CS724 Class Notes

Steve Revilak

January 2009 – May 2009

Part 1

Matrix and Vector Norms

1.1 Lecture – 2/2/2009

In this course, we'll study applications of linear algebra algorithms. We we start with linear spaces and matrices.

1.1.1 Matrices

Consider the problem of text retrieval. We have D, a collection of documents and T, a collection of terms. D is often called a *corpus*. The task is to retrieve documents $\in D$ that contain a user-specified set of terms.

The set of documents and terms can be described as a matrix M. In M, we use one row per document, one column per term. Within the matrix M,

$$m_{ij} = \begin{cases} 1 & \text{if } t_j \in d_i \\ 0 & \text{otherwise} \end{cases}$$

Suppose we are given a retrieval query Q. Q consists of a set of terms $\{t_{i1}, \ldots, t_{il}\}$. Q is very similar to a row in the matrix M, so we can write Q as a vector $\mathbf{q} = (q_1, \ldots, q_m)$, where

$$q_j = \begin{cases} 1 & \text{if } t_j \text{ occurs in } \mathbf{q} \\ 0 & \text{otherwise} \end{cases}$$

The general goal of document retrieval is to find documents that are "close" to the query. Note that we've put *close* in quotes – we have to define what close means (usually in terms of some kind of distance measure). We'll need a way to measure distance in m-dimensional space $\mathbb{R}^m$.

Often, the distance computation can be made easier by decomposing M into smaller matrices. Later in the semester, we'll study ways of doing this. We'll also study techniques for avoiding floating point errors in these kinds of computations.

Given $A\mathbf{x} = \mathbf{b}$, we can solve for $\mathbf{x}$ by multiplying by the inverse of matrix A: $(A^{-1})A\mathbf{x} = (A^{-1})\mathbf{b}$, so that $\mathbf{x} = (A^{-1})\mathbf{b}$. This a standard approach in linear algebra; however, floating point errors can make this procedure more complex.

1.1.2 Applications of Matrices

Pattern Recognition

Suppose we are given the task of identifying handwritten digits. We'll begin with a *training process*, where we collect different examples of written numbers, and digitize them. For the sake of example, suppose we digitize each sample into a 16×16 matrix, where each matrix elements denotes a grayscale value. Figure 1.1 shows three samples that we might collect.

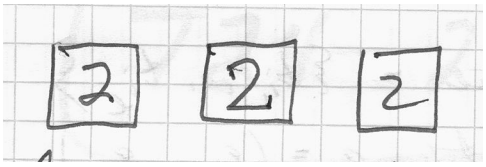

Figure 1.1: Three handwritten samples of the number '2'

These 16×16 samples can also be treated as vectors in $\mathbb{R}^{256}$. How does one go about matching an input digit to our collection of known digits?

Suppose we have 10 subspaces $T_0, \ldots, T_9$, each corresponding to samples of digits $0 \ldots 9$. We obtain an input vector $\mathbf{v}$ by digitizing a handwritten digit. Next, we measure the distance between $\mathbf{v}$ and each subspace T_i. We categorize $\mathbf{v}$ as the i whose subspace T_i produced the smallest distance.

Graph Problems

The internet is a source of many graph problems, where pages correspond to vertices and links correspond to edges.

Given a graph $G = (V, E)$, there are many ways to represent G. A common method is to represent G with an $n \times n$ matrix ($n = |V|$). In this matrix,

$$m_{ij} = \begin{cases} 1 & \text{if there is an edge from } v_i \text{ to } v_j \\ 0 & \text{otherwise} \end{cases}$$

If G is an undirected graph, we can use an incidence matrix representation.

Eigenvalues, spectral clustering, and tensors can reveal many details of a matrix structure. We'll look at some of these techniques later in the course.[1]

1.1.3 Linear Spaces

All linear spaces require a *field*. A field is a set F with two operations: multiplication and addition. The most common fields are $\mathbb{R}$ and $\mathbb{C}$, the sets of real numbers and the set of complex numbers.

For addition, $(F, +)$ forms a commutative group with three properties:

1. $+$ is commutative and associative
2. there is an element (zero) that is neutral with respect to addition.
3. For every $x \in F$ there is a $(-x)$ such that $x + (-x) = 0$.

Similarly, for multiplication $(F, \cdot)$ gives us three properties:

1. multiplication is associative and commutative

[1] Usually we think of matrices as being two dimensional. Tensors are matrices with n-dimensions.

2. there is an element (one) that is neutral with respect to multiplication.
3. If $x \neq 0$ then there exists a an x^{-1} such that $x \cdot x^{-1} = x^{-1} \cdot x = 1$.

The operations $+$ and $\cdot$ are linked by distributive laws: $x \cdot (y + z) = (x \cdot y) + (x \cdot z)$.

Linear spaces are always defined relative to a field.

V is an F-linear space if two operations are defined on V:

1. Addition. For any $\mathbf{x}$, $\mathbf{y} \in V$, we have $\mathbf{x} + \mathbf{y} \in V$.
2. Multiplication between F and V. For any $a \in F$ and $\mathbf{v} \in V$, $a\mathbf{v} \in V$.

Additionally, $(V, +)$ must be a commutative group.

There is also a zero vector, $\mathbf{0}$, such that $0 \cdot \mathbf{v} = \mathbf{0}$.

Vectors also obey distributive laws:

$$a \cdot (\mathbf{x} + \mathbf{y}) = a\mathbf{x} + a\mathbf{y}$$
$$(a + b) \cdot \mathbf{x} = a\mathbf{x} + b\mathbf{x}$$

We will deal mostly with

$\mathbb{R}^n$: the linear space of real numbers
$\mathbb{C}^n$: the linear space of complex numbers

1.1.4 Norms on Linear Spaces

Let V be an R-linear space (a linear space on the set of Reals). An R-norm is

$$\nu \colon V \to \mathbb{R}_{\geq 0}$$

Thus, the norm ν is a mapping between V and non-negative reals. ν must satisfy the following conditions:

1. $\nu(\mathbf{x}) \geq 0$
2. If $\nu(\mathbf{x}) = 0$, then $\mathbf{x} = \mathbf{0}$.
3. $\nu(a\mathbf{x}) = |a| \cdot \nu(\mathbf{x})$. This is the *homogeneity condition*.
4. $\nu(\mathbf{x} + \mathbf{y}) \leq \nu(\mathbf{x}) + \nu(\mathbf{y})$. This is the *triangular inequality*.

Norms on $\mathbb{R}^n$

In $\mathbb{R}^n$, $\nu \colon \mathbb{R}^n \to \mathbb{R}_{\geq 0}$.

If $p \geq 1$, then

$$\nu(\mathbf{x}) = (|x_1|^p + |x_2|^p + \ldots + |x_n|^p)^{\frac{1}{p}}$$

is also a norm.

The most common norm is the *euclidean norm*, where $p = 2$. The Euclidean norm is:

$$\nu_2(\mathbf{x}) = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}$$

How does one prove that a norm is a norm? This can be an interesting problem in itself. In general, we want to prove that, for $p \geq 1$, than $\nu(\mathbf{x})$ meets the three conditions of a norm.

The first condition is usually trivial to prove. For the second condition,

$$\begin{aligned}\nu_p(a\mathbf{x}) &= (|ax_1|^p + |ax_2|^p + \ldots + |ax_n|^p)^{\frac{1}{p}} \\ &= |a| \cdot (|x_1|^p + |x_2|^p + \ldots + |x_n|^p)^{\frac{1}{p}} \\ &= |a| \cdot \nu_p(\mathbf{x})\end{aligned}$$

The third condition is usually tricky. Let p, q be two numbers such that $\frac{1}{p} + \frac{1}{q} = 1$, and note that

$$\begin{aligned}\frac{1}{q} &= 1 - \frac{1}{p} \\ &= \frac{p-1}{p} \\ q &= \frac{p}{p-1}\end{aligned}$$

Figure 1.2 is a graph of $q = \frac{p}{p-1}$.

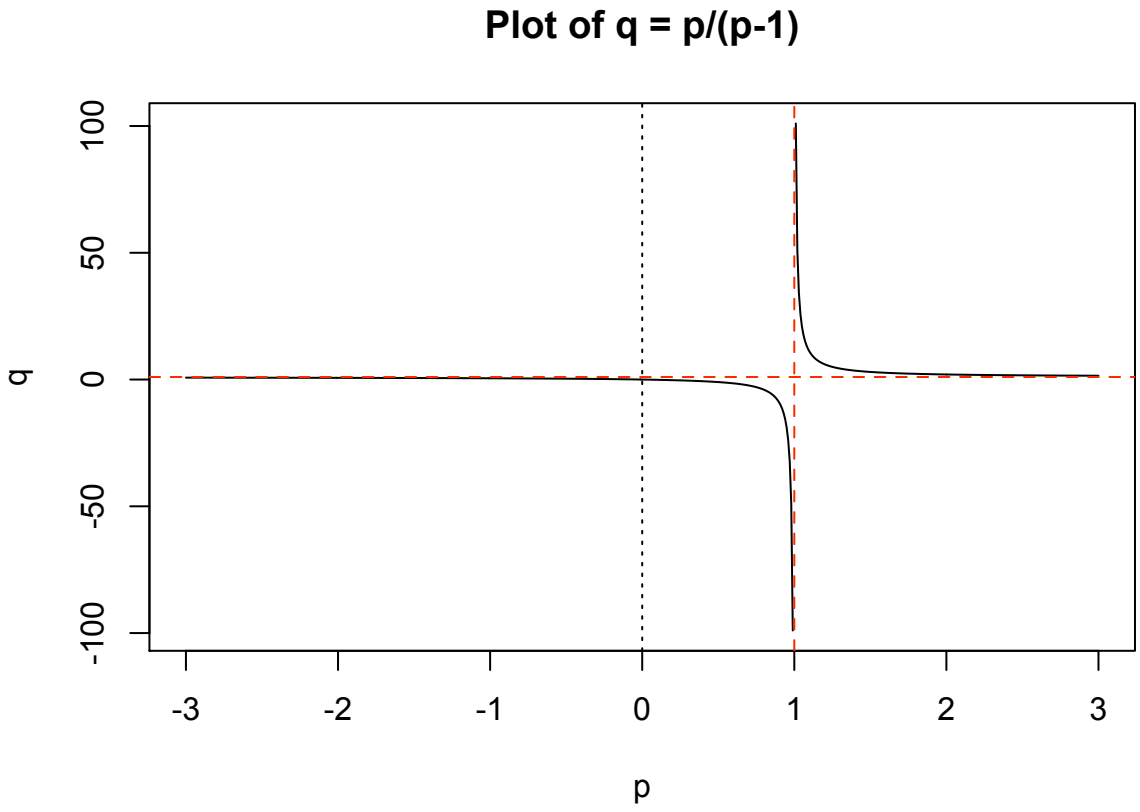

Figure 1.2: Graph of $q = \frac{p}{p-1}$

In Figure 1.2, we can note the following:

- There is a point of discontinuity at $p = 1$.
- $p > 1$ iff $q > 1$

- $p \in (0,1)$ iff $q < 0$.
- $p, q \notin \{0,1\}$.

Suppose a, b are both positions and $p > 1$. We claim that

$$ab < \frac{a^p}{p} + \frac{b^q}{q}.$$

Consider

$$f(x) = \frac{x^p}{p} + \frac{1}{q} - x, \text{ for } x > 0$$

We have

$$\begin{aligned} f(1) &= \frac{1^p}{p} + \frac{1}{q} - 1 \\ &= \frac{1}{q} + \frac{1}{q} - 1 \\ &= 1 - 1 && \text{since } \frac{1}{p} + \frac{1}{q} = 1 \\ &= 0 \end{aligned}$$

The first derivative is

$$\begin{aligned} f'(x) &= x^{p-1} - 1 \\ f'(1) &= 1 - 1 = 0 \end{aligned}$$

The second derivative is

$$\begin{aligned} f''(x) &= (p-1)(x^{p-2}) \\ f''(1) &= p - 1 > 0 && \text{since } p, q > 0 \end{aligned}$$

The second derivative tells us that $f(1)$ is a minimum. Since $f(1) = 0$, and $f(1)$ is a minimum, we know that $f(x) \geq f(1) = 0$. Therefore,

$$\frac{x^p}{p} + \frac{1}{q} - x \geq 0$$

Next, let's point out a few equalities with p and q:

$$\begin{aligned} \frac{p}{q} &= p - 1 \\ \frac{q}{p} &= \frac{1}{p-1} \\ -\frac{q}{p} &= -\frac{1}{p-1} \end{aligned}$$

Now, let's replace x with $ab^{-\frac{1}{p-1}}$. This gives

$$\begin{aligned} &= \frac{a^p b^{-\frac{p}{p-1}}}{p} + \frac{1}{q} - ab^{-\frac{1}{p-1}} \geq 0 \\ &= \frac{a^p b^{-q}}{p} + \frac{1}{q} - ab^{-\frac{q}{p}} \geq 0 && \text{since } -\frac{q}{b} = -\frac{1}{p-1} \\ &= \frac{a^p}{q} + \frac{b^q}{q} - ab^{q-\frac{q}{p}} \geq 0 && \text{multiply by } b^q \\ &= \frac{a^p}{p} + \frac{b^q}{q} - ab \geq 0 && \text{since } q - \frac{q}{p} = 1 \\ \therefore &= \frac{a^p}{p} + \frac{b^q}{q} \geq ab \end{aligned}$$

Suppose we have two sequences of n numbers

$$a_1, \ldots, a_n$$
$$b_1, \ldots, b_n$$

Where all a_i, $b_i > 0$. We can say

$$\sum_{i=1}^{n} a_i b_i \leq \left(\sum_{i=1}^{n} a_i^p \right)^{\frac{1}{p}} \cdot \left(\sum_{i=1}^{n} b_i^q \right)^{\frac{1}{q}} \tag{1.1}$$

When $1/p + 1/q = 1$. Equation (1.1) is called the Hölder Inequality.

A common form of (1.1) is when $p = q = 2$:

$$\sum_{i=1}^{n} a_i b_i \leq \sqrt{\sum_{i=1}^{n} a_i^2} \cdot \sqrt{\sum_{i=1}^{n} b_i^2} \tag{1.2}$$

Equation (1.2) is called the Chaucy-Schwarz Inequality.

1.1.5 Miscellany

- For software, we have the choice of using Matlab or Scilab. (Scilab is very similar to Matlab). Octave is another software package that's worth looking at.
- We will have a makeup class, probably on a Saturday, date TBD. It will most likely be a "double" class.
- Look up "spectral clustering" (what is it?)
- Dig up a linear algebra book, and review the procedure for finding the inverse of a matrix.

1.2 Lecture – 2/4/2009

1.2.1 Vector Norms in $\mathbb{R}^n$

Given the vector $\mathbf{a} = (a_1, \ldots, a_n) \in \mathbb{R}^n$, we would like to prove that $\nu_p(\mathbf{a}) = (|a_1|^p + \ldots + |a_n|^p)^{\frac{1}{p}}$ is a norm for $p \geq 1$.

Recall that a norm must satisfy four properties:

1. $\nu_p(\mathbf{a}) \geq 0$
2. $\nu_p(\mathbf{a}) = 0$ iff $\mathbf{a} = \mathbf{0}$.
3. $\nu_p(b \cdot \mathbf{a}) = |b| \cdot \nu_p(\mathbf{a})$
4. $\nu_p(\mathbf{a} + \mathbf{c}) \leq \nu_p(\mathbf{a}) + \nu_p(\mathbf{c})$ for ever $\mathbf{a}$, $\mathbf{c} \in \mathbb{R}^n$.

The first three properties are usually easy to prove. We will concentrate on the fourth.

Given a parameter $p > 1$, and the condition $\frac{1}{p} + \frac{1}{q} = 1$, we know that $\frac{a^p}{p} + \frac{b^q}{q} \geq ab$. (We proved this in the last lecture.)

Today, we'll prove the Hölder Inequality. Suppose we have vectors $(a_1, \ldots, a_n)$ and $(b_1, \ldots, b_n) \in \mathbb{R}_{>0}$, and $p > 1$. We would like to prove

$$\sum_{i=1}^{n} a_i b_i \leq \left(\sum_{i=1}^{n} a_i^p \right)^{\frac{1}{p}} \cdot \left(\sum_{i=1}^{n} b_i^q \right)^{\frac{1}{q}} \tag{1.3}$$

Suppose we have

$$x_i = \frac{a_i}{(\sum_{i=1}^{n} a_i^p)^{\frac{1}{p}}} \qquad \text{for } 1 \leq i \leq n \tag{1.4}$$

$$y_i = \frac{b_i}{(\sum_{i=1}^{n} b_i^q)^{\frac{1}{q}}} \qquad \text{for } 1 \leq i \leq n \tag{1.5}$$

Then

$$x_i y_i \leq \frac{x_i^p}{p} + \frac{y_i^q}{q} \tag{1.6}$$

We substitute (1.4) and (1.5) back into (1.6):

$$\frac{a_i b_i}{(\sum_{i=1}^{n} a_i^p)^{\frac{1}{p}} \cdot (\sum_{i=1}^{n} b_i^q)^{\frac{1}{q}}} \leq \frac{a_i^p}{\sum_{i=1}^{n} a_i^p} \cdot \frac{1}{p} + \frac{b_i^q}{\sum_{i=1}^{n} b_i^q} \cdot \frac{1}{q} \tag{1.7}$$

Inequality (1.7) holds for $1 \leq i \leq n$.

$$\frac{\sum_{i=1}^{n} a_i b_i}{(\sum_{i=1}^{n} a_i^p)^{\frac{1}{p}} (\sum_{i=1}^{n} b_i^q)^{\frac{1}{q}}} \leq \frac{1}{p} + \frac{1}{q} = 1 \tag{1.8}$$

From (1.8), we can get (1.3). (*How did we get* (1.8)*?*) □

What we've shown so far works for $\mathbf{a}$, $\mathbf{b} \in \mathbb{R}_{\geq 0}$. Now we'd like to develop inequalities that work for negative numbers as well.

Let $\mathbf{a} = (a_1, \ldots, a_n)$ and $\mathbf{b} = (b_1, \ldots, b_n)$ be two vectors of real numbers (they can contain negative numbers).

In this case $(|a_1|, \ldots, |a_n|)$ and $(|b_1|, \ldots, |b_n|)$ are non-negative, so we can apply the previous result from (1.3):

$$\sum_{i=1}^{n} |a_i b_i| \leq \left(\sum_{i=1}^{n} |a_i|^p \right)^{\frac{1}{p}} \left(\sum_{i=1}^{n} |b_i|^q \right)^{\frac{1}{q}} \tag{1.9}$$

The right-hand side of (1.9) is positive, and $|\sum_{i=1}^{n} a_i b_i| \leq \sum_{i=1}^{n} |a_i b_i|$. This gives us the Hölder Inequality in (1.10)

$$\left| \sum_{i=1}^{n} a_i b_i \right| \leq \left(\sum_{i=1}^{n} a_i^p \right)^{\frac{1}{p}} \cdot \left(\sum_{i=1}^{n} b_i^q \right)^{\frac{1}{q}} \tag{1.10}$$

An important special case of (1.10) is when $p = q = 2$; this is called the Cauchy-Schwartz Inequality:

$$\left| \sum_{i=1}^{n} a_i b_i \right| \leq \sqrt{\sum_{i=1}^{n} a_i^2} \cdot \sqrt{\sum_{i=1}^{n} b_i^2} \tag{1.11}$$

In (1.11) notice that the left side is the dot product of $\mathbf{a}$ and $\mathbf{b}$, and the right side is the product of the (Euclidean) norms of $\mathbf{a}$ and $\mathbf{b}$.

1.2.2 Minkovski Inequality

The Minkovski Inequality is useful for proving the fourth property of norms. We will prove that

$$\left(\sum_{i=1}^{n} |a_i + b_i|^p \right)^{\frac{1}{p}} \leq \left(\sum_{i=1}^{n} |a_i|^p \right)^{\frac{1}{p}} + \left(\sum_{i=1}^{n} |b_i|^p \right)^{\frac{1}{p}} \tag{1.12}$$

Note that (1.12) is equivalent to $\nu_p(\mathbf{a} + \mathbf{b}) = \nu_p(\mathbf{a}) + \nu_p(\mathbf{b})$.

The Minkovski Inequality depends only on p where $p \geq 1$.

For $p = 1$, the inequality is trivial:

$$\sum_{i=1}^{n} |a_i + b_i| \leq \sum_{i=1}^{n} |a_i| + \sum_{i=1}^{n} |b_i| \tag{1.13}$$

so let's assume that $p > 1$.

For a moment, assume positive a_i and b_i. We can do the following rearrangement:

$$\begin{aligned}
&\sum_{i=1}^{n} (a_i + b_i)^p \\
=&\sum_{i=1}^{n} (a_i + b_i)^{p-1} \cdot (a_i + b_i) \\
=&\sum_{i=1}^{n} a_i \cdot (a_i + b_i)^{p-1} + \sum_{i=1}^{n} b_i \cdot (a_i + b_i)^{p-1} && \text{distribute } a_i + b_i
\end{aligned}$$

Now, consider

$$\sum_{i=1}^{n} u_i v_i \leq \left(\sum_{i=1}^{n} u_i^p\right)^{\frac{1}{p}} \cdot \left(\sum_{i=1}^{n} v_i^q\right)^{\frac{1}{q}} \tag{1.14}$$

with $\frac{1}{p} + \frac{1}{q} = 1$ and $q = \frac{p}{p-1}$.

Take (1.14) and let $u_i = a_i$ and $v_i = (a_i + b_i)$. This gives

$$\begin{aligned}\sum_{i=1}^{n} a_i(a_i + b_i)^{p-1} &\leq \left(\sum_{i=1}^{n} a_i^p\right)^{\frac{1}{p}} \cdot \left(\sum_{i=1}^{n} (a_i + b_i)^{(p-1)q}\right)^{\frac{1}{q}} \\ &\leq \left(\sum_{i=1}^{n} a_i^p\right)^{\frac{1}{p}} \cdot \left(\sum_{i=1}^{n} (a_i + b_i)^p\right)^{\frac{1}{q}} \end{aligned} \tag{1.15}$$

Related, we have

$$\sum_{i=1}^{n} b_i(a_i + b_i)^{p-1} \leq \left(\sum_{i=1}^{n} b_i^p\right)^{\frac{1}{p}} \cdot \left(\sum_{i=1}^{n} (a_i + b_i)^p\right)^{\frac{1}{q}} \tag{1.16}$$

Therefore, we can put (1.15) and (1.16) together to say

$$\sum_{i=1}^{n} (a_i + b_i)^p \leq \left(\sum_{i=1}^{n} a_i^p\right)^{\frac{1}{p}} \cdot \left(\sum_{i=1}^{n} (a_i + b_i)^p\right)^{\frac{1}{q}} + \left(\sum_{i=1}^{n} b_i^p\right)^{\frac{1}{p}} \cdot \left(\sum_{i=1}^{n} (a_i + b_i)^p\right)^{\frac{1}{q}} \tag{1.17}$$

$$\leq \left[\left(\sum_{i=1}^{n} a_i^p\right)^{\frac{1}{p}} + \left(\sum_{i=1}^{n} b_i^p\right)^{\frac{1}{p}}\right] \cdot \left(\sum_{i=1}^{n} (a_i + b_i)^p\right)^{\frac{1}{q}} \tag{1.18}$$

$$\left(\sum_{i=1}^{n} (q_i + b_i)^p\right)^{p-\frac{1}{q}} \leq \left(\sum_{i=1}^{n} a_i^p\right)^{\frac{1}{p}} + \left(\sum_{i=1}^{n} b_i^p\right)^{\frac{1}{p}} \tag{1.19}$$

Therefore, $\nu_p(\mathbf{a} + \mathbf{b}) \leq \nu_p(\mathbf{a}) + \nu_p(\mathbf{b})$ for $p \geq 1$.

1.2.3 Special Cases of Norms

Let's consider $p = 1$:

$$\nu_1(\mathbf{a}) = (|a_1| + \ldots + |a_n|) \tag{1.20}$$

Equation (1.20) is the *Manhattan Norm*. Suppose we had a point (a_1, a_2). The distance from the origin to (a_1, a_2) would be $(a_1 + a_2)$.

$p = 2$ is another common norm. For $p = 2$, we have

$$\nu_2(\mathbf{a}) = \sqrt{a_1^2 + \ldots + a_n^2} \tag{1.21}$$

Assuming a point (a_1, a_2) the distance from the origin to (a_1, a_2) is $\sqrt{a_1^2 + a_2^2}$ – the normal Euclidean distance.

Now, what will happen to $\nu_p(\mathbf{a})$ as $p \to \infty$?

$$\begin{aligned}\lim_{p\to\infty} \nu_p(\mathbf{a}) &= \max_{1\le i\le n} |a_i| \left[\left(\frac{|a_1|}{\max|a_i|}\right)^p + \ldots + \left(\frac{|a_n|}{\max|a_i|}\right)^p\right]^{\frac{1}{p}} \\ &= \max_{1\le i\le n} |a_i|\end{aligned}$$

Given a point (a_1, a_2), $\nu_\infty(\mathbf{a})$ will be $\max\{a_1, a_2\}$ – whichever "leg" is longest. This is called the *Canberra Norm*.

1.2.4 Relationship Between Norms

Note that as p increases, $\nu_p(\mathbf{a})$ strictly decreases; p and $\nu_p(\mathbf{a})$ have an inverse relationship.

In other words, suppose we have two norms: $\nu_p(\mathbf{a})$ and $\nu_s(\mathbf{a})$. If $p > s$, the we have $\nu_p(\mathbf{a}) < \nu_s(\mathbf{a})$.

This also implies that for $p \ge 1$, $\nu_1(\mathbf{a})$ is an upper limit on the norm of $\mathbf{a}$, and $\nu_\infty(\mathbf{a})$ is a lower limit on the norm of $\mathbf{a}$.

Consider the log of ν_p

$$\ln \nu_p = \frac{\ln(|a_1|^p + \ldots + |a_n|^p)}{p}$$

For simplification, let $|a_i| = c_i$. This gives

$$\ln \nu_p = \frac{\ln(c_1^p + \ldots + c_n^p)}{p}$$

Now, let's take the first derivative of $\ln \nu_p$.

$$f'(p) = \frac{\frac{c_1^p \ln c_1 + \ldots + c_n^p \ln c_n}{c_1^p + \ldots c_n^p} \cdot p - \ln(c_1^p + \ldots + c_n^p)}{p^2}$$

We would like to prove that

$$\frac{c_1^p \ln c_1 + \ldots + c_n^p \ln c_n}{c_1^p + \ldots + c_n^p} \cdot p \le \ln(c_1^p + \ldots + c_n^p)$$

We can manipulate this a little:

$$\begin{aligned}&\frac{c_1^p \ln c_1 + \ldots + c_n^p \ln c_n}{c_1^p + \ldots + c_n^p} \cdot p \le \ln(c_1^p + \ldots + c_n^p) \\ =&\frac{p \cdot (c_1^p \ln c_1 + \ldots + c_n^p \ln c_n)}{c_1^p + \ldots + c_n^p} \le \ln(c_1^p + \ldots + c_n^p) && \text{move } p \text{ to numerator} \\ =&\frac{c_1^p \ln c_1^p + \ldots + c_n^p \ln c_n^p}{c_1^p + \ldots + c_n^p} \le \ln(c_1^p + \ldots + c_n^p) && \text{multiply } p \text{ through} \\ =&\frac{z_1 \ln z_1 + \ldots + z_n \ln z_n}{z_1 + \ldots + z_n} \le \ln(z_1 + \ldots + z_n) && \text{substitute } z_i = c_i^p \\ =&z_1 \ln z_1 + \ldots + z_n \ln z_n \le (z_1 + \ldots + z_n) \cdot \ln(z_1 + \ldots + z_n)\end{aligned}$$

It's pretty easy to see that the inequality holds for the last line.

1.2.5 Spheres in $\mathbb{R}^n$

Let $\mathbf{a}$ be a point, and let r be a distance. The *closed sphere* centered at $\mathbf{a}$ with radius r is given by

$$B_p(\mathbf{a}, r) = \{\mathbf{x} \in \mathbb{R}^n \mid \nu_p(\mathbf{a} - \mathbf{x}) \leq r\} \tag{1.22}$$

The *open sphere* centered at $\mathbf{a}$ with radius r is given by

$$C_p(\mathbf{a}, r) = \{\mathbf{x} \in \mathbb{R}^n \mid \nu_p(\mathbf{a} - \mathbf{x}) < r\} \tag{1.23}$$

The closed sphere includes the "surface" of the sphere; the open sphere does not.

For the sake of example, let $\mathbf{a} = (0, 0)$ and let $r = 1$.

For the 1-norm, we have

$$B_1((0,0), r) = \{(x_1, x_2) \in \mathbb{R}^2 \mid (x_1 + x_2) \leq 1\}$$

This sphere is shown in Figure 1.3.

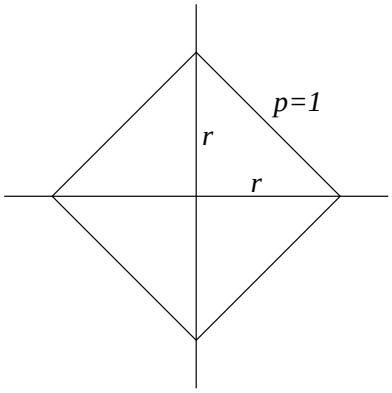

Figure 1.3: Sphere for $B_1(r)$

For the 2-norm, we have

$$B_2((0,0), r) = \{(x_1, x_2) \in \mathbb{R}^2 \mid \sqrt{x_1^2 + x_2^2} \leq 1\}$$

This is an ordinary circle, as shown in Figure 1.4

For the ∞-norm, we have

$$B_\infty((0,0), r) = \{(x_1, x_2) \in \mathbb{R}^2 \mid \max\{|x_1|, |x_2|\} \leq 1\}$$

This sphere is shown in Figure 1.5.

1.2.6 Notation

So far, we've used the notation $\nu_p(\mathbf{x})$ to denote the p-norm of $\mathbf{x}$. Another common notation is $\|\mathbf{x}\|_p$.

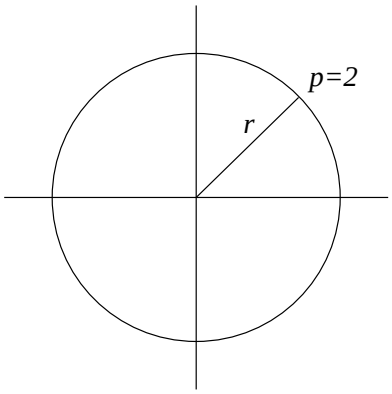

Figure 1.4: Sphere for $B_2(r)$

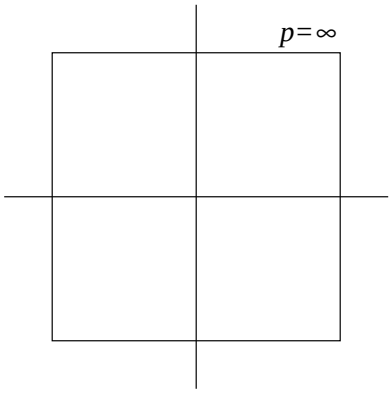

Figure 1.5: Sphere for $B_\infty(r)$

1.2.7 Distances and Metrics

An important application for norms is measuring the distance between patterns (i.e., vectors) in $\mathbb{R}^n$.

Suppose we have a set S. A *distance* is a mapping $d\colon S \times S \to \mathbb{R}_{\geq 0}$.

A distance that meets the following four conditions is called a *metric*.

1. $d(x, x) = 0$
2. If $d(x, y) = 0$, then $x = y$
3. $d(x, y) = d(y, x)$
4. $d(x, y) \leq d(x, z) + d(z, y)$

A distance that satisfies 1–3 (but not 4) is called a *dissimilarity*.

Dissimilarities can be tricky to work with – a distance that does not obey the triangular inequality can behave in non-intuitive ways.

Consider the three points in Figure 1.6. We can see that x is close to y, and we can see that x is close to z. From the Figure, it *looks* like y and z are close together, but a dissimilarity would allow y and z to be very far apart.

A simple, and very standard way of defining distances is

$$d(\mathbf{x}, \mathbf{y}) = \nu(\mathbf{x} - \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\| \tag{1.24}$$

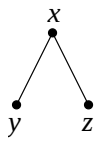

Figure 1.6: Three points: x, y, z

1.2.8 Misc

Our first set of handouts will be available sometime this week. Check the course web site.

1.3 A Quick Review of Linear Algebra Basics

A quick recap of some basic linear algebra:

The 2-norm is

$$\|\mathbf{x}\| = \sqrt{x_1^2 + \ldots + x_n^2}$$

The *dot-product* of two vectors $\mathbf{x}$ and $\mathbf{x}$ is

$$\mathbf{x} \cdot \mathbf{y} = (x_1 y_1) + (x_2 y_2) + \ldots + (x_n y_n)$$

$\mathbf{x}$ and $\mathbf{y}$ must have the same number of elements. Also note that the dot product is a scalar quantity.

The *angle* between two vectors $\mathbf{x}$ and $\mathbf{y}$ is

$$\cos\theta = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|}$$

So, $\cos\theta$ is the dot product of $\mathbf{x}$ and $\mathbf{y}$ divided by the product of the norms of the vectors.

The Cauchy-Schwarz Inequality is

$$|\mathbf{x} \cdot \mathbf{y}| \leq \|\mathbf{x}\| \cdot \|\mathbf{y}\|$$

Matrix multiplication is associative: $A(BC) = (AB)C$, for matrices A, B, C.

Matrix multiplication is not commutative: $AB \neq BA$ (in the general case).

Given a matrix A, the *inverse matrix* of A is A^{-1}. A^{-1} is a matrix such that $A(A^{-1}) = I$.

One method of finding A^{-1}:

- Augment A with the identity matrix I.
- Perform elementary row operations on A until you can turn A into I. The elementary row operations that turn A into I will turn I into A^{-1}.

The elementary row operations are:

- Multiply a row by a scalar
- Add one row to another
- Exchange two rows.

Example: suppose we'd like to find the inverse of

$$A = \begin{bmatrix} 2 & 3 \\ 5 & 8 \end{bmatrix}$$

We'd do this as follows:

$$
\begin{aligned}
&\left[\begin{array}{cc|cc} 2 & 3 & 1 & 0 \\ 5 & 8 & 0 & 1 \end{array}\right] \\
=&\left[\begin{array}{cc|cc} 10 & 15 & 5 & 0 \\ 10 & 16 & 0 & 2 \end{array}\right] && r_1 = r_1 * 5,\ r_2 = r_2 * 2 \\
=&\left[\begin{array}{cc|cc} 10 & 15 & 5 & 0 \\ 0 & 1 & -5 & 2 \end{array}\right] && r_2 = r_2 - r_1 \\
=&\left[\begin{array}{cc|cc} 10 & 15 & 5 & 0 \\ 0 & 15 & -75 & 30 \end{array}\right] && r_2 = r_2 * 15 \\
=&\left[\begin{array}{cc|cc} 10 & 0 & 80 & -30 \\ 0 & 15 & -75 & 30 \end{array}\right] && r_1 = r_1 - r_2 \\
=&\left[\begin{array}{cc|cc} 1 & 0 & 8 & -3 \\ 0 & 1 & -5 & 2 \end{array}\right] && r_1 = r_1/10,\ r_2 = r_2/15
\end{aligned}
$$

This tells us that

$$A^{-1} = \begin{bmatrix} 8 & -3 \\ -5 & 2 \end{bmatrix}$$

And we can verify that

$$AA^{-1} = \begin{bmatrix} 2 & 3 \\ 5 & 8 \end{bmatrix} \cdot \begin{bmatrix} 8 & -3 \\ -5 & 2 \end{bmatrix} = \begin{bmatrix} 16-15 & -6+6 \\ 40-40 & -15+16 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Given a 2×2 matrix A, the *determinant* of A is

$$\det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = (a_{11}a_{22}) - (a_{21}a_{12})$$

For a 3×3 matrix A, the determinant is

$$
\begin{aligned}
&\det \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \\
=&(a_{11}a_{22}a_{33}) + (a_{21}a_{32}a_{13}) + (a_{31}a_{12}a_{23}) - (a_{31}a_{22}a_{13}) - (a_{32}a_{23}a_{11}) - (a_{33}a_{21}a_{12})
\end{aligned}
$$

We're multiplying groups of three terms on each diagonal. We add products downhill from left to right, and subtract products uphill from left to right.

1.4 Lecture – 2/9/2009

Given a vector $\mathbf{a} = \{a_1, \ldots, a_n\}$, we have proved that $\nu_p(\mathbf{a}) = (|a_1|^p + \ldots + |a_n|^p)^{\frac{1}{p}}$ is a norm on $\mathbb{R}^n$ for $p \geq 1$.

Now, we'd like to turn our attention from vector norms to matrix norms.

1.4.1 Matrix Norms

Let A be a matrix in $\mathbb{R}^{m \times n}$ (where m is the number of rows, and n is the number of columns). We can also consider A as

$$A = \{a_1, \ldots, a_n\} \times \{a_1, \ldots, a_m\} \to \mathbb{R} \qquad \text{or,}$$
$$A(i, j) = a_{ij}$$

The first representation treats A as the "join" of two vectors; the second representation treats A as a function of two arguments.

Traditional linear algebra deals with matrices of real numbers. We can also deal with matrices of complex numbers, i.e., $A \in \mathbb{C}^{m \times n}$.

An *eigenvalue* for a matrix A is a number λ such that

$$A\mathbf{x} = \lambda\mathbf{x} \tag{1.25}$$

Traditionally, one finds eigenvalues by solving

$$\det(A - \lambda I_n) = 0 \tag{1.26}$$

Assuming that A is an $n \times n$ matrix, Equation (1.26) is a polynomial of degree n; there are n roots, and some of these roots can be complex. Thus for $A \in \mathbb{R}^{n \times n}$, we can have $\lambda \in \mathbb{C}$. However, if $A \in \mathbb{C}^{n \times n}$, then $\lambda \in \mathbb{C}$. In a sense, matrices of complex numbers are more general than matrices of real numbers.

1.4.2 Complex Numbers and Matrix Operations

A *complex number* z has the form $z = a + ib$. The *conjugate* of z is $\bar{z} = a - ib$. There is an equivalent *polar form* of z:

$$z = |z|(\cos\theta + i\sin\theta)$$
$$z = |z| \cdot e^{i\theta}$$

Given a matrix $A \in \mathbb{R}^{m \times n}$, A' is the *transpose* of A.

$$A'(ij) = A(ji) \tag{1.27}$$

The transpose turns rows into columns. If $A \in \mathbb{R}^{n \times m}$ then A' is in $\mathbb{R}^{m \times n}$.

A' works when A consists of real numbers. If A contains complex numbers, then the analogous operation is A^{H}. We form A^{H} from A by transposing A, and taking the conjugate of each a_{ij}. For example:

$$A = \begin{pmatrix} 1+i & 2 \\ 2-3i & 0 \end{pmatrix} \tag{1.28}$$

$$A^{\mathsf{H}} = \begin{pmatrix} 1-i & 2+3i \\ 2 & 0 \end{pmatrix} \tag{1.29}$$

A^{H} plays the same role for complex numbers that A' plays for reals.

Also, if $A = A'$, then A must be a square matrix. For example:

$$A = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$$
$$A' = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$$

In this case, we call A a *symmetric matrix*.

For complex numbers, we can have $A = A^{\mathsf{H}}$. If $A = A^{\mathsf{H}}$, then A is a *Hermitian Matrix* (named after the French mathematician Hermite).

A^{H} is a generalization of transpose. A^{H} works for reals as well as complex numbers (with reals, there's just no imaginary part).

1.4.3 Operations on Matrices

Of course, we can add two matrices $A + B$, or multiply a matrix by a scalar $a \cdot A$.

The set of matrices of a particular structure is a linear space. $\mathbb{C}^{m \times n}$ is a $\mathbb{C}$-linear space of dimension $m \times n$.

There is a mapping from matrices to vectors:

$$\mathbb{C}^{m \times n} \rightarrow \mathbb{C}^{mn} \tag{1.30}$$

This is called *vectorizing* the matrix. For example:

$$A = \begin{pmatrix} 1+i & 2 \\ 2-3i & 0 \end{pmatrix} \tag{1.31}$$
$$\mathrm{vec}(A) = \begin{pmatrix} 1+i & 2 & 2-3i & 0 \end{pmatrix} \tag{1.32}$$

Because we can map matrices to vectors, many vector manipulations can be applied to matrices. (But we'll see a few cases where this doesn't quite work out.)

1.4.4 Matrix Norms

Let A, B be two square matrices. A matrix norm should satisfy the properties of a vector norm, namely

1. $\|A\| \geq 0$
2. $\|A\| = 0$ iff $A = 0$
3. $\|A + B\| \leq \|A\| + \|B\|$
4. $\|\mathbf{a}B\| \leq \|\mathbf{a}\| \cdot \|B\|$

Additionally, we are interested in norms that satisfy a fifth property

5. $\mu(AB) \leq \mu(A) \cdot \mu(B)$

We'll use $|||A|||$ to denote a true matrix norm (one which satisfies all five properties).

Some (but not all) vector norms are matrix norms.

1-norm for Matrices

For a vector $\mathbf{a}$, $\nu_1(\mathbf{a}) = \sum_{i=1}^{n} |a_i|$. For a matrix $A \in \mathbb{R}^{m \times n}$, we define $\nu_1(A)$ as

$$\nu_1(A) = \sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}| \tag{1.33}$$

$\nu_1(A)$ is certainly a vector norm. Is it a matrix norm as well? In other words, is $\nu_1(AB) \leq \nu_1(A) \cdot \nu_1(B)$? We will show that $\nu_1(A)$ is also a matrix norm.

Let $A \in \mathbb{C}^{m \times n}$ and let $B \in \mathbb{C}^{n \times p}$. Then

$$\nu_1(AB) = \sum_{i=1}^{m} \sum_{j=1}^{p} \sum_{k=1}^{n} |a_{ik} b_{kj}| \tag{1.34}$$

We would like to prove that

$$\sum_{i=1}^{m} \sum_{j=1}^{p} \sum_{k=1}^{n} |a_{ik} b_{kj}| \leq \left(\sum_{n=1}^{m} \sum_{v=1}^{n} |a_{uv}| \right) \cdot \left(\sum_{s=1}^{n} \sum_{t=1}^{p} |b_{st}| \right) \tag{1.35}$$

First, we manipulate the left side of the inequality.

$$\sum_{i=1}^{m} \sum_{j=1}^{p} \sum_{k=1}^{n} |a_{ik} b_{kj}| \qquad a_{ik} \text{ and } b_{kj} \text{ are non-negative} \tag{1.36}$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{p} \sum_{k=1}^{n} |a_{ij}| \cdot |b_{kj}| \tag{1.37}$$

$$\leq \sum_{i=1}^{m} \sum_{j=1}^{p} \sum_{k',k''=1}^{n} |a_{ik'}| \cdot |b_{k''j}| \qquad \leq, \text{ because we add more } k\text{'s} \tag{1.38}$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{p} \sum_{k'=1}^{n} \sum_{k''=1}^{n} |a_{ik'}| \cdot |b_{k''j} \qquad \text{Because } k', k'' \text{ are independent} \tag{1.39}$$

$$= \sum_{i=1}^{m} \sum_{k'=1}^{n} |a_{ik'}| \cdot \sum_{j=1}^{p} \sum_{k''=1}^{n} |b_{k''j}| \qquad \text{Rearrange terms} \tag{1.40}$$

$$= \|A\| \cdot \sum_{j=1}^{p} \sum_{k''=1}^{n} |b_{k''j}| \qquad \text{By definition of } \|A\| \tag{1.41}$$

$$= \|A\| \cdot \|B\| \qquad \text{By definition of } \|B\| \tag{1.42}$$

$$\tag{1.43}$$

This proves (1.35). □

2-norm for Matrices

$\nu_2(A)$ also works as a matrix norm.

$$\nu_2(A) = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2} \tag{1.44}$$

$$\nu_2(AB) \leq \nu_2(A) \cdot \nu_2(B) \qquad \text{condition (5) for matrix norms} \tag{1.45}$$

$$\left(\sum_{i=1}^{m} \sum_{j=1}^{p} \left| \sum_{k=1}^{n} a_{ik} b_{kj} \right|^2 \right)^{\frac{1}{2}} \leq \left(\sum_{i=1}^{m} \sum_{k=1}^{n} |a_{ik}|^2 \right)^{\frac{1}{2}} \cdot \left(\sum_{k=1}^{n} \sum_{j=1}^{p} |b_{kj}|^2 \right)^{\frac{1}{2}} \tag{1.46}$$

Equation (1.46) can be shown with the Chaucy-Schwarz inequality. □

We can also prove (1.46) with the same kind of approach we used for $\nu_1(A)$. Let's start by expanding $\nu_2(AB) \leq \nu_2(A) \cdot \nu_2(B)$

$$\sqrt{\sum_{i=1}^{m} \sum_{j=1}^{p} \sum_{k=1}^{n} |a_{ik} b_{kj}|^2} \leq \sqrt{\sum_{i=1}^{m} \sum_{k=1}^{n} |a_{ik}|^2} \cdot \sqrt{\sum_{k=1}^{n} \sum_{j=1}^{p} |b_{kj}|^2} \tag{1.47}$$

We can combine the radicals in the right hand side of (1.47)

$$\sqrt{\sum_{i=1}^{m} \sum_{j=1}^{p} \sum_{k=1}^{n} |a_{ik} b_{kj}|^2} \leq \sqrt{\sum_{i=1}^{m} \sum_{k=1}^{n} |a_{ik}|^2 \cdot \sum_{k=1}^{n} \sum_{j=1}^{p} |b_{kj}|^2} \tag{1.48}$$

Next, let's work with the left side of (1.48)

$$\sqrt{\sum_{i=1}^{m} \sum_{j=1}^{p} \sum_{k=1}^{n} |a_{ik} b_{kj}|^2} \qquad \text{left side of (1.48)} \tag{1.49}$$

$$\leq \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{p} \sum_{k'k''=1}^{n} |a_{ik'} b_{k''j}|^2} \qquad \text{disassociate } k \text{ into } k' \text{ and } k'' \tag{1.50}$$

$$= \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{p} \sum_{k'=1}^{n} \sum_{k''=1}^{n} |a_{ik'} b_{k''j}|^2} \qquad \text{write } k', k'' \text{ summations explicitly} \tag{1.51}$$

$$= \sqrt{\sum_{i=1}^{m} \sum_{k'=1}^{n} |a_{ik'}|^2 \cdot \sum_{j=1}^{p} \sum_{k''=1}^{n} |b_{k''j}|^2} \qquad \text{rearrange terms} \tag{1.52}$$

(1.52) now the same as (1.47). □

∞-norm for Matrices

What about $\nu_\infty(A)$? $\nu_\infty(A)$ is *not* a matrix norm. We prove this by providing a counter-example.

Following our vector norm definition, $\nu_\infty(A)$ is

$$\nu_\infty(A) = \max_{i,j} |a_{ij}| \tag{1.53}$$

In other words, $\nu_\infty(A)$ is the largest absolute value in A.

Let A and B be

$$A = \begin{pmatrix} a & a \\ a & a \end{pmatrix}$$
$$B = \begin{pmatrix} b & b \\ b & b \end{pmatrix}$$
$$\nu_\infty(A) = a$$
$$\nu_\infty(B) = b$$

The product of A and B is

$$AB = \begin{pmatrix} a & a \\ a & a \end{pmatrix} \begin{pmatrix} b & b \\ b & b \end{pmatrix}$$
$$= \begin{pmatrix} 2ab & 2ab \\ 2ab & 2ab \end{pmatrix}$$
$$\nu_\infty(AB) = 2ab$$

ν_∞ is not a matrix norm because

$$\nu_\infty(AB) \not\leq \nu_\infty(A) \cdot \nu_\infty(B)$$
$$2ab \not\leq a \cdot b$$

To summary, we've shown that ν_1 and ν_2 are matrix norms, but ν_∞ is not a matrix norm.

1.4.5 Supremum and Maximum

What is the difference between a *supremum* and a *maximum*? What follows is an example to illustrate the difference; it's not a formal definition.

Suppose we have the function $f(x) = x^2$ where the domain of x is $x \in (0, 1)$. The domain of x is the open interval $(0, 1)$; so x can get arbitrarily close to zero, but we never have $x = 0$. Likewise, x can get arbitrarily close to one, but we never have $x = 1$.

Thus, $0 < x < 1$ and $0 < f(x) < 1$.

For this range of x, 1 is the least upper bound of $f(x) = x^2$. 1 is the supremum of $f(x)$. $f(x)$ can get arbitrarily close to 1, but we'll never have $f(x) = 1$.

By contrast, consider $g(x) = x^2$, where the domain of x is the closed interval $[0, 1]$. Here, we have $0 \leq x \leq 1$. We can have $x = 1$, and we can have $g(x) = 1$. In this case 1 is the maximum of $g(x)$.

1 is also the supremum of $g(x)$; there just happens to be a value of x that produces the supremum.

1.4.6 Back to Matrix Norms

Suppose $A \in \mathbb{C}^{m \times n}$, let $\nu(\mathbf{x})$ be a vector norm, and $\mu(A)$ is a matrix norm. Let $\mu(A)$ be

$$\mu(A) = \sup\{\nu(A\mathbf{x}) \mid \mathbf{x} \in \mathbb{C}^n, \mathbf{x} \neq 0, \text{ and } \nu(\mathbf{x}) \leq 1\} \tag{1.54}$$

Why is $\mu(A)$ compatible with our rule for matrix product: $\mu(AB) \leq \mu(A) \cdot \mu(B)$?

First, we note that

$$\begin{aligned}\nu(AB\mathbf{x}) &= \nu(AB\mathbf{x}) \\ &= \nu\left(A \cdot \frac{B\mathbf{x}}{\nu(B\mathbf{x})}\right) \cdot \nu(B\mathbf{x})\end{aligned}$$

$$\begin{aligned}\mu(AB) &= \sup\{\nu(AB\mathbf{x}) \mid \mathbf{x} \in \mathbb{C}^n \text{ and } \nu(\mathbf{x}) \leq 1\} \\ &= \max\{\nu(AB\mathbf{x}) \mid \mathbf{x} \in \mathbb{C}^n \text{ and } \nu(\mathbf{x}) = 1\} \\ &= \max\left\{\nu\left(A \cdot (\frac{B\mathbf{x}}{\nu(B\mathbf{x})})\right) \cdot \nu(B\mathbf{x}) \mid \nu(\mathbf{x}) = 1\right\} \\ &\leq \nu(A) \cdot \max\{\nu(B\mathbf{x}) \mid \nu(\mathbf{x}) = 1\} \\ &= \mu(A) \cdot \mu(B)\end{aligned}$$

1.4.7 Miscellany

- Download first handout from `http://www.cs.umb.edu/~dsim/`
- First homework assignment will probably be posted later this week.
- Dust of the calculus books and review complex numbers.
- Find out what a *spectral norm* is

1.5 A Review of Complex Numbers

This material is taken from Stewart's Calculus, 3rd Edition*, Appendix H, pages A46 – A52.*

A *complex number* has the form $z = a + bi$ where a and b are real numbers and $i = \sqrt{-1}$.

The point (a, b) can be plotted on a *Argand Plane*, which is a two dimension plane where the real part appears on the x-axis, and imaginary part appears on the y-axis. For example, the complex number $z = 1 + 2i$ can be plotted as the point $(1, 2)$, as shown in Figure 1.7.

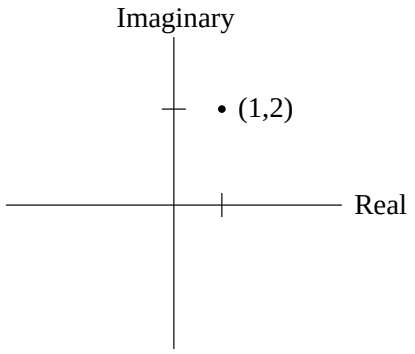

Figure 1.7: Plot of $(1, 2)$ for $z = 1 + 2i$

Given a pair of complex numbers, $z_1 = a + bi$ and $z_2 = c + di$ we say that $z_1 = z_2$ if $a = c$ and $b = d$. Two complex numbers are equal if their real parts are equal, and their imaginary parts are equal.

To add a pair of complex numbers:

$$\begin{aligned}(a + bi) + (c + di) &= a + bi + c + di \\ &= (a + c) + (b + d)i\end{aligned}$$

To multiply a complex number by a scalar:

$$k \cdot (a + bi) = ka + bki$$

To multiply a pair of complex numbers:

$$\begin{aligned}(a + bi) \cdot (c + di) &= ac + bci + adi + bdi^2 \\ &= ac + bci + adi - bd && \text{since } i^2 = -1 \\ &= (ac - bd) + (bc + ad)i\end{aligned}$$

The *conjugate* of a complex number $z = a + bi$ is given by $\overline{z} = a - bi$.

Properties of Conjugates:

$$\begin{aligned}\overline{z + w} &= \overline{z} + \overline{w} \\ \overline{zw} &= \overline{z} \cdot \overline{w} \\ \overline{z^n} &= \overline{z}^n\end{aligned}$$

To divide a pair of complex numbers, we multiply the numerator and denominator by the complex

conjugate of the denominator. Example:

$$\begin{aligned}\frac{-1+3i}{2+5i} &= \frac{-1+3i}{2+5i}\cdot\frac{2-5i}{2-5i}\\ &= \frac{-2+6i+5i-15i^2}{4+10i-10i-25i^2}\\ &= \frac{-2+11i+15}{4+25}\\ &= \frac{13+11i}{29}\end{aligned}$$

The *modulus* or *absolute value* of a complex number $z = a + bi$ is z's distance from the origin:

$$|z| = \sqrt{a^2+b^2}$$

1.5.1 Polar Forms of Complex Numbers

Figure 1.7 showed how a complex number z can be represented as a point. We can also represent z by an angle θ and a distance from the origin r.

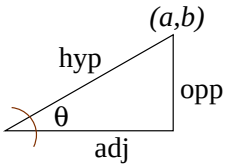

Figure 1.8: Triangle with angle θ

Recall that

$$\begin{aligned}\sin\theta &= \frac{\text{opp}}{\text{hyp}}\\ \cos\theta &= \frac{\text{adj}}{\text{hyp}}\\ \tan\theta &= \frac{\text{opp}}{\text{adj}}\end{aligned}$$

For complex numbers, a is "adj" and b is "opp", so

$$\begin{aligned}a &= r\cos\theta\\ b &= r\sin\theta\end{aligned}$$

and we can represent $z = a + bi$ as

$$\begin{aligned}z &= (r\cos\theta) + (r\sin\theta i)\\ &= r(\cos\theta + i\sin\theta)\end{aligned} \tag{1.55}$$

The angle θ is called the *argument* of z, and we write $\theta = \arg(z)$. Note that $\arg(z)$ is not unique; any two arguments of z differ by multiples of 2π.

Theorem 1.5.1 (DeMoivre's Theorem): If $z = r(\cos\theta + i\sin\theta)$ and n is a positive integer, then

$$\begin{aligned}z^n &= [r(\cos\theta + i\sin\theta)]^n\\ &= r^n(\cos n\theta + i\sin n\theta)\end{aligned}$$

1.5.2 Complex Exponentials

We can have e^z where z is a complex number.

e^z has the same properties the normal exponential function. In particular, $e^{z_1+z_2} = e^{z_1}e^{z_2}$.

If $z = iy$ and y is a real number, then

$$e^{iy} = \cos y + i \sin y$$

and

$$\begin{aligned} e^{x+iy} &= e^x e^{iy} \\ &= e^x(\cos y + i \sin y) \end{aligned}$$

1.6 Lecture – 2/11/2009

1.6.1 Matrix Norms

(In this section, assume we work with square $n \times n$ matrices.)

Let A be an $n \times n$ matrix, and let $\mathbf{x}$ be a vector $\in \mathbb{R}^n$.

If $\mu(A) = \max\{\nu(A\mathbf{x}) \mid \nu(\mathbf{x}) \leq 1\}$, then $\mu(A)$ would be compatible with the product condition for matrix norms: $\mu(AB) \leq \mu(A) \cdot \mu(B)$.

Consider the following two formulas:

$$S_1 = \{\mathbf{x} \in \mathbb{R}^n \mid \nu(\mathbf{x}) \leq 1\} \tag{1.56}$$
$$S_2 = \{\mathbf{x} \in \mathbb{R}^n \mid \nu(\mathbf{x}) = 1\} \tag{1.57}$$

Clearly $S_2 \subseteq S_1$; so, $\max(S_2) \leq \max(S_1)$.

One can think of S_1 as defining a sphere, and S_2 as defining the "skin" of the sphere (but not the sphere's interior).

We would like to prove that $\mu(A) = \max\{\nu(A\mathbf{x}) \mid \nu(\mathbf{x}) = 1\}$.

Note that

$$\max\{\nu(A\mathbf{x}) \mid \nu(\mathbf{x}) = 1\} \leq \max\{\nu(A\mathbf{x}) \mid \nu(\mathbf{x}) \leq 1\}$$

There is a $\mathbf{z}$ such that $\nu(\mathbf{z}) \leq 1$ and $\nu(A\mathbf{z}) = \nu(A)$.

$$\nu(A\mathbf{z}) = \nu(\mathbf{z}) \cdot \nu\left(A \cdot \frac{\mathbf{z}}{\nu(\mathbf{z})}\right) \tag{1.58}$$
$$\leq \nu\left(A \cdot \frac{\mathbf{z}}{\nu(\mathbf{z})}\right) \qquad \text{since } \nu(\mathbf{z}) \leq 1 \tag{1.59}$$
$$\leq \max\{\nu(A\mathbf{x}) \mid \nu(\mathbf{x}) = 1\} \tag{1.60}$$

Why is $\mu(AB) \leq \mu(A) \cdot \mu(B)$?

$$\mu(AB) = \max\{\nu(AB\mathbf{x}) \mid \nu(\mathbf{x}) \leq 1\} \qquad \text{By definition} \tag{1.61}$$
$$= \max\{\nu(A \cdot B\mathbf{x}) \mid \nu(\mathbf{x}) \leq 1\} \tag{1.62}$$
$$= \max\{\nu(B\mathbf{x}) \cdot \nu\left(A\frac{B\mathbf{x}}{\nu(B\mathbf{x})}\right) \mid \nu(\mathbf{x}) \leq 1\} \qquad \text{Note } \nu(A\mathbf{z}) \text{ in (1.58)} \tag{1.63}$$
$$\leq \mu(A) \cdot \max\{\nu(B\mathbf{x}) \mid \nu(\mathbf{x}) \leq 1\} \tag{1.64}$$

We get (1.64) from

$$\mu(A) = \max\{\nu(A\mathbf{x}) \mid \nu(\mathbf{x}) = 1\}$$

Finally, note that the following three definitions of $\mu(A)$ are equivalent:

$$\mu(A) = \max\{\nu(A\mathbf{x}) \mid \nu(\mathbf{x}) = 1\}$$
$$= \max\{\nu(A\mathbf{x}) \mid \nu(\mathbf{x}) \leq 1\}$$
$$= \sup\{\frac{\nu(A\mathbf{x})}{\nu(\mathbf{x})} \mid \nu(\mathbf{x}) \neq 0\}$$

1.6.2 1-Norm and ∞-Norm for Matrices

We'd like to look at the matrix norms $|||\cdot|||_1$ and $|||\cdot|||_\infty$.

If we treat a matrix like a vector, then we have

$$\|A\|_\infty = \max_{i,j} |a_{ij}| \tag{1.65}$$

(1.65) is a vector norm, but it is *not* a true matrix norm. (Doesn't meet the product condition.)

1-Norm for Matrices

Let's examine

$$|||A|||_1 = \max\{\|A\mathbf{x}\| \mid \|\mathbf{x}\| \le 1\} \qquad \text{Note: } A\mathbf{x} \text{ is a vector} \tag{1.66}$$

Let's treat A as a set of column vectors, so that $A = (\mathbf{a}_1, \ldots, \mathbf{a}_n)$. The product $A\mathbf{x}$ looks like this:

$$\begin{aligned} A\mathbf{x} &= \big(\mathbf{a}_1, \ldots, \mathbf{a}_n\big) \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \\ &= \mathbf{a}_1 x_1 + \mathbf{a}_2 x_2 + \ldots + \mathbf{a}_n x_n \end{aligned}$$

This allows us to write

$$\begin{aligned} |||A|||_1 &= \max\{\|\mathbf{a}_1 x_1 + \ldots \mathbf{a}_n x_n\|_1 \mid \|\mathbf{x}\| \le 1\} \\ &\le \max\{|x_1|\|\mathbf{a}_1\| + \ldots + |x_n|\|\mathbf{a}_n\| \mid \|\mathbf{x}\|_1 \le 1\} \end{aligned}$$

Because $\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i| \le 1$, we know that each $x_i \le 1$.

If we replace each $\|\mathbf{a}_i\|$ with $\max \|\mathbf{a}_i\|$, then we will get something larger. Therefore

$$|||A|||_1 \le \max_i \|\mathbf{a}_i\| \tag{1.67}$$

Next, given $|||A|||_1 = \max\{\|A\mathbf{x}\|_1 \mid \|\mathbf{x}\|_1 \le 1\}$, we would like to show that

$$\|\mathbf{a}_i\|_1 \le |||A|||_1 \qquad \text{for every } i \tag{1.68}$$

Let $\mathbf{e}_i$ be a column vector, having a 1 in the i-th position, and zeros everywhere else. The largest element in $\mathbf{e}_i$ is one, so $\|e_i\|_1 = 1$.

If we multiply $A\mathbf{e_i}$, the the result is the i-th column vector of A. Therefore $|||A|||_1 \ge \|A\mathbf{e}_i\|_1$.

In conclusion,

$$|||A|||_1 = \max_i \sum_{j=1}^n |a_{ji}| \tag{1.69}$$

(Note: a_{ji} not a typo)

$|||A|||_1$ is the largest sum of column vectors in A.

Equivalently, we can say that $|||A|||_1$ is the largest column norm of A.

∞-Norm for Matrices

We claim that

$$\|\!|A|\!\|_\infty = \max\{\|A\mathbf{x}\|_\infty \mid \|x\|_\infty \leq 1\} \tag{1.70}$$

The product $A\mathbf{x}$ is a vector, so

$$\|A\mathbf{x}\|_\infty = \max_i \left| \sum_{j=1}^{n} a_{ij} x_j \right| \tag{1.71}$$

$$\leq \max_i \sum_{j=1}^{n} |a_{ij}| \cdot |x_j| \tag{1.72}$$

$$\leq \max_i \sum_{j=1}^{n} |a_{ij}| \qquad \text{since all } |x_j| \leq 1. \tag{1.73}$$

Therefore, $\|\!|A|\!\|_\infty \leq \max_i \sum_{j=1}^{n} |a_{ij}|$.

Note that $\|\!|A|\!\|_\infty$ is the largest *row* norm of A.

Next, we'd like to show

$$\max_i \sum_{j=1}^{n} |a_{ij}| \leq \|\!|A|\!\|_\infty \tag{1.74}$$

For $A = 0$, this clearly works. So, assume that A contains some element that is > 0.

Let us focus on row $\mathbf{p}_i = (a_{p1}, a_{p2}, \ldots, a_{pn})$

We define a vector $\mathbf{z}$, such that

$$z_j = \begin{cases} \frac{|a_{pj}|}{a_{pj}} & \text{if } a_{pj} \neq 0 \\ 1 & \text{otherwise} \end{cases} \tag{1.75}$$

Each element of $\mathbf{z} \in \{-1, 1\}$. Therefore $\|\mathbf{z}\|_\infty = 1$.

If we sum on row p_i,

$$\sum_{j=1}^{n} |a_{pj}| = \sum_{j=1}^{n} |a_{pj} \cdot z_j| \tag{1.76}$$

$$\leq \left| \sum_{j=1}^{n} a_{pj} \cdot z_j \right| \tag{1.77}$$

$$\leq \max_i \left| \sum_{j=1}^{n} a_{pj} \cdot z_j \right| \tag{1.78}$$

$$\leq \|\!|A|\!\|_\infty \tag{1.79}$$

Therefore

$$\|\!|A|\!\|_\infty = \max \sum_{j=1}^{n} |a_{ij}| \qquad \boxed{\text{Largest sum of absolute values in rows}} \tag{1.80}$$

1.6.3 Matrices With Complex Components

For a matrix A of real numbers, A' is the transpose of A.

For a matrix A of complex numbers, A^{H} is the equivalent.

$A^{\mathsf{H}} = (\overline{a}_{ji})$ for $a_{ij} \in \mathbb{C}$.

A^{H} is called the *Hermitian adjoint* of A.

$A \in \mathbb{C}^{n \times n}$ is a Hermitian matrix if $A = A^{\mathsf{H}}$.

If all $a_{ij} \in \mathbb{R}$, then $A = A'$ means that A is a *symmetric matrix*.

The (column) vector $\mathbf{x} = (x_1, \ldots, x_n)^T$ is a matrix with n rows and one column. Therefore $x \in \mathbb{C}^{n \times 1}$. Similarly, the row vector $(x_1, \ldots, x_n) \in \mathbb{C}^{1 \times n}$.

Suppose we have two vectors $\mathbf{u}$, $\mathbf{v} \in \mathbb{C}^{n \times 1}$. ($\mathbf{uv}$ are column vectors.) The product

$$\mathbf{u}^{\mathsf{H}}\mathbf{v} = \overline{u}_1 v_i + \ldots + \overline{u}_n v_n$$

is a single number. $\mathbf{u}^{\mathsf{H}}\mathbf{v}$ is the unique singular value.

My contrast, the product

$$\mathbf{u}\mathbf{v}^{\mathsf{H}} = \begin{pmatrix} u_1\overline{v}_1 & u_1\overline{v}_2 & \ldots & u_1\overline{v}_n \\ u_2\overline{v}_2 & u_2\overline{v}_2 & \ldots & u_2\overline{v}_n \\ \vdots & \vdots & \ddots & \vdots \\ u_n\overline{v}_2 & u_n\overline{v}_2 & \ldots & u_n\overline{v}_n \end{pmatrix}$$

is an $n \times n$ matrix.

Let $A \in \mathbb{C}^{n \times n}$ be a matrix. The matrix

$$H_1 = \frac{A + A^{\mathsf{H}}}{2} \tag{1.81}$$

is like the real component of A (e.g., $z + \overline{z} = re(z)$). For example

$$\begin{aligned} A &= \begin{pmatrix} 1+1i & 2+2i \\ 3+3i & 4+4i \end{pmatrix} \\ A + A^{\mathsf{H}} &= \begin{pmatrix} 1+1i & 2+2i \\ 3+3i & 4+4i \end{pmatrix} + \begin{pmatrix} 1-1i & 3-3i \\ 2-2i & 4-4i \end{pmatrix} \\ &= \begin{pmatrix} 2 & 5-1i \\ 5-1i & 8 \end{pmatrix} \end{aligned}$$

Now consider H_2:

$$\begin{aligned} H_2 &= \frac{A - A^{\mathsf{H}}}{2i} \\ H_2^{\mathsf{H}} &= \frac{A^{\mathsf{H}} - A}{-2i} \end{aligned}$$

Both H_2 and H_2^{H} are Hermitian matrices.

For ordinary complex numbers, note that

$$\begin{aligned} z - \overline{z} &= (a + ib) - (a - ib) \\ &= 0a + 2ib \\ &= 2ib \end{aligned}$$

A matrix is *normal* if $A^{\mathsf{H}}A = AA^{\mathsf{H}}$. (Or, A is normal if $A = A^{\mathsf{H}}$.)

A matrix A is *unitary* if $A^{\mathsf{H}}A = AA^{\mathsf{H}} = I$.

Equation (1.82) is a common example of a normal matrix.

$$\begin{aligned} A &= \begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix} \\ A' &= \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix} \\ AA' &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \end{aligned} \tag{1.82}$$

Unitary matrices are a generalization of matrix rotation.

1.6.4 Misc

- hw1 is coming soon. Keep an eye on the course website
- No class on Monday (school holiday)
- Dust of a linear algebra book, and try to remember how matrix rotation works.

1.7 Matrix Miscellany

Notes from Introduction to Linear Algebra *by Géza Schay, pages 74–76 and 44–45.*

1.7.1 Invertible Matrices

Theorem 1.7.1: An $n \times n$ matrix is *invertible* if and only if $A\mathbf{x} = \mathbf{b}$ has a solution for every n-vector $\mathbf{b}$.

This implies two things

- Any single $\mathbf{b}$ has a unique solution.
- If there is a $\mathbf{b}$ for which $A\mathbf{x} = \mathbf{b}$ has no solution, then A is not invertible.

Theorem 1.7.2: An $n \times n$ matrix A is invertible if and only if $A\mathbf{x} = \mathbf{b}$ has a unique solution for some n-vector $\mathbf{b}$ and then also for all n-vectors $\mathbf{b}$.

A special case of this occurs when $\mathbf{b} = \mathbf{0}$:

Theorem 1.7.3: An $n \times n$ matrix is invertible if and only if $A\mathbf{x} = \mathbf{0}$ has only the trivial solution.

1.7.2 Singular Matrices

> A square matrix that does not have the nice properties of the theorems above is called *singular*.

Theorem 1.7.4: A $n \times n$ matrix is singular if and only if it has any (and then all) of the following properties:

1. A is not invertible.
2. The rank of A is less than n
3. A is not row equivalent to I
4. $A\mathbf{x} = \mathbf{b}$ has no solution for some $\mathbf{b}$.
5. Even if $A\mathbf{x} = \mathbf{b}$ has a solution for a given $\mathbf{b}$, that solution is not unique.
6. The homogenous equation $A\mathbf{x} = \mathbf{0}$ has nontrivial solutions.

Definition 1.7.5 (Rank of a Matrix): The number r of nonzero rows of an echelon matrix U obtained by the forward-elimination pahse of the Gaussian elimination algorithm from a matrix A is called the *rank* of A.

Definition 1.7.6 (Echelon Form): A matrix is said to be in *echelon form*, or to be an *echelon matrix* if it has a staircase-like pattern characterized by these properties:

1. The all-zero rows (if any) are at the bottom.
2. Calling the leftmost nonzero entry of each nonzero row a "leading entry", we have the leading entry in each lower row to the right of the leading entry in every higer row.
   (Put another way, all entries below a leading entry are zero.)

1.8 Scilab notes – 2/18/2009

`rref` turns a matrix into reduced echelon form. For example

```
-->A
 A  =
  - 1.  - 2.  - 1.  - 1.    1.
  - 1.  - 2.    0.    3.  - 1.
    1.    2.    1.    1.    1.
    0.    0.    2.    8.    2.

-->b
 b  =
    9.
    1.
  - 5.
  - 4.

-->rref([A b])
 ans  =
    1.    2.    0.  - 3.    0.  - 3.
    0.    0.    1.    4.    0.  - 4.
    0.    0.    0.    0.    1.    2.
    0.    0.    0.    0.    0.    0.
```

Back substitution gives one solution for $A\mathbf{x} = \mathbf{b}$: $\mathbf{x} = (-1, -1, -4, 0, 2)$. We can verify this as follows:

```
-->x
 x  =
  - 1.  - 1.  - 4.    0.    2.

-->A * x'
 ans  =
    9.
    1.
  - 5.
  - 4.
```

A simpler way to solve $A\mathbf{x} = \mathbf{b}$ is

```
-->x2 = A \ b
rank defficient. rank =     3

 x2  =
    0.
  - 3.
    0.
  - 1.
    2.

-->A * x2
 ans  =
    9.
    1.
  - 5.
  - 4.
```

1.8.1 A few R notes

R's `solve()` routine is limited to square matrices.

Basic matrix multiplication in R

```
> M <- matrix(nrow=4, ncol=5, byrow=T,
              c(-1, -2, -1, -1,  1,
                -1, -2,  0,  3, -1,
                 1,  2,  1,  1,  1,
                 0,  0,  2,  8,  2));
> M
     [,1] [,2] [,3] [,4] [,5]
[1,]   -1   -2   -1   -1    1
[2,]   -1   -2    0    3   -1
[3,]    1    2    1    1    1
[4,]    0    0    2    8    2

> b = c(9, 1, 5, -4)
> b
[1]  9  1  5 -4

> M %*% t(x)
     [,1]
[1,]    9
[2,]    1
[3,]   -5
[4,]   -4
```

`%*%` is R's matrix multiplication operator.

Part 2

Matrices

2.1 Lecture – 2/18/2009

The main topics of the lecture will be matrices, subspaces associated with matrices, and matrix rank.

2.1.1 Subspaces of $\mathbb{R}^n$ and $\mathbb{C}^n$

A *subspace* is a set $T \subseteq \mathbb{R}^n$ where T itself is a linear space.

Note 2.1.1: To give a concrete example, the set of vectors in $\mathbb{R}^3$ constitutes a linear space. The set of three-element vectors of the form (a, a, b) are a subspace of $\mathbb{R}^3$. Vectors of the form (a, a, b) are closed under addition, scalar multiplication, and they posses the other prerequisite requirements of being a subspace. □

Requirements for a subspace:

- if $\mathbf{x}$ and $\mathbf{y}$ belong to T, then $a\mathbf{x} + b\mathbf{y} \in T$, for all scalars a, $b \in \mathbb{R}$ (or in $\mathbb{C}$).

We can give an alternate form of this requirement:

- If $\mathbf{x}, \mathbf{y} \in T$, then $\mathbf{x} + \mathbf{y} \in T$. (Closed under addition)
- For all $\mathbf{x} \in T$, and for all $a \in \mathbb{R}$ (or $\in \mathbb{C}$), $a\mathbf{x} \in T$. (Closed under scalar multiplication).

$\mathbb{R}^n$ is a subspace. The smallest subspace of $\mathbb{R}^n$ is the origin, $\mathbf{0}^n$.

Suppose we have a set U, where $U \subseteq \mathbb{R}^n$. Note that we're just calling U a subset, not a subspace.

Let $\mathbf{x} = a_1\mathbf{u}_1 + \ldots + a_p\mathbf{u}_p$, so that $\mathbf{x}$ is a *linear combination* of $\mathbf{u}_1, \ldots, \mathbf{u}_p \in U$.

For any set U, let $\langle U \rangle$ denote the set of all linear combinations of U. $\langle U \rangle$ is a subspace of $\mathbb{R}^n$.

$\langle U \rangle$ is called the *span* of U.

Let

$$\mathbf{x} = a_1\mathbf{u}_1 + \ldots + a_p\mathbf{u}_p$$
$$\mathbf{y} = b_1\mathbf{v}_1 + \ldots + b_q\mathbf{v}_q$$

where $\mathbf{u}_1, \ldots, \mathbf{u}_p, \mathbf{v}_1, \ldots, \mathbf{v}_q \in T$.

If $\langle U \rangle$ coincides with $\mathbb{R}^n$, then we say that U *generates* $\mathbb{R}^n$.

2.1.2 Linear Independence

Definition 2.1.2 (Linear Independence): A set of vectors W is *linearly independent* if

$$a_1\mathbf{w}_1 + \ldots + a_p\mathbf{w}_p = \mathbf{0} \tag{2.1}$$

has only the solution $a_1 = a_2 = \ldots = a_p = 0$. If (2.1) has a solution where some $a_i \neq 0$, then W is not linearly independent, and we say that W is *linearly dependent*.

A set which is a generator for a space and is linearly independent is called a *basis*. In other words, a basis

- Must be a generator. The set of all linear combinations must generate the space.
- The members of the basis must be linearly independent.

Equivalently, we can say

- A basis must span the space. (minimally spanning)
- A basis must be linearly independent. (maximally independent).

2.1.3 Dimension of a Vector Space

Let V be a vector space. We denote the *dimension* of V as $\dim(V)$.

$\dim(V)$ is the cardinality of the base. For example $\dim(\mathbb{R}^n) = n$.

2.1.4 Basis Vectors

Let $\mathbf{e}_i$ be a vector with one in position i and zeros otherwise. The set of vectors $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ is a basis vector for $\mathbb{R}^n$.

How does this work? Consider

$$\begin{aligned}
\mathbf{e}_1 &= (1, 0, 0, 0, 0, \ldots) \\
\mathbf{e}_2 &= (0, 1, 0, 0, 0, \ldots) \\
\mathbf{e}_3 &= (0, 0, 1, 0, 0, \ldots) \\
\mathbf{e}_4 &= (0, 0, 0, 1, 0, \ldots) \\
&\ldots
\end{aligned}$$

Note that each $\mathbf{e}_i$ has a 1 in a different place. The only way to turn an $\mathbf{e}_i$ into the zero vector is to multiply $\mathbf{e}_i$ by zero. Thus, the set of $\mathbf{e}_i$ are linearly independent.

Now, why does the set of vectors $\mathbf{e}_i$ span $\mathbb{R}^n$? Let's look at a simple example in $\mathbb{R}^3$. Suppose we want to form the vector (a, b, c). We can form (a, b, c) through the linear combination

$$\begin{aligned}
&= a \cdot \mathbf{e}_1 + b \cdot \mathbf{e}_2 + c \cdot \mathbf{e}_3 \\
&= a \cdot (1, 0, 0) + b \cdot (0, 1, 0) + c \cdot (0, 0, 1) \\
&= (a, 0, 0) + (0, b, 0) + (0, 0, c) \\
&= (a, b, c)
\end{aligned}$$

for any values a, b, and c.

In any linear space, there is a basis.

Theorem 2.1.3: Let V be a vector space and let $\{\mathbf{b}_1, \ldots, \mathbf{b}_l\}$ be a set of linearly independent vectors. There is a basis B of V such that $\{\mathbf{b}_1, \ldots, \mathbf{b}_l\} \subseteq B$.

Therefore, one can build a set of basis vectors by extending a linearly independent set. During this process

- You maintain linear independence. (You have to, since the basis vectors must be linearly independent.)
- Generate the entire space. (You have to, since the basis must be a generator for the space.)

2.1.5 Null Spaces

Let A be a rectangular matrix, $A \in \mathbb{C}^{m \times n}$. We denote the *null space* of A as null(A).

$$\text{null}(A) = \{\mathbf{x} \in \mathbb{C}^n \mid A\mathbf{x} = \mathbf{0}\} \tag{2.2}$$

null(A) is the set of $\mathbf{x}$ vectors such that $A\mathbf{x}$ forms the zero vector.

null(A) is a subspace of $\mathbb{R}^n$.

Let $\mathbf{x}$ and $\mathbf{y}$ be two vectors such that $\mathbf{x}, \mathbf{y} \in \text{null}(A)$. We have

$$\begin{aligned} A\mathbf{x} &= \mathbf{0} \\ A\mathbf{y} &= \mathbf{0} \\ A\mathbf{x} + A\mathbf{y} &= A(\mathbf{x} + \mathbf{y}) = \mathbf{0} \\ A(a\mathbf{x}) &\in \text{null}(A) \end{aligned}$$

Therefore null(A) is a real subspace.

2.1.6 Range of a Matrix

We denote the *range* of the matrix A as range(A).

$$\text{range}(A) = \{A\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n\} \subseteq \mathbb{R}^m \tag{2.3}$$

Some texts call this the *image* of A, or Im(A).

The range is the set of vectors $\mathbf{y} \in \mathbb{R}^m$ that can be obtained by multiplying $A\mathbf{x}$ for any $\mathbf{x} \in \mathbb{R}^n$.

Note 2.1.4: range(A) seems like an analog to the range of a function $y = f(x)$. □

range(A) is also a subspace.

Let

$$\begin{aligned} \mathbf{u} &= A\mathbf{x} \qquad \text{and} \\ \mathbf{v} &= A\mathbf{y} \end{aligned}$$

Then

$$\begin{aligned} \mathbf{u} + \mathbf{v} &= A(\mathbf{x} + \mathbf{y}) \\ a\mathbf{u} &= A(a\mathbf{x}) \end{aligned}$$

The number of vectors in this subspace is equal to the *rank* of the matrix.

$$\text{rank}(A) = \dim(\text{range}(A))$$

The rank is also the number of linearly independent columns in A.

Note 2.1.5: For any matrix A, the number of linearly independent columns is equal to the number of linearly independent rows. □

Let's express the matrix A as a set of column vectors:

$$A = (\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_n)$$

Then

$$\begin{aligned} A\mathbf{x} &= (\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_n) \cdot \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \\ &= x_1\mathbf{c}_1 + x_2\mathbf{c}_2 + \ldots + x_n\mathbf{c}_n \end{aligned}$$

The range of a matrix is a subspace of the columns generated by the matrix.

2.1.7 Some Properties of Matrices

Let A be a matrix $A \in \mathbb{C}^{m \times n}$.

n is the dimension of the null space of A, plus the rank of A:

$$n = \dim(\text{null}(A) + \text{rank}(A)) \tag{2.4}$$

For all matrices $A \in \mathbb{R}^{m \times n}$, null$(A)$ is a subspace of $\mathbb{R}^n$. Therefore

$$\dim(\text{null}(A)) < n \qquad \text{is this } < \text{ or } \leq ?$$

(Note: n is the number of columns)

Theorem 2.1.6: Suppose $\mathbf{u}_1, \ldots, \mathbf{u}_k$ is a basis for null(A). This basis can be extended to a basis of $\mathbb{R}^n$. Therefore, $\dim(\text{null}(A)) = k$. (Note that this is similar to, but not the same as, Theorem 2.1.3.)

Let's look at the set of vectors

$$\mathbf{u}_1, \ldots, \mathbf{u}_k, \mathbf{u}_{k+1}, \ldots, \mathbf{u}_n$$

$\mathbf{u}_1, \ldots, \mathbf{u}_k$ is the basis for null(A) mentioned earlier; $\mathbf{u}_{k+1}, \ldots, \mathbf{u}_n$ are the extra vectors that, when combined with $\mathbf{u}_1, \ldots, \mathbf{u}_k$, will form a basis for $\mathbb{R}^n$. The extra vectors $\mathbf{u}_{k+1}, \ldots, \mathbf{u}_n$ are shaped by the matrix A, so that

$$A\mathbf{u}_{k+1}, \ldots, A\mathbf{u}_n \in \mathbb{R}^m \tag{2.5}$$

We have

$$n = \dim(\text{null}(A)) + \text{rank}(A) \tag{2.6}$$

where $\dim(\text{null}(A))$ comes from the first k vectors, and rank(A) comes from the last $n - k$ vectors.

The vectors $\mathbf{u}_{k+1}, \ldots, \mathbf{u}_n$ are a basis for the range of A.

Let $\mathbf{w} \in \text{range}(A) \subseteq \mathbb{R}^m$. $\mathbf{w}$ can be written as $A\mathbf{z}$ where $\mathbf{z} \in \mathbb{R}^n$, and $\mathbf{z}$ is a linear combination of the vectors $\mathbf{u}_i$.

$$\mathbf{z} = a_1\mathbf{u}_1 + a_2\mathbf{u}_2 + \ldots + a_k\mathbf{u}_k + a_{k+1}\mathbf{u}_{k+1} + \ldots + a_n\mathbf{u}_n$$

Given $\mathbf{w} = A\mathbf{z}$, we can expand $\mathbf{w}$

$$\mathbf{w} = A\mathbf{z} = a_1 A\mathbf{u}_1 + a_2 A\mathbf{u}_2 + \ldots a_k A\mathbf{u}_k + a_{k+1} A\mathbf{u}_{k+1} + \ldots + a_n A\mathbf{u}_n$$

Suppose

$$\beta_{k+1} A\mathbf{u}_{k+1} + \ldots + \beta_n A\mathbf{u}_n = \mathbf{0} \tag{2.7}$$

is a linear combination. If all $\beta_i = 0$, then $\mathbf{u}_{k+1}, \ldots, \mathbf{u}_n$ are linearly independent.

Also, if (2.7) holds, then so will

$$A(\beta_{k+1}\mathbf{u}_{k+1} + \ldots + \beta_n \mathbf{u}_n) = \mathbf{0} \tag{2.8}$$

Which means that

$$(\beta_{k+1}\mathbf{u}_{k+1} + \ldots + \beta_n \mathbf{u}_n) \in \text{null}(A)$$

Therefore

$$\alpha_1\mathbf{u}_1 + \ldots + \alpha_k\mathbf{u}_k + \beta_{k+1}\mathbf{u}_{k+1} + \ldots + \beta_n\mathbf{u}_n = \mathbf{0} \tag{2.9}$$

2.1.8 Matrix Rank

Let I_n be an $n \times n$ identity matrix. $\text{rank}(I_n) = n$. All columns of the identity matrix are linearly independent.

Let $\mathbf{u}$, $\mathbf{v}$ be two vectors, such that $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$.

$$\mathbf{u} = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}$$

$$\mathbf{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$$

The product

$$\mathbf{u}'\mathbf{v} = u_1 v_1 + u_2 v_2 + \ldots + u_n v_n$$

is a scalar.

However, the product

$$\mathbf{u}\mathbf{v}' = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} \cdot \begin{pmatrix} v_1 & \ldots & v_n \end{pmatrix} = \begin{pmatrix} u_1 v_1 & u_1 v_2 & \ldots & u_1 v_n \\ u_2 v_1 & u_2 v_2 & \ldots & u_2 v_n \\ \vdots & \vdots & \ddots & \vdots \\ u_n v_1 & u_n v_2 & \ldots & u_n v_n \end{pmatrix}$$

in a $n \times n$ matrix, and this $n \times n$ matrix has **rank one**.

Why does $\text{rank}(\mathbf{u}\mathbf{v}') = 1$? We can write $\mathbf{u}\mathbf{v}'$ as

$$\begin{pmatrix} v_1\mathbf{u} & v_2\mathbf{u} & \ldots v_n\mathbf{u} \end{pmatrix}$$

Each column of $\mathbf{u}\mathbf{v}'$ is a scalar multiple of the vector $\mathbf{u}$. Therefore, they cannot be linearly independent.

2.1.9 Sylvester's Theorem

Sylvester's Theorem tries to express the rank of the product of two matrices using the ranks of the individual matrices.

$$\text{rank}(AB) = \text{rank}(B) - \dim(\text{null}(A) \cap \text{range}(B)) \qquad \text{Sylvester's Theorem} \qquad (2.10)$$

Let's say that $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$. Then $\text{null}(A) \in \mathbb{R}^n$ and $\text{range}(B) \in \mathbb{R}^n$. Therefore, we can intersect $\text{null}(A) \cap \text{range}(B)$.

For any matrix, the number of linearly independent rows is equal to the number of linearly independent columns. Thus, for a matrix $B \in \mathbb{R}^{n \times p}$, we know that $\text{rank}(B) \leq \min(n, p)$.

If $\text{rank}(B) = \min(n, p)$, then we say that B has *full rank*.

Let $\mathbf{u}_1, \ldots, \mathbf{u}_k$ be a basis for $\text{null}(A) \cap \text{range}(B)$, and recall that $\text{null}(A) \cap \text{range}(B) \in \mathbb{R}^n$.

$\mathbf{u}_1, \ldots, \mathbf{u}_k \in \text{range}(B)$. The basis will have as many vectors as $\text{rank}(B)$.

Let $\mathbf{u}_1, \ldots, \mathbf{u}_k, \mathbf{u}_{k+1}, \ldots, \mathbf{u}_l$ be a basis for $\text{range}(B)$. We will have $\text{rank}(B) = l$. Therefore, we need to prove that $\text{rank}(AB) = l - k$.

From $\mathbf{u}_{k+1}, \ldots, \mathbf{u}_l \in \mathbb{R}^n$, we will be able to produce a basis for $\text{range}(AB)$.

Put another way, we need to prove that $A\mathbf{u}_{k+1}, \ldots, A\mathbf{u}_l$ is a basis for $\text{range}(AB)$.

Let

$$\begin{aligned}
\mathbf{u}_{k+1} &= B\mathbf{v}_{k+1} \\
\mathbf{u}_{k+2} &= B\mathbf{v}_{k+2} \\
&\vdots \\
\mathbf{u}_l &= B\mathbf{v}_l
\end{aligned}$$

where $\mathbf{v}_i$ are vectors in $\mathbb{R}^p$.

We can write

$$\begin{aligned}
&A\mathbf{u}_{k+1}, \ldots, A\mathbf{u}_l \\
=&AB\mathbf{v}_{k+1}, \ldots, AB\mathbf{v}_l
\end{aligned}$$

and all of these vectors are in $\text{range}(AB)$.

We need to show that

- These vectors are linearly independent, and
- These vectors generate $\text{range}(AB)$

We will finish this proof in our next lecture.

2.1.10 A Preview of things to Come

Any matrix can be written as a sum of rank one matrices. This is called Singular Value Decomposition (SVD).

SVD allows us to work with singular values, and the corresponding matrices of rank one. Therefore, we can reduce the size of the data set, without sacrificing accuracy.

Given messy data, many rank-one matrices represent noise; it is often beneficial to discard them.

2.2 Logistics

- We will have a makeup class at 10:00 am on Saturday 2/28/2009.
- hw1 is posted, along with the second course handout

2.3 Notes on Linear Spaces

These notes come from Chapter 4 of *Linear Algebra with Applications* 4th edition, by Gareth Williams, pub. Jones and Bartlett, 2001.

2.3.1 Definition of a Vector space

A vector space in $\mathbb{R}^n$ is a set of elements (vectors) for which two operations are defined: addition and multiplication. The space $\mathbb{R}^n$ is closed under these operations.[1]

Definition 2.3.1 (Vector Space): A vector space V satisfies the following conditions

$\mathbf{u}+\mathbf{v} \in V$	closure under addition
$c\mathbf{v} \in V$	closure under scalar multiplication
$\mathbf{u}+\mathbf{v} = \mathbf{v}+\mathbf{u}$	commutative properties
$\mathbf{u}+(\mathbf{v}+\mathbf{w}) = (\mathbf{u}+\mathbf{v})+\mathbf{w}$	
$\mathbf{u}+\mathbf{0} = \mathbf{u}$	
$\forall\ \mathbf{u} \in V, \exists\ (-\mathbf{u}) \mid \mathbf{u}+(-\mathbf{u}) = \mathbf{0}$	$\mathbf{u}$ has an additive inverse
$c(\mathbf{u}+\mathbf{v}) = c\mathbf{u}+c\mathbf{v}$	scalar multiplication axioms
$(c+d)\mathbf{u} = c\mathbf{u}+d\mathbf{u}$	
$c(d\mathbf{u}) = (cd)\mathbf{u}$	
$\mathbf{1u} = \mathbf{u}$	

□

For example, M_{22} the set of 2×2 matrices is a vector space.

2.3.2 Linear Subspaces

Definition 2.3.2 (subspace): Let V be a vector space and let U be a nonempty subset of V. U is said to be a *subspace* of V if U is closed under addition and scalar multiplication. □

Example 2.3.3: Consider elements in $\mathbb{R}^3$ of the form (a, a, b). This set is closed under addition and scalar multiplication. Therefore, it is a subspace.

Example 2.3.4: Consider elements of $\mathbb{R}^3$ of the form $(a, 0, 0)$. This set is also closed under addition and scalar multiplication. Therefore, it is a subspace.

Example 2.3.5: Consider elements of $\mathbb{R}^3$ of the form (a, a^2, b). This set is not closed under addition, so it is not a subspace. As a counter-example, $(2, 4, 3) + (1, 1, 2) = (3, 5, 5)$ and $(3, 5, 5) \notin (a, a^2, b)$.

Theorem 2.3.6: Let U be a subspace of the vector space V. U must contain the zero vector of V. (In other words, all $\mathbf{u} \in U$ must satisfy $\mathbf{0u} = \mathbf{0}$.) □

Example 2.3.7: Vectors of the form $(a, a, a+2)$ are not a subspace, since $(a, a, a+2)$ cannot contain the zero vector.

Definition 2.3.8 (Linear Combination): Let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be vectors in the vector space V. We say that $\mathbf{u} \in V$ is a *linear combination* of $\mathbf{v}_1, \ldots, \mathbf{v}_n$ if there exist scalars $c_1, \ldots, c_n$ such that

$$\mathbf{u} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \ldots c_n\mathbf{v}_n$$

[1] According to `http://en.wikipedia.org/wiki/Linear_space`, the term *vector space* and *linear space* mean the same thing.

□

The problem of determining when a vector $\mathbf{u}$ is a linear combination of $\mathbf{v}_1, \ldots, \mathbf{v}_n$ becomes a problem of solving linear equations.

Example 2.3.9: Is $(-1, 1, 5)$ a linear combination of $(1, 2, 3)$, $(0, 1, 4)$, $(2, 3, 6)$? To determine this, we solve the system of equations:

$$\begin{aligned} c_1 + 0c_2 + 2c_3 &= -1 \\ 2c_2 + c_2 + 3c_3 &= 1 \\ 3c_1 + 4c_2 + 6c_3 &= 5 \end{aligned}$$

In this case, there is a solution: $(c_1, c_2, c_3) = (1, 2, -1)$.

Note: the solution need not be unique.

Definition 2.3.10 (span): The set of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is said to *span* a vector space V if every element in V can be expressed as a linear combination of $\mathbf{v}_1, \ldots, \mathbf{v}_n$.

A *spanning set* of vectors defines the linear space V, since every member of V can be expressed in terms of the spanning set. □

Theorem 2.3.11: Let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be vectors in a vector space V, and let U be the set consisting of all linear combination of $\mathbf{v}_1, \ldots, \mathbf{v}_n$. Then U is a subspace spanned be the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$.

U is said to be the vector space *generated by* $\mathbf{v}_1, \ldots, \mathbf{v}_n$. (By definition, every vector in U can be written as a linear combination of $\mathbf{v}_1, \ldots, \mathbf{v}_n$. □

2.3.3 Linear Independence

Definition 2.3.12 (linear independence): The set of vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ in a vector space V is said to be *linearly dependent* if there exists scalars $c_1, \ldots, c_n$ (not all zero) such that

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \ldots + c_n\mathbf{v}_n = \mathbf{0}$$

The set of vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is *linearly independent* if

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \ldots + c_n\mathbf{v}_n = \mathbf{0}$$

can only be satisfied when $c_1 = c_2 = \ldots = c_n = 0$.

Example 2.3.13: Suppose we'd like to show that $\{(3, -2, 2), (3, -1, 4), (1, 0, 5)\}$ is linearly independent. To do this, we'd need to show that the system of equations

$$\begin{aligned} 3c_1 + 3c_2 + c_3 &= 0 \\ -2c_1 - c_2 + 0c_3 &= 0 \\ 2c_1 + 4c_2 + 5c_3 &= 0 \end{aligned}$$

Has a unique solution of $c_1 = c_2 = c_3 = 0$.

Similarly, if we wanted to show linear dependence in the set $\{(1, 2, 3), (-2, 1, 1), (8, 6, 10)\}$ we need to show that the system of equations

$$\begin{aligned} c_1 - 2c_2 + 8c_3 &= 0 \\ 2c_1 + c_2 + 6c_3 &= 0 \\ 3c_1 + c_2 + 10c_3 &= 0 \end{aligned}$$

has a solution with some $c_i \neq 0$.

Theorem 2.3.14: A set consisting of two or more vectors in a vector space is linearly dependent if and only if it is possible to express one of the vectors as a linear combination of the other vectors. □

The set $\{\mathbf{v}_1, \mathbf{v}_2\}$ is linearly dependent if and only if it is possible to write one vector as a scalar multiple of the other vector.

In $\mathbb{R}^2$, two vectors are linearly dependent if they lie on the same line.

The set $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ is linearly dependent if and only if it is possible to write one of the vectors as a linear combination of the other two vectors. For example, suppose $\mathbf{v}_1$ and $\mathbf{v}_2$ were linearly independent, but $\mathbf{v}_3$ was dependent on $\mathbf{v}_1$ and $\mathbf{v}_2$. $\mathbf{v}_1$ and $\mathbf{v}_2$ would define a plane, and $\mathbf{v}_3$ would lie inside that plane.

Theorem 2.3.15: Let V be a vector space. Any set of vectors in V that contains the zero vector is linearly dependent.

Consider the set $\{\mathbf{0}, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$. From this set, the identity

$$c_1\mathbf{0} + c_2\mathbf{v}_2 + \ldots + c_n\mathbf{v}_n = \mathbf{0}$$

has a solution with $c_1 = 1, c_2 = 0, \ldots c_n = 0$. Because the identity is true with some $c_i \neq 0$, the set of vectors is linearly dependent. □

Theorem 2.3.16: Let the set $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ be linearly dependent in a vector space V. Any set of vectors in V that contains these vectors will also be linearly dependent. □

This theorem can be proven as follows. Since the set $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is linearly dependent there is a solution to

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \ldots + c_n\mathbf{v}_n = \mathbf{0}$$

where some $c_i \neq 0$.

Now consider the set of vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_n, \mathbf{v}_{n+1}, \ldots, \mathbf{v}_p\}$. If we set $c_{n+1} \ldots c_p = 0$, then the equation

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \ldots + c_n\mathbf{v}_n + 0\mathbf{v}_{n+1} + \ldots 0\mathbf{v}_p = \mathbf{0}$$

also has a solution with some $c_i \neq 0$.

2.3.4 Bases and Dimensions

Definition 2.3.17 (basis): A finite set of vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is called a *basis* for a vector space V if the set spans V and is linearly independent.

Intuitively, a basis is a way of characterizing a vector space V, since any $\mathbf{v} \in V$ can be expressed as a linear combination of the basis vectors.

Definition 2.3.18 (standard basis): The set of vectors

$$\begin{aligned}
\mathbf{e}_1 &= (1, 0, 0, 0, \ldots, 0) \\
\mathbf{e}_2 &= (0, 1, 0, 0, \ldots, 0) \\
&\vdots \\
\mathbf{e}_n &= (0, 0, 0, 0, \ldots, 1)
\end{aligned}$$

is a basis for $\mathbb{R}^n$. This basis is called the *standard basis* for $\mathbb{R}^n$.

Theorem 2.3.19: Let $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ be a basis for a vector space V. If $\{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$ is a set of $> n$ vectors in V, then the set $\{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$ is linearly dependent. □

Above, note that any $\mathbf{w}$ can be expressed as a linear combination of vectors in $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ (since $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is a basis).

Theorem 2.3.20: Any two bases for a vector space V consist of the same number of vectors.

Definition 2.3.21 (dimension): If a vector space V has a basis consisting of n vectors, then the *dimension* of V is said to be n. We write $\dim(V)$ for the dimension of V. □

Theorem 2.3.22: Let V be a vector space of dimension n.

If $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is a set of linearly independent vectors in V, then S is a basis for V.

If $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is a set of n vectors that spans V, then S is a basis for V. □

2.4 Lecture - 2/23/2009

2.4.1 Rank of Matrices

Suppose we have a matrix $A \in \mathbb{C}^{m \times n}$. The null space, range, and rank of A are as follows.

$$\text{null}(A) = \{\mathbf{x} \in \mathbb{C}^n \mid A\mathbf{x} = \mathbf{0}\} \tag{2.11}$$
$$\text{range}(A) = \{A\mathbf{t} \mid \mathbf{t} \in \mathbb{C}^n\} \tag{2.12}$$
$$\text{rank}(A) \stackrel{\text{def}}{=} \dim(\text{range}(A)) \tag{2.13}$$

For $A \in \mathbb{C}^n$, we also have

$$n = \dim(\text{null}(A)) + \text{rank}(A) \tag{2.14}$$

2.4.2 Sylvester's Theorem

Sylvester's Theorem is

$$\text{rank}(AB) = \text{rank}(B) - \dim(\text{null}(A) \cap \text{range}(B)) \tag{2.15}$$

Let's prove (2.15).

Recall that $\text{rank}(B) = \dim(\text{range}(B))$.

Also note that $\text{null}(A) \in \mathbb{C}^n$ and $\text{range}(B) \in \mathbb{C}^n$. Therefore $\text{null}(A) \cap \text{range}(B)$ is a also a subspace of $\mathbb{C}^n$.

Figure 2.1 shows the relationship between the different linear spaces mentioned in Sylvester's Theorem.

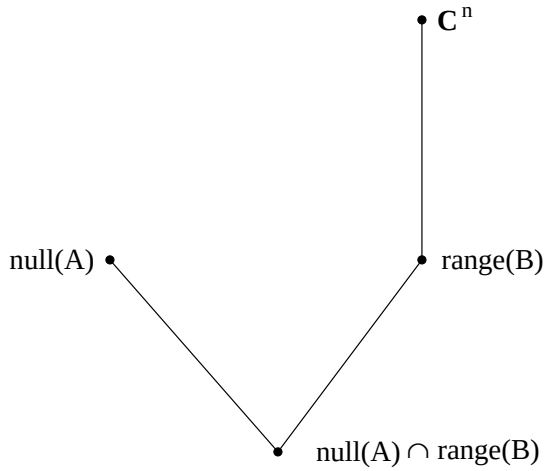

Figure 2.1: Linear Spaces in Sylvester's Theorem

Let's say that $\mathbf{u}_1, \ldots, \mathbf{u}_k$ is a basis for the linear space $\text{null}(A) \cap \text{range}(B)$. We can expand $\mathbf{u}_1, \ldots, \mathbf{u}_k$ to be a basis for $\text{range}(B)$:

$$\text{range}(B) = \mathbf{u}_1, \ldots, \mathbf{u}_k, \mathbf{u}_{k+1}, \ldots, \mathbf{u}_l$$

We have $\text{rank}(B) = l$, and $\dim(\text{null}(A) \cap \text{range}(B)) = k$. We would like to show that $\text{rank}(AB) = l - k$.

We would like to prove that range(AB) has a basis of $A\mathbf{u}_{k+1}, \ldots, A\mathbf{u}_l$. To do this, we need to show that (a) $A\mathbf{u}_{k+1}, \ldots, A\mathbf{u}_l$ is linearly independent and (b) $A\mathbf{u}_{k+1}, \ldots, A\mathbf{u}_l$ spans AB.

Claim: we claim that $A\mathbf{u}_{k+1}, \ldots, A\mathbf{u}_l$ is linearly independent.

Consider the equation

$$c_1 A\mathbf{u}_{k+1} + \ldots + c_{l-k} A\mathbf{u}_l = \mathbf{0} \tag{2.16}$$

If $A\mathbf{u}_{k+1}, \ldots, A\mathbf{u}_l$ is linearly independent, then (2.16) has only a trivial solution where $c_1 = c_2 = \ldots = c_{l-k} = 0$.

We can rearrange (2.16) as

$$A(c_1\mathbf{u}_{k+1} + \ldots + c_{l-k}\mathbf{u}_l) = \mathbf{0} \tag{2.17}$$

In (2.17), note that $c_1\mathbf{u}_{k+1} + \ldots + c_{l-k}\mathbf{u}_l = \text{null}(B)$.

So

$$c_1\mathbf{u}_{k+1} + \ldots + c_{l-k}\mathbf{u}_l = d_1\mathbf{u}_1 + \ldots + d_k\mathbf{u}_k \tag{2.18}$$
$$d_1\mathbf{u}_1 + \ldots + d_k\mathbf{u}_k - c_1\mathbf{u}_{k+1} - \ldots - c_{l-k}\mathbf{u}_l = \mathbf{0} \tag{2.19}$$

Therefore

$$d_1 = d_2 = \ldots = d_k = -c_1 = \ldots = -c_{l-k} = 0$$

Therefore $A\mathbf{u}_{k+1}, \ldots, A\mathbf{u}_l$ are linearly independent.

Claim: we claim that $A\mathbf{u}_{k+1}, \ldots, A\mathbf{u}_l$ spans AB.

Recall that $\mathbf{u}_{k+1}, \ldots, \mathbf{u}_l \in \text{range}(B)$. Therefore there is a set of vectors W such that

$$\begin{aligned}
\mathbf{u}_{k+1} &= B\mathbf{w}_1 \\
\mathbf{u}_{k+2} &= B\mathbf{w}_2 \\
&\vdots \\
\mathbf{u}_l &= B\mathbf{w}_{l-k}
\end{aligned}$$

Therefore, the following are all in range(AB)

$$\begin{aligned}
A\mathbf{u}_{k+1} &= AB\mathbf{w}_1 \\
A\mathbf{u}_{k+2} &= AB\mathbf{w}_2 \\
&\ldots \\
A\mathbf{u}_l &= AB\mathbf{w}_{l-k}
\end{aligned}$$

Let $\mathbf{t}$ be a vector $\mathbf{t} \in \text{range}(AB)$. We can write $\mathbf{t} = AB\mathbf{s}$ for some $\mathbf{s} \in \mathbb{C}^n$.

Since $\mathbf{t} = (AB)\mathbf{s}$, we have $\mathbf{t} = A(B\mathbf{s})$. Let

$$\begin{aligned}
B\mathbf{s} &= e_1\mathbf{u}_1 + \ldots + e_k\mathbf{u}_k + e_{k+1}\mathbf{u}_{k+1} + \ldots + e_l\mathbf{u}_l \\
AB\mathbf{s} &= e_1 A\mathbf{u}_1 + \ldots + e_k A\mathbf{u}_k + e_{k+1}A\mathbf{u}_{k+1} + \ldots + e_l A\mathbf{u}_l
\end{aligned} \tag{2.20}$$

In (2.20), note that $\mathbf{u}_1, \ldots, \mathbf{u}_k \in \text{null}(A)$. All of the $e_i A\mathbf{u}_i$ terms go to zero, leaving

$$AB\mathbf{s} = e_k A\mathbf{u}_k + e_{k+1}A\mathbf{u}_{k+1} + \ldots + e_l A\mathbf{u}_l \tag{2.21}$$

Therefore, $A\mathbf{u}_{k+1}, \ldots, A\mathbf{u}_l$ spans AB.

2.4.3 Matrix Inverses

Let A be a matrix $A \in \mathbb{C}^{n \times n}$. A^{-1} is the *inverse* of A. If A^{-1} exists, then we say that A is *invertible*.

Characteristics of invertible matrices:

- A is a square matrix
- A is non-singular
- $\det(A) \neq 0$.
- $\text{rank}(A) = n$. All columns of A are linearly independent, and all rows of A are linearly independent.
- $\text{null}(A) = \{\mathbf{0}\}$
- $\text{range}(A) = \mathbb{C}^n$

Suppose we have three matrices A, P, and Q:

$$A \in \mathbb{C}^{m \times n}$$
$$P \in \mathbb{C}^{m \times m}$$
$$Q \in \mathbb{C}^{n \times n}$$

Where P, Q are invertible. Then

$$\text{rank}(A) = \text{rank}(PA) = \text{rank}(AQ) = \text{rank}(PAQ)$$

By Sylvester's theorem

$$\text{rank}(PA) = \text{rank}(A) - \dim(\text{null}(P) \cap \text{range}(A))$$

Because P is invertible, we have $\dim(\text{null}(P) \cap \text{range}(A)) = \dim(\{\mathbf{0}\} \cap \text{range}(A)) = 0$. Therefore $\text{rank}(PA) = \text{rank}(A)$.

Now let's examine $\text{rank}(AQ)$.

$$\begin{aligned}\text{rank}(AQ) &= \text{rank}(Q) - \dim(\text{null}(A) \cap \text{range}(Q)) \\ &= \text{rank}(Q) - \dim(\text{null}(A)) \\ &= n - \dim(\text{null}(A)) \\ &= \text{rank}(A)\end{aligned}$$

Therefore, $\text{rank}(PAQ) = \text{rank}(AQ) = \text{rank}(A)$.

Multiplying a matrix A by an invertible matrix does not change its rank.

2.4.4 Frobenius Inequality

Let A, B, C be three conformant matrices

$$A \in \mathbb{C}^{m \times n}$$
$$B \in \mathbb{C}^{n \times p}$$
$$C \in \mathbb{C}^{p \times q}$$

What is $\text{rank}(ABC)$.

Let's start with two applications of Sylvester's Theorem

$$\text{rank}(ABC) = \text{rank}(BC) - \dim(\text{null}(A) \cap \text{range}(BC)) \tag{2.22}$$
$$\text{rank}(AB) = \text{rank}(A) - \dim(\text{null}(A) \cap \text{range}(B)) \tag{2.23}$$

We cross-multiply (2.22) and (2.23):

$$\begin{aligned}&\operatorname{rank}(ABC) + \operatorname{rank}(B) - \dim(\operatorname{null}(A) \cap \operatorname{range}(B)) \\ &\quad = \operatorname{rank}(AB) + \operatorname{rank}(BC) - \dim(\operatorname{null}(A) \cap \operatorname{range}(BC))\end{aligned}$$

$$\begin{aligned}&\operatorname{rank}(ABC) + \operatorname{rank}(B) \\ &\quad = \operatorname{rank}(AB) + \operatorname{rank}(BC) + \dim(\operatorname{null}(A) \cap \operatorname{range}(B)) - \dim(\operatorname{null}(A) \cap \operatorname{range}(BC))\end{aligned}$$

Both $\dim(\operatorname{null}(A) \cap \operatorname{range}(BC))$ and $\dim(\operatorname{null}(A) \cap \operatorname{range}(B))$ are ≥ 0.

Also, $\operatorname{range}(B) > \operatorname{range}(BC)$. Therefore

$$\operatorname{rank}(ABC) + \operatorname{rank}(B) \leq \operatorname{rank}(AB) + \operatorname{rank}(BC) \tag{2.24}$$

Equation (2.24) is the *Frobenius Inequality*.

2.4.5 Spectral Theory of Matrices

Say we have a matrix $A \in \mathbb{C}^{n \times n}$, and a vector $\mathbf{x} \in \mathbb{C}^n$.

We are looking for $\mathbf{x} \neq \mathbf{0}$ such that $A\mathbf{x} = \lambda\mathbf{x}$. For these $\mathbf{x}$, we have $A\mathbf{x}$ co-linear with $\mathbf{x}$.

We rearrange

$$\begin{aligned}A\mathbf{x} &= \lambda\mathbf{x} \\ A\mathbf{x} &= \lambda I_n \mathbf{x} \\ (A - \lambda I_n)\mathbf{x} &= \mathbf{0}\end{aligned}$$

(Above, I_n denotes the identity matrix of dimensions $n \times n$.)

Say we have a matrix D such that $D\mathbf{x} = \mathbf{0}$. If D is invertible, then $\mathbf{x} = \mathbf{0}$ is the only solution to $D\mathbf{x} = \mathbf{0}$.

Therefore, to have a non-trivial solution to $D\mathbf{x} = \mathbf{0}$, D must be non-invertible, and $\det(D) = 0$.

Going back to $(A - \lambda I)\mathbf{x} = \mathbf{0}$, we want $(A - \lambda I)$ to be a non-invertible matrix with $\det(A - \lambda I) = 0$.

Suppose A is

$$A = \begin{pmatrix} a_{11} & a_{21} \\ a_{21} & a_{22} \end{pmatrix}$$

Then

$$\begin{aligned}\det(A - \lambda I) &= \det\left(\begin{pmatrix} a_{11} & a_{21} \\ a_{21} & a_{22} \end{pmatrix} - \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}\right) && (2.25) \\ &= \det\begin{pmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{pmatrix} && (2.26) \\ &= (a_{11} - \lambda)(a_{22} - \lambda) - a_{12}a_{21} && (2.27) \\ &= \lambda^2 - \lambda(a_{11} + a_{22}) + a_{11}a_{22} - a_{12}a_{21} && (2.28)\end{aligned}$$

Equation (2.28) is a polynomial of degree two, so there are two solutions to λ.

$$\begin{aligned}\lambda_1 + \lambda_2 &= a_{11} + a_{22} \\ &= \det(A) \\ &= \operatorname{trace}(A) \qquad \textit{(why is this?)}\end{aligned}$$

The *trace* of a matrix A, denoted trace(A) is the sum of the diagonal elements $a_{11} + a_{22} + \ldots + a_{nn}$.

$$\sum_{i=1}^{n} \lambda_i = \text{trace}(A)$$
$$\prod_{i=1}^{n} \lambda_i = (-1)^n \cdot \det(A)$$

The λ_i are called *eigenvalues* and the $\mathbf{x}$ are called *eigenvectors* (that correspond to λ_i.

The study of eigenvalues and eigenvectors is called the *Spectral Theory of Matrices*.

The *spectrum* of A is

$$\text{spec}(A) = (\lambda_1, \ldots, \lambda_n) \tag{2.29}$$

2.4.6 Generalizing Eigenvalues and Eigenvectors

We've looked at eigenvalues and eigenvectors in terms of square matrices. Next, we'd like to generalize them to rectangular matrices.

Suppose we have $A \in \mathbb{C}^{m \times n}$. Then $A^{\mathsf{H}} \in \mathbb{C}^{n \times m}$.

We claim that $\text{rank}(A) = \text{rank}(A^{\mathsf{H}}A)$. (Note that $A^{\mathsf{H}}A \in \mathbb{C}^{n \times n}$.)

Note that we assumed A was rectangular. Therefore, we can "attach" square matrices to A, while still preserving rank(A).

Recall that rank is linked to the dimension of the nullspace

$$n = \text{rank}(A) + \dim(\text{null}(A))$$
$$n = \text{rank}(A^{\mathsf{H}}A) + \dim(\text{null}(A^{\mathsf{H}}A)) \qquad \text{since } A^{\mathsf{H}}A \text{ is square}$$

We only need to prove that $\text{null}(A) = \text{null}(A^{\mathsf{H}}A)$.

If $A\mathbf{x} = \mathbf{0}$ then $A^{\mathsf{H}}\mathbf{x} = \mathbf{0}$. Therefore $\text{null}(A) \subseteq \text{null}(A^{\mathsf{H}}A)$. Next, let us show the opposite direction.

Let $\mathbf{x} \in \text{null}(A^{\mathsf{H}}A)$.

$$A^{\mathsf{H}}A\mathbf{x} = \mathbf{0}$$
$$\mathbf{x}^{\mathsf{H}}A^{\mathsf{H}}A\mathbf{x} = 0 \qquad \text{this is a scalar zero}$$
$$(A\mathbf{x})^{\mathsf{H}}(A\mathbf{x}) = 0$$

Recall that

$$\mathbf{x} \cdot \mathbf{x}^T = x_1x_1 + x_2x_2 + \ldots + x_nx_n$$
$$= \|\mathbf{x}\|^2$$
$$\mathbf{x}^{\mathsf{H}} \cdot \mathbf{x} = \overline{x}_1x_1 + \overline{x}_2x_2 + \ldots + \overline{x}_nx_n$$
$$= \|\mathbf{x}\|^2$$

Therefore

$$\|A\mathbf{x}\|^2 = 0$$
$$A\mathbf{x} = \mathbf{0}$$

and $\mathbf{x} \in \text{null}(A)$.

Therefore $\text{null}(A) = \text{null}(A^{\mathsf{H}}A)$. So, $\text{rank}(A) = \text{rank}(A^{\mathsf{H}}A)$.

Say that A is a matrix $A \in \mathbb{C}^{n \times n}$.

The solution to $\det(A - \lambda I) = 0$ is a polynomial of degree n. We call n the *algebraic multiplicity* of λ, written $\text{algm}(\lambda, A)$.

The set $\{\mathbf{x} \mid A\mathbf{x} = \lambda\mathbf{x}\}$ is a subspace of $\mathbb{R}^n$. λ can change the magnitude of $\mathbf{x}$, but not the direction of $\mathbf{x}$.

This subspace is called the *invariant subspace* of λ in A. Denote this subspace by $S_A(\lambda)$.

$\dim(\S_A(\lambda))$ is the *geometric multiplicity* of λ in A; we denote this as $\text{geom}(\lambda, A)$.

The geometric multiplicity of λ in A is always smaller than the algebraic multiplicity of λ in A. We will prove this in our next lecture.

2.4.7 Logistics

- Start getting familiar with Scilab. In particular, experiment with the functions for computing eigenvalues and eigenvectors; functions for QR decomposition; and functions for Singular value decomposition.
- In the next few weeks, we will start to do experiments that will require these kinds of computations. One of these areas should be the topic of our in-class presentations.

2.5 Lecture – 2/25/2009

2.5.1 The Origin Space

The smallest linear space is $\{\mathbf{0}\}$ (the zero vector, which is the origin of $\mathbb{R}^n$). Let $V = \{\mathbf{0}\}$.

V is a perfectly good linear space – it's closed under addition and scalar multiplication. However, what is the basis of V?

The basis cannot contain vectors with non-zero elements: such vectors would generate points outside of V.

The basis also cannot contain $\mathbf{0}$, since $\mathbf{0}$ is not linearly independent. Consider:

$$c_1 \cdot \mathbf{x} = \mathbf{0} \tag{2.30}$$

If $\mathbf{x} = \mathbf{0}$, then this equation has non-trivial solutions. Therefore, the basis of V cannot contain $\mathbf{0}$.

As it turns out, the basis of V is $\emptyset$; which implies that $\dim(V) = 0$.

This seems a little odd, but it's an exception to the rule.

2.5.2 Singular Value Decomposition

The main topics of this lecture will be singular values and singular vectors of a matrix A. In general, we will assume that A is a rectangular matrix in $\mathbb{C}^{m \times n}$.

Definition 2.5.1 (Singular Value): We say that σ is a *singular value* of A if there are two vectors $\mathbf{x} \in \mathbb{C}^n$ and $\mathbf{y} \in \mathbb{C}^m$ such that

$$A\mathbf{x} = \sigma\mathbf{y} \qquad \text{and}$$
$$A^{\mathsf{H}}\mathbf{y} = \sigma\mathbf{x}$$

We say that $\mathbf{x}$ is the left singular vector that corresponds to σ.

We say that $\mathbf{y}$ is the right singular vector that corresponds to σ. □

Among other things, we would like to relate σ, $\mathbf{x}$, and $\mathbf{y}$ to eigenvalues and eigenvectors. Note that

$$A^{\mathsf{H}}A\mathbf{x} = \sigma A^{\mathsf{H}}\mathbf{y} \tag{2.31}$$
$$= \sigma^2\mathbf{x} \tag{2.32}$$

$\mathbf{x}$ is an eigenvector of AA^{H} that corresponds to σ^2.

If we have $A\mathbf{x} = \sigma\mathbf{y}$ and $A^{\mathsf{H}}\mathbf{y} = \sigma\mathbf{x}$, then $\mathbf{x}$ is an eigenvector of $A^{\mathsf{H}}A$.

Similarly, for $\mathbf{y}$ we have

$$AA^{\mathsf{H}}\mathbf{y} = \sigma A\mathbf{x}$$
$$\sigma^2\mathbf{y}$$

In summary

$$A^{\mathsf{H}}A\mathbf{x} = \sigma^2\mathbf{x}$$
$$AA^{\mathsf{H}}\mathbf{y} = \sigma^2\mathbf{y}$$

2.5.3 Hermetian Matrices

Let $B = A^{\mathsf{H}}A$. B is a square, $n \times n$ matrix. We have

$$\begin{aligned} B &= A^{\mathsf{H}}A \\ B^{\mathsf{H}} &= A^{\mathsf{H}}A \\ B &= B^{\mathsf{H}} \end{aligned}$$

Therefore, B is a Hermetian matrix. All values of a Hermetian matrix are real numbers.

If B consists only of reals, then B is a *symmetric matrix*.

Suppose we have $B\mathbf{w} = \lambda\mathbf{w}$. We would like to show that $\lambda \in \mathbb{R}$.

$$\begin{aligned} B\mathbf{w} &= \lambda\mathbf{w} \\ \mathbf{w}^{\mathsf{H}}B\mathbf{w} &= \lambda\mathbf{w}^{\mathsf{H}}\mathbf{w} \\ \mathbf{w}^{\mathsf{H}}B^{\mathsf{H}}\mathbf{w} &= \lambda\|\mathbf{w}\|^2 && \text{since } B = B^{\mathsf{H}} \\ \overline{\lambda}\|\mathbf{w}\|^2 &= \lambda\|\mathbf{w}\|^2 \end{aligned}$$

Therefore $\overline{\lambda} = \lambda$ and $\lambda \in \mathbb{R}$.

If a matrix is a Hermetian matrix, then the eigenvalues are real numbers.

2.5.4 Normal and Unitary Matrices

We are interested in matrices such that $AA^{\mathsf{H}} = A^{\mathsf{H}}A$. These are called *normal matrices*.

If $AA^{\mathsf{H}} = A^{\mathsf{H}}A = I_n$, then we say that A is a *unitary matrix*.

Suppose $A = (\mathbf{c}_1\ \mathbf{c}_2\ \ldots\ \mathbf{c}_n)$, where $\mathbf{c}_i$ are the columns in A. The Hermetian adjoint of A is

$$A^{\mathsf{H}} = \begin{pmatrix} \mathbf{c}_1^{\mathsf{H}} \\ \mathbf{c}_2^{\mathsf{H}} \\ \vdots \\ \mathbf{c}_n^{\mathsf{H}} \end{pmatrix} \tag{2.33}$$

If A is a unitary matrix, then

$$AA^{\mathsf{H}} = \begin{pmatrix} \mathbf{c}_1 & \mathbf{c}_2 & \ldots & \mathbf{c}_n \end{pmatrix} \cdot \begin{pmatrix} \mathbf{c}_1^{\mathsf{H}} \\ \mathbf{c}_2^{\mathsf{H}} \\ \vdots \\ \mathbf{c}_n^{\mathsf{H}} \end{pmatrix} = I_n \tag{2.34}$$

Therefore, for all column vectors $\mathbf{c}_i$:

$$\begin{aligned} \mathbf{c}_i \cdot \mathbf{c}^{\mathsf{H}} &= 1 \\ \|\mathbf{c}_i\|^2 &= 1 && \text{euclidean norm is one} \\ \|\mathbf{c}_i\| &= 1 && \text{1-norm is one} \end{aligned}$$

Let $\mathbf{c}_i$ and $\mathbf{c}_j$ be any two distinct columns of a unitary matrix A. We have $\mathbf{c}_i\mathbf{c}_j^{\mathsf{H}} = 0$. This tells us that $\mathbf{c}_i$ and $\mathbf{c}_j$ are orthogonal. All columns of A are perpendicular unit vectors.

2.5.5 Small Note on Vector norms

Given $\mathbf{a} \in \mathbb{C}^n$:

$$\|\mathbf{a}\| = \sqrt{|a_1^2| + |a_2^2| + \ldots + |a_n^2|} \tag{2.35}$$

$$\mathbf{a}^{\mathsf{H}}\mathbf{a} = \begin{pmatrix} \overline{a_1} & \overline{a_2} & \cdots & \overline{a_n} \end{pmatrix} \cdot \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} \tag{2.36}$$

$$= \|\mathbf{a}\|_2^2 \tag{2.37}$$

Therefore $\mathbf{a}^{\mathsf{H}}\mathbf{a} = \|\mathbf{A}\|_2^2$.

2.5.6 Singular Value Decomposition

Theorem 2.5.2: Let A be a matrix $A \in \mathbb{C}^{m \times n}$. There exists two unitary matrices U and V such that $U \in \mathbb{C}^{m \times m}$ and $V \in \mathbb{C}^{n \times n}$ such that $A = UDV^{\mathsf{H}}$, where D is a diagonal matrix $D = \text{diag}(\sigma_1, \ldots \sigma_p) \in \mathbb{C}^{m \times n}$. □

Assume $\sigma_1, \ldots, \sigma_p$ are singular values with

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p \geq 0$$

Definition 2.5.3 (Diagonal Matrix): A diagonal matrix $D = \text{diag}(\sigma_1, \ldots, \sigma_p)$ is a matrix where $d_{ii} = \sigma_i$, and all other elements are zero. □

Example 2.5.4: For example,

$$D = \begin{pmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \\ 0 & 0 & 0 \end{pmatrix}$$

is a 4×3 diagonal matrix.

Another example: the matrix

$$\text{diag}(3, 1) = \begin{pmatrix} 3 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

is a diagonal matrix in $\mathbb{C}^{3,5}$. □

Definition 2.5.5 (Full Rank): Let L be a matrix $L \in \mathbb{C}^{m \times n}$. If $\text{rank}(L) = \min(m, n)$ then L is a *full rank* matrix.

If $\text{rank}(L) < \min(m, n)$ then L is a *degenerate* matrix. □

In general, diagonal matrices are degenerate (i.e., they do not have full rank).

We say that $A = UDV^{\mathsf{H}}$ is a *singular value decomposition*.

For the diagonal matrix D, we have $\text{rank}(D) = p$. The rank is the number of nonzero elements in D.

Also, $\text{rank}(A) = p$. This is the number of non-zero singular values.

Unlike other forms of matrix decomposition, SVD works for *every* matrix A. To demonstrate this, we will give an inductive proof that SVD is possible for every $m \times n$ matrix A.

Let $q = \min(m, n)$, $q \geq 1$ be the smallest dimension of a matrix. If $q = 1$ then we have a matrix A with dimension $(m \times 1)$, $(1 \times n)$, or (1×1).

Base Case: $q = 1$.

For the base case of $q = 1$, we'll treat A as an $(m \times 1)$, single column matrix.

$$A = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix}$$

We would like to prove that every such A has a decomposition

$$A = UDV^{\mathsf{H}} \tag{2.38}$$

$$\begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix} = U \cdot \begin{pmatrix} \sigma_1 \\ \vdots \\ \end{pmatrix} \cdot v \qquad \text{note: } v \text{ is a scalar} \tag{2.39}$$

$\mathbf{a}^{\mathsf{H}}\mathbf{a}$ has the same singular value decomposition as $\mathbf{a}\mathbf{a}^{\mathsf{H}}$.

For a vector $\mathbf{a}$, $\mathbf{a}^{\mathsf{H}} \cdot \mathbf{a} = \|\mathbf{a}\|_2^2$. Therefore (2.39) becomes

$$\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix} = U \cdot \begin{pmatrix} \|\mathbf{a}\|_2 \\ 0 \\ 0 \\ \end{pmatrix} \cdot v \qquad \text{Since } \|a\|_2 = \sigma_1 \tag{2.40}$$

Next, let's take (2.40), and expand U:

$$\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix} = \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \dots & \mathbf{u}_m \end{pmatrix} \cdot \begin{pmatrix} \|\mathbf{a}\|_2 \\ 0 \\ 0 \\ \end{pmatrix} \cdot v \qquad U = (\mathbf{u}_1 \ \dots) \tag{2.41}$$

$$= \mathbf{u}_1 \cdot \|\mathbf{a}\| \cdot v \tag{2.42}$$

In (2.42), $\|\mathbf{a}\|$ is the only non-zero term in D (since D is a diagonal matrix). Therefore, when we multiply $U \cdot D$, all column vectors of U go to zero, with the exception of the first column.

For the base case, we can take $v = 1$, and $\mathbf{u}_1$ to be

$$\mathbf{u}_1 = \begin{pmatrix} \frac{a_1}{\|\mathbf{a}\|} \\ \frac{a_1}{\|\mathbf{a}\|} \\ \dots \\ \frac{a_m}{\|\mathbf{a}\|} \end{pmatrix} \tag{2.43}$$

This takes care of the base case. If A consists of a single column, then the singular value is the euclidean norm of the vector columns.

In general, we want $U = (\mathbf{u}_1 \ \mathbf{u}_2 \ \dots \ \mathbf{u}_m)$ to be perpendicular unit vectors.

Inductive Case: Let's move on to the inductive case, where A contains more than one column (or more than one row).

Take A such that $q = \min(m, n)$.

Let $\mathbf{u}_1$ be a unit vector that is the eigenvector of AA^{H}, with σ_1 as the eigenvalue of AA^{H}.

Let

$$\mathbf{v}_1 = \frac{1}{\sigma_1} A^{\mathsf{H}} \mathbf{u}_1$$

We claim that $(\mathbf{u}_1, \mathbf{v}_1)$ are a pair of singular vectors corresponding to the singular value σ_1.

We have

$$\begin{aligned} A\mathbf{v}_1 &= \frac{1}{\sigma_1} AA^{\mathsf{H}} \mathbf{u}_1 \\ &= \frac{1}{\sigma_1} \sigma_1^2 \mathbf{u}_1 \\ &= \sigma_1 \mathbf{u}_1 \end{aligned}$$

and

$$A^{\mathsf{H}} \mathbf{u}_1 = \sigma_1 \mathbf{v}_1 \qquad \text{note: } AA^{\mathsf{H}} \mathbf{u}_1 = \sigma_1^2 \mathbf{u}_1$$

$AA^{\mathsf{H}} \in \mathbb{C}^{m \times n}$, $\mathbf{u}_1 \in \mathbb{C}^m$, and $\mathbf{v}_1 \in \mathbb{C}^n$.

$\mathbf{u}_1$ and $\mathbf{v}_1$ will be the first rows of their respective unitary matrices.

Given matrices

$$U = \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \dots & \end{pmatrix} \tag{2.44}$$

$$V = \begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \dots & \end{pmatrix} \tag{2.45}$$

$$\tag{2.46}$$

Let U_1 be U with the $\mathbf{u}_1$ removed, and let V_1 be V with $\mathbf{v}_1$ removed:

$$U_1 = \begin{pmatrix} \mathbf{u}_2 & \dots & \end{pmatrix} \tag{2.47}$$

$$V_1 = \begin{pmatrix} \mathbf{v}_2 & \dots & \end{pmatrix} \tag{2.48}$$

$$\tag{2.49}$$

Given the construction of U_1 and V_1, we can write U, V as follows:

$$U = (\mathbf{u}_1 \; U_1) \tag{2.50}$$

$$V = (\mathbf{v}_1 \; V_1) \tag{2.51}$$

$$V^{\mathsf{H}} = \begin{pmatrix} \mathbf{v}_1^{\mathsf{H}} \\ V_1^{\mathsf{H}} \end{pmatrix} \tag{2.52}$$

Note that $D = U^{\mathsf{H}} AV$. We would like to find $U^{\mathsf{H}} AV$.

Given our constructs for U_1 and V_1, we can write $U^{\mathsf{H}} AV$ as

$$U^{\mathsf{H}} AV = \begin{pmatrix} \mathbf{u}_1^{\mathsf{H}} \\ U_1^{\mathsf{H}} \end{pmatrix} \cdot A \cdot \begin{pmatrix} \mathbf{v}_1 & V_1 \end{pmatrix} \qquad \text{separate } \mathbf{u}_1, \mathbf{v}_1. \tag{2.53}$$

$$= \begin{pmatrix} \mathbf{u}_1^{\mathsf{H}} \\ U_1^{\mathsf{H}} \end{pmatrix} \cdot \begin{pmatrix} A\mathbf{v}_1 & AV_1 \end{pmatrix} \qquad \text{Multiply } A, V \tag{2.54}$$

$$= \begin{pmatrix} \mathbf{u}_1^{\mathsf{H}} A\mathbf{v}_1 & \mathbf{u}_1^{\mathsf{H}} AV_1 \\ U_1^{\mathsf{H}} A\mathbf{v}_1 & U_1^{\mathsf{H}} AV_1 \end{pmatrix} \tag{2.55}$$

$A\mathbf{v}_1 = \sigma_1 \mathbf{u}_1$ and $AV_1 = \sigma_1 \mathbf{u}_1$.

As a result, (2.55) becomes

$$= \begin{pmatrix} \sigma_1 & 0 & \dots \\ 0 & U_1^{\mathsf{H}} A V_1 & \dots \\ \vdots & \vdots & \ddots \end{pmatrix} \tag{2.56}$$

In (2.56), with the exception of σ_1, the first row and first column are zero. The rest of the matrix consists of $U_1^{\mathsf{H}} A V_1$.

$U_1^{\mathsf{H}} A V_1$ is exactly like $U^{\mathsf{H}} A V$, but $U_1^{\mathsf{H}} A V_1$ has one less row and one less column. $U_1^{\mathsf{H}} A V_1$ are three matrices with dimensions

$$(m - 1 \times m) \cdot (m \times n) \cdot (n \times n - 1)$$

We continue SVD by taking the process that we applied to $U^{\mathsf{H}} A V$, and applying it to $U_1^{\mathsf{H}} A V_1$.

This completes the inductive step for SVD.

2.5.7 Logistics

- Make up class at 10:00 this Saturday. Meet outside our normal room.
- Look for a third handout (on SVD) in the not-too-distant future.
- Start thinking about presentation topics. Image clustering is a good application of SVD. Looking at the way Scilab implements SVD would be another good topic.

2.6 Lecture – 2/28/2009

2.6.1 Example of SVD

We'd like to do a singular value decomposition of the matrix A.

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

A is a 3×2 matrix. When we decompose $A = UDV^{\mathsf{H}}$, our matrices will have the following dimensions

$$\underset{(3\times 2)}{A} = \underset{(3\times 3)}{U} \cdot \underset{(3\times 2)}{D} \cdot \underset{(2\times 2)}{V^{\mathsf{H}}}$$

Note: (a) D has the same dimensions as A, and (b) U and V^{H} are square matrices, and conformant to multiplication with D.

First, we find AA^{H} and $A^{\mathsf{H}}A$:

$$AA^{\mathsf{H}} = \begin{pmatrix} 0 & 1 \\ 1 & 1 \\ 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} 0 & 1 & 1 \\ 1 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 1 \end{pmatrix} \qquad \text{used to find } U \tag{2.57}$$

$$A^{\mathsf{H}}A = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} 0 & 1 \\ 1 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \qquad \text{used to find } V \tag{2.58}$$

Next, we need to find the eigenvalues of AA^{H} and $A^{\mathsf{H}}A$.

Recall that for a matrix B, one finds eigenvalues by solving for $\det(B - \lambda I) = 0$. For AA^{H}, this is

$$\begin{aligned}
&= \det(AA^{\mathsf{H}} - \lambda I) \\
&= \det \begin{pmatrix} 1-\lambda & 1 & 0 \\ 1 & 2-\lambda & 1 \\ 0 & 1 & 1-\lambda \end{pmatrix} \\
&= (1-\lambda)(2-\lambda)(1-\lambda) + 0 + 0 - 0 - (1-\lambda) - (1-\lambda) \\
&= (2 - 2\lambda - \lambda + \lambda^2)(1-\lambda) - 1 + \lambda - 1 + \lambda \\
&= (2 - 3\lambda + \lambda^2 - 2\lambda + 3\lambda^2 - \lambda^3) - 2 + 2\lambda \\
&= (-\lambda^3 + 4\lambda^2 - 5\lambda + 2) - 2 + 2\lambda \\
&= (-\lambda^3 + 4\lambda^2 - 3\lambda) \\
&= -\lambda(\lambda^2 - 4\lambda + 3) \\
&= -\lambda(\lambda - 1)(\lambda - 3)
\end{aligned}$$

This gives $(\lambda_1, \lambda_2, \lambda_3) = (0, 1, 3)$.

Note: for a 3×3 matrix, we have three eigenvalues.

Next, we find eigenvalues for $A^{\mathsf{H}}A$.

$$\begin{aligned}
&= \det(A^{\mathsf{H}}A - \lambda I) \\
&= \det \begin{pmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{pmatrix} \\
&= (2-\lambda)(2-\lambda) - 1 \\
&= 4 - 4\lambda + \lambda^2 - 1 \\
&= \lambda^2 - 4\lambda + 3 \\
&= (\lambda - 1)(\lambda - 3)
\end{aligned}$$

Here $(\lambda_1, \lambda_2) = (1, 3)$.

In this case, $A^{\mathsf{H}}A$ has two eigenvalues and AA^{H} has three eigenvalues. The eigenvalues of AA^{H} are a superset of the eigenvalues of $A^{\mathsf{H}}A$. This will always be the case when $A^{\mathsf{H}}A$ and AA^{H} have different dimensions: the eigenvalues of the smaller matrix will be a subset of the eigenvalues of the larger matrix.

Next, we find the eigenvectors corresponding to the eigenvalues.

We find Eigenvectors by taking each eigenvalue λ, and solving for $(B - \lambda I)\mathbf{x} = \mathbf{0}$.

We are also interested in *unit eigenvectors*, so that $\|\mathbf{v}\|_2 = 1$.

For matrix AA^{H}:

- For $\lambda = 0$, an eigenvector (v_1, v_2, v_3) has $v_1 = v_3$ and $v_2 = -v_3$. A unit eigenvector of this form is $(\frac{\sqrt{3}}{3}, -\frac{\sqrt{3}}{3}, \frac{\sqrt{3}}{3})$.

  A way to check this in Scilab: `rref([(A * A') [0 0 0]'])`

- for $\lambda = 1$, an eigenvector (v_1, v_2, v_3) has the form $v_2 = 0$, and $v_1 + v_3 = 0$. A unit eigenvector of this form is $(\frac{\sqrt{2}}{2}, 0, -\frac{\sqrt{2}}{2})$.

  To check in Scilab: `rref( [((A * A') - eye(3,3)) [0 0 0]']`.

- for $\lambda = 3$, an eigenvector (v_1, v_2, v_3) has the form $v_1 = v_3$ and $v_2 = 2v_3$. A unit eigenvector of this form is $(\frac{\sqrt{6}}{6}, \frac{\sqrt{6}}{3}, \frac{\sqrt{6}}{6})$.

  To check in scilab: `rref( [((A * A') - 3*eye(3,3)) [0 0 0]'])`

For matrix $A^{\mathsf{H}}A$:

- For $\lambda = 1$, an eigenvector has $v_1 = -v_2$. A unit eigenvector of this form is $(\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2})$.

  Scilab: `rref([([A' * A] - 1 * eye(2,2)) [0 0]'])`

- For $\lambda = 3$, an eigenvector has the form $v_1 = v_2$. A unit eigenvector of this form is $(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})$.

  Scilab: `rref([([A' * A] - 3 * eye(2,2)) [0 0]'])`

The diagonal matrix D will be $\text{diag}(\sqrt{3}, \sqrt{1})$. D has square roots of the non-zero eigenvalues of AA^{H} and $A^{\mathsf{H}}A$ in descending order.

The matrix AA^{H} has dimensions 3×3. This becomes the basis for the matrix U. Below, the eigenvectors appear in columns.

$$\vec{\lambda} = \begin{pmatrix} 3 & 1 & 0 \end{pmatrix}$$

$$U = \begin{pmatrix} \frac{\sqrt{6}}{6} & \frac{\sqrt{2}}{2} & \frac{\sqrt{3}}{3} \\ \frac{\sqrt{6}}{3} & 0 & -\frac{\sqrt{3}}{3} \\ \frac{\sqrt{6}}{6} & -\frac{\sqrt{2}}{2} & \frac{\sqrt{3}}{3} \end{pmatrix}$$

The matrix $A^{\mathsf{H}}A$ is used to find V. The eigenvalues and eigenvectors for V (*not* V^{H}) are

$$\vec{\lambda} = \begin{pmatrix} 3 & 1 \end{pmatrix}$$
$$V = \begin{pmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{pmatrix}$$

The final decomposition is:

$$A = UDV^{\mathsf{H}}$$
$$A = \begin{pmatrix} \frac{\sqrt{6}}{6} & \frac{\sqrt{2}}{2} & \frac{\sqrt{3}}{3} \\ \frac{\sqrt{6}}{3} & 0 & -\frac{\sqrt{3}}{3} \\ \frac{\sqrt{6}}{6} & -\frac{\sqrt{2}}{2} & \frac{\sqrt{3}}{3} \end{pmatrix} \cdot \begin{pmatrix} \sqrt{3} & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{pmatrix}$$

If we multiply UDV^{H}, we do indeed get A.

2.6.2 SVD, Eigenvalues and Eigenvectors with Scilab

In Scilab, our previous example is trivial:

```
-->A
 A  =
    0.    1.
    1.    1.
    1.    0.

-->[U,D,V] = svd(A)
 V  =
  - 0.7071068  - 0.7071068
  - 0.7071068    0.7071068

 D  =
    1.7320508    0.
    0.           1.
    0.           0.

 U  =
  - 0.4082483    0.7071068    0.5773503
  - 0.8164966    7.456E-17  - 0.5773503
  - 0.4082483  - 0.7071068    0.5773503
```

Scilab chose different unit eigenvectors than we did (in this case, they differ only by sign). There's also a slight rounding error (`7.456E-17` instead of zero).

Scilab also makes it easy to compute eigenvalues and eigenvectors.

```
-->[x1, x2] = spec(A * A')
 x2  =
    1.928E-16    0.    0.
    0.           1.    0.
    0.           0.    3.

 x1  =
    0.5773503  - 0.7071068    0.4082483
  - 0.5773503  - 1.367E-17    0.8164966
    0.5773503    0.7071068    0.4082483

-->[x1, x2] = spec(A' * A)
 x2  =
    1.    0.
    0.    3.

 x1  =
  - 0.7071068    0.7071068
    0.7071068    0.7071068
```

In Scilab, `A'` appears to compute the Hermetian adjoint when given a matrix $A \in \mathbb{C}^{m \times n}$.

2.6.3 Some properties of D

Let's examine some properties of the diagonal matrix D.

We already know that

- D is a diagonal matrix
- Elements along the diagonal appear in descending order
- D has the same dimensions as A

Suppose we multiply D and a vector $\mathbf{x} \in \mathbb{R}^n$. Assume that D has p non-zero elements on the diagonal, and that $p \leq n$. The product $D\mathbf{x}$ has the form

$$D\mathbf{x} = \begin{pmatrix} \sigma_1 & 0 & \dots & 0 & \dots & 0 \\ 0 & \sigma_2 & & & & 0 \\ \vdots & & \ddots & & & 0 \\ 0 & & & \sigma_p & & 0 \\ \vdots & & & & \ddots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

$$= \begin{pmatrix} \sigma_1 x_1 \\ \sigma_2 x_2 \\ \vdots \\ \sigma_p x_p \end{pmatrix}$$

The matrix 2-norm of D is

$$|||D|||_2 = \max\{\|D\mathbf{x}\|_2 \mid \|x\|_2 = 1\}$$

The largest element of D is $d_{11} = \sigma_1$, so $|||D|||_2 = \sigma_1$.

Since D is diagonal, note that $\|D\mathbf{x}\|_2$ is

$$\|D\mathbf{x}\|_2 = \sqrt{\sigma_1 x_1 + \sigma_2 x_2 + \ldots + \sigma_p x_p}$$

Frobenius norm of D is

$$\|D\|_F = \sqrt{\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_p^2}$$

Normally, the Frobenius norm would be computed as

$$\|D\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2}$$

This computation is simplified for D, since $\{\sigma_1, \ldots, \sigma_p\}$ are the only non-zero elements.

(Note: the notations $\|D\|_F$ and $\|D\|_2$ mean the same thing: a 2-norm of the vectorization of D.)

$\|\!|A|\!\|_2 = \sigma_1$ is the largest singular value for A. This is called the *spectral radius* of A.

$\|D\|_F = \sqrt{\sigma_1^2 + \sigma_2^2 + \ldots \sigma_p^2}$. These are the non-zero singular values of A.

2.6.4 Norms of Unitary Matrices

U is a unitary matrix when

$$UU^{\mathsf{H}} = U^{\mathsf{H}}U = I_n$$

For unitary matrices,

$$\|\!|UA|\!\|_2 = \|\!|A|\!\|_2$$
$$\|UA\|_F = \|A\|_F$$

For a unitary matrix U:

$$(UA)_{ij} = \sum_{k=1}^{n} u_{ik}a_{kj}$$

This is *not* true for matrices in general. It works for unitary matrices, because $\sum_{i=1}^{n} u_{ik}u_{kj} = 0$ if $k \neq j$.

2.6.5 Note about Norm Notations

- $\|\mathbf{v}\|_2$ is the Euclidean norm of the vector $\mathbf{v}$.
- $\|A\|_F = \|A\|_2$ is the Euclidean norm applied to the vectorization of the matrix A.
- $\|\!|A|\!\|_2 = \max\{\|A\mathbf{x}\|_2 \mid \|\mathbf{x}\|_2 = 1\}$ is a true matrix norm. The RHS of this equation is a definition. We call this the "matrix norm induced by $\|\cdot\|_2$."

2.6.6 Some properties of UDV^{H}

Let's say we have computed $A = UDV^{\mathsf{H}}$. The result has the following form:

$$A = UDV^{\mathsf{H}} \tag{2.59}$$

$$= \begin{pmatrix} u_{11} & \dots & u_{1m} \\ \vdots & \ddots & \vdots \\ u_{m1} & \dots & u_{mm} \end{pmatrix} \cdot \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \dots & \dots & \sigma_p \end{pmatrix} \cdot \begin{pmatrix} \overline{v_{11}} & \dots & \overline{v_{n1}} \\ \vdots & \ddots & \vdots \\ \overline{v_{1n}} & \dots & \overline{v_{nn}} \end{pmatrix} \tag{2.60}$$

$$= \begin{pmatrix} u_{11} & \dots & u_{1m} \\ \vdots & \ddots & \vdots \\ u_{m1} & \dots & u_{mm} \end{pmatrix} \cdot \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \dots & \dots & \sigma_p \end{pmatrix} \cdot \begin{pmatrix} \mathbf{v}_1^{\mathsf{H}} \\ \mathbf{v}_2^{\mathsf{H}} \\ \vdots \\ \mathbf{v}_n^{\mathsf{H}} \end{pmatrix} \quad \text{treat } V^{\mathsf{H}} \text{ as row vectors} \tag{2.61}$$

$$= \begin{pmatrix} u_{11} & \dots & u_{1m} \\ \vdots & \ddots & \vdots \\ u_{m1} & \dots & u_{mm} \end{pmatrix} \cdot \begin{pmatrix} \sigma_1 \mathbf{v}_1^{\mathsf{H}} \\ \sigma_2 \mathbf{v}_2^{\mathsf{H}} \\ \vdots \\ \sigma_p \mathbf{v}_p^{\mathsf{H}} 0 \\ \vdots \\ 0 \end{pmatrix} \quad \text{Find } DV^{\mathsf{H}} \tag{2.62}$$

$$= \begin{pmatrix} \mathbf{u}_1 & \dots & \mathbf{u}_n \end{pmatrix} \cdot \begin{pmatrix} \sigma_1 \mathbf{v}_1^{\mathsf{H}} \\ \sigma_2 \mathbf{v}_2^{\mathsf{H}} \\ \vdots \\ \sigma_p \mathbf{v}_p^{\mathsf{H}} 0 \\ \vdots \\ 0 \end{pmatrix} \quad \text{Treat } U \text{ as column vectors} \tag{2.63}$$

$$= \sigma_1 \mathbf{u}_1 \mathbf{v}_1^{\mathsf{H}} + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^{\mathsf{H}} + \ldots + \sigma_p \mathbf{u}_p \mathbf{v}_p^{\mathsf{H}} \tag{2.64}$$

In (2.62), the zero elements of D simplify the computation of DV^{H}.

Given a column vector $\mathbf{u}$ and a row vector $\mathbf{v}^{\mathsf{H}}$, the product $\mathbf{u}\mathbf{v}^{\mathsf{H}}$ is a matrix (of rank one).

Line (2.64) is really the essence of singular value decomposition. It's a set of rank one matrices, each of which is scaled by some σ_i; and when we add them together, we recover the original matrix A.

2.7 Lecture – 3/4/2009

2.7.1 Singular Value Decomposition

Today, we'll take another look at the SVD example from the last class, and examine a few points in detail.

Given the matrix

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

we would like to find the singular value decomposition $A = UDV^{\mathsf{H}}$, where U and V are *unitary* matrices, and D is a diagonal matrix.

The dimensions of these matrices are

$$\underset{m \times n}{A} = \underset{m \times m}{U} \cdot \underset{m \times n}{D} \cdot \underset{n \times n}{V^{\mathsf{H}}}$$

Let's begin by making a few observations. Given matrix $M \in \mathbb{C}^{p \times q}$ and $M^{\mathsf{H}} \in \mathbb{C}^{q \times p}$, the matrix products MM^{H} and $M^{\mathsf{H}}M$ have the same non-zero eigenvalues. The square roots of these eigenvalues will become the diagonal of our matrix D (the $\sigma_1, \ldots, \sigma_p$ terms).

With respect to A,

- the columns of V are the eigenvectors of $A^{\mathsf{H}}A$
- the columns of U are the eigenvectors of AA^{H}

As we saw last time,

$$AA^{\mathsf{H}} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 1 \end{pmatrix} \qquad (\lambda_1, \lambda_2, \lambda_3) = (3, 1, 0)$$

$$A^{\mathsf{H}}A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \qquad (\lambda_1, \lambda_2) = (3, 1)$$

There are many eigenvectors for any given eigenvalue. Recall that

$$\{\mathbf{y} \mid Y\mathbf{y} = \lambda\mathbf{y}\}$$

is a linear space (an eigenspace?).

We want to construct U and V from unit eigenvectors. These unit eigenvectors are not unique. Thus, we will often have options when selecting eigenvectors, and we'll need to select them such that $A = UDV^{\mathsf{H}}$ holds.

Let's look at an example to illustrate this.

From AA^{H}, two eigenvalues give us four different unit eigenvectors:

$$\mathbf{v}_1 = \alpha_1 \cdot \begin{pmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{pmatrix}$$

$$\mathbf{v}_2 = \alpha_2 \cdot \begin{pmatrix} \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{pmatrix}$$

For α_1, $\alpha_2 \in \{-1, 1\}$.

Our choice of eigenvectors for V (i.e., our choice of $\alpha_i = 1$ or $\alpha_1 = -1$) will influence the set of eigenvectors used to construct U. These eigenvectors *must* be chosen so that $A = UDV^{\mathsf{H}}$.

For U, the matrix AA^{H} gives us two choices for each eigenvector:

$$\mathbf{u}_1 = \beta_1 \cdot \begin{pmatrix} \frac{\sqrt{6}}{6} \\ \frac{\sqrt{6}}{3} \\ \frac{\sqrt{6}}{6} \end{pmatrix}$$

$$\mathbf{u}_2 = \beta_2 \cdot \begin{pmatrix} \frac{\sqrt{2}}{0} \\ 0 \\ -\frac{\sqrt{2}}{2} \end{pmatrix}$$

$$\mathbf{u}_3 = \beta_3 \cdot \begin{pmatrix} \frac{\sqrt{3}}{3} \\ -\frac{\sqrt{3}}{3} \\ \frac{\sqrt{3}}{3} \end{pmatrix}$$

Above, $\beta_i \in \{-1, 1\}$.

Unlike the eigenvectors, the values in D are unique. These values are intrinsic to the matrix A.

Then choosing eigenvectors, it's generally best to start with the *smaller* of U, V. In this case, we'll start with V, and those choices will dictate U.

For V, let's pick $\alpha_1 = 1$ and $\alpha_2 = 1$. We find the U vectors as follows:

$$\mathbf{u}_1 = \frac{1}{\sqrt{3}} A\mathbf{v}_1$$

$$\mathbf{u}_2 = \frac{1}{\sqrt{1}} A\mathbf{v}_2$$

(for $\mathbf{u}_3$, we can choose either 1 or -1).

For example, to choose $\mathbf{u}_2$:

$$\begin{aligned} \mathbf{u}_2 &= \frac{1}{\sqrt{1}} A\mathbf{v}_2 \\ &= \begin{pmatrix} 0 & 1 \\ 1 & 1 \\ 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{pmatrix} \\ &= \begin{pmatrix} -\frac{\sqrt{2}}{2} \\ 0 \\ \frac{\sqrt{2}}{2} \end{pmatrix} \end{aligned}$$

Thus, we must choose $\beta_2 = 1$.

For this example, our decomposition is

$$\begin{pmatrix} 0 & 1 \\ 1 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{6}}{6} & -\frac{\sqrt{2}}{2} & \frac{\sqrt{3}}{3} \\ \frac{\sqrt{6}}{3} & 0 & -\frac{\sqrt{3}}{3} \\ \frac{\sqrt{6}}{6} & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{3} \end{pmatrix} \cdot \begin{pmatrix} \sqrt{3} & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{pmatrix} \tag{2.65}$$

In (2.65) notice what happens when we multiply the third column of U by D – it goes to $\mathbf{0}$. (The column corresponds to a zero eigenvalue). Therefore, there's no harm in throwing away the third column of U, and the last row of D.

$$\begin{pmatrix} 0 & 1 \\ 1 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{6}}{6} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{6}}{3} & 0 \\ \frac{\sqrt{6}}{6} & \frac{\sqrt{2}}{2} \end{pmatrix} \cdot \begin{pmatrix} \sqrt{3} & 0 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{pmatrix} \tag{2.66}$$

Equation (2.66) is called the *reduced SVD*. Let's manipulate this equation a little.

$$\begin{pmatrix} 0 & 1 \\ 1 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{6}}{6} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{6}}{3} & 0 \\ \frac{\sqrt{6}}{6} & \frac{\sqrt{2}}{2} \end{pmatrix} \cdot \begin{pmatrix} \sqrt{3} \cdot \frac{\sqrt{2}}{2} & \sqrt{3} \cdot \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{pmatrix} \qquad \text{multiply } DV^{\mathsf{H}} \tag{2.67}$$

$$= \begin{pmatrix} \frac{\sqrt{6}}{6} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{6}}{3} & 0 \\ \frac{\sqrt{6}}{6} & \frac{\sqrt{2}}{2} \end{pmatrix} \cdot \begin{pmatrix} \sigma_1 \mathbf{v}_1^{\mathsf{H}} \\ \sigma_2 \mathbf{v}_2^{\mathsf{H}} \end{pmatrix} \qquad \text{treat } DV^{\mathsf{H}} \text{ as row vectors} \tag{2.68}$$

$$= \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 \end{pmatrix} \cdot \begin{pmatrix} \sigma_1 \mathbf{v}_1^{\mathsf{H}} \\ \sigma_2 \mathbf{v}_2^{\mathsf{H}} \end{pmatrix} \qquad \text{treat } U \text{ as column vectors} \tag{2.69}$$

$$= \sigma_1 \mathbf{u}_1 \mathbf{v}_1^{\mathsf{H}} + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^{\mathsf{H}} \tag{2.70}$$

Equation (2.70) shows A as the sum of two matrices. Each of these matrices has *rank one*.

Recall that σ_1 is our largest singular value. The matrix $\sigma_1 \mathbf{u}_1 \mathbf{v}_1^{\mathsf{H}}$ is the best approximation of A that can be obtained from a rank one matrix. (We'll prove this shortly).

$$\sigma_1 \mathbf{u}_1 \mathbf{v}_1^{\mathsf{H}} = \sqrt{3} \cdot \begin{pmatrix} \frac{\sqrt{6}}{6} \\ \frac{\sqrt{6}}{3} \\ \frac{\sqrt{6}}{6} \end{pmatrix} \cdot \begin{pmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix} \tag{2.71}$$

$$= \begin{pmatrix} \frac{\sqrt{2}}{2} \\ \sqrt{2} \\ \frac{\sqrt{2}}{2} \end{pmatrix} \cdot \begin{pmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix} \tag{2.72}$$

$$= \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} \tag{2.73}$$

Equation (2.73) is not equal to A, but it's close. This is the closest approximation to A that we can get from a rank 1 matrix.

2.7.2 Norms and Unitary Matrices

In this section, we're going to discuss $|||A|||_2$ and $\|A\|_F$, as they relate to unitary matrices.

Recall that the Frobenius norm is

$$\|A\|_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}|^2}$$

The Frobenius norm is invariant under multiplication by a unitary matrix.

The same holds for $|||A|||_2$.

For a unitary matrix U, we have $UU^{\mathsf{H}} = U^{\mathsf{H}}U = I$. Therefore,

$$|||UA|||_2 = |||A|||_2$$
$$\|UA\|_F = \|A\|_F$$

For an $n \times n$ matrix A, the *trace* of the matrix trace(A) is the sum of A's diagonal elements:

$$\text{trace}(A) = \sum_{i=1}^{n} a_{ij}$$

trace(A) is equal to the sum of A's eigenvalues.

Let B be the matrix $B = AA^{\mathsf{H}}$. Let's look at the elements of B:

$$\begin{aligned}
B_{ij} &= \sum_{k=1}^{n} A_{ik}(A^{\mathsf{H}})_{kj} \\
B_{ii} &= \sum_{k=1}^{n} A_{ik}(A^{\mathsf{H}})_{ki} \\
&= \sum_{k=1}^{n} A_{ik}\overline{A}_{ik} \\
&= \sum_{k=1}^{n} |A_{ik}|^2 \\
\sum_{i=1}^{n} B_{ii} &= \sum_{i=1}^{n}\sum_{k=1}^{n} |A_{ik}|^2 \\
&= \text{trace}(B) \\
&= \|A\|_F^2
\end{aligned}$$

Therefore,

$$\text{trace}(AA^{\mathsf{H}}) = \|A\|_F^2 = \text{trace}(A^{\mathsf{H}}A)$$

For a Unitary matrix U,

$$\begin{aligned}
\|UA\|_F^2 &= \text{trace}((UA)^{\mathsf{H}}UA) \\
&= \text{trace}(A^{\mathsf{H}}U^{\mathsf{H}}UA) \\
&= \text{trace}(AA^{\mathsf{H}}) \\
&= \|A\|_F
\end{aligned}$$

The Frobenius norm of UA is the same as the Frobenius norm of A.

For our diagonal matrix D,

$$\|D\|_2 = \sigma_1$$
$$\|D\|_F = \sqrt{\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_p^2}$$

and

$$\|A\|_2 = \sigma_1$$

2.7.3 Using SVD to Approximate A

Suppose we have an $m \times n$ matrix A, and rank(A) = k. We would like to find a matrix X, such that rank(X) $\leq k$ and X is a "best approximation" of A.

Of course, we need a notion of what a "best approximation" is. It's natural to judge this with a distance measure. For example, below are two good approximations.

$$\|A - X\|_2$$
$$\|A - X\|_F$$

SVD helps us to choose such an X.

2.7.4 Matrix Approximation with SVD

Say we have $A = UDV^{\mathsf{H}}$ for an $m \times n$ matrix A.

$$A = \begin{pmatrix}\mathbf{u}_1 & \mathbf{u}_2 & \dots \mathbf{u}_m\end{pmatrix} \cdot \begin{pmatrix}\sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & & \\ \vdots & & \ddots & \\ 0 & & & \sigma_p\end{pmatrix} \cdot \begin{pmatrix}\mathbf{v}_1^{\mathsf{H}} \\ \mathbf{v}_2^{\mathsf{H}} \\ \vdots \\ \mathbf{v}_n^{\mathsf{H}}\end{pmatrix} \tag{2.74}$$

$$= \begin{pmatrix}\mathbf{u}_1 & \mathbf{u}_2 & \dots \mathbf{u}_m\end{pmatrix} \cdot \begin{pmatrix}\sigma_1\mathbf{v}_1^{\mathsf{H}} \\ \sigma_2\mathbf{v}_2^{\mathsf{H}} \\ \vdots \\ \sigma_p\mathbf{v}_n^{\mathsf{H}} \\ 0 \\ \vdots\end{pmatrix} \qquad \text{terms } p+1 \dots n \text{ go to zero in } DV^{\mathsf{H}} \tag{2.75}$$

$$= \begin{pmatrix}\mathbf{u}_1 & \mathbf{u}_2 & \dots \mathbf{u}_p\end{pmatrix} \cdot \begin{pmatrix}\sigma_1\mathbf{v}_1^{\mathsf{H}} \\ \sigma_2\mathbf{v}_2^{\mathsf{H}} \\ \vdots \\ \sigma_p\mathbf{v}_n^{\mathsf{H}}\end{pmatrix} \qquad \text{terms } p+1 \dots m \text{ go to zero in } U \tag{2.76}$$

$$= \sigma_1\mathbf{u}_1\mathbf{v}_1^{\mathsf{H}} + \sigma_2\mathbf{u}_2\mathbf{v}_2^{\mathsf{H}} + \dots + \sigma_p\mathbf{u}_p\mathbf{v}_p^{\mathsf{H}} \tag{2.77}$$

Notice what's happening above. D is an $m \times n$ matrix, but $d_{ii} = 0$ for $i > p$. The lower right terms of D have the effect of zeroing out some of the rows of V^{H}, and some of the columns of U. This leaves us with at most p singular value terms.

Suppose we retain only the first k terms of (2.77) (for $k < p$):

$$X = \sigma_1\mathbf{u}_1\mathbf{v}_1^{\mathsf{H}} + \sigma_2\mathbf{u}_2\mathbf{v}_2^{\mathsf{H}} + \dots + \sigma_k\mathbf{u}_k\mathbf{v}_k^{\mathsf{H}} \tag{2.78}$$

Claim: In (2.78), X is a matrix of rank k, *and* X is the best approximation of A having rank k.

We'll justify this claim below.

Let's start by making an observation. Consider the space $\mathbb{R}^l$.

Let U and V be two subspaces of $\mathbb{R}^l$. The intersection $U \cap V$ is always a subspace. At the very least, $\mathbf{0} \in U \cap V$, and $\mathbf{0}$ is a perfectly good subspace.

If $\dim(U) + \dim(V) > l$, then $U \cap V$ must contain a nonzero element. Why? Let $\mathbf{u}_1, \dots, \mathbf{u}_p$ be a basis for U, let $\mathbf{v}_1, \dots, \mathbf{v}_q$ be a basis for V, and let $p + q > l$.

Suppose

$$\mathbf{x} = a_1\mathbf{u}_1 + a_2\mathbf{u}_2 + \dots + a_p\mathbf{u}_p$$
$$\mathbf{x} = b_1\mathbf{v}_1 + b_2\mathbf{v}_2 + \dots + a_q\mathbf{v}_q$$

Given $p + q > l$, is it possible to have $\mathbf{x} \neq \mathbf{0}$? The answer is yes, because if we have

$$a_1\mathbf{u}_1 + \dots + a_p\mathbf{u}_p - (b_1\mathbf{v}_1 + \dots + a_q\mathbf{v}_q) = \mathbf{0}$$

There are $> l$ terms. Because there are $> l$ terms, this cannot be linearly dependent in $\mathbb{R}^l$.

Now that we've made this observation, let's return to the main problem of proving that X is the best approximation of A having rank k.

Let

$$r_k = \inf\{\|\|A - X\|\|_2 \mid \text{rank}(X) \leq k\} \tag{2.79}$$

We prove that a lower bound is achieved if

$$X = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^{\mathsf{H}} + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^{\mathsf{H}} + \ldots + \sigma_k \mathbf{u}_k \mathbf{v}_k^{\mathsf{H}}$$

Note that $A - X$ is

$$A - X = \sum_{i=k+1}^{p} \sigma_i \mathbf{u}_i \mathbf{v}_i^{\mathsf{H}}$$

Therefore, σ_{k+1} is the largest singular value in $A - X$. This tells us that $\|\|A - X\|\| = \sigma_{k+1}$.

Let Z be an $m \times n$ matrix ($m \geq n$) such that $\text{rank}(Z) \leq k$. Z and A have the same dimensions. We have

$$\begin{aligned} \dim(\text{null}(Z) + \text{rank}(Z)) &= n \\ \dim(\text{null}(Z)) &\geq n - k \end{aligned}$$

Let T be the subspace generated by $\langle \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{k+1} \rangle$. We have $\dim(T) = k + 1$.

Since $\dim(\text{null}(Z)) \geq n - k$ and $\dim(T) = k + 1$, we have

$$\dim(T) + \dim(\text{null}(Z)) \geq n$$

Therefore, there is a vector $\mathbf{x}$ such that $\mathbf{x} \in T$, $\mathbf{x} \in \text{null}(X)$ and $\mathbf{x} \neq \mathbf{0}$. (This is similar to the $\mathbb{R}^l$ argument we made earlier.)

Therefore,

$$\begin{aligned} \mathbf{x} &= \sum_{i=1}^{k+1} a_i \mathbf{v}_i \\ \|\mathbf{x}\|_2 &= \sum_{i=1}^{k+1} |a_i|^2 \end{aligned}$$

Next, let's examine $\mathbf{x}$. Recall that $\mathbf{x} \neq \mathbf{0}$ and $\mathbf{x} \in \text{null}(X)$.

$$\begin{aligned} (A - X)\mathbf{x} &= A\mathbf{x} - X\mathbf{x} \\ &= A\mathbf{x} && \text{since } \mathbf{x} \in \text{null}(X) \\ &= A \sum_{i=1}^{k+1} a_i \mathbf{v}_i \\ &= \sum_{i=1}^{k+1} a_i (A\mathbf{v}_i) \\ &= \sum_{i=1}^{k+1} a_i \sigma_i \mathbf{u}_i \end{aligned}$$

$$\begin{aligned} \|\|(A - X)\mathbf{x}\|\|_2 &= \sum_{i=1}^{k+1} |a_i|^2 \sigma_i^2 \\ &\geq \sigma_{k+1}^2 \sum_{i=1}^{k+1} |a_i|^2 \\ &= \sigma_{k+1}^2 \|\mathbf{x}\|_2^2 \end{aligned}$$

Therefore, $|||A - X|||_2 \geq \sigma_{k+1}$ and (2.79) holds.

2.8 Report Topics

After looking at several papers, these seem to be the most interesting.

- *Empirical Software Change Impact Analysis using Singular Value Decomposition*, by Mark Sherriff and Laurie Williams, from the 2008 International Conference on Software Testing, Verification, and Validation.

  `http://ieeexplore.ieee.org/iel5/4539516/4539517/04539554.pdf?arnumber=4539554`.

- *Network traffic analysis using singular value decomposition and multiscale transforms*, by Sastry, Rawat, Pujari, and Gulati. From *Information Sciences* Volume 177, Issue 23, 1 December 2007, Pages 5275-5291.

  `http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6V0C-4KJV21S-1&_user=10&_rdoc=1&_fmt=&_orig=search&_sort=d&view=c&_acct=C000050221&_version=1&_urlVersion=0&_userid=10&md5=875d892433d217d136b10ee391980e43`

How to get a set of GnuPG commit information:

```
svn co svn://cvs.gnupg.org/gnupg/trunk gnupg
cd gnupg
svn log --verbose > svn-log.txt
```

GnuPG has approximately 714 files. `svn log` lists 4945 revisions, since 1997-11-18, giving us 11-1/2 years of revision history.

2.9 Misc. Scilab notes

Reading and writing matrices:
`http://www.scilab.org/product/man/index.php?module=fileio&page=fscanfMat.htm`

```
fd=mopen(TMPDIR+'/Mat','w');
mfprintf(fd,'Some text.....\n');
mfprintf(fd,'Some text again\n');
a=rand(6,6);
for i=1:6 ,
 for j=1:6, mfprintf(fd,'%5.2f ',a(i,j));end;
 mfprintf(fd,'\n');
end
mclose(fd);
a1=fscanfMat(TMPDIR+'/Mat')
```

An even simpler example

```
a = rand(10,1)
fprintfMat("foo2", a, "%10.8f")
a1 = fscanfMat("foo2");
```

2.10 Lecture – 3/9/2009

In prior classes, we've looked at SVD: $A = UDV^{\mathsf{H}}$. Today, we'll look at another form of matrix decomposition.

First, we need a few definitions.

Definition 2.10.1 (Positive Matrix): We say that A is a *positive matrix* if all $a_{ij} > 0$.

Definition 2.10.2 (Positive Semi-Definite): If $\mathbf{x} \neq \mathbf{0}$ and $\mathbf{x}'A\mathbf{x} \geq 0$, then we say that the matrix A is *positive semi-definite*.

Definition 2.10.3 (Positive Definite): If $\mathbf{x} \neq \mathbf{0}$ and $\mathbf{x}'A\mathbf{x} > 0$, then we say that the matrix A is *positive definite*.

Note that the term "positive matrix" has nothing in common with the terms "positive semi-definite" or "positive definite". Positive matrix is a completely separate concept.

Consider a 2×2 matrix A:

$$A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$$

A is a positive matrix if $a > 0$, $b > 0$, and $c > 0$.

What would make A a positive semi-definite matrix? A would have to fit the inequality

$$\begin{pmatrix} x_1 & x_2 \end{pmatrix} \cdot \begin{pmatrix} a & b \\ b & c \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \geq 0 \tag{2.80}$$

For every $\mathbf{x} \neq \mathbf{0}$.

Let's multiply (2.80) out

$$\begin{pmatrix} x_1 & x_2 \end{pmatrix} \cdot \begin{pmatrix} a & b \\ b & c \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \tag{2.81}$$

$$= \begin{pmatrix} ax_1 + bx_2 & bx_1 + cx_2 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \tag{2.82}$$

$$= ax_1^2 + 2bx_1x_2 + cx_2^2 \tag{2.83}$$

$$= x_1^2 \cdot \left(a + 2b\frac{x_2}{x_1} + c\left(\frac{x_2}{x_1}\right)^2 \right) \qquad \text{assume } x_1 \neq 0 \tag{2.84}$$

In order for the trinomial in (2.84) to be positive, we need $a > 0$ and $b^2 - ac < 0$.

What we've shown so far applies to real symmetric matrices.

Let's say that A is a Hermetian matrix (i.e., $A = A^{\mathsf{H}}$). Can we extend this property to Hermetian matrices?

If we take $\mathbf{x} \in \mathbb{C}^n$ and compute $\mathbf{x}^{\mathsf{H}}A\mathbf{x}$, then A must be a real matrix, since $\mathbf{x}^{\mathsf{H}}A\mathbf{x} = \mathbf{x}^{\mathsf{H}}A^{\mathsf{H}}\mathbf{x}$.

A Hermetian matrix is positive semi-definite if $\mathbf{x}^{\mathsf{H}}A\mathbf{x} \geq 0$ for $\mathbf{x} \neq \mathbf{0}$.

A Hermetian matrix is positive definite if $\mathbf{x}^{\mathsf{H}}A\mathbf{x} > 0$ for $\mathbf{x} \neq \mathbf{0}$.

2.10.1 Cholesky Decomposition

Every positive definite matrix A can be written as

$$A = R^{\mathsf{H}} R \tag{2.85}$$

Where R is an upper triangular matrix, and R^{H} is a lower triangular matrix. This is the *Cholesky Decomposition*.

2.10.2 Submatrix

Suppose we have a square matrix A. The submatrix

$$B = A\left[\begin{smallmatrix} i_1,\ldots,i_h \\ j_1,\ldots,j_k \end{smallmatrix}\right] \tag{2.86}$$

Forms B out of rows $i_1, \ldots, i_h$ and columns $j_1, \ldots, j_k$.

If the rows and columns are the same, then $A\left[\begin{smallmatrix} i_1,\ldots,i_h \\ i_1,\ldots,i_h \end{smallmatrix}\right]$ is a *principal submatrix* of A, and

$$\det A\left[\begin{smallmatrix} i_1,\ldots,i_h \\ i_1,\ldots,i_h \end{smallmatrix}\right] \tag{2.87}$$

is a *principal minor* of A.

Any element on the diagonal of A is a minor.

Claim 2.10.4: If A is a positive definite matrix, then all principal minors are positive. □

Claim 2.10.5: If A is a positive semi-definite matrix, then all principal minor are non-negative. □

Let's prove Claim 2.10.4 for positive definite matrices.

We know that $\mathbf{x}^{\mathsf{H}} A\mathbf{x} > 0$ for every vector $\mathbf{x} \neq \mathbf{0}$.

Let us choose $\mathbf{x}$ such that

$$x_i = \begin{cases} 0 & \text{if } i \notin \{i_1, \ldots, i_n\} \\ x_i & \text{otherwise} \end{cases}$$

The product $\mathbf{x}^{\mathsf{H}} A\mathbf{x}$ is

$$\begin{aligned} \mathbf{x}^{\mathsf{H}} A\mathbf{x} &= \sum_{p=1}^{1} \sum_{q=1}^{1} a_{pq} \overline{x}_p x_q \\ &= \hat{\mathbf{x}}^{\mathsf{H}} B \hat{\mathbf{x}} \end{aligned}$$

Where $\hat{\mathbf{x}}$ consists of the elements of $\mathbf{x}$ that correspond to $\{i_1, \ldots, i_h\}$ and B is a positive definite submatrix of A.

If A is a positive definite matrix, then all principal submatrices of A are also positive definite.

Note: if all $a_{ii} > 0$ then A *might* be positive definite. However, if some $a_{ii} \leq 0$ then A is *definitely not* positive definite.

2.10.3 Preservation of Positive Definite Matrices Under Multiplication

Theorem 2.10.6: Say $A \in \mathbb{C}^{n \times n}$ is a positive definite matrix, and S is another matrix in $\mathbb{C}^{n \times n}$. If A is a positive definite matrix, then $S^{\mathsf{H}} AS$ is positive semi-definite. □

To prove this, we must show that

$$\mathbf{x}^{\mathsf{H}} S^{\mathsf{H}} A S \mathbf{x} \geq 0$$

For $\mathbf{x} \neq \mathbf{0}$.

Note that $\mathbf{x}^{\mathsf{H}} S^{\mathsf{H}} A S \mathbf{x} = (S\mathbf{x})^{\mathsf{H}} A S \mathbf{x}$.

We want to show that $\text{rank}(S^{\mathsf{H}} A S) = \text{rank}(S)$.

If $S\mathbf{x} = \mathbf{0}$, then $\mathbf{x} \in \text{null}(S)$. Therefore,

$$(S^{\mathsf{H}} A S)\mathbf{x} = S^{\mathsf{H}} A (S\mathbf{x}) = \mathbf{0}.$$

so, $\mathbf{x} \in \text{null}(S^{\mathsf{H}} A S)$.

$$\begin{aligned} S^{\mathsf{H}} A S \mathbf{x} &= \mathbf{0} \\ \mathbf{x}^{\mathsf{H}} S^{\mathsf{H}} A S \mathbf{x} &= \mathbf{0} \\ (S\mathbf{x})^{\mathsf{H}} A (S\mathbf{x}) &= \mathbf{0} && \text{because } \mathbf{x} \in \text{null}(S) \end{aligned}$$

The null spaces for S and $S^{\mathsf{H}} A S$ are the same – therefore, their ranks are the same.

In the previous example, $S^{\mathsf{H}} A S$ was a positive semi-definite matrix. Under what conditions would $S^{\mathsf{H}} A S$ be positive definite?

$$\mathbf{x}^{\mathsf{H}} S^{\mathsf{H}} A S \mathbf{x} > 0 \qquad \text{if } S^{\mathsf{H}} A S \text{ was positive definite}$$

If $\text{null}(S^{\mathsf{H}} A S) = \mathbf{0}$ and $\text{rank}(S) = n$, then $S^{\mathsf{H}} A S$ will be positive definite.

2.10.4 Cholesky Decomposition

Suppose we start with a positive definite Hermetian matrix $A \in \mathbb{C}^{n \times n}$. We can write:

$$A = \begin{pmatrix} a_{11} & \mathbf{a}^{\mathsf{H}} \\ \mathbf{a} & B \end{pmatrix} \qquad \text{since } A \text{ is Hermetian} \tag{2.88}$$

Because A is positive definite, we know that $a_{ii} > 0$.

Now, let's multiply the following:

$$\begin{pmatrix} \sqrt{a_{11}} & \mathbf{0} \\ \frac{1}{\sqrt{a_{11}}}\mathbf{a} & I_{n-1} \end{pmatrix} \cdot \begin{pmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & B - \frac{1}{a_{11}}\mathbf{a}\mathbf{a}^{\mathsf{H}} \end{pmatrix} \cdot \begin{pmatrix} \sqrt{a_{11}} & \frac{1}{\sqrt{a_{11}}}\mathbf{a}^{\mathsf{H}} \\ \mathbf{0} & I_{n-1} \end{pmatrix} \tag{2.89}$$

$$= \begin{pmatrix} \sqrt{a_{11}} & \mathbf{0} \\ \frac{1}{\sqrt{a_{11}}}\mathbf{a} & B - \frac{1}{a_{11}}\mathbf{a}\mathbf{a}^{\mathsf{H}} \end{pmatrix} \cdot \begin{pmatrix} \sqrt{a_{11}} & \frac{1}{\sqrt{a_{11}}}\mathbf{a}^{\mathsf{H}} \\ \mathbf{0} & I_{n-1} \end{pmatrix} \tag{2.90}$$

$$= \begin{pmatrix} a_{11} & \mathbf{a}^{\mathsf{H}} \\ \mathbf{a} & B \end{pmatrix} \tag{2.91}$$

In (2.89) the first matrix is S, and the last matrix is S^{H}.

Every element on the diagonal of $S^{\mathsf{H}} > 0$. Therefore $\text{rank}(S^{\mathsf{H}}) = n$. Also, S is positive definite.

The middle matrix is positive definite, and the middle matrix is a principal submatrix.

By the inductive hypothesis, $B - \frac{1}{a_{11}}\mathbf{a}\mathbf{a}^{\mathsf{H}}$ can be written as $P^{\mathsf{H}} P$, where P is an upper triangular matrix.

Let A_1 be

$$\begin{aligned} A_1 &= B - \frac{1}{a_{11}}\mathbf{a}\mathbf{a}^{\mathsf{H}} \\ &= \begin{pmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & P^{\mathsf{H}} \end{pmatrix} \\ \therefore\ A &= \begin{pmatrix} \sqrt{a_{11}} & \mathbf{0} \\ \frac{1}{\sqrt{a_{11}}}\mathbf{a} & I \end{pmatrix} \cdot \begin{pmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & P \end{pmatrix} \end{aligned}$$

Given a matrix A that is positive (but *not* positive definite), it would be nice if we could decompose A into

$$\underset{(n \times n)}{A} = \underset{(n \times k)}{W} \cdot \underset{(k \times n)}{H}$$

for some small value of k, where W and H are both positive matrices. *However* we cannot do this in all cases.

But we can try to find W, H, such that $\|A - WH\|_F$ if small. If $\|A - WH\|_F$ is small, then WH will be a reasonably good approximation of A.

2.10.5 Miscellany

- Look for papers by Lee Seung
- Look for a new handout on Cholesky Decomposition

In upcoming classes, we will talk about the Gram-Schmidt algorithm, which takes a linear space and derives a set of orthogonal basis vectors.

2.11 Lecture – 3/11/2009

2.11.1 QR Decomposition

Today, we will discuss QR-Decomposition. QR-Decomposition is a way of factoring A into two matrices (Q and R), where

- Q is an orthogonal matrix, and
- R is an upper triangular matrix.

We'll start with some preliminary material.

2.11.2 The Gram-Schmidt Algorithm

Let $U \in \mathbb{R}^n$ be a subspace (other than the trivial subspace). U has a set of basis vectors that generate the subspace.

Let $\dim(U) = n$, and let $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m\}$ be a set of basis vectors for U.

We know two things about this set of basis vectors:

1. $\langle \mathbf{u}_1, \ldots, \mathbf{u}_m \rangle = U$, and
2. the set $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ is linearly independent.

Given the linear space U with basis vectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$, we can construct another set of basis vectors $\{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$, such that

$$\begin{aligned} \|\mathbf{w}\|_i &= 1 && \text{each vector is a unit vector} \\ \mathbf{w}_i^{\mathsf{H}} \mathbf{w}_j &= 0 \quad \text{if } i \neq j && \text{each vector is othogonal} \\ \langle \mathbf{w}_1, \ldots, \mathbf{w}_k \rangle &= \langle \mathbf{u}_1, \ldots, \mathbf{u}_k \rangle \text{ for } 1 \leq k \leq m \end{aligned}$$

The third line says that the first k vectors of $\mathbf{w}$ span the same subspace as the first k vectors of $\mathbf{u}$.

The Gram-Schmidt algorithm takes the $\mathbf{u}$ vectors as input, and produces the $\mathbf{w}$ vectors as output.

The set $W = \{\mathbf{w}_1, \mathbf{w}_2, \ldots\}$ is constructed iteratively.

$$\mathbf{w}_1 = \frac{1}{\|\mathbf{u}_1\|} \cdot \mathbf{u}_1$$

Let's say we've built $\{\mathbf{w}_1, \ldots, \mathbf{w}_{k-1}\}$, and we wish to build $\mathbf{w}_k$.

$$\mathbf{w}_k = \alpha_j \left(\mathbf{u}_k - \sum_{j=1}^{k-1} (\mathbf{u}_k, \mathbf{w}_j) \cdot \mathbf{w}_j \right) \tag{2.92}$$

Above, $(\mathbf{u}_k, \mathbf{w}_j) \cdot \mathbf{w}_j$ denotes projection.

Note 2.11.1 (Projection): The *projection* of a vector $\mathbf{v}$ onto a vector $\mathbf{u}$ is

$$\frac{\mathbf{v} \cdot \mathbf{u}}{\mathbf{u} \cdot \mathbf{u}} \cdot \mathbf{u} \tag{2.93}$$

Given what we have above,

$$(\mathbf{v}, \mathbf{u}) = \frac{\mathbf{v} \cdot \mathbf{u}}{\mathbf{u} \cdot \mathbf{u}} \tag{2.94}$$

where '$\cdot$' denotes dot product. □

Let's project each side of (2.92) by $\mathbf{w}_i$:

$$(\mathbf{w}_k, \mathbf{w}_i) = \alpha_k \left(\sum_{j=1}^{k-1} (\mathbf{u}_k, \mathbf{w}_j)(\mathbf{w}_j, \mathbf{w}_i) \right) \tag{2.95}$$

$$= \alpha_k (\mathbf{u}_k, \mathbf{w}_i) \tag{2.96}$$

Above, the summation term goes to zero because all $\mathbf{w}_i$ are perpendicular.

The purpose of α_k is to scale $\mathbf{w}_k$ to a unit vector:

$$\alpha_k = \frac{1}{\|\mathbf{u}_k - \sum_{j=1}^{k-1} (\mathbf{u}_k, \mathbf{w}_j)\mathbf{w}_j\|} \tag{2.97}$$

The $\mathbf{w}$ vectors generate the same subspace as the $\mathbf{u}$ vectors.

Suppose we have

$$\langle \mathbf{w}_1, \ldots, \mathbf{w}_{k-1} \rangle = \langle \mathbf{u}_1, \ldots, \mathbf{u}_{k-1} \rangle$$

Because $\mathbf{w}_k$ is orthogonal to the other $\mathbf{w}_i$ vectors, we will have

$$\langle \mathbf{w}_1, \ldots, \mathbf{w}_k \rangle = \langle \mathbf{u}_1, \ldots, \mathbf{u}_k \rangle$$

2.11.3 Applications of the Gram-Schmidt Algorithm

How is the Gram-Schmidt algorithm useful? Suppose we start off with a 3×3 random matrix A with $\text{rank}(A) = 3$. The three column vectors of A are linearly independent, but not necessarily orthogonal.

With the Gram-Schmidt algorithm, we can construct a matrix A' where the column vectors of A' span the same linear space as A, *and* the column vectors of A' are orthogonal.

Let $A \in \mathbb{C}^{m \times n}$ where $m \geq n$ and $\text{rank}(A) = n$. A is a *full rank* matrix.

The range of A is generated by its columns

$$A = \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \ldots & \mathbf{u}_n \end{pmatrix}$$

and each $\mathbf{u}_i \in \mathbb{C}^{m \times 1}$ (a column vector).

The columns $\mathbf{u}_i$ are the basis for range(A).

Gram Schmidt alows us to product a set of vectors $\{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$ such that

- $\|\mathbf{w}_i\| = 1$
- $\mathbf{w}_i \cdot \mathbf{w}_j = 0$ if $i \neq j$
- $\langle \mathbf{u}_1, \ldots, \mathbf{u}_n \rangle = \langle \mathbf{w}_1, \ldots, \mathbf{w}_n \rangle$
- The first k $\mathbf{u}$ vectors can be expressed as linear combinations of the first k $\mathbf{w}$ vectors.

Let

$$\begin{aligned} \mathbf{u}_1 &= r_{11}\mathbf{w}_1 \\ \mathbf{u}_2 &= r_{12}\mathbf{w}_1 + r_{22}\mathbf{w}_2 \\ &\vdots \\ \mathbf{u}_n &= r_{1n}\mathbf{w}_1 + \ldots + r_{nn}\mathbf{w}_n \end{aligned}$$

Let's create a matrix $Q = (\mathbf{w}_1, \ldots, \mathbf{w}_n)$ where each $\mathbf{w}_i$ is a column vector. We have $\mathbf{w}_i \in \mathbb{C}^m$ and $Q \in \mathbb{C}^{m \times n}$. Q has the same dimension as the original matrix, A.

Claim 2.11.2: Q is a unitary matrix. ($Q \cdot Q^{\mathsf{H}} = I_n$).

If we write this out,

$$Q \cdot Q^{\mathsf{H}} = I_n$$

$$\begin{pmatrix} \mathbf{w}_1 & \mathbf{w}_2 & \dots & \mathbf{w}_n \end{pmatrix} \cdot \begin{pmatrix} \mathbf{w}_1^{\mathsf{H}} \\ \mathbf{w}_2^{\mathsf{H}} \\ \vdots \\ \mathbf{w}_n^{\mathsf{H}} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & & \\ \vdots & & \ddots & \\ 0 & & & 1 \end{pmatrix}$$

This holds because

- $\mathbf{w}_i \mathbf{w}_i^{\mathsf{H}} = 1$ (unit vectors)
- $\mathbf{w}_i \mathbf{w}_j^{\mathsf{H}} = 0$ for $i \neq j$ (orthogonal vectors)

Therefore $Q \cdot Q^{\mathsf{H}} = I_n$.

Definition 2.11.3 (Column Orthogonal Matrix): A *column orthogonal matrix* (or an *orthonormal* matrix) is to real numbers as a unitary matrix is to complex numbers.

In the world of real numbers, Q is an orthonormal matrix. Q is also the analog of a rotation matrix.

We can express A in terms of Q as follows:

$$\begin{aligned} A &= \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \dots & \mathbf{u}_n \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{w}_1 & \mathbf{w}_2 & \dots & \mathbf{w}_n \end{pmatrix} \cdot \begin{pmatrix} r_{11} & r_{12} & r_{13} & \dots & r_{1n} \\ 0 & r_{22} & r_{23} & \dots & r_{2n} \\ 0 & 0 & r_{33} & \dots & r_{3n} \\ \vdots & & & \ddots & \vdots \\ 0 & \dots & & & r_{nn} \end{pmatrix} \end{aligned} \tag{2.98}$$

In (2.98), the first RHS matrix is Q and the second RHS matrix is R.

We can choose QR such that $r_{ii} > 0$ and $\text{rank}(R) = n$.

We know that $\text{rank}(A) \leq \min(\text{rank}(Q), \text{rank}(R))$. $\text{rank}(A) = n$ and $\text{rank}(Q) = n$; therefore, $\text{rank}(R) = n$.

Earlier, we noted that

$$\mathbf{w}_k = \alpha_k \left(\mathbf{u}_k - \sum_{j=1}^{k-1} (\mathbf{u}_k, \mathbf{w}_j) \mathbf{w}_j \right)$$

Each $r_{ii} = \mathbf{u}_i \cdot \mathbf{w}_i$.

All r_{ii} can be made positive. (We can always make r_{ii} positive by changing $\mathbf{w}_i$ to $(-\mathbf{w}_i)$.

2.11.4 Another Strategy for QR

It's also possible to form a QR Decomposition where Q is a square matrix.

$$\underset{(m \times n)}{A} = \underset{(m \times m)}{Q} \cdot \underset{(m \times n)}{R}$$

2.11.5 Householder Matrices

Say we have a vector $\mathbf{v} \in \mathbb{C}^n$.

The product $\mathbf{v} \cdot \mathbf{v}^{\mathsf{H}}$ is an $n \times n$ rank one matrix.

Let us choose $\mathbf{v}$ such that $\|\mathbf{v}\| = \sqrt{2}$. (In this section, $\|\mathbf{v}\|$ implicitly means $\|\mathbf{v}\|_2$, the Euclidean norm.)

The *Householder Matrix* H_v is

$$H_v = I - \mathbf{v}\mathbf{v}^{\mathsf{H}} \tag{2.99}$$

H_v is a unitary matrix. H_v is also known as the *Householder Reflection*.

Let's manipulate (2.99) a little.

$$\begin{aligned}
H_v &= I_n - \mathbf{v}\mathbf{v}^{\mathsf{H}} \\
H_v^{\mathsf{H}} &= I_n - \mathbf{v}\mathbf{v}^{\mathsf{H}} && \text{since } H_v \text{ is unitary} \\
H_v H_v^{\mathsf{H}} &= (I - \mathbf{v}\mathbf{v}^{\mathsf{H}})(I - \mathbf{v}\mathbf{v}^{\mathsf{H}}) && \text{multiply each side by } H_v \\
&= I - \mathbf{v}\mathbf{v}^{\mathsf{H}} - \mathbf{v}\mathbf{v}^{\mathsf{H}} + \mathbf{v}\mathbf{v}^{\mathsf{H}}\mathbf{v}\mathbf{v}^{\mathsf{H}} \\
&= I - \mathbf{v}\mathbf{v}^{\mathsf{H}} - \mathbf{v}\mathbf{v}^{\mathsf{H}} + \mathbf{v}(\mathbf{v}^{\mathsf{H}}\mathbf{v})\mathbf{v}^{\mathsf{H}} && \text{assoc. property} \\
&= I - \mathbf{v}\mathbf{v}^{\mathsf{H}} - \mathbf{v}\mathbf{v}^{\mathsf{H}} + 2\mathbf{v}\mathbf{v}^{\mathsf{H}} && \text{since } \|\mathbf{v}\| = \sqrt{2} \\
&= I
\end{aligned}$$

Claim 2.11.4: If $\|\mathbf{x}\| = \|\mathbf{y}\|$, then there is a Householder matrix H_v such that $H_v\mathbf{x} = \mathbf{y}$. H_v rotates $\mathbf{x}$ until it becomes $\mathbf{y}$.

Note 2.11.5: For any vector $\mathbf{x}$, we can find a vector $\mathbf{y}$ such that $\|\mathbf{x}\| = \|\mathbf{y}\|$ and $\mathbf{y}$ has one non-zero element, in y_1.

Let's work with Claim 2.11.4.

$$\begin{aligned}
(I - \mathbf{v}\mathbf{v}^{\mathsf{H}})\mathbf{x} &= \mathbf{y} \\
\mathbf{x} - \mathbf{v}\mathbf{v}^{\mathsf{H}}\mathbf{x} &= \mathbf{y} \\
\mathbf{x} - \mathbf{y} &= \mathbf{v}\mathbf{v}^{\mathsf{H}}\mathbf{x} \\
&= \mathbf{v}(\mathbf{v}^{\mathsf{H}}\mathbf{x})
\end{aligned}$$

Note that $\mathbf{v}^{\mathsf{H}}\mathbf{x}$ is a scalar. Let $\frac{1}{\alpha} = \mathbf{v}^{\mathsf{H}}\mathbf{x}$.

$$\begin{aligned}
\mathbf{x} - \mathbf{y} &= \mathbf{v}(\mathbf{v}^{\mathsf{H}}\mathbf{x}) \\
\mathbf{x} - \mathbf{y} &= \mathbf{v}\frac{1}{\alpha} \\
\mathbf{v} &= \alpha(\mathbf{x} - \mathbf{y})
\end{aligned}$$

From this, we can say

$$\begin{aligned}
\sqrt{2} &= \|\mathbf{v}\| \\
&= \alpha \cdot \|\mathbf{x} - \mathbf{y}\| \\
\alpha &= \frac{\sqrt{2}}{\|\mathbf{x} - \mathbf{y}\|} \\
\mathbf{v} &= \sqrt{2} \cdot \frac{\mathbf{x} - \mathbf{y}}{\|\mathbf{x} - \mathbf{y}\|}
\end{aligned}$$

Suppose we have a vector $\mathbf{x}$, and the basis vector $\mathbf{e}_1$

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

The vector $\mathbf{e}_1 \cdot \|\mathbf{x}\|$ has the same norm as $\mathbf{x}$, and one nonzero component.

Next, let's introduce a variable s, where $s \in \{-1, 1\}$.

$$\|\mathbf{x}\| = \| \ s \cdot \mathbf{e}_1 \cdot \|\mathbf{x}\| \ \| \tag{2.100}$$

H_{v_1} is a matrix where

$$H_{v_1} = \sqrt{2} \cdot \frac{\mathbf{x} - s\mathbf{e}_1\|\mathbf{x}\|}{\|\mathbf{x} - s\mathbf{e}_1\|\mathbf{x}\|\|}$$

What does s do? We want to choose $s \in \{-1, 1\}$ such that $s\mathbf{e}_1\|\mathbf{x}\|$ is being added to $\mathbf{x}$, and not subtracted from $\mathbf{x}$. Subtraction tends to cause numerical instability, while addition tends to prevent numerical instability.

The matrix H_v is called a reflector because the vector $\mathbf{x}$ is reflected with respect to $\mathbf{v}$ to produce the vector $\mathbf{y}$. This is illustrated in Figure 2.2.

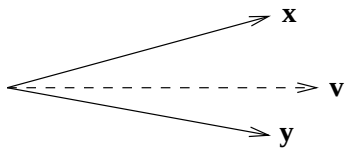

Figure 2.2: $\mathbf{x}$ reflected by $\mathbf{v}$ to produce $\mathbf{y}$

2.11.6 Logistics

- hw2 is posted, due 3/25/2009
- There is a new handout posted on the course web site

2.12 Example of the Gram Schmidt Algorithm

Here is an example of the Gram-Schmidt algorithm. This example was adapted from *Linear Algebra with Applications*, pg. 245.

I've seen two presentations of Gram-Schmidt that break the process into two steps:

- Finding orthogonal vectors (but not necessarily unit vectors).
- Turning the orthogonal vectors into unit vectors (which makes them orthonormal.)

The algorithm we discussed in class does both steps together – but it is possible to perform them separately.

Suppose we have three vectors:

$$\mathbf{u}_1 = (1, 2, 0, 3)$$
$$\mathbf{u}_2 = (4, 0, 5, 8)$$
$$\mathbf{u}_3 = (8, 1, 5, 6)$$

We want to find an orthonormal set of vectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ that span the same space.

Finding $\mathbf{v}_1$.

This part is easy:

$$\begin{aligned}\mathbf{v}_1 &= \frac{1}{\|\mathbf{u}_1\|} \cdot \mathbf{u}_1 \\ &= \frac{1}{\sqrt{14}} \cdot (1, 2, 0, 3) \\ &= \left(\frac{1}{\sqrt{14}}, \frac{2}{\sqrt{14}}, 0, \frac{3}{\sqrt{14}}\right)\end{aligned}$$

Finding $\mathbf{v}_2$.

$$\begin{aligned}\hat{\mathbf{v}}_2 &= \mathbf{u}_2 - (\mathbf{u}_2, \mathbf{v}_1)\mathbf{v}_1 \\ &= (4, 0, 5, 8) - \frac{(4, 0, 5, 8) \cdot \left(\frac{1}{\sqrt{14}}, \frac{2}{\sqrt{14}}, 0, \frac{3}{\sqrt{14}}\right)}{\left(\frac{1}{\sqrt{14}}, \frac{2}{\sqrt{14}}, 0, \frac{3}{\sqrt{14}}\right) \cdot \left(\frac{1}{\sqrt{14}}, \frac{2}{\sqrt{14}}, 0, \frac{3}{\sqrt{14}}\right)} \left(\frac{1}{\sqrt{14}}, \frac{2}{\sqrt{14}}, 0, \frac{3}{\sqrt{14}}\right) \\ &= (4, 0, 5, 8) - \frac{\frac{4}{\sqrt{14}} + 0 + 0 + \frac{24}{\sqrt{14}}}{\frac{1}{14} + \frac{4}{14} + 0 + \frac{9}{14}} \left(\frac{1}{\sqrt{14}}, \frac{2}{\sqrt{14}}, 0, \frac{3}{\sqrt{14}}\right) \\ &= (4, 0, 5, 8) - \frac{28}{\sqrt{14}} \left(\frac{1}{\sqrt{14}}, \frac{2}{\sqrt{14}}, 0, \frac{3}{\sqrt{14}}\right) \\ &= (4, 0, 5, 8) - \left(\frac{28}{14}, \frac{56}{14}, 0, \frac{84}{14}\right) \\ &= (4, 0, 5, 8) - (2, 4, 0, 6) \\ &= (2, -4, 5, 2)\end{aligned}$$

$$\begin{aligned}
\mathbf{v}_2 &= \frac{1}{\|\hat{\mathbf{v}}_2\|}\hat{\mathbf{v}}_2 \\
&= \frac{1}{\sqrt{49}}(2,-4,5,2) \\
&= \frac{1}{7}(2,-4,5,2) \\
&= \left(\frac{2}{7},-\frac{4}{7},\frac{5}{7},\frac{2}{7}\right)
\end{aligned}$$

Finding $\mathbf{v}_3$

This starts to get a little ugly, but what the hey ...

$$\begin{aligned}
\hat{\mathbf{v}}_3 &= \mathbf{u}_3 - (\mathbf{u}_3,\mathbf{v}_1)\mathbf{v}_1 - (\mathbf{u}_3,\mathbf{v}_2)\mathbf{v}_2 \\
&= \mathbf{u}_3 - (\mathbf{u}_3,\mathbf{v}_1)\mathbf{v}_1 - \frac{(8,1,5,6)\cdot\left(\frac{2}{7},-\frac{4}{7},\frac{5}{7},\frac{2}{7}\right)}{\left(\frac{2}{7},-\frac{4}{7},\frac{5}{7},\frac{2}{7}\right)\cdot\left(\frac{2}{7},-\frac{4}{7},\frac{5}{7},\frac{2}{7}\right)}\left(\frac{2}{7},-\frac{4}{7},\frac{5}{7},\frac{2}{7}\right) \\
&= \mathbf{u}_3 - (\mathbf{u}_3,\mathbf{v}_1)\mathbf{v}_1 - \frac{\frac{16}{7}-\frac{4}{7}+\frac{25}{7}+\frac{12}{7}}{\frac{4}{49}+\frac{16}{49}+\frac{25}{49}+\frac{4}{49}}\left(\frac{2}{7},-\frac{4}{7},\frac{5}{7},\frac{2}{7}\right) \\
&= \mathbf{u}_3 - (\mathbf{u}_3,\mathbf{v}_1)\mathbf{v}_1 - \frac{\frac{49}{7}}{\frac{49}{49}}\left(\frac{2}{7},-\frac{4}{7},\frac{5}{7},\frac{2}{7}\right) \\
&= \mathbf{u}_3 - (\mathbf{u}_3,\mathbf{v}_1)\mathbf{v}_1 - 7\left(\frac{2}{7},-\frac{4}{7},\frac{5}{7},\frac{2}{7}\right) \\
&= \mathbf{u}_3 - (\mathbf{u}_3,\mathbf{v}_1)\mathbf{v}_1 - (2,-4,5,2) \\
&= \mathbf{u}_3 - \frac{(8,1,5,6)\cdot(\frac{1}{\sqrt{14}},\frac{2}{\sqrt{14}},0,\frac{3}{\sqrt{14}})}{(\frac{1}{\sqrt{14}},\frac{2}{\sqrt{14}},0,\frac{3}{\sqrt{14}})\cdot(\frac{1}{\sqrt{14}},\frac{2}{\sqrt{14}},0,\frac{3}{\sqrt{14}})}(\tfrac{1}{\sqrt{14}},\tfrac{2}{\sqrt{14}},0,\tfrac{3}{\sqrt{14}}) - (2,-4,5,2) \\
&= \mathbf{u}_3 - \frac{\frac{8}{\sqrt{14}}+\frac{2}{\sqrt{14}}+0+\frac{18}{\sqrt{14}}}{(\frac{1}{14}+\frac{4}{14}+0+\frac{9}{14})}(\tfrac{1}{\sqrt{14}},\tfrac{2}{\sqrt{14}},0,\tfrac{3}{\sqrt{14}}) - (2,-4,5,2) \\
&= \mathbf{u}_3 - \frac{28}{\sqrt{14}}(\tfrac{1}{\sqrt{14}},\tfrac{2}{\sqrt{14}},0,\tfrac{3}{\sqrt{14}}) - (2,-4,5,2) \\
&= \mathbf{u}_3 - \frac{28}{\sqrt{14}}(\tfrac{1}{\sqrt{14}},\tfrac{2}{\sqrt{14}},0,\tfrac{3}{\sqrt{14}}) - (2,-4,5,2) \\
&= \mathbf{u}_3 - (\tfrac{28}{14},\tfrac{56}{14},0,\tfrac{84}{14}) - (2,-4,5,2) \\
&= \mathbf{u}_3 - (2,4,0,6) - (2,-4,5,2) \\
&= (8,1,5,6) - (2,4,0,6) - (2,-4,5,2) \\
&= (6,-3,5,0) - (2,-4,5,2) \\
&= (4,1,0,-2)
\end{aligned}$$

$$\begin{aligned}
\mathbf{v}_3 &= \frac{1}{\|\hat{\mathbf{v}}_3\|}\hat{\mathbf{v}}_3 \\
&= \frac{1}{\sqrt{21}}(4,1,0,-2) \\
&= \left(\frac{4}{\sqrt{21}},\frac{1}{\sqrt{21}},0,-\frac{2}{\sqrt{21}}\right)
\end{aligned}$$

Thus, our final solution is

$$\mathbf{v}_1 = \left(\frac{1}{\sqrt{14}}, \frac{2}{\sqrt{14}}, 0, \frac{3}{\sqrt{14}}\right)$$
$$\mathbf{v}_2 = \left(\frac{2}{7}, -\frac{4}{7}, \frac{5}{7}, \frac{2}{7}\right)$$
$$\mathbf{v}_3 = \left(\frac{4}{\sqrt{21}}, \frac{1}{\sqrt{21}}, 0, -\frac{2}{\sqrt{21}}\right)$$

2.13 Matrix Diagonalization – 3/16/2009

This material comes from pages 292–295 of *Linear Algebra with Applications*, but it can also be found in most linear algebra texts.

Definition 2.13.1 (Similar): If A and B are square matrices of the same size, then B is said to be *similar* to A if there exists an invertible matrix C such that $B = C^{-1}AC$. The transformation of the matrix A to the matrix B is called a *similarity transformation*.

Definition 2.13.2 (Diagonalizable): A square matrix A is said to be *diagonalizable* if there exists a matrix C such that $D = C^{-1}AC$ is a diagonal matrix.

Theorem 2.13.3: Let A be an $n \times n$ matrix. If A has n linearly independent eigenvectors, then A is diagonalizable. The matrix C, whose columns consist of n linearly independent eigenvectors can be used in a similarity transformation $C^{-1}AC$ to give a diagonal matrix D. The diagonal elements of D will be the eigenvalues of A.

Also, if A is diagonalizable, then A has n linearly independent eigenvectors. □

Let's look at an example for $A = \begin{pmatrix} -4 & -6 \\ 3 & 5 \end{pmatrix}$.

The eigenvalues and eigenvectors of A are

$$\mathbf{v}_1 = \begin{pmatrix} -1 \\ 1 \end{pmatrix} \qquad \text{for } \lambda_1 = 2$$

$$\mathbf{v}_2 = \begin{pmatrix} -2 \\ 1 \end{pmatrix} \qquad \text{for } \lambda_2 = -1$$

Since $C = (\mathbf{v}_1 \mathbf{v}_2)$, we diagonalize A as

$$\begin{aligned} D &= \begin{pmatrix} -1 & -2 \\ 1 & 1 \end{pmatrix}^{-1} \begin{pmatrix} -4 & -6 \\ 3 & 5 \end{pmatrix} \begin{pmatrix} -1 & -2 \\ 1 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 2 \\ -1 & -1 \end{pmatrix} \begin{pmatrix} -4 & -6 \\ 3 & 5 \end{pmatrix} \begin{pmatrix} -1 & -2 \\ 1 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 2 & 0 \\ 0 & -1 \end{pmatrix} \end{aligned}$$

2.13.1 Similar matrices and Matrix Powers

If A is similar to the matrix D under the transformation $C^{-1}AC$, then

$$A^k = CD^kC^{-1} \tag{2.101}$$

Finding D^k for a square diagonal matrix is easy: you just raise each d_{ii} to the k-th power.

2.14 Positive Definite Matrices – 3/18/2009

These notes come from Howard Anton and Chris Rorres, *Elementary Linear Algebra: Applications version*, 9th edition, John Wiley and Sons, 2005. Pages 481–485.

A portion of chapter 9 is devoted to quadratic forms. For example,

$$a_1x_1^2 + a_2x_2^2 + a_3x_3^2 + a_4x_1x_2 + a_5x_1x_3 + a_6x_2x_3 \tag{2.102}$$

is a quadratic form. Quadratic forms can be written as $\mathbf{x}'A\mathbf{x}$. For example, a three variable equation like (2.102) can be written as

$$\begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \begin{pmatrix} a_1 & a_4/2 & a_5/2 \\ a_4/2 & a_2 & a_6/2 \\ a_5/2 & a_6/2 & a_3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \tag{2.103}$$

The coefficients of squared terms appear on the diagonal; the off-diagonal elements come from cross product terms. Note that the matrix A in (2.103) is symmetric.

Theorem 2.14.1: Let A be a symmetric $n \times n$ matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. If $\mathbf{x}$ is constrained so that $\|\mathbf{x}\| = 1$ (i.e., $\mathbf{x}$ is a unit vector), then

(a) $\lambda_1 \geq \mathbf{x}'A\mathbf{x} \geq \lambda_n$, and
(b) $\mathbf{x}'A\mathbf{x} = \lambda_1$ if $\mathbf{x}$ is an eigenvector corresponding to λ_1; $\mathbf{x}'A\mathbf{x} = \lambda_n$ is $\mathbf{x}$ is an eigenvector corresponding to λ_n.

□

Theorem 2.14.2: A symmetric matrix A is positive definite if and only if all of the eigenvalues of A are positive. □

I'll summarize the proof below.

Suppose A is positive definite. Then for any eigenvalue of A, we have $A\mathbf{x} = \lambda\mathbf{x}$. Therefore

$$\begin{aligned} 0 &\leq \mathbf{x}'A\mathbf{x} \\ &= \mathbf{x}'\lambda\mathbf{x} && \text{since } A\mathbf{x} = \lambda\mathbf{x} \\ &= \lambda\mathbf{x}'\mathbf{x} \\ &= \lambda\sum_{i=1}^{n} x_i^2 \end{aligned}$$

In the last line, λ must be > 0 because the summation $\sum_{i=1}^{n} x_i^2 > 0$.

Correspondingly, assume all of the eigenvalues of A are positive. We can scale any vector $\mathbf{x}$ to be a unit vector $\hat{\mathbf{x}}$. Since $\hat{\mathbf{x}}$ is a unit vector, Theorem 2.14.1 tells us that $\hat{\mathbf{x}}'A\hat{\mathbf{x}} \geq \lambda_n > 0$. Therefore A is positive definite.

Definition 2.14.3 (Principal Submatrix): Let A be an $n \times n$ matrix. The *principal submatrices* of A are the matrices formed by the first r rows and columns of A, for $1 \leq r \leq n$.

For example, the principal submatrices of an $n \times n$ matrix A are

$$A_1 = \begin{pmatrix} a_{11} \end{pmatrix}$$

$$A_2 = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

$$A_3 = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

$$\vdots$$

$$A_n = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}$$

Theorem 2.14.4: A symmetric matrix A is positive definite if and only if the determinant of every principal submatrix is positive.

We looked at the if part of this theorem in class, but not the only if part.

Finally, Anton and Rorres mention two techniques for eliminating cross-product terms in quadratic forms: Lagrange's Reduction, and Kronecker's Reduction. Anton and Rorres don't describe these techniques. Instead, they say that the techniques can be found in "more advanced texts".

2.14.1 Proving Theorem 2.14.1

In order to prove Theorem 2.14.1, we'll need some background material. (This material also comes from Anton and Rorres.)

Definition 2.14.5 (Inner Product): If $\mathbf{u} = \{u_1, \dots, u_n\}$ and $\mathbf{v} = \{v_1, \dots, v_n\}$ are two vectors in $\mathbb{R}^n$, then the inner product of $\mathbf{u}$ and $\mathbf{v}$ is

$$\langle u, v \rangle = u_1 v_1 + u_2 v_2 + \ldots + u_n v_n \tag{2.104}$$

(Anton and Rorres, page 296.) □

Theorem 2.14.6: If $S = \{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ is an orthonormal basis for an inner product space V, and $\mathbf{u}$ is any vector in V, then

$$\mathbf{u} = \langle \mathbf{u}, \mathbf{v}_1 \rangle \mathbf{v}_1 + \langle \mathbf{u}, \mathbf{v}_2 \rangle \mathbf{v}_2 + \ldots + \langle \mathbf{u}, \mathbf{v}_n \rangle \mathbf{v}_n \tag{2.105}$$

(Anton and Rorres, theorem 6.3.1, page 319.) □

Theorem 2.14.7 (Properties of Symmetric Matrices): If A is an $n \times n$ matrix, then the following statements are equivalent:

1. A is orthogonally diagonalizable
2. A has an orthnormal set of n eigenvectors
3. A is symmetric

(Anton and Rorres, Theorem 7.3.1, page 381.) □

Now, we can give a proof of Theorem 2.14.1. (From Anton and Rorres, pages 484–485.)

Since A is symmetric, the eigenvectors of A are an orthonormal basis for $\mathbb{R}^n$. Let $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ be such a basis, where each $\mathbf{v}_i$ corresponds to the eigenvalue λ_i. For any $\mathbf{x} \in \mathbb{R}^n$, we have

$$\begin{aligned}
\mathbf{x} &= \langle \mathbf{x}, \mathbf{v}_1 \rangle \mathbf{v}_1 + \langle \mathbf{x}, \mathbf{v}_2 \rangle \mathbf{v}_2 + \ldots + \langle \mathbf{x}, \mathbf{v}_n \rangle \mathbf{v}_n \\
A\mathbf{x} &= \langle \mathbf{x}, \mathbf{v}_1 \rangle A\mathbf{v}_1 + \langle \mathbf{x}, \mathbf{v}_2 \rangle A\mathbf{v}_2 + \ldots + \langle \mathbf{x}, \mathbf{v}_n \rangle A\mathbf{v}_n \\
&= \langle \mathbf{x}, \mathbf{v}_1 \rangle \lambda_1 \mathbf{v}_1 + \langle \mathbf{x}, \mathbf{v}_2 \rangle \lambda_2 \mathbf{v}_2 + \ldots + \langle \mathbf{x}, \mathbf{v}_n \rangle \lambda_n \mathbf{v}_n \\
&= \lambda_1 \langle \mathbf{x}, \mathbf{v}_1 \rangle \mathbf{v}_1 + \lambda_2 \langle \mathbf{x}, \mathbf{v}_2 \rangle \mathbf{v}_2 + \ldots + \lambda_n \langle \mathbf{x}, \mathbf{v}_n \rangle \mathbf{v}_n
\end{aligned}$$

From this, it follows that

$$\begin{aligned}
\|\mathbf{x}\|^2 &= \langle \mathbf{x}, \mathbf{v}_1 \rangle^2 + \langle \mathbf{x}, \mathbf{v}_2 \rangle^2 + \ldots + \langle \mathbf{x}, \mathbf{v}_n \rangle^2 \\
&= 1 && \mathbf{x} \text{ is a unit vector}
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{x}' A\mathbf{x} &= \langle \mathbf{x}, A\mathbf{x} \rangle \\
&= \lambda_1 \langle \mathbf{x}, \mathbf{v}_1 \rangle^2 + \lambda_2 \langle \mathbf{x}, \mathbf{v}_2 \rangle^2 + \ldots + \lambda_n \langle \mathbf{x}, \mathbf{v}_n \rangle^2 \\
&\leq \lambda_1 \langle \mathbf{x}, \mathbf{v}_1 \rangle^2 + \lambda_1 \langle \mathbf{x}, \mathbf{v}_2 \rangle^2 + \ldots + \lambda_1 \langle \mathbf{x}, \mathbf{v}_n \rangle^2 && \lambda_1 \text{ is largest} \\
&= \lambda_1 (\langle \mathbf{x}, \mathbf{v}_1 \rangle^2 + \langle \mathbf{x}, \mathbf{v}_2 \rangle^2 + \ldots + \langle \mathbf{x}, \mathbf{v}_n \rangle^2) \\
&= \lambda_1
\end{aligned}$$

Therefore, $\lambda_1 \geq \mathbf{x}' A\mathbf{x}$.

The proof that $\lambda_n \leq \mathbf{x}' A\mathbf{x}$ is similar.

Part 3

Applications of Singular Value Decomposition

3.1 Lecture – 3/23/2009

3.1.1 A Review of SVD

SVD allows us to factor an $m \times n$ matrix A into $A = UDV^{\mathsf{H}}$, where U and V are unitary matrices, and D is a diagonal matrix.

$D = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_p)$ contains the p singular values of A. Also, $\operatorname{rank}(A) = p$.

We've also seen that

- A can be written as the sum of p rank one matrices, where each rank one matrix has the form $\sigma_i \mathbf{u}_i \mathbf{v}_i^{\mathsf{H}}$.
- The eigenvalues of AA^{H} and $A^{\mathsf{H}}A$ are non-negative.
- The rank one matrices $\sigma_i \mathbf{u}_i \mathbf{v}_i^{\mathsf{H}}$ are all orthogonal to each other.

$\mathbb{C}^{m \times n}$ is itself a linear space (closed under addition and scalar multiplication.)

Given $\mathbf{u}_i \mathbf{v}_i^{\mathsf{H}}$ and $\mathbf{u}_j \mathbf{v}_j^{\mathsf{H}}$, their scalar product is

$$\begin{aligned}&(\mathbf{u}_i \mathbf{v}_i^{\mathsf{H}})^{\mathsf{H}} (\mathbf{u}_j \mathbf{v}_j^{\mathsf{H}}) \\ &= \mathbf{v}_i \mathbf{u}_i^{\mathsf{H}} \mathbf{u}_j \mathbf{v}_j^{\mathsf{H}}\end{aligned}$$

Since each rank one matrix is orthogonal,

$$\mathbf{u}_i^{\mathsf{H}} \mathbf{u}_j = 0 \text{ if } i \neq j$$

Therefore

$$\mathbf{v}_i \mathbf{u}_i^{\mathsf{H}} \mathbf{u}_j \mathbf{v}_j^{\mathsf{H}} = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$$

Also,

$$\begin{aligned}\|\mathbf{u}_i \mathbf{v}_i^{\mathsf{H}}\|_2 &= 1 \\ \|\mathbf{u}_i \mathbf{v}_i^{\mathsf{H}}\|_F &= 1\end{aligned}$$

3.1.2 Vector Model of Document Retrieval

The vector model of document retrieval was invented by G. Salton. We start with a *corpus* of documents $K = (T, D)$. T is a set of terms, and D is a set of documents.

Each document $d \in D$ is itself a sequence of terms.

$$d_i = (t_{i_1}, t_{i_2}, \ldots)$$

Document terms do not include stop words. Document terms may also be *stemmed*. (For example, "activities" and "activity" might be stemmed to "activ".)

We can represent the corpus K as a matrix A, where each row is a term and each column is a document. If $|T| = m$, and $|D| = n$, then A is an $m \times n$ matrix, where a_{ij} gives the frequency of term t_i in the document $\mathbf{d}_j$.

If there are m terms, then each document $\mathbf{d}_i$ is a vector in $\mathbb{R}^m$.

The simplest form of retrieval works as follows:

- We start with a query $\mathbf{q}$, which is a vector of length m.

$$\mathbf{q} = (q_1, q_2, \ldots, q_m)$$
$$q_i = \begin{cases} 1 & \text{if term } t_i \text{ appears in } \mathbf{q} \\ 0 & \text{otherwise} \end{cases}$$

The distance $\|\mathbf{d} - \mathbf{q}\|$ doesn't work well. $\mathbf{d}$ is a vector of frequencies, and $\mathbf{q}$ is a binary vector. The two aren't compatible.

The next best thing we can use is a dissimilarity measure. Cosine is a common dissimilarity measure:

$$0 \leq \cos(\mathbf{d}_j, \mathbf{q}) \leq 1$$

The closer that $\cos(\mathbf{d}_j, \mathbf{q})$ is to one, the better the match.

With a cosine dissimilarity, the set of documents retrieved will be

$$\{\mathbf{d} \mid \cos(\mathbf{d}, \mathbf{q}) > \tau\}$$

for some threshold τ.

But the cosine dissimilarity doesn't work well either, for two reasons:

Synonymy Several (lexically) different terms can have the same meaning. For example, "doctor" and "physician".

Polysemy The same (lexical) word can have different meanings.

Retrieval systems based on the vector model tend to be very noisy.

3.1.3 Latent Semantic Indexing

Latent Semantic Indexing is one solution to vector-based document retrieval.

Suppose we have two documents: one document uses the term "doctor" and the other document uses the term "physician". Although these documents use synonyms, chances are that they will have a large number of other terms in common.

Real noise does not have a preference for any direction. Noise tends to be uniformly distributed.

Suppose we use SVD to approximate the document/term matrix A:

$$A \approx \sigma_1 \mathbf{u}_1 \mathbf{v}_1^{\mathsf{H}} + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^{\mathsf{H}} + \ldots$$

by dropping terms associated with the smallest singular values, we'll get rid of a lot of noise.

- Noise is equally distributed among the p components of A.
- If we renounce the low weight components, we'll get rid of a lot of noise, and we'll still have a pretty good approximation of the documents.

These two points are *engineering assumptions*. Engineering assumptions are validated by experimentation, not by formal proofs.

An Example

Let's work out an example in Scilab. For this example, we'll use three documents and five terms.

```
A  =
   1.    0.    0.
   0.    1.    0.
   1.    1.    1.
   1.    1.    0.
   0.    0.    1.
```

To get an SVD decomposition of A:

```
-->[U,D,V] = svd(A)

 V  =
  - 0.6571923    0.2609565    0.7071068
  - 0.6571923    0.2609565  - 0.7071068
  - 0.3690482  - 0.9294103  - 2.351E-16

 D  =
   2.3582945    0.           0.
   0.           1.1993528    0.
   0.           0.           1.
   0.           0.           0.
   0.           0.           0.

 U  =
  - 0.2786727    0.2175811    0.7071068  - 0.5313508    0.3044114
  - 0.2786727    0.2175811  - 0.7071068  - 0.5313508    0.3044114
  - 0.7138349  - 0.3397643  - 2.544E-16  - 0.1098849  - 0.6024328
  - 0.5573454    0.4351621  - 3.490E-17    0.6412357    0.2980214
  - 0.1564894  - 0.7749265  - 2.195E-16    0.1098849    0.6024328
```

Let's compute two approximations:

$$B_1 = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^{\mathsf{H}}$$
$$B_2 = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^{\mathsf{H}} + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^{\mathsf{H}}$$

```
-->B1 = D(1,1) * U(:,1) * V(:,1)'

 B1  =
   0.4319017    0.4319017    0.2425356
```

```
    0.4319017    0.4319017    0.2425356
    1.1063391    1.1063391    0.6212678
    0.8638034    0.8638034    0.4850713
    0.2425356    0.2425356    0.1361966

-->B2 = B1 + D(2,2) * U(:,2) * V(:,2)'

 B2  =
    0.5          0.5          4.996E-16
    0.5          0.5        - 1.665E-16
    1.           1.           1.
    1.           1.           3.331E-16
    3.608E-16  - 2.776E-17    1.
```

Notice how the first two columns of B_2 have "evened out".

Let's say that we have a query $\mathbf{q}$

$$\mathbf{q} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}$$

```
-->q = [1 0 0 1 0]'

 q  =
    1.
    0.
    0.
    1.
    0.
```

Let's compute cosine differences between $\mathbf{q}$ and the three columns of B_2. We use the formula

$$\cos(\mathbf{q}, \mathbf{b}) = \frac{\mathbf{q}' \cdot \mathbf{b}}{\|\mathbf{q}\|\|\mathbf{b}\|}$$

```
--> c1 = (q' * B2(:,1))/(norm(q) * norm(B2(:,1)))

 c1  =
    0.6708204

--> c2 = (q' * B2(:,2))/(norm(q) * norm(B2(:,2)))

 c2  =
    0.6708204

-->c3 = (q' * B2(:,3))/(norm(q) * norm(B2(:,3)))

 c3  =
    4.163E-16
```

With $\tau = 0.67$, we'd get documents $\mathbf{b}_1$ and $\mathbf{b}_2$, but not $\mathbf{b}_3$.

3.1.4 Other Rank-one Decompositions

Say we are given an $m \times n$ matrix A.

For any vectors $\mathbf{x}$ and $\mathbf{y}$, $\mathbf{x}\mathbf{y}^{\mathsf{H}}$ is a matrix of rank one.

$A\mathbf{x}\mathbf{y}^{\mathsf{H}}A$ is also a matrix of rank one. This comes from the Wedderburn Theorem.

Let $B = A - k$. Is there a k such that $\text{rank}(B) = \text{rank}(A) - 1$? The answer turns out to be "yes". And of course, we can apply the same procedure to B, giving a matrix C where $\text{rank}(C) = \text{rank}(B) - 1 = \text{rank}(A) - 2$.

This sounds a lot like what SVD does, but it's not the same as SVD. SVD gives you the closest approximation for a given rank, but SVD is not the only way to get such approximations.

For example, you might want to decompose $A = UV$ where $\|\|A - UV\|\|$ is minimal, and U, V are positive matrices. Different decompositions can allow you to meet different goals.

3.1.5 Logistics

- Our exam will be a take-home exam, given at the end of April.

3.2 Notes on Text Mining – 3/25/2009

These notes come from Chapter 11 of *Matrix Methods in Data Mining and Pattern Recognition* by Lars Eldén.

Text Mining is the process of extracting useful information from large and often unstructured collections of text. Text mining is closely related to the subject of information retrieval.

Eldén's chapter draws example from an IR test collection called Medline.

A *term* is a keyword that carries information about the contents of a document.

An *inverted index* is a list of documents that contain a particular term (or set of terms).

Stop words are words that one can find in nearly any document. Stop words are not interesting for text mining or information retrieval.

Stemming is the process of reducing each word that is conjugated or suffixed. For consistency, any stemming algorithm should be applied to the stop word set.

Document term indexes are often weighted. A common weighting is

$$a_{ij} = f_{ij} \log(\tfrac{n}{n_i})$$

where

f_{ij} is the number of times that term i appears in document j.
n_i is the number of documents that contain term i, and
n is the number of documents.

Thus, when a word appears in few documents, it's weighted more heavily in the documents that contain it.

Document term matrices tend to be very sparse. It's not unusual for 99% of the cells to be zero.

Query matching is "good" when the intersection between relevant and retrieved documents is large, and the number of irrelevant retrieved documents is small.

Precision is

$$P = \frac{D_r}{D_t} \tag{3.1}$$

where

D_r is the number of relevant documents retrieved, and
D_t is the total number of documents retrieved

Recall is

$$R = \frac{D_r}{N_r} \tag{3.2}$$

where

D_r is the number of relevant documents retrieved, and
N_r is the total number of relevant documents in the database.

For cosine measures,

- A large tolerance τ tends to give high precision and low recall.
- A low tolerance τ tends to give low precision and high recall.

Latent Semantic Indexing (LSI) is "based on the assumption that there is some underlying latent semantic structure in the data ... that is corrupted by the wide variety of terms used."

Clustering (e.g., k-means) is another method for finding low-rank approximations of a document term matrix. One clusters documents on the basis of having similar terms.

3.3 Lecture – 3/25/2009

Today, we will begin to look at Principal Component Analysis (PCA). PCA is heavily used in the field of statistics; we'll look at it in more of an algebraic sense.

3.3.1 A Review of Hermetian Matrices

An matrix $A \in \mathbb{C}^{n \times n}$ is *Hermetian* if $A = A^{\mathsf{H}}$.

If $A \in \mathbb{R}^{n \times n}$ and $A' = A$, then we say that A is *symmetric*.

The eigenvalues of a Hermetian matrix are real numbers. If A is Hermetian then $\text{spec}(A) \in \mathbb{R}$. ($\text{spec}(A)$ is the *spectrum* of A – the set of A's eigenvalues.)

We can prove that $\text{spec}(A) \in \mathbb{R}$ as follows:

$$A\mathbf{v} = \lambda\mathbf{v} \tag{3.3}$$

$$\mathbf{v}^{\mathsf{H}}A^{\mathsf{H}} = \lambda\mathbf{v}^{\mathsf{H}} \quad \text{take Hermetian adjoint of each side} \tag{3.4}$$

$$\mathbf{v}^{\mathsf{H}}A^{\mathsf{H}}A\mathbf{v} = \lambda^2\mathbf{v}\mathbf{v}^{\mathsf{H}} \quad \text{“square” each side} \tag{3.5}$$

$$\|A\mathbf{v}\|_2^2 = \lambda^2\|\mathbf{v}\|_2^2 \tag{3.6}$$

$$\lambda^2 = \frac{\|A\mathbf{v}\|^2}{\|\mathbf{v}\|^2} \tag{3.7}$$

λ^2 is non-negative, so λ is real.

In the world of real numbers, if A is symmetric, then A has real eigenvalues.

Symmetric matrices can be diagonalized. If A is a Hermetian matrix, then there is a unitary matrix U such that $A = U^{\mathsf{H}}DU$, where D is a diagonal matrix.

This is a result of the *Schur Triangularization Theorem*.

If $A \in \mathbb{C}^{n \times n}$, then A can be factored to $A = U^{\mathsf{H}}TU$, where T is an upper-triangular matrix and U is a unitary matrix. (Recall: a *unitary* matrix is one where $U^{\mathsf{H}}U = UU^{\mathsf{H}} = I$.)

If A is Hermetian, then we can prove T is Hermetian.

If T is both Hermetian and Upper Triangular, then we can prove that T is diagonal.

Note:

$$A = U^{\mathsf{H}}TU \tag{3.8}$$

$$UA = UU^{\mathsf{H}}TU \quad \text{left-multiply both sides by } U \tag{3.9}$$

$$UAU^{\mathsf{H}} = UU^{\mathsf{H}}TUU^{\mathsf{H}} \quad \text{right-multiply both sides by } U^{\mathsf{H}} \tag{3.10}$$

$$UAU^{\mathsf{H}} = ITUU^{\mathsf{H}} \quad \text{since } UU^{\mathsf{H}} = I \tag{3.11}$$

$$UAU^{\mathsf{H}} = ITI \tag{3.12}$$

$$UAU^{\mathsf{H}} = T \tag{3.13}$$

If $A \in \mathbb{R}^{n \times n}$ is symmetric, then we can factor $A = U^{\mathsf{H}}DU$ where U is an *orthonormal* matrix.

An orthonormal matrix is a unitary matrix with real components.

3.3.2 Similar Matrices

Definition 3.3.1 (Similar Matrices): A and B are *similar matrices* if there exists a matrix P such that $A = P^{-1}BP$. We say that A is similar to B, or $A \sim B$.

If two matrices are similar, then they have the same characteristic polynomial, and the same spectrum.

For a diagonal matrix D, the spectrum (eigenvalues) appear along the diagonal:

$$\sum \text{spec}(D) = \text{trace}(D)$$

3.3.3 An Introduction to Principal Component Analysis

First, we'll present some of the intuition behind PCA.

Suppose we perform an experiment n times, and our experiment measures p variables.

We can plot these points in three-dimensional space, but the results will tend to be divergent. If the data points are scattered, then it is difficult to detect trends in the data.

Sometimes, we can discover linear relationships by rotating the data points in space. The different perspective makes the relationship more obvious.

PCA allows us to find the directions where most of the data scattering occurs.

Suppose we have a set of experiments, $\{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n\}$. (The superscripts denote the experiment number – they're not exponents.)

The result of each experiment is a column vector of measurements: $\{v_1, v_2, \ldots, v_p\}$.

Let's treat variables as rows, and experiments as columns. The set of all experimental results is

$$X = \begin{pmatrix} \mathbf{x}^1 & \mathbf{x}^2 & \ldots & \mathbf{x}^n \end{pmatrix} \qquad \text{or equivalently}$$

$$= \begin{pmatrix} x_1^1 & x_1^2 & \ldots & x_1^n \\ x_2^1 & x_2^2 & \ldots & x_2^n \\ \vdots & \vdots & \ddots & \vdots \\ x_p^1 & x_p^2 & \ldots & x_p^n \end{pmatrix}$$

X is a data matrix with p rows for variables, and n columns for experiments.

Statisticians would call X a *sample*.

The *mean sample* is $\tilde{\mathbf{x}}$.

$$\tilde{\mathbf{x}} = \frac{1}{n}(\mathbf{x}^1 + \mathbf{x}^2 + \ldots + \mathbf{x}^n) \tag{3.14}$$

Given a set of vectors $\{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n\}$ and a vector $\mathbf{z} \in \mathbb{R}^p$, the *inertia* of the sample with respect to $\mathbf{z}$ is

$$I_z = \sum_{i=1}^{n} \|\mathbf{x}^i - \mathbf{z}\|^2 \tag{3.15}$$

We are often interested in the difference in inertia between $\mathbf{z}$ and $\tilde{\mathbf{x}}$, $I_z - I_{\tilde{x}}$.

Let's examine $I_z - I_{\tilde{x}}$.

$$\begin{aligned}
I_z - I_{\tilde{x}} &= \sum_{i=1}^{n} (\mathbf{x}^i - \mathbf{z})'(\mathbf{x}^i - \mathbf{z}) - \sum_{i=1}^{n} (\mathbf{x}^i - \tilde{\mathbf{x}})'(\mathbf{x}^i - \tilde{\mathbf{x}}) \\
&= -\mathbf{z}' \sum_{i=1}^{n} \mathbf{x}^i - \left(\sum_{i=1}^{n} \mathbf{x}^i \right)' \mathbf{z} + n\mathbf{z}'\mathbf{z} + \tilde{\mathbf{x}}' \sum_{i=1}^{n} \mathbf{x}^i + \left(\sum_{i=1}^{n} \mathbf{x}^i \right)' \tilde{\mathbf{x}} - n\tilde{\mathbf{x}}'\tilde{\mathbf{x}} \\
&= -\mathbf{z}' n\tilde{\mathbf{x}} - n\tilde{\mathbf{x}}'\mathbf{z} + n\mathbf{z}'\mathbf{z} + n\tilde{\mathbf{x}}'\tilde{\mathbf{x}} + n\tilde{\mathbf{x}}'\tilde{\mathbf{x}} - n\tilde{\mathbf{x}}'\tilde{\mathbf{x}} \\
&= n\mathbf{z}'\mathbf{z} + n\tilde{\mathbf{x}}'\tilde{\mathbf{x}} - 2n\mathbf{z}'\tilde{\mathbf{x}} \\
&= n(\mathbf{z}'\mathbf{z} + \tilde{\mathbf{x}}'\tilde{\mathbf{x}} - 2\mathbf{z}'\tilde{\mathbf{x}}) \\
&= n\|\mathbf{z} - \tilde{\mathbf{x}}\|_2^2
\end{aligned}$$

The equation

$$I_z - I_{\tilde{x}} = n\|\mathbf{z} - \tilde{\mathbf{x}}\|^2 \tag{3.16}$$

is called Huygen's Formula.

For a set of points, the mean sample has the smallest inertia.

The mean sample is the vector $\tilde{\mathbf{x}}$ such that X has minimal inertial relative to $\tilde{\mathbf{x}}$.

If $\tilde{\mathbf{x}} = \mathbf{0}$, then the set of points is *centered*.

We can always center a set of points via translation:

$$(\mathbf{x}^1 - \tilde{\mathbf{x}}, \mathbf{x}^2 - \tilde{\mathbf{x}}, \ldots, \mathbf{x}^n - \tilde{\mathbf{x}}) \tag{3.17}$$

3.3.4 Covariance Matrix of X

Assume that X is centered. The *covariance matrix* of X is

$$\text{cov}(X) = \frac{1}{n-1} XX' \tag{3.18}$$

In terms of matrix dimensions, this is

$$\underset{(p \times p)}{\text{cov}(X)} \quad = \quad \underset{(p \times n)}{X} \quad \cdot \quad \underset{(n \times p)}{X'} \tag{3.19}$$

The format of $\text{cov}(X)$ depends on the number of variables observed, *not* on the number of experiments done.

Properties of $\text{cov}(X)$:

- $\text{cov}(X)$ is symmetric.
- $\text{cov}(X)$ has non-negative eigenvalues.
- $\text{cov}(X)$ is orthogonally diagonalizable
- $\text{cov}(X) = U'DU$ for an orthonormal matrix U

$(\text{cov}(X))_{ii}$ give the variance of row $\mathbf{v}_i$.

$(\text{cov}(X))_{ij}$ is the variance between $\mathbf{v}_i$ and $\mathbf{v}_j$.

trace$(\text{cov}(X))$ is the total variance.

3.3.5 Project/Presentation Note

Project presentations should fit into a 75 minute class. Assume one hour for the presentation, with the remaining time for question and answer.

3.4 Lecture – 3/30/2009

3.4.1 A review of PCA

Let's begin by reviewing PCA material from our last lecture.

We start with a sample matrix $X = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$. (Note: this time we are using subscripts – not superscripts – to denote different experiments.) Each experiment is a vector $\mathbf{x}_i \in \mathbb{R}^p$.

The columns of X are experiments; the rows of X are variables.

We also have a mean vector $\tilde{\mathbf{x}}$.

$$\tilde{\mathbf{x}} = \tfrac{1}{n}(\mathbf{x}_1 + \mathbf{x}_2 + \ldots + \mathbf{x}_n) \tag{3.20}$$

If $\tilde{\mathbf{x}} = \mathbf{0}$, then we say the matrix X is *centered*.

Inertia

The inertia of X with respect to $\mathbf{z}$ is

$$I_z(X) = \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{z}\|_2^2 \tag{3.21}$$

When $\mathbf{z} = \tilde{\mathbf{x}}$, the inertia is at a minimum.

The principal directions of X are those directions where the inertia is minimal.

3.4.2 Orthonormal Basis

Suppose $U \in \mathbb{R}^n$ is a linear space. U is closed under addition and scalar multiplication.

The set $\{a\mathbf{u} \mid a \in \mathbb{R}\}$ is a one-dimensional subspace.

The set $\{a\mathbf{u} + b\mathbf{v} \mid a, b \in \mathbb{R}\}$ is a two-dimensional subspace (assuming that $\mathbf{u}$, $\mathbf{v}$ are linearly independent).

Every subspace contains the origin.

Every subspace has a set of orthonormal basis vectors.

If U is a subspace, then an orthonormal basis for U is the set of vectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ such that $\|\mathbf{u}_i\| = 1$ and $\mathbf{u}_i \cdot \mathbf{u}_j' = 0$ if $i \neq j$.

$\mathbf{u}_i \cdot \mathbf{u}_j' = 1$ if $i = j$ (since $\|\mathbf{u}_i\| = 1$).

3.4.3 Projection

We use $\text{proj}_U(\mathbf{x})$ to denote the projection of $\mathbf{x}$ onto the subspace U. Note that we must have $\text{proj}_U(\mathbf{x}) \in U$.

Note 3.4.1: In this section, we'll be denoting projection as

$$(\mathbf{x}, \mathbf{u})\mathbf{u} = \mathbf{x}\mathbf{u}' \cdot \mathbf{u} \tag{3.22}$$

I'm used to thinking of projection as

$$\frac{\mathbf{x} \cdot \mathbf{u}}{\mathbf{u} \cdot \mathbf{u}} \cdot \mathbf{u} \tag{3.23}$$

where $\cdot$ denotes dot product. But, since we're talking about orthonormal vectors $\mathbf{u} \cdot \mathbf{u} = 1$, so (3.22) is really equivalent to (3.23). □

Because $\text{proj}_U(\mathbf{x}) \in U$, we can write $\mathbf{x}$ as a linear combination of the orthonormal basis vectors of U.

$$\text{proj}_U(\mathbf{x}) = a_1\mathbf{u}_1 + a_2\mathbf{u}_2 + \ldots a_m\mathbf{u}_m \tag{3.24}$$

The projection $\text{proj}_U(\mathbf{x})$ should also have minimal distance from $\mathbf{x}$. In other words, we want to minimize $\|\mathbf{x} - \text{proj}_U(\mathbf{x})\|_2^2$.

We can write $\|\mathbf{x} - \text{proj}_U(\mathbf{x})\|_2^2$ as

$$\begin{aligned}
&(\mathbf{x} - (a_1\mathbf{u}_1 + \ldots + a_m\mathbf{u}_m))' \cdot (\mathbf{x} - (a_1\mathbf{u}_1 + \ldots + a_m\mathbf{u}_m)) \\
=&\|\mathbf{x}\|_2^2 - a\mathbf{u}_1'\mathbf{x} - \ldots a_m\mathbf{u}_m'\mathbf{x} - a_1\mathbf{x}'\mathbf{u}_1 - \ldots - a_m\mathbf{x}'\mathbf{u}_m + \sum_{i=1}^{m} a_i^2 \\
=&\|\mathbf{x}\|_2^2 - 2\sum_{i=1}^{m} a_i(\mathbf{x}, \mathbf{u}_i) + \sum_{i=1}^{m} a_i^2
\end{aligned}$$

If we take partial derivatives, we would like

$$\frac{\partial d}{\partial a_i} = 0$$

The partial derivative with respect to a_i is

$$\begin{aligned}
&-2(\mathbf{x}, \mathbf{u}_i) + 2a_i \\
\therefore\ &a_i = (\mathbf{x}, \mathbf{u}_i)
\end{aligned}$$

Therefore, projection is

$$\text{proj}_U(\mathbf{x}) = \sum_{i=1}^{m} (\mathbf{x}, \mathbf{u}_i)\mathbf{u}_i \tag{3.25}$$

(But see also Note: 3.4.1.)

Let B be a matrix of orthonormal column vectors,

$$B = \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \ldots & \mathbf{u}_m \end{pmatrix}$$

Each $\mathbf{u}_i \in \mathbb{R}^n$, so B is an $n \times m$ matrix.

Let's look at the product $B'B$. Since $\mathbf{u}_i$ are orthonormal column vectors, we have

$$\mathbf{u}_i \cdot \mathbf{u}_j' = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$$

Therefore $B'B$ is

$$\begin{aligned} B'B &= \begin{pmatrix} \mathbf{u}_1' \\ \mathbf{u}_2' \\ \vdots \\ \mathbf{u}_m' \end{pmatrix} \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \dots & \mathbf{u}_m \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix} \\ &= I_m && \text{an } m \times m \text{ matrix} \end{aligned}$$

Now, let's look at BB':

$$BB' = \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \dots & \mathbf{u}_m \end{pmatrix} \begin{pmatrix} \mathbf{u}_1' \\ \mathbf{u}_2' \\ \vdots \\ \mathbf{u}_m' \end{pmatrix} \qquad \text{an } n \times n \text{ matrix}$$

We claim that $BB'\mathbf{x} = \text{proj}_U(\mathbf{x})$. Why? Recall that B consists of orthonormal column vectors. Therefore,

$$\begin{aligned} BB'\mathbf{x} &= \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \dots & \mathbf{u}_m \end{pmatrix} \begin{pmatrix} \mathbf{u}_1' \\ \mathbf{u}_2' \\ \vdots \\ \mathbf{u}_m' \end{pmatrix} \mathbf{x} \\ &= \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \dots & \mathbf{u}_m \end{pmatrix} \begin{pmatrix} \mathbf{u}_1'\mathbf{x} \\ \mathbf{u}_2'\mathbf{x} \\ \vdots \\ \mathbf{u}_m'\mathbf{x} \end{pmatrix} \\ &= a_1\mathbf{u}_1 + a_2\mathbf{u}_2 + \ldots + a_n\mathbf{u}_m \end{aligned}$$

If $m = 1$, then $\text{proj}_{\langle U \rangle}(\mathbf{x}) = \mathbf{u}\mathbf{u}'\mathbf{x}$.

Also, BB' is an idempotent matrix. We prove this as follows:

$$\begin{aligned} (BB')^2 &= BB'(BB') \\ &= B(B'B)B' && \text{assoc. property} \\ &= BIB' && \text{since } B'B = I \text{ (see above)} \\ &= BB' \end{aligned}$$

Therefore $(BB')^2 = BB'$.

3.4.4 PCA and Covariance Matrices

Recall that PCA starts with a matrix $X = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_n \end{pmatrix}$, and a mean vector $\tilde{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$.

Let's define a matrix $\hat{X}$ as

$$\hat{X} = \begin{pmatrix} \mathbf{x}_1 - \tilde{\mathbf{x}} & \mathbf{x}_2 - \tilde{\mathbf{x}} & \dots & \mathbf{x}_n - \tilde{\mathbf{x}} \end{pmatrix}$$

The mean of $\hat{X}$ will be zero, and $\hat{X}$ is centered.

$\hat{X}$ allows us to define the covariance matrix of X.

$$\text{cov}(X) = \frac{1}{n-1}\hat{X}\hat{X}' \tag{3.26}$$

If X is already centered, then $\hat{X} = X$ and $\text{cov}(X) = \frac{1}{n-1}XX'$.

Given n experiments and p variables, $\text{cov}(X)$ is a $p \times p$ matrix. It depends on the number of variables – not on the number of experiments.

$(\text{cov}(X))_{ii}$ gives the variance of variable v_i.

$$\begin{aligned}(\text{cov}(X))_{ii} &= \frac{1}{n-1}\sum_{j=1}^{n}(x_{ij} - \tilde{x}_i)^2 \\ &= \text{var}(v_i)\end{aligned}$$

$(\text{cov}(X))_{ik}$ gives the covariance between v_i and v_k.

The correlation coefficient of v_i, v_k is

$$\text{corr}(v_i, v_k) = \frac{(\text{cov}(X))_{ik}}{\sqrt{\text{var}(v_i)\,\text{var}(v_k)}}$$

We have $-1 \leq \text{corr}(v_i, v_k) \leq 1$. A value of one denotes strong positive correlation; a value of negative one denotes strong negative correlation; a value of zero means that v_i, v_k are uncorrelated (but not necessarily independent).

$\text{cov}(X)$ is a symmetric matrix. Therefore $\text{cov}(X) = \text{cov}(X)'$. $\text{cov}(X)$ has real eigenvalues, and we can diagonalize it.

$\text{cov}(X)$ is symmetric. Therefore, there is an orthonormal $p \times p$ matrix P such that $D = P\,\text{cov}(X)P'$, where D is a diagonal matrix.

Orthonormal matrices are the real-number equivalent of unitary matrices. Because P is orthonormal, $P'P = PP' = I$.

By the above

$$\begin{aligned}D &= P\,\text{cov}(X)P' \\ &= \frac{1}{n-1}PXX'P'\end{aligned}$$

Let $Z = PX$. Z is linearly related to X.

$$\begin{aligned}D &= \frac{1}{n-1}PXX'P' \\ &= \frac{1}{n-1}ZZ' \qquad \text{because } Z = PX \qquad = \text{cov}(Z) \\ \therefore\ \text{cov}(Z) &= P\,\text{cov}(X)P'\end{aligned}$$

This tells us that $\text{cov}(Z)$ is a diagonal matrix. The variables in $Z = PX$ are uncorrelated (their correlation coefficients are zero). However, the variables in Z are not necessarily independent.

$$\begin{aligned}\sum_{i=1}^{n}(\text{cov}(X))_{ii} &= \text{trace}(\text{cov}(X)) \\ &= \text{ the total variance} \\ \text{trace}(\text{cov}(X)) &= \text{trace}(\text{cov}(Z))\end{aligned}$$

Z and X are *similar* matrices. They have the same characteristic polynomial; therefore they have the same trace.

The columns of P are the principal directions of X.

Given a matrix X, the inertial of $\text{proj}_{\langle \mathbf{z} \rangle}(X)$ reaches extreme values on the principal components. ($\mathbf{z}$ is one of the principal directions.) This is one of the optimality theorems for PCA.

3.4.5 To-Do

- Look at PCA functions in Scilab and R.
- hw3 and the next handout are on the course web site.

3.5 Notes on PCA – 4/1/2009

Some these notes come from one of the course handouts. There are a few bits from *Matrices and Transformations* by Anthony J. Pettofrezzo, Dover, 1978. There are a few of my own observations as well.

Claim 3.5.1: The matrix XX' is symmetric.

Proof

$$\begin{aligned} XX' &= (X''X')' && \text{by properties of transpose} \\ &= (XX')' && \text{since } X'' = X \end{aligned}$$

$\therefore XX'$ is symmetric. □

By the above reasoning, the matrix

$$\text{cov}(X) = \frac{1}{n-1} XX'$$

is also symmetric. Because $\text{cov}(X)$ is symmetric, $\text{cov}(X)$ is orthonormally diagonalizable, which is to say that there are matrices P, D such that

$$D = P' \,\text{cov}(X) P$$

Above,

- D is a diagonal matrix, and
- P is an orthonormal matrix. (i.e., $PP' = P'P = I$)

In D, each diagonal element d_i gives the variance associated with the i principal component.

The columns of P are the eigenvectors of $\text{cov}(X)$. These columns are the principal components of X.

Principal components "explain" the sources of total variance. Thus, the largest diagonal elements of D are associated with the greatest sources of variance.

3.5.1 SVD and PCA

There is a singular value decomposition associated with PCA.

$$\begin{aligned} P' \,\text{cov}(X) P &= D \\ PP' \,\text{cov}(X) P &= PD && \text{left multiply by } P \\ PP' \,\text{cov}(X) PP' &= PDP' && \text{right multiply by } P' \\ I \,\text{cov}(X) I &= PDP' && P \text{ is orthonormal, so } PP' = I \\ \text{cov}(X) &= PDP' \end{aligned}$$

This tells us the following

- The columns of P are the eigenvectors of $\text{cov}(X)$ (noted above).
- Each $d_i \in D$ is an eigenvalue of $\text{cov}(X)$
- Each variance $d_i \in D$ is the square of a singular value of X.
- $\text{cov}(X)$ can be written as the sum of rank-one matrices. Each rank-one matrix $d_i \mathbf{p}_i \mathbf{p}_i'$ corresponds to a principal component.

So, it seems that PCA is really just a special case of SVD.

3.6 Lecture – 4/1/2009

Today, we'd like to look at $I_O(\text{proj}_{\mathbf{u}}(X))$. This is the inertia of X projected onto $\mathbf{u}$, with respect to the origin (we use I_O to represent inertia with respect to the origin).

Let A be a matrix $A \in \mathbb{R}^{n \times n}$, and let $\mathbf{x}$ be a vector in $\mathbb{R}^n$. We define a function $f(\mathbf{x})$:

$$f(\mathbf{x}) = \mathbf{x}' A \mathbf{x}$$

or equivalently

$$f(x_1, \ldots, x_n) = \begin{pmatrix} x_1 & \ldots & x_n \end{pmatrix} A \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

To find the extremums of $f(\mathbf{x})$ we need to find the column vector of points where the partial derivatives $\frac{\partial f}{\partial x_i}$ are zero. (This set of partial derivatives is called the *gradient* of f.)

$$\text{grad}(f) = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix} \qquad \text{gradient is a vector}$$

$$\text{grad}(f) = \mathbf{0} \qquad \text{gives extremum points}$$

For illustration, if $n = 3$, $\mathbf{x}' A \mathbf{x}$ looks like this:

$$\begin{aligned} \mathbf{x}' A \mathbf{x} &= \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \\ &= \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \begin{pmatrix} x_1 a_{11} + x_2 a_{12} + x_3 a_{13} \\ x_1 a_{21} + x_2 a_{22} + x_3 a_{23} \\ x_1 a_{31} + x_2 a_{32} + x_3 a_{33} \end{pmatrix} \\ &= x_1(x_1 a_{11} + x_2 a_{12} + x_3 a_{13}) + x_2(x_1 a_{21} + x_2 a_{22} + x_3 a_{23}) + x_3(x_1 a_{31} + x_2 a_{32} + x_3 a_{33}) \end{aligned}$$

We can also write $f(\mathbf{x})$ as a summation

$$f(\mathbf{x}) = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i a_{ij} x_j$$

The partial derivative for x_k looks like this:

$$\frac{\partial f}{\partial x_k} = \begin{pmatrix} x_1 & x_2 & \ldots & x_n \end{pmatrix} \begin{pmatrix} a_{1k} \\ \vdots \\ a_{nk} \end{pmatrix} + \begin{pmatrix} a_{1k} & \ldots & a_{nk} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

The column a_{ik} vector represents the k-th column of A. The row a_{ik} vector represents the k-th row of A.

Therefore,

$$\text{grad}(f) = \text{grad}(\mathbf{x}'A\mathbf{x}) = (A' + A)\mathbf{x} \tag{3.27}$$

Now, let's return to $I_O(\text{proj}_{\mathbf{u}}(X))$.

Assuming that $\|\mathbf{u}\|_2 = 1$, we'd like to find $\mathbf{u}$ such that $I_O(\text{proj}_{\mathbf{u}}(X))$ is at an extremum.

Last class, we noted that

$$\text{proj}_{\mathbf{x}}(\mathbf{x}) = \mathbf{u}\mathbf{u}'\mathbf{x}$$

For the inertia with respect to the origin,

$$\begin{aligned}
I_O(\text{proj}_{\mathbf{u}}(X)) &= \sum_{i=1}^{n} \|\text{proj}_{\mathbf{u}}(\mathbf{x}_i)\|_2^2 \\
&= \sum_{i=1}^{n} \|\mathbf{u}\mathbf{u}'\mathbf{x}\|_2^2 \\
&= \sum_{i=1}^{n} \mathbf{x}_i'\mathbf{u}\mathbf{u}' \cdot \mathbf{u}\mathbf{u}'\mathbf{x}_i \\
&= \sum_{i=1}^{n} \mathbf{x}_i'\mathbf{u}(\mathbf{u}'\mathbf{u})\mathbf{u}'\mathbf{x}_i \\
&= \sum_{i=1}^{n} \mathbf{x}_i'\mathbf{u}\mathbf{u}'\mathbf{x}_i && \text{since } \mathbf{u}'\mathbf{u} = 1 \\
&= \sum_{i=1}^{n} \mathbf{u}'\mathbf{x}_i\mathbf{x}_i'\mathbf{u}
\end{aligned}$$

Note that $\mathbf{u}'\mathbf{x}_i\mathbf{x}_i'\mathbf{u}$ fits nicely with our earlier formula for finding the gradient of $\mathbf{x}'A\mathbf{x}$.

The condition we want to meet is

$$\text{grad}\left(I_O(\text{proj}_{\mathbf{u}}(X) + \lambda(1 - \mathbf{u}'\mathbf{u}))\right) = 0 \tag{3.28}$$

(The term $\lambda(1 - \mathbf{u}'\mathbf{u})$ is a Lagrange coefficient.)

For the right term of (3.28),

$$\begin{aligned}
\text{grad}(1 - \mathbf{u}'\mathbf{u}) &= \text{grad}(1 - \mathbf{u}'I\mathbf{u}) \\
&= -(I' + I)\mathbf{u} \\
&= -2\mathbf{u}
\end{aligned}$$

For the left term of (3.28), let $B_i = \mathbf{x}_i\mathbf{x}_i'$. B_i is a symmetric matrix.

$$\begin{aligned}
\text{grad}\left(\sum_{i=1}^{n} \mathbf{u}'\mathbf{x}_i\mathbf{x}_i'\mathbf{u}\right) &= \text{grad}\left(\sum_{i=1}^{n} \mathbf{u}'B_i\mathbf{u}\right) \\
&= \sum_{i=1}^{n} (B_i' + B_i)\mathbf{u} && \text{from equation (3.27)} \\
&= \sum_{i=1}^{n} 2B_i\mathbf{u}
\end{aligned}$$

Therefore, our condition is

$$2\sum_{i=1}^{n} B_i\mathbf{u} - 2\lambda\mathbf{u} = 0 \qquad \text{or equivalently}$$

$$\sum_{i=1}^{n} B_i\mathbf{u} = \lambda\mathbf{u}$$

3.6.1 Relation Between SVD and PCA

We know that $\text{cov}(X)$ is a symmetric matrix. Therefore it can be written as

$$\text{cov}(X) = d_1\mathbf{u}_1\mathbf{u}_1' + d_2\mathbf{u}_2\mathbf{u}_2' + \ldots + d_p\mathbf{u}_p\mathbf{u}_p'$$

where $\mathbf{u}_1, \ldots, \mathbf{u}_p$ are orthogonal vectors.

3.7 More Notes on Positive Definite Matrices

These notes come from Steven J. Leon, *Linear Algebra with Applications*, 6th edition, pub. Prentice Hall, 2002, pp. 390–403.

Let $f(\mathbf{x})$ be the function

$$f(\mathbf{x}) = \mathbf{x}'A\mathbf{x} \tag{3.29}$$

We say that f is

definite if $f(\mathbf{x})$ takes on one sign for all $\mathbf{x} \neq \mathbf{0}$.

indefinite if $f(\mathbf{x})$ takes on values that differ in sign.

positive definite if $f(\mathbf{x}) > 0$ for all $\mathbf{x} \neq \mathbf{0}$.

positive semi-definite if $f(\mathbf{x}) \geq 0$ for all $\mathbf{x} \neq \mathbf{0}$.

negative definite if $f(\mathbf{x}) < 0$ for all $\mathbf{x} \neq \mathbf{0}$.

negative semi-definite if $f(\mathbf{x}) \leq 0$ for all $\mathbf{x} \neq \mathbf{0}$.

Theorem 3.7.1: Let A be a real symmetric $n \times n$ matrix. Then A is positive definite if and only if all of its eigenvalues are positive. □

If the eigenvalues of A are all negative, then A is negative definite and $-A$ is positive definite.

Theorem 3.7.1 allows us to establish additional properties of positive definite matrices.

Property 1 If A is a symmetric positive definite matrix, then A is non-singular.

Property 2 If A is a symmetric positive definite matrix, then $\det(A) > 0$.

Property 3 If A is a symmetric positive definite matrix, then the leading principal submatrices $A_1, \ldots, A_n$ if A are all positive definite.

Property 4 If A is a symmetric positive definite matrix, then A can be reduced to an upper triangular form (Gaussian Elimination) without exchanging rows, and the pivot elements will all be positive.

Property 5 If A is a symmetric positive definite matrix, then A can be factored into a product LDL', where L is lower triangular with 1's along the diagonal, and D is a diagonal matrix whose diagonal entries are all positive.

Property 6 If A is a symmetric positive definite matrix, then A can be factored into a product LL', where L is lower triangular with positive diagonal elements. This is a Cholesky factorization of A.

Property 7 A can be factored in $A = B'B$ for some non-singular matrix B. (Note: this is essentially the same thing as a Cholesky decomposition; let $B = L'$.)

The following statements are equivalent:

1. A is positive definite.
2. The leading principal submatrices of A are positive definite.
3. A can be reduced to upper triangular form without swapping rows, and the pivot elements will all be positive. (i.e., we can do Gaussian Elimination by repeatedly replacing a row by its sum with the multiple of another row.)
4. A has a Cholesky decomposition $A = LL'$ where L is lower triangular with positive diagonal entries.
5. A can be factored into $A = B'B$ for some non-singular matrix B (let $B = L'$).

3.8 Lecture – 4/6/2009

3.8.1 hw2 in review

One of hw2's problems was the following:

> Let $A \in \mathbb{C}^{n \times n}$ be a matrix such that $A^2 = A$. Prove that $\text{rank}(A) + \text{rank}(I - A) = n$.

Some people tried to solve the problem by doing manipulations with A^{-1}. That is not a valid approach. If $\text{rank}(A) \neq n$ then A^{-1} won't even exist.

Here is a correct solution. We know that

$$n = \text{rank}(A) + \dim(\text{null}(A)) \tag{3.30}$$

Therefore, we can prove $\text{rank}(A) + \text{rank}(I - A) = n$ by proving that $\text{range}(I - A) = \text{null}(A)$.

Let $\mathbf{x} \in \text{range}(I - A)$. Then there is some vector $\mathbf{t}$ such that $\mathbf{x} = (I - A)\mathbf{t}$. With that as a starting point,

$$\begin{aligned}
\mathbf{x} &= (I - A)\mathbf{t} && \text{from above} \\
A\mathbf{x} &= A(I - A)\mathbf{t} && \text{left multiply by } A \\
A\mathbf{x} &= (A - A^2)\mathbf{t} && \text{distribute RHS} \\
A\mathbf{x} &= (A - A)\mathbf{t} && \text{since } A^2 = A \\
A\mathbf{x} &= 0^{n \times n}\mathbf{t} \\
A\mathbf{x} &= \mathbf{0}
\end{aligned}$$

Therefore $\mathbf{x} \in \text{range}(I - A)$ implies $\mathbf{x} \in \text{null}(A)$.

Going in the opposite direction. Suppose that $\mathbf{x} \in \text{null}(A)$. We have

$$\begin{aligned}
(I - A)\mathbf{x} &= I\mathbf{x} - A\mathbf{x} \\
&= I\mathbf{x} - \mathbf{0} && \text{since } \mathbf{x} \in \text{null}(A) \\
&= I\mathbf{x} \\
&= \mathbf{x}
\end{aligned}$$

Therefore $\mathbf{x} \in \text{null}(A)$ implies $\mathbf{x} \in \text{range}(I - A)$.

3.8.2 Least Squares Method

Let X be a sample matrix, so that $X = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_n \end{pmatrix}$.

Let's assume that our experiments have two variables, so $\mathbf{x}_i \in \mathbb{R}^2$, and $\mathbf{x}_i = \begin{pmatrix} x_{1i} \\ x_{2i} \end{pmatrix}$.

If treat each vector $\mathbf{x}_i$ as a point in two dimensions, using the coordinates (x_{1i}, x_{2i}), then we can plot the results of several experiments on a graph. Figure 3.1 shows an example of such a graph.

In Figure 3.1, each point is an $\mathbf{x}_i$, and we drawn a "best fit" line $y = \alpha x + \beta$ through the set of points. The distance between $y = \alpha x + \beta$ and a point $\mathbf{x}_i$ is a *residual*.

Our goal is to find α, β that best predict y from x. The line acts as a model for our data.

How do we choose such a line?

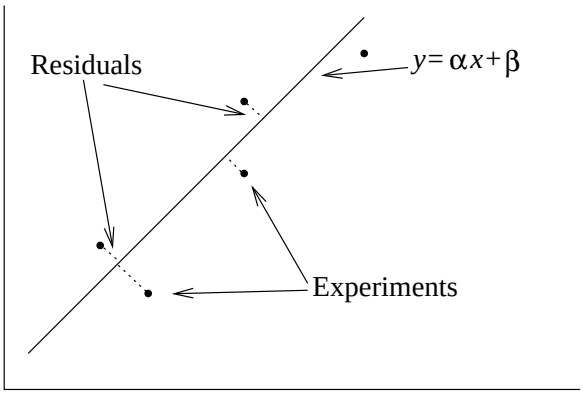

Figure 3.1: Graph of Experimental Data with linear model

One approach is to minimize the sum of the squares of the residuals:

$$\begin{aligned} x_{21} &= \alpha x_{11} + \beta \\ x_{22} &= \alpha x_{12} + \beta \\ &\vdots \\ x_{2n} &= \alpha x_{1n} + \beta \end{aligned}$$

Here, we have n equations and two indeterminants. It is an overdetermined system, so it will be very difficult to find an exact solution.

We can also express this system of equations in matrix form:

$$\begin{pmatrix} x_{11} & 1 \\ x_{12} & 1 \\ \vdots & \\ x_{1n} & 1 \end{pmatrix} \cdot \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} x_{21} \\ x_{22} \\ \vdots \\ x_{2n} \end{pmatrix} \tag{3.31}$$

Given $A \in \mathbb{C}^{m \times n}$ where $m \geq n$ (or even $m \gg n$), we can find an $\mathbf{x}$ such that $\|A\mathbf{x} - \mathbf{b}\|_2$ is minimal. This is the next best thing to an exact solution.

In our case, we are seeking α, β such that the 2-norm of (3.31) is minimal:

$$\left\| \begin{pmatrix} x_{11} & 1 \\ x_{12} & 1 \\ \vdots & \\ x_{1n} & 1 \end{pmatrix} \cdot \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} x_{21} \\ x_{22} \\ \vdots \\ x_{2n} \end{pmatrix} \right\|_2 \tag{3.32}$$

If we are minimizing a norm, then we are also minimizing the square of the norm. Therefore, we can also attempt to minimize $\|A\mathbf{x} - \mathbf{b}\|_2^2$. For us, this will be equivalent to minimizing equation (3.33).

$$\left[\begin{pmatrix} \alpha & \beta \end{pmatrix} \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ 1 & 1 & \dots & 1 \end{pmatrix} - \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \end{pmatrix} \right] \times \left[\begin{pmatrix} x_{11} & 1 \\ x_{12} & 1 \\ \vdots & \\ x_{1n} & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} - \begin{pmatrix} x_{21} \\ x_{22} \\ \vdots \\ x_{2n} \end{pmatrix} \right] \tag{3.33}$$

Minimizing (3.33) amounts to minimizing the sum of the squares of the residuals.

Assume $A \in \mathbb{R}^{n \times m}$ and $\mathbf{x}, \mathbf{b} \in \mathbb{R}^n$. Our minimization problem is

$$\begin{aligned}\|A\mathbf{x} - \mathbf{b}\|_2^2 &= (A\mathbf{x} - \mathbf{b})'(A\mathbf{x} - \mathbf{b}) \\ &= (\mathbf{x}'A' - \mathbf{b}')(A\mathbf{x} - \mathbf{b}) \\ &= \mathbf{x}'A'A\mathbf{x} - \mathbf{b}'A\mathbf{x} - \mathbf{x}'A'\mathbf{b} - \mathbf{b}'\mathbf{b} \\ &= \mathbf{x}'A'A\mathbf{x} - 2\mathbf{b}'A\mathbf{x} + \mathbf{b}'\mathbf{b}\end{aligned}$$

We would like to have

$$\text{grad}(\mathbf{x}'A'A\mathbf{x} - 2\mathbf{b}'A\mathbf{x} + \mathbf{b}'\mathbf{b}) = 0$$

We achieve this when

$$\begin{aligned}2A'A\mathbf{x} - 2\mathbf{b}'A &= \mathbf{0} \qquad \text{or equivalently, when} \\ A'A\mathbf{x} &= \mathbf{b}'A\end{aligned}$$

Assume that A is a full-rank matrix, such that $\text{rank}(A) = n$. Our original system was overdetermined, but $A'A\mathbf{x} = A'\mathbf{b}$ is solvable.

Note that we've assumed $A \in \mathbb{R}^{m \times n}$, with $m \gg n$. $A' \in \mathbb{R}^{n \times m}$, so $A'A \in \mathbb{R}^{n \times n}$. This is a much smaller matrix to work with.

We will solve this by using a QR decomposition.

3.8.3 QR Decomposition

We have $A \in \mathbb{R}^{m \times n}$. For now assume that A has full rank; $\text{rank}(A) = n$.

We will decompose A as

$$A = Q \begin{bmatrix} R \\ O \end{bmatrix}$$

Where

- $Q \in \mathbb{R}^{m \times m}$ and Q is orthonormal.
- R is an $n \times n$ matrix
- O is an $(m-n \times n)$ matrix. (O pads the space below R with zeros, to make the matrix multiplication conformant.

Because A has full rank, $A'A$ also has full rank (see Section 3.8.5, for a discussion of why this is the case). Because $A'A$ has full rank, we know that $A'A$ is non-singular, and $A'A\mathbf{x} = A'\mathbf{b}$ has a unique solution.

$A'A$ is

$$\begin{aligned}A'A &= \begin{bmatrix} R' & O^{n \times m-n} \end{bmatrix} Q' \cdot Q \begin{bmatrix} R \\ O^{m-n \times n} \end{bmatrix} \\ &= \begin{bmatrix} R' & O^{n \times m-n} \end{bmatrix} I \begin{bmatrix} R \\ O^{m-n \times n} \end{bmatrix} \qquad Q \text{ orthonormal, so } Q'Q = I\end{aligned}$$

We want to minimize

$$A\mathbf{x} - \mathbf{b} = Q \begin{bmatrix} R \\ O^{m-n \times n} \end{bmatrix} \mathbf{x} - \mathbf{b} \tag{3.34}$$

$$= Q \begin{bmatrix} R \\ O^{m-n \times n} \end{bmatrix} \mathbf{x} - QQ'\mathbf{b} \qquad \text{note: } QQ' = I \tag{3.35}$$

$$= Q \left(\begin{bmatrix} R \\ O^{m-n \times n} \end{bmatrix} \mathbf{x} - Q'\mathbf{b} \right) \tag{3.36}$$

Recall that multiplying a vector by an orthonormal matrix will not change the norm; multiplication by an orthonormal matrix only rotates a vector, without changing its length.

In other words, for an orthonormal matrix Q,

$$\|Q\mathbf{w}\| = \|\mathbf{w}\| \tag{3.37}$$

Therefore, starting with

$$Q\left(\begin{bmatrix} R \\ O^{m-n\times n} \end{bmatrix}\mathbf{x} - Q'\mathbf{b}\right)$$

we can minimize

$$\left\|\begin{bmatrix} R \\ O^{m-n\times n} \end{bmatrix}\mathbf{x} - Q'\mathbf{b}\right\|_2^2$$

since multiplication by Q will not change the norm.

Let's write $Q'\mathbf{b}$ as a pair of vectors: $Q'\mathbf{b} = \begin{pmatrix}\mathbf{b}_1 \\ \mathbf{b}_2\end{pmatrix}$. This gives

$$\begin{aligned}
&\left\|\begin{pmatrix} R \\ O^{m-n\times n} \end{pmatrix}\mathbf{x} - Q'\mathbf{b}\right\|_2^2 && \text{from above} \\
&= \left\|\begin{pmatrix} R \\ O \end{pmatrix} - \begin{pmatrix}\mathbf{b}_1 \\ \mathbf{b}_2\end{pmatrix}\right\|_2^2 \\
&= \left\|\begin{pmatrix} R\mathbf{x} - \mathbf{b}_1 \\ -\mathbf{b}_2 \end{pmatrix}\right\|_2^2 \\
&= \|R\mathbf{x} - \mathbf{b}_1\|_2^2 + \|\mathbf{b}_2\|_2^2
\end{aligned}$$

$\|\mathbf{b}_2\|_2^2$ is a constant, so we only need to solve $\|R\mathbf{x} - \mathbf{b}_1\|_2^2$.

So far, we've assumed that A is full rank. What happens if A is not full rank? We'll look at this in our next lecture.

3.8.4 Gradients

Recall that $\text{grad}(\mathbf{x}'C\mathbf{x}) = (C + C')\mathbf{x}$.

Also note that the gradient $\text{grad}(\mathbf{x}'\mathbf{d}) = \mathbf{d}$. Why?

$$\begin{aligned}
\frac{\partial \mathbf{x}'\mathbf{d}}{\partial \mathbf{x}_1} &= d_1 \\
\frac{\partial \mathbf{x}'\mathbf{d}}{\partial \mathbf{x}_2} &= d_2 \\
&\cdots \\
\frac{\partial \mathbf{x}'\mathbf{d}}{\partial \mathbf{x}_n} &= d_n
\end{aligned}$$

Therefore $\text{grad}(\mathbf{x}'\mathbf{d}) = \mathbf{d}$.

3.8.5 $A'A$, Rank, and Positive Definiteness

We can show that $\text{rank}(A'A) = \text{rank}(A)$. Recall that

$$n = \text{rank}(A) + \dim(\text{null}(A))$$
$$n = \text{rank}(A'A) + \dim(\text{null}(A'A))$$

If $\text{null}(A) = \text{null}(A'A)$, then we know that $\text{rank}(A) = \text{rank}(A'A)$.

Let $\mathbf{z}$ be a vector $\mathbf{z} \in \text{null}(A)$; we have $A\mathbf{z} = \mathbf{0}$. But then $\mathbf{z} \in \text{null}(A'A)$ because $A'A\mathbf{z} = A'\mathbf{0} = \mathbf{0}$.

Conversely, let $\mathbf{z} \in \text{null}(A'A)$. We have

$A'A\mathbf{z} = \mathbf{0}$	since $\mathbf{z} \in \text{null}(A'A)$
$\mathbf{z}'A'A\mathbf{z} = 0$	since $\mathbf{z}'\mathbf{0} = 0$
$(A\mathbf{z})'A\mathbf{z} = 0$	transpose LHS
$\|A\mathbf{z}\| = 0$	

$\|A\mathbf{z}\| = 0$ only holds for the zero vector; therefore $A\mathbf{z} = \mathbf{0}$ and $\mathbf{x} \in \text{null}(A)$. □

We can also show that $A'A$ is positive semi-definite. $A'A$ is symmetric, so $A'A$ will be positive semi-definite if $\mathbf{x}'A'A\mathbf{x} \geq 0$.

$$\begin{aligned}\mathbf{x}'A'A\mathbf{x} &= (A\mathbf{x})'A\mathbf{x} \\ &= \|A\mathbf{x}\|_2^2\end{aligned}$$

If $\mathbf{x} \neq \mathbf{0}$, then $\|A\mathbf{x}\|_2^2$ must be ≥ 0. Thus, $A'A$ is positive semi-definite. □

If $A\mathbf{x} = \mathbf{0}$ implies that $\mathbf{x} = \mathbf{0}$, then $A'A$ is positive definite.

3.9 More Notes on Matrix Diagonalization – 4/8/2009

These notes come from Gareth Williams, *Linear Algebra with Applications*, 6th edition, Jones & Bartlett, 2008, pages 292–298.

3.9.1 General Diagonalization

Let A be an $n \times n$, invertible matrix. Because A is invertible, A is non-singular and $\text{rank}(A) = n$.

We can write find a diagonal matrix D that is *similar* to A by using the decomposition

$$D = C^{-1}AC \tag{3.38}$$

Similar matrices have the same eigenvalues. Therefore $\text{spec}(D) = \text{spec}(A)$.

The matrix C consists of linearly independent eigenvectors of A. If A is diagonalizable, then A has n linearly independent eigenvectors.

The diagonal elements of D will be the eigenvalues of A.

If $\text{rank}(A) < n$, then A is not diagonalizable.

3.9.2 Diagonalization of Symmetric Matrices

If A is a symmetric $n \times n$ matrix, then the eigenvalues of A are real numbers, and A has n linearly independent eigenvectors.

If a matrix C is orthogonal, then $C^{-1} = C'$.

For a symmetric matrix, $D = C^{-1}AC$ becomes $D = C'AC$.

Definition 3.9.1 (Orthogonally Diagonalizable): A square matrix A is said to be *orthogonally diagonalizable* if there exists an orthogonal matrix C such that $D = C'AC$ is a diagonal matrix. □

The set of orthogonally diagonalizable matrices is, in fact, the set of symmetric matrices.

Theorem 3.9.2: Let A be a square matrix. A is orthogonally diagonalizable if and only if A is a symmetric matrix.

We can form $D = C'AC$ as follows:

1. Find a basis for each eigenspace of A
2. Find an orthonormal basis for each eigenspace. (Use the Gram-Schmidt algorithm if necessary.)
3. Let C be the matrix whose columns are these orthonormal vectors.
4. The matrix $D = C'AC$ will be a diagonal matrix.

Conversely, suppose A is orthogonally diagonalizable, such that $D = C'AC$. From $D = C'AC$, we know that $A = CDC'$

$$\begin{aligned}
A' &= (CDC')' \\
&= ((CD)C')' \\
&= C''(CD)' \\
&= CD'C' \\
&= CDC' && D \text{ is diagonal, so } D' = D \\
&= A
\end{aligned}$$

Therefore, A must be symmetric. □

Finally, if A is positive semi-definite, then the eigenvalues of A will be ≥ 0. Therefore the diagonal elements of D will be *positive* real numbers.

3.10 Lecture – 4/8/2009

In our last lecture, we discussed the problem of finding (approximate) solutions to the system $A\mathbf{x} = \mathbf{b}$, when A had full rank. The goal was to find a $\mathbf{x}$ that minimized $\|A\mathbf{x} - \mathbf{b}\|$.

Today, we will discuss the case where A does not have full rank.

3.10.1 Least Squares Approximation, Continued

Let $A \in \mathbb{R}^{m \times n}$ where $m \gg n$, and let $\text{rank}(A) = r$.

We can decompose A as follows:

$$A = U \begin{pmatrix} R & O \\ O & O \end{pmatrix} V' \tag{3.39}$$

In (3.39),

- U is an $m \times m$ orthonormal matrix
- R is an $r \times r$ matrix ($r = \text{rank}(A)$)
- The middle term has dimensions $m \times n$
- V is an $n \times n$ orthonormal matrix.

Because U, V are orthonormal, we'll have

$$UU' = U'U = I_m$$
$$VV' = V'V = I_n$$

The equation we seek to minimize is

$$\|A\mathbf{x} - \mathbf{b}\|_2 = \left\| U \begin{pmatrix} R & O \\ O & O \end{pmatrix} V'\mathbf{x} - \mathbf{b} \right\|_2 \tag{3.40}$$

In (3.40), $\mathbf{b}$ is a vector $\mathbf{b} \in \mathbb{R}^m$.

We continue to manipulate (3.40) as follows:

$$\|A\mathbf{x} - \mathbf{b}\|_2 = \left\| U \begin{pmatrix} R & O \\ O & O \end{pmatrix} V'\mathbf{x} - \mathbf{b} \right\|_2 \quad \text{from above} \tag{3.41}$$

$$= \left\| U \begin{pmatrix} R & O \\ O & O \end{pmatrix} V'\mathbf{x} - UU'\mathbf{b} \right\|_2 \quad \text{since } UU' = I \tag{3.42}$$

$$= \left\| U \left(\begin{pmatrix} R & O \\ O & O \end{pmatrix} V'\mathbf{x} - U'\mathbf{b} \right) \right\|_2 \quad \text{factor out } U \tag{3.43}$$

$$= \left\| \begin{pmatrix} R & O \\ O & O \end{pmatrix} V'\mathbf{x} - U'\mathbf{b} \right\|_2 \quad U\text{, orthonormal, does not change the norm} \tag{3.44}$$

Let $V'\mathbf{x} = \mathbf{y}$, so that $\mathbf{x} = V\mathbf{y}$. Substituting this into (3.44) gives

$$\left\| \begin{pmatrix} R & O \\ O & O \end{pmatrix} \mathbf{y} - U'\mathbf{b} \right\|_2 \tag{3.45}$$

In (3.45), when we multiply $\begin{pmatrix} R & O \\ O & O \end{pmatrix} \mathbf{y}$ the "bottom" of the resulting vector will be zeros. Because R is an $r \times r$ matrix, the first r elements of $\mathbf{y}$ will survive, but the bottom $n - r$ elements will not.

Let's break $\mathbf{y}$ into two parts: those that will survive and those that will not:

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} \qquad \mathbf{y}_1 \in \mathbb{R}^r,\ \mathbf{y}_2 \text{ goes to zero} \tag{3.46}$$

Substituting (3.46) into (3.45) gives

$$\left\| \begin{pmatrix} R\mathbf{y}_1 \\ O \end{pmatrix} - U'\mathbf{b} \right\|_2 \tag{3.47}$$

We can break up $U'\mathbf{b}$ in the same fashion: the first r elements, and the last $n-r$ elements.

$$U'\mathbf{b} = \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix} \qquad \mathbf{b}_1 \in \mathbb{R}^r,\ \mathbf{b}_2 \in \mathbb{R}^{n-r} \tag{3.48}$$

Substituting (3.48) into (3.47) gives

$$\left\| \begin{pmatrix} R\mathbf{y}_1 \\ O \end{pmatrix} - \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix} \right\|_2 \tag{3.49}$$

$$= \left\| \begin{pmatrix} R\mathbf{y}_1 - \mathbf{b}_1 \\ -\mathbf{b}_2 \end{pmatrix} \right\|_2 \tag{3.50}$$

$$= \|R\mathbf{y}_1 - \mathbf{b}_1\|_2^2 + \|\mathbf{b}_2\|_2^2 \tag{3.51}$$

$$= A\mathbf{x} + \mathbf{b} \tag{3.52}$$

In (3.50), the $\|\mathbf{b}_2\|_2^2$ term is a constant. So, we only need to be concerned with solving $R\mathbf{y} = \mathbf{b}_1$.

Note that R has rank r, and R is a non-singular matrix.

To find $\mathbf{x}$,

$$\mathbf{x} = V\mathbf{y} \qquad \text{since } \mathbf{y} = V'\mathbf{x} \tag{3.53}$$

$$= V \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{w} \end{pmatrix} \qquad \text{break } \mathbf{y} \text{ into two components} \tag{3.54}$$

$$= \begin{pmatrix} V\mathbf{y}_1 \\ V\mathbf{w} \end{pmatrix} \tag{3.55}$$

In (3.55), $\mathbf{w}$ is arbitrary. Finding $\mathbf{y}_1$ gives us a solution for $\mathbf{x}$.

3.10.2 Scaling

Suppose we have a data set with high dimensionality, and we'd like to present this information in an understandable way. In general, people find low-dimension spaces easier to understand than high-dimension spaces.

Therefore, we'd like to transform points in a high dimensional space to points in a low dimensional space. In general, this is a difficult thing to do – changing dimensionality distorts the distances between points.

Say we have (S, δ) where S is a set of objects, and δ is a dissimilarity on S.

A *dissimilarity* is a function $\delta\colon S \times S \to \mathbb{R}_{\geq 0}$ that satisfies two properties.

- $\delta(s, s) = 0$. (definiteness)
- $\delta(s, t) = \delta(t, s)$. (symmetry)

Dissimilarities are rather weak measures:

- A dissimilarity does not satisfy the triangular inequality.
- $\delta(u, v) = 0$ does not imply that $u = v$.

Scaling starts with a dissimilarity space (S, δ), and tries to represent objects in S as objects in $\mathbb{R}^p$, where p is "small".

Each object $s \in S$ becomes a point $x \in \mathbb{R}^p$.

Say we have a *scaling function* $f\colon S \to \mathbb{R}^p$. f should preserve, to the greatest extent possible, the original dissimilarity measure. In other words, for s_i, $s_j \in S$, we would like

$$d(f(s_i), f(s_j)) \approx \delta(s_i, s_j)$$

(Above, d is a distance.)

There are several types of scaling. Two common types are *metric scaling* and *non-metric scaling*.

- **Metric scaling**. Metric scaling tries to preserve as much of the original dissimilarity as possible.
- **Non-metric scaling**. Non-metric scaling tries to preserve the relative order in the original dissimilarity.

  Suppose we are given $\{x, y, u, v\} \in S$ and $\{f(x), f(y), f(u), f(v)\} \in \mathbb{R}^p$. If $\delta(x, y) < \delta(u, v)$, then non-metric scaling tries to preserve this as $d(f(x), f(y)) < d(f(u), f(v))$.

We will focus on metric scaling.

Set $S = \{s_1, \ldots, s_n\}$ be a dissimilarity space. We'd like to map the objects in S to vectors, using the lowest dimension possible.

$$f\colon S \to \mathbb{R}^p \qquad \text{for } p \ll n$$

Let $d_{ij} = \delta(s_i, s_j)$. d_{ij} can be specified in terms of an $n \times n$ matrix D. This is our original dissimilarity measure.

Let $f(s_i) = \mathbf{x}_i$. f is our scaling function to vectors in $\mathbb{R}^p$. Ideally, we'd like the scaled distances to match the original dissimilarity.

$$\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 = d_{ij}^2 \tag{3.56}$$

We would also like our vectors to be centered around the origin, such that $\sum_{i=1}^n \mathbf{x}_i = \mathbf{0}$.

For now, we'll assume that such a mapping f exists.

Let's multiply out (3.56).

$$\begin{aligned} d_{ij}^2 &= \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 && (3.57)\\ &= (\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j) && (3.58)\\ &= \mathbf{x}_i'\mathbf{x}_i - \mathbf{x}_j'\mathbf{x}_i - \mathbf{x}_i'\mathbf{x}_j + \mathbf{x}_j'\mathbf{x}_j && (3.59)\\ &= \|\mathbf{x}_i\|_2^2 - 2\mathbf{x}_i'\mathbf{x}_j + \|\mathbf{x}_j\|_2^2 && (3.60) \end{aligned}$$

We can sum d_{ij} over i and j:

$$\sum_{i=1}^n d_{ij}^2 = \sum_{i=1}^n \|\mathbf{x}_i\|_2^2 + n\|\mathbf{x}_j\|_2^2 \qquad \text{See note below} \tag{3.61}$$

$$\sum_{j=1}^n d_{ij}^2 = n\|\mathbf{x}_i\|_2^2 + \sum_{j=1}^n \|\mathbf{x}_j\|_2^2 \tag{3.62}$$

Note: In (3.61) and (3.62), $2\mathbf{x}_i'\mathbf{x}_j$ disappears, since our system of vectors is centered around the origin. Also note that $\sum_{i=1}^n \|\mathbf{x}_i\|_2^2 = \sum_{j=1}^n \|\mathbf{x}_j\|_2^2$. Therefore,

$$\sum_{i=1}^n \sum_{j=1}^n d_{ij}^2 = 2n \sum_{i=1}^n \|\mathbf{x}_i\|_2^2 \tag{3.63}$$

Suppose we are given vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$. The *Gram matrix* of the sequence $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is

$$\begin{pmatrix} \mathbf{x}_1'\mathbf{x}_1 & \ldots & \mathbf{x}_1'\mathbf{x}_n \\ \vdots & \ddots & \vdots \\ \mathbf{x}_n'\mathbf{x}_1 & \ldots & \mathbf{x}_n'\mathbf{x}_n \end{pmatrix} \qquad \text{Gram Matrix} \tag{3.64}$$

In (3.64), each term is the scalar product of two vectors.

The Gram Matrix is

- positive definite, if $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are linearly independent.
- positive semi-definite, if $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are not linearly independent.

We would like to compute the Gram Matrix of vectors representing the points in $\mathbb{R}^p$.

Let's work with $\mathbf{x}_i'\mathbf{x}_j$.

$$\mathbf{x}_i'\mathbf{x}_j = \frac{1}{2}\left(\|\mathbf{x}_i\|_2^2 + \|\mathbf{x}_j\|_2^2 - d_{ij}^2\right) \tag{3.65}$$

Equation (3.65) is the *cosine theorem*.

Let's take an $\mathbf{x}_i$ term from (3.61) and and $\mathbf{x}_j$ term from (3.62).

$$\mathbf{x}_i'\mathbf{x}_j = \frac{1}{2}\left[\frac{1}{n}\left(\sum_{j=1}^n d_{ij}^2 - \sum_{j=1}^n \|\mathbf{x}_j\|_2^2\right) + \frac{1}{n}\left(\sum_{i=1}^n d_{ij}^2 - \sum_{i=1}^n \|\mathbf{x}_i\|_2^2\right) - d_{ij}^2\right] \tag{3.66}$$

Noting that $\sum_{i=1}^n \|\mathbf{x}_i\|_2^2 = \frac{1}{2n}\sum_{i=1}^n \sum_{j=1}^n d_{ij}$, we continue with

$$= \frac{1}{2}\left[\frac{1}{n}\sum_{j=1}^n d_{ij}^2 + \frac{1}{n}\sum_{i=1}^n d_{ij}^2 - \frac{2}{n}\cdot\frac{1}{n}\left(\sum_{i=1}^n \sum_{j=1}^n d_{ij}^2\right) - d_{ij}^2\right] \tag{3.67}$$

$$= \frac{1}{2}\left[\frac{1}{n}\left(\sum_{j=1}^n d_{ij}^2 + \sum_{i=1}^n d_{ij}^2\right) - \frac{2}{n^2}\left(\sum_{i=1}^n \sum_{j=1}^n d_{ij}^2\right) - d_{ij}\right] \tag{3.68}$$

If $X = \begin{pmatrix}\mathbf{x}_1 & \mathbf{x}_2 & \ldots & \mathbf{x}_n\end{pmatrix}$, then the Gram matrix is $X'X$. Therefore, we can compute $X'X$ starting from a dissimilarity (S, δ).

$X'X$ is a symmetric matrix, and $\text{rank}(X'X) = \text{rank}(X) = p$.

$X \in \mathbb{R}^{p \times n}$ and $p \ll n$.

We can decompose $X'X$ as

$$X'X = V \cdot \text{diag}(\sigma_1, \ldots, \sigma_p) \cdot V' \tag{3.69}$$

where the dimension are

$$\underset{(n\times n)}{X'X} = \underset{(n\times p)}{V} \cdot \underset{(p\times p)}{\text{diag}(\sigma_1,\ldots,\sigma_p)} \cdot \underset{(p\times n)}{V'}$$

Note that

$$X'X = V\cdot \text{diag}(\sigma_1,\ldots,\sigma_p)\cdot V' \tag{3.70}$$
$$= V\cdot \text{diag}(\sqrt{\sigma_1},\ldots,\sqrt{\sigma_p})\cdot \text{diag}(\sqrt{\sigma_1},\ldots,\sqrt{\sigma_p})\cdot V' \tag{3.71}$$
$$\therefore\ X = \text{diag}(\sqrt{\sigma_1},\ldots,\sqrt{\sigma_p})\cdot V' \tag{3.72}$$

So far, we've assumed that f exists, whereby $X'X$ is a positive definite matrix.

If f does not exist, then $X'X$ will not be positive definite. However, we'll be able to modify an $X'X$ so that it is positive definite. We'll look at this process in our next lecture.

3.11 Misc. Notes on Eigenvalues – 4/13/2009

> The material in this section comes from Howard Anton, *Elementary Linear Algebra*, 9th edition, Wiley, 2005. Pages 360–366.

The *characteristic equation* of a matrix A is $\det(\lambda I - A) = \mathbf{0}$. λ that satisfy this equation are eigenvalues.

Theorem 3.11.1: If A is an $n \times n$ matrix, then the eigenvalues of A are the entries on the main diagonal of A. □

Theorem 3.11.2: If k is a positive integer, λ is an eigenvalue of a matrix A, and $\mathbf{x}$ is the corresponding eigenvector, then λ^k is an eigenvalue of A^k and $\mathbf{x}$ is a corresponding eigenvector. □

Theorem 3.11.3: A square matrix A is invertible if and only if $\lambda = 0$ is not an eigenvalue of A. □

Implications of Theorem 3.11.3 A has full rank iff $\lambda = 0$ is not an eigenvalue of A. Conversely, if A does not have full rank, then $\lambda = 0$ will be an eigenvalue of A.

Theorem 3.11.4: If A is an $n \times n$ matrix, and $T_A : \mathbb{R}^n \to \mathbb{R}^n$ is a multiplication by A, then the following statements are equivalent.

1. A is invertible
2. $A\mathbf{x} = \mathbf{0}$ has only the trivial solution.
3. The reduced row-echelon form of A is I_n.
4. A is expressible as a product of elementary matrices.
5. $A\mathbf{x} = \mathbf{b}$ is consistent for every $n \times 1$ matrix $\mathbf{b}$.
6. $\det(A) \neq 0$.
7. The range of $T_A = \mathbb{R}^n$
8. T_A is one-to-one.
9. The column vectors of A are linearly independent.
10. The row vectors of A are linearly independent.
11. The column vectors of A span $\mathbb{R}^n$.
12. The row vectors of A span $\mathbb{R}^n$.
13. The column vectors of A form a basis for $\mathbb{R}^n$.
14. The row vectors of A form a basis for $\mathbb{R}^n$.
15. A has rank n.
16. $\text{rank}(\text{null}(A)) = 0$.
17. $A'A$ is invertible
18. $\lambda = 0$ is not an eigenvalue of A.

3.12 Lecture – 4/13/2009

3.12.1 A Brief Note About Positive (Semi) Definite Matrices

The positive definiteness of a matrix A has nothing to do with whether or not A is symmetric. A can be positive definite and asymmetric.

Suppose

$$\mathbf{x}'A\mathbf{x} \geq 0 \qquad \text{and}$$
$$\mathbf{x}'A'\mathbf{x} \geq 0$$

Adding these two equations gives

$$\mathbf{x}'\frac{1}{2}(A + A')\mathbf{x} \geq 0$$

and the matrix $\frac{1}{2}(A + A')$ *is* symmetric.

Claim 3.12.1: If $\mathbf{x}'A\mathbf{x} \geq 0$, there is always a symmetric matrix B such that

- $\mathbf{x}'A\mathbf{x} = \mathbf{x}'B\mathbf{x}$, and
- B is symmetric □

Say we treat $\mathbf{x}'A\mathbf{x}$ as a function of $\mathbf{x}$. Even if A is not symmetric, we can always find a matrix B that is.

3.12.2 Principal Component Regression

In statistics, the term *regression* has a specific meaning. Usually the term is used in the context of "linear regression", where we are trying to model a set of points with a straight line that best represents those points.

Principal Component Regression (PCR) is similar to the least squares method that we looked at recently.

Suppose we have a matrix $A \in \mathbb{R}^{n \times n}$ and a vector $\mathbf{b} \in \mathbb{R}^n$. We would like to express $\mathbf{b}$ as a linear combination of A's columns. In order for this to be possible, we must have $\mathbf{b} \in \text{range}(A)$.

We'd like to express $\mathbf{b}$ as a linear combination (or as an approximate linear combination) of the principal components of A. (For example, we might want to limit ourselves to the first k principal components.)

SVD and PCA

Let's assume $A \in \mathbb{C}^{p \times n}$.

We've seen how SVD is related to PCA. For SVD

$$A = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^{\mathsf{H}} + \ldots + \sigma_l \mathbf{u}_l \mathbf{v}_l^{\mathsf{H}} \qquad l = \text{rank}(A) \tag{3.73}$$
$$= UDV^{\mathsf{H}} \tag{3.74}$$
$$= \begin{pmatrix} \mathbf{u}_1 & \ldots & \mathbf{u}_p \end{pmatrix} \begin{pmatrix} \sigma_1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \sigma_p \end{pmatrix} \begin{pmatrix} \mathbf{v}_1^{\mathsf{H}} \\ \vdots \\ \mathbf{v}_n^{\mathsf{H}} \end{pmatrix} \tag{3.75}$$

In (3.75), U is $(p \times p)$, D is $(p \times n)$ and V is $(n \times n)$.

For PCA, we have

$$\text{cov}(A) = \frac{1}{n-1} AA^{\mathsf{H}} \tag{3.76}$$

Multiplying AA^{H} gives

$$AA^{\mathsf{H}} = (\sigma_1 \mathbf{u}_1 \mathbf{v}_1^{\mathsf{H}} + \ldots + \sigma_\ell \mathbf{u}_\ell \mathbf{v}_\ell^{\mathsf{H}}) \times (\sigma_1 \mathbf{v}_1 \mathbf{u}_1^{\mathsf{H}} + \ldots + \sigma_\ell \mathbf{v}_\ell \mathbf{u}_\ell^{\mathsf{H}}) \tag{3.77}$$

$$= \sigma_1^2 \mathbf{u}_1 \mathbf{v}_1^{\mathsf{H}} + \ldots + \sigma_\ell \mathbf{u}_\ell \mathbf{v}_\ell \tag{3.78}$$

We get (3.78) because each matrix $\mathbf{u}_i \mathbf{v}_i^{\mathsf{H}}$ is orthogonal. (If $i \neq j$, the multiplication makes the terms go to zero).

In general, we have

$$(AA^{\mathsf{H}})\mathbf{u}_k = \sigma_k^2 \mathbf{u}_k$$

and

$$\frac{1}{n-1}(AA^{\mathsf{H}})\mathbf{u}_k = \frac{\sigma_k^2}{n-1}\mathbf{u}_k$$

The principal components are the left singular vectors of A.

Let $A = UDV^{\mathsf{H}}$. Let's look at $\mathbf{u}_i \mathbf{v}_i^{\mathsf{H}}$ and $\mathbf{u}_j \mathbf{v}_j^{\mathsf{H}}$.

$$(\mathbf{u}_i \mathbf{v}_i^{\mathsf{H}})^{\mathsf{H}}(\mathbf{u}_j \mathbf{u}_j) = 0 \qquad \text{if } i \neq j$$

The matrices associated with each singular value are orthogonal.

Now Back to Principal Component Regression

We would like to regress $\mathbf{b}$ as a linear combination of $A\mathbf{x}$. Our goal is to minimize $\|A\mathbf{x} - \mathbf{b}\|_2$.

Let $A \in \mathbb{R}^{p \times n}$ and $b \in \mathbb{R}^n$.

Let $V = \{\mathbf{v}_1 \ldots \mathbf{v}_n\}$ be the set of column vectors of A.

We seek $\mathbf{x} \in \text{range}(\mathbf{v}_1 \ldots \mathbf{v}_k)$ for $k \leq n$.

Let $V_k = \begin{pmatrix} \mathbf{v}_1 & \ldots & \mathbf{v}_k \end{pmatrix}$. V_k consists of k of A's columns.

$\mathbf{y} \in \text{range}(V_k)$ means that $\mathbf{x} = V_k \mathbf{y}$ for some vector $\mathbf{y}$.

Also VV_k' is

$$VV_k' = \begin{pmatrix} \mathbf{v}_1 \\ \vdots \\ \mathbf{v}_n' \end{pmatrix} \begin{pmatrix} \mathbf{v}_1 & \ldots & \mathbf{v}_k \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 0 \end{pmatrix}$$

$$= \begin{pmatrix} I_k \\ O_{n-k,k} \end{pmatrix}$$

Using these decompositions, we can write

$$\begin{aligned}
\|A\mathbf{x}-\mathbf{b}\|_2 &= \|UDV'V_k\mathbf{y}-UU'\mathbf{b}\|_2 && \mathbf{x}=V_k\mathbf{y}, \text{ and } UU'=I && (3.79)\\
&= \|U(DV'V_k\mathbf{y}-U'\mathbf{b})\|_2 && \text{factor } U && (3.80)\\
&= \|DV'V_k\mathbf{y}-U'\mathbf{b}\| && U \text{ orthonormal. Doesn't change norm} && (3.81)\\
&= \left\| \begin{pmatrix} \sigma_1 y_1 \\ \vdots \\ \sigma_k y_k \\ 0 \\ \vdots \\ 0 \end{pmatrix} - \begin{pmatrix} \mathbf{u}_1'\mathbf{b} \\ \vdots \\ \mathbf{u}_p'\mathbf{b} \end{pmatrix} \right\|_2 && && (3.82)\\
&= \sum_{i=1}^{k} \|\sigma_i y_i - \mathbf{u}_i'\mathbf{b}\|_2^2 + \sum_{i=k+1}^{p} \|\mathbf{u}_i'\mathbf{b}\|_2^2 && && (3.83)
\end{aligned}$$

Therefore, we want

$$y_i = \frac{\mathbf{u}_i'\mathbf{b}}{\sigma_i}$$

Once we have $\mathbf{y}$, we can recover $\mathbf{x}$.

$$\begin{aligned}
\mathbf{x} &= \begin{pmatrix} \mathbf{v}_1 & \dots & \mathbf{v}_k \end{pmatrix} \begin{pmatrix} \frac{\mathbf{u}_1'\mathbf{b}}{\sigma_1} \\ \vdots \\ \frac{\mathbf{u}_k'\mathbf{b}}{\sigma_k} \end{pmatrix} \\
&= \mathbf{v}_1\frac{\mathbf{u}_1\mathbf{b}}{\sigma_1} + \ldots + \mathbf{v}_k\frac{\mathbf{u}_k\mathbf{b}}{\sigma_k}
\end{aligned}$$

3.12.3 Raleigh-Ritz Ratios

Say we have a matrix A. Where the eigenvalues of A are concerned, we have $A\mathbf{x}=\lambda\mathbf{x}$.

Similarly, $\mathbf{x}^{\mathsf{H}}A\mathbf{x}=\lambda\mathbf{x}^{\mathsf{H}}\mathbf{x}$.

$\mathbf{x}^{\mathsf{H}}\mathbf{x}=\|\mathbf{x}\|_2^2$ and

$$\lambda = \frac{\mathbf{x}^{\mathsf{H}}A\mathbf{x}}{\mathbf{x}^{\mathsf{H}}\mathbf{x}}$$

The ratio $\frac{\mathbf{x}^{\mathsf{H}}A\mathbf{x}}{\mathbf{x}^{\mathsf{H}}\mathbf{x}}$ is called the *Raleigh-Ritz Ratio*.

If $\mathbf{x}^{\mathsf{H}}\mathbf{x}=1$, then geometrically, $\mathbf{x}$ forms a unit sphere.

If A is Hermetian, then $\lambda=\mathbf{x}^{\mathsf{H}}A\mathbf{x}$ is a real number.

Say the eigenvalues of A are $\lambda_1 \le \lambda_2 \le \ldots \le \lambda_n$.

Theorem 3.12.2 (Raleigh-Ritz Theorem): If we have a Hermetian matrix, then

$$\lambda_1\mathbf{x}^{\mathsf{H}}\mathbf{x} \le \mathbf{x}^{\mathsf{H}}A\mathbf{x} \le \lambda_n\mathbf{x}^{\mathsf{H}}\mathbf{x}$$

This theorem bounds $\mathbf{x}^{\mathsf{H}}A\mathbf{x}$ with respect to A's eigenvalues. □

For a Hermetian matrix A, we can factor $A=UDU^{\mathsf{H}}$, where

- U is a unitary matrix ($UU^{\mathsf{H}} = U^{\mathsf{H}}U = I$), and
- D is a diagonal matrix, whose diagonal elements are the eigenvalues of A.

Therefore

$$A = UDU^{\mathsf{H}} \tag{3.84}$$

$$\mathbf{x}^{\mathsf{H}}A\mathbf{x} = \mathbf{x}^{\mathsf{H}}UDU^{\mathsf{H}}\mathbf{x} \tag{3.85}$$

$$= \mathbf{y}^{\mathsf{H}}D\mathbf{y} \qquad \text{let } \mathbf{y} = U^{\mathsf{H}}\mathbf{x} \tag{3.86}$$

$$= \begin{pmatrix}\overline{y_1} & \dots & \overline{y_n}\end{pmatrix}\begin{pmatrix}\lambda_1 & \dots & 0\\ \vdots & \ddots & \vdots\\ 0 & \dots & \lambda_n\end{pmatrix}\begin{pmatrix}y_1\\ \vdots\\ y_n\end{pmatrix} \tag{3.87}$$

$$= \lambda_1 y_1\overline{y_1} + \ldots + \lambda_n y_n\overline{y_n} \tag{3.88}$$

$$= \lambda_1|y_1|^2 + \ldots + \lambda_n|y_n|^2 \tag{3.89}$$

In line (3.89), each term is a real number.

Let $\mathbf{x} = U\mathbf{y}$. Since U is a unitary matrix,

$$\|\mathbf{y}\|_2^2 = |y_1|^2 + \ldots + |y_n|^2 \tag{3.90}$$

$$= \|\mathbf{x}\|_2^2 \tag{3.91}$$

and

$$\lambda_1\|\mathbf{y}\|_2^2 \le \|\mathbf{y}\|_2^2 \le \lambda_n\|\mathbf{y}\|_2^2 \tag{3.92}$$

$$\therefore \lambda_1\|\mathbf{x}\|_2^2 \le \mathbf{x}^{\mathsf{H}}A\mathbf{x} \le \lambda_n\|\mathbf{x}\|_2^2 \tag{3.93}$$

As a result of the Raleigh-Ritz theorem, we can say that the eigenvalues of a matrix A depend continuously on A. This property does *not* apply to the eigenvectors – it only applies to the eigenvalues.

Suppose we have two matrices A and B, such that $B - A = E$. In this context E represents the error between A and B.

These matrices will have the eigenvalues:

B: $\beta_1 \le \beta_2 \le \ldots \le \beta_n$
A: $\alpha_1 \le \alpha_2 \le \ldots \le \alpha_n$
E: $\epsilon_1 \le \epsilon_2 \le \ldots \le \epsilon_n$

The difference between A and B is bounded by the eigenvalues ϵ_i of E.

3.12.4 The Courant-Fisher Theorem

Let $A \in \mathbb{C}^{n\times n}$ is a Hermetian matrix, and let $W = \{\mathbf{w}_1, \ldots, \mathbf{w}_k\}$ be a set of vectors.

Theorem 3.12.3 (Courant-Fisher Theorem):

$$\lambda_{k+1} = \min_{\mathbf{w}}\max_{\mathbf{x}}\{\mathbf{x}^{\mathsf{H}}A\mathbf{x} \mid \|\mathbf{x}\| = 1 \text{ and } \mathbf{w}_i'\mathbf{x} = 0 \text{ for } 1 \le i \le k\}$$

or equivalently,

$$\lambda_{k+1} = \min_{\mathbf{w}}\max_{\mathbf{x}}\{\mathbf{x}^{\mathsf{H}}A\mathbf{x} \mid \|\mathbf{x}\| = 1 \text{ and } \mathbf{x} \in \langle\mathbf{w}_1, \ldots, \mathbf{w}_k\rangle^{\perp}\}$$

Geometrically, we start with a circle. Then, we find the largest $\mathbf{x}^{\mathsf{H}}A\mathbf{x}$ that cuts through the circle. Then we vary the circle to find the smallest $\mathbf{w}$.

3.12.5 Logistics

- **hw3**. For hw3 problem 3, we can assume that the matrix B is symmetric.
- **Presentations**. Our presentations should make some use of software – Scilab, Matlab, Octave, etc. In addition to presenting material from our papers, we should perform experiments based on the papers, and present the results of those experiments.

3.13 Lecture – 4/15/2009

3.13.1 Notes on hw3 prob 4

In this problem we have a matrix $A \in \mathbb{R}^{n\times 2}$.

First, note that $A'A$ is

$$\begin{aligned} A'A &= \begin{pmatrix} \mathbf{u}^{\mathsf{H}} \\ \mathbf{v}^{\mathsf{H}} \end{pmatrix} \begin{pmatrix} \mathbf{u} & \mathbf{v} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{u}^{\mathsf{H}}\mathbf{u} & \mathbf{v}^{\mathsf{H}}\mathbf{u} \\ \mathbf{u}^{\mathsf{H}}\mathbf{v} & \mathbf{v}^{\mathsf{H}}\mathbf{v} \end{pmatrix} \\ &= \begin{pmatrix} \|\mathbf{u}\|_2^2 & (\mathbf{u},\mathbf{v}) \\ (\mathbf{u},\mathbf{v}) & \|\mathbf{v}\|_2^2 \end{pmatrix} \end{aligned}$$

(Above, $(\mathbf{u},\mathbf{v})$ is the inner product of vectors $\mathbf{u}$ and $\mathbf{v}$).

The characteristic polynomial of this $A'A$ is

$$\begin{aligned} &\begin{vmatrix} \|\mathbf{u}\|_2^2 - \lambda & (\mathbf{u},\mathbf{v}) \\ (\mathbf{u},\mathbf{v}) & \|\mathbf{v}\|_2^2 - \lambda \end{vmatrix} = 0 \\ =&\lambda^2 - \lambda(\|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_{@}^2) + \|\mathbf{u}\|_2^2\|\mathbf{v}\|_2^2 - (\mathbf{u},\mathbf{v})^2 \end{aligned}$$

We can make several observations about this.

One observation: by the Cauchy-Schwarz inequality, the eigenvectors are both real numbers, or both imaginary numbers.

3.13.2 Courant-Fisher Theorem

Say we deal with vectors $\mathbf{w} \in \mathbb{R}^n$. Let W be a set of k vectors: $W = \{\mathbf{w}_1, \ldots, \mathbf{w}_j\}$.

Let A be a Hermetian matrix. We know that the eigenvalues of A are real numbers.

Finally, we know that $\mathbf{x}^{\mathsf{H}}A\mathbf{x}$ is a real number. Regardless of our choice for $\mathbf{x}$, $\mathbf{x}^{\mathsf{H}}A\mathbf{x}$ will lie between the smallest and largest eigenvalue of A. (This is a result of the Raleigh-Ritz Theorem.)

We would like $\mathbf{x}$ to be such that $\mathbf{x} \in \langle W\rangle^{\perp}$. $\mathbf{x}$ should lie on a hyperplane perpendicular to W.

Let us write the eigenvalues of A in descending order:

$$\begin{aligned} &\text{spec}(A) = \{\lambda_1, \ldots, \lambda_n\} \\ &\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \end{aligned}$$

The Courant-Fisher Theorem states that

$$\lambda_{k+1} = \min_{W} \max_{\mathbf{x}} \{\mathbf{x}^{\mathsf{H}}A\mathbf{x} \mid \|\mathbf{x}\|_2 = 1 \text{ and } \mathbf{x} \in \langle W\rangle^{\perp}\} \tag{3.94}$$

In (3.94), k is the number of vectors in W.

First, we'd like to show that

$$\lambda_{k+1} \leq \min_{W} \max_{\mathbf{x}} \{\mathbf{x}^{\mathsf{H}}A\mathbf{x} \mid \|\mathbf{x}\|_2 = 1 \text{ and } \mathbf{x} \in \langle W\rangle^{\perp}\} \tag{3.95}$$

We are seeking $\mathbf{x}$. We want $\|\mathbf{x}\|_2 = 1$, $\mathbf{x} \in \langle W\rangle^{\perp}$, and $\mathbf{x}^{\mathsf{H}}A\mathbf{x} \geq \lambda_{k+1}$.

If $\mathbf{x} \in \langle W \rangle^{\perp}$, then each $\mathbf{w}^{\mathsf{H}}\mathbf{x} = 0$. In other words,

$$\begin{aligned} \mathbf{w}_1^{\mathsf{H}}\mathbf{x} &= 0 \\ &\vdots \\ \mathbf{w}_k^{\mathsf{H}}\mathbf{x} &= 0 \end{aligned} \tag{3.96}$$

A is a Hermetian matrix. Therefore, we can diagonalize A as $A = U^{\mathsf{H}}TU$, where

- U is a unitary matrix. ($U^{\mathsf{H}}U = UU^{\mathsf{H}} = I$)
- T is a diagonal matrix. The diagonal elements of T are the eigenvalues of A.

Let $\mathbf{x} = U^{\mathsf{H}}\mathbf{y}$. We can re-write the linear system in (3.96) as

$$\begin{aligned} \mathbf{w}_1^{\mathsf{H}}U^{\mathsf{H}}\mathbf{y} &= 0 \\ &\vdots \\ \mathbf{w}_k^{\mathsf{H}}U^{\mathsf{H}}\mathbf{y} &= 0 \\ y_{k+1} &= 0 \\ &\vdots \\ y_n &= 0 \end{aligned} \tag{3.97}$$

The system in (3.97) has infinitely many solutions. Let us choose a solution where $\|\mathbf{y}\|_2 = 1$.

Note that $\|U^{\mathsf{H}}\mathbf{y}\|_2 = \|\mathbf{y}\|_2 = 1$. Since U is unitary, it doesn't change the norm. Therefore $\|\mathbf{x}\|_2 = 1$.

Let's plug some of these substitutions back into $\mathbf{x}^{\mathsf{H}}A\mathbf{x}$.

$$\begin{aligned} \mathbf{x}^{\mathsf{H}}A\mathbf{x} &= \mathbf{y}^{\mathsf{H}}UAU^{\mathsf{H}}\mathbf{y} && \text{since } \mathbf{x} = U^{\mathsf{H}}\mathbf{y} \\ &= \mathbf{y}^{\mathsf{H}}T\mathbf{y} && \text{since } A = U^{\mathsf{H}}TU \text{ and } UAU^{\mathsf{H}} = T \\ &= \mathbf{y}^{\mathsf{H}} \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_n \end{pmatrix} \mathbf{y} && \text{write out } T \\ &= \lambda_1 y_1^2 + \ldots + \lambda_n y_n^2 && \lambda_i \text{ in descending order} \\ &= \lambda_1 y_1^2 + \ldots + \lambda_k y_k^2 && \text{since } \lambda_{k+1} \ldots \lambda_n = 0 \\ &\geq \lambda_{k+1} && \text{since } \lambda_k \geq \lambda_{k+1} \end{aligned}$$

The minimum is achieved when W's vectors are linearly independent. In this case, we will achieve equality.

Let's see how to choose W so that we get equality.

We chose $\mathbf{w}_1, \ldots, \mathbf{w}_k$ as follows:

$$\begin{aligned} \mathbf{w}_1 &= U\mathbf{e}_1 && \mathbf{e}_i \text{ are standard basis vectors} \\ &\vdots \\ \mathbf{w}_k &= U\mathbf{e}_k \end{aligned}$$

$\mathbf{w}_1, \ldots, \mathbf{w}_k$ are the first k columns of U.

We want

$$\begin{aligned}\mathbf{w}_1^{\mathsf{H}}\mathbf{x} &= 0\\ &\vdots\\ \mathbf{w}_k^{\mathsf{H}}\mathbf{x} &= 0\end{aligned}$$

$$\begin{aligned}\mathbf{e}_1^{\mathsf{H}}(U^{\mathsf{H}}\mathbf{x}) &= 0\\ &\vdots\\ \mathbf{e}_k^{\mathsf{H}}(U^{\mathsf{H}}\mathbf{x}) &= 0\end{aligned}$$

And $y_1 = y_2 = \ldots = y_k = 0$.

Let $\mathbf{y} = U^{\mathsf{H}}\mathbf{x}$. The first k components of $\mathbf{y}$ are fixed, but we are free to choose the last $n-k$ components.

Again, we diagonalize A as $A = U^{\mathsf{H}}TU$, and

$$\begin{aligned}\mathbf{y} &= U^{\mathsf{H}}\mathbf{x}\\ \mathbf{x} &= U\mathbf{y}\\ \mathbf{x}^{\mathsf{H}} &= \mathbf{y}^{\mathsf{H}}U^{\mathsf{H}}\end{aligned}$$

We can manipulate $\mathbf{x}^{\mathsf{H}}A\mathbf{x}$.

$$\begin{aligned}\mathbf{x}^{\mathsf{H}}A\mathbf{x} &= \mathbf{y}^{\mathsf{H}}U^{\mathsf{H}}AU\mathbf{y} && \text{since } \mathbf{x} = U\mathbf{y}\\ &= \mathbf{y}^{\mathsf{H}}T\mathbf{y} && \text{since } U^{\mathsf{H}}AU = T\\ &= \sum_{i=1}^{n}\lambda_i y_i^2\\ &= \lambda_{k+1}y_{k+1}^2 + \ldots + \lambda_n y_n^2 && \text{since first } k \text{ terms are zero}\\ &= 1 && \text{since } \|y\|_2 = 1\end{aligned}$$

This is achieved if $y_{k+1} = 1$ and $k_{k+2} = \ldots = y_n = 0$.

3.13.3 Applications of The Courant-Fisher Theorem

Suppose we have two matrices, $A, B \in \mathbb{C}^{n\times m}$, and let $B - A = E$. E measures the error between A and B.

Suppose A, B, E have the following eigenvalues:

A: $\alpha_1 \geq \alpha_2 \geq \ldots \geq \alpha_n$
B: $\beta_1 \geq \beta_2 \geq \ldots \geq \beta_n$
E: $\epsilon_1 \geq \epsilon_2 \geq \ldots \geq \epsilon_n$

We can prove that $\beta_i - \alpha_i$ varies between ϵ_n and ϵ_1:

$$\epsilon_n \leq \beta_i - \alpha_i \leq \epsilon_i$$

Since $B - A = E$, we also have $B = A + E$. By the Courant-Fisher Theorem,

$$\beta_k = \min_W \max_{\mathbf{x}}\{\mathbf{x}^{\mathsf{H}}B\mathbf{x} \mid \|\mathbf{x}\|_2 = 1 \text{ and } \mathbf{x} \in \langle \mathbf{w}_1, \ldots, \mathbf{w}_{k-1}\rangle^{\perp}\} \tag{3.98}$$

$$= \min_W \max_{\mathbf{x}}\{\mathbf{x}^{\mathsf{H}}A\mathbf{x} + \mathbf{x}^{\mathsf{H}}E\mathbf{x} \mid \|\mathbf{x}\|_2 = 1 \text{ and } \mathbf{x} \in \langle \mathbf{w}_1, \ldots, \mathbf{w}_{k-1}\rangle^{\perp}\} \quad \leq \max_{\mathbf{x}}\{\mathbf{x}^{\mathsf{H}}A\mathbf{x} + \mathbf{x}^{\mathsf{H}}E\mathbf{x} \mid \|\mathbf{x}\|_2 = 1 \text{ and } \mathbf{x} \in \langle \tag{3.99}$$

In (3.98) and (3.99), $W = \{\mathbf{w}_1, \ldots, \mathbf{w}_{k-1}\}$.

As before, we have

$$\begin{aligned} A &= U^{\mathsf{H}} D U \\ &= U^{\mathsf{H}} \begin{pmatrix} \alpha_1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \alpha_n \end{pmatrix} U \end{aligned}$$

We choose $\mathbf{w}_i = U\mathbf{e}_i$. With this choice, we will have

$$\begin{aligned} \mathbf{e}_1^{\mathsf{H}} U^{\mathsf{H}} \mathbf{x} &= 0 \\ &\vdots \\ \mathbf{e}k - 1^{\mathsf{H}} U^{\mathsf{H}} \mathbf{x} &= 0 \end{aligned}$$

Let $\mathbf{y} = U^{\mathsf{H}}\mathbf{x}$. Substitution gives

$$\begin{aligned} \mathbf{e}_1^{\mathsf{H}} \mathbf{y} &= 0 \\ &\vdots \\ \mathbf{e}_{k-1}^{\mathsf{H}} \mathbf{y} &= 0 \end{aligned}$$

Therefore, $y_1 = \ldots = y_{k-1} = 0$.

U is a unitary matrix, so $\|\mathbf{x}\|_2 = \|\mathbf{y}\|_2 = 1$, and $\|\mathbf{y}\|_2$ comes from $y_k, \ldots, y_n$.

Since $\mathbf{y} = U^{\mathsf{H}}\mathbf{x}$, we also have $\mathbf{x} = U\mathbf{y}$ and $\mathbf{x}^{\mathsf{H}} = \mathbf{y}^{\mathsf{H}} U^{\mathsf{H}}$.

$$\begin{aligned} \mathbf{x}^{\mathsf{H}} A \mathbf{x} &= \mathbf{y}^{\mathsf{H}} U^{\mathsf{H}} A U \mathbf{y} \\ &= \mathbf{y}^{\mathsf{H}} \begin{pmatrix} \alpha_1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \alpha_n \end{pmatrix} \mathbf{y} \\ &= \sum_{i=k}^{n} \alpha_i y_i^2 \\ &\leq \alpha_k \end{aligned}$$

Therefore

$$\begin{aligned} \mathbf{x}^{\mathsf{H}} A \mathbf{x} &\leq \alpha_k && \text{and} \\ \beta_k &\leq \alpha_k + \epsilon_1 && \text{and} \\ \alpha_k + \epsilon_n &\leq \beta_k \end{aligned}$$

This is called the *stability property of eigenvalues*.

Given $B - A = E$, if E's eigenvalues are small, then B and A are pretty close.

3.13.4 Condition Numbers

Let A be a matrix, and assume that the inverse A^{-1} exists.

We can solve $A\mathbf{x} = \mathbf{b}$ by solving $\mathbf{x} = A^{-1}\mathbf{b}$.

From a the standpoint of numerical stability, $\mathbf{x} = A^{-1}\mathbf{b}$ tends to be a bad approach.

The condition number of a matrix is the ratio of the largest to the smallest eigenvalue

$$\frac{\sigma_1}{\sigma_n} \qquad \text{condition number}$$

If the condition number is large, then the matrix is numerically unstable.

3.13.5 High-dimensional Spaces

In this section, we'll see how our intuition of two-dimensional and three-dimensional spaces does not hold for larger dimensions.

Spheres

What is the volume of a sphere of radius r?

$$V_1 = 2r \qquad \text{in one dimension}$$
$$V_2 = \pi r^2 \qquad \text{in two dimensions}$$
$$V_3 = \frac{4}{3}\pi r^3 \qquad \text{in three dimensions}$$
$$V_n = \frac{\pi^{\frac{n}{2}} r^n}{\Gamma(\frac{n}{2}+1)} \qquad \text{in } n \text{ dimensions}$$

Γ denotes the Euler Gamma Function:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$$
$$\Gamma(n+1) = n!$$
$$\Gamma(\tfrac{1}{2}) = \sqrt{\pi}$$
$$\Gamma(x+1) = x\Gamma(x)$$
$$\Gamma(\tfrac{3}{2}) = \frac{1}{2}\sqrt{\pi}$$

For example, in two dimensions, we have

$$V_2 = \frac{\pi r^2}{\Gamma(2)} = \frac{\pi r^2}{1} = \pi r^2$$

Suppose we have two concentric spheres in $\mathbb{R}^3$, as shown in Figure 3.2. In Figure 3.2, the outer sphere

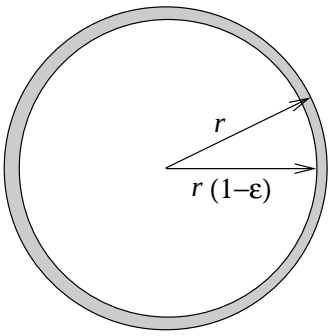

Figure 3.2: Two concentric spheres in $\mathbb{R}^3$

has radius r, the inner sphere has radius $r(1-\epsilon)$ and the width of the shaded area is ϵ.

Let V be the volume of the outer sphere, and V' be the volume of the inner sphere. The ratio between

the volumes is

$$\frac{V - V'}{V} = \frac{\frac{\pi^{\frac{n}{2}} r^n}{\Gamma(\frac{n}{2}+1)} - \frac{\pi^{\frac{n}{2}}(1-\epsilon)^n}{\Gamma(\frac{n}{2}+1)}}{\frac{\pi^{\frac{n}{2}} r^n}{\Gamma(\frac{n}{2}+1)}}$$
$$= 1 - (1-\epsilon)^n$$

As $n \to \infty$, $1 - (1-\epsilon)^n$ tends to 1.

Therefore, as the number of dimensions grows, most of the area of the sphere becomes concentrated at its outer edge.

Rectangles

Suppose we have a unit cube, like the one shown in Figure 3.3.

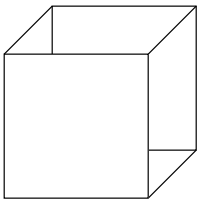

Figure 3.3: Unit Cube in $\mathbb{R}^3$

What is the length of d, the longest diagonal in the cube?

$d_2 = \sqrt{2}$	in two dimensions
$d_3 = \sqrt{3}$	in three dimensions
$d_n = \sqrt{n}$	in n dimensions

As the number of dimensions grows, the distance from the center to a face of the cube remains 0.5. But, the distance between two opposite vertices grows. In high-dimensional spaces, the cube ends up looking something like Figure 3.4.

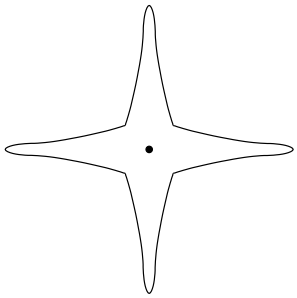

Figure 3.4: Crude Drawing of a Cube in High-Dimensional Space

As another example, consider the two squares shown in Figure 3.5. Suppose we were to grab the upper right corner of the shaded square, and pull it along the diagonal, in the direction of the arrow. How far would we have to move the corner in order to double the column of the shaded square?

As n becomes large, a minuscule change along the diagonal will produce a very big change in volume.

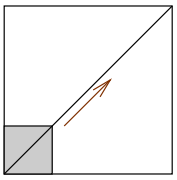

Figure 3.5: Two rectangles

3.14 PCA Tutorials

Some PCA tutorials:

- `http://www.snl.salk.edu/~shlens/pub/notes/pca.pdf`
- `http://www.iiap.res.in/astrostat/`. Click "R Tutorials", then "Principal Component Analysis"
- Chapter 12 of Maindonald and Braun, *Data Analysis and Graphics Using R*, 2nd edition, Cambridge University Press, 2007 also discusses PCA. The discussion is informal, but there's R code that one can try out.

3.15 PCA Notes – 5/14/2009

These notes come from one of the aforementioned PCA tutorials, `http://www.snl.salk.edu/~shlens/pub/notes/pca.pdf`. The treatment is less rigorous than our class discussions, but it helps to glue the concepts together.

3.15.1 A Motivation for PCA

As before, we start with a sample matrix X. Each row of X represents a variable, and each column of X represents a measurement (sample).

If we always knew what we were doing, we'd never measure redundant information, and all of our basis of measurement would be orthogonal. In the real world, things seldom work out so well.

The main goal of PCA is to identify a small number of basis dimensions that best describe our data. Knowing this, we can determine which parts of our data are truly useful, and which parts are just noise.

Our original data was gathered according to some basis. With PCA we ask the question, "Is there another basis, which is a *linear transformation* of the original basis, that best expresses the data"?

For the sake of illustration, consider $Y = TX$. Here

- X is our original data
- T is a linear transformation that is applied to X. (A change of basis).
- Y is a new representation of our data (having undergone a change of basis).

Geometrically, T is a matrix that rotates and stretches the data, thereby turning X into Y. To put it another way, Y is a projection of X onto the basis T.

3.15.2 Variance and Covariance

A covariance matrix can help us identify redundancies in our data set. Consider two vectors, $\mathbf{a}$ and $\mathbf{b}$:

$$\mathbf{a} = (a_1, \ldots, a_n)$$
$$\mathbf{b} = (b_1, \ldots, b_n)$$

The individual variances of $\mathbf{a}$, $\mathbf{b}$ are

$$\sigma^2_{\mathbf{a}} = \frac{1}{n}\sum_{i=1}^{n} a_i^2$$
$$\sigma^2_{\mathbf{b}} = \frac{1}{n}\sum_{i=1}^{n} b_i^2$$

The covariance between $\mathbf{a}$ and $\mathbf{b}$ is

$$\sigma^2_{\mathbf{ab}} = \frac{1}{n}\sum_{i=1}^{n} a_i b_i$$

Covariance measure the degree of linear relationship between two variables.

- A large positive covariance denotes a strong positive correlation.
- A large negative covariance denotes a strong negative correlation.
- A covariance of zero denotes completely uncorrelated data.

Note that

$\sigma^2_{\mathbf{ab}} = 0$ IFF $\mathbf{a}$, $\mathbf{b}$ are completely uncorrelated

$\sigma^2_{\mathbf{ab}} = \sigma^2_{\mathbf{a}}$ IFF $\mathbf{a} = \mathbf{b}$

As a dot product computation, the covariance of $\mathbf{a}$, $\mathbf{b}$ is

$$\sigma^2_{\mathbf{ab}} = \frac{1}{n}\mathbf{a}'\mathbf{b}$$

A *covariance matrix* is $\text{cov}(X) = \frac{1}{n}XX'$.

- The diagonal terms of $\text{cov}(X)$ give the variance of individual variables.
- Large elements on the diagonal correspond to "interesting" structures in the data.
- The off-diagonal terms of $\text{cov}(X)$ give the covariance between pairs of variables.
- Large off-diagonal elements correspond to redundant variables.

3.15.3 The Goal of PCA

PCA has two goals:

- To minimize redundancy (large covariance)
- To maximize the signal to noise ratio in the data (large variance)

We can factor $\text{cov}(X) = PDP'$, or equivalently $D = P'\,\text{cov}(X)P$. Here

- D is a diagonal matrix that contains the eigenvalues of $\text{cov}(X)$.
- P is an orthonormal matrix whose columns are the eigenvectors of $\text{cov}(X)$.

For PCA, the columns of P are the principal components. P provides our change of basis.

Going back to our earlier discussion:

$$Z = P'X \tag{3.100}$$

Z contains the "new variables" from the change of basis to X. Z is the projection of X onto the new basis P.

3.15.4 Assumptions That PCA Makes

PCA makes the following assumptions:

- PCA assumes *linearity*. This allows us to change basis through a linear transformation.
- PCA assumes that large variances indicate important structures in the data.
- PCA assumes that the principal components (columns op P) are orthogonal.

There are a few notable cases where PCA fails to work well:

- PCA can fail if the data has non-linear relationships. In some cases, one can work around this by applying a transformation to the data, so that it is linear.

An alternative to PCA is *Independent Component Analysis* or ICA. ICA does not assume linearity in the data. ICA reduces the data to a set of dimensions that are statistically independent.

Part 4

Class Projects

Our last few classes consisted of student project presentations. What follows is a collection of random notes – things that I thought were interesting at the time.

4.1 Tyler's Presentation – 4/22/2009

The basis of PCA

- PCA was first used to discover sources of variance.
- PCA played an important role in the behavioral sciences, in particular the development of IQ.

Intelligence is a difficult thing to define, but people generally agree that there are several "aspects" to intelligence. This is reflected in standardized tests: most standardized tests have a verbal component and a quantitative reasoning component. People in the behavioral sciences were interested in how "aspects" of intelligence contributed to overall "general intelligence".

Factor Analysis is a cousin to PCA.

PCA and factor analysis are similar tests, but they were developed with different motivations.

PCA was developed by mathematicians. Factor analysis was developed by psychologists.

PCA assumes zero error. Factor analysis assumes an explicit error.

Tyler used *scree plots* in several of his examples. I hadn't heard this term before.

Scree plots are used in conjuction with Principal component analysis. Given n eigenvalues, you plot $1 \ldots n$ on the x axis, and the corresponding λ_i on the y axis. The slope of the plot helps you to determine which eigenvalues to retain.

One explanation: `http://janda.org/workshop/factor%20analysis/SPSS%20run/SPSS08.htm`

4.2 Nicks's Presentation – 4/27/2009

Nick is doing research in Chinese character recognition.

I didn't realize that this was such a challenging problem. There are two forms of Chinese characters: traditional and simplified. In generate for each simplified character, there are n traditional characters that mean the same thing (or about the same thing).

Stroke count is one way to measure the complexity of a character. However, *stroke* refers to the way a character is drawn – not individual lines in the character. To give an example in english, "L" consists of two lines, but only one stroke. "T" consists of two strokes.

In his work, Nick used something called the "Box Counting Dimension".

Suppose you draw a free shape on a piece of graph paper, and count the number of boxes that the line passes through. The box counting dimension comes from analyzing how the number of boxes changes as the grid is made coarser or finer.

4.3 Kahn's Presentation – 4/29/2009

Kahn talked about the role of eigenvalues in Google's page rank computation.

The rank of page p is influenced by the rank of pages q_i that link to p. Likewise, the ranks of each q_i are influenced by the ranks of r_j that link to q_i. As a result, Pagerank is an iterative process – you repeat the process until the ranks stabilize.

This kind of phenemenon can also be observed in stock markets. If company A owns stock in company B, then the value of B has an influence on the value of A.

The stock market allows circular relationships. A can own stock in B, B can own stock in C, and C can own stock in A. Cycles of lenght three are okay. Cycles of length two are bad (think Enron.)

There is a branch of mathematics that studies adjacency matrices of graphs. This is called *spectral theory of graphs.*

Chakrabari has a good book on web mining.

4.4 Fransesco's Presentation – 5/4/2009

Fransesco's presentation dealt with non-negative matrix factorization.

The idea is to take a matrix V, and factor it as

$$\underset{(n \times m)}{V} = \underset{(n \times r)}{W} \cdot \underset{(r \times m)}{H}$$

Where W and H are non-negative matrices.

For example, if V was $(n \times 4)$, W was $(n \times 2)$ and H was (2×4), then we'd be taking four characteristics from V $(m = 4)$, and representing them as two composite characteristics $(r = 2)$.

Non-negative numbers are more intuitive in certain applications. The factorization is additive, and it tends to produce a better representation of natural phenemenon.

For example, suppose our matrix was actually a grayscale image, and we wanted to represent components of that image. Non-negative matrices make more sense for this (i.e., what is the significance of a negative grayscale value?).

Non-negative factorization gives you "parts" that can be summed to a whole, and each part tends to have its own real-world significance.

GNU Free Documentation License

Version 1.3, 3 November 2008

<http://fsf.org/>

Preamble

The purpose of this License is to make a manual, textbook, or other functional and useful document "free" in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or noncommercially. Secondarily, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of "copyleft", which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

1. APPLICABILITY AND DEFINITIONS

This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The "**Document**", below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as "**you**". You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A "**Modified Version**" of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A "**Secondary Section**" is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document's overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The "**Invariant Sections**" are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The "**Cover Texts**" are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A "**Transparent**" copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not "Transparent" is called "**Opaque**".

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The "**Title Page**" means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, "Title Page" means the text near the most prominent appearance of the work's title, preceding the beginning of the body of the text.

The "**publisher**" means any person or entity that distributes copies of the Document to the public.

A section "**Entitled XYZ**" means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as "**Acknowledgements**", "**Dedications**", "**Endorsements**", or "**History**".) To "**Preserve the Title**" of such a section when you modify the Document means that it remains a section "Entitled XYZ" according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

2. VERBATIM COPYING

You may copy and distribute the Document in any medium, either commercially or noncommercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

3. COPYING IN QUANTITY

If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

4. MODIFICATIONS

You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

A. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.

B. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.

C. State on the Title page the name of the publisher of the Modified Version, as the publisher.

D. Preserve all the copyright notices of the Document.

E. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.

F. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.

G. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.

H. Include an unaltered copy of this License.

I. Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the Document, create one stating the title, year, authors,

and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.

J. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.

K. For any section Entitled "Acknowledgements" or "Dedications", Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.

L. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.

M. Delete any section Entitled "Endorsements". Such a section may not be included in the Modified Version.

N. Do not retitle any existing section to be Entitled "Endorsements" or to conflict in title with any Invariant Section.

O. Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version's license notice. These titles must be distinct from any other section titles.

You may add a section Entitled "Endorsements", provided it contains nothing but endorsements of your Modified Version by various parties—for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

5. COMBINING DOCUMENTS

You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled "History" in the various original documents, forming one section Entitled "History"; likewise combine any sections Entitled "Acknowledgements", and any sections Entitled "Dedications". You must delete all sections Entitled "Endorsements".

6. COLLECTIONS OF DOCUMENTS

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

7. AGGREGATION WITH INDEPENDENT WORKS

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an "aggregate" if the copyright resulting from the compilation is not used to limit the legal rights of the compilation's users beyond what the individual works permit. When the Document is included in an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document's Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

8. TRANSLATION

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled "Acknowledgements", "Dedications", or "History", the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

9. TERMINATION

You may not copy, modify, sublicense, or distribute the Document except as expressly provided under this License. Any attempt otherwise to copy, modify, sublicense, or distribute it is void, and will automatically terminate your rights under this License.

However, if you cease all violation of this License, then your license from a particular copyright holder is reinstated (a) provisionally, unless and until the copyright holder explicitly and finally terminates your license, and (b) permanently, if the copyright holder fails to notify you of the violation by some reasonable means prior to 60 days after the cessation.

Moreover, your license from a particular copyright holder is reinstated permanently if the copyright holder notifies you of the violation by some reasonable means, this is the first time you have received notice of violation of this License (for any work) from that copyright holder, and you cure the violation prior to 30 days after your receipt of the notice.

Termination of your rights under this section does not terminate the licenses of parties who have received copies or rights from you under this License. If your rights have been terminated and not permanently reinstated, receipt of a copy of some or all of the same material does not give you any rights to use it.

10. FUTURE REVISIONS OF THIS LICENSE

The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See http://www.gnu.org/copyleft/.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License "or any later version" applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation. If the Document specifies that a proxy can decide which future versions of this License can be used, that proxy's public statement of acceptance of a version permanently authorizes you to choose that version for the Document.

11. RELICENSING

"Massive Multiauthor Collaboration Site" (or "MMC Site") means any World Wide Web server that publishes copyrightable works and also provides prominent facilities for anybody to edit those works. A public wiki that anybody can edit is an example of such a server. A "Massive Multiauthor Collaboration" (or "MMC") contained in the site means any set of copyrightable works thus published on the MMC site.

"CC-BY-SA" means the Creative Commons Attribution-Share Alike 3.0 license published by Creative Commons Corporation, a not-for-profit corporation with a principal place of business in San Francisco, California, as well as future copyleft versions of that license published by that same organization.

"Incorporate" means to publish or republish a Document, in whole or in part, as part of another Document.

An MMC is "eligible for relicensing" if it is licensed under this License, and if all works that were first published under this License somewhere other than this MMC, and subsequently incorporated in whole or in part into the MMC, (1) had no cover texts or invariant sections, and (2) were thus incorporated prior to November 1, 2008.

The operator of an MMC Site may republish an MMC contained in the site under CC-BY-SA on the same site at any time before August 1, 2009, provided the MMC is eligible for relicensing.