2. x and y are in the same column. Then x and y are together in the two blocks determined by the other two squares in their column.
3. x and y are in different rows and in different columns. Then x and y are together in two blocks, one determined by the square at the intersection of the row of x and the column of y, the other determined by the intersection of the column of x and the row of y. The following array, where the blocks are those determined by the squares marked with an asterisk (*), is illustrative :



Therefore, in all the above 3 scenario $\lambda = 2$. So the design is a BIBD.

2.11 References and further reading

4. Combinational Design: Constructions and Analysis, Douglas R. Stinson, Springer.
5. Introductory Combinatorics, Richard A. Brualdi-Fifth Edition, Pearson Education Inc.
6. Linear Algebra and its Application,Third Edition,David C.Lay, Pearson Education Inc.
7. Combinatorial Methods with Computer Application, Jonathan L. Gross, Pearson Education Inc.

Unit 3 ❑ Symmetry Groups and Color Patterns

Structure

3.0 Objectives

3.1 Introduction

3.2 Review of Permutation Group

3.3 Groups of Symmetry and action of a group on a set

3.4 Colouring and Colouring Patterns

3.5 Polya's theorem and pattern inventory

3.6 Generating functions for non-isomorphic graphs.

3.7 Summary

3.8 Exercises

3.9 Reference and Further reading

3.0 Objectives

The main objective of the present chapter is to study groups of symmetry and action of a group on a set, colouring and colouring patterns, Polya theorem and pattern inventory, generating functions for non-isomorphic graphs.

3.1 Introduction

This chapter presents some interesting applications of abstract algebra to practical real-world problems. Whereas many applications of calculus are presented in undergraduate courses, usually no such applications are given in courses on abstract algebra. The object of this book is to fill this lacuna. It is hoped that this will make the study of abstract algebra more interesting and meaningful, especially for those whose interest in algebra is not confined to mere abstract theory.

3.2 Review of Permutation Group

Definition 3.2.1

Let S be a non-empty set. A per mutation of S is a bijective mapping of S onto itself. A group is called a permutation group on S if the elements of the group are some permutations of S and the group operation is the composition of two maps.

Example 3.2.1.1

Let X be a non-empty set and let Sx be the set of all bijective functions of X on to itself. Since e, being the identity map on Xi is bijective, e S. Thus Sx = . By usual compositions of function s oi it can be easily verified that (Sx, o) forms a group and the group is said to be the **permutation group** on X. Let us now consider permutation on a finite set. Let In denotes the finite set {1, 2, 3, 4, ..., n}. Any permutation on In is a bijectivemap from In to itself. Let Sn be the set of all permutations on In. Then (Sn, o) forms a group under usual compositions of functions o. This group is said to be the symmetric group on n elements. It is easy to see that $|S_n| = n!$.

Example 3.2.1.2

Let us consider the group S3, the elements of the group, all the permutations on I_3. As the number of bijective functions of I_3 on to itself is 6, we have $|S_3| = 6$. Now

$S_3 = \{e, \alpha, \beta, \gamma, \delta, \sigma\}$, where $e = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}$, $\alpha = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}$, $\beta = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix}$, $\gamma = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}$,

$\delta = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}$, $\sigma = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}$. It can be easily verified that under the compositions or product of the elements of S_2 viz

$\delta o \gamma = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix} o \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix} = \alpha, (S_2, o)$ will form a group.

Example 3.2.1.3

If $n \in Z^+$ such that n ≥ 3, then the symmetric group S_n is a non commutative group.

Definition 3.2.2

A permutation α on I_n = {1, 2, 3, ..., n} is called a k-cycle or cycle of length k if there exists distinct elements $i_1, i_2, \ldots\ldots, i_k$ in I_n such that

$\alpha(i_1) = i_2, \alpha(i_2) = i_3, \alpha(i_3) = i_4, \ldots\ldots\ldots, \alpha(i_{k-1}) = i_k,$

$\alpha(i_k) = i_1$ and $\alpha(x) = x \ \forall x \in I_n \setminus \{i_1, i_2, i_3, \ldots\ldots\ldots, i_k\}$.

A k-cycle with k = 2 is called a transposition.

Example 3.2.2.1

If $\alpha = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 5 & 2 & 4 & 1 & 6 \end{pmatrix}$ is a permutation on I_6 = {1, 2, 3, 4, 5, 6} such that

$$\alpha(1) = 3, \alpha(3) = 2, \alpha(2) = 5, \alpha(5) = 1, \alpha(4) = 4, \alpha(6) = 6.$$

Hence α is a 4 -cycle and it is denoted by (1325).

Definition 3.2.3

Two cycles $(i_1, i_2, i_3, \ldots\ldots, i_k)$ and $(j_1, j_2, j_3 \ldots\ldots, j_t)$ of S_n are said to bed is joint if $\{i_1, i_2, i_3, \ldots, i_k\} \cup \{j_1, j_2, j_3, \ldots, j_i\} = \phi$.

Theorem 3.2.3.2

Prove that product of two disjoint cycles in S_n is commutative.

Proof. Let $\alpha = (i_1\ i_2 \ldots . i_k)$ and $\beta = (j_1\ j_2 \ldots . j_k)$ be two disjoint cycles.

Claim : $(\alpha\ \beta)(x) = (\beta\ \alpha)(x)\ \ \forall x \in I_n$.

Now,

$$\begin{array}{lll} i_1 \mapsto i_2 & & j_1 \mapsto j_2 \\ i_2 \mapsto i_3 & & j_2 \mapsto j_3 \\ \vdots & & \vdots \\ \alpha : i_{t-1} \mapsto i_t & \beta : & j_{k-1} \mapsto j_k \\ i_t \mapsto i_1 & & j_k \mapsto i_1 \\ \text{when } y \notin \{i_1, i_2, \ldots\ldots, i_t\} & & \text{when } y \notin \{j_1, j_2, \ldots\ldots, j_k\} \end{array}$$

Suppose x is neither i_1 ; i_2 ; ....., i_k ; nor j_1, j_2, ..., j_k. Then $\alpha(x) = x$ and $\beta(x) = x$. Hence,

$$(\alpha\beta)(x) = \alpha\big(\beta(x)\big) = \alpha(x) = x,$$

$$(\beta\alpha)(x) = \beta\big(\alpha(x)\big) = \beta(x) = x.$$

Suppose now that x is one of i_1 ; i_2 ; ....., i_t. Hence, $x \notin \{j_1, j_2, \ldots, j_k\}$. Then $\beta x = x$ and $\alpha(x)$ is one of i_1, i_2 ; ....., i_t. Hence, $(\alpha\beta)(x) = \alpha\big(\beta(x)\big) = \alpha(x)$ and $(\beta\alpha)(x) = \beta\big(\alpha(x)\big) = \beta(x)$ as $\alpha(x) \notin \{j_1, j_2, \ldots\ldots, j_k\}$. Similarly, if x is one of j_1, j_2, ..., j_k, then $x \notin \{i_1, i_2, \ldots\ldots, i_k\}$, proceeding as above it can be shown that $(\alpha\beta)(x) = \beta(x) = (\beta\alpha)(x)$. Hence, $(\alpha\beta)(x) = (\beta\alpha)(x)\ \forall x \in I_n$.

Theorem 3.2.3.3

Any non-identity permutation $\alpha \in S_n (n \geq 2)$ can be expressed as a product of disjoint cycles, where each cycle is of length ≥ 2.

Proof. Let $\alpha \in S_n, n \geq 2$. We begin by considering $1, \alpha, \alpha(1), \alpha^2(1), \ldots..$ till we find the smallest positive integer r such that $\alpha r = 1$. This gives us a r-cycle, say α_1 such that

$$\alpha_2 = (1\alpha(1)\alpha^2(1).....\alpha^{r-1}(1)).\ ]$$

Let $i \in I_n$ such that i is the smallest integer, that does not appear in α_1. Then we consider $\alpha(i)$, $\alpha^2(1)$, ... so on until we come across the smallest positive integer s such that $\alpha^s(i) = i$. Evidently this gives us an s-cycle, say α^2, such that

$$\alpha_2 = (1\alpha(i)\alpha^2(i) ..... \alpha^{s-1}(i)).$$

Before proceeding further, it is to be noted that α^1, α^2 been constructed must have disjoint cycles, *i.e.*

$$(1\alpha(1)\alpha^2(1) ....\alpha^{r-1}(1)) \bigcap (1\alpha(i)\alpha^2(i).....\alpha^{e-1}(i)) = \phi .$$

Indeed otherwise, if $\alpha^p(i) = \alpha^k(1)$, for some p, k with $1 \le p \le s$, $1 \le k \le r$, then we must have $\alpha^{p+1}(i) = \alpha(\alpha^p(i)) = \alpha(\alpha^k(1)) = \alpha(\alpha^{k+1}(1))$ and so on, which in turn implies that $\alpha(\alpha^{p+i}(i)) = \alpha(\alpha^{k+t}(1))$ for $t = 1, 2....$ Now $\exists\ t$ such that $p + t = s$. Hence, for this t, $i = \alpha^s(i) = \alpha^{P+t}(i) = \alpha^{k+t}(1)$. This appears that i appears in α_1, a contradiction to the choice of i. Now if,

$$\{1,\alpha(1),\alpha^2(1), ..., \alpha^{r-1}(1)\} \bigcup \{1,\alpha(i),\alpha^2(i), ..., \alpha^{s-1}(i)\} \neq I_n,$$

then we consider the smallest number of In not appearing in the left hand side union above and continue the same process as before to construct the cycle α_3. Since, *In* is finite, the aforesaid process must terminate after finite steps, with some cycle, say, α_m. From the denition of the cycles α_1, α_2, ..., α_m, it follows that $\alpha = \alpha_1$ o α_2, ..., α_m, it follows that $\alpha = \alpha_1$ o α_2 o .... o α_m.

The uniqueness of the decomposition is treated as an exercise.

Example 3.3.3.4

Any non-identity permutation $\alpha \in S_n (n \ge 2)$ is either a transposition or can be expressed as a product of transposition. (Prove !)

Definition 3.2.4

Any non-identity permutation $\alpha \in S_n$ is called an even permutation if α can be expressed as aproduct of an evern number of transpositions and a permutation $\alpha \in S_n$ is called an odd permutation if a can be expressed as a product of an odd number of transpositions. The set of all even permutations in S_n forms a group and this group is known to be the afternating group of degree n, denoted by A_n.

It is to be noted that identity permutation is an even permutation. (Justify)

3.3 Groups of Symmetry

In this section we consider the application of group theory to study the symmetries of a geometrical figure (in a plans or three-dimensional space).

Definition 3.3.1

Suppose X is a two or three dimensional figure. Then any symmetry of X is defined to be a rigid motion that maps the figure to itself.

We write $Sym(X)$ to denote the set of all symmetries of X. Obviously, $Sym(X)$ is a subset of Sn, the set of all permutations of all verticles of X. It can be easitly checke that the composition of two symmetries is a symmetry and invese of a symmetry is a symmetry. So the following result is obvious.

Theorem 3.3.1

$Sym(X)$ is a subgroup of S_x.

Remark 3.3.1

The symmetric group Sx is defined for any non-empty set X, but Sym(X) is defined only for a figure.

Consider the symmetries of a polygon P. (By P we mean the set of points constituting the polygon.) Let V be the set of vertices of the polygon. It is clear from geometrical consideration that any symmetry of the polygon must map a vertex to a vertex.Thus ρ determines a symmetry $\bar{\rho}$ of the set V. Convesley given any symmetry $\bar{\rho}$ of V, it determines uniquely a symmetry ρ of the polygon that coincides with ρ on the verticlses. Hence we can identify the symmetries of the polygon with the symmetries of the set of its vertices. In other words, speaking more formally, the group of symmetries of P is isomorphic with the group of symmetries of V. Let us now consider a regular polygon of n sides ($n \geq 3$). Let us label the vertices in counterclockwise order as 1, 2, ..., n. Consider any symmetry ρ of the set of vertices. Suppose a maps vertex 1 to vertex i. Then a must take vertex 2 to a vertex adjacent to i – i.e., either $i + 1$ or $i - 1$. Once $\rho(1)$ and $\rho(2)$ are fixed, the mapping p is completely determined by the fact that it preservesthe distance between very two points. So if ρ maps 2 to $i + 1$, then it must map 3, 4, ... to $i + 2$, $i + 3$, ..., respectively. On the other hand, if ρ maps 2 to $i + 1$, then it must map 3, 4, ... to $i - 2$, $i - 3$, ..., respectively. Thus there are exactly two symmetries viz ρ_i and ρ_i, that take vertex 1 to i. These are given by

$$p_i = \begin{pmatrix} 1 & 2 & 3 & .... & n \\ i & i+1 & i+2 & .... & i-1 \end{pmatrix}$$

$$Q_i = \begin{pmatrix} 1 & 2 & 3 & .... & n \\ i & i-1 & i-2 & .... & i+1 \end{pmatrix}$$

Thus we see that a regular polygon of n sides has a in all $2n$ symmetries viz ρ_i, Q_i, $i = l, ..., n$.

The mapping ρ_i preserves the cyclic order of the verticss, but Q_i reverses the cyclic order. Geometrically, ρ_i represents a rotation of the polygon abut its cerner through an angle $\frac{2\pi(i-1)}{n}$, and Q_i represents a reflection in the diameter lying midway between vertices 1 and i. (By a rotation, we mean a rotation of the polygon in its own plane. A reflection in a diameter is equivalent to a rotation abou the diameter through an angle π, but this rotation takes place in the third dimension and not in the plane of the polygon.) It is obvious that a ρ_i is the identity permutation, and Q_i represents reflection in the diameter through vertex 1. The identity permutation is equivalent to a rotation through an angle 2π.

The $2n$ symmetris ρ_i, Q_i, i = 1, ..., n, can be expressed in terms of two basic symmetries. We write $\alpha = \rho_2, \beta = Q_i$, so

$$\alpha = \begin{pmatrix} 1 & 2 & \ldots & n \\ 2 & 3 & \ldots & 1 \end{pmatrix}, \beta = \begin{pmatrix} 1 & 2 & \ldots. & n \\ 1 & n & \ldots. & 2 \end{pmatrix}$$

Geometrically, a represents a rotation through angle $\frac{2\pi}{n}$ and moves each vertex i to i + 1. For any integer m = 1, ..., n, α^{m-1} represents a rotation angle $\frac{2\pi(m-l)}{n}$; hence α^{m-1} = ρm. Further $\alpha^{m-1}\beta(1) = \alpha^{m-1}(1) = m$ and $\alpha^{m-1}\beta(2) = \alpha^{m-1}(n) = m-1$. Since a symmetry is determined uniquely by its effect on vertices 1 and 2, it follows that $\alpha^{m-1}\beta = Q_m$. Thus the $2n$ symmetries are given by $\alpha^{m-1}\beta, m = 1, 2, ..., n$. It is clear that $\alpha^m = e$ and $\beta^m = e$. Further, consider $\beta\alpha$: $\beta\alpha(1) = \beta(2) = n$ and $\beta\alpha(2) = \beta(3) = n - 1$. Hence $\beta\alpha = Q_n = \alpha^{n-1}\beta$. Thus we have proved the following result.

Theorem 3.3.2

The group G of symmetries of a regular polygon of n sides is given by

G = {e, α, ..., α^{n-1}, β, $\alpha\beta$, ..., $\alpha^{n-1}\beta$}

where a represents a rotation through an angle $\frac{2\pi}{n}$, and β represents reflection in a diameter through a vertex. Moreover, the following relations hold in the group G :

$$\alpha^n = e, \beta^2 = e, \beta\alpha = \alpha^{n-1}\beta.$$

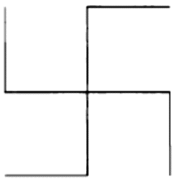


Figure 1 : The symmetry group of the figure is C_4.

Any group of $2n$ elements that has the same structure as the group G above is called a dihedral group of degree n and denoted by D_n. That is, we have the following definition.

Definition 3.3.2

A dihedral group of degree n, written D_n, is a group of order $2n$ given by

$$D_n = \{e, a, ..., a^{n-1}, b, ab, ..., a^{n-1}b\}$$

with the following defining relations :

$$a^n = e, b^2 = e, ba = a^{n-1}b.$$

We have thus shown that the group of symmetries of a regualr polygon of n sides ($n >$) is a dihedral grop of degree n. We shall shortly explain the geopmetrical intepretation of the dihedral groups D_1 and D_2 asgroups of symmetries. For the present, let us observe that $D_1 = e, b$, with $b^2 = e$, is a cyclic group of order 2. Further, $D_2 = \{e, a, b, ab\}$ has the defining relations $a^2 = e$, $b^2 = e$, and $ba = ab$. Hence D_2 is identical with a Klein's 4-group.

If we interpret the elements of the dihedral group D_n, $n > 2$, as permutations of the set $\{1, ..., n\}$ of vertices of a regular polygon, then D_n is a subgroup of the symmetric group S_n. In particular, D_3 has six elements and hence $D_3 = S_3$. For $n > 3$, Dn is a proper subgroup of S_n.

An equilateral triangle is a regular polygon of three sides. Hence its group of symmetries is D_3. Any permutation of the vertices of an equilateral triangle is a symmetry of the triangle. Hence S_3 and D_3 are the same.

Consider now an isosceles (but not equilateral) triangle. It has ony one symmetry b in addition to the identity permutation-namely, the one given by reflection in the median bisecting the angle between the two equal sides. So the group G of symmetries of an isosceles triangle is given by $G = \{e, \beta\}$, with $\beta^2 = e$. As noted above, G is a dihedral group of degree 1.

Consider next the symmetris of a rectangle (other than a square). It is easily seen that there are only three symmetries α, β, γ in addition to e. Geometrically, these represent a rotation through an angle π and refelction in the lines thoruhg the centre and parallel to the sides of the rectangle. Labeling the vertices as 1, 2, 3, 4 in order, we can write these symmetries as permutations of the vertices as follows :

$$\alpha = (13)(24), \beta = (12)(34), \gamma = (14)(23).$$

It is easily verified $\alpha\beta = \gamma = \beta\alpha$. Hence the group of symmetries of a rectangle is given by G = $\{e, \alpha, \beta, \alpha\beta\}$, with the defining relations $\alpha^2 = e$, $\beta^2 = e$, $\beta\alpha = \alpha\beta$. So G is a dihedral grop of degree 2. We can summarize the results proved above with few examples : The group of symmetries of an isosceles triangle is D_1. The group of symmetries of a rectangle is D_2. For any $n > 2$, the group of symmetries of a regular polygon of n sides is D_n. The dihedral group Dn. has a subgroup $C_n = \{e, a,$

..., a^{n-1}} that, geometrically, consists of all rotational symmetries of the polygon.

The symmetry group of a geometric figure may be infinite. For example, a circle has infinitely many symmetries. It can be shown that if the symmetry group G of a plane figure is finite, then G is either D_n or C_n for some n.

Example 3.3.1

Let us now consider the symmetries of three-dimensional geometric objects. Consider first a regular tetrahedron.Let the vertices be labeled as 1, 2, 3, 4. As in the case of an equlateral triangle, the distance between every two vertices of a regular tetrahedron is the same. Hence every permutation of the vertices is a symmetry. Therefore the symmetry group of a regular tetrahedron is S4. How many of the 24 permutations in S_4 are rotations ? It is clear that by a suitable rotation we can take vertex 1 to any vertex i = 1, 2, 3, 4. Having done that, we can rotate the tetrahedron about an axis through the new position of vertex 1 through angles $0, \frac{2\pi}{3}$, and $\frac{4\pi}{3}$ to obtain three symmetries. Thus there are in all 4 × 3 = 12 rotational symmetries. They form a subgroup of the group of all symmetries of the tetrahedron. The following table gives the 12 rotational symmetries of a regular tetrahedron as permutations of the vertices and their geometric description as rotations. The edge $i - j$ denotes the edge joining vertices i and j.

Permutations	**Axis and Angle of Rotation**
(1) = e	any axis,rotation through angle 2π
(234), (243)	axis through vertex 1, angles $\frac{2\pi}{3}$ and $\frac{4\pi}{3}$
(134), (143)	axis through vertex 2, angles $\frac{2\pi}{3}$ and $\frac{4\pi}{3}$
(124), (142)	axis through vertex 3, angles $\frac{2\pi}{3}$ and $\frac{4\pi}{3}$
(123),(132)	axis through vertex 4, angles $\frac{2\pi}{3}$ and $\frac{4\pi}{3}$
(12) (3 4)	axis through middle points of edges 1-2 and 3-4, angle π
(1 3) (2 4)	axis through middle points of edges 1-3 and 2-4, angle π
(14) (23)	axis through middle points of edges 1-4 and 2-3, angle π

Example 3.3.2

Let us now consider the symmetries of a cube. Let the vertices of the cube be labeled 1, 2, ..., 8 such that vertices 2, 3, and 4 are adjacent to vertex 1. Let $\rho \in S_8$ be a symmetry. Suppose ρ takes 1 to i. Then must take 2, 3, and 4 to the three vertices adjacent to i, which can be done in 6 ways. Hence there are $8 \times 6 = 48$ symmetries in all. Of these, 24 are rotational symmetries. The vertex 1 can be taken to any vertex i by a rotation, and then we can rotate the cube around the diameter through the new position of vertex 1 to obtain three symmetries.

We can arrive at the same result by considering the symmetries of the cube as per mutations of its six faces. By a rotation, face 1 can be taken to face i (i = 1, ..., 6). Having done that, we can rotate the cube around the diameter perpendicular to the new position of the face 1 through angles $\frac{\pi}{2}$ $\pi$and $\frac{3\pi}{2}$ and obtain three symmetries. Hence there are $6 \times 4 = 24$ rotational symmetries.The following table gives a geometric description of the various types of rotational symmetries of a cube and their numbers.

Axis and Angle of Rotation	Number
any axis,angle 2π	1
axis through opposite vertices, angles $\frac{2\pi}{3}$ and $\frac{4\pi}{3}$	8
axis through centers of opposite faces, angles $\frac{\pi}{2}, \pi$, and $\frac{4\pi}{3}$	9
axis through middle points of opposite edges, angle π	6

Action of a group on a set

Definition 3.3.3

Let (G, o) be a group and A be a non empty set, then G is said to act on A if there exists a function $*: G \times A \mapsto A$ satisfying

- $(g_1 \text{ o } g_2) * a = g_i * (g_2 * a)$
- $e * a = a \ \forall \ g_1, g_2 \in G$ and $a \in$ A.

The mapping * is called a group action of G(or a G action) on A and A is called a G-set andwe express it by saying G acts on A. Similarly we can define action of G on A on the right by considering the map from A × G → A satisfying

- $a * (g_1 \circ g_2) = (a * g_1) * g_2$
- $a * e = a \forall g_1, g_2 \in G$ and $a \in A$.

Example 3.3.3

Let G be a group and A be any non empty set. Define * : G × A → A A such that $g * a = a$, $\forall a \in A, g \in G$. It can be easily veried that under *, G acts on A and A is a G-set.

Example 3.3.4

Let (G, o) be any group and take $A = G$. Define * by $g * a = g \circ a$, $g \in G$, $a \in A = G$. Then * is a group action since

$$(g_1 \circ g_2) * a = (g_1 \circ g_2) \circ a = g_1 \circ (g_2 \circ a) = g_3 * (g_2 \circ a) = g_1 * (g_2 * a),$$

$$e * a = e \circ a = a, \forall g_1, g_2 \in G, a \in A.$$

Example 3.3.5

Let (G o) be any group and take $A = G$. Define * by $g * a = a \circ g^{-1}$, $g \in G$, $a \in A$. Then

$$(g_1 \circ g_2) * a = a \circ (g_1 \circ g_2)^{-1} = (a \circ g_2^{-1}) \circ g_1^{-1} = g_1 * (g_2 * a),$$

$$e * a = a \circ e^{-1} = a.$$

Hence, * is a group action and is sometimes called as regular action of G on itself.

Example 3.3.6

Given any nonempty set X, let G be a subgroup of the symmetric group S_X. For any $g \in$ G and $x \in X$, we define $g * x = g(x)$. Then it follows from conditions in Definition (3.3.3) that * is an action of G on X. We say in this case that G acts naturally on X. In particular, if G is the group of symmetries of a set X of points in space, then G acts naturally or cannonically on X.

Theorem 3.3.3

If a group (G, o) acts on a set X, then G determines a subgroup of S_X that is homomorphic image of S_X.

Proof. Let G be a group acting on a set X. Given $g \in G$, we define a map $\eta_g : X \to X$ by the rule $\eta_g(x) = g * x$. Then η_g is a per mutation of X. Let $x, y \in X$, then $\eta_g(x) = \eta_g(y) \Rightarrow g * x = g * y \Rightarrow g^{-1} * (g * x) = g^{-1} * (g * y) \Rightarrow (g^{-1} \circ g) * x = (g^{-1} \circ g) * y \Rightarrow e * x = e * y \Rightarrow x = y$. Hence, η_g is injective. Further, given $y \in X$, let $x = g^{-1} * y$. Then, $\eta_g(x) = g * x = g * (g^{-1} \circ y) = (g^{-1} * g) * y =$

$e * y$. Hence, η_g is surjective. Consider now the mapping $\phi : G \to S_X$ defined as $g \to \eta_g$. Let $g, h \in$ G. Then for all $x \in X$, $\eta_{goh}(x) = (g \circ h) * x = g * (h * x) = \eta_g(\eta_h(x)) = (\eta_g \eta_h)(x)$. Hence, $\phi(g \circ h) = \phi(g)\phi(h)$ which proves ϕ is a homomorphism. Hence, *Im* ϕ is a subgroup of S_X and a homomorphic image of G.

Kernel of an Action

Definition 3.3.4

Let $* : G \times A \to A$ be a group action, then Kernel of * is defined to be the set

$$Ker(*) = \{g \in G | \ g * a = a \ , \ \forall a \in A\}$$

It can be easily verified that $Ker(*)$ is a subgroup of G.

Orbits And Stabilizers

Definition 3.3.5

Let (G, o) be a group acting on a set A under *. Let $a \in A$ be any fixed element. Then the set

$$Stab(a) = G_a = \{g \in G | \ g * a = a\}$$

is called the stabilizer of a in G. Then indeed G_a is a subgroup of G(verify!).

Definition 3.3.6

Let G be a group acting on a set A under *. For any $a \in A$, let

$$Orb(a) = Ga = \{x \in A | \ x = g * a \text{ for some } g \in G_g = \{g * a | \ g \in G\}.$$

Then Ga is called an orbit of a under G.

Remark 3.3.2

Since $e * a = a (\in Ga)$, orbit of any element of A is a non-empty subset of A.

Problem 3.3.1

Let (G, o) act on a set A under *. For any $a, b \in A$, define $a \sim b$ iff $\exists \ g \in G$, such that $a = g * b$. Show that is an equivalence relation and for any $a \in A$, the equivalence class of a (denoted by [a]) is the orbit of a in G.

Solution 3.3.1

- Reflexivity follows as $e * a = a \ , \ \forall a \in A$, thus $a \sim a \ , \ \forall a \in A$.
- For symmetry, assume that $a \sim b \Rightarrow \exists \ g \in G$, *s.t.* $a = g * b$. Now $g^{-1} * a = g^{-1} * (g * b) = (g^{-1} \circ g) * b = e * b = b \Rightarrow b \sim a$.
- Transitivity is also true. (verify!)

Hence, $\sim$ is an equivalence relation.

Let $a \in A$ be any element, then equivalence class of a is given by

$$[a] = \{x \in A \mid x \sim A\} = \{x \in A \mid x = g * A \text{ for some } g \in G\}$$

$= \{g * a \mid g \in G\}$

which is the orbit of a under G.

Theorem 3.3.4

Let (G, o) be a group acting on a set X and $x \in$ X. Then the index of the subgroup G_x in G is $[G : G_x] = |Gx|$.

Proof. As usual, $G = G_x$ denote the set of all left cosets of G_x in G. We define the map $\phi : G = Gx \to Gx$ by $\phi(gG_x) = g * x$.

Claim : ϕ is bijective. If $gG_x = hG_x$, then $g^{-1}h \in S$, hence $g^{-1}hx = x$ and therefore $g * x = g(g^{-1}hx) = h * x$. This shows that ϕ is well-dened. If $\phi(gG_x) = \phi(hG_x)$, then $g * x = h * x$ and hence $g^{-1}hx = x$, which implies $g^{-1}h \in G_x$, which implies $gG_x = hG_x$ whence ϕ is injective. If $y \in Gx$, then $y = g * x = \phi(gG_x)$ for some $g \in G$. Hence, ϕ is surjective implies ϕ is bijective. Hence the proof.

Theorem 3.3.5

(BURNSIDE THEOREM) : Let G be a finite group acting on a finite set X. Then the number κ of orbits in X under G is given by

$$\kappa = \frac{1}{|G|}\sum_{g \in G} F(g).$$

where $F(g) = |Fix(g) = \{x \in X | g * x = x\}$ is the number of elements in X that are fixed by g.

Proof. We count in two ways the number of ordered pairs $(g, x) \in G \times X$ such that g fixes x. We take

$$P = \{(g, x) \in G \times X | gx = x\}$$

If $g * x = x$, then $g \in G_x$ and $x \in Fix(g)$. Hence, given $x \in X$, the number of elements that fix x equals $|G_x|$. On the other hand, given $g \in G$, the number of elements in X that are fixed by g equals $|Fix(g)|$. Hence,

$$(3.1) \qquad \sum_{x \in X} |G_x| = |P| = \sum_{g \in G} |Fix(g)|.$$

Since, the index of the subgroup G_x in G is $|G : G_x| = |Gx|$, therefore,

$$|Gx| = [G : G_x] = \frac{|G|}{|G_x|}.$$

Hence,

$$(3.2) \qquad \sum_{x \in X} |G_x| = |G| \sum_{x \in X} \frac{1}{|G_x|}$$

Now for any orbit $T \in X/G$, we have

$$\sum_{x \in T} \frac{1}{|G_x|} = \sum_{x \in T} \frac{1}{|T|} = |T| \frac{1}{|T|} = 1.$$

Therefore, since the orbits in X form a partition of X,

$$\sum_{x \in X} \frac{1}{|Gx|} = \sum_{T \in x/G} \sum x \in T \frac{1}{|Gx|} = |X/G| = \kappa .$$

Hence, using (3.1) and (3.2), we obtain

$$\kappa = \sum_{x \in X} \frac{1}{|Gx|} = \frac{1}{|G|} \sum_{x \in X} |G_x| = \frac{1}{|G|} \sum_{g \in G} F(g) .$$

3.4 Colorings and Color Patterns

Pattern is repetition of line, form, shape, texture, etc. Since color can create form, shape, texture it may similarly create pattern. Pattern can be symmetric or a symmetric.In art and design color pattern plays very significant role. Color patterns can be found in nature as well as are manmade. Suppose that you are given a graph G with n vertices and are asked to paint its vertices such that no adjacent vertices have the same color. What is the minimum number of colors that you would require? This constitutes a coloring problem. Having painted the vertices, you can group them into different sets-one set consisting of all red vertices, another of blue, and so forth. This is a partitioning problem. The coloring and partitioning can, of course, be

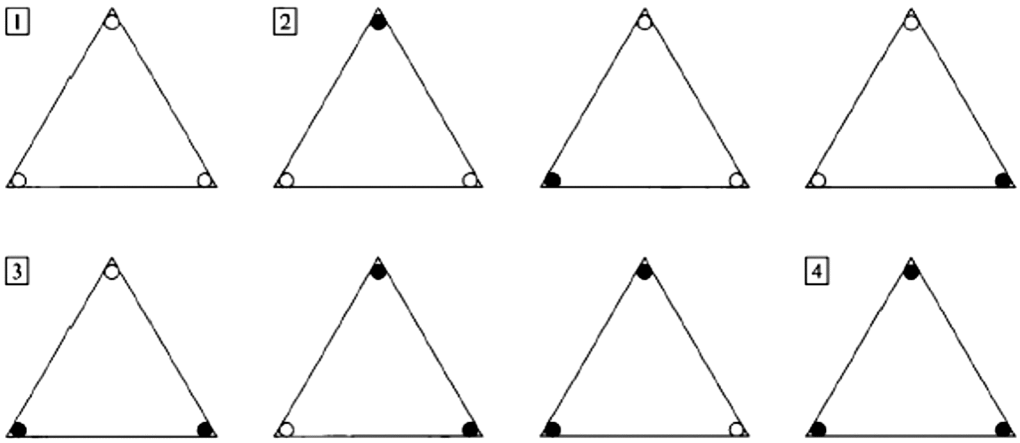


Figure 2 : Equilateral Triangle

performed on edges or vertices of a graph. The coloring and partitioning of vertices (or edges) is not performed out of mere playfulness. Partitioning is applicable to many practical problems such as coding theory, partitioning of logic in digital computers and state reduction of sequential machines.

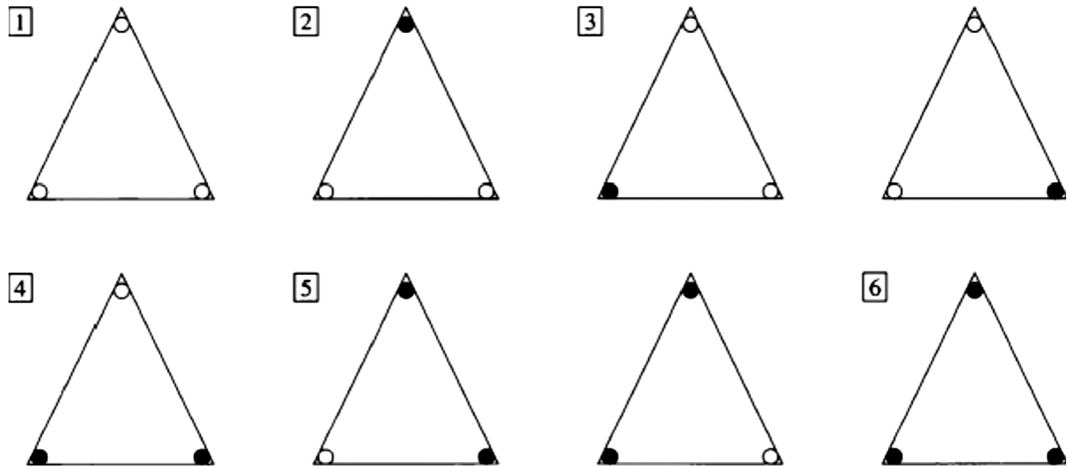


Figure 3 : Isosceles Triangle

Suppose we color each vertex of an equilateral triangle white or black. Then there are 8 ways in which the three vertices can be colored. Let us refer to them as color assignments or colorings. We say that two color assignments are equivalent (or have the same pattern) if one of them can be obtained from the other by rotating the triangle through an appropriate angle or flipping it over. The second operation-namely, flipping over-is equivalent to reflection in some mirror line. We then find that the eight color assignments fall into four distinct patterns, as shown here.

If we consider an isosceles triangle, then we find that the eight color assignments fall into six distinct patterns.

Finally, if we consider a triangle whose sides are all of unequal lengths, then no two colorings are equivalent, and hence all eight colorings are distinct patterns. On the basis of the above examples let us formulate a general problem.

Let $|S| = n$, whose elements are specified points or parts of some given geometric figure. Let $|T|$ = m, be the set of m colors. If each element in S is assigned a color from the set T. The total number of ways in which such color assignments can be made is m×m×..... ×m (n times) = mn. The problem is to find the number of distinct patterns in which these mn color assignments can be considered. To identify distinct color patterns, we may use either the weaker conditions *i.e.* both rotations and reections or the finer conditions *i.e.* only rotations.The number of distinct color

patterns under the coarser (weak) conditions is less than or equal to the number under the finer (strong) conditions (verify!). From the examples discussed above, the number of distinct color patterns (*i.e.* mn) depends both on the numbers *m, n* and on the symmetry properties of the under lying geometric figure. The symmetry property possessed by the figure is directly proportional to the number of equivalent pairs of colorings and inversely proportional to the number of distinct color patterns.

Let X denote the set of all m-colorings of S. Let G be a group of symmetries of the set S. The group G acts naturally on the set S. Therefore G also acts on the set X. Given $g \in 2\ G$ and $x \in X$, gx represents the color assignment obtained by performing on the coloring x the symmetry operation (rotation or reflection) represented by g. Two color assignments x and y are equivalent if and only if $y = gx$ for some $g \in G$. Hence all color assignments that are equivalent to x lie in the orbit of x under G. Each orbit represents a color pattern.Thus by virtue of the action of the group G, the number of distinct patterns is equal to the number of orbits in the set X. This number is given by Burnside the orem. Suppose the points in the set S are coplanar. Then the group of symmetries of S is either a dihedral group

$$D_q = \{e, \alpha, \ldots\ldots\ldots, \alpha^{q-1}, \beta, \alpha\beta, \ldots\ldots\ldots, \alpha^{q-1}\beta\}$$

where α represents a rotation through an angle $\frac{2\pi}{q}$ and β is the reflection or its cyclic subgroup $C_q = \{e, \alpha, \alpha^{q-1}\}$ consisting of all rotations in D_q.

Problem 3.4.1

Each vertex of an equilateral triangle is colored by one of m given colors. Find the number of distinct patterns among all possible colorings.

Solution 3.4.1

Since each vertex can be colored in m ways, the total number of color assignments is m^3. The group G of symmetries of an equilateral triangle is the dihedral group of degree 3 ; that is, $G = D_3 = \{e, \alpha, \alpha^2, \beta, \alpha\beta, \alpha^2\beta\}$ where α represents a rotation through angle $\frac{2\pi}{q}$ and β is a reflection in a diameter. Since, every color assignment is invariant under the identity e; hence $F(e) = m^3$. To find the number of color assignments invariant under the other elements of G, let us number the vertices 1, 2, and 3. Then α takes vertex 1 to 2, 2 to 3, and 3 to 1. If a color assignment is invariant under α, the nall three vertices must have the same color. This common color can be any one of the m given colors. Hence there are m color assignments that are invariant under α, so $F(\alpha) = m$. The same reason in g applies to α^2, so $F(\alpha^2) = m$. If β is the reection in the diameter passing through vertex 1. Then β takes vertex 2 to 3 and 3 to 2. If a color assignment is

in variant under β, then vertices 2 and 3 must have the same color, so vertices 1 and 2 can have arbitrary colors. Hence the number of color assignments invariant under β is m^2. The same argument holds for the other two reflections $\alpha\beta$ and $\alpha^2\beta$. Hence F(β) = F($\alpha\beta$) = F($\alpha^2\beta$) = m^2. Using Burnside the orem, the number of patterns (that is, the number of orbits under G) is

$$\kappa = \frac{1}{|G|}\{F(e)+F(\alpha)+F(\alpha^2)+F(\beta)+F(\alpha\beta)+F(\alpha^2\beta)\}$$

$$= \frac{1}{6}(m^3+3m^2+2m).$$

To find the number of patterns under the finer criterion of rotations only, we take the group of rotations H is

$$\kappa' = \frac{1}{|H|}(F(e)+F(\alpha)+F(\alpha^2) = \frac{1}{3}(m^3+m^2).$$

In the particular case of only two colors, on putting $m = 2$ in the above results, we obtain $\kappa = \kappa' = 4$. In the case $m = 3$, we have $\kappa = 10, \kappa' = 11$.

Problem 3.4.2

A rectangular dining table seats six persons, two along each longer side and one on each shorter side. A colored napkin, having one of m given colors, is placed for each person. Find the number of distinct patterns among all possible color assignments.

Solution 3.4.2

For rectangle, the group of symmetries is

$$G = D_2 = \{e, \alpha, \beta, \alpha\beta\}$$

where α is a rotation through angle π, and β is a reflection. Let us take to be the reflection in the line through the center parallel to the longer side of the rectangle. Then represents the reection in the line parallel to the shorter side. Let us number the six napkins on the dinning table as follows

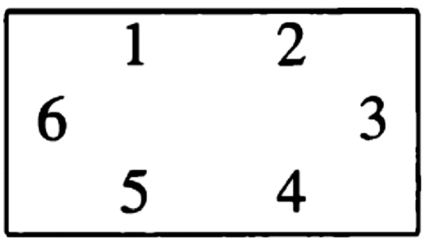


Figure 4 : Dining Table

Then takes napkin 1 to 4, 2 to 5, 3 to 6, and viceversa. If a color assignment is invariant under the rotation, then the napkins 1 and 4 must have the same color, 2 and 5 must have the same color, and 3 and 6 must have the same color. So we can assign arbitrary colors to napkins 1, 2, and 3. Hence the number of color assignments in variant under is m^3. Now β keep napkins 3 and 6 fixed, takes 14 to 5, 2 to 4, and viceversa. If a color assignment is invariant under β, then napkins 1 and 5 must have the same color, and 2 and 4 must have the same color. So we can assign arbitrary colors to napkins 1, 2, 3, and 6. Hence the number of color assignments invariant under is m^4. By a similar reasoning, we find that the number of color assignments in variant under is m^3. Therefore, by virtue of Burnside theorem, the number of patterns is

$$\kappa = \frac{1}{|G|}\{F(e)+F(\alpha)+F(\alpha\beta)\}$$

$$= \frac{1}{6}(m^6+m^4+2m^3).$$

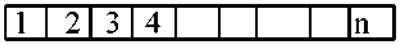
The number of patterns under the finer criterion of rotations only is

$$\kappa' = \frac{1}{2}\{F(e)+F(\alpha)\} = \frac{1}{2}(m^6+m^4).$$

In the particular case of two colors, we have $\kappa = 24, \kappa' = 36$.

Problem 3.4.3

A straight necktie in the form of along rectangular strip is divided in to n bands of equal width parallel to the shorter side. Each band is colored by one of m given colors. Find the number of ties with distinct patterns.



Solution 3.4.3

The group of symmetries of a rectangle is the dihedral group D_2. But in the present case the reection in the line parallel to the longer side doesn't play any role. The relevant group here is $G = D_1 = \{e, \alpha\}$, where α may represent a rotation through angle π', or a reflection in the line through the center parallel to the shorter side of the rectangle. (The two operations are equivalent in this case.) If a color assignment is invariant under α, then the bands 1 and n must have the same color, the bands 2 and $n - 1$ must have the same color, and so on. In general, the bands i and $n + 1 - i$ must have the same color. If n is even, we can assign arbitrary colors to bands 1, ..., $\frac{n}{2}$ hence $F(\alpha) = m^{\frac{n}{2}}$. But if n is odd, the bands 1, ..., $\frac{n+1}{2}$ can

be assigned arbitrary colors. (The $\frac{n+1}{2}$ th band is the band in the middle.) Hence $F(\alpha) = m^{\frac{n}{2}}$. Therefore, by Burnside theorem, the number of patterns is

$$\kappa = \frac{1}{|G|}\{F(e)+F(\alpha)\}$$

$$= \begin{cases} \frac{1}{2}(m^n + m^{\frac{n}{2}}), & \text{if } n \text{ is even;} \\ \frac{1}{2}(m^n + m^{\frac{(n+1)}{2}}), & \text{if } n \text{ is odd.} \end{cases}$$

Theorem 3.4.1

Let G be a finite group acting on a finite set S, let C be a finite set of m elements, and let $X = C^S$ be the set of mappings from S to C. Let $g \in G$ be fixed. Then the number of elements in X(*w. r. t g*) is given by $F(g) = m^{\lambda(g)}$ where λ(g) is the number of disjoint cycles (including cycles of length 1) in the cycle decomposition of the permutation η_g of S induced by g. Consequently, the number of orbits in X under the action of G is given by

$$\kappa = \frac{1}{|G|}\sum_{g\in G} m^{\lambda(g)}.$$

Proof. Let $g \in G$ and $f \in X$. If $g * f = f$, then $f(s) = (g * f)(s) = f(g^{-1}(s)) \ \forall s \in S$. Hence, $f(g^{-1}(s)) = f(g{-}1(g(s))) = f(s), \ \forall s \in S$.

Conversely, if $f(g(s)) = f(s), \ \forall s \in S$, then $(g * f)(s) = f(g^{-1}(s)) = f(gg^{-1}(s)) = f(s) \forall s \in S$; hence $g * f = f$. Thus $f \in Fix(g) \Leftrightarrow f(g(s)) = f(s) \ \forall s \in S$.

Let η_g be the permutation of S determined by g *i.e.* $\forall s \in S$; $\eta_g(s) = gs$. Let $\eta_g = \alpha_1\alpha_2 \ldots\ldots \alpha_\lambda$ be the decomposition of η_g into disjoint cycles. Any cycle in this decomposition is of the form $\alpha = (agag^2a \ldots\ldots g^{r-1}a)$. If $f \in$ Fix(g), then $f(a) = f(ga)$ = $\ldots\ldots$ $f(g^{r-1}a)$; hence f is constant on the elements in the cycle α. This holds for every cycle α, in the decomposition of η_g.

Conversely, if f is constant on every cycle α_i, then $f(g(s)) = f(s), \ \forall s \in S$. Hence $f \in$ Fix(g) if and only if f is constant on each cycle in the decomposition of η_g.

Let $f \in$ Fix(g) and $f_1, f_2, f_3, \ldots\ldots\ldots, f_\lambda \in$ C be the values of f in the cycles α_1 α_2 $\ldots\ldots\alpha_\lambda$ respectively. Then $f_1, f_{12}, f_3, \ldots\ldots, f_\lambda$ can be chosen in m different ways. Hence, $|Fix(G)| = m^\lambda$. This proves the first part of the theorem. The second part of the theorem follows by Burnside theorem.

3.5 Polya's Theorem and Pattom Inventory

Polya's Counting Theory is a spectacular tool that allows us to count the number of distinct items given a certain number of colors or other characteristics. Basic questions we might ask are, "How many distinct squares can be made with blue or yellow vertices?" or "How many necklaces with n beads can we create with clear and solid beads? We will count two objects as'the same' if they can be rotated or flipped to produce the same configuration. While these questions may seem uncomplicated, there is a lot of mathematical machinery behind them.Thus, in addition to counting all possible positions for each weight, we must be sure to not recount the configuration again if it is actually the same as another. We can use Burnside's Lemma to enumerate the number of distinct objects. However, sometimes we will also want to know more information about the characteristics of these distinct objects. Polya's Counting Theory is uniquely useful because it will act as a picture function-actually producing a polynomial that demonstrates what the die rent configurations are, and how many of each exist. As such, it has numerous applications.Some that will be explored include chemical isomer enumeration, graph theory and music theory.

Let G be a group acting on a set X. Let $R = Q[t_1, t_2, \ldots\ldots, t_q]$ be the set of all polynomial sin some given in determinates $t_1, t_2, \ldots\ldots, t_q$ with rational coeficients. A mapping $\omega : X \mapsto R$ is called a weight function on X under G if $\omega(g(x)) = \omega(x)$; $\forall\, g \in G; x \in X$. If this condition holds, then every element in the orbit $T = Orb(x) = Gx$ has the same weight $\omega(x)$. The common weight of all elements in an orbit T is called the weight of T and written as $\omega(T)$. The following theorem, known as weighted Burnside theorem, is a generalization of **Theorem(3.3.5)**.

Theorem 3.5.1. Let G be a finite group acting on a finite set X. Let $\omega : X \to R$ be a weight function on X under G. Then the sum of the orbits in X under G is

$$\sum_{T \in X/G} \omega(T) = \frac{1}{|G|} \sum_{g \in G} \sum_{x \in Fix(g)} \omega(x)$$

Proof. From Burnside theorem, let

$$P = \{(g, x) \in G \times X \mid gx = x\}.$$

Now,
$$S = \sum_{(g,x) \in P} \omega(gx)$$

can be computed in two ways. On one way,

$$(5.1) \quad S = \sum_{g \in G} \sum_{x \in Fix(g)} \omega(gx) = \sum_{g \in G} \sum_{x \in Fix(g)} \omega(x)$$

On another way,

(5.2) $$S = \sum_{x \in X} \sum_{g \in Stab(x)} \omega(gx) = \sum_{x \in X} |Stab(x)| \omega(x).$$

Since, the index of the subgroup G_x in G is $|G : G_x| = |Gx|$, we obtain

$$|Stab(x)| = \frac{|G|}{Orb(x)}.$$

Moreover, by Theorem (3.3.5), for any orbit $T \in X/G$, we have

$$\sum_{x \in T} \frac{1}{|Orb(x)|} = 1$$

Thus,
$$\sum_{x \in X} |Stab(x)| \omega(x) = \frac{|G|}{Orb(x)} \omega(x)$$

$$= |G| \omega(T) \sum_{x \in T} \frac{1}{|Orb(x)|} = |G| \omega(T).$$

Therefore, since the orbits form a partition of X, we have

$$\sum_{x \in X} |Stab(x)| \omega(x) = \sum_{T \in X/G} \sum_{x \in T} |Stab(x) \omega(x)| = |G| \sum_{T \in X/G} \omega(T).$$

Hence, from equations (5.1) and (5.2),we obtain

$$\sum_{T \in X/G} \omega(T) = \frac{1}{|G|} \sum_{g \in G} \sum_{x \in Fix(g)} \omega(x)$$

Remark 3.5.1

In particular, if we take $\omega(x) = 1$ for every $x \in X$ in Theorem 3.5.1, we recover the original Burnside theorem. Then $\omega(T) = 1$ for every orbit T, and so the left-hand side gives the number κ of orbits.

Theorem 3.5.2

(Polya's Theorem) Let G be a finite group acting on a finite set S having n elements, let C be a finite nonempty set, and let $X = C^S$ be the set of all mappings from $S \to C$. Let $u : C \to R$, and let $w : X \to R$ be the weight function on X under G induced by u. Then the sum of the weights of the orbits in X under G is

$$\sum_{T \in X/G} \omega(T) = \frac{1}{|G|} \sum_{g \in G} \prod_{i=1}^{n} \left(\sum_{c \in C} (u(c))^i \right)^{\lambda_i(g)}$$

where $\lambda_i(g)$ is the number of cycles of length $i (i = 1, \ldots, n)$ in the cycle decomposition of the permutation η_g of S induced by g.

Proof. Let $g \in G$ and $\eta_g = \alpha_1 \alpha_2 \ldots \alpha_r$ be the cycle decomposition of η_g. Let $f \in X$. Then by Theorem (3.4.1), $f \in Fix(g) \Leftrightarrow f$ is constant in each cycle α_j. Let f_j be the value of f on the cycle α_j $(j = 1, 2, \ldots r)$. Then,

$$\omega(f) = \prod_{s \in S} u(f(s)) = \prod_{j=1}^{r} (u(f_j))^{|\alpha_j|}$$

where α_j denotes length of the cycle α_j. Hence,

$$\sum_{f \in Fix(g)} \omega(f) = \sum_{f_1 \in C} \ldots \ldots \sum_{f_r \in C} \prod_{j=1}^{r} (u(f_j))^{|\alpha_j|}$$

Changing the order of summation and multiplication on the right-hand side of the above equation, we have

$$\sum_{f \in Fix(g)} \omega(f) = \prod_{j=1}^{r} \sum_{c \in C} (u(c))^{|\alpha_j|} = \prod_{i=1}^{n} \left(\sum_{c \in C} (u(c))^{i} \right)^{\lambda_i(g)}$$

where $\lambda_i(g)$ is the number of cycles of length i in the cycle decomposition of η_g $(i = 1, \ldots\ldots, n)$. Hence, by the weighted Burnside the orem, the sum of the weights of the orbits in X is

$$\sum_{T \in X/G} \omega(T) = \frac{1}{|G|} \sum_{g \in G} \sum_{x \in Fix(g)} \omega(f)$$

$$= \frac{1}{|G|} \sum_{g \in G} \prod_{i=1}^{n} \left(\sum_{c \in C} (u(c))^{i} \right)^{\lambda_i(g)}$$

Remark 3.5.2

In particular, if in Polya's theorem we take $u(c) = 1$ for each $c \in C$, then were cover Theorem 3.5.1.

Corollary 3.5.1

Prove that

$$(5.3) \quad \sum_{T \in X/G} \omega(T) = \frac{1}{|G|} \sum_{g \in G} \prod_{i=1}^{n} (t_1^i + t_2^i + t_3^i \ldots\ldots + t_m^i)^{\lambda_i(g)}$$

where symbol shave their usual meaning.

Proof. Taking $u(c) = 1$; $\forall$ C in Polya's theorem, we have $w(T) = 1$, $\forall T \in X = G$ and further assuming $|C| = m$, we have

$$\prod_{i=1}^{n}\left(\sum_{c\in C}(u(c))^i\right)^{\lambda_i(g)} = \prod_{i=1}^{n}(m)^{\lambda_i(g)} = m^{\lambda(g)}$$

where $\lambda_{(g)} = \sum_{i=1}^{n}\lambda_i(g)$ is the total number of cycles in the decomposition of η_g.

Now, let C = $\{c_1, c_2, \ldots\ldots, c_m\}$ be the set of m colors and $R = Q|t_1, t_2, \ldots\ldots, t_q|$ be the set of all polynomials in some given indeterminates $t_1, t_2, \ldots\ldots, t_q$ with rational coeficients. Let ω be the weight function on X induced by the mapping $u : C \to R$ with $u(c_i) = t_i$, $i = 1, 2, 3, \ldots\ldots, m$. Then applying Polya's theorem, we have the desired result.

Pattern Inventory

Consider a color assignment $f : S \to C$ in which the colors $c_1, c_2, \ldots\ldots, c_m$ occur with frequencies $\beta_1, \beta_2, \ldots\ldots, \beta_m$, where $\beta_i \in Z^+(i = 1,2,3,\ldots\ldots,m)$ such that $\beta_1 + \beta_2 + \ldots\ldots + \beta_m = n$. So β_i (i = 1, 2, 3, ..., m) is the number of elements $s \in S$ such that $f(c_i) = s$. Therefore the weight of f and hence also the weight of the orbit T containing f are

$$\omega(T) = w(f) = \prod_{s\in S} u(f(s)) = t_1^{\beta_i}\ldots\ldots t_m^{\beta_m}$$

Hence the sum of the weights of the orbits in X is equal to

$$\sum_{T\in X/G}\omega(T) = \sum p(\beta_1, \beta_2, \ldots\ldots\ldots\ldots, \beta_m)^{t_1^{\beta_1}\ldots\ldots t_m^{\beta_m}} \quad (6.1)$$

where $p(\beta_1, \beta_2, \ldots\ldots, \beta_m)$ denotes the number of orbit shaving the same weight $t_1^{\beta_1}\ldots\ldots t_m^{\beta_m}$, and the summation on the right-hand side is overall m-tuples $(\beta_1, \beta_2, \ldots\ldots, \beta_m)$ of non negative integers such that $\sum_{i=1}^{m}\beta_i = n$. Equivalently, in terms of colorings, $p(\beta_1, \beta_2, \ldots\ldots, \beta_m)$ is the number of patterns in which the colors $c_1, c_2, \ldots\ldots, c_m$ occur with frequencies $\beta_1, \beta_2, \ldots\ldots, \beta_m$ respectively. The polynomial on the right-hand side of equation 6.1 is a homogeneous polynomial of

degree n in m indeterminates. We denote it by $P_{X/G}(t_1, t_2, \ldots\ldots, t_m)$ and refer to it as the pattern inventory of orbits in X under G. Thus the number of patterns with given color frequencies $\beta_1, \beta_2, \ldots\ldots, \beta_m$ is the coeficient of the monomial $t_1^{\beta_1}\ldots\ldots t_m^{\beta_m}$ in the pattern inventory polynomial $P_{X/G}(t_1, t_2, \ldots\ldots, t_m)$. Equating the two expressions for $\sum\omega(T)$ in equations 5.3 and 6.1,we obtain

$$(6.2) \quad P_{X/G}(t_1,t_2,\ldots\ldots,t_m) = \frac{1}{|G|}\sum_{g\in G}\prod_{i=1}^{n}(t_1^i + t_2^i + t_2^i \ldots\ldots + t_m^i)^{\lambda_i(g)}.$$

Since $P_{X/G}(t_1, t_2, \ldots\ldots, t_m)$ is a homogeneous polynomial of degree n in $t_1, t_2, \ldots\ldots, t_m$, the fact is also clear from the right-hand side in equation (6.2). Since η_g is a permutation of a set of n elements, the sum of the lengths of the cycles in the decomposition of η_g is equal to n, so $\sum i\lambda_i(g) = n$. Hence every term in the summation in equation (6.2) is of degree n. Further, $P_{X/G}(t_1, t_2, \ldots\ldots, t_m)$ is symmetric in the indeterminates $t_1, t_2, \ldots\ldots, t_m$.

Cycle Index Polynomial

The result of Polya's theorem and the formula for the pattern inventory can be expressed in a more compact and elegant form by introducing the concept of the cycle index polynomial of a permutation group. Given a permutation σ of a set of n elements, the cycle index of σ is defined to be the n-tuple $(\lambda_1, \lambda_2, \lambda_3, \ldots\ldots, \lambda_n)$, where λ_i denotes the number of cycles of length i in the cycle decomposition of σ.

Definition 3.5.1

Given a permutation group G of degree n, the cycle index polynomial of G is defined to be the polynomial Z_G in n in determinates $r_1, r_2, r_3 \ldots\ldots, r_n$ given by

$$\mathbb{Z}_G(r_1, r_2, r_3\ldots\ldots, r_n) = \frac{1}{|G|}\sum_{g\in G}\prod_{i=1}^{n} r_i^{\lambda_i(g)}$$

where λ_i(g)(i = 1, 2, 3, ..., n), denotes the number of cycles of length i in the cycle decomposition of the permutation. The sum of the coeficients in the polynomial Z_G is 1. This is a consequence of the fact that the number of terms in the summation on the right-hand side is equal to $|G|$.

Theorem 3.5.3

Let G be a finite group acting on a finite set S having n elements, and let $Z_G(r_1, r_2, r_3, \ldots\ldots, r_n)$ be the cycle index polynomial of G acting on S. Let $|C| = m$ and $X = C_S$ be the set of all mappings from $S \rightarrow C$. Then

1. The number κ of orbits in X under G is given by

$$\kappa = \mathbb{Z}_G(m, m, m, \ldots\ldots, m).$$

2. The pattern inventory of orbits in X under G is given by

$$P_{X/G}(t_1, t_2, \ldots\ldots, t_m) = \mathbb{Z}_G(t_1 + t_2 + \ldots\ldots + t_m, t_i^2 + t_2^2 + \ldots\ldots + t_m^2, t_i^n + t_2^n + \ldots. + t_m^n).$$

Proof. Exercise

Problem 3.5.1

Find the cycle index polynomial of S_3.

Solution 3.5.1

Here, $S_3 = \{(1), (12), (13), (23), (123), (132)\}$. The identity permutation (1) = (1) (2) (3) has the cycle index (3, 0, 0). The three permutations (12), (13), and (23) all have the index (1, 1, 0). The remaining two permutations both have the index (0, 0, 1). Hence the cycle index polynomial of S_3 is

$$\mathbb{Z}_{S_3} = \frac{1}{6}(r_1^3 + 3r_1r_2 + 2r_3).$$

3.6 Generating Functions For Non-isomorphic Graphs

Intuitively, a graph consists of a set of points and a set of lines such that each line joins a pair of points. In this section we consider an application of Polya's theorem in graph theory. Informally, a graph consists of a set of vertices of which some pairs (possibly all or none) are joined by line segments or arcs. A formal definition is given below.

Definition 3.6.1

A graph G is an ordered pair $G = (V, E)$, where V is a finite non empty set and $E \subseteq V^2$, where for any set V, $V^2 = V \times V = \{(u, v) | u, v \in V, u 6 = v)$. The elements of V are called vertices of the graph G, and the elements of E are called its edges.

Two vertices a, b in a graph are said to be adjacent if the pair (a, b) is an edge. A graph G is commonly represented by a diagram in which the vertices are shown as points or small circles, and two vertices a and b are joined by a segment or an are if and only if (a, b) is an edge. The positions of the vertices in the diagram and the shapes of the arcs joining the vertices are immaterial. For example, the three diagrams below, though they look quite different from one another, represent the same graph $G = (V, E)$ with

$$V = \{a, b, c, d\}, \quad E = \{(a, b), (a, c), (a, d), (b, c), (c, d)\}$$

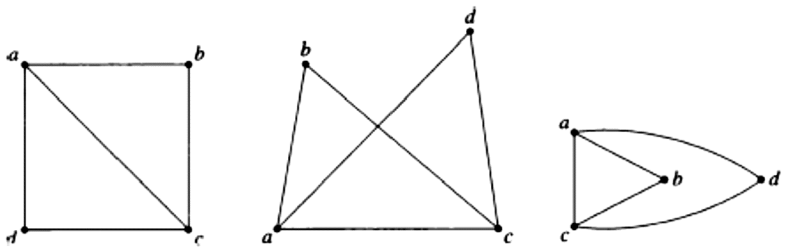


Figure 5 : Graph

A less trivial and more interesting example is provided by the two diagrams below, which represent the same graph, known as the Petersen graph.

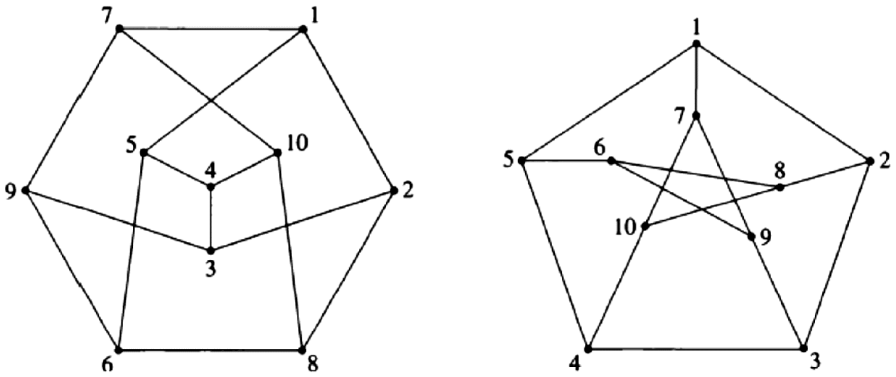


Figure 6 : Peterson Graph

Here, $|V^2| = \binom{n}{2} = \frac{n(n-1)}{2} \Rightarrow$ there are $2^{|V^2|}$ distinct subsets of V^2. Hence the number of distinct possible graphs on a given set V with n vertices is $\frac{n(n-1)}{2}$. But among these, there are several cases of graphs that are essentially a like and can be obtained from one another by permuting the vertices. Such graphs are said to be isomorphic. Now we are in a position to define:

Definition 3.6.2

Two graphs $G = (V, E)$ and $G' = (V', E')$ are said to be isomorphic if $\exists$ a bijective mapping $f : V \to V'$ such that $\forall\ (a, b) \in V^2$, $(a, b) \Leftrightarrow E\ (f(a), f(b)) \in E'$.

For example, the graphs represented by the two diagrams below are seen to be isomorphic on taking $V = \{1, 2, 3, 4\}$ and $V' = \{a, b, c, d\}$, the mapping $f : V \to V'$ given by $1 \to a$, $2 \to c$, $3 \to b$, $4 \to d$,

It follows from the definition that two graphs are isomorphic if and only if the irrespective diagrams can be obtained from each other by relabeling the vertices.

Let us now focus our attention to those graphs that are not isomorphic. Our problem is to find the maximum number of non isomorphic graphs with a given number n of vertices. For small values of n, one can find this number by drawing all possible unlabeled diagrams with n vertices. For example,we can easily see that there are just four different diagrams of a graph with three vertices, as shown below. Hence there are only four non isomorphic graphs with three vertices. But this method is not practical for large values of n.

As we shall see shortly, this problem is closely related to that of counting color patterns and so can be tackled by applying the theory that we developed in the foregoing sections. Let X be the set of all graphs on a given set V of

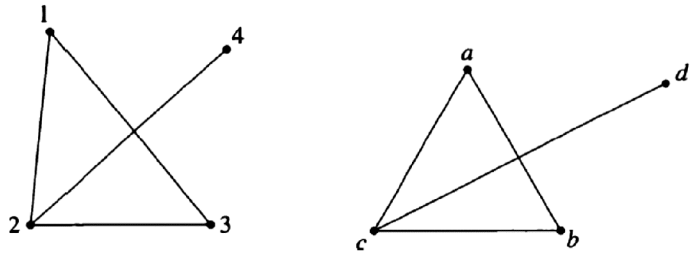


Figure 7 : Isomorphic Graph

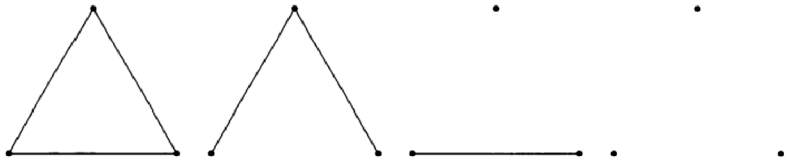


Figure 8 : NonIsomorphic Graph

n vertices, and let $K = (V, V^2)$ be the graph in which every pair $(a, b) \in V^2$ is an edge. Given a graph $G = (V, E)$, let us color the edges in the diagram of the graph K with two colors, say black and white, as follows : If $(a, b) \in E$, we color the edge (a, b) in K with black, otherwise with white. Thus each graph G on the vertex set V determines a coloring of the edges of K with two colors. It is clear that this gives a one-to-one correspondence between the set X of graphs on V and the set of all 2-colorings of the edges in the diagram of K. In other words, we can identify X with the set of all two-color assignments of the set V^2. Without loss of generality, we write $V = \{1, ..., n\}$ and let S_n be the group of all permutations of V. The natural action of S_n on V induces an action on V^2 by the rule $\sigma(a, b) = (\sigma(a), \sigma(b)) \ \forall \sigma \in S_n$ and $(a, b) \in V^2$. Two graphs $G = (V, E)$ and $G' = (V, E')$ are isomorphic if and only if

the corresponding 2-colorings of V^2 are in the same orbit under the action of the group S_n. So the non-isomorphic graphs on the vertex set V correspond to the orbits in X under S_n. It follows that the number of non-isomorphic graphs on V is equal to the number κ of orbits in X under S_n. By Theorem 3.4.1, we have

$$\kappa = \frac{1}{n!}\sum_{\sigma \in S_n} 2^b,$$

where λ_σ is the number of cycles (including cycles of length 1) in the cycle decomposition of the permutation of V^2 induced by σ.

Theorem 3.6.1

The generating function $f_n(x)$ for the non-isomorphic graphs on n vertices is given by

$$f_n(x) = \mathbb{Z}(1+x, 1+x^2, \ldots\ldots, 1+x^N),$$

where, $N = |V^2|$ and $Z(r_1, r_2, \ldots\ldots, r_N)$, is the cycle index polynomial of the group S_n acting on the set V^2, where $V = \{1, 2, 3, \ldots\ldots, n\}$.

Proof. By virtue of Polya's theorem, we can find the number $g(n, m)$ of non-isomorphic graphs on V having m edges, $m = 0, 1, \ldots\ldots, N$, where $N = |V^2|$. Let $Z(r^1, r^2, \ldots\ldots, r^N)$ be the cycle index polynomial of the group S_n acting on the set V^2. Then the pattern inventory of the orbits in X under S_n is the polynomial

$$P(t_1, t_2) = Z(t_1 + t_2, t^2_1 + t^2_2, \ldots\ldots, t^N_1 + t^N_2).$$

The coefficient of the monomial $t^m t^{N-m}$ in this polynomial gives the number $g(n, m)$ of non-isomorphic graphs having m edges. Since $P(t_1, t_2)$ is symmetric in t_1, t_2, it follows that $g(n, m) = g(n, N - m)$. Putting $t_1 = 1$, $t_2 = x$, we obtain the function

$$f_n(x) = P(1, x) = Z(1 + x, 1 + x^2, \ldots\ldots, 1 + x^N).$$

The coeficient of x^m in the polynomial $f_n(x)$ gives the number $g(n, m)$ of non-isomorphic graphs on n vertices having m edges, so

$$f_n(x) = \sum_{m=0}^{N} g(n,m)x^N.$$

The polynomial $f_n(x))$ is called the generating function for the non-isomorphic graphs on n vertices. Since $g(n, m) = g(n, N - m)$, it follows that $f_n(x)$ is a reciprocal polynomial.

Remark 3.6.1

To compute the cycle index polynomial $Z(r_1, r_2, \ldots\ldots, r_N)$ of S_n acting on V^2 where $V = \{1, 2, 3, \ldots\ldots, n\}$ it is not necessary to consider every in S_n. Suppose σ,

σ have the same cycle structure. Then they are conjugate elements in the group S_n. Therefore the permutations induced by them on the set V^2 are also conjugate and have the same cycle structure.

Problem 3.6.1

Find the generating function f_4 for the non-isomorphic graphs on four vertices.

Solution 3.6.1

Here V = {1, 2, 3, 4} and therefore V^2 = {(1, 2), (l, 3), (l, 4), (2, 3), (2, 4), (3, 4)} For the sake of conve- nience, let us write the pair *a, b as ab*. Then we can write V^2 as V^2 = {12, 13, 14, 23, 24, 34} The following table shows a typical permuation $\sigma \in S_4$ for each possible cycle structure, the number #(σ) of permutations that have the same cycle structure as σ, the permutation σ induced by σ on V^2, the number of cycles in the decomposition of σ, and the monomial contributed by σ to the cycle index poynomial.

σ	$\#(\sigma)$	$\bar{\sigma}$	$\lambda(\bar{\sigma})$	$\prod y_i^{\lambda_i(\bar{\sigma})}$
$e=(1)$	1	$e=(12)$	6	y_1^6
(1 2)	6	(13 23)(14 24)	4	$y_1^2y_2^2$
(1 2)(3 4)	3	(13 24)(14 23)	4	$y_1^2y_2^2$
(1 2 3)	8	(12 23 13)(14 24 34)	2	y_3^2
(1 2 3 4)	6	(12 23 34 14)(13 24)	2	y_2y_4

Figure 9: Calculation

Here the cycle index polynomial of S_4 acting on the set V^2 is given by

$$\mathbb{Z}(y_1, y_2, y_3, y_4, y_5, y_6) = \frac{1}{24}(y_1^6 + 9y_1^2y_2^2) + 8y_3^2 + y_2y_4.$$

So,

$$f_4(x) = \frac{1}{24}\{(1+x)^6 + (1+x)^2(1+x^2)^2 + 8(1+x^3)^2 + 6(1+x^2)(1+x^4)\}.$$

Simplifying, we obtain

$$f_4(x) = 1 + x + 2x^2 + 3x^3 + 2x^4 + x^5 + x^6$$

The total number of non-isomorphic graphs is $f_4(1)$ = 11. Or equivalently, by using Theorem 3.4.1, we can obtain κ = 11. Here are the diagrams of the 11 non-isomorphic graphs on four vertices.

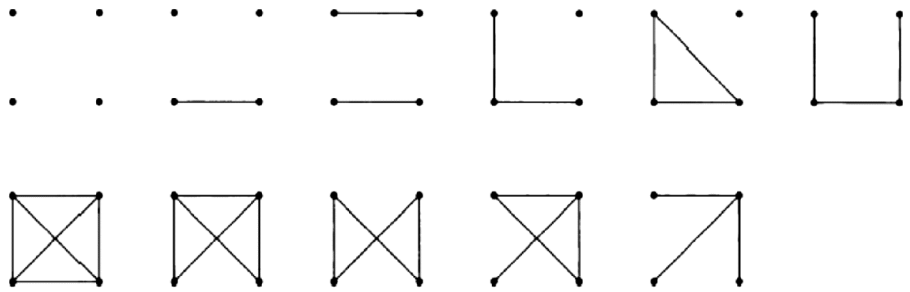


Figure 10 : Nonisomorphic

Remark 3.6.2

In the above problem, the cycle index polynomial $Z(r_1, r_2, ..., r_6)$ of S_4 acting on V^2 where $V = \{1, 2, 3, 4\}$ is replaced by $Z(y_1, y_2, ..., y_6)$.

3.7 Summary

The unit highlights on permutation groups, groups of symmetry and action of a group on a set. The unit also introduces the concept of coloring and color patterns. The learner can understand the use of Polya's theorem to solve problems related to counting the patterns. The unit also includes the concepts of pattern inventory, generating functions for non-isomorphic graphs which will motivate the learners to increase their knowledge in their future courses.

3.8 Exercises

1. Product of two cycles may not be a cycle. Justify.
2. The cycles (2435) and (168) are disjoint cycles where as the cycles (4532) and (138) are not disjoint.
3. Any non-identity permutation $\alpha \in S_n$ is either an even permutation or an odd permutation but never both.
4. Let (G, o) be any group and take $A = G$. Define * by $g * a = g \text{ o } a \text{ o } g^{-1}$, $g \in G$, $a \in A$. Prove that is a group action and is called action by conjugation.
5. Suppose each vertex of a regular hexagon is colored by one of m given colors. Find the number of distinct patterns among all colorings.
6. Suppose each vertex of a regular hexagon is colored by one of m given colors. Find the number of distinct patterns among all colorings.
7. Find the generating function for the non-isomorphic graphs on five vertices and draw their diagrams.

3.9 References and Further Reading

[1] Nagpaul, S.R., Jain, S.K., Topics in Applied Abstract Algebra, The Brooks/ Cole Series in Advanced Mathematics.

[2] Sen, M.K., Ghosh, S., Mukhopadhyay, P., Topics in Abstract Algebra, Universities Press, Mathematics.

[3] Khanna, V.K., Bhambri, S.K., A course in Abstract Algebra, 5th Edition, Vikas Publishing House Pvt. Ltd.

Unit 4 ❑ Application of Linear Transformations

Structure

4.0 Objectives

4.1 Introduction

4.2 Fibonacci Numbers

4.3 Incidence Models

4.4 Differential equations.

4.5 Least squares methods

4.5.1 Approximate solutions of system of linear equations

4.5.2 Approximate inverse of an m × n matrix

4.5.3 Solvinga matrix equation using its normal equation

4.5.4 Finding functions that approximate data.

4.6 Linear Algorithms

4.6.1 LDU Factorization

4.6.2 The Row Reduction Algorithm and its Inverse

4.6.3 Back and Forward Substitution : Solving $Ax = y$

4.6.4 Approximate Inverse and Projection Algorithms

4.7 Summary

4.8 Exercise

4.9 References

4.0 Objectives

The main objective of the present unit is to study the various aspects on Applications of Linear Transformations viz Fibonacci numbers, incidence models, Least squares methods and Linear algorithms.

4.1 Introduction

The main goal of the unit is to help students master the basic concepts and skills they will use later in their careers. The topics here follow the recommendations of the linear algebra curriculum study group, which are based on a careful investigation of the real needs of the students and a consensus among professionals in many

disciplines that use linear algebra. Hopefully, this course will be one of the most useful and interesting mathematics classes taken by undergraduates.

4.1 Fibonacci Numbers

The ancient Greeks attributed a mystical and a esthetical signicance to what is called a Golden Section. The Golden section is the division of a line segment in to two parts such that the smaller on a to the larger one b is $b : (a + b)$ *i.e.*

$$\frac{a}{b}+\frac{b}{a+b}.$$

Hence $a^2 + ab - b^2 = 0$. In particular, if $b = 1$, then $a^2 + a - 1 = 0$. Thus, $a = \frac{-1\pm\sqrt{5}}{2}$ The particular value $a = \frac{\sqrt{5}-1}{2}$ is said to be the golden mean.

We introduce the sequence of numbers as the Fibonacci numbers, name being derived from the Italian Mathematician Leonard o Fibanacci, who lived in Pisa. This problem gives rise to the sequence of numbers *viz* $a_0 = 0$, $a_1 = 1$, $a_2 = 1$, $a_3 = 2$, $a_4 = 3$, ..., ... where the $a_{n+1} = a_n + a_{n-1}$, $n \geq 1$. This sequence is known as the Fibonacci sequence and its terms are said to be the Fibonacci Numbers. Now the question arises : Is it possible to find a simple formulae for finding an as a function of n ? The answer is in armative sense as demonstrated below. The approach we take is by means of 2×2 matrices.

We write down a sequence of vectors built up from the Fibonacci numbers as follows :

$$v_0=\begin{pmatrix}a_0\\a_1\end{pmatrix} \qquad v_1=\begin{pmatrix}a_1\\a_2\end{pmatrix} \qquad v_2=\begin{pmatrix}a_2\\a_3\end{pmatrix} \qquad v_3=\begin{pmatrix}a_3\\a_4\end{pmatrix}$$

.....

$$v_n=\begin{pmatrix}a_n\\a_{n+1}\end{pmatrix}, \forall n \geq 1.$$

Now for all $n \geq 1$,

$$a_n = 0a_{n-1} + a_n$$

$$a_{n+1} = a_{n-1} + a_n.$$

This can be written in matrix notation as

$$\begin{pmatrix}a_n\\a_{n+1}\end{pmatrix}=\begin{pmatrix}0&1\\1&1\end{pmatrix}\begin{pmatrix}a_{n-1}\\a_n\end{pmatrix}$$

i.e. $v_n = Av_{n-1}, \forall n \geq 1$ *where* $A = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$

Thus A carries each term of the sequence into its success or in the sequence. Going back to the beginning, we have

$v_1 = Av_0$

$v_2 = Av_1 = A^2v_0$

$v_3 = Av_2 = A^3v_0$,

...

$v_{n+1} = Av_n = A^{n+1}v_0$.

If we know a_0, a_1, then using the relation $A^nv_0 = v_n$ we can find the formula for a_n, which is

$$a_n = \text{the nearest integer to } \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^n.$$

[for details refer to [4]].

There are several natural variants of the Fibonacci numbers that one could introduce. To begin with, we need not begin the proceedings with $a_0 = 0$, $a_1 = 1$, $a_2 = 1$, ..., ..., we could start with

$$a_0 = a,\ a_1 = b,\ a_2 = a + b,\ \ldots\ldots,\ \ldots\ldots,\ a_{n+1} = a_n + a_{n-1}.$$

The same matrix $\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$, the same characteristic roots, and the same characteristic vectors w_1, w_2 (say) arise. The only change we have to make is to express $\begin{pmatrix} a \\ b \end{pmatrix}$ as combination of w_1, w_2 . The rest of the argument follows as above.

A second variant might be :

$a_0 = 0$, $a_1 = 1$, $a_2 = c$, ..., ..., $a_{n+1} = ca_n + da_{n-1}$,

where c and d are fixed integers. The change from the argument above would be that we use the matrix $B = \begin{pmatrix} 0 & 1 \\ d & c \end{pmatrix}$; it is to be noted that

$$B\begin{pmatrix} a_{n-1} \\ a_n \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ d & c \end{pmatrix}\begin{pmatrix} a_{n-1} \\ a_n \end{pmatrix} = \begin{pmatrix} a_n \\ da_{n-1} + ca_n \end{pmatrix} = \begin{pmatrix} a_{n-1} \\ a_n \end{pmatrix}$$

If the characteristic roots of B are distinct, to find the formulae for a_n we must

find the characteristic roots and associated vectors of B, express the first vector $\begin{pmatrix}0\\1\end{pmatrix}$ as a combination of these characteristic vectors, and proceed as before.

Example 4.2.1. Suppose that

$$a_0 = 0,\ a_1 = 1,\ a_2 = 1,\ a_3 = a_2 + 2a_1 = 3,\ \ldots\ldots,\ \ldots\ldots$$
$$a_n = a_{n-1} + 2a_{n-2},\ n \geq 2.$$

Thus, $B = \begin{pmatrix}0 & 1\\2 & 1\end{pmatrix}$. The characteristic roots(eigen values) of B are 2, –1.

If $w_1 = \begin{pmatrix}1\\2\end{pmatrix}, w_2 = \begin{pmatrix}-1\\1\end{pmatrix}, then\ Bw_1 = 2w_1, Bw_2 = -w_2$. Also, $\begin{pmatrix}0\\1\end{pmatrix} = \frac{1}{3}w_1 + \frac{1}{3}w_2$, hence

$$v_n = B^n v_0 = \frac{1}{3}B^n w_1 + \frac{1}{3}B^n w_2 = \frac{2^n}{3}w_1 + \frac{(-1)^n}{3}w_2 = \begin{pmatrix}\frac{2^n}{3} + \frac{(-1)^{n+1}}{3}\\ \frac{2^{n+1}}{3} + \frac{(-1)^{n+2}}{3}\end{pmatrix}.$$

Thus $a_n = \dfrac{2^n + (-1)^{n+1}}{3}, n \geq 0$

4.3 Incidence Models

Models used to determine a local prices of goods transported among various cities, and certain other models used to determine electrical potentials at nodes of a network of electrical currents and to determine displacements at nodes of a mechanical structures under stress, all have one things in common. When these models are stripped of the trappings that go with the particular model, which is left is an incidence diagram as shown below :

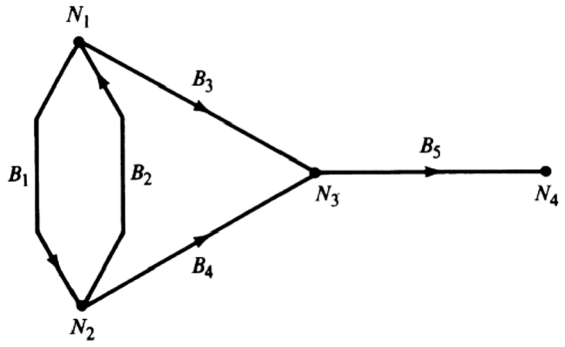


Figure 1: Incidence Diagram

What exactly is an incidence diagram? It consists of certain number of m nodes and a certain number n of oriented branches, each of which begins and ends in different nodes, such that every node is the beginning or ending of same branch. Now the question is how incidence diagram be represented as mathematical model ? Considering to each incidence diagram is an incidence matrix such that each column has one entry whose value is 1, one entry whose value is -1, and 0 for all other entries. To form this matrix, let there be one row for each node N_r and one column for each branch B_s. If node N_r is the beginning of branch B_s, let the (r, s)th entry of the matrix be 1. Otherwise let the (r, s) th entry be 0, indicating that N_r is neither the beginning nor the ending of the branch B_s. The matrix of the foregoing diagram is

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

Conversely, given an $m \times n$ incidence matrix T, $\exists$ a corresponding incidence diagram with m nodes $N_1, N_2, \ldots\ldots, N_m$ corresponding to the rows of T and n oriented branches $B_1, B_2, \ldots\ldots, B_n$ corresponding to the columns of T. If rth row has -1 in the column s, then the node N_r, is the beginning of the branch B_s. On the other hand if the rth row has 1 in the column s, then the node N_r, is the end of the branch B_s. For instance, the matrix T given by

$$T = \begin{pmatrix} 1 & 1 & -1 & 0 & 0 \\ 1 & -1 & 0 & -1 & 0 \\ 0 & 0 & 1 & 1 & -1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

is an incidence matrix whose incidence diagram is illustrated here.

Example

In a transportation model, the nodes N_r of our incidence diagram represent cities and the branches B_s transportation routes between the cities. Suppose, bisleri is produced and consumed in the cities N_r and transported along the routes B_s such that :

1. For each city N_r, if we add the rates at which bisleri is transported along the routes B_s heading to the city N_r and then subtract therates F_s for routes B_s heading out of city N_r, we get the difference G_r between the rate of production and the rate of consumption in city N_r. So for city N_2 since route B_1 heads in whereas routes B_2 and B_4 heads out, $G_2 = F_1 - F_2 - F_4$.
2. If we denote P_r to be the price of bisleri (per bottle) in city N_r and E_s the price at the end of the route B_s minus the price at the beginning of the route B_s, $\exists$ a positive constant R_s such that $R_s F_s = E_s$. The constant R_s reflects the distance

and other resistance to flow along the route B_s that will increase the price difference E_s. For $s = 3$, the price at the beginning of B_3 is P_1, such that $R_3F_3 = E_3$, where $E_3 = P_3 - P_1$.

4.4 Differential Equations

Finding the set of all solutions to the homogeneous differential equation *viz*

$$\frac{d^n v}{dt^n} + a_{n-1}\frac{d^{n-1}v}{dt^{n-1}} + \ldots\ldots\ldots + a_0 v = 0.$$

is really the same as finding the nullspace of a linear transformation. The reason is that the mapping T that maps a function v of a real variable t *i.e.*, $v(t)$ to its derivative Tv = is linear on a vector space F_n of n times differentiable complex valued functions v over t. Here the derivative of the function

$$v(t) = a(t) + b(t)i$$

is defined as, $\frac{dv}{dt} = \frac{da}{dt} + \frac{db}{dt}i.$

Let $f(x) = x^n + a_{n-1}\, x^{n-1} + \ldots\ldots\ldots + a_0$; $f(T)$ being linear on F_n we can express the set of solutions to

$$\frac{d^n v}{dt^n} + b_{n-1}\frac{d^{n-1}v}{dt^{n-1}} + \ldots\ldots\ldots + b_0 v = 0.$$

as the null-space

$$W = \{v \in F_n \mid f(T)v = 0\}$$

To determine the null space W of $f(T)$, since $f(T)v = 0 \Rightarrow Tf(T)v = f(T)T(v) = 0$, therefore $T(W) \subset$ W. Let $v \in W$. Let V be the subspace of W spanned over C by the functions v, Tv, ……., $T^{n-1}v$. Since, $f(T)v = 0$, $T^n v$ is a linear combinations of $\{v, Tv, \ldots\ldots, T^{n-1}v\}$, so $f(T) \subset V$. It follows that T maps v, Tv, ……, $T^{n-1}v$ into V, *i.e.* $TV \subset$ V.

Theorem 4.4.1. Let T be a linear transformations of a finite dimensional vector space V over a field F of complex numbers and let $f(T)V = 0$, where $f(x) = (x - a)^{m1} + (x - a_2)^{m2} + \ldots\ldots + (x - a_k)^{mk}$. Then V is the direct sum

$$V = V_{a_1}(T) \oplus V_{a_2}(T) \oplus \ldots\ldots V_{a_t}(T),$$

where $V_{ar}(T)$ are the generalized characteristic subspaces

$$V_{a_r}(T) = \{v \in V \mid (T - a_r I)^e v = 0, \textit{ for some } e\}.$$

Proof. We assume that $V \neq \phi$, take b to be the characteristic root of T, and $(x - b) \mid f(x)$. Taking $S = T - bI$, we have $V = X \oplus Y$, where X is the generalised null subspace

$$X = \{v \in 2\ V | S^e v = 0, \text{ for some } e\}$$

of S and $Y = \bigcap_{e=1}^{\infty} S^e v$. To prove this we use the equations

$$X = \{v \in V \mid S^e v = 0\}, \quad Y = S^e v$$

for dim (V). Since, $S^e X = 0$, $S^e Y = S^{2e} V = S^e V = Y$ for such e, the mapping defined by S^e on X is 0 and the mapping defined by S^e on Y is 1 – 1 and onto. So, $X \cap Y = \{0\}$. Now let e = dim V and take $v \in V$. Choose $w \in Y$ such that $S^e v = S^e w$. Then $v = (v - w) + w$. Since, $S^e(v - w) = S^e v - S^e w = 0$, we get that $v \in X + Y$. Since, we know that $X \cap Y = \{0\}$, we conclude that $V = X \oplus Y$. The subspace X of V is non zero since b is a characteristic root of T. If $X = V$, then $(x - b) \mid f(x) \Rightarrow V = V_b(T)$ and we are done. Otherwise, both X and Y are nonzero subspaces of lower dimension than V. Since T maps the subspaces X and Y into themselves and $f(T)X = 0$, $f(T)Y = 0$, by induction, we get the desired decomposition

$$Y = Y_{a_1}(T) \oplus Y_{a_2}(T) \ldots\ldots\ldots \oplus Y_{a_k}(T).$$

Moreover, we also have $X = V_b(T)$, where $(x - b) \mid f(x)$ and $Y_b(T) = \{0\}$, as $S = T - I$ is 1 – 1 on Y. Renumbering the a_r so that $b = a_1$, we get

$$V = X \oplus Y = V_{a_2}(T) \oplus V_{a_2}(T) \oplus \ldots\ldots\ldots \oplus V_{a_k}(T).$$

Theorem 4.4.2. The vector space W of n-times differentiable complex valued solutions v to

$$\frac{d^n v}{dt^n} + b_{n-1}\frac{d^{n-1} v}{dt^{n-1}} + \ldots\ldots\ldots\ldots + b_0 v = 0.$$

as the null-space

$$W = \{v \text{ e } F_n \mid f(T)v = 0\}$$

is n-dimensional over C. Prove that the basis for W is

$$\{t^{n_r} e^{a_r t} \mid 1 \leq r \leq k, 0 \leq n_r \leq m_r - 1\}$$

where

$x^n + b_{n-1}x^{n-1} + \ldots\ldots + b_a$, factors as $(x - a_1)^{m_1}(x - a_2)^{m_2} \ldots\ldots\ldots (x - a_k)^{m_k}$.

Proof. Writing v as $v = v_1 + \ldots\ldots\ldots + v_k$ with $v_r \in V_{ar}(T), \forall r$, we have

$$(T - a_r I)^{mr} v_r = 0.$$

For some fixed r, let us denote $a_r = a$, $v_r = w$ and $m_r = m$. Since, $T(e^{at}u) = ae^{at}u + e^{at}Tu$ for any differentiable function u, we have

$$(T - aI)(e^{at}u) = e^{at}Tu.$$

Applying $(T - aI)$ like this for $(m - 1)$ more times,we also have $(T - a, I)^m (e^{at}u) = e^{at}T^m u$.

Replacing u by $e^{-at}w$, we have

$$(T - aI)^m (e^{at}e^{-at}w) = e^{at}T^m (e^{-at}w), \text{ i.e.}$$

$$(T - aI)^m w = e^{at}T^m (e^{-at}w).$$

So, $(T - aI)^m w$ is equivalent to $T^m(e^{-at}w) = 0$ which in turn is equivalent to the condition that $e^{-at}w$ is a polynomial $p(t)$ of degree less than m. But then $w = pe^{at}$ holds implies w is a linear combination of e^{at}, te^{at}, ..., $t^{m-1}e^{at}$.

Conversely, the functions e^{at}, te^{at}, ..., t^{m-1} eat are solutions to the differential equations $(T - aI)^m = 0$. Since, $(x - a)^m = (x - a_r)^m r$ is a factor of $f(x)$, they are also the solutions to the differential equations $f(T)v = 0$.

Since, each solution $v \in W$ is a sum of functions v_r, each of which is a linear combination of

$$\{e^{a_r t}, te^{a_r t}, ..., t^{m_r - 1}e^{a_r t}\},$$

we can say that the set

$$\{t^{n_r}e^{a_r}t \mid 1 \le r \le k, 0 \le n_r \le m_r - 1\}$$

spans W. So, W is finite dimensional vector space. As, T a linear transformation of W, we have $f(T)W = 0$ and

$$W = W_{a_1}(T) \oplus W_{a_2}(T) \oplus ..... \oplus W_{a_k}(T),$$

where are generalized characteristic subspaces introduced in Theorem 3.4.1 Since, the functions $e^{ar}t$, $te^{ar}t$, ..., $t^{mr}-1$ $e^{ar}t$ are linearly independent elements of $W_{a_r}(T)$, $\forall_r$, the set

$$\{t^{n_r}e^{a_r}t \mid 1 \le r \le k, 0 \le n_r \le m_r - 1\}$$

is linearly independent.

Theorem 4.4.3. For any $T \in M^n(C)$ (collection of all $n \times n$ order matrices with complex entries), $x_0 \in C^n$, and $t_0 \in \mathbb{R}$, $x(t) = e^{(t-t_a)T}x_0$ is a unique solution of matrix differential equation $x'(t) = Tx(t)$ such that $x(t_0) = x_0$.

Proof. Taking $f(t) = e^{tT}x_0$, we differentiate the series

$$f(t) = x_0 + \frac{tT}{1!}x_0 + \frac{t^2T^2}{2!}x_0 + \ldots\ldots\ldots + \frac{t^kT^k}{k!}x_0 + \ldots\ldots\ldots$$

term by term, getting its derivative

$$f'(t) = 0 + tx_0 + \frac{tT^2}{1!}x_0 + \ldots\ldots\ldots + \frac{t^{k-1}T^k}{(k-1)!}x_0 + \ldots\ldots\ldots$$

$$= T\left(tx_0 + \frac{tT^2}{1!}x_0 + \ldots\ldots\ldots + \frac{t^{k-1}T^k}{(k-1)!}x_0 + \ldots\ldots\ldots\right) = Te^{tT}x_0.$$

Since, $e^{tT}x_0$ has derivative $e^{tT}x_0$ and $e^{(t-t_0)T}x_0 = e^{tT}e^{-t_0T}x_0$, the derivative of the function $x(t) = e^{(t-t_0)T}x_0$ is $Tx(t)$(Verify). Since, $x(t_0) = x_0$, this proves the existence of the solution. it remains to prove that the solution is unique.

Taking $u(t)$ denote any other solution to the equation $x'(t) = Tx(t)$. Setting, $v = u - e^{(t-t_0)T}x_0$, we have

$$v(t_0) = 0. \text{ (verify)}$$

To prove $u = e^{(t-t_0)T}x_0$, its suffices to show that $v = 0$, using $v(t_0) = 0$. To the contrary, let v be non-zero.Taking $d > 0$ such that T^d is linearly independent on $T^0, T^1, \ldots\ldots, T^{d-1}$, the vector function T^dv is a linear combination of $v, Tv, \ldots\ldots, T^{d-1}v$. It follows that the linear span V of $\{v, Tv, \ldots\ldots, T^{d-1}v\}$ over is mapped to itself by T. Let w be the characteristic vector of T in V such that $Tw = aw$, for some scalar a. Since, $w \in V$, $v' = Tv$ and $v(t_0) = 0$, w satisfies the conditions $w' = Tw$ and $w(t_0) = 0$. So, $w' = aw$. Since the solutions to the dieffrential equations

$$w'_r(t) = awr\ (t)$$

are of the form $w_r(t) = w_re^{a(t-t0)}$ for some constants r_r, w is of the form

$$\begin{pmatrix} w_1e^{a(t-t_0)} \\ w_2e^{a(t-t_0)} \\ \ldots\ldots \\ w_ne^{a(t-t_0)} \end{pmatrix}.$$

Since, $w(t_0) = 0 = \begin{pmatrix} w_1 \\ w_2 \\ .... \\ w_n \end{pmatrix}$, it follows $w = 0$, which contradicts the fact of being w as the characteristic vector and therefore, non-zero. So, our hypothesis that $v \neq 0$ leads to contradiction. Hence the proof.

Theorem 4.4.4. For any t_0 and v_0, v_1, ..., v_{n-1}, the differential equation

$$\frac{d^n v}{dt^n} + b_{n-1}\frac{d^{n-1}v}{dt^{n-1}} + ... + b_0 v = 0,$$

has a unique solution v such that $v(r)(t_0) = v_r$ for $0 < r < n-1$, namely, $v = u_1$,

where $u = e^{(t-t_0)T}\begin{pmatrix} v_0 \\ v_1 \\ .... \\ v_n \end{pmatrix}$

Proof. Consider the system of equations

$$\begin{aligned} u_1' &= u_2 \\ u_2' &= u_3 \\ ... &= ... \\ u_{n-1}' &= u_n \\ u_n' &= -b_0 u_1 - .... - b_{n-1}u_n \end{aligned}$$

of n dierential equations in n unknowns represented by the matrix equation $u' = Tu$, where T is

$$\begin{pmatrix} 0 & & -b_0 \\ 1 & & .... \\ 0 & & .... \\ .... & \ddots & .... \\ 0 & ....0 & 1-b_{n-1} \end{pmatrix}$$

the companion matrix to the poly nomial $b_0 + ... + b_{n-1}x^{n-1} + x^n$ with the same coefficients as the differential equations. The condition $u' = Tu \Rightarrow v = u_1$ satisfies the conditions

$$u_1 = v^{(0)},\ u_2 = v^1,\ ....,\ u_n = v^{n-1}.$$

But then,

$$v(n) = u_n' = -b_0 u_1 - ..... -b_{n-1}u_n,\ i.e.$$

$$v(n) = -b_0 v^{(0)} - \ldots\ldots - b_{n-1} v^{(n-1)}.$$

so by virtue of Theorem 4.4.3, gives the desired result.

4.5 Least Squares Methods

4.5.1 Approximate solutions of systems of linear equations

The system of m linear equations in n variables with matrix equation $Ax = y$ has a solution x if and only if y is in the column space of A. What shall we do with an equation $Ax = y$ such as

$$\begin{pmatrix} 6 & 3 \\ 2 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$$

which has no solution? We can always find an approximate solution x by replacing y by the vector y' in the column space of A nearest to y and solving $Ax = y'$ instead. In case of the equation

$$\begin{pmatrix} 6 & 3 \\ 2 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \end{pmatrix},$$

where A = $\begin{pmatrix} 6 & 3 \\ 2 & 1 \end{pmatrix}$ and $y = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$, the column space of A is the span R $\begin{pmatrix} 3 \\ 1 \end{pmatrix}$ of $\begin{pmatrix} 3 \\ 1 \end{pmatrix}$.

So, y' is the multiple $y' = t\begin{pmatrix} 3 \\ 1 \end{pmatrix}$ of $\begin{pmatrix} 3 \\ 1 \end{pmatrix}$ such that the length $\|y - y'\| = \left\|\begin{pmatrix} 2 \\ 2 \end{pmatrix} - t\begin{pmatrix} 3 \\ 1 \end{pmatrix}\right\|$ is as small as possible. we can represent y, y' and W pictorially as follows:

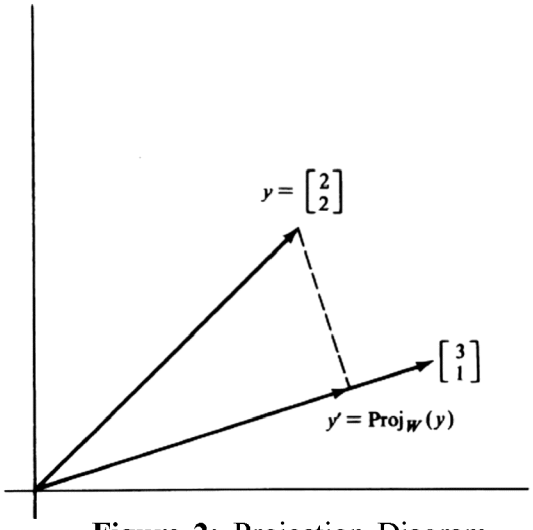


Figure 2: Projection Diagram

To get an explicit expression for the y' in the column space W of A nearest to y, we need to define :

Definition 4.5.1.1. Let W be a subspace of R^n, so that $R^n = W \oplus W^{\perp}$, and write $y \in R^n$ as $y_1 + y_2$, where $y_1 \in W$ and $y_2 \in W^{\perp}$. Then y_1 is called the projection of y on W and is denoted by $y_1 = \text{Proj}_W(y)$. For $W^{\perp}$ refer to [2].

For the equation $\begin{pmatrix} 6 & 3 \\ 2 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$,

we take W to be the span $\mathbb{R}\begin{pmatrix} 3 \\ 1 \end{pmatrix}$ of $\begin{pmatrix} 3 \\ 1 \end{pmatrix}$. Then $W^{\perp}$ is the span of $\begin{pmatrix} -1 \\ 3 \end{pmatrix}$. So we get that

$$\begin{pmatrix} 2 \\ 2 \end{pmatrix} = \frac{4}{5}\begin{pmatrix} 3 \\ 1 \end{pmatrix} + \frac{2}{5}\begin{pmatrix} -1 \\ 3 \end{pmatrix}.$$

the first term of which is $Proj_w \begin{pmatrix} 2 \\ 2 \end{pmatrix} = \frac{4}{5}\begin{pmatrix} 3 \\ 1 \end{pmatrix}$.

To minimise the length,

$$\|y - y'\| = \left\| \begin{pmatrix} 2 \\ 2 \end{pmatrix} - t\begin{pmatrix} 3 \\ 1 \end{pmatrix} \right\|$$

$$= \left\| \frac{4}{5}\begin{pmatrix} 3 \\ 1 \end{pmatrix} + \frac{2}{5}\begin{pmatrix} -1 \\ 3 \end{pmatrix} - t\begin{pmatrix} 3 \\ 1 \end{pmatrix} \right\|$$

$$= \left\| \left(\frac{4}{5} - t \right)\begin{pmatrix} 3 \\ 1 \end{pmatrix} + \frac{2}{5}\begin{pmatrix} -1 \\ 3 \end{pmatrix} \right\|.$$

Since, $\begin{pmatrix} 3 \\ 1 \end{pmatrix}$ and $\begin{pmatrix} -1 \\ 3 \end{pmatrix}$ are orthogonal we must take $t = \frac{4}{5}$ to eliminate the term involving $\begin{pmatrix} 3 \\ 1 \end{pmatrix}$.

So the element y' of W nearest to $\begin{pmatrix} 2 \\ 2 \end{pmatrix}$ is $y' = Proj_W \begin{pmatrix} 2 \\ 2 \end{pmatrix} = \frac{4}{5}\begin{pmatrix} 3 \\ 1 \end{pmatrix}$. From the theorem of Pythagoras, we have that if $y = y_1 + y_2$, where y_1 and y_2 are orthogonal, then $\|y\|^2 = \|y_1\|^2 + \|y_2\|^2$, since

$$\begin{aligned}\|y\|^2 &= < y_1 + y_2, y_1 + y_2 > \\ &= < y_1, y_1 > + < y_1, y_2 > + < y_2, y_2 > \\ &= < y_1, y_1 > + < y_2, y_2 > \\ &= \|y_1\|^2 + \|y_2\|^2.\end{aligned}$$

Using this we have,

Taking $y = y_1 + y_2$, with $y_1 \in W$ and $y_2 \in W^{\perp}$ such that $y_1 = Proj_w(y)$. Then for $w \in W$, the distance $\|y - w\|^2 = \|(y_1 - w) + y_2^2\|$. By Pythagoras theorem, we have

$$\|y_1 - w\|^2 + \|y_2\|^2.$$

which is minimal if and only if $w = y_1 = Proj_w(y)$. Thus we are in a position to state,

Theorem 4.5.1.1. Let W be a subspace of R^n. Then the element y' of W nearest to $y \in R^n$ is $y' = Proj_w(y)$.

Given a vector y and a subspace W, the method of going from y to the vector $y' = Proj_w(y)$ in W nearest to y is sometimes called the method of least squares since the sum of the squares $y - y'$ is there by minimized.

Next how to compute $Proj_w$ is evident from the following theorem.

Theorem 4.5.1.2. Let W be a subspace of R^n with orthogonal basis $w_1, w_2, \ldots\ldots, w_k$ and let $y \in R^n$. Then

$$Proj_w(y) = \frac{< y_1 w_1 >}{< w_1, w_2 >} w_1 + \ldots.. + \frac{< y, w_k >}{< w_k, w_k >} w_k.$$

Proof. Let $$y_1 = \frac{< y, w_1 >}{< w_1, w_2 >} w_1 + \ldots.. + \frac{< y, w_k >}{< w_k, w_k >} w_k.$$

Then we have,

$$< y - y_1, w_j > = < y, w_j > - \frac{< y, w_j >}{< w_j, w_j >} < w_j, w_j >, \text{ for } 1 \le j \le k.$$

So, $(y - y_1) \in 2\ W$ and $y_1 = Proj_w(y)$.

The geometrical statement of the above theorem is as follows : the projection of y on the span W of mutually orthogonal vectors $w_1, w_2, \ldots\ldots, w_k$ equals to the sum of the projection of y on the lines $Rw_1, \ldots\ldots, Rw_k$.

Problem 4.5.1.1. Suppose that we have a supply of 5000 units of A, 4000 units of B and 2000 units of C, materials used in manufacturing products are P and Q, if we ask:

if each units of P uses 2 units of A, 0 units of B and 0 units of C, and each units

of Q uses 3 units of A, 4 units of B and 1 unit of C, how many units p and q of P and Qn should we make if we want to use up the entire supply?

Solution 4.5.1.1. The system is represented by the equation

$$\begin{pmatrix} 2 & 3 \\ 0 & 4 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} p \\ q \end{pmatrix} = \begin{pmatrix} 5000 \\ 4000 \\ 2000 \end{pmatrix}$$

Since, the vector $\begin{pmatrix} 5000 \\ 4000 \\ 2000 \end{pmatrix}$ is not a linear combination of $p\begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix} + q\begin{pmatrix} 3 \\ 4 \\ 1 \end{pmatrix}$ of the columns of $\begin{pmatrix} 2 & 3 \\ 0 & 4 \\ 0 & 1 \end{pmatrix}$ there is no exact solution $\begin{pmatrix} p \\ q \end{pmatrix}$ So, we get an approximate solution $\begin{pmatrix} p \\ q \end{pmatrix}$ by finding those values of p and q for which the distance from $p\begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix} + q\begin{pmatrix} 3 \\ 4 \\ 1 \end{pmatrix}$ to $\begin{pmatrix} 5000 \\ 4000 \\ 2000 \end{pmatrix}$ is as small as possible. To fullfil the purpose we first find the vector in the space W of linear combinations $p\begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix} + q\begin{pmatrix} 3 \\ 4 \\ 1 \end{pmatrix}$ that is closest to $\begin{pmatrix} 5000 \\ 4000 \\ 2000 \end{pmatrix}$. This vector is the projection $Proj_w\begin{pmatrix} 5000 \\ 4000 \\ 2000 \end{pmatrix}$ of $\begin{pmatrix} 5000 \\ 4000 \\ 2000 \end{pmatrix}$ on W. Computing this by the formulae $Proj_w(v) = < v, w_1 > w_1 + <v, w_2> w_2$ of theorem 4.1.2, where w_1 and w_2 are the orthonormal basis $w_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$, $w_2 = \frac{1}{c}\begin{pmatrix} 0 \\ 4 \\ 1 \end{pmatrix}$, where $c = \sqrt{17}$ of W, we get,

$$\left\langle \begin{pmatrix} 5000 \\ 4000 \\ 2000 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \right\rangle \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + \left\langle \begin{pmatrix} 5000 \\ 4000 \\ 2000 \end{pmatrix}, \frac{1}{c}\begin{pmatrix} 0 \\ 4 \\ 1 \end{pmatrix} \right\rangle \frac{1}{c}\begin{pmatrix} 0 \\ 4 \\ 1 \end{pmatrix}, \quad i.e.$$

$$Proj_{\mathrm{w}} \begin{pmatrix} 5000 \\ 4000 \\ 2000 \end{pmatrix} = 5000\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + \left(\frac{1}{c}16{,}000\right) + \left(\frac{1}{c}2{,}000\right)$$

$$= 5000\begin{pmatrix}1\\0\\0\end{pmatrix} + \frac{18000}{17}\begin{pmatrix}0\\4\\1\end{pmatrix}. \quad (4.1)$$

To get p and q amounts to expressing equation (4:1) as a linear combination

$$\left(2500 - \left(\frac{3}{2}\right)\left(\frac{18000}{17}\right)\right)\begin{pmatrix}2\\0\\0\end{pmatrix} + \frac{18000}{17}\begin{pmatrix}3\\4\\1\end{pmatrix}$$

of $\begin{pmatrix}2\\0\\0\end{pmatrix}$ and $\begin{pmatrix}3\\4\\1\end{pmatrix}$. So we get

$$p = 2500 - \left(\frac{3}{2}\right)\left(\frac{18000}{17}\right) = 911.76$$

$$q = \frac{18000}{17} = 1058.82.$$

Thus our approximate solution is $\begin{pmatrix}p\\q\end{pmatrix} = \begin{pmatrix}911.76\\1058.82\end{pmatrix}$. Thus using 911.76 units of P and 1058.82 units of Q, the vector representing supplies used is

$$911.76\begin{pmatrix}2\\0\\0\end{pmatrix} + 1058.82\begin{pmatrix}3\\4\\1\end{pmatrix} = Proj_w\begin{pmatrix}5000\\4000\\2000\end{pmatrix}$$

$$= 5000\begin{pmatrix}1\\0\\0\end{pmatrix} + \frac{18000}{17}\begin{pmatrix}0\\4\\1\end{pmatrix} = \begin{pmatrix}5000\\4235.29\\1058.82\end{pmatrix}$$

So we exactly used 5000 units of A, 4235.29 units of B and 1058.82 units of C.

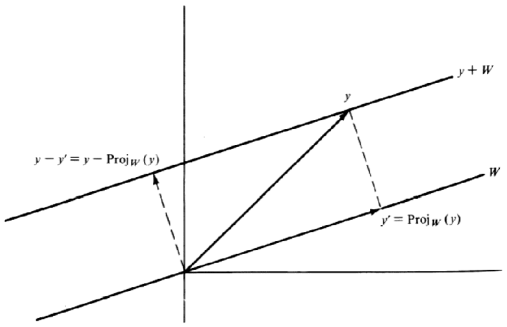
In the above example, we found an approximate solution in the following sense:

Definition 4.5.1.2. For any $m \times n$ matrix A with real entries, an approximate solution x to an equation $Ax = y$ is a solution to the equation $Ax = Proj_{A(Rn)}(y)$.

In the above *e.g.*, W is the column space $A(\mathrm{R}^2)$ of A and we found $Proj_{\mathrm{A(R2)}}\begin{pmatrix}5000\\4000\\2000\end{pmatrix}$ by using the formulae $Proj_{A(R2)}(v) = \langle v,w_1\rangle\, w_1 + \langle v,w_2\rangle\, w_2$ of Theorem 4.5.1.2, where w_1 and w_2 are the orthonormal basis of $A(\mathrm{R}^2)$. Now to find the shortest approximate solution x to $Ax = y$, we use the following lemma *viz.*

Lemma 4.5.1.1. Let W be a subspace of R^n and $y \in R^n$. Then the element of $y + W = \{y + w | w \in W\}$ of shortest length is $y - Proj_w(y)$.

Proof. The length $||y + w||$ of $y + w$ is the distance from y to $-w$. To minimize this distance from all $w \in W$, we take $-w = Proj_w(y)$, by Theorem 4.5.1.2. Hence the



proof.

Figure 3:

By virtue of above lemma, we get the shortest element of $v + N$ by replacing v by $v - Proj_N(v) = Proj_N(v)$, where N = Nullspace(A). Since, $N^{\perp}$ is the column space of transpose A^T of A, $Proj_{N\perp}(v)$ is just the projection $Proj_{AT(Rm)}(v)$ of v on the column space of A^T. So we can find the shortest approximate solution x to $Ax = y$ as follows:

1. Find one approximate solution v to the equation $Av = y$ by any method (*e.g.*, by the one given above)

2. Replace the approximate solution v by $x = Proj_{AT(Rm)}(v)$.

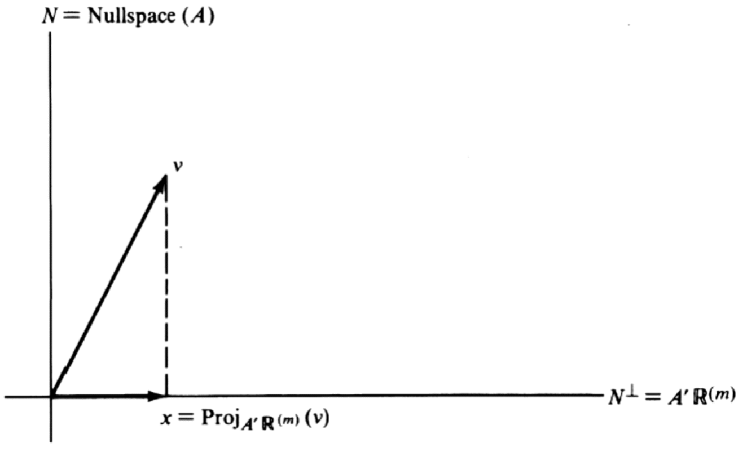


Figure 4 :

Problem 4.5.1.2. Find the best approximate solution to

$$\begin{pmatrix} 2 & 3 & 5 \\ 0 & 4 & 4 \\ 0 & 1 & 1 \end{pmatrix} x = \begin{pmatrix} 5000 \\ 4000 \\ 2000 \end{pmatrix}.$$

Solution 4.5.1.2. Since, the column space of $A = \begin{pmatrix} 2 & 3 & 5 \\ 0 & 4 & 4 \\ 0 & 1 & 1 \end{pmatrix}$ = column space of $W = \begin{pmatrix} 2 & 3 \\ 0 & 4 \\ 0 & 1 \end{pmatrix}$ refer to the Problem 4.5.1.1, the approximate solution $\begin{pmatrix} p \\ q \end{pmatrix} = \begin{pmatrix} 911.76 \\ 1058.82 \end{pmatrix}$ to the equation $\begin{pmatrix} 2 & 3 \\ 0 & 4 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} p \\ q \end{pmatrix} = \begin{pmatrix} 5000 \\ 4000 \\ 2000 \end{pmatrix}$, which satisfies the equation $\begin{pmatrix} 2 & 3 \\ 0 & 4 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} p \\ q \end{pmatrix} =$ $Proj_{A(\mathrm{R3})} \begin{pmatrix} 5000 \\ 4000 \\ 2000 \end{pmatrix}$, leads to the approximate solution

$v = \begin{pmatrix} p \\ q \\ 0 \end{pmatrix} = \begin{pmatrix} 911.76 \\ 1058.82 \\ 0 \end{pmatrix}$ to the equation $\begin{pmatrix} 2 & 3 \\ 0 & 4 \\ 0 & 1 \end{pmatrix} v = \begin{pmatrix} 5000 \\ 4000 \\ 2000 \end{pmatrix}$

as it satisfies the equation $\begin{pmatrix} 2 & 3 \\ 0 & 4 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} p \\ q \end{pmatrix} = \mathrm{Pr}oj_{A(\mathbb{R}^3)} \begin{pmatrix} 5000 \\ 4000 \\ 2000 \end{pmatrix}$, leads to the approximate solution.

$$v = \begin{pmatrix} p \\ q \\ 0 \end{pmatrix} = \begin{pmatrix} 911.75 \\ 1058.82 \\ 0 \end{pmatrix} \text{ to the equation } \begin{pmatrix} 2 & 3 \\ 0 & 4 \\ 0 & 1 \end{pmatrix} v = \begin{pmatrix} 5000 \\ 4000 \\ 2000 \end{pmatrix}$$

as it satisfies the equation $\begin{pmatrix} 2 & 3 & 5 \\ 0 & 4 & 4 \\ 0 & 1 & 1 \end{pmatrix} x = Proj_{A(\mathbb{R}^3)} \begin{pmatrix} 5000 \\ 4000 \\ 3000 \end{pmatrix}$. So, to get the best approximate solution, v is replaced by $v = Proj_{\mathrm{N}\perp}(v)$,

where $v = \begin{pmatrix} 911.76 \\ 1058.82 \\ 0 \end{pmatrix}, N = \text{Nullspace} \begin{pmatrix} 2 & 3 & 5 \\ 0 & 4 & 4 \\ 0 & 1 & 1 \end{pmatrix}$

Since, $N = \left\{ rw \mid w = \begin{pmatrix} -1 \\ -1 \\ -1 \end{pmatrix}, r \text{ being any scaler} \right\}$ (verify), we replace v by

$$v = Proj_N(v) = v - \frac{< v, w >}{< w, w >} w$$

$$= \begin{pmatrix} 911.76 \\ 1058.82 \\ 0 \end{pmatrix} - \frac{-911.76 - 1058.82}{3} \begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix} = \begin{pmatrix} 254.90 \\ 401.96 \\ 656.86 \end{pmatrix}$$

The approximate solution $\begin{pmatrix} 911.76 \\ 1058.82 \\ 0 \end{pmatrix}$ has length 1397.29. New approximate solution $\begin{pmatrix} 254.90 \\ 401.96 \\ 656.86 \end{pmatrix}$ has length 811.18, a substantial decrease in length.

Theorem 4.5.1.3. Let A bean mn matrix with real entries and let $y \in \mathbb{R}^m$. Then $Ax = y$ has unique best approximate solution x. Necessary and sufficient conditions that x be the best approximate solution to $Ax = y$ are

1. $Ax = Proj_{A(\mathbb{R}^n)}(y)$,

2. x is in the column space of $A^{\perp}$, transpose of A.

Proof. Suppose that u and v satisfies the above two conditions stated in the theorem. Then the vector $w = v - u$ is in the nullspace N of A, since

$$Aw = Av - Au = Proj_{A(\mathbb{R}^n)}(y) = Proj_{A(\mathbb{R}^n)}(y) = 0$$

Moreover, $w \perp N$, since u and v are in the column space of A^T. Since, $w \in N$, $w \perp w \Rightarrow w = 0 \Rightarrow u = v$.

4.5.2 The Approximate Inverse of an $m \times n$ matrix.

Few well-known results :

1. Any $m \times n$ matrix A such that the equation $Ax = y$ has unique solution $x \in \mathbb{R}^n$ for every $y \in \mathbb{R}^m$ is an invertible $n \times n$ matrix.

2. For any $m \times n$ matrix A with real entries the equation $Ax = y$ has the unique best approximate solution $x \in R^n$ for every $y \in R^m$.

What, the, should we able to say ?

Let us denote the shortest approximate solution to $Ax = y$ by $A^-(y), \forall y \in \mathbb{R}^m$.

By Theorem 4.5.1.3, this means that $A^-(y)$ is an unique vector $v \in \mathrm{R}^n$ such that

1. $Av = Proj_{A(\mathbb{R}^n)}(y)$,

2. v is in the column space of A^{T}, transpose of A.

Theorem 4.5.2.1. Prove that the mapping $\mathrm{A}^- : \mathrm{R}^m \to \mathrm{R}^n$ is linear.

Proof. Let $y, z \in \mathrm{R}^m$ and $c \in \mathrm{R}$, we have

$$A(A^-(y) + \mathrm{A}^-(z) = AA^-(y) + AA^-(z)$$

$$= Proj_{A(\mathbb{R}^n)}(y) + Proj_{A(\mathbb{R}^n)}(\mathrm{z})$$

$$= Proj_{A(\mathbb{R}^n)}(y + z)(\textit{Verify!})$$

2. $A^-(y) + A^-(z)$ is in the column space of A^{T}, as $A^-(y)$ and $A^-(z)$ belongs.

It follows that $\mathrm{A}^-(y + z) = A^-(y) + A^-(z)$. By similar reasoning we have,

1. $A(cA^-(y)) = cAA\text{–}(y) = c\,Proj_{A(\mathbb{R}^n)}(y) = Proj_{A(\mathbb{R}^n)}(cy)$,

2. $cA^-(y)$ is in the column space of A^{T}.

From this it follows, $A^-(cy) = cA^-(y)$.

Since, every linear map has a matrix representation, therefore, A^- is $n \times m$ matrix.

Definition 4.5.2.1. For any $m \times n$ matrix A with real entries, we call the $n \times m$ matrix A^- the approximate inverse of A, since $A^-(y)$ is the shortest approximate solution x to $Ax = y$, $\forall y$. The approximate inverse is also called the Pseudo Inverse.

Theorem 4.5.2.2. Prove that $AA^- = Proj_{A(\mathbb{R}^n)}$ and $A^-A = Proj_{AT(\mathbb{R}^n)}$.

Proof. Since, A– maps y to the shortest x suchthat $Ax = Proj_{AT(\mathbb{R}^n)}$, AA^- maps y to $Proj_{A(\mathbb{R}^n)}(y)$. And since A maps x or y where upon A^- maps y to $Proj_{AT(\mathbb{R}^n)}(x)$, A^-A maps x to $Proj_{AT(\mathbb{R}^n)}(x)$.

Problem 4.5.2.1. Find the approximate inverse A– of the 3×2 matrix $A = \begin{pmatrix} 2 & 3 \\ 0 & 4 \\ 0 & 1 \end{pmatrix}$ and use it to find the shortest approximate solution $A^- \begin{pmatrix} 5000 \\ 4000 \\ 2000 \end{pmatrix}$ *of* $\begin{pmatrix} 2 & 3 \\ 0 & 4 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} p \\ q \end{pmatrix} = \begin{pmatrix} 5000 \\ 4000 \\ 2000 \end{pmatrix}$.

Solution 4.5.2.1. The projections of $\begin{pmatrix}1\\0\\0\end{pmatrix},\begin{pmatrix}0\\1\\0\end{pmatrix},\begin{pmatrix}0\\0\\1\end{pmatrix}$ on the column space of $\begin{pmatrix}2&3\\0&4\\0&1\end{pmatrix}$ are $\begin{pmatrix}1\\0\\0\end{pmatrix},\begin{pmatrix}0\\\frac{16}{17}\\\frac{4}{17}\end{pmatrix},\begin{pmatrix}0\\\frac{4}{17}\\\frac{16}{17}\end{pmatrix}$ and the solutions to the equations

$$\begin{pmatrix}2&3\\0&4\\0&1\end{pmatrix}x=\begin{pmatrix}1\\0\\0\end{pmatrix};\begin{pmatrix}2&3\\0&4\\0&1\end{pmatrix}x=\begin{pmatrix}0\\\frac{16}{17}\\\frac{4}{17}\end{pmatrix};\begin{pmatrix}2&3\\0&4\\0&1\end{pmatrix}x=\begin{pmatrix}0\\\frac{4}{17}\\\frac{16}{17}\end{pmatrix}\text{ are}$$

$$\begin{pmatrix}\frac{1}{2}\\0\end{pmatrix},\begin{pmatrix}-\frac{16}{17}\\\frac{4}{17}\end{pmatrix},\begin{pmatrix}-\frac{3}{34}\\\frac{1}{17}\end{pmatrix}.$$

Since, column space of the transpose $\begin{pmatrix}2&3\\0&4\\0&1\end{pmatrix}$ is R^2, the projections of $\begin{pmatrix}\frac{1}{2}\\0\end{pmatrix},\begin{pmatrix}-\frac{6}{17}\\\frac{4}{17}\end{pmatrix},\begin{pmatrix}-\frac{3}{34}\\\frac{1}{17}\end{pmatrix}$ on the column space of the transpose of $\begin{pmatrix}2&3\\0&4\\0&1\end{pmatrix}$ are just $\begin{pmatrix}\frac{1}{2}\\0\end{pmatrix},$ $\begin{pmatrix}-\frac{6}{17}\\\frac{4}{17}\end{pmatrix},\begin{pmatrix}-\frac{3}{14}\\\frac{1}{17}\end{pmatrix}$ themselves. So, theapproximate inverse of $\begin{pmatrix}2&3\\0&4\\0&1\end{pmatrix}$ is $\begin{pmatrix}\frac{1}{2}&-\frac{6}{17}&-\frac{3}{14}\\0&\frac{4}{17}&\frac{1}{17}\end{pmatrix}$ and the shortest approximate solution of

$$\begin{pmatrix}2&3\\0&4\\0&1\end{pmatrix}\begin{pmatrix}p\\q\end{pmatrix}=\begin{pmatrix}5000\\4000\\2000\end{pmatrix}is\begin{pmatrix}p\\q\end{pmatrix}=\begin{pmatrix}911.76\\1058.82\end{pmatrix}$$

4.5.3 Solving a matrix equation using its normal equation

Till now, we have found approximate solutions to $Ax = y$ and computed approximate inverse of A directly from the definitions. Now the question arises, are there any better methods ? Fortunately, we can find the approximate solutions x to $Ax = y$ by finding solutions x to the corresponding normal equation $A^TAx = A^Ty$. A^T being the transpose of A. In most applications, finding solutions to $A^TAx = A^Ty$ is easier than finding the

approximate solutions to $Ax = y$ directly. One reason for this is that A^TA is an $n \times n$ matrix when A is an $m \times n$ matrix. So, if $n < m$, which is true in most applications. A^TA is an $n \times n$ matrix which is of smaller size than A which is an $m \times n$ matrix. Another fact is that A^TA is symmetric, so similar to a diagonal matrix.

Why can we find approximate solutions x to $Ax = y$ by finding the solutions x to the corresponding normal equation $A^TAx = A^Ty$?

Theorem 4.5.3.1. Let $y \in R^m$. Then,

1. The approximate solutions x to $Ax = y$ are just the solutions x to the corresponding normal equations $A^TAx = A^Ty$,

2. $x = A^-y \Leftrightarrow A^TAx = A^Ty$ and x is in the column space of A^T

Proof. The condition $Ax = Proj_{A(Rn)}(y)$ on the element Ax of the column space A is that $y - Ax$ be orthogonal to the column space of A, *i.e.*, that $A^T(y - Ax) = 0$. But this is just the condition that s be a solution of $A^TAx = A^Ty$. Further, that x to be the shortest approximate solution $Ax = y$ is, by Theorem 4.5.1.3, equivalent to requiring that x be in the column space of A^T. Hence the proof.

Theorem 4.5.3.2. If the columns of A are linearly independent, then the matrix A^TA is invertible.

Proof. Since, A^TA is a square matrix, it is invertible if and only if its nullspace is 0. Let us choose x in such away that $A^TA = 0$, its suffices to show that $x = 0$. Multiplying by x^T, we have $0 = x^TA^TAx = (Ax)^T(Ax)$, implies $||Ax|| = 0 \Rightarrow Ax = 0$. Since, the columns of A are linearly independent, it follows that $x = 0$.

Corollary 4.5.3.1. Suppose that the columns of A are linearly independent. Then for any $y \in R^m$, there is unique approximate solution x to $Ax = y$, namely, $x = (A^TA)^{-1}A^Ty$.

Proof. If the columns of A are linearly independent, we know that A^TA is invertible (by theorem 4.3.2). Since, by theorem 4.3.1, x is an approximate solution of $Ax = y$ if and only if x is a solution of $A^TAx = A^Ty$, it follows that x is an approximate solution of $Ax = y$ if and only if $x = (A^TA)^{-1}A^Ty$.

It is very useful to have the explicit formulae $A^- = (A^TA)^{-1}A^T$ for the approximate inverse of A in the case that the columns of A are linearly independent. Is there such a formulae in general ? The answer to this is that, for a rectangular matrix $[A]_{m \times n}$, where $n < m$ and the columns are linearly independent, the approximate inverse of A denoted by A^- is defined by $A^- = (A^TA)^{-1}A^T$.

Example 4.5.3.1. Since, the columns of $A = \begin{pmatrix} 2 & 3 \\ 0 & 4 \\ 0 & 1 \end{pmatrix}$ are linearly independent, the

approximate inverse of A is

$$A^{-} = \left[\begin{pmatrix}2 & 0 & 0\\ 3 & 4 & 1\end{pmatrix}\begin{pmatrix}2 & 3\\ 0 & 4\\ 0 & 1\end{pmatrix}\right]^{-1}\begin{pmatrix}2 & 0 & 0\\ 3 & 4 & 1\end{pmatrix} = \begin{pmatrix}4 & 6\\ 6 & 26\end{pmatrix}^{-1}\begin{pmatrix}2 & 0 & 0\\ 3 & 4 & 1\end{pmatrix}$$

$$= \frac{1}{34}\begin{pmatrix}13 & -3\\ -3 & 2\end{pmatrix}\begin{pmatrix}2 & 0 & 0\\ 3 & 4 & 1\end{pmatrix} = \begin{pmatrix}\frac{1}{2} & -\frac{6}{17} & -\frac{3}{34}\\ 0 & \frac{4}{17} & \frac{1}{17}\end{pmatrix}$$

which agrees with our calculation of A^{-} of the preceding section.

4.5.4 Finding Functions that approximate data

In an experiment having an input variable x and output variable y, we get output values $y_0, \ldots\ldots, y_m$ corresponding to the input values $x_0, \ldots\ldots, x_m$ from data generated in an experiment.We then seek to find a useful functional approximation to the mapping $y_r = f(x_r)r = 1, 2, \ldots\ldots, m$, *i.e.* we want to find a function such as $y = ax^2 + bx + c$ such that y_r and $ax^2_r + bx_r + c$ are equal or nearly equal for $r = 1, 2, \ldots\ldots, m$. If we seek a functional approximation of order n, *i.e.* a function of the form $y = p(x) = c_0 + c_1x + \ldots\ldots + c_nx^n$, how do we choose the coefficients c_r? We write down the equations

$$c_0 + c_1x_0 + \ldots\ldots + c_nx_0^n = y_0$$
$$\vdots$$
$$c_0 + c_1x_m + \ldots\ldots + c_nx_m^n = y_m$$

considering x^s_r as the entries of the coefficient matrix A and the c_r as the unknowns. Thus we find the approximate solution

$$\begin{pmatrix}c_0\\ \vdots\\ c_n\end{pmatrix} \text{ to } \begin{pmatrix}1 & x_0 & \ldots & \ldots & x_0^n\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ 1 & x_m & \ldots & \ldots & x_m^n\end{pmatrix}\begin{pmatrix}c_0\\ \vdots\\ c_n\end{pmatrix} = \begin{pmatrix}y_0\\ \vdots\\ y_m\end{pmatrix},$$ for example by calculating the approximate inverse A^{-} of $A = \begin{pmatrix}1 & x_0 & \ldots & \ldots & x_0^n\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ 1 & x_m & \ldots & \ldots & x_m^n\end{pmatrix}$ and letting $\begin{pmatrix}c_0\\ \vdots\\ c_n\end{pmatrix} = A^{-}\begin{pmatrix}y_0\\ \vdots\\ y_m\end{pmatrix}$. The functional approximation $y = p(x) = c_0 + c_1x + \ldots\ldots + c_nx^n$ for the mapping $y_r = f(x_r)r$

= 1, 2, ..., m that we obtain in this way is the polynomial $p(x)$ of degree n for which the sum of the squares $(y_0 - p(x_0))^2 + ... + (y_m - p(x_m))^2$ is as small as possible.This method of finding functional approximation is often called the method of least squares.

Example 4.5.4.1. In a time study experiment that we conduct to find a functional relationship between the duration x of the tea break (in minutes) and the value y(in lakhs of rupees) of the work performed the same day by a group of employees, tea breaks of $x_0 = 10, x_1 = 15, x_2 = 21, x_3 = 5$

minutes duration and the values $y_0 = 10, y_1 = 14, y_2 = 13, y_3 = 10$

of lakhs of rupees worth of work performed were observed on the four successive days of the experiment. We decide to analyze the data by two ways viz first-order approximation and then use it to get second order approximation.

1. To begin the first-order approximation, we first get the general approximate solution to

$$\begin{pmatrix} 1 & x_0 \\ \vdots & \vdots \\ 1 & x_m \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \end{pmatrix} = \begin{pmatrix} y_0 \\ \vdots \\ y_m \end{pmatrix}$$

This is

$$\begin{pmatrix} c_0 \\ c_1 \end{pmatrix} = A^T \begin{pmatrix} y_0 \\ \vdots \\ y_m \end{pmatrix} = (A^T A)^{-} A^T \begin{pmatrix} y_0 \\ \vdots \\ y_m \end{pmatrix}, \text{ where A} = \begin{pmatrix} 1 & x_0 \\ \vdots & \vdots \\ 1 & x_m \end{pmatrix}, \textit{ i.e.}$$

$$\begin{pmatrix} c_0 \\ c_1 \end{pmatrix} = \left[\begin{pmatrix} 1 & \dots & 1 \\ x_0 & \dots & x_m \end{pmatrix} \begin{pmatrix} 1 & x_0 \\ \vdots & \vdots \\ 1 & x_m \end{pmatrix} \right]^{-1} \begin{pmatrix} 1 & \cdots & 1 \\ x_0 & \cdots & x_m \end{pmatrix} \begin{pmatrix} y_0 \\ \vdots \\ y_m \end{pmatrix}$$

$$= \begin{pmatrix} m+1 & 1.x \\ x.1 & x.x \end{pmatrix}^{-1} \begin{pmatrix} 1.y \\ x.y \end{pmatrix} or \begin{pmatrix} c_0 \\ c_1 \end{pmatrix} = \begin{pmatrix} m+1 & 1.x \\ x.1 & x.x \end{pmatrix}^{-1} \begin{pmatrix} 1.y \\ x.y \end{pmatrix}.$$

where $1 = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ and $u \,.\, v$ denotes $< u, v >$. In the time study $m = 3$ and we have

$x \,.1 = 10 + 15 + 21 + 5 = 51$

$x \,. x = 100 + 225 + 441 + 25 = 791$

$1 \,. y = 10 + 14 + 13 + 10 = 47$

$x \,.\, y = 100 + 210 + 273 + 50 = 633.$

So,

$$\begin{pmatrix} c_0 \\ c_1 \end{pmatrix} = \begin{pmatrix} m+1 & 1.x \\ x.1 & x.x \end{pmatrix}^{-1} \begin{pmatrix} 1.y \\ x.y \end{pmatrix} = \begin{pmatrix} 4 & 51 \\ 51 & 791 \end{pmatrix}^{-1} \begin{pmatrix} 47 \\ 633 \end{pmatrix}$$

$$= \frac{1}{563} \begin{pmatrix} 791 & -51 \\ -51 & 4 \end{pmatrix}^{-1} \begin{pmatrix} 47 \\ 633 \end{pmatrix} = \begin{pmatrix} 8.69 \\ 0.24 \end{pmatrix}$$

and we find that the linear function $y = 8.69 + 0.24x$ is the approximation of first order. Let's see how it approximates :

x	10	15	21	
actual y	10	14	13	10
approximating $y = 8.69 + 0.24x$	11.57	12.29	13.73	9.89

2. For the second order approximation, the general solution to

$$\begin{pmatrix} 1 & x_0 & x_0^2 \\ \vdots & \vdots & \vdots \\ 1 & x_m & x_m^2 \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} y_0 \\ \vdots \\ y_m \end{pmatrix}$$

is $\begin{pmatrix} c_0 \\ c_1 \\ c_2 \end{pmatrix} = A^{-} \begin{pmatrix} y_0 \\ \vdots \\ y_m \end{pmatrix} = (A^T A)^{-} A^T \begin{pmatrix} y_0 \\ \vdots \\ y_m \end{pmatrix}$, *where* $A = \begin{pmatrix} 1 & x_0 & x_0^2 \\ \vdots & \vdots & \vdots \\ 1 & x_m & x_m^2 \end{pmatrix}$, *i.e.*

$$\begin{pmatrix} c_0 \\ c_1 \\ c_2 \end{pmatrix} \text{ is given by } \left[\begin{pmatrix} 1 & \cdots & 1 \\ x_0 & \cdots & x_m \\ x_0^2 & \cdots & x_m^2 \end{pmatrix} \begin{pmatrix} 1 & x_0 & x_0^2 \\ \vdots & \vdots & \vdots \\ 1 & x_m & x_m^2 \end{pmatrix} \right]^{-1} \begin{pmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_m \\ x_m^2 & \cdots & x_m^2 \end{pmatrix} \begin{pmatrix} y_0 \\ \vdots \\ y_m \end{pmatrix}$$

But then $\begin{pmatrix} c_0 \\ c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} m+1 & 1.x & 1.x^2 \\ x.1 & x.x & x.x^2 \\ x^2.1 & x^2.x & x^2.x^2 \end{pmatrix}^{-1} \begin{pmatrix} 1.y \\ x.y \\ x^2.y \end{pmatrix}, 1 = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$

So, $\begin{pmatrix} c_0 \\ c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 4 & 51 & 791 \\ 51 & 791 & 13,761 \\ 791 & 13,761 & 255231 \end{pmatrix}^{-1} \begin{pmatrix} 47 \\ 633 \\ 10,133 \end{pmatrix}$.

For these c_0, c_1, c_2, the approximation of second order is the quadratic polynomial $y = c_0 + c_1x + c_2x^2$.

Theorem 4.5.4.1. The matrix $A = \begin{pmatrix} 1 & x_0 & \cdots & x_0^n \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & \cdots & x_n^n \end{pmatrix}$ is invertible if and only if the $x_0, x_1, \ldots\ldots, x_n$ are all distinct.

Proof. If two of the x_r are same then the two rows will be identical and hence, the matrix will not be invertible. Suppose, conversely, the matrix A is not invertible. Then the system of equations

$$c_0 + c_1x_0 + \ldots + c_nx^n{}_m = y_0$$
$$c_0 + c_1x_m + \ldots + c_nx^n{}_m = y_m$$

has a non-zero solution $\begin{pmatrix} c_0 \\ \vdots \\ c_n \end{pmatrix}$, *i.e.* $\exists$ a non-zero polynomial $p(x) = c_0 + c_1x + \ldots\ldots$ $+ c_nx^n$ of degree n which vanishes at all of the $n + 1$ numbers $x_0, x_1, \ldots\ldots, x_n$. Since the polynomial $p(x)$ of degree n has atmost n roots, two of the x_r are equal.

4.6 Linear Algorithms

Now that we have the theory and applications behind us, we ask : How can we instruct a computer to carry out the computations? So that we can get into the subject enough to get a glimpse of what it is about, we restrict ourselves to a single computational problem-but one of great importance the problem of solving a matrix-vector equation $Ax = y$ for x exactly or approximately, where the entries of A, x, and y are to be real. To instruct a computer to find x for a given A and y, we simply find a mathematical expression for x and devise an unambiguous recipe or step by step process for computing x from the expression. We call such a process analgorithm. How do we choose a particular algorithm to compute x? Many factors may be involved, depending on the uses that will be made of the algorithm. Here we want to be able to solve $Ax = y$ without knowing anything in advance about the size or nature of the matrix A or vectors y that we are given as input.

The algorithm that plays the most central role in this section, the row reduction algorithm discussed in section (5.2) may come about as close to satisfy ing all four of the foregoing properties as one could wish. It is simple and works for all matrices.At the same time it is very fast and since it uses virtually no memory except

the memory that held the original matrix, it is very efficient in its use of memory. This algorithm, which row reduces memory representing an ordinary matrix into memory representing that matrix in a special factored format, has another very desirable feature. It is reversible ; that is, there is a corresponding inverse algorithm to found orow reduction. This means that we can restore a matrix in its factored format back to its original unfactored format without performing cumbersome multiplications. To solve the equation $Ax = y$ exactly, we use the row reduction algorithm to replace A in memory by three very special matrices L, D, U whose product, in the case where no row interchanges are needed, is $A = LDU$. If rank $A = r$, L is an $m \times r$ matrix which is essentially lower triangular with 1's on the diagonal (its rows may have to be put in another order to make L truly lower triangular), D is an invertible $r \times r$ matrix, and U is an echelon $r \times n$ matrix. The entries of these factors, except for the known 0's and 1's, are nicely stored together in the memory that had been occupied by A. So that we can get them easily whenever we need them, we mark the locations of the r diagonal entries of D.

Example 4.6.1. When we apply the row reduction algorithm to the matrix

$$A = \begin{pmatrix} 6 & 12 & 18 & 0 \\ 2 & 9 & 6 & 10 \\ 3 & 7 & 10 & 7 \end{pmatrix}.$$

from which we extract the matrices

$$L = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{3} & 1 & 0 \\ \frac{1}{2} & \frac{1}{5} & 1 \end{pmatrix}, D = \begin{pmatrix} 6 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 1 \end{pmatrix}, U = \begin{pmatrix} 1 & 2 & 3 & 0 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 5 \end{pmatrix}$$

Solving $Ax = y$ for x now is reduced to solving $LDUx = y$ for x. To get x, we simply solve $Lv = y$, $Dw = v$ and $Ux = w$. These equations are easy to solve (or to determine to be unsolvable) because of the special nature of the matrices L, D, U. So in this way, we get our solution x to $Ax = y$, if a solution exists. When we are done,we can reverse the row reduction to restore the matrix A. This amounts to multiplying the factors L, D, U to get her in a very efficient way, getting back A = LDU. This is of great importance, since we may, with in a small time frame, want to go back and forth many times between the unfactored and factored formats of the matrix. After discussing this, in Sections 4.5.2 through 4.5.3 we go on to develop an algorithm for solving $Ax = y$ approximately. This algorithm also finds exact solutions, if they exist, but not as efficiently as does the algorithm for finding exact solutions. Here the algorithm replaces the matrix A by a factorization $A^{-} = J^{T}K$ of its approximate inverse, where J^{T}, in the case where no row interchanges are needed, is upper triangular. Solving $Ax = y$ approximately for x now is reduced to

solving $x = J^T Ky$, which we solve in two steps $z = Ky$ and $x = J^T z$. Finally, we illustrate how these algorithms can be used to build a computer program for solving $Ax = y$ and finding the exact or approximate inverse of a matrix A.

4.6.1 The LDU Factorization of A

For any $m \times n$ matrix A and vector $y \in R^{(m)}$, solving the matrix-vector equation $Ax = y$ for $x \in R^{(n)}$ which amounts to reducing the augmented matrix $[A, y]$ to an echelon matrix $[U, z]$ and solving $Ux = z$ instead. Let's look a gain at the reason for this, so that we can improve our methods. If M is the product of the inverses of the elementary matrices used during the reduction, a matrix that we can build and store during the reduction, then M is an invertible $m \times m$ matrix and $[MU, Mz] = [A, y]$. Since $Ux = z$ if and only if $MUx = Mz$, x is a solution of $Ux = z$ if and only if x is a solution of $Ax = y$. From this comes something quite useful. If we reduce A to its echelon form U, gaining M during the reduction, then $A = MU$ and for any right-hand-side vector y that we may be given, we can solve $Ax = y$ in two steps as follows:

(i) Solve $Mz = y$ for z.

(ii) Solve $Ux = z$ for x.

Putting these two steps together, we see that the x we get satisfies

$$Ax = MUx = Mz = y.$$

If no interchanges were needed during the reduction, M is an invertible lower triangular matrix. So since U is an echelonmatrix,bothequations $Mz = y$ and $Ux = z$ are easy to solve, provided that solutions exist.We have already seen how to solve $Ux = z$ for x by back substitution, given z. And we can get z from $Mz = y$, for a given y, using a reversed version of back substitution which we call forward substitution. Of course, if interchanges are needed during the reduction, we must also keep track of them and take them into account.

How do we find and store M so that $A = MU$? Let's look at few examples:

Example 4.6.1.2. Let's take $A = \begin{pmatrix} 6 & 12 & 8 & 0 \\ 2 & 9 & 6 & 10 \\ 3 & 7 & 10 & 7 \end{pmatrix}$ (unfactored format for the matrix A), and row reduce it to an echelon matrix U. As we reduce A, we keep track of certain non zero pivotentries a_{pq} and for each of them, we store each multiplier $\frac{a_{tq}}{a_{pq}}(t > p)$ as entry (t, q) after the (t, q) entry has been changed to zero as a result of performing the operation $Add\left(t, p; -\frac{a_{tq}}{a_{pq}}\right)$ means Row (t) + $\left(-\frac{a_{tq}}{a_{pq}} Row(p)\right)$. The

row operations $Add\left(2,1;-\frac{1}{3}\right), Add\left(3,1;-\frac{1}{2}\right), Add\left(3,2;-\frac{1}{5}\right)$ reduce A to the upper triangular matrix $V = \begin{pmatrix} 6 & 12 & 18 & 0 \\ 0 & 5 & 0 & 10 \\ 0 & 0 & 1 & 5 \end{pmatrix}$. If we write the pivot entries used during the reduction in bold face, the successive matrices encountered in the reduction are

$$\begin{pmatrix} \mathbf{6} & 12 & 18 & 0 \\ 2 & \mathbf{9} & 6 & 10 \\ 3 & 7 & \mathbf{10} & 7 \end{pmatrix}, \begin{pmatrix} \mathbf{6} & 12 & 18 & 0 \\ \frac{1}{3} & \mathbf{5} & 0 & 10 \\ 3 & 7 & \mathbf{10} & 7 \end{pmatrix}, \begin{pmatrix} \mathbf{6} & 12 & 18 & 0 \\ \frac{1}{3} & \mathbf{5} & 0 & 10 \\ \frac{1}{2} & \mathbf{1} & \mathbf{1} & \mathbf{7} \end{pmatrix}, \begin{pmatrix} \mathbf{6} & 12 & 18 & 0 \\ \frac{1}{3} & \mathbf{5} & 0 & 10 \\ \frac{1}{2} & \frac{1}{5} & \mathbf{1} & \mathbf{5} \end{pmatrix}$$

These matrices successively displace A in memory. The upper triangular part of the last one,

$$\begin{pmatrix} 6 & 12 & 8 & 0 \\ \frac{1}{3} & 5 & 0 & 10 \\ \frac{1}{2} & \frac{1}{5} & 1 & 5 \end{pmatrix} \text{(LVfactored format for } A\text{)},$$

holds $V = \begin{pmatrix} 6 & 12 & 18 & 0 \\ 0 & 5 & 0 & 10 \\ 0 & 0 & 1 & 5 \end{pmatrix}$ where as the lower part of it holds the lower entries of the matrix of multipliers $L = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{3} & 1 & 0 \\ \frac{1}{2} & \frac{1}{5} & 0 \end{pmatrix}$. *Our Claim : $A = LV$.* Of course, it is easy to compute the product to check that $A = LV$. To see why $A = LV$, however, note also that if we were to apply the same operations $Add\left(2,1;-\frac{1}{3}\right)$, $Add\left(3,1;-\frac{1}{2}\right)$, $Add\left(3,2;-\frac{1}{3}\right)$, to $L = I$, implies that L can be gotten by applying their inverses in the opposite order to I. So

$$L = Add\left(2,1;\frac{1}{3}\right) Add\left(3,1;\frac{1}{2}\right) Add\left(3,2;\frac{1}{5}\right) I$$

$$LV = Add\left(2,1;\frac{1}{3}\right) Add\left(3,1;\frac{1}{2}\right) Add\left(3,2;\frac{1}{5}\right) V = A.$$

Going on, we can factor $V = \begin{pmatrix} 6 & 12 & 18 & 0 \\ 0 & 5 & 0 & 10 \\ 0 & 0 & 1 & 5 \end{pmatrix}$ by taking the matrix of pivots $D = \begin{pmatrix} 6 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ and the echelon matrix $U = D^{-1} = \begin{pmatrix} 1 & 2 & 3 & 0 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 5 \end{pmatrix}$. So we get the factorizations $V = DU$ and $A = LDU$:

$$\begin{pmatrix} 6 & 12 & 18 & 0 \\ 0 & 5 & 0 & 10 \\ 0 & 0 & 1 & 5 \end{pmatrix} = \begin{pmatrix} 6 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 & 0 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 5 \end{pmatrix}$$

$$\begin{pmatrix} 6 & 12 & 18 & 0 \\ 0 & 5 & 0 & 10 \\ 0 & 0 & 1 & 5 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{3} & 1 & 0 \\ \frac{1}{2} & \frac{1}{5} & 0 \end{pmatrix} \begin{pmatrix} 6 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 & 0 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 5 \end{pmatrix}$$

Of course, we could do this directly, starting from where we left off with the matrix $\begin{pmatrix} 6 & 12 & 18 & 0 \\ \frac{1}{3} & 5 & 0 & 10 \\ \frac{1}{2} & \frac{1}{5} & 1 & 5 \end{pmatrix}$ Simply divide the upper entries (entries above the main diagonal) of V, row by row, by the pivot entry of the same row to get the upper entries of U, to get

$\begin{pmatrix} 6 & 12 & 18 & 0 \\ \frac{1}{3} & 5 & 0 & 10 \\ \frac{1}{2} & \frac{1}{5} & 1 & 5 \end{pmatrix}$ (LDU factored format for A).

So not only have we factored $A = MU$, but our M comes to us in the factored form $M = LD$, enabling us to store it in factored form by storing L and D.

From our example we see how to find and store M so that $A = MU$. In fact, M comes to us in a factored form $M = LD$, so that $A = LDU$, and the factors L, D, U are stored efficiently during the reduction. In the general case, things go the same way. If no interchanges are needed, we can reduce A to an upper triangular matrix V using only the elementary row operations $Add\left(t, p; \frac{a_{tq}}{a}\right)$, At the stage where we have a non zero entry a in row p and column q, the pivot entry, and use the operation

$Add\left(t, p; -\frac{a_{tq}}{a}\right)$, to make the (t, q) entry 0 for $t > p$, the (t, q) entry becomes available to us for storing the multiplier $\frac{a_{tq}}{a}$. Letting L be the corresponding lower triangular matrix of multipliers, consisting of the multipliers a, $\frac{a_{tq}}{a}$ (with $t > p$) used in the reduction, below the diagonal, and 1's on the diagonal, we get $A = LV$. Why? In effect, to get V we are multiplying A by the element ary matrices corresponding to $Add\left(t, p; -\frac{a_{tq}}{a}\right)$; and we are multiplying I by their inverses, in reverse order, to get L. To see this, just compute the product of their inverses in reverse order, which has the same effect as writing the same multipliers in the same places, but starting from the other end and working forward. For example, if $L = \begin{pmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 4 & 5 & 1 \end{pmatrix}$, it is the product

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 4 & 5 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 4 & 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 5 & 1 \end{pmatrix}.$$

Theorem 4.6.1.1. If no interchanges take place in the reduction of an $m \times n$ matrix A to an echelon matrix U, then $A = LDU$, where L is the lower triangular matrix of multipliers, D is the matrix of pivots, and U is the echelon matrix.

Proof. If we get $V = E_k \ldots\ldots E_1 A$, then we get $A = E_1^{-1} \ldots\ldots E_k^{-1} = LV$. After we get $A = LV$, we go on to factor V as $V = DD^{-1}V = DU$, where D is the diagonal matrix of pivots whose diagonal entry in row p is 1 if row p of V is 0 and a if a is the first nonzero entry of row p of V, and where U is the echelon matrix $D^{-1}v$. Then we can rewrite the product $A = LV$ as $A = LDU = MD$, where $M = LD$.

We can further simplify the factorization $A = LDU$ by throwing away parts of the matrices that are not needed. Letting r be the rank of A, we throw away all but the first r columns of L, all but the first r rows and columns of D, and all but the first r rows of U. Then we still have $A = LDU$, but now L is a lower triangular $m \times r$ matrix with 1's on the diagonal, D is an invertible diagonal $r \times r$ matrix, and U is an $r \times n$ echelon matrix.

4.6.2 The Row Reduction Algorithm and its Inverse

In order to give the algorithm, we must describe how to store an $m \times n$ matrix A in memory and how to perform and keep track of row interchanges. Of course, we must have an $m \times n$ array Memory (R, C) of real numbers in the memory of the computer, to hold the entries of A. So that we do not need to actually move entries when we perform a row or column interchange, we just keep track of the rows and columns by making and updating lists Row and Col of their rows and columns in memory. So if we load the 4×6 matrix into memory with the Row = [1, 2, 3, 4] and Col = [1, 2, 3, 4, 5, 6], we can keep track of Row, Col, and the entries, held in the array Memory (R, C), by the following 4×6 matrix structure A :

$$A = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 3 & 4 & 5 \\ 1 & 3 & 4 & 5 & 5 & 5 \\ 3 & 3 & 3 & 3 & 3 & 3 \\ 2 & 2 & 2 & 2 & 2 & 2 \end{pmatrix} \end{array}$$

All we mean by this is that the matrix A that we loaded in the computer was loaded by setting up the two lists Row = [1, 2, 3, 4] and Col = [1, 2, 3, 4, 5, 6], and putting the entries of A into the array Memory (R, C) according to the lists Row and Col. In this case, Row and Col indicate that the usual order should be used, so the entries occur in the array Memory (R, C) in the same order as they occur in A. So giving A is the same as giving the lists Row and Col and the array Memory (R, C). After loading A in memory in this way, suppose that we first interchange rows 3 and 4, then rows 4 and 2, then columns 2 and 4. The matrix A undergoes the following changes :

To make corresponding changes in the matrix structure A, we keep updating the lists :

Row = [1, 2, 3, 4] (after interchanging rows 3 and 4)

Row = [1, 4, 2, 3] (then after interchanging rows 4 and 2)

Col = [1, 4, 3, 2, 5, 6] (then after interchanging columns 2 and 4)

Let's look at the matrix structure A, which represents A as it undergoes the corresponding transformations :

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 3 & 4 & 5 \\ 1 & 3 & 4 & 5 & 5 & 5 \\ 3 & 3 & 3 & 3 & 3 & 3 \\ 2 & 2 & 2 & 2 & 2 & 2 \end{pmatrix} \end{array}, \quad \begin{array}{c} \\ 1 \\ 2 \\ 4 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 3 & 4 & 5 \\ 1 & 3 & 4 & 5 & 5 & 5 \\ 3 & 3 & 3 & 3 & 3 & 3 \\ 2 & 2 & 2 & 2 & 2 & 2 \end{pmatrix} \end{array}, \quad \begin{array}{c} \\ 1 \\ 4 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 3 & 4 & 5 \\ 1 & 3 & 4 & 5 & 5 & 5 \\ 3 & 3 & 3 & 3 & 3 & 3 \\ 2 & 2 & 2 & 2 & 2 & 2 \end{pmatrix} \end{array}, \quad \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 4 & 3 & 2 & 5 & 6 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 3 & 4 & 5 \\ 1 & 3 & 4 & 5 & 5 & 5 \\ 3 & 3 & 3 & 3 & 3 & 3 \\ 2 & 2 & 2 & 2 & 2 & 2 \end{pmatrix} \end{array}.$$

Definition 4.6.2.1. An m × n matrix structure A consists of Row, Col, and Memory (R, C), where Row is a 1–1 onto function from {1, ..., m} to itself, Col is a 1 – 1 on to function from {1, ...,n} to itself, and Memory (R, C) is an $m \times n$ array of real numbers.

When we write Row = [1, 4, 2, 3], we mean that Row is the mapping Row (1) = 1, Row (4) = 2, Row (2) = 3, Row (3) = 4 from {1, 2, 3, 4} to itself. Similarly, writing Col = [1, 4, 3, 2, 5, 6] means that Col is the mapping from {1, 2, 3, 4, 5, 6} to itself such that Col(s) = t, where s is in position t in the list {1, 4, 3, 2, 5, 6]. So since 4 is in position 2, Col(4) = 2. The 4 × 6 matrix structure consisting of Row = [1, 4, 2, 3], Col = [1, 4, 3, 2, 5, 6], and on 4 × 6 array Memory (R, C) is just

$$\begin{array}{c} \\ 1 \\ 4 \\ 2 \\ 3 \end{array}\begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 3 & 4 & 5 \\ 1 & 3 & 4 & 5 & 5 & 5 \\ 3 & 3 & 3 & 3 & 3 & 3 \\ 2 & 2 & 2 & 2 & 2 & 2 \end{pmatrix} \end{array}$$

Matrices get put in, or taken from, matrix structures according to

Definition 4.6.2.2. The $m \times n$ matrix A corresponding to the matrix structure A consisting of the lists Row, Col, and the $m \times n$ array Memory (R, C) is the $m \times n$ matrix A whose ($r \times s$) entry A_{rs} is given by the formula

$$A_{rs} = \text{Memory (Row}(r)\text{; Col}(s)\text{):}$$

For example, the 4 × 6 matrix A corresponding to the 4 × 6 matrix structure A described earlier is the matrix

$A = A_{rs}$ = Memory (Row (r), Col(s)).

which can easily be read from A when we write it out to look at :

$$\begin{array}{c} \\ 1 \\ 4 \\ 2 \\ 3 \end{array}\begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 3 & 4 & 5 \\ 1 & 3 & 4 & 5 & 5 & 5 \\ 3 & 3 & 3 & 3 & 3 & 3 \\ 2 & 2 & 2 & 2 & 2 & 2 \end{pmatrix} \end{array}, \qquad A = \begin{pmatrix} 0 & 3 & 0 & 1 & 4 & 5 \\ 3 & 3 & 3 & 3 & 3 & 3 \\ 2 & 2 & 2 & 2 & 2 & 2 \\ 1 & 5 & 4 & 3 & 5 & 5 \end{pmatrix}.$$

For example A_{43} = Memory (Row(4), Col(3)) = Memory(2, 3) is read from the matrix structure by going to the row of memory marked 4 (which is row 2 of memory) and to the column of memory marked 3 (which is column 3 of memory) and getting the entry $A_{43} = 4$ in that row and column. Of course, our objective in all of this been to

represent $m \times n$ matrices by $m \times n$ matrix structures and row and column interchanges on $m \times n$ matrices by corresponding operations on $m \times n$ matrix structures.

Definition 4.6.2.3. To interchange rows (respectively, columns) p and q of an $m \times n$ matrix structure A consisting of Row, Col, and Memory, just interchange the values of Row(p) and Row(q) [respectively, Col(p), Col(q)].

In our example above, we interchanged rows 4 and 2 when Row was Row = [1, 2, 4, 3]. The result there was that the values Row(4) = 3 and Row(2) = 2 of Row = [1, 2, 4, 3] were interchanged, resulting in the new list Row = [1, 4, 2, 3], the new values 2 and 3 for Row(4) and Row(2) having been obtained by interchanging the old ones, 3 and 2. We can now give the row reduction algorithm. This algorithm is reducing a matrix to an echelon matrix, but we've made some important changes and added some new features :

1. Our operations are performed on an $m \times n$ matrix structure rather than an $m \times n$ matrix, to make it easy to perform them and keep track of row interchanges.
2. Where '0' occurs there, we now say 'less in absolute value than e (where e is a fixed small positive value which depends on the computer to be used).
3. Instead of looking for the 'first nonzero value if any' in the rest of a given column, we look for the first value that is the largest in absolute value' in rest of that column.
4. As we reduce to the echelon matrix U, we keep track of the pivot entries and use the freed memory on and below them to store the entries of D and L. In particular, we record the number of pivot entries in the variable rank A. Of these changes, (2) and (3) lead to increased numerical stability. In other words, these changes are important if we prefer not to divide by numbers so small as to lead to serious errors in the computations. The others enable us to construct, store, and retrieve the factors, L, D, and U of A. They also enable us to reverse the algorithm and restore A.

Of course, the operations on the entries Ars of A performed in this algorithm are really performed as operations on Row, Col, and the array Memory (R, C), the correspondence of entries being A_{rs} = Memory (Row(r), Col(s)). This algorithm does not involve column interchanges and, infact, neither to the other algorithms considered. So, henceforth we take Col to be the identity list Col(s) = s and we do not label columns of a matrix structure.

Algorithm to row reduce an $m \times n$ matrix A to an echelon matrix U

Starting with (p, q) = (1, 1) and continuing as long as $p < m$ and $q < n$, do the following :

1. Get the first largest (in absolute value) (p', q) entry

$$a = A_{p'q} = \text{Memory (Row}(p'), \text{Col}(q)), \text{ of } A \text{ for } p' > p.$$

2. If its absolute value is less than, then we decrease the value of p by 1 (so later, when we increase p and q by 1, we try again in the same row and next column), but otherwise we call it a pivot entry and we do the following :

(a) We record that the (p, q) entry is the pivot entry in row p, we do this using a function Pivot list by setting Pivot list $(p) = q$.

(b) If $p' > p$, we interchange rows p and p' (by interchanging the values of Row (p) and Row (p')).

(c) For each row t with $t > p$, we perform the elementary row operation

$$Add(t, p; -\frac{A_{tq}}{a})(add -\frac{A_{tq}}{a} \text{ times row } p \text{ to row } t),$$

where A_{tq} is the current (t, q) entry of A; in doing this, we do not disturb the area in which we have already stored multipliers; we then record the operation by writing the multipler $\frac{A_{tq}}{a}$ as entry (t, q) of A ; (since we know that there should be a 0 there, we lose no needed information when we take over this entry as storage for our growing record).

(d) We perform the elementary row operation

$$\text{Multiply } \left(p; \frac{1}{a}\right) \text{ (divide the entries of row } p \text{ by } a);$$

we then record the operation by writing the divisor $d_p = a$ as entry (p, q) of A; (sinc we know there should be a 1 there, we lose no needed information when we take over this entry for our growing record).

3. We increase the values of p and q by 1 (on to the next row and column...). After all this has been done, we record that row $p - 1$ was the last nonzero row by setting the value rank $A = p - 1$.

This algorithm changes a matrix structure representing A in unfactored form at to a matrix structure representing A in LDU factored format. Since we keep track of the pivot sand the number r = rank A of pivots, we can get the entries of the echelon matrix U, the diagonal matrix D, and the lower triangular matrix L.

1. L is the $m \times r$ matrix whose (p, q) entry is

$$A_{pPivotList(q)} = \text{Memory (Row}(p), \text{Col(PivotList}(q)), \text{ for } p > q$$

$= 1$, for $p = q$

$= 0$, for $p < q$.

2. D is the $r \times r$ diagonal matrix with (q, q) entry

$A_{qPivotList(q)}$ = Memory(Row(q), Col(PivotList(q)); for $1 < q < r$.

3. U is the $r \times n$ echelon matrix whose $(p \times q)$ entry is

$A_{pPivotList(q)}$ = Memory (Row(p), Col(PivotList(q)), for $p > q$

$= 1$, for $p = q$

$= 0$, for $p < q$.

The $A = LDU$ factorization of the Section 4.6.1 is then replaced by a factorization $A = PLDU$, where P is the permutation matrix corresponding to the list Row, defined by

4. P is the $m \times m$ matrix whose (p, q) entry is 1 if p = Row(q) and 0 otherwise.

Example 4.6.2.1. Let $A = \begin{pmatrix} 3 & 7 & 10 & 7 \\ 6 & 12 & 18 & 0 \\ 2 & 9 & 6 & 10 \end{pmatrix}$ be represented by the matrix structure

$$A = \begin{matrix} 1 \\ 2 \\ 3 \end{matrix}\begin{pmatrix} 3 & 7 & 10 & 7 \\ 6 & 12 & 18 & 0 \\ 2 & 9 & 6 & 10 \end{pmatrix} \text{(unfactored format for } A\text{)},$$

with row = [1, 2, 3]. Then the row operations

$$\text{Interchange } (1, 2),\ Add\left(2,1;-\frac{1}{2}\right),\ Add\left(3,1;-\frac{1}{3}\right),$$

$$\text{Interchage } (2, 3),\ Add\left(3,2;-\frac{1}{5}\right)$$

reduce the matrix structure A to

$$V = \begin{matrix} 3 \\ 1 \\ 2 \end{matrix}\begin{pmatrix} 0 & 0 & 1 & 5 \\ 6 & 12 & 18 & 0 \\ 0 & 5 & 0 & 10 \end{pmatrix}$$

which represents the upper triangular matrix V = How do we get this, and what is the multiplier matrix? Writing the pivots in bold face, as in the earlier example, the successive matrices encountered in the reduction are :

$$\begin{matrix} 1 \\ 2 \\ 3 \end{matrix}\begin{pmatrix} 3 & 7 & 10 & 7 \\ 6 & 12 & 18 & 0 \\ 2 & 9 & 6 & 10 \end{pmatrix}, \begin{matrix} 2 \\ 1 \\ 3 \end{matrix}\begin{pmatrix} 3 & 7 & 10 & 7 \\ 6 & 12 & 18 & 0 \\ 2 & 9 & 6 & 10 \end{pmatrix}, \begin{matrix} 2 \\ 1 \\ 3 \end{matrix}\begin{pmatrix} \frac{1}{2} & 1 & 1 & 7 \\ 6 & 12 & 18 & 0 \\ 2 & 9 & 6 & 10 \end{pmatrix}$$

$$\begin{matrix}2\\1\\3\end{matrix}\begin{pmatrix}\frac{1}{2} & 1 & 1 & 7\\ 6 & 12 & 18 & 0\\ \frac{1}{3} & 5 & 0 & 10\end{pmatrix},\quad \begin{matrix}3\\1\\2\end{matrix}\begin{pmatrix}\frac{1}{2} & 1 & 1 & 7\\ 6 & 12 & 18 & 0\\ \frac{1}{3} & 5 & 0 & 10\end{pmatrix},\quad \begin{matrix}3\\1\\2\end{matrix}\begin{pmatrix}\frac{1}{2} & \frac{1}{5} & 1 & 5\\ 6 & 12 & 18 & 0\\ \frac{1}{3} & 5 & 0 & 10\end{pmatrix}$$

Now A has been reduced to

$$V = \begin{matrix}3\\1\\2\end{matrix}\begin{pmatrix}0 & 0 & 1 & 5\\ 6 & 12 & 18 & 0\\ 0 & 5 & 0 & 10\end{pmatrix}$$

and A to $$V = \begin{pmatrix}6 & 12 & 18 & 0\\ 0 & 5 & 0 & 10\\ 0 & 0 & 1 & 5\end{pmatrix}$$

and the matrix structure of multipliers is $L = \begin{matrix}3\\1\\2\end{matrix}\begin{pmatrix}\frac{1}{2} & \frac{1}{5} & 1\\ 1 & 0 & 0\\ \frac{1}{3} & 1 & 0\end{pmatrix}$,

with corresponding matrix $L = \begin{pmatrix}1 & 0 & 0\\ \frac{1}{2} & 1 & 0\\ \frac{1}{2} & \frac{1}{5} & 1\end{pmatrix}$

The matrix P is obtained by listing rows 1, 2, 3 of the identity matrix as the rows 3, 1, 2 of memory *i.e.*, P is the matrix

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

obtained from the matrix structure $L = \begin{matrix}3\\1\\2\end{matrix}\begin{pmatrix}\frac{1}{2} & \frac{1}{5} & 1\\ 1 & 0 & 0\\ \frac{1}{2} & 1 & 0\end{pmatrix}$

by removing the row labels and the multipliers.

Theorem 4.6.2.1: Suppose that A is represented by the matrix structure with Row = [1,...., m], which is reduced to its L.D.U. format with list Row updated during reduction to record the affect of interchanges. Then A is obtained by performing the product $PLDU$, where P,L,D,U are as described above.

4.6.3 Back and Forward Substitution : Solving $Ax = y$

Now that we can use the row reduction algorithm to go from the $m \times n$ matrix A to the matrices L, D, U and the list Row, which was built from the interchanges

during reduction, we ask : How do we use *P, L, D, U,* and Row to solve $Ax = y$? By Theorem 4.6.2.1, $A = PLDU$. So we can break up the problem of solving $Ax = y$ into parts, namely solving $Pu = y$, $Lv = u$, $Dw = v$, and $Ux = w$ for u, v, w, x. The only thing that is new here is solving $Pu = y$ for u. So, let's look at this in the case of the preceding example. There we have Row = (3, 1, 2) and $P = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$ is the corresponding permutation matrix. So, solving we obtain

$$\begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix},$$

for $\begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}$, we get $\begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} y_3 \\ y_2 \\ y_1 \end{pmatrix} = \begin{pmatrix} y_{Row(1)} \\ y_{Row(2)} \\ y_{Row(3)} \end{pmatrix}$, What this means is that we need make only one alteration in our earlier solution of $Ax = y$, namely, replace $\begin{pmatrix} y_2 \\ y_3 \\ y_1 \end{pmatrix} by \begin{pmatrix} y_{Row(1)} \\ y_{Row(2)} \\ y_{Row(3)} \end{pmatrix}$, So,we now have the

§ Algorithm for solving $Ax = y$ for x

Use the row reduction algorithm to get the matrices *L, D, U* and the list Row. Given a particular y in the column space of A, we do the following :

1. Solve $Lv = y$ by the forward substitution formula

$$v_p = y_{Row(p)} - \sum_{j=1}^{p-1} Lp_j v_j,$$

for $p = 1$ to the rank of A.

2. Solve $Dw = v$ by the formula $w_p = \dfrac{v_p}{D_{pp}}$

for $p = 1$ to the rank of A.

3. Solve $Ux = v$ by the back substitution equations :

$x_q = w_p - \sum_{j=q+1}^{n}$, if column q contains a pivot entry ; or

x_q = 1 ; if column q contains a no pivot entry

for $1 < q < n$.

When y is not in the column space of A, the above algorithm for solving $Ax = y$ exactly cannot be used. Instead, we can solve $Ax = y$ approximately by solving the normal equation $A^TAx = A^Ty$ exactly by the above algorithm. Since A^Ty is in the column space of A^TA, by our earlier discussion of the normal equation, this is always possible. So we have

§ Algorithm forsolving $Ax = y$ for x approximately:

We will use the algorithm for solving $A^TAx = A^Ty$ for x exactly. If we need to solve $Ax = y$ approximately for many different vectors y, it is more efficient first to find A^- and then to use it to get $x = A^-y$ for each y. In the next section, we give an algorithm for finding A^- for any A. We now turn to the important special case when the columns of A are linearly independent. In this case, we can solve the normal equation $A^TAx = A^Ty$ and find A^- efficiently by a simple algorithm involving the ***Gram-Schmidt Orthogonalization*** Process. From the columns v_1, v_2, ..., v_k of A, the Gram-Schmidt Orthogonalization Process gives us orthogonal vectors w_1, w_2, ..., w_k, where

$$w_s = v_s - \frac{< v_s, w_{s-1} >}{< w_{x-1}, w_{s-1} >} w_{s-1} - .... - \frac{< v_s, w_1 >}{< w_1, w_1 >} w_1$$

or $$v_s = w_s + \frac{< v_s, w_{s-1} >}{< w_{s-1}, w_{s-1} >} w_{s-1} + ... + \frac{< v_s, w_1 >}{< w_1, w_1 >} w_1$$

for $1 < s < k$. Letting $u_i = \left(\frac{1}{|w_1|}\right) w_1, ..., u_k = \left(\frac{1}{|w_k|}\right) w_k, b_{ss} = |w_s|$ and setting

$$b_{rs} = \frac{< v_s, w_r >}{< w_s, w_r >} |w_r|$$

for r < s, $1 < s < k$, we can rewrite this as

$$v_s = b_{1s}u_1 + ... + b_{s-1}, u_{s-1} + b_{ss}u_s$$

for $1 < s < k$. Letting Q be the $m \times k$ matrix whose columns are the ortho normal vectors u_1, u_2, ..., u_k and R be the $k \times k$ matrix whose (r, s) entry is b_{rs} for $r < s$ and 0 for $r > s$, these equations imply that

$$A = QR \text{ (Prove!)}$$

This is the so-called QR factorization of A as product of a matrix Q with orthonormal columns and an invertible upper triangular matrix R. (We leave it as an exercise for the reader to show that there is only one such factorization of A). So applying the Gram-

Schmidt Orthogonalization process to the columns of A to get orthonormal vectors u_1, u_2, ..., u_k in the above manner gives us the QR factorization $A = QR$.

e.g., if $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$, we get the orthogonal vectors

$$w_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

$$w_2 = \begin{pmatrix} 2 \\ 4 \end{pmatrix} - \frac{1.2+3.4}{1.1+3.3}\begin{pmatrix} 1 \\ 3 \end{pmatrix} = \begin{pmatrix} \frac{2}{5} \\ -\frac{2}{5} \end{pmatrix}$$

whose lengths are $|w_1| = \sqrt{10}$ and $|w_2| = \frac{1}{5}\sqrt{10}$. From these, we get the orthonormal vectors

$$u_1 = \frac{1}{\sqrt{10}}\begin{pmatrix} 1 \\ 3 \end{pmatrix}$$

$$u_2 = \frac{1}{\sqrt{10}}\begin{pmatrix} 3 \\ -1 \end{pmatrix}.$$

The equations $v_s = b_{1s}u_1 + ... + b_{s-1s}u_{s-1} + b_{ss}u_s \ (1 \le s \le k)$

are then $$\sqrt{10}u_1 = \begin{pmatrix} 1 \\ 3 \end{pmatrix}$$

$$\frac{7}{5}\sqrt{10}u_1 + \frac{1}{5}\sqrt{10}u_2 = \begin{pmatrix} 2 \\ 4 \end{pmatrix},$$

the matrices Q, R are $Q = \frac{1}{\sqrt{10}}\begin{pmatrix} 1 & 3 \\ 3 & -1 \end{pmatrix}$ and $\sqrt{10}\begin{pmatrix} 1 & \frac{7}{5} \\ 0 & \frac{1}{5} \end{pmatrix}$ and the QR factorization of $\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ is

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = \left(\frac{1}{\sqrt{10}}\begin{pmatrix} 1 & 3 \\ 3 & -1 \end{pmatrix}\right)\left(\sqrt{10}\begin{pmatrix} 1 & \frac{7}{5} \\ 0 & \frac{1}{5} \end{pmatrix}\right).$$

Given the QR factorization $A = QR$ for a matrix A with independent columns, the normal equation $A^TAx = A^Ty$,

can be solved for x easily and efficiently. Replacing A by QR in the normal equation

$$A^TAx = ATy,$$

can be solved for x easily and efficiently. Replacing A by QR in the normal equation $A^TAx = A^Ty$, we get $R^TQ^TQRx = R^TQ^Ty$. Since R and R^T are invertible and the columns of Q are orthonormal, this simplies to

$Rx = Q^Ty$,

$x = \mathrm{R}^{-1}Q^Ty$

(Prove!). Since R is upper triangular and invertible, the equation $Rx = Q^Ty$ can be solved for x using the back substitution equations

$$b_{qq}x_q = z_q - \sum_{j=q+1}^{k} b_{qj}x_j,$$

where Z_q is entry q of $Q^Ty(1 < q < k)$. Moreover, since R is invertible, there is only one such solution x. So x is the shortest approximate solution to $Ax = y$, from which it follows that $\mathrm{A}^- = R^{-1}Q^T$. Computing the inverse R^{-1} can be done very easily, since R is upper triangular. One simply finds the columns $c_1, c_2, \ldots, c_k$ of R^{-1} as the solutions c_s to the equations $Rc_s = e_s$ (column s of the $k \times k$ identity matrix) using back substitution equations :

$$b_{qq}c_{qs} = 0 - \sum_{j=q+1}^{k} b_{qj}c_{js}, \text{ for } q \neq s$$

$$b_{qq}c_{qs} = 1 - \sum_{j=q+1}^{k} b_{sj}c_{js}.$$

for each s with $1< s < k$. As it turns out, cjs = 0 for j > s. So the above back substitution equations simplify to the equations

$$b_{qq}c_{qs} = -\sum_{j=q+1}^{s} b_{qj}c_{js} \text{ for } q < s$$

$$b_{ss}c_{qs} = 1 \quad c_{ss} = 0, \text{ for } q < s$$

for $1 < s < k$. We summarize all of this by formulating the following algorithms.

§ Algorithm for solving $Ax = y$ for x approximately when the columns of A are linearly independent :

I. Use the QR factorization $A = QR$ to get the equation $Rx = Q^Ty$ (which replaces the normal equation $A^TAx = A^Ty$).

II. Solve $Rx = Q^T y$ for xx by the back substitution equations

$$b_{ss}x_y = z_q - \sum_{j=q+1}^{k} b_{qj}x_j$$

where z_q is entry q of $Q^T y$ for $1 < q < k$.

§ Algorithm for finding R^{-1} for an invertible upper triangular matrix R.

Letting $b_{rr}c_{ra}$ denote the (r, s) entry of an invertible upper triangular $k \times k$ matrix R, the entries c_{rs} of R^{-1} are determined by the back substitution equations

$$b_{qq}c_{qs} = -\sum_{j=r+1}^{s} b_{rj}c_{js}, \text{ for } r < s$$

$$b_{ss}c_{ss} = 1$$

$$c_{ss} = 0, \text{ for } r > s$$

for $1 < s < k$.

§ Algorithm for finding A^- when the columns of A are linearly independent :

(i) Use the QR factorization $A = QR$ to get Q and R.

(ii) Find R^{-1} by the above algorithm.

(iii) Then $A^- = R^{-1}Q^T$.

4.6.4 Approximate Inverse And Projection Algorithms

In section (4) we saw how to find the approximate solution x to an equation $Ax = y$, where A is an m n real matrix. To do this efficiently for each of a larger number of different y, we should first get A^- and then compute x as $x = A^-y$. How do we get A^-? In principle, we can get A^- by using the methods of section (4) to calculate each of its columns A^-e_s (where e_s is column s of the identity matrix) as an approximate solution xs to the equation $Ax_s = e_s$. However, there are more efficient methods, which are based on Theorem 4.6.2.1 and diagonalization of a symmetric matrix. Unfortunately, however, these methods are also some what complicated. To avoid the complications, we have worked out an efficient new method for finding A^- which uses only elementary row operations. This method, a variation of the method for finding the inverse of an invertible matrix, is based on two facts. The first of these is that the matrix $Proj_{ATR(m)}$ is just J^TJ where J is gotten by row reducing A^TA to an orthonormalized matrix in the sense of

Definition 4.6.4.1. An orthonormalized matrix is an $m \times n$ matrix J satisfying the following conditions :

(i) Each non zero row of J has length 1.

(ii) Any two different nonzero rows of J are orthogonal.

We can always row reduce a matrix to an orthonormalized upper triangular matrix. How ? First, reduce it to an echelon matrix. Then orthonormalize the nonzero rows in the reverse order r, ..., 1 by performing the following row operations on A for each value of k from r down to 1 :

(i) Multiply $\left(k, \frac{1}{u_k}\right)$, where u_k is the current length of row k.

(ii) For each value of q from $k-1$ down to 1, add $(q, k, -v_{kq})$, where v_{kq} is the inner product of the current rows k and q.

How do we show that $Proj_{AT(\mathbb{R}^m)}$ equals J^TJ ? First, we need some preliminary tools.

Definition 4.6.4.2. A matrix $P \in M_n(\mathbb{R})$ is a projection if $P = P^T$ and $P^2 = P$.

Theorem 4.6.4.1. If P is a projection, then $P = Proj_{P(\mathbb{R}^n)}$.

Proof. Let us denote the column space of P by W. Let v e R^n and write $v = v_1 + v_2$, where $v_1 \in W, v_2 \in W^{\perp}$. Then $v_1 = Pu$ for some u, so that $Pv_1 = P_2u = Pu = v_1$. Thus $Pv_1 = v_1$. Letting u now represent an arbitrary element of R^n, $Pu \in W$ such that Pu and v_2 are orthogonal.This implies that

$$0 = (Pu)^Tv_2 = u^TP^Tv_2 = u^TPv_2.$$

But then Pv_2 is orthogonal to u for all $u \in Rn$, which implies that $Pv_2 = 0$. It follows that $P_v = Pv_1 + Pv_2$ for all v, *i.e.* $P = Proj_w(v)$.

Theorem 4.6.4.2. The column spaces of A^TA and A^T are the same.

Proof. Certainly, the column space of A^TA is contained in the column space of A^T. Since the dimensions of the column spaces of A^TA and A^T are the ranks of A^TA and A, respectively, and since these are equal as we just saw, it follows that the column spaces of A^TA and A^T are equal.

Theorem 4.6.4.3. Let A be a real $m \times n$ matrix. Then for any orthonormalized matrix J that is row equivalent to A, $Proj_{AT(\mathbb{R}^m)}$ and the columns of $I - J^TJ$ span the null space of A.

Proof. Since J is an orthonormalized matrix, we get $(JJ^T)J = J$. But then $J^TJJ^TJ = J^TJ$. Since $(J^TJ)^T$ and $(J^TJ)^2 = J^TJ$, J^TJ is a projection and $J^TJ = Proj_{JTJ(\mathbb{R}^n)}$. Since J and A are row equivalent, J^T and A^T have the same column spaces. So, by Theorem (4.6.4.2), J^TJ ; J^T; A^T have the same column spaces. But then $J^TJ = Proj_{JTJ(\mathbb{R}^n)} = Proj_{AT(\mathbb{R}^n)}$. It follows that the columns of $I - J^TJ$ span the null-space of A, since :

(i) The nullspace of A is $(A^{\mathrm{T}}\mathbb{R}^{(m)})^{\mathrm{T}}$ (Prove!).

(ii) $(A^{\mathrm{T}}\mathbb{R}^{(m)})^{\mathrm{T}} = (J^{\mathrm{T}}J\mathbb{R}^{(n)})^{\perp} = I - J^{T}J\mathbb{R}^{(n)}$ since $J^{\mathrm{T}}J$ is a projection (Prove!).

Lemma 4.6.4.1. Let $P \in \mathrm{M}_{n}(R)$ satisfy the equation $PP^{\mathrm{T}} = P^{\mathrm{T}}$. Then P is a projection.

Proof. Since $P^{\mathrm{T}} = PP^{\mathrm{T}}$, $P = P^{\mathrm{T}}$. But then $P = P^{\mathrm{T}} = PP^{\mathrm{T}} = PP = P^{2}$, so P is a projection.

Theorem 4.6.4.4. Let A bean $m \times n$ real matrix, and let M be an invertible $n \times n$ matrix such that $J = MA^{\mathrm{T}}A$ is orthonormalized. Then $A^{-} = J^{\mathrm{T}}MA^{\mathrm{T}}$.

Proof. Let $B = J^{\mathrm{T}}MA^{\mathrm{T}}$. We claim first that

(i) $BA = J^{\mathrm{T}}J = Proj_{AT(\mathbb{R}^{n})}$.

(ii) $BAB = B$.

(iii) $AB = Proj_{AT(\mathbb{R}^{n})}$.

[I] Since B = $J^{\mathrm{T}}MA^{\mathrm{T}} = J^{\mathrm{T}}J$, and since $A^{\mathrm{T}}A$ and A| have the same column space by Theorem (4.6.4.2), (i) follows from Theorem (4.6.4.3).

[II] For (ii), we first use the equation $JJ^{\mathrm{T}}J = J$ from the proof of Theorem 4.6.4.3 to get the equation $J^{\mathrm{T}}JJ^{\mathrm{T}} = J^{\mathrm{T}}$. Then $BAB = J^{\mathrm{T}}JB = J^{\mathrm{T}}JJ^{\mathrm{T}}MA^{\mathrm{T}}$.

[III] For (iii.), we first show that AB is a projection. By Lemma 4.6.4.1 it suffices to show that $(AB)(AB)^{\mathrm{T}} = (AB)^{\mathrm{T}}$, which follows from the equations

$$AB(AB)^{T} = (AJ^{T}MA^{T})(AM^{T}JA^{T}) = AJ^{T}(MA^{T}A)M^{T}JA^{T}$$

$$= AJ^{T}JM^{T}JA^{T} = (AJ^{T}J)M^{T}JA^{T} = AM^{T}JA^{T} = (AB)^{T}.$$

Here we use the fact that since $JJA^{\mathrm{T}} = A^{\mathrm{T}}$, by Theorem 4.6.4.3, $AJ^{\mathrm{T}}J = \mathrm{A}$. (Prove!) Finally, the column space of A contains AB $\mathbb{R}^{(m)}$, which in turn contains $ABAR^{(n)}$ and $A(BA\,\mathbb{R}^{(n)})$, which in turn is $A(A^{\mathrm{T}}\mathbb{R}^{(m)})$ by (i). Since AA^{T} and A have the same column space by Theorem 4.6.4.2, it follows that all these spaces are actually equal *i.e.*

$$A\mathbb{R}^{(n)} = AB\mathbb{R}^{(m)} = ABA\mathbb{R}^{(n)} = A(A^{T}\mathbb{R}^{(m)}) = A\mathbb{R}^{(n)}.$$

So A and AB have the same column spaces. But then the projection AB is just $AB = Proj_{A(\mathbb{R}^{n})}$.

To show that $B = A^{-}$, let $x = By$. Then from (iii) we get that $Ax = ABy = \mathrm{Proj}_{AR^{(n)}}\, y$, and from (i) and (ii) we get that $\mathrm{Proj}_{AT\mathbb{R}^{(n)}}\, x = BAx = BABy = By = x$. So x is the shortest solution to $Ax = \mathrm{Proj}_{AT\mathbb{R}^{(n)}}\, y$ and $x = A^{-}y$. Since this is true for all y, we get that $B = \mathrm{A}^{-}$.

§ Algorithm to compute the projection $\mathbf{Proj}_{AT\mathbb{R}^{(n)}}$ for a real $m \times n$ matrix A

(i) Row reduce A^TA to an orthonormalized matrix $J = MA^TA$.

(ii) Then $Proj_{AT(\mathbb{R}^m)}$ is J^TJ.

§ Algorithm to compute the nullspace of a real m × n matrix A

(iii) Then the columns of $I - J^TJ$. span the nullspace of A.

§ Algorithm to compute the approximate inverse of a real m × n matrix A

(iv) Then A^- is J^TK, where $K = MA^T$.

§ Algorithm to compute the projection $Proj_{AT(\mathbb{R}^n)}$ for a real $m \times n$ matrix A

(v) Then $Proj_{AT(\mathbb{R}^n)}$ is AA^T.

§ Algorithm to find all approximate solutions of $Ax = y$

(vi) The shortest approximate solution to $Ax = y$ is $x = A^-y$, which we have by (iv).

(vii) Every approximate solution is $x + w$, where $w \in (I - J^TJ)\mathbb{R}^{(n)}$, by (iii).

It is instructive to look at some examples.

Example 4.6.4.1. To compute the projection of R^2 on to the column space of $\begin{pmatrix}1\\2\end{pmatrix}$, row reduce $\begin{pmatrix}1\\2\end{pmatrix}\begin{pmatrix}1\\2\end{pmatrix}^T = \begin{pmatrix}1 & 2\\ 2 & 4\end{pmatrix}$ to its orthonormalized echelon form $J = \begin{pmatrix}\frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}}\\ 0 & 0\end{pmatrix}$. Then the projection is $J^TJ = \begin{pmatrix}\frac{1}{5} & \frac{2}{5}\\ \frac{2}{5} & \frac{4}{5}\end{pmatrix}$.

Problem 4.6.4.1. Calculate A^- for $A = \begin{pmatrix}2 & 3\\ 0 & 4\\ 0 & 1\end{pmatrix}$

Solution 4.6.4.1. Since $A^TA = \begin{pmatrix}4 & 6\\ 6 & 26\end{pmatrix}$, we row reduce $\begin{pmatrix}4 & 6 & 2 & 0 & 0\\ 6 & 26 & 3 & 4 & 1\end{pmatrix}$ to the echelon form

$\begin{pmatrix}0 & 1.5 & 0.5 & 0 & 0\\ 0 & 1 & 0 & \frac{4}{17} & \frac{1}{17}\end{pmatrix}$ We then apply the operation Add $(1, 2, -v_{21})$ where v_{21} is the inner product 1.5 of (l, 1.5) and v = (0, 1), getting $\begin{pmatrix}1 & 0 & 0.5 & -\frac{5}{17} & -\frac{1.5}{17}\\ 0 & 1 & 0 & \frac{4}{17} & \frac{1}{17}\end{pmatrix}$. Since $J = I$, $A^- = IK = K = \begin{pmatrix}0.5 & -\frac{5}{17} & -\frac{1.5}{17}\\ 0 & \frac{4}{17} & \frac{1}{17}\end{pmatrix}$.

Problem 4.6.4.2. Calculate A^- for $A = \begin{pmatrix} 2 & 3 & 5 \\ 0 & 4 & 4 \\ 0 & 1 & 1 \end{pmatrix}$.

Solution 4.6.4.2. Here the columns of A are linearly dependent. Since, $A^TA = \begin{pmatrix} 4 & 6 & 10 \\ 6 & 26 & 32 \\ 10 & 32 & 42 \end{pmatrix}$ we row reduce $\begin{pmatrix} 4 & 6 & 10 & 2 & 0 & 0 \\ 6 & 26 & 32 & 3 & 4 & 1 \\ 10 & 32 & 42 & 5 & 4 & 1 \end{pmatrix}$ to the row reduced eschelon form as

$$\begin{pmatrix} 4 & 6 & 10 & 2 & 0 & 0 \\ 6 & 26 & 32 & 3 & 4 & 1 \\ 10 & 32 & 42 & 5 & 4 & 1 \end{pmatrix} \xrightarrow{\frac{1}{2}R_2} \begin{pmatrix} 2 & 3 & 5 & 1 & 0 & 0 \\ 6 & 26 & 32 & 3 & 4 & 1 \\ 10 & 32 & 42 & 5 & 4 & 1 \end{pmatrix} \xrightarrow[R_2 - 5R_1]{R_2 - 3R_1} \begin{pmatrix} 2 & 3 & 5 & 1 & 0 & 0 \\ 0 & 17 & 17 & 0 & 4 & 1 \\ 0 & 17 & 17 & 0 & 4 & 1 \end{pmatrix}$$

$$\xrightarrow{R_3 - R_2} \begin{pmatrix} 2 & 3 & 5 & 1 & 0 & 0 \\ 0 & 17 & 17 & 0 & 4 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \xrightarrow{\frac{1}{17}R_2} \begin{pmatrix} 2 & 3 & 5 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & \frac{4}{17} & \frac{1}{17} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

To avoid further fractions, we orthogonalize (2, 3, 5) and (0, 1, 1) directly, by the operation Add (1, 2, –4), where 4 was chosen as the inner product 8 of (2, 3, 5) and (0, 1, 1) divided by the inner product 2 of (0, 1, 1) and (0, 1, 1). We then get

$$\begin{pmatrix} 2 & -1 & 1 & 1 & -\frac{16}{17} & -\frac{4}{17} \\ 0 & 1 & 1 & 0 & \frac{4}{17} & \frac{1}{17} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

To normalize the orthogonal vectors (2, –1, 1), (0, 1, 1) to vectors of length 1, we apply the operations Multiply $\left(1, \frac{1}{\sqrt{6}}\right)$ and Multiply $\left(2, \frac{1}{\sqrt{2}}\right)$, getting

$$\begin{pmatrix} .8165 & -.4082 & .4082 & .4082 & -.3842 & -.0961 \\ 0 & .7071 & .7071 & 0 & .1664 & .0416 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

So,

$$A^- = J^TK = \begin{pmatrix} .8165 & -.4082 & .4082 & .4082 \\ 0 & .7071 & .7071 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} .4082 & -.3842 & -.0961 \\ 0 & .1664 & .0416 \\ 0 & 0 & 0 \end{pmatrix}$$

$$= \begin{pmatrix} .3333 & -.3137 & -0784 \\ -.1667 & .2745 & .0686 \\ .1667 & -.0391 & -.0098 \end{pmatrix}$$

4.7 Summary

The present unit deals with Fibonacci numbers incidence models and differential equations. The learners can explain how Least Squares Methods can povide approximate solutions of system of linear equations. They can also appreciate the method to find out the approximate inverse of non-square matrix. The unit also shows how to solve a system of linear equation using different linear algorithms like row reduction, matrix factorization, matrix inverse and Projection algorithm. They may use the concepts of this unit while specializing in their future course.

4.8 Exercises

1. Find α_n if

   a) $\alpha_n = 0$, $\alpha_1 = 4$, $\alpha_2 = 4, \ldots = \alpha_{n-1} + \alpha_{n-2}$ for $n \geq 2$.

   b) $\alpha_0 =$, $\alpha_1 = 1, \ldots, \alpha_n = 4\alpha_{n-1}, \ldots = \alpha_{n-1} + 30_{n-2}$ for $n \geq 2$.

   c) $\alpha_0 = 1$, $\alpha_1 = 3, \ldots, \alpha_n = 4\alpha_{n-1} + 3_{n-2}$ for $n \geq 2$.

   d) $\alpha_0 = 1$, $\alpha_1 = 1, \ldots, \alpha_n = 2\left(\alpha_{n-1} + \alpha_{n-2}\right)$ for $n \geq 2$.

2. Show directly that $\frac{1}{\sqrt{5}}\left\{\left[\left(1+\sqrt{5}\right)/2\right]^n - \left[\left(1-\sqrt{5}\right)/2\right]^n\right\}$ is an integer and is positive.

3. Draw the incidence diagra, for $T = \begin{pmatrix} 1 & -1 & 0 & 1 \\ -1 & 0 & 1 & 0 \\ 0 & 1 & -1 & -1 \end{pmatrix}$

4. Show that if D is an incidence diagram with m nodes that is not connected (some node cannot be reached from the first node). then the corresponding incidence matrix has rank less than m-1.

5. Find all approximate solution $\begin{pmatrix} 2 & 1 & 5 \\ 0 & 2 & 2 \\ 0 & 0 & 0 \end{pmatrix} x = \begin{pmatrix} 2 \\ 0 \\ 2 \end{pmatrix}$

6. Find the shortest approximate solution to $\begin{pmatrix} 2 & 1 & 5 \\ 0 & 2 & 2 \\ 0 & 0 & 0 \end{pmatrix} x = \begin{pmatrix} 2 \\ 0 \\ 2 \end{pmatrix}$

7. Find the appropriate inverse of $A = \begin{pmatrix} 1 & 1 \\ 0 & 2 \\ 1 & 1 \end{pmatrix}$ and use it to find an approximate solution to $Ax = \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}$

8. Find LDU for the matrices $\begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 4 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 1 & 2 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}$

9. If the matrix A is symmetric, that is, $A = A'$, and no interchanges take place when the row reduction algorithm is applied to A, show that in the resulting factorization A = LDU, L is the transpose of U.

10. Show that if QR = ST, where Q and S are $m \times k$ matrics, then Q = S and R = T.

4.9 References

[1] Lay, David C., Linear Algebra and its Applications, Pearson.

[2] Friedberg, S., Insel, A., Spence, L., Linear Algebra, Pearson New International Edition.

[3] Zhang, F, Matrix Theory: Basic Results and Techniques, Second Edition, Universitext, Springer.

[4] Strang, Gilbert, Linear Algebra and its Applications, Cengage Publications.

Unit 5 ❑ Matrix Theory

Structure

5.0 Objectives

5.1 Introduction

5.2 Special types of matrices

- **5.2.1 Idempotent, nilpotent, involution and projection**
- **5.2.2 Tri-diagonal matrices**
- **5.2.3 Circulant matrices**
- **5.2.4 Vandermonde matrices**
- **5.2.5 Hadamard matrices**
- **5.2.6 Permutation and doubly stochastic matrices**
  - **Frobenius konig theorem**
  - **Birkhoff theorem**

5.3 Positive Semi-definite matrices

- **5.3.1 Positive semi-definite matrices**
- **5.3.2 Square root of a positive semi-definite matrix**
- **5.3.3 A pair of positive semi-definite matrices**
- **5.3.4 Simultaneous diagonalization**

5.4 Symmetric matrices and quadratic forms

- **5.4.1 Diagonalization of symmetric matrices**
- **5.4.2 Quadratic Forms**
- **5.4.3 Constrained motion**
- **5.4.4 The singular value decomposition**
  - **Applications of the Singular Value Decomposition**
- **5.4.5 Applications to image processing and statistics**

5.5 Summary

5.6 Exercises

5.7 References

5.0 Objectives

The main objective of the matrix theory is to deal with various articles related with matrices and their applications in various areas.

5.1 Introduction

Modern work in matrix theory confined to either linear or algebraic techniques. The subject has a great deal of interaction with combinatories, group theory, graph theory, operator theory, and other mathematical disciplines. Matrix theory is still one of the richest branches of mathematics. This unit contains articles covering various types of Special types of matrices. Positive Semi-definite matrices and Symmetric matrices and quadratic forms.

5.2 Special type of matrices

5.2.1 Idempotence, Nilpotence, Involution, and Projections

We first present three types of matrices that have simple structures under similarity : idempotent matrices, nilpotent matrices, and involutions. We then turn attention to orthogonal projection matrices.

Definition 5.2.1.1. A square matrix A is said to be idempotent, or a projection, if

$$A^2 = A$$

nilpotent if for some $k \in Z^+$, $A^k = 0$;

and involutary if $A^2 = I$

where symbols have their usual meaning.

Theorem 5.2.1.1. Let A be a square complex matrix of order n. Then

1. A is idempotent if and only if A is similar to a diagonal matrix of the form diag (1, ..., 1, 0, ..., 0).

2. A is nilpotent if and only if all the eigenvalues of A are zero.

3. A is involutary if and only if A is similar to a diagonal matrix of the form diag (1, ..., 1, – 1, ..., –1).

Proof. 1. Necessary Part : Let $A = P^{-1}(J_1 \oplus ..... \oplus J_k)P$ be a Jordan decomposition of A. Then for each $i = 1, 2, ..., k$

$A^2 = A \Rightarrow J^2_i = J_i$.

if J is a Jordan block and if $J^2 = J$, then J must be of size 1; that is, J is a number. The assertion then follows. The sufficiency part is quite obvious.

2. Necessary Part : Consider the Jordan decomposition of A as

$$A = U^{-1}\begin{pmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} U.$$

where U is an n-square unitary matrix.

If $A^k = 0$, then each λ_i^k, and A has only zero eigenvalues. For sufficiency part, it is trivial to verify by computation that $A^n = 0$ if all the eigenvalues of A are equal to zero.

3. Exercise !

Theorem 5.2.1.2. Let A and B be nilpotent matrices of the same size. If A and B commute, then $A + B$ is nilpotent.

Proof. Let $A^m = 0$ and $B^n = 0$. On computation, we have

$$(A + B)^{m+n} = 0,$$

for each term in the expansion of $(A + B)^{m+n}$ is A^{m+n}, is B^{m+n}, or contains A^sB^t, $s > m$ or $t > n$. In any case, every term vanishes.

By choosing a suitable basis for C^n, we can interpret Theorem 5.2.1.1(1) as follows. A matrix A is a projection if and only if C^n can be decomposed as

(1.1) $$C^n = W_1 \oplus W_2$$

where W_1 and W_2 are subspaces such that for all $w_1 \in W_1$, $w_2 \in W_2$,

$Aw_1 = w_1$, $Aw_2 = 0$.

Thus, if $w = w_1 + w_2 \in C^n$, where $w_1 \in W_1$ and $w_2 \in W_2$, then

$Aw = Aw_1 + Aw_2 = w_1$.

Such a w_1 is called the projection of w on W_1. Here,

$$W_1 = Im\ \mathrm{A},\ W_2 = KerA = Im(\mathrm{I} - A).$$

Using this and Theorem 1.1.1(1), we are in a position to state the following theorem :

Theorem 5.2.1.3. For any $A \in M_n$, the following statements are equivalent:

(a) A is a projection matrix ; that is, $A^2 = A$.

(b) $C^n = ImA + KerA$ with $Ax = x$ for all $x \in Im\ A$.

(c) $KerA = Im(I - A)$.

(d) $rank\ A + rank\ (I - A) = n$.

(e) $Im\ A \cap Im(I - A) = \{0\}$.

We now turn our attention to orthogonal projection matrices. A square complex matrix A is called an orthogonal projection if

$$A^2 = A = \mathrm{A}^* \text{ (conjugate transpose of } A).$$

For orthogonal projection matrices, the subspaces

$W_1 = Im\ A$, $W_2 = Im(I - A)$

in (1.1) are orthogonal ; i.e., for all $w_1 \in W_1$ and $w_2 \in W_2$,

(1.2) $< w_1, w_2 > = 0.$

Since, $< w_1, w_2 > = < Aw_1, w_2 > = < w_1, Aw_2 >< w_1, Aw_2 > 0$,

therefore, $< Ax, (I - A)x > = 0, \ \forall x \in C^n$.

Theorem 5.2.1.4. For any $A \in M_n$, the following statements are equivalent :

(a) A is an orthogonal projection matrix ; *i.e.* $A^2 = A = A$.

(b) $A = U^*\ diag(1, ..., 1, 0 ....., 0)U$ or some unitary matrix U.

(c) $||x - Ax|| < ||x - Ay||$ for every x and y in C^n.

(d) $A^2 = A$ and $||Ax|| < ||x||$ for every $x \in C^n$.

(e) $A = A^*A$

Proof. (a) $\Leftrightarrow$ (b) : Let(a) holds. Because A is Hermitian, by the spectral decomposition theorem viz "Let A bean n-square complex matrix with eigenvalues λ_1, λ_2, ..., λ_n. Then A is normal if and only if A is unitarily diagonalizable *i.e.*, there exists a unitary matrix U such that

$$V^* AV = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_n)$$

In particular, A is Hermitian if and only if the λi are all real and is positive semidefinite if and only if the i are all nonnegative", we have $A = V$ diag $(\lambda_1, \lambda_2, ..., \lambda_n)$ V for some unitary matrix V, where the i are the eigenvalues of A. However, A is idempotent and thus has only eigenvalues 1 and 0 according to the previous theorem. It follows that

$$A = U^* diag(1, ..., 1, 0, ..., 0)U,$$

where rank $A = r$ and U is some unitary matrix. The converse part is obvious. (a)$\Leftrightarrow$(c) : Let (a) holds. Let A be a northogonal projection. We have the decomposition (1.1) with the orthogonality condition (1.2). Let $x = x_1 + x_2$, where $x_1 \in W_1, x_2 \in W_2$ and $< w_1, w_2> = 0$. Analogously, $y = y_1 + y_2\ y_1 \in W_1, y_2 \in W_2$. Now, as $x_1 \in W_1, y_1 \in W_1) \ x_1 - y_1 \in W_1$ and $W_1 \perp W_2$. Since, $<u, v> = 0 \rightarrow \|u\|^2 + \|v\|^2 = \|u + v\|^2$, we obtain

$$\| x - Ax \|^2 = \| x_2 \|^2 \leq \| x_2 \|^2 + \| x_1 - y_1 \|^2 = \| x_2 + (x_1 - y_1) \|^2 = \| x - Ay \|^2$$

It suffices to show that the decomposition (1.1) with the orthogonality condition (1) holds, where $W_1 = Im\ A$ and $W_2 = Im(I - A)$. Now,

$x = Ax + (I - A)x, \ \forall x \in C^n \Rightarrow C^n = Im\ A + Im(I - A)$.

Claim : $\forall x, y, x \in Im\ A, y \in Im(I - A) \Rightarrow <x, y> = 0$.

On the contrary, let us suppose that for some $x, y \in C_n, < (I - A)x, Ay > 6 = 0$.

We show that $\exists\ z \in C^n$ such that

$$\|x - Az\| < \|x - Ax\|;$$

which is a contradiction to the given condition(*c*). Let for some $x,\ y \in C^n$, $< (I - A)x, Ay > = \beta(= 0)$: We may assume that $\beta < 0$, otherwise replace x by e^{i0x}, where $\theta \in R$ is such that $e^{i\theta\beta} < 0$.

Let $Z_x = x - \in y$, where $\theta > 0$. Then

$$\| x - Ax_e \|^2 = \| (x - Ax) + (Ax - Az_\in) \|$$

$$= \| x - Ax \|^2 + \| Ax - Ax_\in \|^2 + 2 < (I - A)x, A(x - z_\in) >$$

$$= \| x - Ax \|^2 + \| Ax - Ax_\in \|^2 + 2 \in < (I - A)x, Ay >$$

$$= \| x - Ax \|^2 + \in^2 \| Ay \|^2 + 2 \in \beta$$

As, $\beta < 0$ we have $\in^2 \| Ay \|^2 + 2 \in \beta < 0$ for some small enough $\in$, which results in a contradiction to the assumption in (c).

(a) $\Rightarrow$ (d) : If A is an orthogonal projection matrix, then the orthogonality condition (1.2) holds. Thus, $< Ax, (I - A)x > = 0$

and

$$\| Ax \|^2 \leq \| Ax \|^2 + \| (I - A)x \|^2 = \| Ax + (I - A)x \|^2 = \| x \|^2 .$$

(d) $\Rightarrow$ (e): If $A \neq A^* A$, *i.e.* $(A^* - I)A \neq 0$ or $A^*(I - A) \neq 0$, then by Theorem 5.2.1.1(1), rank $(I - A) < n$ and dim $Im(I - A) < n$.

We show that $\exists x(\neq 0)$ such that

$< x, (I - A)x > = 0$, but $(I - A)x\ 6 = 0$.

Thus, for this x,

$$\| Ax \|^2 = \| x - (I - A)x \|^2 = \| x \|^2 + \| (I - A)x \|^2 > \| x \|^2,$$

which contradicts the condition $\|Ax\| < \|x\|$ for every $x \in C^n$.

To show the existence of such a vector x, it suffuces to show that $\exists x(\neq 0)$ such that

$$x \in (\mathrm{Im}(I - A))^\perp, but \notin Ker(I - A), i.e.\ (\mathrm{Im}(I - A))^\perp \subset Ker(I - A).$$

We know that

$$\dim Im(I - A) + \dim K\ \mathrm{er}(I - A) = n$$

and $\mathbb{C}^n = \mathrm{Im}(I - A) \oplus (\mathrm{Im}(I - A))^\perp$

Now, if $(\mathrm{Im}(I-A))^{\perp} \subseteq Ker(I-A), then\ (\mathrm{Im}(I-A))^{\perp} = Ker(I-A)\ as$

It follows,

$$Im(I-A) = Im(I-A^*)(\text{Verify !})$$

Hence, $I - A = I - A$ and A is Hermitian, which proves(e).

(e) => (a): If $A = A^*A$, implies A is Hermitian. Thus,

$$A = A^*A = AA = A^2.$$

5.2.2 Tridiagonal Matrices

One of the frequently used techniques in determinant computation is recursion. We illustrate this method by computing the determinant of a tridiagonal matrix and go on studying the eigenvalues of matrices of this kind. A square tridiagonal matrix of order n is a matrix with entries $t_{ij} = 0$ whenever $|i-j| > 1$. The determinant of a tridiagonal matrix can be calculated inductively. For simplicity, we consider the special tridiagonal matrix

(1.3) :
$$T_n = \begin{pmatrix} a & b & & & 0 \\ c & a & b & & \\ & c & a & b & \\ & & \ddots & \ddots & \ddots \\ & & & c & a & b \\ 0 & & & & c & a \end{pmatrix}$$

Theorem 5.2.2.1. Let T_n be defined as in (1.3).Then,

$$\det T_n = \begin{cases} a^n, & if\ bc = 0; \\ (n+1)\left(\frac{a}{2}\right)^n, & if\ a^2 = bc; \\ \frac{(\alpha^{n+1} + \beta^{n+1})}{\alpha - \beta}, & if\ a^2 \neq bc, \end{cases}$$

where, $\alpha = \dfrac{a + \sqrt{a^2 - bc}}{2}$, $\beta = \dfrac{a - \sqrt{a^2 - bc}}{2}$.

Proof. On expanding the determinant along the first row of the matrix in (1.3), we obtain the recursive formula as

(1.4) $\det T_n = a \det T_{n-1} - bc \det T_{n-2}$.

If $bc = 0$; then $b = 0$ or $c = 0$ and from (1.3), we obtain det $T_n = a^n$.

If $bc \neq 0$, let and be the solutions of $x^2 - ax + bc = 0$. Then,

$$\alpha + \beta = a,\ \alpha\beta = bc.$$

Now, we know that,

$$a^2 - 4bc = (\alpha - \beta)^2.$$

From the recursive formula (1.4), we have

$\det T_n - \alpha \det T_{n-1} = \beta(\det T_{n-1} - \alpha \det T_{n-2})$, *and*

$\det T_n - \beta \det T_{n-1} = \alpha(\det T_{n-1} - \beta \det T_{n-2})$.

Let us denote,

$f_n = \det T_n - \alpha \det T_{n-1}$, *and*

$g_n = \det T_n - \beta \det T_{n-1}$.

Then,

$$f_n = \beta f_{n-1}, g_n = \alpha g_{n-1},$$

with (by a simple computation),

$$f_2 = \beta^2, g_2 = \alpha^2,$$

Hence, $f_n = \beta^n$, $g_n = \alpha^n$, *i.e.*

(1.5) $\det T_n - \alpha \det T_{n-1} = \beta^n, \det T_n - \beta \det T_{n-1} = \alpha^n$.

Using T_{n+1} in (1.4) and subtracting, we obtain

$$\det T_n = \frac{(\alpha^{n+1} + \beta^{n+1})}{\alpha - \beta}, \text{ if } \alpha \neq \beta.$$

If $\alpha = \beta$, then by induction we have,

$$(n+1)\left(\frac{a}{2}\right)^n.$$

Theorem 5.2.2.2. If T_n is a tridiagonal matrix defined as in (1.3) with $a, b, c \in$ R and $bc > 0$, then the eigenvalues of T_n are all real and have eigenspaces of dimension one.

Proof. The first half follows from the argument prior to the theorem. For the second part, it is sufficient to prove that each eigenvalue has only one eigen vector upto a factor.

Let $x = (x_1, \ldots, x_n)$ be an eigenvector of T_n corresponding to the eigenvalue λ. Then $(\lambda I - T_n)x = 0, x \neq 0$,

or equivalently,

$$(\lambda - a)x_1 - bx_2 = 0$$

$$-cx_1 + (\lambda - a)x_{n-1} - bx_n = 0$$

$$-cx_{n-1}(\lambda - a)x_n = 0.$$

As $b = 0$, x_2 is determined by x_1 in the first equation, so are x_3, ..., x_n successively by x_2, x_3 and soon. If x_1 is replaced by kx_1, then x_2, x_3, ..., x_n become kx_2, kx_3, ..., kx_n, and the eigenvector is unique up to a factor.

Remark 5.2.2.1. The theorem is in fact true for a general tridiagonal matrix when a_i is real and $b_i c_i > 0$ for each i.

5.2.3 Circulant Matrices

An n-square circulant matrix is a matrix of the form

$$(1.6) \qquad \begin{pmatrix} C_0 & C_1 & C_2 & .... & C_{n-1} \\ C_{n-1} & C_n & C_1 & .... & C_{n-2} \\ C_{n-2} & C_{n-1} & C_n & .... & C_{n-3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ C_1 & C_2 & C_3 & .... & C_0 \end{pmatrix}$$

where C_0, C_1, C_2, ..., C_{n-1} are complex numbers.

Example 5.2.3.1.

$$N = \begin{pmatrix} 1 & 2 & 3 & .... & n \\ n & 1 & 2 & .... & n-1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 3 & 4 & 5 & .... & 2 \\ 2 & 3 & 4 & .... & 1 \end{pmatrix}$$

and

$$(1.7) \qquad \overset{\vee}{P} = \begin{pmatrix} 0 & 1 & 0 & .... & 0 \\ 0 & 0 & 1 & .... & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & .... & 1 \\ 1 & 0 & 0 & .... & 0 \end{pmatrix}$$

are circulant matrices. Note that $\overset{\vee}{P}$ is also a permutation matrix. We refer to this $\overset{\vee}{P}$ as the $n \times n$ primary permutation matrix.

Theorem 5.2.3.1. An n-square matrix C is circulant if and only *if*

$$C = \overset{\vee}{P}C\overset{\vee}{P}^{T},$$

where P is the $n \times n$ primary permutation matrix.

Proof. The proof follows by direct verication.

Theorem 5.2.3.2. Let C be a circulant matrix in the form (1.6), and let $f(\lambda) = C_0 + C_1\lambda + ... + C_{n-1}\lambda^{n-1}$.

Then,

(i) $C = f(\overset{\vee}{P})$, where $\overset{\vee}{P}$ is the $n \times n$ primary permutation matrix.

(ii) C is a normal matrix ; i.e. $C^* C = CC^*$.

(iii) The eigenvalues of C are $f(w^k)$; $k = 0, 1, 2, ..., n - 1$.

(iv) det $C = f(\omega^0) f(\omega^1) f(\omega^2) ..... f(\omega^{n-1})$, where ω denotes nth root of unity.

(v) F^*CF is a diagonal matrix, where F is the unitary matrix with the (i, j) entry equal to

$$\frac{1}{\sqrt{n}}\omega^{(i-1)(j-1)}, i, j = 1,2,....., n.$$

Proof. (i) Verify by direct computation.

(ii) is due to the fact that if matrices A and B commute, so do $p(A)$ and $q(B)$, where p and q are any polynomials (Verify). Here, $\bar{P}\hat{P}^* = \bar{P}^* \hat{P}$.

(iii) The characteristic polynomial of P is

$$\det(\lambda I - P) = \lambda^n - 1 = \prod_{k=0}^{n-1}(\lambda - \omega^k).$$

Thus, the eigenvalues of $\overset{\vee}{P}$ and $\overset{\vee}{P}^i$ are, respectively, ω^k and ω^{ik}, $k = 1, 2, ..., n - 1$. It follows that the eigenvalues of $C = f(\overset{\vee}{P})$ are $f(\omega^k)$; $k = 0, 1, 2, ..., n - 1$ and that

$$\det C = \prod_{k=0}^{n-1} f(\omega^k).$$

(iv) Same as the (iii).

(v) Let $x_k = (1, \omega, \omega^k, \omega^{2k} ....., \omega^{(n-1)k}, 1)$, $k = 0, 1, 2, ..., n - 1$. Then,

$$\overset{\vee}{P}x_k = (\omega, \omega^k, \omega^{2k}....., \omega^{(n-1)k}, 1)^T = \omega^k x_k ;$$

and $$Cx_k = f(\overset{\vee}{P})x_k = f(\omega^k)x_k,$$

i.e. x_k are the eigenvectors of $\overset{\vee}{P}$ and C corresponding to the eigenvalues ω^k and $f(\omega^k)$, respectively, $k = 0, 1, ..., n-1$. However, because

$$< x_i, x_j >= \prod_{k=0}^{n-1} \omega^{jk}\omega^{ik} = \sum_{k=0}^{n-1} \omega^{(i-j)k} = \begin{cases} 0, & i \neq j; \\ n, & i = j, \end{cases}$$

we have that

$$\left\{ \frac{1}{\sqrt{n}} x_0, \frac{1}{\sqrt{n}} x_1, ..., \frac{1}{\sqrt{n}} x_{n-1} \right\}$$

is an orthonormal basis for C^n. Thus, we get a unitary matrix as

$$F = \begin{pmatrix} 1 & 1 & 1 & .... & 1 \\ 1 & \omega & \omega^2 & .... & \omega^{n-1} \\ 1 & \omega^2 & \omega^4 & .... & \omega^{2(n-1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \omega^{n-1} & \omega^{2(n-1)} & .... & \omega^{(n-1)(n-1)} \end{pmatrix}$$

such that $F^*CF = diag(f(\omega^0)f(\omega^0).....f(\omega^{n-1}))$.

That F is aunitary matrix is veried by a direct computation.

Remark 5.2.3.1. The unitary matrix F, called a Fourier matrix, is independent of C.

5.2.4 Vandermonde Matrices

Denition 5.2.4.1. An n-square Vandermonde matrix., denoted by $V_n(a_1), a_2, ..., a_n$ or simply V is a matrix of the form

$$\begin{pmatrix} 1 & 1 & 1 & ..... & 1 \\ a_1 & a_2 & a_3 & ..... & a_n \\ a_1^2 & a_2^2 & a_3^2 & ..... & a_n^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_1^{n-1} & a_2^{n-1} & a_3^{n-1} & ..... & a_n^{n-1} \end{pmatrix}$$

Vandermonde matrices play a role in many places such as interpolation problems in Numerical Analysis and solving systems of linear equations. We consider the determinant and the inverse of a Vandermonde matrix in this section.

Theorem 5.2.4.1. Let $V_n(a_1, a_2, ..., a_n)$ be a Vandermonde matrix.Then and $V_n(a_1, a_2, ..., a_n)$ is invertible if and only if all the a_i are distinct.

Proof. We proceed with the proof by induction. There is nothing to show if $n = 1, 2$. Let $n > 3$. Suppose the assertion is true when the size of the matrix is $n - 1$. For the case of n, subtracting row i multiplied by a_1 from row $i + 1$, for i going down from $n - 1$ to 1, we have

$$\det V = \begin{vmatrix} 1 & 1 & 1 & .... & 1 \\ 0 & a_2 - a_1 & a_2 - a_1 & .... & a_n - a_1 \\ 0 & a_2(a_2 - a_1) & a_3(a_2 - a_1) & .... & a_n(a_n - a_1) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & a_2^{n-2}(a_2 - a_1) & a_3^{n-2}(a_3 - a_2) & ..... & a_n^{n-2}(a_n - a_1) \end{vmatrix}$$

$$= \begin{vmatrix} a_2 - a_1 & a_2 - a_1 & ..... & a_n - a_1 \\ a_2(a_2 - a_1) & a_2(a_2 - a_1) & ..... & a_n(a_n - a_1) \\ \vdots & \vdots & \vdots & \vdots \\ a_2^{n-2}(a_2 - a_1) & a_2^{n-2}(a_3 - a_1) & ..... & a_n^{n-2}(a_n - a_1) \end{vmatrix}$$

$$= \prod_{j=2}^{n} (a_2 - a_1) \det V_{n-1}(a_2, ..., a_n)$$

$$= \prod_{j=2}^{n} (a_3 \prod_{2 \le i < j \le n} (a_j - a_1), \text{ (by (hypothesis)}$$

$$= \prod_{1 \le i < j \le n} (a_j - a_1).$$

It is readily seen that the Vandermonde matrix is singular if and only if atleast two of the a_i are equal.

Theorem 5.2.4.2. For any integers $k_1 < k_2 < ..... < k_n$, the quotient

$$\frac{\det V_n(k_1, k_2, ..., k_n)}{\det V_n(1, 2, ..., n)},$$

is an integer.

Proof. Let f_i be any monic polynomial of degree i for $i = 1, 2, ..., n$. The additive property of determinants shows that

$$(1.8) \qquad \begin{vmatrix} 1 & 1 & 1 & .... & 1 \\ f_1(k_1) & f_1(k_2) & f_1(k_3) & .... & f_1(k_n) \\ f_2(k_1) & f_2(k_2) & f_2(k_3) & .... & f_2(k_n) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ f_{n-1}(k_{n-1}) & f_{n-1}(k_2) & f_{n-1}(k_3) & .... & f_{n-1}(k_n) \end{vmatrix}$$

is the same as det $V_n(k_1, k_2, ..., k_n)$. By taking, for any integer a,

$f_i(a) = a(a-1)(a-2) ..... (a - i + 1) = i!\binom{a}{i}$.

we see that f_i (a) is divisible by$(i-1)!$. Factoring out $(i-1)!$ from row i, $i = 1, 2, ..., n$, we see that the determinant in (1.8), thus det $V_n(k_1, k_2, ..., k_n)$, is divisible by the product $\prod_{i=1}^{n}(i-1)!$. The proof is complete, for $\prod_{i=1}^{n}(i-1)! = \det V_n(1, 2, ..., n)$.

We now turn our attention to the inverse of a Vandermonde matrix. Consider the polynomial in x given by the product

$$p(x) = (x + a_1)(x + a_2) ..... (x + a_n),$$

where $a_1, a_2, ..., a_n$ are constants. Expand $p(x)$ as a polynomial

$p(x) = s_0x^n + s_1x^{n-1} + s_2x^{n-2} + ... + s_{n-1}x + s_n$;

where $s_0 = 1$ and for each $k = 1, 2, ..., n$,

$$s_k = s_k(a_1, a_2, ..., a_n) = \sum_{1 \le p_1 \le p_2 \le ... \le p_k \le n} \prod_{q=1}^{k} a_{pq}.$$

We refer to s_k , depending on $a_1, a_2, ..., a_n$, as the k-th elementary symmetric function of $a_1, a_2, ..., a_n$. (for details refer to [4])

Theorem 5.2.4.3. Suppose that a_i, $i = 1, 2, ..., n$ are distinct. Then $V_n(a_1, a_1a_2, ..., a_n)^{-1} = (\alpha_{ij})$ where for each pair of i and j,

$$\alpha_{ij} = \frac{(-1)^{i+j}\sum_{p_1<....<p_{n-1}}\prod_{q=1,pq\neq i}^{n-j} a_{p_q}}{\prod_{k=1,k\neq 1}^{n}(a_k - a_i)}$$

Proof. Recall from elementary linear algebra ([3]) that the entries of the inverse of the matrix V are the cofactors of order $n - 1$ divided by det V i.e.,

$$V^{-1} = \left(\frac{1}{\det V}c_{ij}\right)^T.$$

where c_{ij} is the cofactor of the (i, j)-entry of V. Now we compute the cofactors c_{ij}. Let V_k be the matrix obtained from V by deleting row $k + 1$ (the kth powers) and adjoining as a new nth row the nth powers of the a_i. We show

(1.9) $\det V_k = s_{n-k} \det V.$

Augment V with the nth powers of the a_i as the $(n + 1)$th row and with $(1, -x, (-x)^2, ..., (-x)^n)$ as the first column. Denote the resulting matrix by W. Then W is a Vandermonde matrix and

$$\det W = (x + a_1)(x + a_2) ..... (x + a_n) \det V$$
$$= (x_n + s_1 x^{n-1} + ... + s^{n-1}x + s_n) \det V.$$

(1.10)

Expanding det W along the first column, we have

(1.11) $\det W = \det V_0 + x \det V_1 + ... + x^n \det V.$

By comparing equations (1.10) and (1.11), we obtain identity (1.9). Each cofactor c_{ij} is a determinant of order $n - 1$ in the same form as det V_k. Let $V(a_j)$ and $s_k(a_j)$ denote, the $(n - 1)$ square Vandermonde matrix and the kth elementary symmetric function of $a_1, a_2, ..., a_n$ without a_j. Using equation (1.9) we have respectively,

$$c_{ij} = (-1)^{i+j} \det V(i \mid j)$$

$$= (-1)^{i+j} s_{(n-1)-(i-1)}(a_j) \det V(a_j)$$

$$= (-1)^{i+j} s_{(n-1)}(\bar{a}_j) \det V(\hat{a}_j)$$

Hence,

$$\frac{1}{\det V} c_{ij} = \frac{(-1)^{i+j} s_{(n-i)}(\hat{a}_j) \det V(\hat{a}_j)}{\prod_{t>s}(a_t - a_s)}$$

$$= \frac{(-1)^{i+j} s_{(n-1)}(\hat{a}_1)}{\prod_{s<j}(a_j - a_s) \prod_{j<i}(a_i - a_j)}$$

$$= \frac{(-1)^{i+j} s_{(n-i)}(\hat{a}_i)}{\prod_{k=1, k\neq j}^{n}(a_k - a_j)}$$

$$= \frac{(-1)^{i+j} \sum_{p_1<...<p_{n-j}} \prod_{q=1, p_q \neq i}^{n-j} a_{p_q}}{\prod_{k=1, k\neq i}^{n}(a_k - a_i)}$$

$$\alpha_{ij} = \frac{(-1)^{i+j} \sum_{p_1<....<p_{n-1}} \prod_{q=1, p_q \neq i}^{n-i} a_{p_q}}{\prod_{k=1, k\neq i}^{n}(a_k - a_i)}$$

§ Application of Vandermonde Matrices

An interesting application follows : Let $A \in M_n$. Then,

$$A^n = 0,\ trA^k = 0,\ k = 1, 2, ..., n$$

Because $A^n = 0$, A is nilpotent, thus, A has only zero eigenvalues; so does A^k for each k. For the other way round, let the eigenvalues of A be $\lambda_1, \lambda_2, ..., \lambda_n$. Then the trace identities imply

$$\lambda_1 + \lambda_2 + ... + \lambda_n = 0$$
$$\lambda^2_1 + \lambda^2_2 + ... + \lambda^2_n = 0$$
$$... = ...$$
$$\lambda^n_1 + \lambda^n_2 + ... + \lambda^n_n = 0,$$

rewritten as

$$V_n(\lambda_1, \lambda_2, ..., \lambda_n)(\lambda_1, \lambda_2, ..., \lambda_n)^T = 0.$$

If all of the λ_i are distinct, then by the preceding theorem the Vandermonde matrix is non singular and the system of equations in $\lambda_1, \lambda_2, ..., \lambda_n$ has only the trivial solution $\lambda_1 = \lambda_2 = ... = \lambda_n = 0$. If some of the i are identical, for instance, $\lambda_1 = \lambda_2$ and $\lambda_2 = \lambda_3 = ... = \lambda_n$ are distinct, we then write the system as

$$V_{n-1}(\lambda_2, ..., \lambda_v)(2\lambda_2, ..., \lambda_v)^T = 0$$

A similar argument will result in $\lambda_2 = \lambda_3 = ... = \lambda_n = 0$.

This idea applies to the interpolation problem of finding a polynomial $f(x)$ of degree at most $n - 1$ satisfying

$f(x_i) = y_i,\ i = 1, 2, ..., n,$

where x_i and y_i are given constants [7].

Exercise 5.2.1.1. Let $x_1, x_2, ..., x_n$ be different numbers. Show that for any set of n numbers $y_1, y_2, ..., y_n$ there exists a polynomial $f(x)$ of degree atmost $n - 1$ such that

$$f(x_i) = y_i,\ i = 1, 2, ..., n,$$

In particular, for any numbers $\lambda_1, \lambda_2, ..., \lambda_n$, there exist polynomials $g(x)$, and $h(x)$ if each $\lambda_i > 0$, of degree at most $n - 1$ such that

$$g(\lambda i) = \lambda_i,\ h(\lambda_i) = \sqrt{\lambda_i}, i = 1,2,3,....,n.$$

5.2.5 Hadamard matrices

Definition 5.2.5.1. An n-square matrix A is called a Hadamard matrix if each entry of A is 1 or -1 and if the rows or columns of A are orthogonal *i.e.*,

$$AA^T = I_n,\ A^TA = I_n.$$

Exercise 5.2.5.1. Prove that for any real matrix of order n, $AA^T = I_n$ and $A^TA = I_n$ are equivalent.

Example 5.2.5.1. The following are two examples of Hadamard matrices *viz*

$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \end{pmatrix}$$

If A is a Hadamard matrix, then so is AP for any matrix P with entries ± 1 satisfying $PP^T = I$. Thus, one may change the -1 in the first row of A to $+1$ by multiplying an appropriate matrix P with diagonal entries ± 1. There is only one 2×2 Hadamard matrix of this kind. Can one construct a 3×3 Hadamard matrix? The answer to this question is the following theorem *viz*

Theorem 5.2.5.1. Let $n > 2$. A necessary condition for an n-square matrix A to be a Hadamard matrix is that n is a multiple of 4.

Proof. **Case I** Let $A = (a_{ij})$ bean n-square Hadamard matrix. The entries of A are ± 1, the equation $AA^T = nI$ yields

$$\sum_{k=1}^{n} a_{ik}a_{jk} = \begin{cases} 0, & \text{if } i = j; \\ n, & \text{if } i = j. \end{cases}$$

Upon computation, we have

$$\sum_{k=1}^{n} (a_{ik} - a_{2k})(a_{1k} + a_{3k}) = \sum_{k=1}^{n} a_{1k}^2 + \sum_{k=1}^{n} a_{1k}a_{2k} + \sum_{k=1}^{n} a_{1k}a_{3k} + \sum_{k=1}^{n} a_{2k}a_{3k}$$

$$= \sum_{k=1}^{n} a_{1k}^2 = n.$$

The possible values for $a_{1k} + a_{2k}$ and $a_{1k} + a_{3k}$ are + 2, 0, –1. Thus, each term in the summation

$$\sum_{k=1}^{n} (a_{1k} + a_{2k})(a_{1k} + a_{3k})$$

must be +4, 0, or –4. It follows that n is divisible by 4 *i.e.* 4 | n.

Case II Let P be an n-square matrix with main diagonal entries 1 or –1 such that the first row of AP consists entirely of +1. Here, AP is also a Hadamard matrix. Since the second and third rows of AP are orthogonal to the first row, they must each have the same number, say r, of + 1s and –1s. Thus $n = 2r$ is an even number.

Let $n^{+}-$ be the number of columns of AP that contain a + 1 of row 2 and a –1 of row 3. Similarly, we define $n^{-}{}_{+}$, $n^{+}{}_{+}$ and $n^{-}{}_{-}$. Then

$$n_{+}^{+} + n_{-}^{+} = n_{+}^{+} + n_{+}^{-} = n_{-}^{-} + n_{-}^{+} = r.$$

Thus,

$$n_{+}^{+} = n_{-}^{-} \text{ and } n_{-}^{+} = n_{+}^{-}.$$

The orthogonality of rows 2 and 3 implies that

$$n_{+}^{+} = n_{-}^{-} = n_{-}^{+} + n_{+}^{-} \Rightarrow n_{-}^{+} = n_{-}^{+} \Rightarrow n = 2r = 4n_{+}^{+}$$

is a multiple of **4**.

Theorem 5.2.5.2. If A is a Hadamard matrix, then so is

$$\begin{pmatrix} A & A \\ A & -A \end{pmatrix} \tag{1.12}$$

By this theorem, Hadamard matrices H_n of order 2n can be generated recursively by defining

$$H_1 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \; {}_1H_n = \begin{pmatrix} H_{n-1} & H_{n-1} \\ H_{n-1} & -H_{n-1} \end{pmatrix}, n \geq 2. \tag{1.13}$$

Theorem 5.2.5.3. Let H_n be defined as in (1.13). Then H_n has eigenvalues $+2^{\frac{n}{2}}$ and $-2^{\frac{n}{2}}$ each of multiplicity 2^{n-1}, and an eigenvector x_n corresponding to the positive eigenvalue $2^{\frac{n}{2}}$.

Proof. The proof is done by induction on n. The case of $n = 1$ was discussed just prior to the theorem. Now for $n > 2$, we have

$$\det(\lambda I - H_n) = \begin{vmatrix} \lambda I - H_n & -H_{n-1} \\ -H_{n-1} & \lambda I + H_n \end{vmatrix}$$

$$= \det(\lambda I - H_{n-1})(\lambda I + H_{n+1}) - H_{n-1}^2)$$

$$= \det(\lambda^2 I - 2H_{n-1}^2)$$

$$= \det(\lambda I - \sqrt{2}H_{n-1})\det(\lambda I + \sqrt{2}H_{n-1}).$$

Thus each eigenvalue μ of H_{n-1} generates two eigenvalues $+\sqrt{2}\mu$ of H_n. The assertion then follows by the induction hypothesis, for H_{n-1} has eigenvalues $+2^{\frac{(n-1)}{2}}$ and

$-2^{\frac{(n-1)}{2}}$ each of multiplicity 2^{n-1}. To see the eigenvector part, we observe that, by induction again,

$$H_n x_n = \begin{pmatrix} H_{n-1} & H_{n-1} \\ H_{n-1} & -H_{n-1} \end{pmatrix} \begin{pmatrix} x_{n-1} \\ (-+\sqrt{2})x_{n-1} \end{pmatrix}$$

$$= \begin{pmatrix} \sqrt{2}H_{n-1}x_{n-1} \\ (2-\sqrt{2})H_{n-1}x_{n-1} \end{pmatrix} = 2^{\frac{n}{2}} \begin{pmatrix} x_{n-1} \\ (-+\sqrt{2})x_{n-1} \end{pmatrix} = 2^{\frac{n}{2}} x_n.$$

Let J_n denote the n-square matrix whose entries are all equal to 1. We give a lower bound for the size of a Hadamard matrix that contains a J_n as a submatrix.

Theorem 5.2.5.4. If A is an n-square Hadamard matrix that contains a J_n as a submatrix, then $m > n^2$.

Proof. We may assume by permutation that A is partitioned as

$$A = \begin{pmatrix} J_n & X \\ Y & Z_3 \end{pmatrix}. \tag{1.14}$$

where Z_s is an s-square matrix of entries 1, and $s = m - n$. Since A is a Hadamard matrix of size $m = n + s$, we have

$AA^T = (n + s)I_m$,

which implies, by using the block form (1.14) of A, that

$J^2_n + XX^T = (n + s)I_n$.

Thus,

$XX^T = (n + s)I_n - nJ_n$. (1.15)

The eigenvalues of the right-hand matrix in (1.15) are

$$n+s-n^2,\ n+s,\ \ldots..,\ n+s$$

However, XX^T is positive semidefinite, and thus has nonnegative eigenvalues. Therefore, $n + s - n^2 > 0$ or $m > n^2$.

5.2.6 Permutation and Doubly Stochastic Matrices

Our goal in this section is to show that every permutation matrix is a direct sum of primary permutation matrices under permutation similarity and that every doubly stochastic matrix is a convex combination of permutation matrices.

A square matrix is called a permutation matrix if each row and column of the matrix has exactly one 1 and all other entries are 0. It is easy to see that there are n! permutation matrices of size n. Furthermore, the product of two permutation matrices

of the same size is a permutation matrix, and if P is a permutation matrix, then P is invertible, and $P^{-1} = P^{T}$.

A n-square A is said to be reducible if there exists a permutation matrix P such that

$$P^{T}AP = \begin{pmatrix} B & C \\ O & D \end{pmatrix} \tag{1.16}$$

where B and D are square matrices of order atleast 1. A matrix is said to be irreducible if it is not reducible. Here a matrix of order 1 is considered to be irreducible. The matrix $P^{T}AP$ in equation (1.16) is similar to A through the permutation matrix P. We say that they are permutation similar. It is obvious that the diagonal entries of irreducible permutation matrices are all equal to 0, but not viceversa. For example,

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

Theorem 5.2.6.1. Every reducible permutation matrix is permutation similar to a direct sum of irreducible permutation matrices.

Proof. Let A bean n-square reducible permutation matrix, as in equation (1.16). The matrix C in this case must be zero, for otherwise, let B be $r \times r$ and D be $s \times s$, where $r + s = n$. Then B contains r 1s (in columns) and D contains s 1's (in rows). If C contained a 1, then A would have at least $r + s + 1 = n + 1$ 1's, a contradiction. The assertion then follows by the induction on B and D.

We now show that every n-square irreducible permutation matrix is permutation similar to the $n \times n$ primary permutation matrix

$$P = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ 1 & 0 & 0 & \dots & 0 \end{pmatrix} \tag{1.17}$$

Theorem 5.2.6.2. A primary permutation matrix is irreducible.

Proof. Suppose the $n \times n$ primary permutation matrix P is reducible. Let $S^{T}PS = J_1 \quad J_k$, $k > 2$, where S is some permutation matrix and the J_i are irreducible matrices with order less than n. The rank of $P - I$ is $n - 1$, for $\det(P - I) = 0$ and the submatrix of size $n - 1$ by deleting the last row and the last column from $P - I$ is nonsingular. It follows that

rank $(S^{T}PS - I)$ = rank $(S^{T}(P - I)S = n - 1$.

By using the above decomposition, we obtain

$$\text{rank } (S^{T}PS - I) = \sum_{i=1}^{n} rank(J_i - I) \leq n - k < n - 1.$$

This is a contradiction. The proof is complete.

Theorem 5.2.6.3. A permutation matrix is irreducible if and only if it is permutation similar to a primary permutation matrix.

Proof. Let Q be an $n \times n$ permutation matrix and P the $n \times n$ primary permutation matrix in equation (1.17). If Q is permutation similar to P, then Q is irreducible by the previous theorem. Conversely, suppose that Q is irreducible. We show that Q can be brought to P through simultaneous row and column permutations. Let the 1 of the first row be in the position $(1, i_1)$. Then $i_1 = 1$ since Q is irreducible. If $i_1 = 2$, we proceed to the next step, considering the 1 in the second row. Otherwise, $i_1 > 2$. Permute columns 2 and i_1 so that the 1 is placed in the (1, 2) position. Permute rows 2 and i_1 to get a matrix Q_1. This matrix is permutation similar to Q and also irreducible. If the (2, 3)-entry of Q_{11} is 1, we go on to the next step. Otherwise, let the $(2, i_2)$- entry be 1, $i_2 = 3$. If $i_2 = 1$, then Q_1 would be reducible, for all entries in the first two columns but not in the first two rows equal 0. Thus, $i_2 > 3$. Permute columns 3 and i_2 so that the 1 is in the (2, 3) position. Here the 1 in the (1, 2) position was not affected by the permutations in the second step. Continuing in this way, one obtains the permutation matrix P in the form of equation (1.17). The product of a sequence of permutation matrices is also a permuttation matrix, therefore we have a permutation matrix S such that

$$S^{T}QS = S^{-1}QS = P.$$

Remark 5.2.6.1. Combining the above theorems, we see that every reducible permutation matrix is permutation similar to a direct sum of primary permutation matrices. Moreover, the rank of an n-square irreducible permutation matrix minus I is $n - 1$.

Exercise 5.2.6.1. Let P bean $n \times n$ irreducible permutation matrix. Show that

$$\text{rank } (P - I) = n - \text{I}$$

Theorem 5.2.6.4. Let Q be an n-square permutation matrix. Then Q is irreducible if and only if the eigen-values of Q are 1, ω, ω^2, ..., ω^{n-1}, where ω is an n-th primitive root of unity.

Proof. If Q is irreducible, then Q is similar to the $n \times n$ primary permutation matrix, according to Theorem 1.6.3, which has the eigen-values 1, ω, ω^2, ..., ω^{n-1}, where ω is an n-th primitive root of unity; so does matrix Q.

Conversely, suppose that 1, ω. ω^2, ..., ω^{n-1} are the eigenvalues of Q. Here $\omega^k = 1$ for any $1 < k <$ n since ω is an nth primitive root of unity. If Q is reducible, then we may write

$$S^TQS = J_1 \oplus ..... \oplus J_k,$$

where S is a permutation matrix, and the J_i are primary permutation matrices with order less than n. The eigenvalues of those J_i are the eigenvalues of Q, none of which is an nth primitive root of unity, for the order of every J_i is less than n. This is a contradiction. Thus, Q is irreducible.

We next present a beautiful relation between permutation matrices and doubly stochastic matrices, a type of matrices that plays an important role in statistics and in some other subjects.

Definition 5.2.6.1. A square matrix is said to be doubly stochastic if all entries of the matrix are non negative and the sum of the entries in each row and each column equals 1. Equivalently, a matrix A with nonnegative entries is doubly stochastic if

(1.18) $\qquad e^TA = e^T$, and $Ae = e$, e = (1, 1, 1, ..., 1)^T.

It is readily seen that permutation matrices are doubly stochastic and so is the product of two doubly stochastic matrices. We show that a matrix is a doubly stochastic matrix if and only if it is a convex combination of nite permutation matrices. To prove this, we need a result, which is of interest in its own right.

§ Frobenius–Konig Theorem

Let A bean n-square complex matrix. Then every product of n entries of A taken from distinct rows and columns equals 0, in symbols,

(1.19) $\quad a_{1i_1}, a_{2i_2}, a_{3i_3}, a_{4i_4}, \ldots\ldots\ a_{ni_n} = 0, \ \{i_1, i_2, \ldots\ldots, i_n\} = (1,2,3,\ldots.,n)$

if and only if A containsan $r \times s$ zero submatrix, where $r + s = n + 1$.

Remark 5.2.6.2. First notice that property (1.19) of A will remain true when row or column permutations are applied to A. In other words, an n-square matrix A has property (1.19) if and only if PAQ has the property, where P and Q are any n-square permutation matrices.

Proof. Necessary Part : If all the entries of A are zero, there is nothing to prove. Suppose A has a nonzero entry and consider the submatrix obtained from A by deleting the row and the column that contain the nonzero entry. An application of induction on the $(n - 1) \times (n - 1)$ submatrix results in a zero submatrix of size $p \times q$, where $p + q = (n - 1) + 1 = n$. We thus may write A, by permutation, as

$$A = \begin{pmatrix} B & C \\ O & D \end{pmatrix}.$$

where B is $q \times q$ and D is $p \times p$. Since every product of the entries of A from different rows and columns is 0, this property must be in herited by B or D, say B. Applying the induction to B, we see that B has an $1 \times s$ zero submatrix such that $1 + s = q + 1$. Putting this zero submatrix in the lower-left corner of B, we see that A has an $r \times s$ zero submatrix, where $r = p + 1$ and $r + s = n + 1$.

Sufficient part : We may assume by permutation that the $r \times s$ zero submatrix is in the lower-left corner, and write

$$A = \begin{pmatrix} B & C \\ O & D \end{pmatrix}$$

Because $n \times r = s - 1$, B is of size $(s - 1) \times s$. Thus, the remust be a zero among any s entries taken from the first s columns and any s different rows. Therefore, every product $a_{1i_1}, a_{2i_2}, a_{3i_3}, a_{4i_4}, \ldots\ldots a_{ni_n}$ has to contain a zero factor, hence equals zero.

§ Birkho Theorem

A matrix A is doubly stochastic if and only if it is a convex combination of permutation matrices.

Proof. Necessary Part: We apply induction on the number of zero entries of the doubly stochastic matrices. If A has (at most) $n^2 - n$ zeros, then A is a permutation matrix, and we have nothing to show. Suppose that the doubly stochastic matrices with at least k zeros are convex combinations of permutation matrices. We show that the assertion holds for the doubly stochastic matrices with $k - 1$ zeros. Let A be an n-square doubly stochastic matrix of $k - 1$ zero entries. If every product of the entries of A from distinct rows and columns is zero, then A may be written as, upto permutation,

$$A = \begin{pmatrix} B & C \\ O & D \end{pmatrix}$$

where the zero submatrix is of size $r \times s$ with $r + s = n + 1$. Since the entries in each column A add upto 1, the sum of all entries of B equals s. Similarly, by considering rows, the sum of all entries of D is r. Thus, the sum of all entries of A would be atleast $r + s = n + 1$. This is impossible, for the sum of all entries of A is n. Therefore, some product $a_{1i_1}, a_{2i_2}, a_{3i_3}, a_{4i_4}, \ldots\ldots, a_{ni_n} = 0$. Let P_1 be a permutation matrix with 1 in the positions (j, i,), i = 1, 2, ..., n, and 0 elsewhere. Consider the matrix

$$E = (1 - \delta)^{-1}(A - \delta P_i),$$

where $\delta = \min(a_{1i_1}, a_{2i_2}, a_{3i_3}, a_{4i_4}, \ldots\ldots, a_{ni_n})$. It is readily seen by equation (1.18) that E is also a doubly stochastic matrix and that E has atleast one more zero than A. By the induction hypothesis, there are positive numbers $t_1, t_2, \ldots, t_m$ of sum 1, and permutation matrices $P_2, P_3, \ldots\ldots, P_m$, such that

$$E = t_2P_2 + \ldots\ldots + t_mP_m.$$

It follows that $A = P_1 + (1 - \delta)t_2P_2 + \ldots\ldots + (1 - \delta)t_mP_m$;

where P_i are permutation matrices, and their coefficients are non{negative and sum up to 1.

Sucient Part : Let A be a convex combination of permutation matrices $P_1, P_2, P_3, \ldots\ldots, P_m$ i.e.,

$$A = t_1P_1 + t_2P_2 + \ldots\ldots + t_mP_m$$

where $t_1, t_2, \ldots\ldots, t_m$ are non negative numbers of a sum equal to 1. Then it is easy to see that $e^TA = e^T$ and $Ae = e$, where $e = (1, \ldots\ldots, 1)^T$. By equation (1.18) A is doubly stochastic.

5.3 Positive Semi-definite matrices

5.3.1 Positive Semi-definite matrices

Definition 5.3.1.1. An n-square complex matrix A is said to be positive semi definite or nonnegative definite, written as $A > 0$, if $A = A$ and

(2.1) $$x^*Ax \geq 0, \forall x \in \mathbb{C}^n.$$

A is further called positive definite, symbolized $A > 0$, if the strict in equality in (2.1) holds $\forall$ x(= 0) 2 C^n. It is immediate that if A is an $n \times n$ complex matrix, then

(2.2) $$A > 0 \Leftrightarrow x^*AX \geq 0$$

for every $n \times m$ complex matrix X. (Note that one may augment a vector by zero entries to get a matrix of size $n \times m$.) The spectral decomposition theorem of positive semi-definite matrices best characterizes positive semi-deniteness under unitary similarity.

Theorem 5.3.1.1. An $n \times n$ complex matrix A is positive semi definite if and only if there exists an $n \times n$ unitary matrix U such that

(2.3) $$A = U^*diag(\lambda_1, \ldots\ldots, \lambda_n)U.$$

where the i are the eigenvalues of A and a re all nonnegative. In addition, if $A > 0$ then det $A > 0$. A is positive definite if and only if all the i in (2.3) are positive. Besides, if $A > 0$ then det $A > 0$.

A principal minor is the determinant of a submatrix indexed by the same rows and columns, called a principal submatrix. Positives emidefinite matrices have many interesting and important properties and play a central role in matrix theory.

Theorem 5.3.1.2. Let A bean n-square real symmetric matrix. Then

(i) A is positive definite if and only if the determinant of every leading principal submatrix (leading minor) of A is positive.

(ii) A is positive semi definite if and only if the determinant of every (not just leading) principal submatrix of A is nonnegative.

Proof. Let A_k be a k-square principal submatrix of $A \in M_n$. By permuting rows and columns we may place A_k in the upper-left corner of A. In other words, there exists a permutation matrix P such that A_k is the (1, 1)-block of $P^T AP$. If $A \geq 0$, then (2.1) holds. Thus, for any $x \in R^k$,

$$x^* A_k x = y^* Ay \geq 0, \text{ where } y = P\begin{pmatrix} x \\ 0 \end{pmatrix} \geq R^k.$$

This says that A_k is positive semi definite. Therefore, det $A_k > 0$. The strict in equalities hold for positive definite matrix A.

Converse : It is easy to verify that a minor of a matrix $A \in M_n$ is the determinant of a subsquare matrix of A, then

(2.4) $$\det(\lambda I - A) = \lambda^n - \delta, \lambda^{n-1} + \delta_2 \lambda^{n-2} - \ldots + (-1)^n \det A.$$

where δ_i is the sum of all principal minors of order i, i = 1, 2, 3, ..., n – 1.

In anticipation to the for going fact, every principal submatrix of A has a non-negative determinant, then the polynomial in is given by (2.4), containing no negative zeroes since each δ is non-negative. The case where A is positive definite follows similarly.

As a side product of the proof, we see that A is positive (semi) definite if and only if all of its principal submatrices are positive(semi) definite. It is immediate that $A \geq 0 \Rightarrow a_{ii} \geq 0$ and that $a_{ii}a_{jj} \geq \|a_{ij}\|^2$ for $i = j$ by considering 2-square principal submatrices $\begin{pmatrix} a_{ii} & a_{ij} \\ a_{ji} & a_{jj} \end{pmatrix}$ Thus, if some diagonal entry $a_{ii} = 0$, then $a_{ij} = 0$ for all j, and hence, $a_{hi} = 0$ for all h, in as much as A is Hermitian. We conclude that some diagonal entry $a_{ii} = 0$ if and only if the row and the column containing $a_{ii} = 0$ consist entirely of 0.

Theorem 5.3.1.3. Prove *TFAE* for $A \in M_n(R)$

(i) A is positive semidefinite.

(ii) $A = B^TB$ for some matrix B.

(iii) $A = C^TC$ for some upper-triangular matrix C.

(iv) $A = D^TD$ for some upper-triangular matrix D with non negative diagonal entries (Cholesky factorization).

(v) $A = E^T \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix}$ for some $n \times n$ invertible matrix E and $r \times n$ matrix F, where r is the rank of A (Rank factorization).

Proof. Exercise

Exercise 5.3.1.1. Show that if A is a positive semidefinite matrix, then so are the A, A^T, adj (A) and A^{-1}, if the inverse exists.

Exercise 5.3.1.2. Let A be a positive semi definite matrix. Show that $tr\ A \geq 0$. Equality holds if and only if $A = 0$.

Exercise 5.3.1.3. Let $A \in 2\ M_n$ be positive semi-definite. Show that $(\det A)^{\frac{1}{n}} \leq \frac{1}{n} trA$.

Exercise 5.3.1.4. If $A > 0$ then the Cholesky factorization of A is unique.

Exercise 5.3.1.5. Find a Hermitian matrix A such that the leading minors are all non-negative, but A is not positive semidfienite.

5.3.2 A pair of positive semi-definite matrices

Let A and B be two Hermitian matrices of the same size. If $A \setminus B$ is positive semidefinite, we write $A \geq B$ or $B \geq A$.

It is easy to see that $\geq$ is a partial ordering, referred to as L owner (partial) ordering, on the set of Hermitian matrices, that is,

(i) $A \geq A$ for every Hermitian matrix A.

(ii) If $A \geq B$ and $B \geq A$, then $A = B$.

(iii) If $A \geq B$ and $B \geq C$, then $A \geq C$.

Obviously, $A + B \geq B$ if $A \geq 0$. That in (2.2) of the previous section immediately generalizes as follows.

(2.5) $A \geq 0 \Leftrightarrow X^* AX \geq 0$ for every complex matrix X of appropriate size. If A and B are both positive semidefinite, then $(A^{\frac{1}{2}}) = A^{\frac{1}{2}}$ and thus $A^{\frac{1}{2}} B A^{\frac{1}{2}} \geq 0$

Theorem 5.3.2.1. Let $A > 0$ and $B > 0$ be of the same size. Then

(i) The trace of the product AB is less than or equal to the product of the traces $tr\ A$ and $tr\ B$ i.e., $tr\ (AB) \geq tr\ A\ tr\ B$.

(ii) The eigenvalues of AB are all non negative. Furthermore, AB is positive semidefinite if and only if $AB = BA$.

(iii) If, are the largest eigenvalues of A, B, respectively, then

$$-\frac{1}{4}\alpha\beta \leq AB + BA \leq 2\alpha\beta I.$$

Proof. (i) By unitary similarity, with $A = U^* DU$,

$\text{tr}(AB) = tr\ (U^*DUB) = tr\ (DUBU^*)$,

we may assume that $A = \text{diag}(\lambda_1, \ldots\ldots, \lambda_n)$. Suppose that $b_{11}, \ldots\ldots, b_{nn}$ are the diagonal entries of B. Then

$$\begin{aligned}\text{tr}(AB) &= \lambda_1 b_{11} + \ldots. + \lambda_n b_{nn} \\ &\leq (\lambda_1 + \ldots\ldots + \lambda_n)(b_{11} + \ldots\ldots + b_{nn}) \\ &= tr_A\ tr_B.\end{aligned}$$

(ii) We known that XY and $Y\ X$ have the same eigenvalues if X and Y are square matrices of the same size. Thus, $AB = A^{\frac{1}{2}}(A^{\frac{1}{2}}B)$ has the same eigenvalues as $A^{\frac{1}{2}}B\,A^{\frac{1}{2}}$ which is positive semidefinite. AB is not positive semidefinite in general, since it need not be Hermitian. If A and B commute, however, then AB is Hermitian, for

$$(AB)^* = B^*A^* = BA = AB$$

and thus $AB \geq 0$. Conversely, if $AB \geq 0$, then it is Hermitian, and

$$AB = (AB)^* = B^*A^* = BA:$$

(iii) We assume that $A \neq 0$ and $B \neq 0$. Dividing through the inequalities by, $\alpha\beta$ we see that the statement is equivalent to its case $\alpha = 1$; $\beta = 1$. Thus, its suffices to show that $-\frac{1}{4}I \leq AB + BA \leq 2I$.

It is to be noted that $0 \leq A \leq I \Rightarrow 0 \leq A^2 \leq A \leq I$. It follows that

$$\begin{aligned}0 &= (A + B - \frac{1}{2}I) \\ &= (A+B)^2 - (A+B) + \frac{1}{4}I \\ &= A^2 + B^2 + AB + BA - A - B + \frac{1}{4}I\end{aligned}$$

$$\leq AB + BA + \frac{1}{4}I$$

that is, $AB + BA > -\frac{1}{4}I$. To show $AB + BA \leq 2I$, we compute

$0 \leq (A - B)^2 = A^2 + B^2 - AB - BA \leq 2I \leq AB - BA$:

Theorem 5.3.2.2. Let A and B be n-square positive semi-denite matrices. Then there exists an invertible matrix P such that P^*AP and P^*BP are both diagonal matrices. In addition, if A is nonsingular, then P can be chosen so that $P^*AP = I$ and P^*BP is diagonal.

Proof. Let rank $(A + B) = r$ and S be an on singular matrix so that

$$S^*(A + B)S = \begin{pmatrix} I, & 0 \\ 0 & 0 \end{pmatrix}$$

Conformally partition S*BS as

$$S^* BS = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$$

By (2.5),we have $S^*(A + B)S \geq S^*BS$. This implies

$B_{22} = 0$, $B_{12} = 0_1$ $B_{21} = 0$.

Now for B_{11}, because $B_{11} > 0$, there exists an r-square unitary matrix T such that TB11T is diagonal. Let us put

$$P = S\begin{pmatrix} T & 0 \\ 0 & I.... \end{pmatrix}$$

Then PBP and PBP = P(A+B)P – PBP are both diagonal. If A is invertible, we write A = CC for some matrix C. Consider matrix $(C^{-1})BC^{-1}$. Since it is positive semi-denite,we have a unitary matrix U such that

$$(C^{-1})^* BC^{-1} = UDU^*.$$

where D is a diagonal matrix with nonnegative diagonal entries. Let $P = C^{-1}U$. Then $PAP = I$ and $PBP = D$.

Many results can be derived by reduction of positive semi-definite matrices A and B to diagonal matrices, or further to nonnegative numbers, to which some elementary in equalities may apply. The following two are immediate from the previous theorem by writing $A = P^*D_1P$ and $B = PD_2P$, where P is an invertible matrix, and D_1 and D_2 are diagonal matrices with nonnegative entries.

Theorem 5.3.2.3. Let A > 0, B > 0 be of the same order (> 1). Then
det(A + B) ≥ det A + det B
with equality if and only if A + B is singular or A = 0 or B = 0, and

$$(A+B)^{-1} \leq \frac{1}{4}(A^{-1}+B^{-1})$$

if A and B are nonsingular, with equality if and only if A = B.

Theorem 5.3.2.4. If A > 0, B > 0, then
(i) rank A > rank B,
(ii) det$A$ > detB,
(iii) B^{-1} ≥ if A and B are nonsingular.

Every positive semi-definite matrix has a positive semi-definite square root. The square root is a matrix monotone function for positive semi-definite matrices in the sense that the Lnowner partial ordering is preserved when taking the square root.

Theorem 5.3.2.5. Let A and B be positive semi-definite matrices.Then

$$A \geq B \Rightarrow A^{\frac{1}{2}} \geq B^{\frac{1}{2}}$$

Proof. It may be assumed that A is positive definite by continuity. Let $C = A1$, $D = B1$ and E = C – D. We have to establish $E > 0$. For this purpose, it is sufficient to show that the eigenvalues of E are all nonnegative. It is to be noted that

$$0 > C^2 - D^2 = C^2 - (C-E)^2 = CE + EC - E^2.$$

It follows that CE + EC ≥ 0, for E is Hermitian and $E^2 > 0$. On the otherhand, let be an eigenvalue of E and let u be an eigen vector corresponding to λ. Then λ is real and by (2.1),

$$0 > u^*(CE+EC)u = 2\lambda(u^*Cu).$$

Since $C > 0$, we have λ > 0. Hence $E > 0$; namely, C > D.

Theorem 5.3.2.6. Let A and B be positive semi-definite matrices. Then

$$A > B \Rightarrow A^* > B^r, 0 > r > 1.$$

Exercise 5.3.2.1 Give an example where $A > 0$ and $B > 0$ but AB is not Hermitian.

Exercise 5.3.2.2 Let A, B, C be three n-square positive semidefinite matrices. Give an example showing that there does not necessarily exist an invertible matrix P such that P^*AP, P^*BP, PCP are all diagonal.

Exercise 5.3.2.3 Show that for Hermitian matrices A and B of the same size, $A^2 + B^2 > AB + BA$.

Exercise 5.3.2.4 Let A and B be n-square real symmetric invertible matrices. Show that there exists a real n-square invertible matrix P such that P^TAP and P^TBP are both diagonal if and only if all the roots of $p(x) = \det(xA - B)$ and $q(x) = \det(xB - A)$ are real.

5.3.3 Square root of a positive semi-definite matrix

Every non negative number has a unique non negative square root. The analogous result for positive semi-definite matrices also holds.

Theorem 5.3.3.1. For every $A \geq 0$, there exists a unique $B \geq 0$ so that $B^2 = A$. Further more, B can be expressed as a polynomial in A.

Proof. We may view n-square matrices as linear operators on C^n. The spectral theorem [refer to Theorem 5.4.1.4] ensures the existence of orthonormal eigenvectors $u_1, u_2, ..., u_n$ belonging to the eigenvalues $\lambda_1, \lambda_2, ..., \lambda_n$ of A, respectively. Then $u_1, u_2, ..., u_n$ form an orthonormal basis for C^n and $A(u_i) = \lambda_i u_i$, $\lambda_i > 0$. Define a linear operator B by $B(u_i) =$ for $i = 1, 2, ..., n$. It is routine to check that $B_2(x) = A(x)$ and $< B(x), x > 0$ for all vectors x *i.e.*, $B^2 = A$ and $B > 0$. To show the uniqueness, suppose C is also a linear operator such that $C^2(x) = A(x)$ and

$$< C(x), x >=< x, C(x) >\geq 0$$

for all vectors x. If v is an eigenvector of C : $Cv = \mu v$, then $C^2v = \mu^2 v$, *i.e.;* μ^2 is an eigenvalue of A. Hence, the eigenvalues of C are the non negative square roots of the eigenvalues of A *i.e.*, $\sqrt{\lambda_1}, \sqrt{\lambda_2}, ..., \sqrt{\lambda_n}$.

Choose orthonormal eigenvectors $v_1, v_2, ..., v_n$ corresponding to the eigenvalues $\sqrt{\lambda_1}, \sqrt{\lambda_2}, ..., \sqrt{\lambda_n}$. of C, respectively. Then $v_1, v_2, ..., v_n$ form an ortho normal basis for C^n. Let $u_i = w_1, v_1 + ... + w_{ni}v_n$, $i = 1, 2, ..., n$. On one hand, $C^2(u_i) = A(u_i) = \lambda_i u_i = w_1\lambda_i v_1 + ... + w_{ni}\lambda_i v_n$, however, $C^2(u_i) = w_{1i}\lambda_1 v_1 + ... + w_{ni}\lambda_n v_n$. Because $v_1, v_2, ..., v_n$ are linearly independent, we have $wtii = w_{ti}\lambda_i$ for each t. It follows that $w_{ti}\sqrt{\lambda_1} = w_{ti}\sqrt{\lambda_1}, t = 1,2,....,n.$ Thus,

$$\begin{aligned} Cu &= C(w_1 v_1 + ... + w_{ni}v_n) \\ &= w_{1i}\sqrt{\lambda_i}v_1 + ... + w_{ni}\sqrt{\lambda_i}v_n \\ &= w_{1i}\sqrt{\lambda_i}v_1 + ... + w_{ni}\sqrt{\lambda_i}v_n \\ &= \sqrt{\lambda_i}u_i = B(u_i) \end{aligned}$$

As u_1, u_2, ..., u_n constitute a basis for C^n, we conclude $B = C$. To see that B is a polynomial of A, let p(x) be a polynomial, by interpolation, such that $p(\lambda_i) = \sqrt{\lambda_i}, i = 1,2,....,n$. Then it is easy to verify that p(A) = B.

Remark 5.3.3.1. (i) Such a matrix B is called the square root of A, denoted by $A1$

(ii) A*A is positive semi-definite for every complex matrix A and that the eigenvalues of (A*A) are the singular values of A.

Exercise 5.3.3.1. Let A > 0 and B > 0 be of the same size. Show that $BA^2B < I \Rightarrow B^1\ AB^1 > I$.

Exercise 5.3.3.2. Show by example that $A \geq B \geq 0 \Rightarrow A^2 \geq B^2$.

5.4 Symmetric matrices and quadratic forms

5.4.1 Diagonalization of symmetric matrices

Through out the section, unless otherwise mentioned, all matrices are real. A square matrix A is said to be diagonalizable if A is similar to a diagonal matrix, that is, if $A = PDP^{-1}$ for some invertible matrix P and some diagonal matrix D. The next theorem gives a characterization of diagonalizable matrices and tell show to construct a suitable factorization.

Theorem 5.4.1.1. [3] The Diagonalization Theorem : An n × n matrix A is diagonalizable if and only if A has n linearly independent eigen vectors.

In fact, $A = PDP^{-1}$, with D a diagonal matrix, if and only if the columns of P are n linearly independent eigenvectors of A. In this case, the diagonal entries of D are eigenvalues of A that correspond, respectively, to the eigenvectors in P.

A symmetric matrix is a matrix A such that $A^T = A_i$, A^T being the transpose of A. Such a matrix is necessarily square. Its main diagonal entries are arbitrary, but its other entries occur in pairs on opposite sides of the main diagonal.

Example 5.4.1.1. Lets consider the following matrices:

Symmetric : $\begin{pmatrix} 1 & 0 \\ 0 & -3 \end{pmatrix}, \begin{pmatrix} 0 & -1 & 0 \\ -1 & 5 & 8 \\ 0 & 8 & -7 \end{pmatrix}, \begin{pmatrix} a & b & c \\ b & d & e \\ c & e & f \end{pmatrix}$

Non-symmetric : $\begin{pmatrix} 1 & -3 \\ 3 & 0 \end{pmatrix}, \begin{pmatrix} 1 & -4 & 0 \\ -6 & 1 & -4 \\ 0 & -6 & 1 \end{pmatrix}, \begin{pmatrix} 5 & 4 & 3 & 2 \\ 4 & 3 & 2 & 1 \\ 3 & 2 & 1 & 0 \end{pmatrix}$

To begin the study of symmetric matrices, it is helpful to review the diagonalization process from [2], [3].

Problem 5.4.1.1. Diagonalize the matrix $A = \begin{pmatrix} 6 & -2 & -1 \\ -2 & 6 & -1 \\ -1 & -1 & 5 \end{pmatrix}$

Solution 5.4.1.1. The characteristic equation of A is

$$0 = -\lambda^3 + 17\lambda^2 - 00\lambda + 144 = -(\lambda - 8)(\lambda - 6)(\lambda - 3).$$

For,

$$\lambda = 8 : v_1 = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}; \lambda = 6; v_2 = \begin{pmatrix} -1 \\ -1 \\ 2 \end{pmatrix}; \lambda = 3 : v_3 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

So the normalized (unit) eigenvectors are

$$u_1 = \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{pmatrix}, u_2 = \begin{pmatrix} -\frac{1}{\sqrt{6}} \\ -\frac{1}{\sqrt{6}} \\ \frac{2}{6} \end{pmatrix}, u_3 = \begin{pmatrix} \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \end{pmatrix}.$$

Let P = $\begin{pmatrix} -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ 0 & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{3}} \end{pmatrix}$ D, $\begin{pmatrix} 8 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 3 \end{pmatrix}$ Then, A = PDP–1. Since P is square and has ortho normal columns, *P* is an orthogonal matrix, and P–1 = P^T. The following theorem explains why the eigenvectors in the above problem are orthogonal they correspond to distinct eigenvalues.

Theorem 5.4.1.2. If A is symmetric, then any two eigenvectors from different eigen spaces are orthogonal.

Proof. Let v1 and v2 be eigen vectors that correspond to distinct eigenvalues, say, 1 and 2 . It's sucess to show that $v_1 \quad v_2 = 0$. Now,

$$\lambda_1 v_1 . v_2 = (\lambda_1 v_1)^T v_2 = (Av_1)^T v_2 \quad \because v_1 \text{ is an eigen vector}$$

$$= (v_1^T A^T) v_2 = v_1^T (Av_2) \quad \because A^T = A$$

$$= v_1^T (\lambda_2 v_2) \quad \because \text{ is an eigen vector}$$

$$= \lambda_2 v_1^T v_2$$

$$= \lambda_2 v_1 . v_2$$

Hence, $(\lambda_1 - \lambda_2) v_1 . v_2 = 0 \Rightarrow v_1 . v_2 = 0, \quad \because \lambda_1 - \lambda_2 \neq 0 \Rightarrow \lambda_1 \neq \lambda_2.$

Definition 5.4.1.1. An n n matrix A is said to be orthogonally diagonalizable if 9 an orthogonal matrix *P*(with $P^{-1} = P^T$) and a diagonal matrix D such that

(3.1) $A = PDP^{-1} = PDP^T$.

Such a diagonalization requires *n* linearly independent and ortho normal eigenvectors. When is this possible ? If *A* is orthogonally diagonalizable as in equation (3.1), then

$$A^T = (PDP^T)^T = P^{TT}D^TP^T = PDP^T = A.$$

Thus A is symmetric. The following the orem conversely states that every symmetric matrix is orthogonally diagonalizable.

Theorem 5.4.1.3. An n × n matrix A is orthogonally diagonalizable if and only if A is a symmetric matrix.

Problem 5.4.1.2. Orthogonally diagonalize the matrix $A = \begin{pmatrix} 3 & -2 & 4 \\ -2 & 6 & 2 \\ 4 & 2 & 3 \end{pmatrix}$ whose characteristic equation is

$0 = -\lambda^3 + 12\lambda^2 - 21\lambda - 98 = -(\lambda - 7)^2(\lambda + 2)$.

Solution 5.4.1.2. For,

$$\lambda = 7 : v_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, v_1 = \begin{pmatrix} -y_2 \\ 1 \\ 0 \end{pmatrix}; \lambda = 2 : v_3 = \begin{pmatrix} -1 \\ -y_2 \\ 1 \end{pmatrix}$$

Although v_1 and v_2 are linearly independent, therefore they are not orthogonal. Recall from [3] that the projection of v_2 on to v_1 is $\frac{v_2.v_1}{v_1 - v_2}$, and the component of v_2 orthogonal to v_1 is

$$z_2 = v_2 - \frac{v_2.v_1}{v_1.v_1} = \begin{pmatrix} -\frac{1}{2} \\ 1 \\ \frac{1}{4} \end{pmatrix}$$

Then (v_1, z_2) is an orthogonal set in the eigen space for = 7. Here, z_2 is a linear combination of the eigen vectors v_1 and v_2, so z_2 is in the eigen space. This construction of z_2 is just the Gram-Schmidt process (refer to [3]) Since the eigenspace is two-dimensional (with basis v_1, v_2 the orthogonal set (v_1, z_2) is an orthogonal basis for the eigenspace, by the Basis Theorem. (refer to [3]).

Normalize (v_1, z_2) to obtain the following ortho normal basis for the eigens pace for $\lambda = 7$.

$$u_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ 0 \\ \frac{1}{\sqrt{2}} \end{pmatrix}, u_2 = \begin{pmatrix} -\frac{1}{\sqrt{18}} \\ \frac{4}{\sqrt{18}} \\ \frac{1}{\sqrt{18}} \end{pmatrix}.$$

An ortho normal basis for the eigenspace for $\lambda = -2$ is

$$u_s = \frac{1}{2\| v_s \|} 2v_s = \frac{1}{3}\begin{pmatrix} -2 \\ -1 \\ 2 \end{pmatrix} = \begin{pmatrix} -\frac{2}{3} \\ -\frac{1}{3} \\ \frac{2}{3} \end{pmatrix}.$$

By Theorem 3.1.1, u_3 is orthogonal to the other eigenvectors u_1 and u_2 . Hence $\{u_1, u_2, u_3\}$ is an orthonormal set. Let

$$P = (u_1 \ u_2 \ u_3) = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{15}} & -\frac{2}{3} \\ u & -\frac{4}{\sqrt{18}} & -\frac{1}{3} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{12}} & \frac{2}{3} \end{pmatrix}, D = \begin{pmatrix} 7 & 0 & 0 \\ 0 & 7 & 0 \\ 0 & 0 & -2 \end{pmatrix}$$

Then P orthogonally diagonalizes A, and $A = PDP^{-1}$.

In the above example, the eigenvalue 7 has multiplicity two and the eigenspace is two-dimensional.

The Spectral Theorem

The set of eigen values of a matrix A is sometimes called the spectrum of A, and the following description of the eigenvalues is called a spectral theorem.

Theorem 5.4.1.4. The Spectral Theorem for Symmetric matrices : An $n \times n$ matrix A has the following properties:

(i) A has n real eigenvalues, counting multiplicities.

(ii) The dimension of the eigenspace for each eigenvalue equals the multiplicity of as a root of the characteristic equation.

(iii) The eigenspaces are mutually orthogonal, in the sense that eigen vectors corresponding to different eigen- values are orthogonal.

(iv) A is orthogonally diagonalizable.

Proof. The proof of the theorem follows from [3].

The Spectral Decomposition

Suppose $A = PDP^{-1}$, where the columns of P are ortho normal eigenvectors u_1, u_2, ..., u_n of A and the corresponding eigenvalues λ_1, λ_2, ..., λ_n are in the diagonal matrix D. Then, since $P-1 = P^T$, we obtain

$$(3.2)\quad A = PDP^{-1} = (u_1, u_2, ..., u_n)\begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}\begin{pmatrix} u_1^T \\ \vdots \\ u_n^T \end{pmatrix}$$

$$(3.3)\quad = (\lambda_1 c, ... \lambda_n u_n)\begin{pmatrix} u_1^T \\ \vdots \\ u_n^T \end{pmatrix}.$$

Hence, $A = \lambda_1 u_1 u^T_1 + \lambda_2 u_2 u^T_2 + ... + \lambda_n u_n u^T_n$.

This representation of A is called a spectral decomposition of A because it breaks up A into pieces determined by the spectrum (eigenvalues) of A. Each termin equation (3.2) is an n × n matrix of rank 1.

Problem 5.4.1.3. Construct a spectral decomposition of the matrix *A* that has the orthogonal diagonalization.

$$A = \begin{pmatrix} 7 & 2 \\ 2 & 4 \end{pmatrix} = \begin{pmatrix} \frac{2}{\sqrt{5}} & -\frac{1}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \end{pmatrix}\begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}\begin{pmatrix} \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \\ -\frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \end{pmatrix}$$

Solution 3.1.3. Denote the columns of P by u1 and u2 . Then,

$A = 8u_1u^T_1 + 3u_2u^T_2$.

To verify this decomposition of A, compute

$$u_1 u_1^T = \begin{pmatrix} \frac{2}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} \end{pmatrix}\begin{pmatrix} \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \end{pmatrix} = \begin{pmatrix} \frac{4}{5} & \frac{2}{5} \\ \frac{2}{5} & \frac{1}{5} \end{pmatrix}$$

$$u_1 u_2^T = \begin{pmatrix} -\frac{1}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} \end{pmatrix}\begin{pmatrix} -\frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \end{pmatrix} = \begin{pmatrix} \frac{1}{5} & -\frac{2}{5} \\ -\frac{2}{5} & \frac{4}{5} \end{pmatrix} and$$

$$8u_1u_1^T + 3u_2u_2^T = \begin{pmatrix} \frac{32}{5} & \frac{16}{5} \\ \frac{16}{5} & \frac{8}{5} \end{pmatrix} + \begin{pmatrix} \frac{3}{5} & -\frac{6}{5} \\ -\frac{6}{5} & \frac{12}{5} \end{pmatrix} = \begin{pmatrix} 7 & 2 \\ 2 & 4 \end{pmatrix} = A$$

5.4.2 Quadratic Forms

Quadratic forms, occur frequently in applications of linear algebra to engineering (indesign criteria and opti- mization) and signal processing (as output noise power).

They also arise, for example, in physics (as potential and kinetic energy), differential geometry (as normal curvature of surfaces), economics (as utility functions), and statistics (in confidence ellipsoids). Some of the mathematical background for such applications flows easily from our work on symmetric matrices.

Definition 5.4.2.1. A Quadratic form on R n is a function Q defined on R^n whose value at a vector x in R^n can be computed by an expression of the form $Q(x) = x^TAx$, where A is an n × n symmetric matrix. The matrix A is called the matrix of the quadratic form. The simplest example of a non zero quadratic form is $Q(x) = x^TIx = \|x\|^2$, I being the identity matrix. The following examples show the connection between any symmetric matrix A and the quadratic form x^TAx.

Example 5.4.2.1. Let $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ Then, x^TAx for the matrices $A = \begin{pmatrix} 4 & 0 \\ 0 & 3 \end{pmatrix}$ and A $= \begin{pmatrix} 3 & -2 \\ -2 & 7 \end{pmatrix}$ are $4x^2_1 + 3x^2_2$ and $3x^2_1 - 4x_1x_2 + 7x^2_2$ respectively.

Example 5.4.2.2. For $x \in R^3$, let $Q(x) = 5x^2_1 + 3x^2_2 + 2x^2_3 - x_1x_2 + 8x_2x_3$. Write this quadratic form as x^TAx.

The coefficients of x^2_1, x^2_2, x^2_3 on the diagonal of A. To make A symmetric, the coefficient of x_ix_j for $i = j$ must be split evenly between the (i, j) – and (j, i) –entries in A. The coefficient of x_1x_3 is 0. Thus,

$$Q(x) = x^TAx = (x_i\ x_2\ x_3)\begin{pmatrix} 5 & -\frac{1}{2} & 0 \\ -\frac{1}{2} & 3 & 4 \\ 0 & 4 & 2 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}.$$

Exercise 5.4.2.1. Let $Q(x) = x^2_1 - 8x_1x_2 - 5x^2_2$. Compute the value of $Q(x)$ for $x = \begin{pmatrix} -3 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ -2 \end{pmatrix}$ *and* $\begin{pmatrix} 1 \\ -3 \end{pmatrix}$.

Remark 5.4.2.1. In some cases, quadratic forms are easier to use when they have no cross-product terms— that is, when the matrix of the quadratic form is a diagonal matrix. Fortunately, the cross-product term can be eliminated by making a suitable change of variable.

Change of Variable in a Quadratic Form

If $x \in R^n$ represents a variable vector in, then a change of variable is an equation of the form

(3.4) $x = Py$ or equivalently $y = P^{-1}x$

where P is an invertible matrix and y is a new variable vector in R^n. Here y is the coordinate vector of x relative to the basis of R^n determined by the columns of P. If the change of variable equation (3.4) is made in a quadratic form x^TAx, then

(3.5) $x^TAx = (Py)^TA(Py) = y^TP^TAPy = y^T(P^TAP)y;$

and the new matrix of the quadratic formis P^TAP. Since A is symmetric, Theorem 3.1.2 guarantees that there is an orthogonal matrix P such that P^TAP is a diagonal matrix D, and the quadratic form in equation (3.5) becomes y^TDy. This is the strategy of the next theorem.

Theorem 5.4.2.1. The Principal Axes Theorem : Let A be an n × n symmetric matrix. Then there exists an orthogonal change of variable, $x = Py$, that transforms the quadratic form x^TAx into a quadratic form y^TDy with no cross-product term.

Remark 5.4.2.2. The columns of P in the theorem are called the principal axes of the quadratic form x^TAx. The vector y is the coordinate vector of x relative to the ortho normal basis of R^n given by these principal axes.

Example 5.4.2.3. Make a change of variable that transforms the quadratic form in Example 5.4.2.1 into a quadratic form with no cross-product term.

The matrix of the quadratic form in Exercise 5.4.2.1 is $A = \begin{pmatrix} 1 & -4 \\ -4 & 5 \end{pmatrix}$ The first step is to orthogonally diagonalize A. Its eigenvalues turn out to be $\lambda = 3$ and $\lambda = 7$. Associated unit eigenvectors are

$$\lambda = 3 \; ; \begin{pmatrix} \frac{2}{\sqrt{5}} \\ -\frac{1}{\sqrt{5}} \end{pmatrix}; \lambda = 7; \begin{pmatrix} \frac{1}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} \end{pmatrix}$$

These vectors are automatically orthogonal (because they correspond to distinct eigenvalues) and so provide an orthonormal basis for R^2. Let

$$P = \begin{pmatrix} \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \\ -\frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \end{pmatrix}, \; D = \begin{pmatrix} 3 & 0 \\ 0 & -7 \end{pmatrix}$$

Then, $A = PDP^{-1}$ and $D = P^{-1}AP$. A suitable change of variable is

$$x = Py, \; x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}.$$

Then, $x_1^2 - 8x_ix_2 - 5x_2^2 = x^TAx = (Py)^TA(Py) = y^TP^TAPy = y^TDy$

$$= 3y_i^2 - 7y_i^2 .$$

Remark 5.4.2.3. Example 5.4.2.3 illustrates the Theorem 5.4.2.1.

To illustrate the meaning of the equality of quadratic forms in Example 3.2.3, we can compute $Q(x)$ for $x = \begin{pmatrix} 2 \\ -2 \end{pmatrix}$ using the new quadratic form. Since $x = Py$, $y = P^{-1}x = P^Tx$, so

$$y = \begin{pmatrix} \frac{2}{\sqrt{5}} & -\frac{1}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \end{pmatrix} \begin{pmatrix} 2 \\ -2 \end{pmatrix} = \begin{pmatrix} \frac{6}{\sqrt{5}} \\ -\frac{2}{\sqrt{5}} \end{pmatrix}.$$

Hence, $3y^2_1 - 7y^2_2 = 16$.

The geometrical interpretation of Example 5.4.2.3 is illustrated from Figure 1.

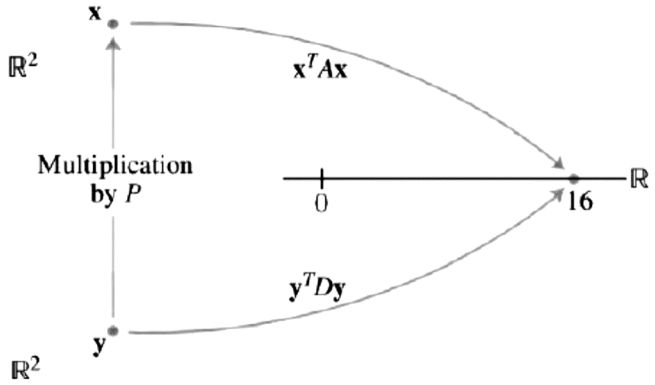


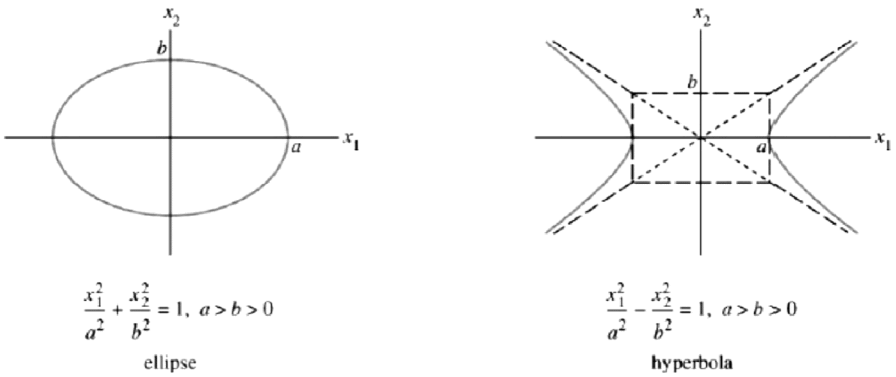
Figure 1 : Change of variable in x^TAx

A Geometric View of Principal Axes

Suppose $Q(x) = x^TAx$, where A is an invertible 2×2 symmetric matrix, and let c be a constant. It can be shown that the set of all $x \in R^2$ that satisfy

(3.6) $x^TAx = c$;

either corresponds to an ellipse (or circle), a hyperbola, two inter secting lines, or a single point, or contains no points at all. If A is a diagonal matrix, the graph is in standard position, such as in Figure 2. If A is not a diagonal matrix, the graph of (3.6) is rotated out of standard position, as in Figure 3. Finding



$\frac{x_1^2}{a^2} + \frac{x_2^2}{b^2} = 1,\ a > b > 0$ ellipse

$\frac{x_1^2}{a^2} - \frac{x_2^2}{b^2} = 1,\ a > b > 0$ hyperbola

Figure 2 : An ellipse and a hyperbola in standard position

the principal axes (determined by the eigenvectors of A) amounts to finding a new coordinate system with respect to which the graph is in standard position.

The hyp here *A* is

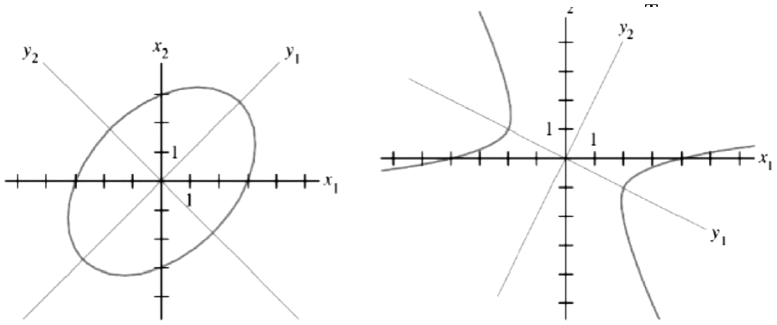


Figure 3 : An ellipse and a hyperbola not in standard position

the matrix in Example 5.4.2.3. The positive y_1-axis in Figure 3(b) is in the direction of the first column of the matrix *P* in Example 5.4.2.3, and the positive y_2-axis is in the direction of the second column of P.

Exercise 5.4.2.2. Find a change of variable that removes the cross-product term from the equation of the ellipse in Figure 3(a).

Classifying QuadraticForms

When A is an n n matrix, thequadraticform $Q(x) = x^TAx$ is a real-valued function with domain R^n. Figure 4 displays the graphs of four quadratic forms with domain R^n. For each point $x = (x_1, x_2)$ in the domain of a quadratic form *Q*, the graph displays the point $((x_1, x_2)\ z)$ where $z = Q(x)$. Notice that except at $x = 0$, the values of $Q(x)$ are all positive in Figure 4(a) and all negative in Figure 4(d). The horizontal cross-sections of the graphs are ellipses in Figures 4(a) and 4(d) and hyperbola sin Figure 4(c).

Definition 5.4.2.2. A quadratic form Q is :

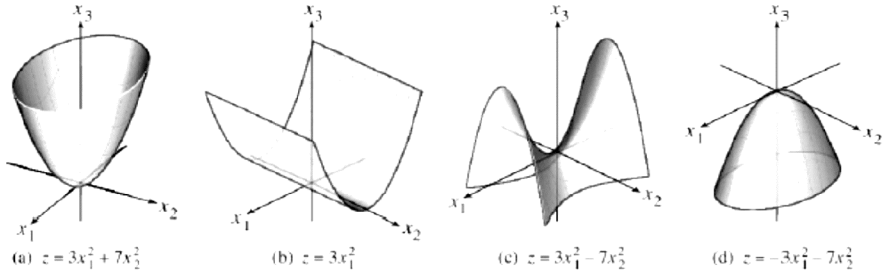


Figure 4 : Graph of quadratic forms

- positive definite if $Q(x) > 0, \forall x \; 6 = 0$
- negative definite if $Q(x) < 0, \forall x \; 6 = 0$
- indenite if $Q(x)$ assumes both positive and negative values.

Also, Q is said to be positive semi definite if $Q(x) \geq 0, \forall x$, and to be negative semidefinite if $Q(x) \leq 0$; $\forall$ x. The quadratic forms in Figure 4(a) and 4(b) are both positive semi definite. Theorem 5.4.2.2 characterizes some quadratic forms interms of eigenvalues. The classification of a quadratic form is often carried over to the matrix of the form. Thus a positive definite matrix A is a symmetric matrix for which the quadratic form x^TAx is positive definite. Other terms, such as positive semi definite matrix, are defined analogously.

5.4.3 Constrained motion

Engineers, economists, scientists, and mathematicians of ten need to find the maximum or minimum value of a quadratic form Q(x) for x in some specified set. Typically, the problem can be arranged so that x varies over the set of unit vectors. This constrained optimization problem has an interesting and elegant solution. In the foregoing examples and the discussion in foregoing sections will illustrate how such problems arise in practice. The requirement that a vector x in R^n be a unit vector can be stated in several equivalent ways :

$$\|x\| = 1, \|x\|^2 = 1, x^Tx = 1$$

and

(3.8) $$x^2_1 + x^2_2 + x^2_3 + ... + x^2_n = 1.$$

The expanded version equation (3.8) of $x^Tx = 1$ is commonly used in applications. When aquadratic form Q has no cross product terms, it is easy to find the maximum and minimum of $Q(x)$ for $x^Tx = 1$.

e.g. Find the maximum and minimum of $Q(x) = 9x^2_1 + 4x^2_2 + 3x^2_3$ subject to the constraint $x^Tx = 1$.

Since, x^2_2 and x^2_3 are non-negative,

$$4x_3^2 \leq 9x_3^2 \quad \textit{and} \quad 3x_3^2 \leq 9x_3^2$$

and hence, $Q(x) = 9x_1^2 + 4x_2^2 + 3x_3^2$

$$= 9x_1^2 + 9x_2^2 + 3x_3^2 = 9(x_1^2 + x_3^2 + x_3^2) = 9$$

whenever $x^2_1 + x^2_2 + x^2_3 + ... + x^2_n = 1$. So the maximum value of $Q(x)$ cannot exceed 9 when x is a unit vector. Further more, $Q(x) = 9$ when $x = (1, 0, 0)$. Thus 9 is the maximum value of $Q(x)$ for $x^Tx = 1$. To find the minimum value of $Q(x)$, observe that

$$= 9x_1^2 \geq 9x_1^2 \quad \textit{and} \quad 4x_2^2 \geq 3x_2^2$$

and hence, $Q(x) \geq 3x^2_1 + 3x^2_2 + 3x^2_3 = 3(x^2_1 + x^2_2 + x^2_3)$ whenever, $x^2_1 + x^2_2 + x^2_3 + ... + x^2_n = 1$. Also, $Q(x) = 3$ when $x_1 = x_2 = 0$, $x_3 = 1$. So, 3 is the minimum value of $Q(x)$ for $x^Tx = 1$.

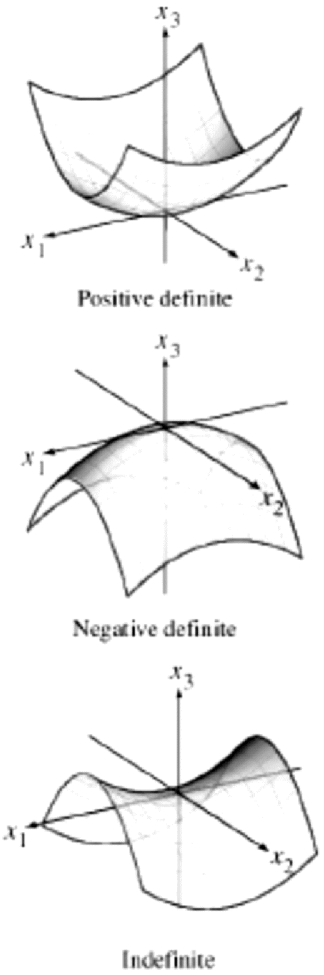


Figure 5

Theorem 5.4.3.1. Quadratic Forms and Eigenvalues : Let A bean $n \times n$ symmetric matrix. Then aquadratic form x^TAx is :

- positive definite if and only if the eigenvalues of A are all positive,
- negative definite if and only if the eigenvalues of A are all negative,
- indefinite if and only if A has both positive and negative eigenvalues.

Proof. By the Principal Axes Theorem, there exists an orthogonal change of variable $x = Py$ such that

(3.7) $Q(x) = x^TAx = y^TDy = \lambda_1 y^2_1 + \lambda_2 y^2_2 + ... + \lambda_n y^2_n$.

where $\lambda_1, \lambda_2, ..., \lambda_n$ are the eigenvalues of A. Since P is invertible, $\exists$ a one-to-one correspondence between all nonzero x and allnonzero y. Thus the values of $Q(x)$ for $x = 0$ coincide with the values of the expression on the right side of (3.7), which is obviously controlled by the signs of the eigenvalues $\lambda_1, \lambda_2, ..., \lambda_n$, in the three ways described in the theorem.

Problem 5.4.3.1. Is $Q(x) = 3x^2_1 + 2x^2_2 + x^2_3 + 4x_1x_2 + 4x_2x_3$ positive definite ?

Solution 5.4.3.1. Because of all the plus signs, this form "looks" positive definite. But the matrix of the form is

$$\begin{pmatrix} 3 & 2 & 0 \\ 2 & 2 & 2 \\ 0 & 2 & 1 \end{pmatrix}$$

and the eigenvalues of A turn out to be 5, 2, and –1. So Q is an indefinite quadratic form, not positive definite. The classification of a quadratic form is often carried over to the matrix of the form. Thus a positive definite matrix A is a symmetric matrix for which the quadratic form x^TAx is positive definite. Other terms, such as positive semidefinite matrix, are defined an alogously.

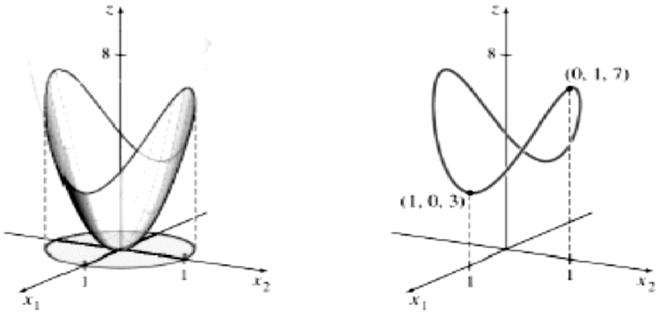
Example 5.4.3.1. Let $A = \begin{pmatrix} 3 & 0 \\ 0 & 7 \end{pmatrix}$ and let $Q(x) = x^TAx$ for $x \in R^2$. Figure 6(1) displays the graph of Q. Figure 6(2) shows only the portion of the graph inside a cylinder; the intersection of the cylinder with the surface is the set of points (x_1, x_2, z) such that $z = Q(x_1, x_2)$ and $x^2_1 + x^2_2 = 1$. The "heights" of these points are the constrained values of $Q(x)$. Geometrically, the constrained optimization problem is to locate the highest and lowest points on the intersection curve. The two highest points on the curve are 7 units above the x_1x_2-plane, occurring where $x_1 = 0$ and $x_2 = \pm 1$. These points correspond to the eigenvalue 7 of A and the eigenvectors $x = (0, 1)$ and $-x = (0,1)$. Similarly, the two lowest points on the curve are 3 units above the x_1x_2-plane. They correspond to the eigenvalue 3 and the eigenvectors (1, 0) and (–1, 0). Every point on the intersection curve in Figure 6(2) has z-coordinate between 3 and 7, and for any number t between 3 and 7, there is a unit vector x such that $Q(x) = t$. In other words, the set of all possible values of x^TAx, for $\|x\| = 1$, is the closed interval $3 \leq t \leq 7$. It can be shown that for any symmetric matrix A, the set of all possible values of x^TAx, for $\|x\| = 1$, is a closed interval on the real axis. Denote the left and right endpoints of this interval by m and M, respectively. That is, let

(3.9) $m = min\{xTAx : \|x\| = 1\}$, $M = max\{x^TAx : \|x\| = 1\}$

Remark 5.4.3.1. The use of minimum and maximum in (3.9), and least and greatest in the theorem, refers to the natural ordering of the real numbers, not to magnitudes.

Theorem 5.4.3.2. Let A be a symmetric matrix, and define m and M as in (3.9). Then M is the greatest eigenvalue λ_1 of A and m is the least eigenvalue of A. The value

of x^TAx is M when x is a unit eigenvector u_1 corresponding to M. The value of x^TAx is m when x is a unit eigenvector corresponding to m.



1 $z = 3x_1^2 + 7x_2^2$.

2 The intersection of $z = 3x_1^2 + 7x_2^2$ and the cylinder $x_1^2 + x_2^2 = 1$.

Figure 6

Proof. Let us orthogonally diagonalize A as PDP^{-1}. We know that

(3.10) $x^TAx = y^TDy$, when $x = Py$.

Also, since, $P^TP = I$ and $\|Py\|^2 = (Py)^T(Py) = y^TP^TPy = y^Ty = \|y\|^2$, therefore,

$$\|x\| = \|Py\| = \|y\| \text{ holds } \forall\, y.$$

In particular, $\|y\| = 1 \Leftrightarrow \|x\| = 1$. Thus, x^TAx and y^TDy assume the same set of values as x and y range over these to fall unit vectors. Suppose that A is a 3×3 matrix with eigenvalues $a > b > c$. Arrange the (eigenvector) columns of P so that $P = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}$ and $D = \begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix}$ Given any unit vector y in R^3 with coordinates y_1, y_2, y_3,

$$\begin{aligned} ay_1^2 &= ay_1^2 \\ by_2^2 &= ay_2^2 \\ cy_3^2 &= ay_3^2 \end{aligned}$$

and obtained

$$\begin{aligned} y^TDy &= ay_1^2 + by_2^2 + cy_3^2 \\ &\le ay_1^2 + ay_2^2 + ay_3^2 \\ &= a(y_1^2 + y_2^2 + y_3^2) \\ &= a\,\|y\|^2 = a. \end{aligned}$$

Thus $M < a$. However, when $y = e_1 = (1, 0, 0)$ then $y^T Dy = a$, so $M = a$ holds. By equation 3.10, x that corresponds to $y = e_1$ is the eigenvector u_1 of A, as

$$x = P_1 = [u_1, u_2, u_3] \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = u_1$$

Thus $M = a = e_1^T De_1 = u^T{}_1 A u_1$

which proves the statement about M. A similar argument shows that m is the least eigenvalue c, and this value of $x^T Ax$ is attained when $x = P_1 = u_3$.

Problem 5.4.3.2 Let $A = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 3 & 1 \\ 1 & 1 & 4 \end{pmatrix}$. Find the maximum value of the quadratic form $x^T Ax$ subject to the constraint $x^T Ax = 1$ and and a unit vector at which this maximum value is attained.

Solution 5.4.3.2 Using Theorem 5.4.3.2, the desired maximum value is the greatest eigenvalue of A. The characteristic equation turns out to be

$$0 = -\lambda^3 + 10\lambda^2 - 27\lambda + 18 = -(\lambda - 6)(\lambda - 3)(\lambda - 1).$$

The greatest eigenvalue is 6. The constrained maximum of $x^T Ax$ is attained when x is a unit eigen vector for = 6. Solve for (A – 6I) = O and find an eigenvector $\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$ and set $u_1 = \begin{pmatrix} \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \end{pmatrix}$

Theorem 5.4.3.3 Let A, λ_1, u_1 be as defined in Theorem 5.4.3.2. Then the maximum value of $x^T Ax$ subject to the constraints

$$x^T x = 1, \; x^T u_1 = 0$$

is the second greatest eigenvalue, λ_2, and this maximum is attained when x is an eigenvector u_2 corresponding to λ_2 .

Proof. Hint: It can be proved by an argument similar to forgoing theorem 3.3.1 in which the theorem reduces to the case where the matrix of the quadratic form is diagonal.

Example 5.4.3.2 Find the maximum value of $9x^2_1 + 4x^2_2 + 3x^2_3$ subject to the constraints $x^Tx = 1$, $x^Tu_1 = 0$; where $u_1 = (1, 0, 0)$

If $x = (x_1, x_2, x_3)$, then the constraint $x^Tu_1 = 0 \Rightarrow x_1 = 0$. For such a unit vector, $x^2_1 + x^2_2 = 1$ and

$$9x^2_1 + 4x^2_2 + 3x^2_3 = 4x^2_1 + 3x^2_2$$
$$\leq 4x^2_1 + 4x^2_2 = 4 :$$

Thus the constrained maximum of the quadratic form does not exceed 4. And this value is attained for $x = (1\ 0, 0)$, which is an eigenvector for the second greatest eigenvalue of the matrix of the quadratic form.

Exercise 5.4.3.1 Let A be the matrix in Problem 5.4.3.2 and let u_1 be a unit eigenvector corresponding to the greatest eigenvalue of A. Find the maximum value of x^TAx subject to the constraints

$$x^Tx = 1, x^Tu_1 = 0.$$

The next theorem generalizes the foregoing one and combining with Theorem (3.3.1), gives a signicant characterization of all the eigenvalues of A.

Theorem 5.4.3.4 Let A be a symmetric $n \times n$ matrix with an orthogonal diagonalisation $A = PDP^{-1}$, where the entries on the diagonal of D are arranged so that $\lambda_1 > \lambda_2 \geq .... \geq \lambda_n$ and where the columns of P are corresponding unit eigenvectors $u_1, u_2, ..., u_n$. Then for $k = 2, 3, 4, ..., n$, the maximum value of $x^Tx = 1$, $x^Tu_1 = 0$, $x^Tu_{k-1} = 0$ is the eigenvalue k and this maximum value is attained at $x = u_k$.

The proof is beyond the scope of the book.

5.4.4 The singular value Decomposition

Unfortunately, as we know, not all matrices can be factored as $A = PDP^T$ with diagonal D. However, a factorization $A = QDP^T$ is possible for any $m \times n$ matrix A!A special factorization of this type, called the singular value decomposition, is one of the most useful matrix factorizations in applied linear algebra. The singular value decomposition is based on the following property of the ordinary diagonalization that can be imitated for rectangular matrices. The absolute values of the eigenvalues of a symmetric matrix A measure the amounts that A stretches or shrinks certain eigenvectors. If $Ax = \lambda x$ and $||x|| = 1$, then

(3.11) $$||Ax|| = ||\lambda x|| = |\lambda|\ ||x|| = |\lambda|$$

If λ_1 is the eigenvalue with the greatest magnitude, then a corresponding unit eigenvector v_1 identies a direction in which the stretching effect of A is greatest. That is, the length of Ax is maximized when $x = v_1$, and $||Av_1|| = |\lambda_1|$ by (3.11). This

description of v_1 and λ_1 has an analogue for rectangular matrices that will lead to the singular value decomposition.

Example 5.4.4.1. If $A = \begin{pmatrix} 4 & 11 & 14 \\ 8 & 7 & -2 \end{pmatrix}$ then the linear transformation $x \longrightarrow Ax$ maps the unit sphere $\{x : \|x\| = 1\}$ in R^3 on to the ellipse in R^2 (refer to Figure 7). Find a unit vector x at which the length $\|Ax\|$ is maximized, and compute this maximum length.

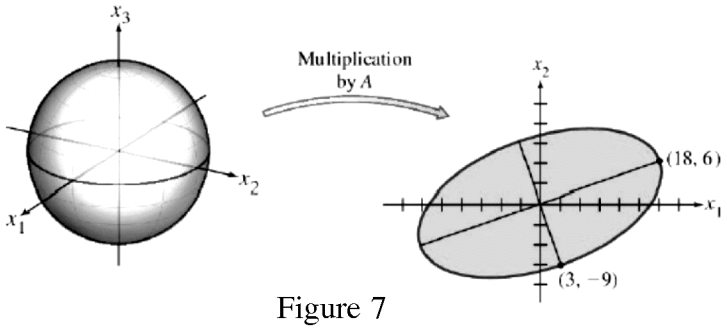


Figure 7

The quantity $\|Ax\|^2$ is maximized at the same x that maximizes $\|Ax\|$, and $\|Ax\|^2$ is easier to study. Here,

$$\|Ax\|^2 = (Ax)^T Ax = x^T A^T Ax = x^T (A^T A)x.$$

Also, A^TA is a symmetric matrix, since $(A^TA)^T = A^TA$. So the problem now is to maximize the quadratic form $x^T(A^TA)x$ subject to the constraint $\|x\| = 1$. The maximum value is the greatest eigenvalue 1 of A^TA (verify!). Also, the maximum value is attained at a unit eigenvector of A^TA corresponding to λ_1.

For this e.g., we have $A^TA = \begin{pmatrix} 80 & 100 & 40 \\ 100 & 170 & 140 \\ 40 & 140 & 200 \end{pmatrix}$ The eigenvalues of A^TA are $\lambda_1 = 360$, $\lambda_2 = 90$, and $\lambda_3 = 0$. Corresponding unit eigenvectors are, respectively,

$$v_1 = \begin{pmatrix} \frac{1}{3} \\ \frac{2}{3} \\ \frac{2}{3} \end{pmatrix}, v_2 = \begin{pmatrix} -\frac{2}{3} \\ -\frac{1}{3} \\ \frac{2}{3} \end{pmatrix}, v_3 = \begin{pmatrix} \frac{2}{3} \\ -\frac{1}{3} \\ \frac{1}{3} \end{pmatrix}.$$

The maximum value of $\|Ax\|^2$ is 360, attained when x is the unit vector v_1. The vector Av_1 is a point on the ellipse in Figure 7 farthest from the origin, namely, $Av_1 = \begin{pmatrix} 18 \\ 6 \end{pmatrix}$. For $\|x\| = 1$, the maximum value of Ax is $A_1 = \sqrt{360} = 6\sqrt{10}$.

§ The Singular Values of an m × n Matrix

Let A bean $m \times n$ matrix. Then A^TA is symmetric and can be orthogonally diagonalized. Let v_1, v_2, ..., v_n be an orthonormal basis for R^n consisting of eigenvectors of A^TA, and let λ_1, ..., λ_n be the associated eigenvalues of A^TA. Then, for $1 < i < n$, we have

(3.12) $$\begin{aligned} \|Av_i\|^2 &= (Av_i)^TAv_i = v_i^TA^TAv_i \\ &= v_i^T(\lambda_i v_i), \ * \ v_i \text{ is an eigen vector of } A^TA \\ &= \lambda_i, \ * \ v_i \text{ is a unit vector.} \end{aligned}$$

So the eigen values of A^TA are all nonnegative. By renumbering, if necessary, we may assume that the eigen values are arranged so that

$$\lambda_1 \geq \lambda_2 \geq ...., \lambda_n \geq 0$$

The singular values of A are the square roots of the eigen values of A^TA, denoted by ξ_1, ξ_2, ..., ξ_n, and they are arranged in decreasing order. That is, $\xi_i = \sqrt{\lambda_i}$. By equation (3.12), the singular values of A are the lengths of the vectors Av_i, $i = 1, 2, ..., n$.

Example 5.4.4.2. Let A be the matrix in Example 5.4.4.1. Since the eigenvalues of A^TA are 360, 90, and 0, the singular values of A are

$$\sigma_1 = 6\sqrt{10}, \sigma_2 = 3\sqrt{10} \ \& \ \sigma_3 = 0.$$

From Example 5.4.4.1, the first singular value of A is the maximum of ||Ax|| over all unit vectors, and the maximum is attained at the unit eigenvector v_1. Theorem (5.4.3.3) shows that the second singular value of A is the maximum of ||Ax|| over all unit vectors that are orthogonal to v_1, and this maximum is attained at the second unit eigenvector, v_2. For the v_2 in Example 5.4.4.1,

$$Av_2 = \begin{pmatrix} 3 \\ -9 \end{pmatrix}.$$

This point is on the minor axis of the ellipse in Figure 8, just as Av_1 is on the major axis. (See Figure 11.) The first two singular values of A are the lengths of the major and minor semiaxes of the ellipse.

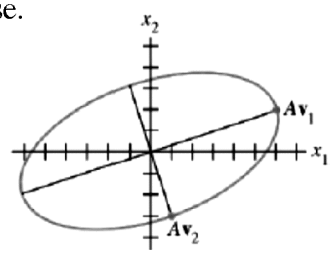


Figure 8

This point is on the minor axis of the ellipse in Figure7, just as Av_1 is on the major axis. (refer to Figure 11.)

Theorem 5.4.4.1. Suppose (v_1, v_2, ..., v_n) is an orthonormal basis of R_n consisting of eigenvectors of A^TA, arranged so that the corresponding eigenvalues of A^TA satisfy λ_1, ..., λ_n, and suppose A has r nonzero singular values. Then $\{Av_1, ..., Av_n$ is an orthogonal basis for Col A, and rank $A = r$.

Proof. Because v_i and $\lambda_j v_j$ are orthogonal for $i = j$, therefore

$$(Av_i)^T(Av_j) = v_i^T A^T v_j = v_i^T(\lambda_j v_j) = 0.$$

Thus $\{Av_1, ..., Av_n\}$ is an orthogonal set. Further more, since the lengths of the vectors Av_1, ..., Av_n are the singular values of A, and as there are r nonzero singular values, $Av_i = 0$ iff $1 \leq i \leq r$. So Av_1, ..., Av_n are linearly independent vectors, and they are in col A. Finally, for any $y \in$ 2 col A-say $y = Ax$– we can express

$x = c_1v_1 + c_2v_2 + ... + c_nv_n$

and

$y = Ax = c_1Av_1 + c_2Av_2 + ... + c_rAv_r + c_{r+1}av_{r+1} + ... + c_nAv_n$

$= c_1Av_1 + c_2Av_2 + ... + c_rAv_r + 0 + ... + 0.$

Thus $y \in$ span $\{Av_1, ..., Av_r\}$, which shows Av_1, ..., Av_n is an orthogonal basis for col A. Hence rank A = dim col A = r.

§ The Singular Value Decomposition

The decomposition of A involves an m × n diagonal matrix of the form

(3.13) $$\Xi = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}$$

where D is an $r \times r$ diagonal matrix for some r not exceeding the smaller of m and n. (If r equals m or n or both, some or all of the zero matrices do not appear.)

Theorem 5.4.4.2. The Singular Value Decomposition : Let A bean m × n matrix with rank r. Then there exists an m × n matrix as in (3.13) for which the diagonal entries in D are the first r singular values of A, ξ_1, ξ_2, ..., $\xi_r > 0$, and there exist an $m \times n$ orthogonal matrix U and an n × n orthogonal matrix V such that $A = U\,\Xi V^T$.

Proof. Let λ_i and v_i be as in Theorem 5.4.4.1, so that Av_i, $i = 1, 2, ..., n$ is an orthogonal basis for Col A. Normalize each Av_i, $i = 1, 2, ..., r$ to obtain an ortho normal basis u_1, u_2, ..., u_r, where

$$u_i = \frac{1}{\| Av_1 \|} Av_1 = \frac{1}{\xi_i} Av_i$$

and

(3.14) $Av_i = \xi_i u_i$

Now extend $\{u_i, i = 1, 2, ..., r\}$ to an orthonormal basis $\{u_i, i = 1, 2, ..., m\}$ of R^m, and let

$U = [u_1 u_2 .... u_m]$ and $V = [v_1 v_2 .... v_n]$.

By construction, U and V are orthogonal matrices. Also, from (3.14),

$$AV = [AV_1 AV_2 ..... v_1 0, ..., 0] = |\xi_i u_1 .....\xi_r u_r 0, ...0|$$

Let D be the diagonal matrix with diagonal entries ξ_1, ..., ξ_r and let be as in (3.13) above. Then

$$U\Xi = [u_1 u_2 ....u_m]\begin{pmatrix} \xi_i & & 0 \\ & \xi_2 \ddots & \\ 0 & & \xi_i \end{pmatrix}$$

$$= [\xi_1 u_1 ....\xi_r u_r 0, ...0] = AV$$

Since V is an orthogonal matrix, $U\ \Xi\ V^T = AV\ V^T = A$.

Remark 5.4.4.1. Any factorization $A = U\Xi V^T$, with U and V orthogonal, Ξ as in equation (3.13), and positive diagonal entries in D_i, is called a singular values decomposition (or *SVD*) of *A*. The matrices *U* and *V* are not uniquely determined by A, but the diagonal entries of are necessarily the singular values of A. The columns of *U* in such a decomposition are called left singular vectors of *A*, and the columns of *V* are called right singular vectors of A.

Example 5.4.4.3. Use the results of Examples 5.4.4.1 and 5.4.4.2 to construct a singular value decomposition of

$$A = \begin{pmatrix} 4 & 11 & 14 \\ 8 & 7 & -2 \end{pmatrix}.$$

A construction can be divided into three steps.

Step 1 Find an orthogonal diagonalization of A^TA : That is, find the eigenvalues of A^TA and a corresponding ortho normal set of eigenvectors. However, for the matrix *A* here, the eigen data for A^TA are provided in Example 5.4.4.1.

Step 2 Set up V and Ξ : Arrange the eigenvalues of A^TA in decreasing order. In Example 3.4.1, the eigenvalues are already listed in decreasing order : 360, 90, and 0.

The corresponding unit eigen vectors, v_1, v_2, v_3 are the right singular vectors of A. Using Example 3.4.1, construct

$$V = [v_1 v_2 v_3] = \begin{pmatrix} \frac{1}{3} & -\frac{2}{3} & \frac{2}{3} \\ \frac{2}{3} & -\frac{1}{3} & -\frac{2}{3} \\ \frac{2}{3} & \frac{2}{3} & \frac{1}{3} \end{pmatrix}$$

The square roots of the eigen values are the singular values :

$$\xi_1 = 6\sqrt{10}, \xi_2 = 3\sqrt{10}, \xi_3 = 0.$$

The nonzero singular values are the diagonal entries of D. The matrix Ξ is the same size as A, with D in its upper left corner and with 0's elsewhere.

$$D = \begin{pmatrix} 6\sqrt{10} & 0 \\ 0 & 3\sqrt{10} \end{pmatrix}, \Xi = [D \ 0] = \begin{pmatrix} 6\sqrt{10} & 0 & 0 \\ 0 & 3\sqrt{10} & 0 \end{pmatrix}$$

Step 3 Construct U : When A has rank r, the first r columns of U are the normalized vectors obtained from Av_i, i = 1, 2, ..., r. In this example, A has two nonzero singular values, so rank A = 2. Recall from Example (5.4.4.2) and the paragraph before Example (5.4.4.2) that $\|Av_1\| = \xi_1$ and $\|Av_2\| = \xi_2$. Thus

$$u_1 = \frac{1}{\xi_1} Av_1 = \frac{1}{6\sqrt{10}} \begin{pmatrix} 18 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{3}{\sqrt{10}} \\ \frac{1}{\sqrt{10}} \end{pmatrix}$$

$$u_2 = \frac{1}{\xi_1} Av_2 = \frac{1}{2\sqrt{10}} \begin{pmatrix} 3 \\ -9 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{10}} \\ \frac{3}{\sqrt{10}} \end{pmatrix}$$

Thus, the singular value decomposition of A is

$$A = \begin{pmatrix} \frac{3}{\sqrt{10}} & \frac{1}{\sqrt{10}} \\ \frac{1}{\sqrt{10}} & \frac{3}{\sqrt{10}} \end{pmatrix} \begin{pmatrix} 6\sqrt{10} & 0 & 0 \\ 0 & 3\sqrt{10} & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{3} & -\frac{2}{3} & \frac{2}{3} \\ \frac{2}{3} & -\frac{1}{3} & -\frac{2}{3} \\ \frac{2}{3} & \frac{2}{3} & \frac{1}{3} \end{pmatrix}$$

Problem 5.4.4.1. Find a singular value decomposition of $A = \begin{pmatrix} 1 & -1 \\ -2 & 2 \\ 2 & -2 \end{pmatrix}$

Solution 5.4.4.1. Step 1 Here $A^TA = \begin{pmatrix} 9 & -9 \\ -9 & 9 \end{pmatrix}$. The eigen values of A^TA are 18 and 0, with corresponding unit eigenvectors

$$v_1 = \begin{pmatrix} \frac{1}{2} \\ -\frac{1}{\sqrt{2}} \end{pmatrix}, v_2 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}.$$

Step 2 These unit vectors form the columns of $V = [v_1\ v_2]$. The singular values are $\xi_1 = \sqrt{18}, \xi_2 = 0.$

Since there is only one nonzero singular value, the "matrix" D may be written as a single number $D = \sqrt{18}$ Hence,

$$\Xi = \begin{pmatrix} D & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} \sqrt{18} & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$$

Step 3 Here, $Av_1 = \begin{pmatrix} \frac{2}{\sqrt{2}} \\ -\frac{4}{\sqrt{2}} \\ \frac{4}{\sqrt{2}} \end{pmatrix}$, $Av_2 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$

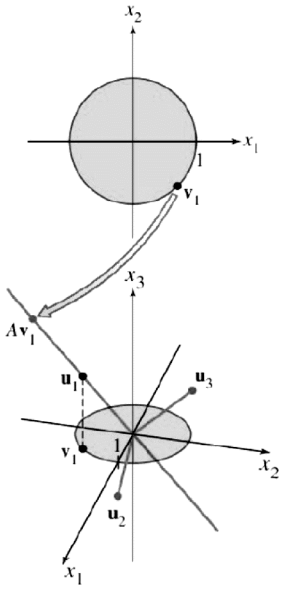


Figure 9

Now, $\|Av_1\| = \xi_1 = \sqrt{18}$ (verify!). Also $\|Av_2\| = \xi_2 = 0$. The only column found for U is,

$$u_1 = \frac{1}{\sqrt{18}} Av_1 = \begin{pmatrix} \frac{1}{3} \\ -\frac{2}{3} \\ \frac{2}{3} \end{pmatrix}$$

The other columns of U are found by extending the set u_1 to an ortho normal basis for R^3. In this case, we need two orthogonal unit vectors u_2 and u_3 that are orthogonal to u_1. (Refer to Figure 9.) Each vector must satisfy $u^T{}_1 x$, which is equivalent to the equation $x_1 - 2x_2 + 2x_3 = 0$. A basis for the solution set of this equation is

$$\omega_1 = \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}, \quad \omega_2 = \begin{pmatrix} -2 \\ 0 \\ 1 \end{pmatrix}.$$

Applying the Gram-Schmidt normalization process to $\{w_1, w_2\}$, one obtain

$$u_2 = \begin{pmatrix} \frac{2}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} \\ 0 \end{pmatrix}, \quad u_3 = \begin{pmatrix} -\frac{3}{\sqrt{45}} \\ \frac{4}{\sqrt{45}} \\ \frac{5}{\sqrt{45}} \end{pmatrix}.$$

Finally taking U = $[u_1\ u_2\ u_3]$, and using steps 2 and 3, we obtain the desired result.

§ Applications of the Singular Value Decomposition

The Singular Value Decomposition (SVD) is often used to estimate the rank of a matrix, as noted above. Several other numerical applications are described briefly below, and an application to image processing is presented in subsection 5.4.5.

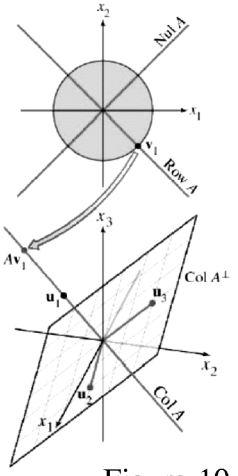


Figure 10

Example 5.4.4.4. (The Condition Number) Most numerical calculations in volving an equation $Ax = b$ are as reliable as possible when the SVD of A is used. The two orthogonal matrices U and V do not affect lengths of vectors or angles between vectors(Theorem 5.4.3.3 in subsection 5.4.3). Any possible in stabilities in numerical calculations are identified in Ξ. If the singular values of A are extremely large or small, roundoff errors are almost inevitable, but an error analysis is aided by knowing the entries in and V. If A is an invertible n × n matrix, then the ratio $\frac{\xi_1}{\xi_n}$ of the largest and smallest singular values gives the condition number of A. Actually, a condition number of A can be computed in several ways, but the definition given here is widely used for studying Ax = b.

Example 5.4.4.5. (Bases for Fundamental Subspaces [7]) Given an SVD for an m × n matrix A, let $u_1, u_2, ..., u_m$ be the left singular vectors, $v_1, ..., v_n$ the right singular vectors, and $\xi_1, \xi_2, ..., \xi_n$ the singular values, and let r be the rank of A. By Theorem 5.4.4.1,

(3.15) $\{u_1, u_2, ..., u_r\}$

is an ortho normal basis for Col A. Also its known that (Col A). = NulA^T. Hence,

(3.16) $\{u_{r+1}, ..., u_m\}$

is an ortho normal basis for Nul A^T. Since $Av_i = \xi_i$ for $1 \leq i \leq n$, and $\xi_i = 0$ if and only if $i > r$, the vectors $v_{r+1}, ..., v_n$ span a subspace of Nul A of dimension n – r. By the Rank Theorem (refer to [3]), we have dim *Nul*A + rank $A = n$ which follows that

(3.17) $\{v_{r+1}, ..., v_n\}$

is an orthonormal basis for Nul A, by the Basis The- orem [refer to [3]]. From (3.15) and (3.17), we have $(Nul\ A^T)^T$ = Col$A$. Interchanging A and A^T, its known that (Nul $A)^T$ = Col A^T + Row A. Hence, from (3.17)

(3.18) $\{v_1, ..., v_r\}$

is an ortho normal basis for Row A.

Remark 5.4.4.1. For the definition of Fundamental subspaces refer to [3].

Theorem 5.4.4.3. The Invertible Matrix Theorem : Let A be an n × n matrix. Then the following statements are each equivalent to the statement that A is an invertible matrix.

- (Col $A)^T = \{0\}$.
- (Nul $A)^T = R_n$.
- Row A = R_n.

- A has n nonzero singular values.

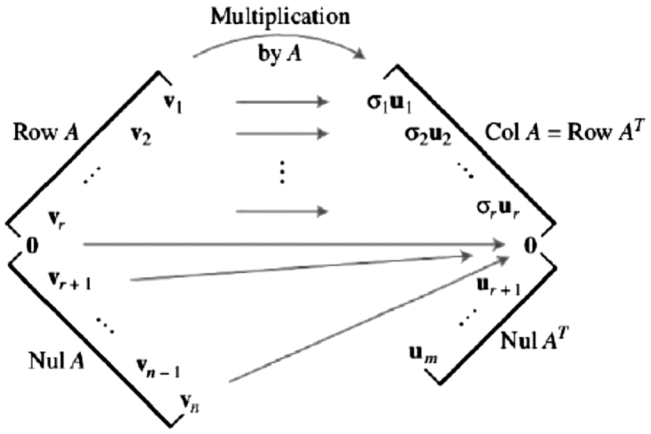


Figure 11

Example 5.4.4.6. (Reduced SVD and the Pseudoinverse of A) When Ξ contains rows or columns of zeros, a more compact decomposition of A is possible. Using the notation established above, let r = rank A, and partition U and V into submatrices whose first blocks contain r columns :

U = [U_r Um–r], where U_r = [u_1 .... u_r]

V = [V_r Vm–r], where V_r = [v_1 .... v_r] Then Ur is $m \times r$ and Vr is $n \times r$. Then partitioned matrix multiplication shows that

(3.19) $$A = [Ur\ U_{m-r}] \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} V^T \\ V_{n-r} \end{pmatrix}$$

This factorization of A is called a reduced singular valued ecomposition of A. Since the diagonal entries in D are nonzero, D is invertible. The following matrix is called the pseudoinverse (also, the Moore-Penrose inverse) of A

(3.20) $Ay = V_r D^{-1} U^T r.$

Example 5.4.4.7. (Least-Squares Solution) Given the equation $Ax = b$, use the pseudo inverse of A in (3.20) to define

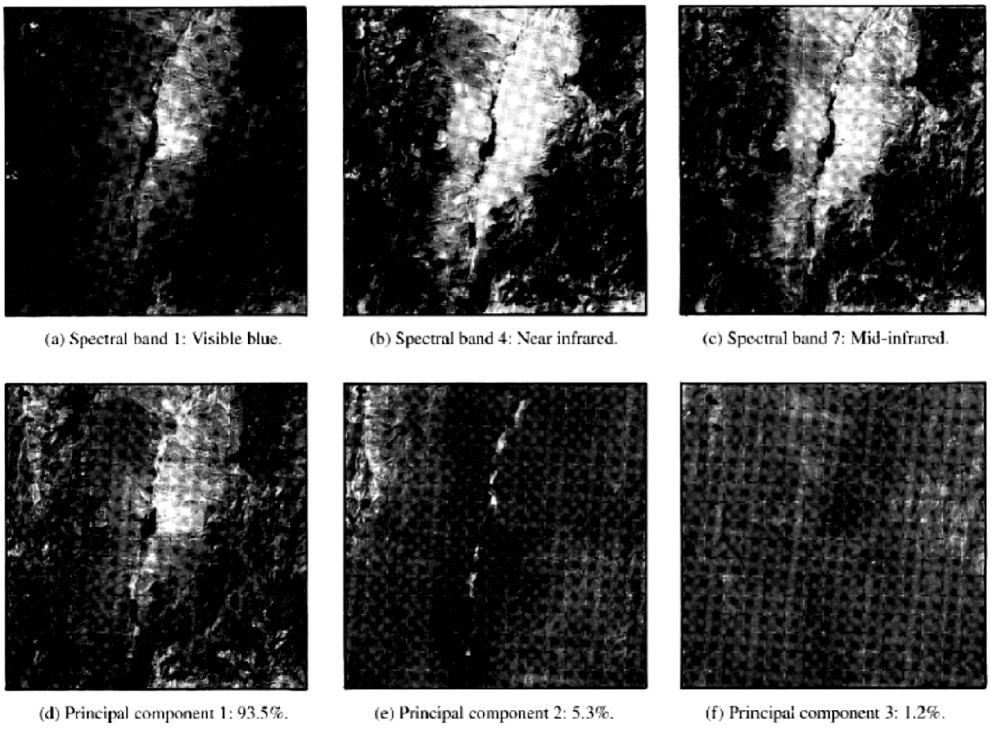
$$x = A^t b = V_r D^{-1} U_r^T b.$$

Now feeding *SVD* in (3.19),

$$\begin{aligned} Ax &= (U_r D V_r^T)(V_r D^{-1} U_r^T b) \\ &= U_r D D^{-1} U_r^T b \\ &= U_T U_r^T b. \end{aligned}$$

It follows from (3.15) that UrU^{T}_{r}b is the ortho gonal projection b of b on to Col A. Thus ^x is a least-squares solution of $Ax = b$. Infact, this x has the smallest length among all least-squares solutions of $Ax = b$.

5.4.5 Applications to image processing and statistics



(a) Spectral band 1: Visible blue. (b) Spectral band 4: Near infrared. (c) Spectral band 7: Mid-infrared.

(d) Principal component 1: 93.5%. (e) Principal component 2: 5.3%. (f) Principal component 3: 1.2%.

Figure 12 : satelite image

The main goal of this section is to explain a technique, called principal component analysis, used to analyze such multivariate data. The calculations will illustrate the use of orthogonal diagonalization and the singular value decomposition. Principal component analysis can be applied to any data that consist of lists of measurements made on a collection of objects or individuals. For instance, consider a chemical process that produces a plastic material. To monitor the process, 300 samples are taken of the material produced, and each sample is subjected to a battery of eight tests, such as melting point, density, ductility, tensile strength, and soon. The laboratory report for each sample is a vectorin R^8, and the set of such vectors forms an 8×300 matrix, called the matrix of observations. Loosely speaking, we can say that the process control data are eight-dimensional. The following example describe data that can be visualized graphically.

Example 5.4.5.1. An example of two-dimensional data is given by a set of weights and heights of *N* college students. Let X_i denote the observation vector in R^2 that lists the weight and height of the ith student. If ω denotes weight and *h* height, then the matrix of observations has the form

$$\begin{pmatrix} w_1 & w_2 & .... & w_n \\ h_1 & h_2 & .... & h_n \end{pmatrix}$$

where, $X_i = (w_i, h_i)$, i = 1, 2, ..., N, the set of observation vectors can be visualized as a two dimensional scatter plot. Refer to Figure 13.

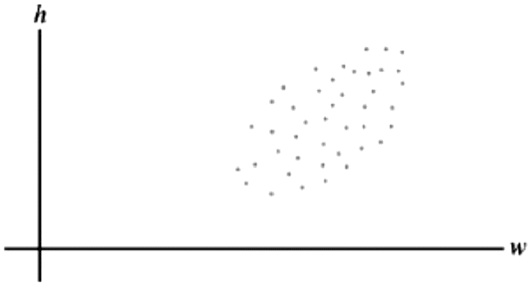


Figure 13 : A scatter plot of spectral data for a satellite image

Example 5.4.5.2. The first three photographs of Railroad Valley, Nevada, shown in the beginning of the section can be viewed as one image of the region, with three spectral components, because simultaneous measurements of the region were made at three separate wave lengths. Each photograph gives different information about the same physical region. For instance, the first pixel in the upper-left corner of each photograph corresponds to the same place on the ground (about 30 meters by 30 meters). To each pixel there corresponds an observation vectorin R^3 that lists the signal intensities for that pixel in the three spectral bands. Typically, the image is 2000 × 2000 pixels, so there are 4 million pixels in the image. The data for the image form a matrix with 3 rows and 4 million columns (with columns 34 arranged in any convenient order). In this case, the "multidimensional" character of the data refers to the three spectral dimensions rather than the two spatial dimensions that naturally belong to any photograph. The data can be visualized as a cluster of 4 million points in R^3, perhaps as in Figure 14.

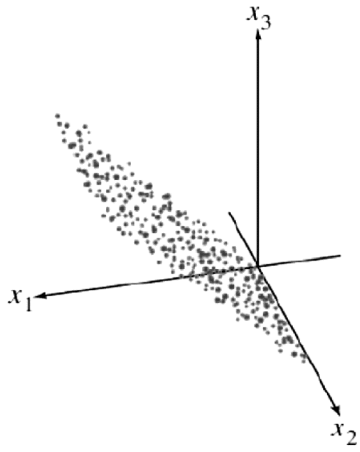


Figure 14 : A scatter plot of spectral data for a satellite image.

§ Mean and Covariance

For principal component analysis, let [X_1, ..., X_N] be a p × N matrix of observations, such as described above. The sample mean, M, of the observation vectors X_i, i = 1, 2, ..., N is given by

$$M = \frac{1}{N}(X_i + ... + X_N)$$

For the data in Figure 13, the sample mean is the point in the "center" of the scatter plot. For k = 1, 2, ..., N. Let

$$\overline{X}_k = X_k - M.$$

The columns of the $p \times N$ matrix

$$B = [\overline{X}_1 .... X_N]$$

have a zero sample mean, and B is said to be in mean-deviation form. When the sample mean is subtracted from the data in Figure13, the resulting scatter plot has the form in Figure 15.

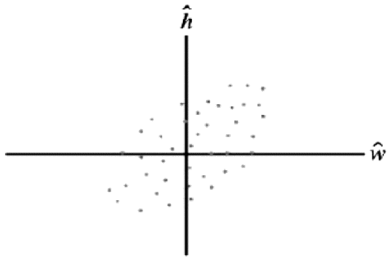


Figure 15 : Weight-height data in mean-deviation form.

The (sample) covariance matrix is the $p \times p$ matrix S defined by

$$S = \frac{1}{N-1} BB^T.$$

Since any matrix of the form BB^T is positive semidefinite, so is S.

Example 5.4.5.3. Three measurements are made on each of four individuals in a random sample from a population. The observation vectors are

$$X_1 = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, X_2 = \begin{pmatrix} 4 \\ 2 \\ 13 \end{pmatrix}, X_3 = \begin{pmatrix} 7 \\ 8 \\ 1 \end{pmatrix}, X_4 = \begin{pmatrix} 3 \\ 4 \\ 5 \end{pmatrix}.$$

Calculate the sample mean and the covariance matrix.

The sample mean is

$$M = \frac{1}{4}\left\{\begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} + \begin{pmatrix} 4 \\ 2 \\ 13 \end{pmatrix} + \begin{pmatrix} 7 \\ 8 \\ 1 \end{pmatrix} + \begin{pmatrix} 8 \\ 4 \\ 5 \end{pmatrix}\right\} = \begin{pmatrix} 5 \\ 4 \\ 5 \end{pmatrix}$$

Subtract the sample mean from $\bar{X}_1, \bar{X}_2, \bar{X}_3, \bar{X}_4$ to obtain

$$\bar{X}_1 = \begin{pmatrix} -1 \\ -2 \\ -4 \end{pmatrix}, \bar{X}_2 = \begin{pmatrix} -1 \\ -2 \\ 8 \end{pmatrix}, \bar{X}_3 = \begin{pmatrix} 2 \\ 4 \\ -4 \end{pmatrix}, \bar{X}_4 = \begin{pmatrix} 3 \\ 0 \\ 0 \end{pmatrix}$$

and $$B = \begin{pmatrix} -4 & -1 & 2 & 3 \\ -2 & -2 & 4 & 0 \\ -4 & 8 & -4 & 0 \end{pmatrix}$$

The sample covariance matrix is

$$S = \frac{1}{3} BB^T = \begin{pmatrix} 10 & 6 & 0 \\ 6 & 8 & -8 \\ 0 & -8 & 32 \end{pmatrix}$$

To discuss the entries in $S = [s_{ij}]$, let X represent a vector that varies over the set of observation vectors and denote the coordinates of X by x_i, $i = 1, 2, \ldots, p$. Then x_1, for example, is a scalar that varies over the set of first coordinates of $[X_1 \ldots X_N]$. For $j = 1, 2, \ldots, p$, the diagonal entry sjj in S is called the variance of x_j. The variance of x_j measures the spread of values of x_j. The total variance of the data is the sum of the variances on the diagonal of S. In general, the sum of the diagonal entries of a square matrix S is called the trace of the matrix, written $tr(S)$. Thus

total variance = $tr(S)$.

The entry s_{ij} in S for $i = j$ is called the covariance of x_i and x_j. Observe that in Example 3.5.3, the covariance between x_1 and x_3 is 0 because the (1, 3) entry in S is 0. Statisticians say that x_1 and x_3 are uncorrelated. Analysis of the multivariate data in X_1 ..... X_N is greatly simplified when most or all of the variables x_1, x_2, ..., x_p are uncorrelated, i.e. when the covariance matrix of X_1 ..... X_N is diagonal or nearly diagonal.

§ Principal Component Analysis

Let us assume the matrix [X_1 ..... X_N] is already in mean-deviation form. The goal of principal component analysis is to find an orthogonal $p \times p$ matrix $P = [u_1 \ ..... \ u_p]$ that determines a change of variable $X = PY$ or

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} = [u_1 \ ..... \ u_p] \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix}$$

with the property that the new variables y_1, ..., y_p are uncorrelated and are arranged in order of decreasing variance.

The unit eigenvectors u_1, ..., u_p of the covariance matrix S are called the principal components of the data (in the matrix of observations). The first principal component is the eigenvector corresponding to the largest eigenvalue of S, the second principal component is the eigenvector corresponding to the second largest eigenvalue, and so on.

The first principal component u_1 determines the new variable y_1 in the following way :

Let c_1, ..., c_p be the entries in u_1. Since u^T_1 is the first row of P^T, the equation $Y = P^TX$ shows that

$$y_1 = u_1^T X = c_1 x_1 + ... + c_p x_p.$$

Thus y_1 is a linear combination of the original variables x_1 ....., x_p, using the entries in the eigenvector u_1 as weights. In a similar fashion, u_2 determines the variable y_2, and soon.

Example 5.4.5.4. The initial data for the multi-spectral image of Railroad Valley (Example 5.4.5.2) consisted of 4 million vectors in R^3. The associated covariance matrix is

$$S = \begin{pmatrix} 2382.78 & 2611.84 & 2136.20 \\ 2611.84 & 3106.47 & 2553.90 \\ 2136.20 & 2553.90 & 2650.71 \end{pmatrix}$$

Find the principal components of the data, and list the new variable determined by the first principal component.

The eigenvalues of S and the associated principal components (the unit eigenvectors) are

$$\lambda_1 = 7614.23 \qquad \lambda_2 = 427.63 \qquad \lambda_3 = 98.10$$

$$u_1 = \begin{pmatrix} 5417 \\ 6295 \\ 5570 \end{pmatrix} \qquad u_2 = \begin{pmatrix} -4894 \\ -3026 \\ -8179 \end{pmatrix} \qquad u_3 = \begin{pmatrix} 6384 \\ -7157 \\ 1441 \end{pmatrix}$$

Using two decimal places for simplicity, the variable for the first principal component is

$$y_1 = 54x_1 + 63x_2 + 56x_3 \ .$$

This equation was used to create Figure 12(d) in the beginning of the section. The variables x_1, x_2, and x_3 are the signal intensities in the three spectral bands. The values of x_1, converted to a grayscale between black and white, produced Figure12(a). Similarly, the values of x_2 and x_3 produced Figure 12(b) and Figure12(c), respectively. At each pixel in Figure 12(d), the grayscale value is computed from y_1, a weighted linear combination of x_1, x_2 and x_3. In this sense, Figure12(d) "displays" the first principal component of the data.

§ Reducing the Dimension of Multivariate Data

Principal component analysis is potentially valuable for applications in which most of the variation, or dynamic range, in the data is due to variations in only a few of the new variables, y_1, ..., y_p. It can be shown that an orthogonal change of variables, $X = PY$, does not change the total variance of the data. (Roughly speaking, this is true because left-multiplication by P does not change the lengths of vectors or the angles between them. This means that if $S = PDP^T$, then

{total variance of x_1, ..., x_p} = {total variance of y_1, ..., y_p} = $tr\ (D) = \lambda_1 + ...+ \lambda_p$

The variance of y_j is λ_j, and the quotient λ_j = trS measures the fraction of the total variance that is "explained" or "captured" by y_j.

Remark 5.4.5.1. The calculations in Exercise 5.4.5.1 will reflect the data that have practically no variance in the third (new) coordinate. The values of y_3 are all close to zero. Geometrically, the data points lie nearly in the plane $y_3 = 0$, and their locations can be determined fairly accurately by knowing only the values of y_1 and y_2. In fact, y_2 also has relatively small variance, which means that the points lie approximately along a line, and the data are essentially one-dimensional.

§ Characterizations of Principal Component Variables

If y_1, ..., y_p a rise from a principal component analysis of a $p \times$ N matrix of observations, then the variance of y_1 is as large as possible in the following sense : If u is any unit vector and if $y = u^TX$, then the variance of the values of y as X varies over the original data X_1, ..., X_N turns out to be u^TSu. By Theorem 5.4.3.4 the maximum value of u^TSu, overall unit vectors u, is the largest eigenvalue 1 of S, and this variance is attained when u is the corresponding eigenvector u_1. In the sameway, Theorem 5.4.3.4 shows that y_2 has maximum possible variance a mongall variables $y = u^T$X that are uncorrelated with y_1. Likewise, y_3 has maximum possible variance among all variables uncorrelated with both y_1 and y_2, and so on.

5.5 Objective Type Questions

Mark each statement True or False. Justify each answer. In each part, A represents an $n \times n$ matrix.

(i) If A is orthogonally diagonalizable, then A is symmetric.

(ii) If A is an orthogonal matrix, then A is symmetric.

(iii) If A is anorthogonal matrix, then $\|Ax\| = \|x\|$ for all $x \in R^n$.

(iv) The principal axes of a quadratic form x^TAx can be the columns of any matrix P that diagonalizes A.

(v) If P is an $n \times n$ matrix with orthogonal columns, then $P^T = P^{-1}$.

(vi) If every coefficient in a quadratic form is positive, then the quadratic form is positive definite.

(vii) If $x^TAx > 0$ for some x, then the quadratic form x^TAx is positive definite.

(viii) By a suitable change of variable, any quadratic form can be changed in to one with no cross-product term.

(ix) A positive definite quadratic form can be changed into a negative definite form by a suitable change of variable $x = Pu$, for some orthogonal matrix P.

(x) An indefinite quadratic form is one whose eigenvalues are not definite.

(xi) If A is $n \times n$, then A and A^TA have the same singular values.

5.6 Summary

The present unit is focused on different special types of matrices : idempotent, nilpotent, involution and proection, tri-diagonal matrices, circulantmatrices. The learners can now explain Vandermonde matrices, Handmard matrices, permutation and doubly stochastic matrices and their use in solving different problems. The unit also introduces the concepts of Positive Semi-definite matrices and the method to find the square root of a positive semi-definite matrix. The unit also includes diagonalization of symmetric matrices, quadratic forms, constrained optimization, singular value decomposition, and applications to image processing and statistics.

5.7 Exercises

1. Compute the various percentages of variance of the Railroad Valley multi-spectral data that are displayed in the principal component Figure 12,(d)–(*f*), shown in the beginning of the section.

2. Find the singular values of the matrix $\begin{pmatrix} 3 & 0 \\ 8 & 3 \end{pmatrix}$

3. Find the SVD of $A = \begin{pmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{pmatrix}$

4. Suppose the factorization below is an SVD of a matrix A, with the entries in U and V rounded to two decimal places.

$$A = \begin{pmatrix} .40 & -.78 & .47 \\ .37 & -.33 & -.87 \\ -.84 & -.52 & -.16 \end{pmatrix} \begin{pmatrix} 7.10 & 0 & 0 \\ 0 & 2.10 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} .30 & -.51 & -.81 \\ .76 & .64 & -.12 \\ .58 & -.58 & .58 \end{pmatrix}$$

   (i) What is the rank of A ?

   (ii) Use this decomposition of A, with no calculations, to write a basis for Col A and a basis for Nul A.

5. Show that the columns of V are eigenvectors of A^TA, the columns of U are eigenvectors of AA^T, and the diagonal entries of are the singular values of A.

6. Show that if P is a northogonal $m \times m$ matrix, then PA has the same singular values as A.

5.8 References

[1] Herstein, I.N., Winter, D.J., A Primer on Linear Algebra, Macmillan Publishing Company, New York.

[2] Lay, David C., Linear Algebra and its Applications, Pearson New International Edition.

[3] Friedberg, S., Insel, A., Spence, L., Linear Algebra, Pearson New International Edition.

[4] Zhang, F., Matrix Theory : Basic Results and Techniques, Second Edition, Universitext, Springer.

[5] Ayres Jr, Frank, Schaum's Theory and Problems of Matrices, Schaum Publishing Co, NewYork.

[6] Gentle, James, Matrix Algebra, Theory, Computations and Applications in Statistics, Springer.

[7] Strang, Gilbert, Linear Algebra and its Applications, Cengage Publications.

[8] Hogben, Leslie, Hand Book on Linear Algebra, Chapman and Hall CRC.

[9] Lax, Peter D., Linear Algebra and its Applications, Wiley Interscience.

NOTES