THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

DEPARTMENT OF STAT. AND OR

REFRESHER COURSE, SUMMER 2019

Linear Algebra

Original Author:
Oleg MAYBA,
UC Berkeley
Department of Statistics,
2006,
DMS Grant No 0130526

Modified By:
Eric Lock (UNC, 2010 & 2011)
Gen Li (UNC, 2012)
Michael Lamm (UNC, 2013)
Wen Jenny Shi (UNC, 2014)
Meilei Jiang (UNC, 2015 & 2016)
Iain Carmichael (UNC, 2017)
Adam Waterbury (UNC, 2018)
Brendan Brown (UNC, 2019)

Instructor:
Brendan Brown
(UNC at Chapel Hill)

August 9, 2019

Contents

1 Introduction **3**

2 Vector Spaces **4**
2.1 Basic Concepts . . . 5
2.2 Special Spaces . . . 8
2.3 Properties and geometry of inner product spaces . . . 12
2.4 Gram-Schmidt Process . . . 15
2.5 Comments on infinite-dimensional spaces . . . 16
Exercises . . . 17

3 Matrices and Matrix Algebra **20**
3.1 Matrix Operations . . . 20
3.2 Special Matrices . . . 22
3.3 The Four Fundamental Spaces . . . 23
3.4 Sample Covariance matrix . . . 25
Exercises . . . 26

4 Projections and Least Squares Estimation **27**
4.1 Projections . . . 27
4.2 Applications to Statistics: Least Squares Estimator . . . 30
Exercises . . . 32

5 Linear functionals, riesz representation and hyperplane separation **34**
5.1 Linear functionals . . . 34
5.2 Hyperplane separation . . . 36

6 Matrix Decompositions **39**
6.1 Determinants . . . 39
6.2 Eigenvalues and Eigenvectors . . . 41
6.3 Complex Matrices . . . 45
6.4 Facts that lead up to the spectral theorem . . . 45
6.5 Spectral Theorem . . . 47
6.6 Examples: Spectral theory . . . 50
Exercises . . . 52

7 Tensors **55**
7.1 Exercises . . . 57

8 Singular Value Decomposition **57**
8.1 Definition . . . 57
8.2 Low rank approximation . . . 60
8.3 A few Applications . . . 60
8.4 Principal Components Analysis . . . 62
8.5 The $25,000,000,000 Eigenvector: the Linear Algebra Behind Google . . . 65

Exercises . . . 65

9 Matrix functions and differentiation **67**
9.1 Basics . . . 67
9.2 Jacobian and Chain Rule . . . 67
9.3 Matrix functions . . . 69
Exercises . . . 70

10 Computation **71**
10.1 Power method . . . 71
10.2 Gradient Descent . . . 73

11 Statistics: Random Variables **74**
11.1 Expectation, Variance and Covariance . . . 74
11.2 Distribution of Functions of Random Variables . . . 76
11.3 Derivation of Common Univariate Distributions . . . 79
11.4 Random Vectors: Expectation and Variance . . . 82
Exercises . . . 84

12 Further Applications to Statistics: Normal Theory and F-test **85**
12.1 Bivariate Normal Distribution . . . 85
12.2 F-test . . . 86
Exercises . . . 88

13 References **89**

1 Introduction

These notes are intended for use in the warm-up camp for incoming UNC STOR graduate students. Welcome to Carolina!

We assume that you have taken a linear algebra course before and that most of the material in these notes will be a review of what you've already known. If some of the material is unfamiliar, do not be intimidated! We hope you find these notes helpful. If not, you can consult the references listed at the end, or any other textbooks of your choice for more information or another style of presentation (most of the proofs on linear algebra part have been adopted from Strang, the proof of F-test from Montgomery et al, and the proof of bivariate normal density from Bickel and Doksum).

Linear algebra is an important and fundamental math tool for probability, statistics, numerical analysis and operations research. Lots of material in this notes will show up in your future study and research. There will be 8 algebraic classes in total. In the second week we may cover some topics in computational linear algebra with applications to statistics and machine learning (PCA, PageRank, Spectra Clustering, etc). See the notes and course schedule on the UNC STOR bootcamp webpage.

2 Vector Spaces

A set V is a **vector space** over $\mathbb{R}$ (or any field) with two operations defined on it:

1. Vector addition, that assigns to each pair of vectors $v_1, v_2 \in V$ another vector $w \in V$ (we write $v_1 + v_2 = w$)
2. Scalar multiplication, that assigns to each vector $v \in V$ and each scalar $r \in \mathbb{R}$ (field) another vector $w \in V$ (we write $rv = w$)

The elements of V are called **vectors** and must satisfy the following 8 conditions $\forall\ v_1, v_2, v_3 \in V$ and $\forall\ r_1, r_2 \in \mathbb{R}$:

1. Commutativity of vector addition: $v_1 + v_2 = v_2 + v_1$
2. Associativity of vector addition: $(v_1 + v_2) + v_3 = v_1 + (v_2 + v_3)$
3. Identity element of vector addition: $\exists$ vector $0 \in V$, s.t. $v + 0 = v$, $\forall\ v \in V$
4. Inverse elements of vector addition: $\forall\ v \in V\ \exists\ -v = w \in V$ s.t. $v + w = 0$
5. Compatibility of scalar multiplication with (field) multiplication: $r_1(r_2 v) = (r_1 r_2)v$, $\forall\ v \in V$
6. Distributivity of scalar multiplication with respect to (field) addition: $(r_1 + r_2)v = r_1 v + r_2 v$, $\forall\ v \in V$
7. Distributivity of scalar multiplication with respect to vector addition: $r(v_1 + v_2) = rv_1 + rv_2$, $\forall\ r \in \mathbb{R}$
8. Identity element of scalar multiplication: $1v = v$, $\forall\ v \in V$

Vector spaces over fields other than $\mathbb{R}$ are defined similarly, with the multiplicative identity of the field replacing 1. We won't concern ourselves with those spaces, except for when we need complex numbers later on. Also, we'll be using the symbol 0 to designate both the number 0 and the vector 0 in V, and you should always be able to tell the difference from the context. Sometimes, we'll emphasize that we're dealing with, say, $n \times 1$ vector 0 by writing $0_{n\times 1}$.

Here are some examples of vector spaces

1. $\mathbb{R}^n$ with usual operations of element-wise addition and scalar multiplication. An example of these operations in $\mathbb{R}^2$ is illustrated in Figure 2.
2. Vector space $F_{[-1,1]}$ of all functions defined on interval $[-1, 1]$, where we define $(f+g)(x) = f(x) + g(x)$ and $(rf)(x) = rf(x)$.

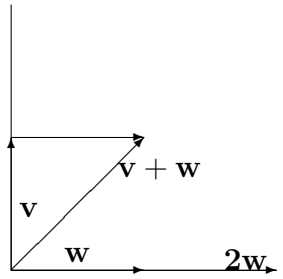

Figure 1: Vector Addition and Scalar Multiplication

2.1 Basic Concepts

Subspace and span We say that $S \subset V$ is a **subspace** of V, if S is closed under vector addition and scalar multiplication, i.e.

1. $\forall s_1, s_2 \in S$, $s_1 + s_2 \in S$
2. $\forall s \in S$, $\forall r \in \mathbb{R}$, $rs \in S$

You can verify that if those conditions hold, S is a vector space in its own right (satisfies the 8 conditions above). Note also that S has to be non-empty; the empty set is not allowed as a subspace.

Examples:

1. A subset $\{0\}$ is always a subspace of a vectors space V.
2. Given a set of vectors $S \subset V$, $\mathbf{span}(S) = \{w : w = \sum_{i=1}^{n} r_i v_i, r_i \in \mathbb{R}, \text{ and } v_i \in S\}$, the set of all linear combinations of elements of S (see below for definition) is a subspace of V.
3. $S = \{(x, y) \in \mathbb{R}^2 : y = 0\}$ is a subspace of $\mathbb{R}^2$ (x-axis).
4. The set of all continuous functions defined on interval $[-1, 1]$ is a subspace of $F_{[-1,1]}$.

For all of the above examples, you should check for yourself that they are in fact subspaces.

Given vectors $v_1, v_2, \ldots, v_n \in V$, we say that $w \in V$ is a **linear combination** of $v_1, v_2, \ldots, v_n$ if for some $r_1, r_2, \ldots, r_n \in \mathbb{R}$, we have $w = \sum_{i=1}^{n} r_i v_i$. If every vector in V is a linear combination of $S = \{v_1, v_2, \ldots, v_n\}$, we have $\text{span}(S) = V$, then we say S **spans** V.

Fact: A set $S \subseteq V$ is a subspace if and only if it is closed under linear combinations.

Linear independence and dependence Given vectors $v_1, v_2, \ldots, v_n \in V$ we say that $v_1, v_2, \ldots, v_n$ are **linearly independent** if $\sum_{i=1}^{n} r_i v_i = 0 \implies r_1 = r_2 = \ldots = r_n = 0$, i.e. the only linear combination of $v_1, v_2, \ldots, v_n$ that produces 0 vector is the trivial one. We say

that $v_1, v_2, \ldots, v_n$ are **linearly dependent** otherwise.

We now prove two results that will be used later and give some practice with linear algebra proofs.

Theorem: Let $I, S \subset V$ be such that I is linearly independent, and S spans V. Then for every $x \in I$ there exists a $y \in S$ such that $\{y\} \cup I \backslash \{x\}$ is linearly independent.

Proof: We prove this result by contradiction. First two two facts that can be easily verified from the definitions above.

1. If a set $J \subset V$ is linearly independent, then $J \cup \{y\}$ is linearly dependent if and only if $y \in \text{span}(J)$.

2. If $S, T \subset V$ with $T \subset \text{span}(S)$ then $\text{span}(T) \subset \text{span}(S)$.

Assume for the sake of contradiction that the claim does not hold. I.e. suppose there there exists a $x \in I$ such that for all $y \in S$ $\{y\} \cup I \backslash \{x\}$ is linearly dependent. Let $I' = I \backslash \{x\}$. Since I is linearly independent it follows that I' is also linearly independent. Then by the first fact above, the fact that $\{y\} \cup I'$ is linearly dependent implies $y \in \text{span}(I')$. Moreover, this holds for all $y \in S$ so $S \subset \text{span}(I')$.

By the second fact we then have that $\text{span}(S) \subset \text{span}(I')$. Now since S spans V it follows that $x \in V = \text{span}(S) \subset \text{span}(I') = \text{span}(I \backslash \{x\})$. This means there exists $v_1, v_2, \ldots, v_n \in I \backslash \{x\}$ and $r_1, r_2, \ldots, r_n \in \mathbb{R}$ such that $0 = x - \sum_{i=1}^{n} r_i v_i$, contradicting the assumption that I linearly independent. $\square$

Corollary: Let $I, S \subset V$ be such that I is linearly independent, and S spans V. Then $|I| \leq |S|$, where $|\cdot|$ denotes the number of elements of a set (possibly infinite).

Proof: If $|S| = \infty$ then the claim holds. Additionally, if $I \subset S$ the claim holds. So assume $|S| = m < \infty$, and $I \not\subset S$.

Now consider the following algorithm. Select $x \in I, x \notin S$. By the theorem above, choose a $y \in S$ such that $I' = \{y\} \cup I \backslash \{x\}$ is linearly independent. Note that $|I'| = |I|$ and that $|I' \cap S| > |I \cap S|$. If $I' \subset S$ then the claim holds and stop the algorithm, otherwise continue the algorithm with $I = I'$.

Now note that the above algorithm must terminate in at most $m < \infty$ steps. To see this, first note that after the m^{th} iteration $S \subset I'$. Next, if the algorithm does not terminate at this iteration $I' \not\subset S$, and there would exist a $x \in I', x \notin S$. But then since S spans V there would exist $v_1, v_2 \ldots, v_n \in S \subset I'$ and $r_1, r_2, \ldots, r_n \in \mathbb{R}$ such that $0 = x - \sum_{i=1}^{n} r_i v_i$ contradicting I' linearly independent. $\square$

In the following definitions, the reader might want to look ahead to the section on normed

and metric spaces. We follow Lax, Functional Analysis, for most infinite-dimensional concepts.

Basis and dimension Suppose $\{v_i\}_{i\in I}$, where I is an arbitrary index set, are vectors in a normed space V that are linearly independent and such that the closure (in the norm topology) of $span\{v_i\}_{i\in I}$ is V. Then we say that the set $\{v_i\}_{i\in I}$ is a **basis** for V whose **dimension** is the cardinality of I.

The span of a finite number of vectors is closed, so if $|I| < \infty$, for example when $V = \mathbb{R}^n$, then we have $V = span\{v_i\}_{i\in I}$.

The following theorem is proved using Zorn's lemma.

Theorem: Every vector space V has a basis.

Theorem: Let S be a basis for V, and let T be another basis for V. Then $|S| = |T|$.

Proof: This follows directly from the above Corollary since S and T are both linearly independent, and both span V. $\square$

We focus on finite-dimensional spaces. The most common infinite-dimensional spaces have countable bases, rather than uncountable ones.

Examples:

1. $S = \{0\}$ has dimension 0.
2. Any set of vectors that includes 0 vector is linearly dependent (why?)
3. If V has dimension n, and we're given $k < n$ linearly independent vectors in V, then we can extend this set of vectors to a basis.
4. Let $v_1, v_2, \ldots, v_n$ be a basis for V. Then if $v \in V$, $v = r_1v_1 + r_2v_2 + \ldots + r_nv_n$ for some $r_1, r_2, \ldots, r_n \in \mathbb{R}$. Moreover, these coefficients are unique, because if they weren't, we could also write $v = s_1v_1 + s_2v_2 + \ldots + s_nv_n$, and subtracting both sides we get $0 = v - v = (r_1 - s_1)v_1 + (r_2 - s_2)v_2 + \ldots + (r_n - s_n)v_n$, and since the v_i's form basis and are therefore linearly independent, we have $r_i = s_i \ \forall i$, and the coefficients are indeed unique.
5. $v_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $v_2 = \begin{bmatrix} -5 \\ 0 \end{bmatrix}$ both span x-axis, which is the subspace of $\mathbb{R}^2$. Moreover, any one of these two vectors also spans x-axis by itself (thus a basis is not unique, though dimension is), and they are not linearly independent since $5v_1 + 1v_2 = 0$

6. $e_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$, $e_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$, and $e_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$ form the standard basis for $\mathbb{R}^3$, since every vector $\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ in $\mathbb{R}^3$ can be written as $x_1e_1 + x_2e_2 + x_3e_3$, so the three vectors span $\mathbb{R}^3$ and their linear independence is easy to show. In general, $\mathbb{R}^n$ has dimension n.

7. Let $\dim(V) = n$, and let $v_1, v_2, \ldots, v_m \in V$, s.t. $m > n$. Then $v_1, v_2, \ldots, v_m$ are linearly dependent.

8. Suppose that we are interested in defining a basis for $\mathcal{C}([0,1])$, the space of all continuous real-valued functions defined on the interval $[0,1]$. What difficulties might arise?

2.2 Special Spaces

Inner product space (Real) A real inner product space is a vector space over $\mathbb{R}$ equipped with a function $f : V \times V \to \mathbb{R}$ (which we denote by $f(v_1, v_2) = \langle v_1, v_2 \rangle$), s.t. $\forall\ v, w, z \in V$, and $\forall r \in \mathbb{R}$:

1. $\langle v, w + rz \rangle = \langle v, w \rangle + r\langle v, z \rangle$ (linearity)
2. $\langle v, w \rangle = \langle w, v \rangle$ (symmetry)
3. $\langle v, v \rangle \geq 0$ and $\langle v, v \rangle = 0$ if and only if $v = 0$ (positive-definiteness)

Typically, we will consider only real vector spaces. However, complex inner product spaces are ubiquitous, so we give a definition.

Inner product space (Complex) A complex inner product space is a vector space over $\mathbb{C}$, the complex numbers, equipped with a function $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{C}$, with the following properties

1. $\langle v, w + \alpha z \rangle = \langle v, w \rangle + \bar{\alpha}\langle v, z \rangle$ (anti-linearity)
2. $\langle w + \alpha z, v \rangle = \langle w, v \rangle + \alpha\langle z, v \rangle$ (linearity)
3. $\langle v, w \rangle = \overline{\langle w, v \rangle}$ (skew symmetry)
4. $\langle v, v \rangle \geq 0$ and $\langle v, v \rangle = 0$ if and only if $v = 0$ (positive-definiteness)

where $\bar{\alpha}$ denotes the conjugate of a complex number α. In some texts, the convention is to make anti-linearity appear in the first argument of the inner product rather than the second.

Examples:

1. In $\mathbb{R}^n$ the standard inner product between two vectors is $x^T y$. Given 2 vectors $x, y \in \mathbb{R}^n$ where $x = [x_1, x_2, \cdots, x_n]^T$ and $y = [y_1, y_2, \cdots, y_n]^T$, we define their inner product

$$\langle x, y \rangle = x^T y := \sum_{i=1}^{n} x_i y_i$$

   You can check yourself that the 3 properties above are satisfied, and the meaning of notation $x^T y$ will become clear from the next section.

2. Given $f, g \in \mathcal{C}([-1, 1])$, we define $\langle f, g \rangle = \int_{-1}^{1} f(x)g(x)dx$. Once again, verification that this is indeed an inner product is left as an exercise.

Normed space The **norm**, or **length**, of a vector v in the vector space V is a function $g : V \to \mathbb{R}$ (which we denote by $g(v) = \|v\|$), s.t. $\forall\ v, w \in V$, and $\forall r \in \mathbb{R}$:

1. $\|rv\| = |r|\|v\|$
2. $\|v\| \geq 0$, with equality if and only if $v = 0$
3. $\|v + w\| \leq \|v\| + \|w\|$ (triangle inequality)

Examples:

1. In $\mathbb{R}^n$, let's define the **length of a vector** $x := \|x\| = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2} = \sqrt{x^T x}$, or $\|x\|^2 = x^T x$. This is called the **Euclidian norm**, or the L_2 norm (denote by $\|x\|_2$). (verify it by yourself)

2. Again in $\mathbb{R}^n$, if we define $\|x\| = |x_1| + \ldots + |x_n|$, it's also a norm called the L_1 norm (denote by $\|x\|_1$). (verify it by yourself)

3. Given $p \geq 1$ and $f \in \mathcal{C}([-1, 1])$, we define $\|f\|_p = \left(\int_{-1}^{1} |f(x)|^p dx\right)^{\frac{1}{p}}$, which is also a norm. It is clear that $||\cdot||_p$ satifies the first two properties of norms, so we will show the triangle inequality (known as Minkowski's Inequality) holds as well.

   Proof: Let $f, g \in \mathcal{C}([-1, 1])$ be nonzero, and note that the function $x \mapsto x^p$ is convex, so for each $t \in [0, 1]$ and $x \in [-1, 1]$,

$$\left(t\frac{|f(x)|}{||f||_p} + (1-t)\frac{|g(x)|}{||g||_p}\right)^p \leq t\left(\frac{|f(x)|}{||f||_p}\right)^p + (1-t)\left(\frac{|g(x)|}{||g||_p}\right)^p.$$

   In particular, since $\lambda \doteq \frac{||f||_p}{||f||_p + ||g||_p} \in [0, 1]$,

$$\left(\lambda\frac{|f(x)|}{||f||_p} + (1-\lambda)\frac{|g(x)|}{||g||_p}\right)^p \leq \lambda\left(\frac{|f(x)|}{||f||_p}\right)^p + (1-\lambda)\left(\frac{|g(x)|}{||g||_p}\right)^p.$$

so

$$\begin{aligned}\frac{1}{(||f||_p + ||g||_p)^p}\int(|f(x)| + |g(x)|)^p dx &= \int\left(\lambda\frac{|f(x)|}{||f||_p} + (1-\lambda)\frac{|g(x)|}{||g||_p}\right)^p dx \\ &\leq \frac{\lambda}{||f||_p^p}\int|f(x)|^p dx + \frac{(1-\lambda)}{||g||_p^p}\int|g(x)|^p dx \\ &= \frac{\lambda}{||f||_p}||f||_p^p + \frac{1-\lambda}{||g||_p}||g||_p^p \\ &= 1.\end{aligned}$$

It follows that

$$\int(|f(x)| + |g(x)|)^p dx \leq (||f||_p + ||g||_p)^p,$$

so it is enough for us to observe that

$$||f+g||_p^p = \int|f(x)+g(x)|^p dx \leq \int(|f(x)| + |g(x)|)^p dx = (||f||_p + ||g||_p)^p.$$

□

4. For any inner product space V, $\|x\| = \sqrt{\langle x, x\rangle}$ defines a norm.

Again, not all vector spaces have norms defined in them. For those with defined norms, they are called the **normed spaces**.

In general, we can naturally obtain a norm from a well defined inner product space. Let $\|v\| = \sqrt{\langle v, v\rangle}$ for $\forall v \in V$, where $\langle\cdot,\cdot\rangle$ is the inner product on the space V. It's not hard to verify all the requirements in the definition of norm (verify it by yourself). Thus, for any defined inner product, there is a naturally derived norm. However, not all normed spaces have a norm which is derived from an inner product.

Many normed spaces have norms which are not derived from an inner product (see the exercises at the end of this chapter).

Metric Space A more general notion of distance on the vector space is the **metric**. In fact, metrics do not require the vector space structure and can be defined on many types of spaces, such as graphs. We limit our focus to vector spaces V to reduce notation.

A metric is a function $d : V \times V \to \mathbb{R}$ such that for $x, y, z \in V$ it satisfies:

1. $d(x, y) = d(y, x)$
2. $d(x, y) \geq 0$, with equality if and only if $x = y$
3. $d(x, y) \leq d(x, z) + d(y, z)$ (triangle inequality)

A space equipped with a metric is called **metric space**. We will assume the reader is familiar with metric spaces and the basic related concepts, such as **topology, completeness, compactness, continuity and separability.**

Hilbert and Banach spaces, and separability A complete inner product space is called Hilbert space. A complete normed space is a Banach space.

We note that a Hilbert space is separable if and only if it has a countable orthonormal basis (see the definitions below). Also, **all finite-dimensional inner product spaces are Hilbert spaces.**
Examples:

1. Define a metric on the vertices of a connected graph as the shortest path length between them.
2. Let $V = [-1, 1]$ and define the discrete metric

$$d(x, y) = \begin{cases} 1 & x = y \\ 0 & x \neq y \end{cases}$$

For any normed space, we can naturally derive a metric as $d(x, y) = \|x - y\|$. This metric is said to be induced by the norm $\|\cdot\|$. However, the opposite is not true: the metric defined on a vector space need not be derived from a norm.

If a metric d on a vector space V satisfies the properties: $\forall x, y, z \in V$ and $\forall r \in \mathbb{R}$,

1. $d(x, y) = d(x + z, y + z)$ (translation invariance)
2. $d(rx, ry) = |r|d(x, y)$ (homogeneity)

then we can define a norm on V by $\|x\| := d(x, 0)$.

To sum up, the relation between the three special spaces is as follows:

$$\text{inner product} \rightarrow \text{norm} \rightarrow \text{metric}.$$

We mean that every inner product on a vector space induces a norm on that vector space, and every norm on a vector space induces a metric on that vector space.

Remark: Sometimes you will come across quantities that similar *like* one of these objects. For example, the Kullback-Leibler divergence (also called relative entropy) is a measure of how similar two probability distributions are. It is called a *distance* which is like a metric, but does not satisfy the triangle inequality. Another example is a *Kernel* which is a lot like an inner product, but it allow us to do statistics in more complicated spaces (that might not even be linear spaces e.g. string, image or graph kernels).

2.3 Properties and geometry of inner product spaces

We say that vectors v, w in an inner product space V are **orthogonal** if $\langle v, w\rangle = 0$. It is denoted as $v \perp w$.

An arbitrary set of vectors $\{v_i\}_{i\in I}$ in V is called **orthonormal** if $x_i \perp x_j$ and $\|x_i\| = 1$ for all $i, j \in I$.

Examples:

1. In $\mathbb{R}^n$ the notion of orthogonality agrees with our usual perception of it. If x is orthogonal to y, then **Pythagorean theorem** tells us that $\|x\|^2 + \|y\|^2 = \|x - y\|^2$. Expanding this in terms of inner products we get:

$$x^T x + y^T y = (x-y)^T(x-y) = x^T x - y^T x - x^T y + y^T y \text{ or } 2x^T y = 0$$

   and thus $\langle x, y\rangle = x^T y = 0$.

2. **Nonzero orthogonal vectors are linearly independent**. Suppose we have $q_1, q_2, \ldots, q_n$, a set of nonzero mutually orthogonal vectors in a finite-dimensional space V, i.e., $\langle q_i, q_j\rangle = 0\ \forall i \neq j$, and suppose that $r_1 q_1 + r_2 q_2 + \ldots + r_n q_n = 0$. Then taking inner product of q_1 with both sides, we have $r_1\langle q_1, q_1\rangle + r_2\langle q_1, q_2\rangle + \ldots + r_n\langle q_1, q_n\rangle = \langle q_1, 0\rangle = 0$. That reduces to $r_1\|q_1\|^2 = 0$ and since $q_1 \neq 0$, we conclude that $r_1 = 0$. Similarly, $r_i = 0\ \forall\ 1 \leq i \leq n$, and we conclude that $q_1, q_2, \ldots, q_n$ are linearly independent.

3. Suppose we have a $n \times 1$ vector of observations $x = [x_1, x_2, \cdots, x_n]^T \in \mathbb{R}^n$. Let $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \in \mathbb{R}$ be the mean of the observations and $c = [x_1 - \bar{x}, x_2 - \bar{x}, \cdots x_n - \bar{x}]^T$ be the vector of mean centered observations. Then c is orthogonal to the vector of ones $1_n = [1, 1, \cdots, 1]^T \in \mathbb{R}^n$, since

$$\begin{aligned} 1_n^T c &= \sum_{i=1}^{n} 1(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\bar{x} \\ &= \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i = 0 \end{aligned}$$

Orthogonal subspace and complement Suppose S, T are subspaces of a possibly infinite-dimensional inner-product space V. Then we say that they are **orthogonal subspaces** if every vector in S is orthogonal to every vector in T. We say that S is the **orthogonal complement** of T in V, if S contains ALL vectors orthogonal to vectors in T and we write $S = T^{\perp}$.

For example, the x-axis and y-axis are orthogonal subspaces of $\mathbb{R}^3$, but they are not orthogonal complements of each other, since y-axis does not contain $[0, 0, 1]^T$, which is perpendicular to every vector in x-axis. However, y-z plane and x-axis ARE orthogonal complements of each other in $\mathbb{R}^3$.

You should prove as an exercise that if $\dim(V) = n$, and $\dim(S) = k$, then $\dim(S^\perp) = n - k$.

Cauchy-Schwarz Inequality: for v and w elements of V, the following inequality holds:

$$\langle v, w\rangle^2 \leq \langle v, v\rangle \cdot \langle w, w\rangle$$

with equality if and only if v and w are linearly dependent.

Proof 1: Suppose that v and w are nonzero, and consider the second degree polynomial

$$p(t) = \langle tw + v, tw + v\rangle = \langle w, w\rangle t^2 + 2\langle w, v\rangle t + \langle v, v\rangle.$$

Note that $p(t)$ is nonnegative for all t, and thus has either one real root or two complex roots. In the first case the zero occurs at $t = -\frac{\langle w,v\rangle}{\langle w,w\rangle}$, so we have

$$\begin{aligned} 0 &= p\left(-\frac{\langle w, v\rangle}{\langle w, w\rangle}\right) \\ &= \langle w, w\rangle \left(\frac{\langle w, v\rangle}{\langle w, w\rangle}\right)^2 + 2\langle w, v\rangle \left(-\frac{\langle w, v\rangle}{\langle w, w\rangle}\right) + \langle v, v\rangle. \end{aligned}$$

Multiplying the expression above by $\langle w, w\rangle$ and simplifying reveals that

$$\langle w, v\rangle^2 = \langle v, v\rangle \cdot \langle w, w\rangle.$$

If there are two complex roots, then it must be the case that

$$\left(2\langle w, v\rangle\right)^2 - 4\langle w, w\rangle \cdot \langle v, v\rangle < 0,$$

which can be simplified to show that

$$\langle w, v\rangle^2 < \langle v, v\rangle \cdot \langle w, w\rangle.$$

If v and w are linearly independent, then $p(t)$ is strictly positive for all real t (as $tw + v$ will be nonzero for all t), which means that both roots are complex, and thus that equality does not hold. If v and w are linearly dependent, then we can find some real t such that $tw + v = 0$, which means that $p(t)$ has one real root, and, as shown above, equality holds. □

Proof 2: Note that $\langle v, 0\rangle = -\langle v, -0\rangle = -\langle v, 0\rangle \;\Rightarrow\; \langle v, 0\rangle = 0, \forall v \in V$. If $w = 0$, the equality obviously holds. If $w \neq 0$, let $\lambda = \frac{\langle v,w\rangle}{\langle w,w\rangle}$. Since

$$\begin{aligned} 0 &\leq \langle v - \lambda w, v - \lambda w\rangle \\ &= \langle v, v\rangle - 2\lambda\langle v, w\rangle + \lambda^2\langle w, w\rangle \\ &= \langle v, v\rangle - \frac{\langle v, w\rangle^2}{\langle w, w\rangle} \end{aligned}$$

□

With Cauchy-Schwarz inequality, we can define the **angle** between two nonzero vectors v and w as:

$$angle(v, w) = \arccos\left(\frac{\langle v, w\rangle}{\sqrt{\langle v, v\rangle \cdot \langle w, w\rangle}}\right)$$

The angle is in $[0, \pi)$. This generates nice geometry for the inner product space. For example, the Pythagorean theorem from Euclidean geometry holds in this abstract setting and is proved by direct computation as shown in the example above for $\mathbb{R}^n$.

Pythagorean theorem: Let V be an inner product space with inner product $\langle \cdot, \cdot \rangle$. For each pair of orthogonal vectors $x, y \in V$,

$$\|x + y\|^2 = \|x\|^2 + \|y\|^2$$

Parallelogram Identity: For each $x, y \in V$, an inner product space,

$$||x + y||^2 + ||x - y||^2 = 2||x||^2 + 2||y||^2.$$

Proof. Calculate

$$\begin{aligned} ||x + y||^2 + ||x - y||^2 &= \langle x + y, x + y\rangle + \langle x - y, x - y\rangle \\ &= \langle x, x\rangle + 2\langle x, y\rangle + \langle y, y\rangle + \langle x, x\rangle - 2\langle x, y\rangle + \langle y, y\rangle \\ &= 2||x||^2 + 2||y||^2. \end{aligned}$$

□

□

Any norm satisfying the Parallelogram Identity is derived from an inner product. The proof is straightforward and uses the **Polarization Identity**, which itself is proven by calculation.

Polarization Identity: Let V be an inner product space and let $||\cdot||$ be the norm derived from the inner product.

If V is over $\mathbb{R}$

$$\langle x, y\rangle = \frac{1}{4}\left(||x + y||^2 - ||x - y||^2\right) \quad \forall x, y \in V,$$

and if V is over $\mathbb{C}$

$$\langle x, y\rangle = \frac{1}{4}\left(||x + y||^2 - ||x - y||^2\right) - \frac{i}{4}\left(||x + iy||^2 - ||x - iy||^2\right) \quad \forall x, y \in V,$$

Law of cosines Thanks to the Cauchy-Schwartz inequality, we may define the following notion of an angle between $x, y \in V$, a real inner product space, by

$$ang(x, y) = \arccos\frac{\langle x, y\rangle}{\|x\|\|y\|} \in [0, \pi]$$

Thus $x \perp y$ has the geometric intepretation that x, y are at right angles to each other. We also recover the law of cosines by expanding $\|x-y\|^2$ and using the equality above, as

$$\|x-y\|^2 = \|x\|^2 + \|y\|^2 - 2\|x\|\|y\|\cos(ang(x,y))$$

2.4 Gram-Schmidt Process

Suppose we're given linearly independent vectors $v_1, v_2, \ldots, v_n$ in an inner product space V. Then we know that $v_1, v_2, \ldots, v_n$ form a basis for the subspace which they span. Then, the **Gram-Schmidt process** can be used to construct an orthogonal basis for this subspace, as follows:

Let $q_1 = v_1$ Suppose v_2 is not orthogonal to v_1. then let rv_1 be the **projection** of v_2 on v_1, i.e. we want to find $r \in \mathbb{R}$ s.t. $q_2 = v_2 - rq_1$ is orthogonal to q_1. Well, we should have $\langle q_1, (v_2 - rq_1)\rangle = 0$, and we get $r = \frac{\langle q_1, v_2\rangle}{\langle q_1, q_1\rangle}$. Notice that the span of q_1, q_2 is the same as the span of v_1, v_2, since all we did was to subtract multiples of original vectors from other original vectors.

Proceeding in similar fashion, we obtain

$$q_i = v_i - \Big(\frac{\langle q_1, v_i\rangle}{\langle q_1, q_1\rangle} q_1 + \ldots + \frac{\langle q_{i-1}, v_i\rangle}{\langle q_{i-1}, q_{i-1}\rangle} q_{i-1}\Big),$$

and we thus end up with an orthogonal basis for the subspace. If we furthermore divide each of the resulting vectors $q_1, q_2, \ldots, q_n$ by its length, we are left with **orthonormal basis**, i.e. $\langle q_i, q_j\rangle = 0 \ \forall i \neq j$ and $\langle q_i, q_i\rangle = 1, \forall i$ (why?). We call these vectors that have length 1 **unit** vectors.

Gram-Schmidt is used to prove the following theorem in the finite and countably-infinite-dimensional cases by induction. In the uncountably infinite-dimensional case one must use Zorn's lemma.

Theorem: Every Hilbert space has an orthonormal basis.

You can now construct an orthonormal basis for the subspace of $F_{[-1,1]}$ spanned by $f(x) = 1, g(x) = x$, and $h(x) = x^2$ (Exercise 2.6 (b)). An important point to take away is that given any basis for finite-dimensional V, if there's an inner product defined on V, we can always turn the given basis into an orthonormal basis.

Example Consider the space $V = L^2(\mathbb{R}, \gamma(dx))$ where $\gamma(dx) = (2\pi)^{-1/2} e^{-x^2/2}$, that is the space of real-valued functions f such that, if X is a standard normal random variable we have [1]

[1] Tehnicality: Elements in V are not in fact functions but equivalence classes of functions equal to each other except possibly on a set B such that $\int_B dx = 0$. Equality in this example should be understood in this sense.

$$\mathbb{E}f(X)^2 = \int_{-\infty}^{\infty} f(x)^2 \gamma(dx) < \infty$$

,

with inner product

$$\langle f, g \rangle = \int_{-\infty}^{\infty} f(x)g(x)\gamma(dx).$$

Define the set of vectors $p_i(x) = x^i$, $i = 0, 1 \ldots$. It can be shown that

1. the polynomials (linear combinations of p_i) are dense in V in the norm metric
2. $\langle f, p_i \rangle = 0$ for all i implies $f = 0$, which implies the closure of $span\{p_i\}$ is V
3. $\{p_i\}$ are linearly independent and so, using the previous statement, form a basis
4. Gram-Schmidt on $\{p_i\}$ (without the final normalization) gives an orthogonal basis for V given by the **Hermite polynomials**, which can be written succinctly as $H_n(x) = (-1)^n e^{x^2/2} \frac{d^n}{dx^n} e^{-x^2/2}$ for $n \geq 1$ and $H_0 = 1$.

2.5 Comments on infinite-dimensional spaces

Much of these notes concerns finite-dimensional spaces. Many results, and the geometric intuition, from such spaces extend to infinite-dimensional Hilbert spaces. But there are some important differences. We mention just a few here, in the case where V is an infinite-dimensional Hilbert space with countable basis $\{v_i\}_1^\infty$.

This first example shows why V is **not the span but the closure of the span of an orthonormal basis.**

Theorem For each infinite sequence of vectors $\{x_i\}_1^\infty \subset V$, there exists a $x_0 \in V$ which is not a finite linear combination of the $\{x_i\}$.

To prove this, we may assume $\{x_i\}$ form an orthonormal for span$\{x_i\}$, since otherwise we could just apply Gram-Schmidt without changing the span of the set. If the span of $\{x_i\}$ is in fact finite-dimensional, then the orthogonal decomposition theorem (see below) and the fact that $\dim V = \infty$ show there must exist x_0 as claimed.

If the span is infinite-dimensional we consider the closure of span$\{x_i\}$ to give a closed subspace and therefore a Hilbert space inheriting the same inner product. Any span of a finite number of x_i is a closed subspace, and therefore there must be an x_0 not contained in it by the orthogonal decomposition theorem.

To be more constructive, one could consider $x_0 = \sum_1^\infty x_i/n$. This is in V since it is summable in squared norm. If x_0 could be written as a finite linear combination $x_0 = \sum_1^k \alpha_i x_i$, say, there would exist an N such that x_N is not part of the sum. By orthogonality $\langle x_0, x_N \rangle = 0$. But $\langle x_0, x_j \rangle = 1/j > 0$ for all $j \geq 1$, a contradiction.

Theorem The unit ball $\{x \in V \mid ||x|| \leq 1\}$ is not compact.

The proof of this second theorem follows from the fact that if $\{v_i\}$ form an orthonormal basis, then the Pythagorean theorem says $||v_n - v_m|| = \sqrt{2}$ for all $n \neq m$. Therefore, no subsequence of $\{v_i\}$ can be Cauchy, and therefore it has no convergent subsequence.

This is in contrast to the case of finite-dimensional Euclidean spaces, where closed and bounded sets are compact.

Theorem Every separable, infinite-dimensional Hilbert V space over $\mathbb{C}$ is isometrically isomorphic to the space $\ell^2 = \{x = (x_1 \dots) \mid x_i \in \mathbb{C}, \sum |x_i|^2 < \infty\}$. In other words, there exists a bijective linear map T such that $\langle Tx, Tx\rangle_{\ell^2} = \langle x, x\rangle_V$.

Exercises

2.1 Let $\ell^p = \left\{ x = (x_1, x_2, \dots) \in \mathbb{R}^{\mathbb{N}} : ||x||_p = \left(\sum_{i=1}^{\infty} |x_i|^p \right)^{\frac{1}{p}} < \infty \right\}$.

1. Suppose $p, q > 1$ such that $1/p + 1/q = 1$. For $x = \{x_i\} \in \ell^p$ and $y = \{y_i\} \in \ell^q$, show $\sum |x_i||y_i| \leq \|x\|_p \|y\|_q$. (Hint: Use Young's inequality, $ab \leq a^p/p + b^q/q$ for $a, b > 0$ and p, q as above.)
2. For $p \geq 1$ and $x, y \in \ell^p$, show $\|x + y\|_p \leq \|x\|_p + \|y\|_p$.
3. Prove $|| \cdot ||_p$ is a norm.
4. Prove that $|| \cdot ||_p$ is derived from an inner product if and only if $p = 2$. (Hint: Use the parallelogram identity and a very simple example).

2.2 Prove: Every separable, infinite-dimensional Hilbert V space over $\mathbb{C}$ is isometrically isomorphic to the space $\ell^2 = \{x = (x_1 \dots) \mid x_i \in \mathbb{C}, \sum |x_i|^2 < \infty\}$. In other words, there exists a bijective linear map T such that $\langle Tx, Tx\rangle_{\ell^2} = \langle x, x\rangle_V$.

Hint: Recall that V is separable if and only if it has a countable orthonormal basis $\{x_i\}_1^\infty$. Consider the map $T(x) = (\langle x, x_1\rangle, \langle x, x_2\rangle \dots)$.

2.3 Show that the space F_0 of all differentiable functions $f : \mathbb{R} \to \mathbb{R}$ with $\frac{df}{dx} = 0$ defines a vector space.

2.4 *Garcia, P.4.27* Let $x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \in \mathbb{R}^3$ and let $||x||_{\alpha,\beta,\gamma} = \alpha|x_1| + \beta|x_2| + \gamma|x_3|$, where $\alpha, \beta, \gamma > 0$. Show that $|| \cdot ||_{\alpha,\beta,\gamma}$ is a norm on $\mathbb{R}^3$, but that it is not derived from an inner product.

2.5 Let d be a metric on a vector space V such that

1. $d(x, y) = d(x + z, y + z)$
2. $d(rx, ry) = |r|d(x, y)$

hold for each $x, y, z \in V$ and each $r \in R$. Prove that $\|x\| \doteq d(x, 0)$ defines a norm on V.

2.6 Verify for yourself that the two conditions for a subspace are independent of each other, by coming up with 2 subsets of $\mathbb{R}^2$: one that is closed under addition and subtraction but NOT under scalar multiplication, and one that is closed under scalar multiplication but NOT under addition/subtraction.

2.7 *Strang, section 3.5 #17b* Let V be the space of all vectors $v = [c_1\, c_2\, c_3\, c_4]^T \in \mathbb{R}^4$ with components adding to 0: $c_1 + c_2 + c_3 + c_4 = 0$. Find the dimension and give a basis for V.

2.8 Let $v_1, v_2, ..., v_n$ be a linearly independent set of vectors in V. Prove that if $n = dim(V)$, $v_1, v_2, ..., v_n$ form a basis for V.

2.9 If $F_{[-1,1]}$ is the space of all continuous functions defined on the interval $[-1, 1]$, show that $\langle f, g\rangle = \int_{-1}^{1} f(x)g(x)dx$ defines an inner product of $F_{[-1,1]}$.

2.10 Parts (a) and (b) concern the space $F_{[-1,1]}$, with inner product $\langle f, g\rangle = \int_{-1}^{1} f(x)g(x)dx$.

(a) Show that $f(x) = 1$ and $g(x) = x$ are orthogonal in $F_{[-1,1]}$

(b) Construct an orthonormal basis for the subspace of $F_{[-1,1]}$ spanned by $f(x) = 1, g(x) = x$, and $h(x) = x^2$.

2.11 If a subspace S is contained in a subspace V, prove that $S^{\perp}$ contains $V^{\perp}$.

2.12 Suppose we have have two sets of paired observations $(x_1, y_1), \ldots, (x_n, y_n)$ and let $x, y \in \mathbb{R}^n$ be the resulting vectors of observations. The *cosine similarity* between these two observations is given by

$$\text{sim}(x, y) := \arccos\left(\frac{x^T y}{||x||||y||}\right)$$

i.e. $\text{sim}(x, y)$ is a measure of the angle between x and y in $\mathbb{R}^n$.

(a) Look up the formula for *Pearson correlation* and write it in terms of the vectors x, y. What is the difference between Pearson correlation and cosine similarity?

(b) Give a geometric interpretation of Pearson correlation. *Hint*: see example 3 from Section 2.3.

2.13

1. Show that every finite-dimensional normed space V over $\mathbb{R}$ is topologically isomorphic with $\mathbb{R}^n$, i.e. there exists a continuous linear map $T : V \mapsto \mathbb{R}^n$ with continuous inverse, where $\mathbb{R}^n$ has the standard Euclidean norm.

2. Show that all norms on V as above are equivalent, in the sense that if $\|x_n\|_1 \to 0$ then $\|x_n\|_2 \to 0$ for any two norms $\|\cdot\|_1$, $\|\cdot\|_2$.

3 Matrices and Matrix Algebra

An $m \times n$ matrix A is a rectangular array of numbers that has m rows and n columns, and we write:

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}$$

For the time being we'll restrict ourselves to real matrices, so $\forall\ 1 \leq i \leq m$ and $\forall\ 1 \leq j \leq n$, $a_{ij} \in \mathbb{R}$. Notice that a familiar vector $x = [x_1, x_2, \cdots x_n]^T \in \mathbb{R}^n$ is just a $n \times 1$ matrix (we say x is a **column vector**.) A $1 \times n$ matrix is referred to as a **row vector**. If $m = n$, we say that A is **square**.

3.1 Matrix Operations

Matrix addition Matrix addition is defined elementwise, i.e. $A + B := C$, where

$$c_{ij} = a_{ij} + b_{ij}$$

Note that this implies that $A + B$ is defined only if A and B have the same dimensions. Also, note that matrix addition is commutative i.e. $A + B = B + A$.

Scalar multiplication Scalar multiplication is also defined element-wise. If $r \in \mathbb{R}$, then $rA := B$, where

$$b_{ij} = ra_{ij}$$

Any matrix can be multiplied by a scalar. Multiplication by 0 results in zero matrix, and multiplication by 1 leaves matrix unchanged, while multiplying A by -1 results in matrix $-A$, s.t. $A + (-A) = A - A = 0_{m \times n}$.

You should check at this point that a set of all $m \times n$ matrices is a vector space with operations of addition and scalar multiplication as defined above.

Matrix multiplication Matrix multiplication is trickier. Given a $m \times n$ matrix A and a $p \times q$ matrix B, AB is only defined if $n = p$. In that case we have $AB = C$, where

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$$

i.e. the i, j-th element of AB is the inner product of the i-th row of A and j-th column of B, and the resulting product matrix is $m \times q$. See http://matrixmultiplication.xyz/ for a nice visualization of matrix multiplication.

You should at this point come up with your own examples of A, B s.t both AB and BA are defined, but $AB \neq BA$. Thus matrix multiplication is, in general, non-commutative.

Trace We also introduce another concept here: for a square matrix A, its **trace** is defined to be the sum of the entries on main diagonal($tr(A) = \sum_{i=1}^{n} a_{ii}$). For example, $tr(I_{n\times n}) = n$. You may prove for yourself (by method of entry-by-entry comparison) that $tr(AB) = tr(BA)$, and $tr(ABC) = tr(CAB)$. It's also immediately obvious that $tr(A + B) = tr(A) + tr(B)$.

Transpose Let A be $m \times n$, then the **transpose** of A is the $n \times m$ matrix A^T, s.t. $a_{ij} = a^T_{ji}$. Now the notation we used to define the inner product on $\mathbb{R}^n$ makes sense, since given two $n \times 1$ column vectors x and y, their inner product $\langle x, y\rangle$ is just $x^T y$ according to matrix multiplication.

Inverse Let $I_{n\times n}$, denote the $n \times n$ **identity** matrix, i.e. the matrix that has 1's down its main diagonal and 0's everywhere else (in future we might omit the dimensional subscript and just write I, the dimension should always be clear from the context). You should check that in that case, $I_{n\times n}A = AI_{n\times n} = A$ for every $n \times n$ A. We say that $n \times n$ A, has $n \times n$ **inverse**, denoted A^{-1}, if $AA^{-1} = A^{-1}A = I_{n\times n}$. If A has inverse, we say that A is **invertible**.

Not every matrix has inverse, as you can easily see by considering the $n \times n$ zero matrix. A square matrix that is not invertible is called **singular** or **degenerate**. We will assume that you are familiar with the use of elimination to calculate inverses of invertible matrices and will not present this material.

The following are some important results about inverses and transposes:

1. $(AB)^T = B^T A^T$

   Proof: Can be shown directly through entry-by-entry comparison of $(AB)^T$ and $B^T A^T$.

2. If A is invertible and B is invertible, then AB is invertible, and $(AB)^{-1} = B^{-1}A^{-1}$.

   Proof: Exercise 3.1(a).

3. If A is invertible, then $(A^{-1})^T = (A^T)^{-1}$

   Proof: Exercise 3.1(b).

4. A is invertible if and only if $Ax = 0 \Longrightarrow x = 0$ (we say that $N(A) = \{0\}$, where $N(A)$ is the nullspace of A, to be defined shortly).

   Proof: Assume A^{-1} exists. Then,

$$\begin{aligned} & Ax = 0 \\ \rightarrow \quad & A^{-1}(Ax) = A^{-1}0 \\ \rightarrow \quad & x = 0. \end{aligned}$$

Now, assume $Ax = 0$ implies $x = 0$. Note that mulitplying A by a vector on the right is a linear combination of the columns of A. By the assumption we get of A are linearly independent and therefore form a basis for $\mathbb{R}^n$ (since there are n of them).

Consider the standard basis vectors $e_1 = [1, 0, \cdots, 0]^T$, $e_2 = [0, 1, \cdots, 0]^T$, ..., $e_n = [0, 0, \cdots, 1]^T$. Since the columns of A for a basis, for each standard basis vector e_i we can find a vector c_i such that $Ac_i = e_i$.

Convince yourself that the matrix C whose columns are given by these c_i is a right inverse of A i.e. $AC = I$. $\square$

Perspectives on matrix-vector multiplication There are a couple different ways to think of multiplying a matrix times a vector. Suppose we have a matrix $A \in \mathbb{R}^{n \times d}$ and a vector $v \in \mathbb{R}^d$. The result of multiplying the matrix A by the vector v on the right is a vector i.e. $Av \in \mathbb{R}^n$.

1. Av gives a linear combination of the columns of A. Let $A_j, j = 1, \ldots, d$ be the columns of A. Then $Av = \sum_{j=1}^{d} v_j A_j$.
2. Av give the dot product between the rows of A and v. Let $a_i, i = 1, \ldots, n$ be the rows of A. Then the ith entry $(Av)_i = a_i^T v$.

We can make similar observations for multiplying A by a vector $u \in \mathbb{R}^n$ on the left by switching the words "row" and "column" i.e. uA is a linear combination of the rows of A.

Perspectives on matrix-matrix multiplication Suppose we have two matrices $A \in \mathbb{R}^{n \times k}$ and $B \in \mathbb{R}^{k \times d}$. Let $C = AB \in \mathbb{R}^{n \times d}$. Denote the columns of the matrix A be given by $A_1, \ldots, A_k$ where $A_j \in \mathbb{R}^n$. Denote the rows of A by $a_1, \ldots, a_n$ where $a_i \in \mathbb{R}^k$. Similarly for B (i.e. capital letter for columns, lower case letter for rows).

1. The jth column of C is given by A times the jth column of B i.e. $C_j = AB_j$. We can apply either of the two matrix-vector interpretations to AB_j.
2. The ith row of C is given by the ith row of A times B i.e. $c_i = a_i B$. Again we can apply the above two matrix-vector interpetations.
3. C is give by the sum of the outer products between the columns of A and the rows of B i.e. $C = A_1 b_1^T + \cdots + A_k b_k^T$.

3.2 Special Matrices

Symmetric Matrix A square matrix A is said to be **symmetric** if $A = A^T$. If A is symmetric, then A^{-1} is also symmetric (Exercise 3.2).

Orthogonal Matrix A square matrix Q is said to be **orthogonal** if $Q^T = Q^{-1}$. You should prove that columns of an orthogonal matrix are orthonormal, and so are the rows. Conversely, any square matrix with orthonormal columns is orthogonal. We note that orthogonal matrices preserve lengths and inner products:

$$\langle Qx, Qy \rangle = x^T Q^T Q y = x^T I_{n\times n} y = x^T y.$$

In particular $\|Qx\| = \sqrt{x^T Q^T Q x} = \|x\|$, which means that Q is an **isometry**. Also, if A, and B are orthogonal, then so are A^{-1} and AB.

Idempotent Matrix We say that a square matrix A is **idempotent** if $A^2 = A$.

Positive Definite Matrix We say that a square matrix A is **positive definite** if $\forall\ n \times 1$ vectors $x \neq 0_{n\times 1}$, we have $x^T A x > 0$. We say that A is **positive semi-definite** (or **non-negative definite** if A is symmetric and $\forall\ n \times 1$ vectors $x \neq 0_{n\times 1}$, we have $x^T A x \geq 0$. You should prove for yourself that every positive definite matrix is invertible (Exercise 3.3)). One can also show that if A is positive definite, then so is A^T (more generally, if A is positive semi-definite, then so is A^T).

These have complex matrix analogs by replacing the transpose with the Hermitian transpose. In the exercises, you will show that if A is a complex matrix, non-negative definiteness implies the matrix is Hermitian. This is why some definitions of positive definiteness for complex matrices include the requirement that it be Hermitian. This is not true in the real case.

Diagonal and Triangular Matrix We say that a square matrix A is **diagonal** if $a_{ij} = 0$ $\forall\ i \neq j$. We say that A is **upper triangular** if $a_{ij} = 0\ \forall\ i > j$. **Lower triangular** matrices are defined similarly.

3.3 The Four Fundamental Spaces

Column Space and Row Space Let A be $m \times n$. We will denote by $col(A)$ the subspace of $\mathbb{R}^m$ that is spanned by columns of A, and we'll call this subspace the **column space** of A. Similarly, we define the **row space** of A to be the subspace of $\mathbb{R}^n$ spanned by rows of A and we notice that it is precisely $col(A^T)$.

Nullspace and Left Nullspace Now, let $N(A) = \{x \in \mathbb{R}^n : Ax = 0\}$. You should check for yourself that this set, which we call **kernel** or **nullspace** of A, is indeed subspace of $\mathbb{R}^n$. Similary, we define the **left nullspace** of A to be $\{x \in \mathbb{R}^m : x^T A = 0\}$, and we notice that this is precisely $N(A^T)$.

The fundamental theorem of linear algebra states:

1. $\dim(col(A)) = r = \dim(col(A^T))$. Dimension of column space is the same as dimension of row space. This dimension is called the **rank** of A.

2. $col(A) = (N(A^T))^\perp$ and $N(A) = (col(A^T))^\perp$. The columns space is the orthogonal complement of the left nullspace in $\mathbb{R}^m$, and the nullspace is the orthogonal complement of the row space in $\mathbb{R}^n$. We also conclude that $\dim(N(A)) = n - r$, and $\dim(N(A^T)) = m - r$.

We will not present the proof of the theorem here, but we hope you are familiar with these results. If not, you should consider taking a course in linear algebra (math 383).

We can see from the theorem, that the columns of A are linearly independent if and only if the nullspace doesn't contain any vector other than zero. Similarly, rows are linearly independent if and only if the left nullspace doesn't contain any vector other than zero.

Solving Linear Equations We now make some remarks about solving equations of the form $Ax = b$, where A is a $m \times n$ matrix, x is $n \times 1$ vector, and b is $m \times 1$ vector, and we are trying to solve for x. First of all, it should be clear at this point that if $b \notin col(A)$, then the solution doesn't exist. If $b \in col(A)$, but the columns of A are not linearly independent, then the solution will not be unique. That's because there will be many ways to combine columns of A to produce b, resulting in many possible x's. Another way to see this is to notice that if the columns are dependent, the nullspace contains some non-trivial vector x^*, and if x is some solution to $Ax = b$, then $x + x^*$ is also a solution. Finally we notice that if $r = m < n$ (i.e. if the rows are linearly independent), then the columns MUST span the whole $\mathbb{R}^m$, and therefore a solution exists for every b (though it may not be unique).

We conclude then, that if $r = m$, the solution to $Ax = b$ always exists, and if $r = n$, the solution (if it exists) is unique. This leads us to conclude that if $n = r = m$ (i.e. A is full-rank square matrix), the solution always exists and is unique. The proof based on elimination techniques (which you should be familiar with) then establishes that a square matrix A is full-rank if and only if it is invertible.

We now give the following results:

1. $\text{rank}(A^TA) = \text{rank}(A)$. In particular, if $\text{rank}(A) = \text{n}$ (columns are linearly independent), then A^TA is invertible. Similarly, $\text{rank}(AA^T) = \text{rank}(A)$, and if the rows are linearly independent, AA^T is invertible.

   Proof: Exercise 3.5.

2. $N(AB) \supset N(B)$

   Proof: Let $x \in N(B)$. Then,

   $$(AB)x = A(Bx) = A0 = 0,$$

   so $x \in N(AB)$. $\square$

3. $col(AB) \subset col(A)$, the column space of product is subspace of column space of A.

   Proof: Note that

$$col(AB) = N((AB)^T)^\perp = N(B^T A^T)^\perp \subset N(A^T)^\perp = col(A). \quad \square$$

4. $col((AB)^T) \subset col(B^T)$, the row space of product is subspace of row space of B.

   Proof: Similar to (3).

3.4 Sample Covariance matrix

Suppose we have a data matrix $X \in \mathbb{R}^{n \times d}$ with n observations and d variables. Let $x_1, \ldots, x_n \in \mathbb{R}^n$ be the observations (rows of X). We define the sample covariance matrix matrix $S \in \mathbb{R}^{d \times d}$ by

$$S := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T$$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \in \mathbb{R}^d$. Note the i$th$ diagonal entry, S_{ii} is the sample variance of the ith variable. The off diagonal entry S_{ij} is the sample covariance between the ith variable and jth variable.

First we state some facts about the sample covariance matrix.

1. S is symmetric.

2. S is positive semi-definite.

   proof: Let $A \in \mathbb{R}^d$ be some matrix given by $A = \sum_{i=1}^{n} y_i y_i^T$ where $y_i \in \mathbb{R}^d$. Note S is in this form with $y_i = x_i - \bar{x}$. Let $v \in \mathbb{R}^d$ by any vector.

$$v^T A v = v^T \left(\sum_{i=1}^{n} y_i y_i^T \right) v \tag{1}$$

$$= \sum_{i=1}^{n} (v^T y_i)(y_i^T v) \tag{2}$$

$$= \sum_{i=1}^{n} (v^T y_i)^2 \geq 0. \tag{3}$$

   Thus any matrix in the form of A is positive semi definite

3. Let $P := \text{ span}(x_1 - \bar{x}, \ldots, x_n - \bar{x}) \subseteq \mathbb{R}^d$ subspace spanned by the mean centered observations and $r := dim(P)$ be the rank of this subspace. Then $\text{rank}(S) = r \leq \min n, d$. Note if we are in the high dimensional, low sample size setting (i.e. $n < d$) the S is automatically not invertible. On the other hand, if $n > d$ and the data are randomly generated then S is almost surely invertible.

4. We can write S in a couple different ways which may be useful in different contexts. Let $X_c \in \mathbb{R}^{n \times d}$ be the data matrix after the columns have mean centered (i.e. the ith row of X_c is given by $x_i - \bar{x}$) and $1_n \in \mathbb{R}^n$ be the vector of ones.

$$
\begin{aligned}
S &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i x_i^T - n\bar{x}\bar{x}^T\right) \\
&= \frac{1}{n-1}\left(X^T X - \frac{1}{n} X^T 1_n 1_n^T X\right) \\
&= \frac{1}{n-1} X_c^T X_c
\end{aligned}
$$

Exercises

3.1 Prove the following results:

(a) If A is invertible and B is invertible, then AB is invertible, and $(AB)^{-1} = B^{-1}A^{-1}$

(b) If A is invertible, then $(A^{-1})^T = (A^T)^{-1}$

3.2 Let $A, B \in \mathbb{R}^{n\times n}$ be orthogonal matrices. Is $A + B$ orthogonal?

3.3 Show that if A is symmetric, then A^{-1} is also symmetric.

3.4 Show that any positive definite matrix A is invertible (think about nullspaces).

3.5 *Horn & Johnson 1.2.2* For $A : n \times n$ and invertible $S : n \times n$, show that $tr(S^{-1}AS) = tr(A)$. The matrix $S^{-1}AS$ is called a **similarity** of A.

3.6 Show that $\text{rank}(A^T A) = \text{rank}(A)$. In particular, if $\text{rank}(A) = \text{n}$ (columns are linearly independent), then $A^T A$ is invertible. Similarly, show that $\text{rank}(AA^T) = \text{rank}(A)$, and if the rows are linearly independent, AA^T is invertible. (Hint: show that the nullspaces of the two matrices are the same).

3.7 Linear regression takes a data matrix $X \in \mathbb{R}^{n\times d}$ (n observations and d variables) and a vector of coefficients $\beta \in \mathbb{R}^d$ then outputs a prediction vector $\hat{Y} = X\beta \in \mathbb{R}^n$. Explain what is happening with the observations/variables of X using the two perspectives on matrix-vector multiplication.

3.8 Let $A \in \mathbb{R}^{n\times d}$ and define the *Frobenius norm* by $\|A\|_F = \sqrt{tr(A^T A)}$. Show this is a norm.

3.9 Show S can be written in the three different ways given in fact 4 in section 3.4.

3.10 Suppose $d > n$ and we have a matrix $A \in \mathbb{R}^{d\times d}$ given by $A = \sum_{i=1}^{n} y_i y_i^T + \beta D$ where $\beta > 0$ is a constant and $D \in \mathbb{R}^{d\times d}$ is a diagonal matrix with strictly positive entries. Show that A is positive definite (i.e. for all $v \neq 0$, $v^T Av > 0$). Argue why this must mean A is invertible.

Now read Section 4.3.1 on *regularized discriminant analysis* from Elements of Statistical Learning (PDF available at https://web.stanford.edu/~hastie/Papers/ESLII.pdf). Pay particular attention to equations 4.13 and 4.14.

4 Projections and Least Squares Estimation

4.1 Projections

In an inner product space, suppose we have n linearly independent vectors $a_1, a_2, \ldots, a_n$ in $\mathbb{R}^m$, and we want to find the projection of a vector b in $\mathbb{R}^m$ onto the space spanned by $a_1, a_2, \ldots, a_n$, i.e. to find some linear combination $x_1a_1 + x_2a_2 + \ldots + x_na_n = b^*$, s.t. $\langle b^*, b - b^* \rangle = 0$. It's clear that if b is already in the span of $a_1, a_2, \ldots, a_n$, then $b^* = b$ (vector just projects to itself), and if b is perpendicular to the space spanned by $a_1, a_2, \ldots, a_n$, then $b^* = 0$ (vector projects to the zero vector).

Hilbert Projection Theorem Assume V is a Hilbert space (complete inner product space) and S is a closed convex subset of V. For any $v \in V$, there exist an unique y in S s.t.

$$y = \arg\min_{x \in S} \|v - x\|$$

The vector y is called the **projection** of the vector v onto the subset S.

Proof: Let $d \doteq \inf_{s \in S} ||v - s||$, and let $\{y_n\} \subseteq S$ be a sequence such that $||y_n - v|| \to d$. In order to see that $\{y_n\}$ is Cauchy, apply the Parallelogram Law to $u_n \doteq v - y_n$ and $w_m \doteq y_m - v$ to see that

$$||y_m - y_n||^2 + ||y_m + y_n - 2v||^2 = 2||y_m - v||^2 + 2||v - y_n||^2.$$

Rearranging the previous identity,

$$\begin{aligned} ||y_m - y_n||^2 &= 2||y_m - v||^2 + 2||v - y_n||^2 - ||y_m + y_n - 2v||^2 \\ &= 2||y_m - v||^2 + 2||v - y_n||^2 - 4\left|\left|\frac{y_m + y_n}{2} - v\right|\right|^2. \end{aligned}$$

Since $y_m, y_n \in S$ and S is convex, $\frac{y_m + y_n}{2} \in S$, and therefore $\left|\left|\frac{y_m + y_n}{2} - v\right|\right|^2 \geq d^2$, so

$$||y_m - y_n||^2 \leq 2||y_m - v||^2 + 2||v - y_n||^2 - 4d^2 \to 0.$$

Since V is complete , $\{y_n\}$ is Cauchy, and S is closed, there is some $y \in S$ such that $y_n \to y$. Additionally $x \mapsto ||x||$ is continuous, so

$$||v - y|| = \lim_{n \to \infty} ||v - y_n|| = d.$$

In order to show that the y is unique, suppose that there is some projection, z, of v onto S such that $y \neq z$. Consider the sequence

$$\{z_n\} = \{y, z, y, z, \ldots\},$$

and note that $\lim_{n \to \infty} ||z_n - v|| = d$ (in particular, $||z_n - v|| = d$ for all $n \geq 1$), so $\{z_n\}$ must be Cauchy, which is a contradiction, since $\{z_n\}$ is not convergent. $\square$

Example: If the set $S = \{x | ||x||_2 \leq 1\}$ then the projection is given by $\frac{x}{\max(1, ||x||)}$.

Corollary: Assume that V is as above, and that S is a closed subspace of V. Then s^* is the projection of $v \in V$ onto S if and only if $\langle v - s^*, s\rangle = 0 \ \ \forall s \in S$.

Proof: Let $v \in V$, and assume that s^* is the projection of v onto S. The result holds trivialy if $s = 0$ so assume $s \neq 0$. Since S is a subspaces, $s^* - ts \in S$. By the Projection Theorem for all $s \in S$ the function $f_s(t) = \|v - s^* + ts\|^2$ has a minimum at $t = 0$. Rewriting this function we see

$$\begin{aligned} f_s(t) &= \|v - s^* + ts\|^2 \\ &= \langle v - s^* + ts, v - s^* + ts\rangle \\ &= \langle ts, ts\rangle + 2\langle v - s^*, ts\rangle + \langle v - s^*, v - s^*\rangle \\ &= t^2\|s\|^2 + 2t\langle v - s^*, s\rangle + \|v - s^*\|^2. \end{aligned}$$

Since this is a quadratic function of t with positive quadratic coefficient, the minimum must occur at the vertex, which implies $\langle v - s^*, s\rangle = 0$.

For the opposite direction, note first that the function $f_s(t)$ will still be minimized at $t = 0$ for all $s \in S$. Then for any $s' \in S$ take $s \in S$ such that $s = s^* - s'$. Then taking $t = 1$ it follows that

$$\|v - s^*\| = f_s(0) \leq f_s(1) = \|v - s^* + s^* - s'\| = \|v - s'\|.$$

Thus by s^* is the projection of v onto S. $\square$

The following facts follow from the Projection Theorem and its Corollary.
Fact 1: The projection onto a closed subspace S of V, denoted by P_S, is a linear operator.

Proof: Let $x, y \in V$ and $a, b \in \mathbb{R}$. Then for any $s \in S$

$$\begin{aligned} \langle ax + by - aP_Sx - bP_Sy, s\rangle &= \langle a(x - P_Sx), s\rangle + \langle b(y - P_Sy), s\rangle \\ &= a\langle x - P_Sx, s\rangle + b\langle y - P_Sy, s\rangle \\ &= a \cdot 0 + b \cdot 0 = 0. \end{aligned}$$

Thus by the Corollary $P_S(ax + by) = aP_Sx + bP_Sy$, and P_S is linear. $\square$

Fact 2: Let S be a closed subspace of V. Then every $v \in V$ can be written uniquely as the sum of $s_1 \in S$ and $t_1 \in S^\perp$.

Proof: That $V \subset S + S^\perp$ follows from the Corollary and taking $s_1 = P_Sv$ and $t_1 = v - P_Sv$ for any $v \in V$. To see that this is unique assume that $s_1, s_2 \in S$ and $t_1, t_2 \in S^\perp$ are such that

$$s_1 + t_1 = v = s_2 + t_2.$$

Then $s_1 - s_2 = t_2 - t_1$, with $s_1 - s_2 \in S$ and $t_2 - t_1 \in S^\perp$, since each is a subspace of V. Therefore $s_1 - s_2, t_2 - t_1 \in S \cap S^\perp$ which implies

$$s_1 - s_2 = t_2 - t_1 = 0 \text{ or } s_1 = s_2 \text{ and } t_1 = t_2. \ \square$$

Fact 3: Let S and V be as above. Then for any $x, y \in V$, $\|x - y\| \geq \|P_S x - P_S y\|$.

Proof: First for any $a, b \in V$,

$$\begin{aligned}\|a\|^2 = \|a - b + b\|^2 &= \langle a - b + b, a - b + b\rangle \\ &= \langle a - b, a - b + b\rangle + \langle b, a - b + b\rangle \\ &= \langle a - b, a - b\rangle + 2\langle a - b, b\rangle + \langle b, b\rangle \\ &= \|a - b\|^2 + \|b\|^2 + 2\langle a - b, b\rangle.\end{aligned}$$

Taking $a = x - y$ and $b = P_S x - P_S y$, Fact 1 and the Corollary imply that $\langle a - b, b\rangle = 0$ and thus

$$\|x - y\|^2 = \|a\|^2 = \|a - b\|^2 + \|b\|^2 \geq \|b\|^2 = \|P_S x - P_S y\|^2. \square$$

Now let us focus on the case when $V = \mathbb{R}^m$ and $S = \text{span}\{a_1, \ldots, a_n\}$ where $a_1, \ldots, a_n$ are linearly independent and $A = [a_1 \ldots a_n]$

Fact 4: Suppose $a_1, \ldots, a_n$ are orthonormal (i.e. mutually orthogonal, unit norm). Then the **projection matrix** is given by $P_S = AA^T = \sum_{i=1}^n a_i a_i^T$.

Fact 5: In general (i.e. the a_i are linearly independent, but not necessarily orthonormal) the projection matrix is given by $P_S = P = A(A^T A)^{-1} A^T$.

Proof: First $S = \text{span}\{a_1, \ldots, a_n\} = col(A)$ and $b \in \mathbb{R}^m$. Then $Pb \in S$ implies that there exists a $x \in \mathbb{R}^n$ such that $Ax = Pb$. The Corollary to the Projection Theorem states that $b - Ax \in col(A)^\perp$. The Theorem on fundemental spaces tells us that $col(A)^\perp = N(A^T)$ and thus

$$A^T(b - Ax) = 0 \Rightarrow A^T Ax = A^T b$$

The linear independence of $\{a_1, \ldots, a_n\}$ implies that $rank(A) = n$, which by previous exercise means $A^T A$ is invertible, so $x = (A^T A)^{-1} A^T b$ and thus $Pb = Ax = A(A^T A)^{-1} A^T b$. $\square$

We follow up with some properties of projection matrices P:

1. P is symmetric and idempotent (what should happen to a vector if you project it and then project it again?).

   Proof: Exercise 4.1(a).

2. $I - P$ is the projection onto orthogonal complement of $col(A)$ (i.e. the left nullspace of A)

   Proof: Exercise 4.1(b).

3. Given any vector $b \in \mathbb{R}^m$ and any subspace S of $\mathbb{R}^m$, b can be written (uniquely) as the sum of its projections onto S and $S^\perp$

   Proof: Assume $dim(S) = q$, so $dim(S^\perp) = m - q$. Let $A_S = [a_1\, a_2\, ...a_q]$ and $A_{S^\perp} = [a_{q+1}\, ...\, a_m]$ be such that $a_1, ..., a_q$ form a basis for S and $a_{q+1}, ..., a_m$ form a basis for $S^\perp$. By 2, if P_S is the projection onto $col(A_S)$ and $P_{S^\perp}$ is the projection onto $col(A_{S^\perp})$, $\forall b \in \mathbb{R}^m$

$$P_S(b) + P_{S^\perp}(b) = P_S(b) + (I - P_S)b = b.$$

   As columns of A_S and $A_{S^\perp}$ are linearly independent, the vectors $a_1, a_2, ..., a_m$ form a basis of $\mathbb{R}^m$. Hence,

$$b = P_S(b) + P_{S^\perp}(b) = c_1a_1 + ... + c_qa_q + c_{q+1}a_{q+1} + ... + c_ma_m$$

   for unique $c_1, ..., c_m$. $\square$

4. $P(I - P) = (I - P)P = 0$ (what should happen to a vector when it's first projected to S and then $S^\perp$?)

   Proof: Exercise 4.1(c).

5. $col(P) = col(A)$

   Proof: Exercise 4.1(d).

6. Every symmetric and idempotent matrix P is a projection.

   Proof: All we need to show is that when we apply P to a vector b, the remaining part of b is orthogonal to $col(P)$, so P projects onto its column space. Well, $P^T(b - Pb) = P^T(I - P)b = P(I - P)b = (P - P^2)b = 0b = 0$. $\square$

7. Let a be a vector in $\mathbb{R}^m$. Then a projection matrix onto the line through a is $P = \frac{aa^T}{\|a\|^2}$, and if $a = q$ is a unit vector, then $P = qq^T$.

8. Combining the above result with the fact that we can always come up with an orthonormal basis for $\mathbb{R}^m$ (Gram-Schmidt) and with the fact about splitting vector into projections, we see that we can write $b \in \mathbb{R}^m$ as $q_1q_1^Tb + q_2q_2^Tb + \ldots + q_mq_m^Tb$ for some orthonormal basis $\{q_1, q_2, \ldots, q_m\}$.

9. If A is a matrix of rank r and P is the projection on $col(A)$, then $tr(P) = r$.

   Proof: Exercise 4.1(e).

4.2 Applications to Statistics: Least Squares Estimator

Suppose we have a linear model, where we model some response as

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ip}\beta_p + \epsilon_i,$$

where $x_{i1}, x_{i2}, \ldots, x_{ip}$ are the values of explanatory variables for observation i, ϵ_i is the error term for observaion i that has an expected value of 0, and $\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients we're interested in estimating. Suppose we have $n > p$ observations. Then writing the above system in matrix notation we have $Y = X\beta + \epsilon$, where X is the $n \times p$ matrix of explanatory variables, Y and ϵ are $n \times 1$ vectors of observations and errors respectively, and $p \times 1$ vector β is what we're interested in. We will furthermore assume that the columns of X are linearly independent.

Since we don't actually observe the values of the error terms, we can't determine the value of β and have to estimate it. One estimator of β that has some nice properties (which you will learn about) is **least squares estimator** (LSE) $\hat{\beta}$ that minimizes

$$\sum_{i=1}^{n} (y_i - \tilde{y}_i)^2,$$

where $\tilde{y}_i = \sum_{i=1}^{p} \tilde{\beta}_j x_{ij}$. This is equivalent to minimizing $\|Y - \tilde{Y}\|^2 = \|Y - X\tilde{\beta}\|^2$. It follows that the **fitted values** associated with the LSE satisfy

$$\hat{Y} = \arg \min_{\tilde{Y} \in col(X)} \|Y - \tilde{Y}\|^2$$

or that $\hat{Y}$ is the projection of Y onto $col(X)$. It follows then from Fact 4 that the the fitted values and LSE are given by

$$\hat{Y} = X(X^T X)^{-1} X^T Y = HY \text{ and } \hat{\beta} = (X^T X)^{-1} X^T Y.$$

The matrix $H = X(X^T X)^{-1} X^T$ is called the **hat matrix**. It is an orthogonal projection that maps the observed values to the fitted values. The vector of **residuals** $e = Y - \hat{Y} = (I - H)Y$ are orthogonal to $col(X)$ by the Corollary to the Projection Theorem, and in particular $e \perp \hat{Y}$.

Finally, suppose there's a column x_j in X that is perpendicular to all other columns. Then because of the results on the separation of projections (x_j is the orthogonal complement in $col(X)$ of the space spanned by the rest of the columns), we can project b onto the line spanned by x_j, then project b onto the space spanned by rest of the columns of X and add the two projections together to get the overall projected value. What that means is that if we throw away the column x_j, the values of the coefficients in β corresponding to other columns will not change. Thus inserting or deleting from X columns orthogonal to the rest of the column space has no effect on estimated coefficients in β corresponding to the rest of the columns.

Recall that the Projection Theorem and its Corollary are stated in the general setting of Hilbert spaces. One application of these results which uses this generality and arrises in STOR 635 and possibly 654 is the interpretation of conditional expectations as projections. Since this application requires a good deal of material covered in the first semester courses,

i.e measure theory and integration, an example of this type will not be given. Instead an example on a simpler class of functions will be given.

Example: Let $V = \mathcal{C}([-1,1])$ with $\|f\|^2 = \langle f, f\rangle = \int_{-1}^{1} f(x)f(x)dx$. Let $h(x) = 1$, $g(x) = x$ and $S = \text{span}\{h, g\} = \{\text{all linear functions}\}$. What we will be interested is calculating $P_S f$ where $f(x) = x^2$.

From the Corollary we know that $P_S f$ is the unique linear function that satisfies $\langle f - P_S f, s\rangle = 0$ for all linear functions $s \in S$. By previous (in class) exercise finding $P_S f$ requires finding constants a and b such that

$$\langle x^2 - (ax+b), 1\rangle = 0 = \langle x^2 - (ax+b), x\rangle.$$

First we solve

$$\begin{aligned} 0 = \langle x^2 - (ax+b), 1\rangle &= \int_{-1}^{1} (x^2 - ax - b)\cdot 1\, dx \\ &= \left(\frac{x^3}{3} - \frac{ax^2}{2} - bx\right)\Big|_{-1}^{1} \\ &= \left(\frac{1}{3} - \frac{a}{2} - b\right) - \left(\frac{-1}{3} - \frac{a}{2} + b\right) \\ &= \frac{2}{3} - 2b \Rightarrow b = \frac{1}{3}. \end{aligned}$$

Next,

$$\begin{aligned} 0 = \langle x^2 - (ax+b), x\rangle &= \int_{-1}^{1} (x^2 - ax - b)\cdot x\, dx \\ &= \int_{-1}^{1} x^3 - ax^2 - bx\, dx \\ &= \frac{x^4}{4} - \frac{ax^3}{3} - \frac{bx^2}{2}|_{-1}^{1} \\ &= \left(\frac{1}{4} - \frac{a}{3} - \frac{b}{2}\right) - \left(\frac{1}{4} + \frac{a}{3} - \frac{b}{2}\right) \\ &= \frac{-2a}{3} \Rightarrow a = 0. \end{aligned}$$

Therefore $P_S f = ax + b = \frac{1}{3}$ $\square$

Exercises

4.1 Prove the following properties of projection matrices:

(a) P is symmetric and idempotent.

(b) $I - P$ is the projection onto orthogonal complement of $col(A)$ (i.e. the left nullspace of A)

(c) $P(I - P) = (I - P)P = 0$

(d) $col(P) = col(A)$

(e) If A is a matrix of rank r and P is the projection on $col(A)$, $tr(P) = r$.

4.2 Show that the best least squares fit to a set of measurements $y_1, \cdots, y_m$ by a horizontal line — in other words, by a constant function $y = C$ — is their average

$$C = \frac{y_1 + \cdots + y_m}{m}$$

In statistical terms, the choice $\bar{y}$ that minimizes the error $E^2 = (y_1 - y)^2 + \cdots + (y_m - y)^2$ is the **mean** of the sample, and the resulting E^2 is the **variance** σ^2.

5 Linear functionals, riesz representation and hyperplane separation

5.1 Linear functionals

Linear functionals, dual space of a normed space The dual space of a real normed space V, denoted V^*, is the real vector space of **continuous linear functionals**, i.e. maps $\ell : V \mapsto \mathbb{R}$ that are linear and continuous in the topology given by the norm metric. In fact, it can be shown that

$$V^* = \{\ell : V \mapsto \mathbb{R} \,|\, \exists\, C \in (0, \infty) \text{ such that } \quad |\ell(x)| \leq C\|x\| \quad \forall x \in V\}$$

In other words, **every continuous linear functional on a normed space is Lipschitz**.

Dual spaces can be defined for arbitrary vector spaces endowed with a topology in which linear operations are continuous. We focus on normed spaces here, and really just on inner product spaces. Often one will consider complex-valued linear functionals, in which case V^* is a vector space over the complex numbers, but we stick to the real case here. All of the results presented here still hold in the complex case.

Example: Suppose that $V = \mathcal{C}([-1, 1] : \mathbb{R})$, the space of real-valued continuous functions defined on $[-1, 1]$. Then the linear transformation $\phi : V \to \mathbb{R}$ given by

$$\phi(f) = \int_{-1}^{1} f(x)dx,$$

is a continuous linear functional. In fact, for any bounded continuous function g

$$\phi_g(f) = \int_{-1}^{1} f(x)g(x)dx,$$

is a continuous linear functional.

The following theorems show that any linear functional on a Hilbert space can be expressed in terms of the inner product. We prove the finite-dimensional case and leave the infinite-dimensional one as an exercise.

The Riesz Representation Theorem (finite dimension): Let V be a finite dimensional inner product space and let $\phi \in V^*$. Then there is a unique $y \in V$ such that

$$\phi(x) = \langle y, x \rangle,$$

for each $x \in V$. Additionally, if $v_1, v_2, \ldots, v_n$ is an orthonormal basis of V, then

$$y = \sum_{i=1}^{n} \phi(v_i)v_i.$$

We refer to y as the **Riesz vector** for ϕ.

Proof. Let $x \in V$ and note that there are scalars $r_1, r_2, \ldots, r_n$ such that

$$x = \sum_{i=1}^{n} r_i v_i.$$

Using the basic properties of orthornormal bases and inner products we can see that

$$r_i = \sum_{j=1}^{n} r_j \langle v_i, v_j \rangle = \left\langle v_i, \sum_{i=1}^{n} r_j v_j \right\rangle = \langle v_i, x \rangle,$$

so

$$x = \sum_{i=1}^{n} \langle v_i, x \rangle v_i.$$

Thus if we define $y = \sum_{i=1}^{n} \phi(v_i) v_i$,

$$\phi(x) = \phi\left(\sum_{i=1}^{n} \langle v_i, x \rangle v_i \right) = \sum_{i=1}^{n} \langle v_i, x \rangle \phi(v_i) = \left\langle \sum_{i=1}^{n} \phi(v_i) v_i, x \right\rangle = \langle y, x \rangle.$$

Recall that the map $(\lambda_1 \ldots \lambda_n) \mapsto x = \sum_1^n \lambda_i x_i$ is one-to-one and onto, i.e. x is uniquely determined by the coefficients in the basis representation, and every $x \in V$ can be written this way for some vector of coefficients. Uniqueness of y follows.

□

Theorem: Dual basis Say V is an inner product space over $\mathbb{R}$ of dimension n. If $\{v_i\}_1^n$ is an orthonormal basis for V,

- There exists a unique set $\ell_1 \ldots \ell_n \in V^*$ such that $\ell_i(v_j) = \delta_{ij}$, where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise.
- $\ell_1, \ldots \ell_n$ is a basis for V^*, so that in particular V and V^* have the same dimension.

Proof. Define $\ell_i(x) = \langle v_i, x \rangle$ for all x. These maps are continuous linear functionals by the Cauchy-Schwartz inequality. The first statement is then a consequence of orthonormality.

We must check that $\ell_i, i = 1 \ldots n$ are linearly independent and span V^*. That each $\phi \in V^*$ is a linear combination of $\{\ell_i\}$ was shown in the previous proof. To check linear independence we must show that

$$\sum_{1}^{n} \alpha_i \ell_i(x) = 0 \quad \forall\, x \quad \Longrightarrow \quad \alpha_i = 0,\ i = 1 \ldots n$$

For such α_i, by definition of ℓ_i we have $\sum_1^n \alpha_i \ell_i(x) = \langle \sum_1^n \alpha_i v_i,\ x \rangle = 0$ for all x. This implies $\sum_1^n \alpha_i v_i = 0$, which shows $\alpha_i = 0$ for all i since $\{v_i\}$ is a basis. □

In the example above V is not finite dimensional, so the theorem above does not apply. However, there is an analogous result for infinite dimensional inner product spaces, proven as an exercise below.

5.2 Hyperplane separation

The following results are stated for finite-dimensional vector spaces over the reals. However, they have analogs in very general linear spaces. See Lax, Functional Analysis, for example.

Theorem (Hyperplane separation V1) Suppose $U, V \subset \mathbb{R}^n$ are non-empty convex sets, at least one of which is open. There exists a non-zero linear functional ℓ and $c \in \mathbb{R}$ such that

$$\ell(x) \leq c \leq \ell(y) \quad \forall\, x \in U,\, y \in V$$

Theorem (Hyperplane separation V2) Suppose $U \subset \mathbb{R}^n$ is a non-empty closed convex set. If $x_0 \notin U$, there exists a non-zero linear functional ℓ and $c_1, c_0 \in \mathbb{R}$ such that

$$\ell(x) < c_1 < c_0 < \ell(x_0) \quad \forall\, x \in U$$

By the Riesz representation theorem, we could equivalently have written the results above in terms of $\ell(x) = \langle v, x \rangle$ for some vector v. There are several other versions of hyperplane separation theorems, including analogous statements for very general vector spaces. See Lax.

For hyperplanes in $\mathbb{R}^n$, we note the following facts about the affine space given by the hyperplane associated with a vector v and constant c, denoted here $H = \{x \mid \langle v, x \rangle = c\}$

- H is a convex set, and vectors along the surface of H are given by $x_1 - x_0$ for $x_1, x_0 \in H$. If $c = 0$, H is a subspace.
- $v/||v||$ is the unit normal to H, that is if $x_0, x_1 \in H$ then $\langle x_1 - x_0, v \rangle = 0$
- To determine the shortest distance $d(x, H)$ from any point x to H, we take an arbitrary point $x_0 \in H$ and compute the length of the projection of $x - x_0$ onto v, which is perpendicular to H by the previous statement.

  Recall the projection of $x - x_0$ onto v is $\langle x - x_0, v \rangle\, v/||v||^2$, and therefore

  $$d(x, H) = ||\langle x - x_0, v \rangle\, v/||v||^2|| = \frac{|\langle x, v \rangle - c\,|}{||v||}$$

  Notice this does not depend on the point x_0 chosen.

Example: Separating hyperplanes for classification This example comes from Hastie, chapter 4.

Suppose you have k data vectors $x_i \in \mathbb{R}^d$ and associated two-class labels $y_i \in \{-1, 1\}$. If the data for vectors in each class can be separated by a hyperplane, we may predict the class label for a new observation x_* based on the region it lies in relative to the hyperplane.

Suppose $I \subset \{1 \dots k\}$ is the index set for data with class label 1. We consider the closed convex hulls of the points $x_i, i \in I$ and $x_j, j \in I^c$. These sets are compact. If they are disjoint, by adapting the theorems above we can show there exists a hyperplane separating them. In fact, there could be infinitely many.

Which separating hyperplane do we chose for our classification model? This is the question resolved by **maximal margin (or optimal) separating hyperplane** problem. Specifically, we look for a solution to

$$\max_{v,c,\, ||v||=1} M \qquad \text{such that} \quad y_i(\langle x_i, v\rangle - c) \geq M \quad \forall i$$

As seen above, if v is normalized to length one, $\langle x_i, v\rangle - c$ gives the **signed** distance of x_i from the hyperplane given by $\langle x, v\rangle = c$. Thus the problem above seeks a hyperplane such that data in class 1 are on one side of the plane, data in class -1 are on the other side of the hyperplane, such that the distance from the data to the plane is maximized. This is a convex optimization problem solved via Lagrange multipliers.

If, however, the data cannot be separated by a hyperplane, we may use a reformulation that allows some points to be misclassified up to a level of error. This method is called a linear **Support Vector Machine**. See Hastie for details.

Exercises

5.1 *Garcia P.5.16* Let $V = \{p \in \mathcal{C}([0,1] : \mathbb{R}) : p(x) = ax^3 + bx^2 + cx + d$ for some $a, b, c, d \in \mathbb{R}\}$ be the space of polynomials of degree three or less defined on $[0, 1]$. Recall the inner product on V given by

$$\langle p, q\rangle = \int_0^1 p(x)q(x)dx$$

for each $p, q \in V$.

1. Consider the map $\phi : V \to \mathbb{R}$ given by

$$\phi(p) = \int_0^1 \sqrt{x}p(x)dx$$

   (a) Show that ϕ is a continuous linear functional.

   (b) Find the Riesz vector for ϕ. You do not need to find the coefficients of the Riesz vector (polynomial) explicitly. See the hint.

   Hint: Find a basis for V, then note you need only show the identity in the riesz representation theorem for basis elements. Set up a linear system of equations and show it **can be** solved, for example using criteria for invertibility given in the chapters below. You do not need to actually solve it, however.

5.2 In this question, you will prove the Riesz representation for real Hilbert spaces V.

We may assume the linear functional $\phi \neq 0$, since otherwise we have $\phi(x) = \langle 0, x \rangle$ for all $x \in V$.

1. Let ϕ be a continuous linear functional on V, and write $Y = \{x \in V \mid \phi(x) = 0\}$ for its nullspace.

   Show Y is a subspace of V that is closed, i.e. such that for $x_n \in Y$ with $\|x_n - x\| \to 0$ for some $x \in V$, then $x \in Y$.

2. Show its orthogonal complement $Y^\perp$ is one-dimensional. (Hint: Use the previous question and the orthogonal decomposition theorem to show there exists a vector z such that $\phi(z) \neq 0$. Use the linearity of ϕ to find an explicit decomposition for any $x \in V$ of the form $x = z_1 + z_2$ where $z_1 \in Y$. Use the uniqueness part of the decomposition theorem to show $z_2 \in Y^\perp$ and $z_2 \in span(z)$.

3. Take $z, z_2 \in Y^\perp$ as in the previous question. Find a $c \in \mathbb{R}$ such that $< cz, z >= \ell(z)$. Argue that this, and the previos question, give the result.

5.3 Let A be an $m \times n$ real matrix and define $K = \{y \in \mathbb{R}^m \mid Ax = y, \quad x \geq 0\}$ where the statement $x \geq 0$ for $x \in \mathbb{R}^n$ is evaluated component-wise.

1. Show K is a closed, convex set.

2. Show that **exactly one** of the following two statements holds : For fixed $y \in \mathbb{R}^m$ non-zero,

   (i) There exists a solution $x \geq 0$ to the equation $Ax = y$

   (ii) There exists a $v \in \mathbb{R}^m$ such that $v^T A \leq 0$ and $\langle v, y \rangle > 0$.

   Hint: Use the hyperplane separation theorem V2.

6 Matrix Decompositions

We will assume that you are familiar with LU and QR matrix decompositions. If you are not, you should look them up, they are easy to master. We will in this section restrict ourselves to eigenvalue-preserving decompositions.

6.1 Determinants

Often in mathematics it is useful to summarize a multivariate phenomenon with a single number, and the determinant is an example of this. It is only defined for square matrices. We will assume that you are familiar with the idea of determinants, and specifically calculating determinants by the method of cofactor expansion along a row or a column of a square matrix. Below we list the properties of determinants of real square matrices. The first 3 properties are defining, and the rest are established from those three properties. In fact, the operation on square matrices which satisfies the first 3 properties must be a determinant.

1. $\det(A)$ depends linearly on the first row.

$$\det\begin{bmatrix} ra_{11}+sa'_{11} & ra_{12}+sa'_{12} & \dots & ra_{1n}+sa'_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} =$$

$$\text{r det}\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} + \text{s det}\begin{bmatrix} a'_{11} & a'_{12} & \dots & a'_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}.$$

2. Determinant changes sign when two rows are exchanged. This also implies that the determinant depends linearly on EVERY row, since we can exhange row$i$ with row 1, split the determinant, and exchange the rows back, restoring the original sign.

3. $\det(I) = 1$

4. If two rows of A are equal, $\det(A) = 0$ (why?)

5. Subtracting a multiple of one row from another leaves determinant unchanged.

   Proof: Suppose $A = \left[a'_1, \cdots, a'_i, \cdots, a'_j, \cdots, a'_n\right]', \tilde{A} = \left[a'_1, \cdots, a'_i - ra'_j, \cdots, a'_j, \cdots, a'_n\right]^T$. Then,

$$\begin{aligned} \det(\tilde{A}) &= \det(\left[a'_1, \cdots, a'_i, \cdots, a'_j, \cdots, a'_n\right]') - r\det\left[a'_1, \cdots, a'_j, \cdots, a'_j, \cdots, a'_n\right]^T \\ &= \det(A) + 0 = \det(A) \end{aligned}$$

   □

6. If a matrix has a zero row, its determinant is 0. (why?)

7. If a matrix is triangular, its determinant is the product of entries on main diagonal

   Proof: Exercise 6.1.

8. $\det(A) = 0$ if and only if A is not invertible (proof involves ideas of elimination)

9. $\det(AB) = \det(A)\det(B)$. In particular, if A is inversible, $\det(A^{-1}) = \frac{1}{det(A)}$.

   Proof: Suppose $\det(B) = 0$. Then B is not invertible, and AB is not invertible (recall $(AB)^{-1} = B^{-1}A^{-1}$, therefore $\det(AB) = 0$. If $\det(B) \neq 0$, let $d(A) = \frac{det(AB)}{det(B)}$. Then,

   (1) For $A_* = [a_{11}^*, a_{12}^*, \cdots, a_{1n}^*] \in \mathbb{R}^n$, let A_i be the i^{th} row of A, $r \in \mathbb{R}$, and A^* be the matrix A but with its first row replaced with A_*. Then,

$$\begin{aligned} d\left(\begin{bmatrix} rA_1 + A_* \\ \vdots \\ A_n \end{bmatrix}\right) &= det\left(\begin{bmatrix} rA_1 + A_* \\ \vdots \\ A_n \end{bmatrix} B\right)(det(B))^{-1} \\ &= \frac{\det\left(\begin{bmatrix} (rA_1 + A_*)B \\ \vdots \\ A_nB \end{bmatrix}\right)}{det(B)} \\ &= \frac{det\left(\begin{bmatrix} rA_1B \\ \vdots \\ A_nB \end{bmatrix}\right) + det\left(\begin{bmatrix} A_*B \\ \vdots \\ A_nB \end{bmatrix}\right)}{det(B)} \\ &= \frac{r \cdot det\,(AB) + det\,(A^*B)}{det(B)} \\ &= r \cdot d(A) + d(A^*). \end{aligned}$$

   Using the same argument for rows $2, 3, \ldots, n$ we see that $d(\cdot)$ is linear for each row.

   (2) Let $A^{i,j}$ be the matrix A with rows i and j interchanged, and WLOG assume $i < j$. Then

$$d(A^{i,j}) = \frac{det\left(\begin{bmatrix} A_1 \\ \vdots \\ A_j \\ \vdots \\ A_i \\ \vdots \\ A_n \end{bmatrix} B\right)}{det(B)} = \frac{det\left(\begin{bmatrix} A_1B \\ \vdots \\ A_jB \\ \vdots \\ A_iB \\ \vdots \\ A_nB \end{bmatrix}\right)}{det(B)} = \frac{det\,((AB)^{i,j})}{det(B)} = \frac{-det(AB)}{det(B)} = -d(A).$$

(3) $d(I) = det(IB)/det(B) = det(B)/det(B) = 1$.

So conditions 1-3 are satisfied and therefore $d(A) = det(A)$. $\square$

10. $\det(A^T) = \det(A)$. This is true since expanding along the row of A^T is the same as expanding along the corresponding column of A.

6.2 Eigenvalues and Eigenvectors

Eigenvalues and Eigenvectors Given a square $n \times n$ matrix A, we say that λ is an **eigenvalue** of A, if for some non-zero $x \in \mathbb{R}^n$ we have $Ax = \lambda x$. We then say that x is an **eigenvector** of A, with corresponding eigenvalue $\lambda \in \mathbb{C}$. Note that even of the entries of A are real, λ may be complex.

Eigenvalue/vectors methods show up all over the place in statistics e.g. see Eigenproblems in Pattern Recognition (http://www.cs.utah.edu/~piyush/teaching/eig_book04.pdf).

Note that

$$Ax = \lambda x \iff (\lambda I - A)x = 0 \iff \lambda I - A \text{ is not invertible} \iff \det(\lambda I - A) = 0$$

The upshot is: if we write $p(\lambda) = \det(\lambda I - A)$ then p is a polynomial whose roots are the eigenvalues of A. Every polynomial has at least one distinct (possibly complex) root therefore, every matrix has at least one eigenvalue. This polynomial characterization of eigenvalues can also be used to show the following theorem.

Theorem:

1. The determinant is the product of the eigenvalues i.e. $\det(A) = \prod_{i=1}^{n} \lambda_i$
2. The trace is the sum of the eigenvalues i.e. $\text{tr}(A) = \sum_{i=1}^{n} \lambda_i$
3. $\det(\lambda I - A) = 0$ if and only if λ is an eigenvalue of A.

Proof. The explanation for 3. is above. Some details will be excluded, but the basic argument is as follows:

1. Using the cofactor expansion method of calculating $p(t) = \det(tI - A)$, note the following:
   (a) $p(t)$ is a degree n polynomial, and the coefficient of t^n is 1.
   (b) The coefficient of t^{n-1} in $p(t)$ is $-\sum_{i=1}^{n} a_{ii}$.
   (c) Use (a) to see that $p(0) = r_0$, and thus that $p(0) = \det(0I - A) = (-1)^n \det(A)$.

2. Observe that $p(t)$ can be factored as $p(t) = \prod_{i=1}^{n}(t-\lambda_i)$, where each λ_i is a (not necessarily distinct) eigenvalue of A.

3. Expand the expression in 2. and use 1. (c) to show that $\mathrm{tr}(A) = \sum_{i=1}^{n} \lambda_i$ and $\det(A) = \prod_{i=1}^{n} \lambda_i$.

□

□

Let's consider how to compute the eigenvalues/vectors of a matrix. First compute the roots of p defined above to find the eigenvalues. Given an eigenvalue λ we can find an eigenvector v by finding a basis for the kernel of $A - \lambda I$. In other words, computing the eigenvalues/vectors of A requires two subroutines: finding the roots of a polynomial and finding the basis for a kernel. The latter can be solved using row reduction (see standard linear algebra references). The former, finding the roots of an arbitrary polynomial, is harder. We can find the roots of a polynomial easily in special cases (e.g. if $n = 2$ then use the quadratic formula) however the general case is not so nice. A result from Galois theory (Abel-Ruffini theorem) says finding the exact roots of an arbitrary polynomial of degree ≥ 5 is essentially impossible.

Theorem 25.1 *from Numerical linear algebra, Trefethen: For any $m \geq 5$, there is a polynomial $p(z)$ of degree m with rational coefficients that has a real root $p(r) = 0$ with the property that r cannot be written using any expression involving rational numbers, addition, subtraction, multiplication, division and kth roots.*

This theorem implies "there could be no computer program that would produce the exact roots of an arbitrary polynomial in a finite number of steps." The same must therefore hold for eigenvalues since one can show computing eigenvalues is equivalent to finding roots of a polynomial. The upshot is that **any eigenvalue solver must be iterative and approximate.** Computing eigenvalues/vectors is an important and well studied problem e.g. see Numerical linear algebra, by Trefethen for (many) algorithms which solve this problem.

You should be able to see that the eigenvalues of A and A^T are the same (why? Do the eigenvectors have to be the same?), and that if x is an eigenvector of A ($Ax = \lambda x$), then so is every multiple rx of x, with same eigenvalue ($Arx = \lambda rx$). In particular, a unit vector in the direction of x is an eigenvector.

If A is an $n \times n$ matrix, then its **spectrum** is the set

$$\sigma_A = \{\lambda : \lambda \text{ is an eigenvalue of } A\}.$$

We are often interested in determining exactly what the spectrum of a particular matrix is. However, sometimes it is enough to find a region containing σ_A. The following result,

which is stated for matrices with complex entries, provides the means to find a region containing σ_A.

Theorem: Let $A \in \mathbb{C}^{n\times n}$. Then

$$\sigma_A \subseteq \bigcup_{i=1}^{n} \left\{ z \in \mathbb{C} : |z - a_{ii}| \leq \sum_{j\neq i} |a_{ij}| \right\}.$$

Proof. Let $\lambda \in \sigma_A$ be an eigenvalue with corresponding eigenvector x. Define another eigenvector, y, corresponding to λ such that $y_i = 1$ and $\max_{j\neq i} |y_j| \leq 1$. Since $Ay = \lambda y$,

$$\lambda = \sum_{j=1}^{n} a_{ij}y_j = \sum_{j\neq i}^{n} a_{ij}y_j + a_{ii}y_i = \sum_{j\neq i}^{n} a_{ij}y_j + a_{ii},$$

and

$$|\lambda - a_{ii}| = \left| \sum_{j\neq i} a_{ij}y_j \right| \leq \sum_{j\neq i} |a_{ij}|.$$

Thus $\lambda \in \left\{ z \in \mathbb{C} : |z - a_{ii}| \leq \sum_{j\neq i} |a_{ij}| \right\}$.

□

□

Recall that if A is an $n \times n$ matrix, then it represents a linear transformation from between two finite dimensional vector spaces, and that A has at least one eigenvalue. The following example shows that the same does not hold if the spaces are infinite dimensional - in particular, linear transformations on infinite dimensional vector spaces need not have any eigenvalues.

Example: Let $\phi : \mathcal{C}([0,1] : \mathbb{R}) \to \mathcal{C}([0,1] : \mathbb{R})$ be given by

$$\phi(f)(x) = \int_0^x f(y)dy.$$

Theorem : Eigenvectors corresponding to distinct eigenvalues are linearly independent.

Proof: Suppose that there are only two distinct eigenvalues (A could be 2×2 or it could have repeated eigenvalues), and let $r_1x_1 + r_2x_2 = 0$. Applying A to both sides we have $r_1Ax_1 + r_2Ax_2 = A0 = 0 \implies \lambda_1r_1x_1 + \lambda_2r_2x_2 = 0$. Multiplying first equation by λ_1 and subtracting it from the second, we get $\lambda_1r_1x_1 + \lambda_2r_2x_2 - (\lambda_1r_1x_1 + \lambda_1r_2x_2) = 0 - 0 = 0 \implies r_2(\lambda_2 - \lambda_1)x_2 = 0$ and since $x_2 \neq 0$, and $\lambda_1 \neq \lambda_2$, we conclude that $r_2 = 0$. Similarly, $r_1 = 0$ as well, and we conclude that x_1 and x_2 are in fact linearly independent. The proof extends to more than 2 eigenvalues by induction. □

Diagonalizable We say a matrix $A \in \mathbb{R}^{n\times n}$ is diagonalizable if there exists some $S \in \mathbb{R}^{n\times n}$ such that $S^{-1}AS = D$ where D is a diagonal matrix (i.e. if A is similar to a diagonal matrix). By the proof above, every matrix that has n DISTINCT eigenvalues is diagonalizable by the proof above. Let S be the matrix whose columns are the eigenvectors, then $AS = \Lambda S$ where Λ is the matrix with the eigenvalues on the diagonal. By the proof above, S is invertible (why?) so we get that A is diagonalizable. Note that some matrices that fail to have n distinct eigenvalues may still be diagonalizable, as we'll see in a moment.

Now suppose that we have $n \times n$ A and for some S, we have $S^{-1}AS = \Lambda$, a diagonal matrix. Then you can easily see for yourself that the columns of S are eigenvectors of A and diagonal entries of Λ are corresponding eigenvalues. So the matrices that can be made into a diagonal matrix by pre-multiplying by S^{-1} and post-multiplying by S for some invertible S are precisely those that have n linearly independent eigenvectors (which are, of course, the columns of S). Clearly, I is diagonalizable ($S^{-1}IS = I$) $\forall$ invertible S, but I only has a single eigenvalue 1. So we have an example of a matrix that has a repeated eigenvalue but nonetheless has n independent eigenvectors.

If A is diagonalizable, calculation of powers of A becomes very easy, since we can see that $A^k = S\Lambda^k S^{-1}$, and taking powers of a diagonal matrix is about as easy as it can get. This is often a very helpful identity when solving recurrent relationships.

Example A classical example is the Fibonacci sequence 1, 1, 2, 3, 5, 8, ..., where each term (starting with 3rd one) is the sum of the preceding two: $F_{n+2} = F_n + F_{n+1}$. We want to find an explicit formula for n-th Fibonacci number, so we start by writing

$$\begin{bmatrix} F_{n+1} \\ F_n \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} F_n \\ F_{n-1} \end{bmatrix}$$

or $u_n = Au_{n-1}$, which becomes $u_n = A^n u_0$, where $A = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$, and $u_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. Diagonalizing A we find that $S = \begin{bmatrix} \frac{1+\sqrt{5}}{2} & \frac{1-\sqrt{5}}{2} \\ 1 & 1 \end{bmatrix}$ and $\Lambda = \begin{bmatrix} \frac{1+\sqrt{5}}{2} & 0 \\ 0 & \frac{1-\sqrt{5}}{2} \end{bmatrix}$, and identifying F_n with the second component of $u_n = A^n u_0 = S\Lambda^n S^{-1} u_0$, we obtain $F_n = \frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^n - \left(\frac{1-\sqrt{5}}{2}\right)^n\right]$
We finally note that there's no relationship between being diagonalizable and being invertible. $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ is both invertible and diagonalizable, $\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$ is diagonalizable (it's already diagonal) but not invertible, $\begin{bmatrix} 3 & 1 \\ 0 & 3 \end{bmatrix}$ is invertible but not diagonalizable (check this!), and $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ is neither invertible nor diagonalizable (check this too).

6.3 Complex Matrices

Complex Matrix We now allow complex entries in vectors and matrices. Scalar multiplicaiton now also allows multiplication by complex numbers, so we're going to be dealing with vectors in $\mathbb{C}^n$, and you should check for yourself that $\dim(\mathbb{C}^n) = \dim(\mathbb{R}^n) = n$ (Is $\mathbb{R}^n$ a subspace of $\mathbb{C}^n$?)[2] We also note that we need to tweak a bit the earlier definition of transpose to account for the fact that if $x = [1, i]^T \in \mathbb{C}^2$, then

$$x^T x = 1 + i^2 = 0 \neq 2 = \|x\|^2.$$

Recall that if $z = a + bi \in \mathbb{C}$ then the *conjugate* of x, is $\bar{z} = a - bi$. The *modulus* of $z \in \mathbb{C}$ is given by $\sqrt{z \cdot \bar{z}} = \sqrt{a^2 + b^2}$.

Returning to vectors/matrices we definite the *conjugate-transpose* or *Hermetian-transpose*. If $M \in \mathbb{C}^{n \times d}$ then the transpose-conjugate, $M^H \in \mathbb{C}^{d \times n}$, is given by taking the transpose of M then taking the conjugate of each entry i.e. $M^H_{ij} = \overline{M}_{ji}$. We now have $x^H x = \|x\|^2$. If $x \in \mathbb{R}^n$, then $x^H = x^T$.

You should check that $(A^H)^H = A$ and that $(AB)^H = B^H A^H$ (you might want to use the fact that for complex numbers $x, y \in \mathbb{C}$, $\overline{x + y} = \bar{x} + \bar{y}$ and $\overline{xy} = \bar{x}\bar{y}$). We say that x and y in $\mathbb{C}^n$ are orthogonal if $x^H y = 0$ (note that this implies that $y^H x = 0$, although it is NOT true in general that $x^H y = y^H x$).

Special Matrices We generalize some of the special matrices for $\mathbb{R}^{n \times n}$ to $\mathbb{C}^{n \times n}$.

1. $A \in \mathbb{C}^{n \times n}$ is *Hermitian* if $A = A^H$. You should check that every real, symmetric real matrix is Hermitian.
2. $A \in \mathbb{C}^{n \times n}$ is *unitary* if $A^H A = A A^H = I$ i.e. $(A^H = A^{-1})$. You should check that real, orthogonal matrix is unitary.
3. $A \in \mathbb{C}^{n \times n}$ is *normal* if it commutes with its Hermitian transpose: $A^H A = A A^H$.You should check that Hermitian and unitary matrices are normal.
4. $A \in \mathbb{C}^{n \times n}$ is *positive definite* if for all $x \in \mathbb{C}^n$, $x^H A x > 0$.

6.4 Facts that lead up to the spectral theorem

We next present some very important results about Hermitian and unitary matrices. These facts will then be used to prove the spectral theorem.

1. If A is Hermitian, then $\forall x \in \mathbb{C}^n$, $y = x^H A x \in \mathbb{R}$.

   Proof: taking the hermitian transpose we have $y^H = x^H A^H x = x^H A x = y$, and the only scalars in $\mathbb{C}$ that are equal to their own conjugates are the reals. $\square$

[2] We refer to a one dimensional real subspace as a line. Why should we call a one dimensional subspace of $\mathbb{C}^n$ a complex plane?

2. If A is Hermitian, and λ is an eigenvalue of A, then $\lambda \in \mathbb{R}$. In particular, all eigenvalues of a symmetric real matrix are real (and so are the eigenvectors, since they are found by elimination on $A - \lambda I$, a real matrix).

   Proof: suppose $Ax = \lambda x$ for some nonzero x, then pre-multiplying both sides by x^H, we get $x^H Ax = x^H \lambda x = \lambda x^H x = \lambda \|x\|^2$, and since the left-hand side is real, and $\|x\|^2$ is real and positive, we conclude that $\lambda \in \mathbb{R}$. $\square$

3. If A is positive definite, and λ is an eigenvalue of A, then $\lambda > 0$.

   Proof: Let nonzero x be the eigenvector corresponding to λ. Then since A is positive definite, we have $x^H Ax > 0 \implies x^H(\lambda x) > 0 \implies \lambda \|x\|^2 > 0 \implies \lambda > 0$. $\square$

4. If A is Hermitian, and x, y are the eigenvectors of A, corresponding to different eigenvalues ($Ax = \lambda_1 x, Ay = \lambda_2 y$), then $x^H y = 0$.

   Proof: $\lambda_1 x^H y = (\lambda_1 x)^H y$ (since λ_1 is real) $= (Ax)^H y = x^H (A^H y) = x^H (Ay) = x^H(\lambda_2 y) = \lambda_2 x^H y$, and get $(\lambda_1 - \lambda_2) x^H y = 0$. Since $\lambda_1 \neq \lambda_2$, we conclude that $x^H y = 0$. $\square$

5. The above result means that if a real symmetric $n \times n$ matrix A has n distinct eigenvalues, then the eigenvectors of A are mutually orthogonal, and if we restrict ourselves to unit eigenvectors, we can decompose A as $Q \Lambda Q^{-1}$, where Q is orthogonal (why?), and therefore $A = Q \Lambda Q^T$. We will later present the result that shows that it is true of EVERY symmetric matrix A (whether or not it has n distinct eigenvalues).

6. Unitary matrices preserve inner products and lengths.

   Proof: Let U be unitary. Then $(Ux)^H (Uy) = x^H U^H U y = x^H I y = x^H y$. In particular $\|Ux\| = \|x\|$. $\square$

7. Let U be unitary, and let λ be an eigenvalue of U. Then $|\lambda| = 1$ (Note that λ could be complex, for example i, or $\frac{1+i}{\sqrt{2}}$).

   Proof: Suppose $Ux = \lambda x$ for some nonzero x. Then $\|x\| = \|Ux\| = \|\lambda x\| = |\lambda| \|x\|$, and since $\|x\| > 0$, we have $|\lambda| = 1$. $\square$

8. Let U be unitary, and let x, y be eigenvectors of U, corresponding to different eigenvalues ($Ux = \lambda_1 x, Uy = \lambda_2 y$). Then $x^H y = 0$.

   Proof: $x^H y = x^H I y = x^H U^H U y = (Ux)^H (Uy) = (\lambda_1 x)^H (\lambda_2 y) = \lambda_1^H \lambda_2 x^H y = \bar{\lambda_1} \lambda_2 x^H y$ (since λ_1 is a scalar). Suppose now that $x^H y \neq 0$, then $\bar{\lambda_1} \lambda_2 = 1$. But $|\lambda_1| = 1 \implies \bar{\lambda_1} \lambda_1 = 1$, and we conclude that $\lambda_1 = \lambda_2$, a contradiction. Therefore, $x^H y = 0$. $\square$

9. A *schur factorization* of a square matrix A is $A = QTQ^H$ where Q is unitary and T is upper triangular. Every square matrix has a Schur factorization.

   Proof: (borrowed from Numerical Linear Algebra, by Trefethen) We prove the result by induction. The base case $n = 1$ is trivial. Assume the result is true for $n - 1 \geq 2$. Let x be an eigenvector of A with corresponding eigenvalue lambda. Assume x is

normalized. Let U be a unitary matrix whose first column is x (Why can we construct such a unitary matrix? *Hint*: Gram-Schmit). We can check that

$$U^H A U = \begin{bmatrix} \lambda & B \\ 0 & C \end{bmatrix}$$

where $B \in \mathbb{C}^{1\times n-1}$ and $C \in \mathbb{C}^{n-1\times n-1}$. By the inductive hypothesis the matrix C has a Schur factorization, $C = VTV^H$ i.e. V is unitary and T is upper triangular. Now let the matrix Q be given by

$$Q := U \begin{bmatrix} 1 & 0 \\ 0 & V \end{bmatrix}.$$

We can now check that Q is a unitary matrix and

$$Q^H A Q = \begin{bmatrix} \lambda & BV \\ 0 & T \end{bmatrix}.$$

is a Schur factorization of A. $\square$.

10. If A is normal, and U is unitary, then $B = U^{-1}AU$ is normal.

    Proof: $BB^H = (U^HAU)(U^HAU)^H = U^HAUU^HA^HU = U^HAA^HU = U^HA^HAU$ (since A is normal) $= U^HA^HUU^HAU = (U^HAU)^H(U^HAU) = B^HB$. $\square$

11. If $n \times n$ A is normal, then $\forall x \in \mathbb{C}^n$ we have $\|Ax\| = \|A^Hx\|$.

    Proof: $\|Ax\|^2 = (Ax)^HAx = x^HA^HAx = x^HAA^Hx = (A^Hx)^H(A^Hx) = \|A^Hx\|^2$. And since $\|Ax\| \geq 0 \leq \|A^Hx\|$, we have $\|Ax\| = \|A^Hx\|$. $\square$

12. If A is normal and A is upper triangular, then A is diagonal.

    Proof: Consider the first row of A. In the preceding result, let $x = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$. Then $\|Ax\|^2 = |a_{11}|^2$(since the only non-zero entry in first column of A is a_{11}) and $\|A^Hx\|^2 = |a_{11}|^2 + |a_{12}|^2 + \ldots + |a_{1n}|^2$. It follows immediately from the preceding result that $a_{12} = a_{13} = \ldots = a_{1n} = 0$, and the only non-zero entry in the first row of A is a_{11}. You can easily supply the proof that the only non-zero entry in the i-th row of A is a_{ii} and we conclude that A is diagonal. $\square$

6.5 Spectral Theorem

First we state the typical version of the spectral theorem.

Spectral Theorem: If $A \in \mathbb{C}^{n\times n}$ is Hermitian then there we can write $A = U\Lambda U^H$ where $U \in \mathbb{C}^{n\times n}$ is unitary and $\Lambda \in \mathbb{R}^{n\times n}$ is diagonal with real entries. Equivalently, every Hermitian matrix has n real eigenvalues and with n linearly independent eigenvectors.

proof: Let $A = UTU^H$ where U is unitary and T is upper triangular be the Schur decomposition of A (exists for every matrix by fact 9). Since A is Hermitian it is also normal; thus fact 10 shows T is normal. Since T is normal and upper triangular is is in fact diagonal (by fact 10).

The Schur decomposition $A = UTU^T$ is thus an eigenvalue decomposition i.e. $AU = UT$ where T is the diagonal matrix of the eigenvalues. Finally, fact 2 says the eigenvalues of A are real thus T is real. $\square$.

Next we state a special case of the spectral theorem for real, symmetric matrices (this is the version that we usually use in statistics).

Spectral Theorem (real case): If $A \in \mathbb{R}^{n\times n}$ is Symmetric then there we can write $A = Q\Lambda Q^T$ where $Q \in \mathbb{R}^{n\times n}$ is orthogonal and $\Lambda \in \mathbb{R}^{n\times n}$ is diagonal with real entries. Equivalently, every real symmetric matrix has n real eigenvalues and with n linearly independent eigenvectors. The real case follows from the above Spectral theorem plus the second part of fact 2.

A natural question to ask is: what is the set of matrices which the spectral theorem applies to? In other words, can we characterize the set of matrices A that are *unitary diagonalizable* i.e. there exists a unitary U and (possibly complex) diagonal Λ such that $A = U\Lambda U^H$? It turns out the answer is exactly the class of normal matrices.

Spectral Theorem (general): $A \in \mathbb{C}^{n\times n}$ is normal if and only if there exists unitary $U \in \mathbb{C}^{n\times n}$ and diagonal $\Lambda \in \mathbb{C}^{n\times n}$ such that $A = U\Lambda U^H$. Equivalently, A is normal if and only if it as an orthonormal set of (possibly complex) eigenvectors.

Here are some facts that are based on the spectral theorem.

1. If $A \in \mathbb{R}^{n\times n}$ is positive definite, it has a square root B, s.t. $B^2 = A$.

   Proof: By the spectral theorem we can write $A = Q\Lambda Q^T$, where all diagonal entries of Λ are positive and Q is orthogonal. Let $B = Q\Lambda^{1/2}Q^T$, where $\Lambda^{1/2}$ is the diagonal matrix that has square roots of main diagonal elements of Λ along its main diagonal, and calculate B^2 (more generally if A is positive semi-definite, it has a square root). You should now prove for yourself that A^{-1} is also positive definite and therefore $A^{-1/2}$ also exists. $\square$

2. If A is idempotent, and λ is an eigenvalue of A, then $\lambda = 1$ or $\lambda = 0$.

   Proof: Exercise 6.4.

There is another way to think about the result of the Spectral theorem. Let $x \in \mathbb{R}^n$ and consider $Ax = Q\Lambda Q^T x$. Then (do it as an exercise!) carrying out the matrix multiplication on $Q\Lambda Q^T$ and letting $q_1, q_2, \ldots, q_n$ denote the columns of Q and $\lambda_1, \lambda_2, \ldots, \lambda_n$ denote the diagonal entries of Λ, we have:

$$Q\Lambda Q^T = \lambda_1 q_1 q_1^T + \lambda_2 q_2 q_2^T + \cdots + \lambda_n q_n q_n^T$$

and so

$$Ax = \lambda_1 q_1 q_1^T x + \lambda_2 q_2 q_2^T x + \ldots + \lambda_n q_n q_n^T x$$

We recognize $q_i q_i^T$ as the projection matrix onto the line spanned by q_i, and thus every $n \times n$ symmetric matrix is the sum of n 1-dimensional projections. That should come as no surprise: we have orthonormal basis $q_1, q_2, \ldots q_n$ for $\mathbb{R}^n$, therefore we can write every $x \in \mathbb{R}^n$ as a unique combination $c_1 q_1 + c_2 q_2 + \ldots + c_n q_n$, where $c_1 q_1$ is precisely the projection of x onto line through q_1. Then applying A to the expression we have $Ax = \lambda_1 c_1 q_1 + \lambda_2 c_2 q_2 + \ldots + \lambda_n c_n q_n$, which of course is just the same thing as we have above.

Variational characterization of eigenvectors The spectral theorem shows that an orthonormal eigenbasis exists for Hermitian matrices. The *Courant-Fischer* theorem characterizes the eigenvectors/values of a Hermitian matrix through an optimization perspective. In this section we focus on real, symmetric matrices, but the discussion is essentially identical for complex matrices.

For a symmetric matrix $A \in \mathbb{R}^{n \times n}$ let the *Rayleigh quotient* be $R : \mathbb{R}^n \to \mathbb{R}$ given by

$$R(x) = \frac{x^T A x}{x^T x}$$

We set $R(0) = 0$. In other words, R computes the quadratic form of x then normalizes by then length of x. If x is an eigenvector with eigenvalue λ then we can see $R(x) = \lambda$.

By the spectral theorem we can create an orthonormal basis, $\{v_1, \ldots, v_n\}$, of eigenvectors of A. Suppose we order the eigenvalues as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. For any $x \in \mathbb{R}^n$ we can write $x = \sum_{i=1}^n a_i v_i$. Therefore we can rewrite R as (check this yourself)

$$R(x) = \frac{\sum_{i=1}^n \lambda_i a_i^2}{\sum_{i=1}^n a_i^2}$$

Notice R is now a weighted average of the eigenvalues.

Suppose that $\lambda_1 > \lambda_2$ i.e. the *leading eigenvalue* is strictly larger than the next eigenvalue. With out loss of generality suppose that x has norm 1 which means $\sum_{i=1}^n a_i^2 = 1$ (why can we do this?). Then the weighted average (and therefore the Rayleigh quotient) is maximized when $a_1 = 1$ and $a_i = 0$ for $i > 1$. In other words, the Rayleigh quotient is maximized when x is the leading eigenvector v_1 (really we should say "proportional to" the leading eigenvector).

We have just shown that the solution to $\max_{x:||x||=1} R(x)$ is given by v_1 in the case that $\lambda_1 > \lambda_2$.[3] In other words, the leading eigenvector gives the direction which maximizes the quadratic form $x^T A x$ when $\lambda_1 > \lambda_2$. What about if $\lambda_1 = \lambda_2$.

If two eigenvectors v_1, v_2 have the same eigenvalue λ then any linear combination of those two eigenvectors gives another eigenvector with the same eigenvalue (since $A(a_1 v_1 + a_2 v_2) = \lambda(a_1 v_1 + a_2 v_2)$). Let $V_1(A) \subseteq \mathbb{R}^n$ be the subspace corresponding to eigenvectors of the leading eigenvalue λ_1. The Rayleigh quotient is maximized by any eigenvector in $V_1(A)$ (why?).

The general Courant-Fischer theorem essentially says that if we restrict the x to lie in the subspace orthogonal to the leading eigenvector (or eigen subsapce V_1) then the Raleigh quotient is maximized by the eigenvector corresponding to the second largest eigenvalue. Similarly for the 3rd, 4th, ...

[3] Restricting x to be norm 1 is an technical detail that makes things a little nicer, but doesn't really change the meaning of this result.

For a statement and discussion of the full Courant-Fischer theorem see wikipedia: https://en.wikipedia.org/wiki/Min-max_theorem and https://en.wikipedia.org/wiki/Rayleigh_quotient.

6.6 Examples: Spectral theory

Convergence of Markov chains (Levin, Peres, Wilmer) In this example, we will use the spectral decomposition to show certain Markov chains reach stability exponentially fast.

A Markov chain, conceptually, models the evolution of some quantity over time, in cases where the next time step's outcome depends only on the previous outcome and not on the entire history. Markov chains, and their continuous-time analogs, are widely used in the physical and social sciences, operations research and simulation. Markov Chain Monte Carlo (MCMC) algorithms are one example.

Let's consider a simple Markov chain: Consider a sequence of random variables $X_1, X_2, \ldots$ each taking one of n possible values, which we can assume without loss of generality are the integers $\{1, 2, \ldots n\}$.

Such Markov chains are defined by a *transition matrix* $\boldsymbol{P} = (p_{ij})_{i,j=1\ldots n}$ defined by the conditional probabilities

$$\text{Prob}(X_1 = j \mid X_0 = i) = p_{ij}$$

We will assume, as is often the case, that

$$\text{Prob}(X_{m+1} = j \mid X_m = i) \qquad m \geq 0$$

In other words, p_{ij} is the probability the outcome takes value, or 'state', j tomorrow if it is in state i today. You should convince yourself that multi-step transition probabilities are given by matrix powers, i.e.

$$p_{ij}^{(k)} = (\boldsymbol{P}^k)_{ij} = \text{Prob}(X_{m+k} = j \mid X_m = i)$$

If $\nu = (\nu_1 \ldots \nu_n)$ is any probability vector ($\nu_i \geq 0$, $\sum \nu_i = 1$) representing the distribution of X_0, which we call the *initial distribution*, we have by definition of conditional probabilities

$$\text{Prob}(X_m = j) = (\nu^T \boldsymbol{P}^m)_j$$

We will make the following additional assumptions about our matrix $\boldsymbol{P}$. This guarantees the chain is irreducible and aperiodic (terms you do not need to know for this course):

$$p_{ij} > 0 \qquad \forall\, i, j$$

A fact from Markov chain theory is that

$$\exists\, \pi = (\pi_1 \ldots \pi_n) \quad \pi^T \boldsymbol{P} = \pi^T, \quad \pi_i > 0\, i = 1 \ldots n$$

π is called the stationary distribution. A standard question in Markov chain theory generally is to ask: **How quickly does**

$$p_{ij}^{(m)} \to \pi_j \quad \forall\, i$$

These types of questions are important if, for example, you are using MCMC to estimate an unknown 'true' distribution for model parameters, in which convergence of the MCMC algorithm implies you have indeed done so.

We make a final assumption (called reversibility):

$$\pi_i p_{ij} = p_{ij}\pi_j \quad \forall\, i, j$$

We can answer the convergence question in the following sequence of steps, which you can work out for yourself by computation.

1. Define the matrix A with entries $a_{ij} = \sqrt{\frac{\pi_i}{\pi_j}} p_{ij}$ and check that the assumptions above imply A is a symmetric positive definite matrix. In other words, $A = D^{1/2}\boldsymbol{P}D^{-1/2}$

2. You can check using the spectral theorem and direct calculation: $\lambda_1 = 1 > \lambda_2$ with $\phi_1 = \sqrt{\pi}$, and $\lambda_n > -1$. In other words there is one eigenvalue at 1 and the rest are in $(-1, 1)$

3. Spectral theorem: $D^{1/2}\boldsymbol{P}D^{-1/2} = A = \Phi\Lambda\Phi^T$, where Φ is the matrix whose columns are $\phi_1, \phi_2 \dots \phi_n$ and Λ is the diagonal matrix of eigenvalues. This implies

   $$\boldsymbol{P} = D^{-1/2}\Phi\Lambda\Phi^T D^{1/2}, \qquad \boldsymbol{P}^m = D^{-1/2}\Phi\Lambda^m\Phi^T D^{1/2}$$

   Note $D^{-1/2}\phi_i = \psi_i$ are eigenvectors for $\boldsymbol{P}$ corresponding to the same eigenvalues, and ψ_1 is the vector of ones.

4. Recall that $0 < \lambda_\star = \max(|\lambda_2|, |\lambda_n|) < 1$. Using the previous statement, you can calculate

   $$\frac{p_{ij}^{(m)}}{\pi_j} = \left(D^{-1/2}\Phi\Lambda^m\Phi^T D^{-1/2}\right)_{ij} = 1 + \sum_{k=2}^{n} \lambda_k^m (D^{-1/2}\Phi)_{ik}(D^{-1/2}\Phi)_{jk}$$

   and therefore by Cauchy-Schwartz

   $$\left|\frac{p_{ij}^{(m)}}{\pi_j} - 1\right| \le \sum_{k=2}^{n} \lambda_k^m |(D^{-1/2}\Phi)_{ik}(D^{-1/2}\Phi)_{jk}| \le \lambda_\star^m \sqrt{\sum_{k=2}^{n} |(D^{-1/2}\Phi)_{ik}|^2 \sum_{k=2}^{n} |(D^{-1/2}\Phi)_{jk}|^2} \le C\lambda_\star^m$$

   for a positive constant C not dependent on i, j.

In other words, since $\lambda_\star \in (0, 1)$, the Markov chain converges to its stationary distribution exponentially fast.

Principal Components Analysis and the Rayleigh quotient The Rayleigh quotient characterization of eigenvalue/vectors is useful for a lot of statistical applications. For example, Principal Components Analysis (PCA) finds directions of maximal variance which you can write out as a maximizing the Rayleigh quotient of the sample covariance matrix. Given a data matrix $X \in \mathbb{R}^{n \times d}$, assuming the columns have been mean centered, the first PCA component can be formulated as the following optimization problem.

$$\min_{w \in \mathbb{R}^d \text{ s.t. } ||w||=1} ||X - Xww^T||_F^2$$

where $||A||_F^2 = tr(A^T A)$ is the Frobenius norm (note you can show $||A||_F^2 =$ sum of the squared entries of A). The term, Xww^T, is the projection of the data onto the subspace spanned by the unit vector w (why?). Therefore $X - Xww^T$ is the "residual" of the projection i.e. how far away the original data are away from the projected data. In other words, the problem above means: find w that minimizes the residuals of the projected data.

A little bit of matrix algebra (exercise 6.6) shows the minimizing above problem of minimizing the residuals is equivalent to maximizing the Rayleigh quotient of the sample covariance matrix i.e.

$$\max_{w \in \mathbb{R}^d \text{ s.t. } ||w||=1} w^T S w$$

where $S := X^T X$ is the sample covariance matrix (recall variables of X have been centered). The upshot is: the first PCA component (often referred to as the first *loading*) is given the leading eigenvector of the sample covariance matrix. The high PC components are given by the remaining eigenvectors of the sample covariance matrix.

At lot of algorithms in statistics and machine learning can be understood as as the eigenvectors/values of various matrices, For example: PCA, least squares/Ridge regression, Canonical Correlation Analysis, Partial Least Squares, Fisher's Linear Discrimination, Spectral Clustering, kernel versions of all of these, etc. An excellent survey is given by Eigenproblems in Pattern Recognition (and can be found at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.128.2674&rep=rep1&type=pdf).

Exercises

6.1 Let $A \in \mathbb{C}^n$. We will show that if A is non-negative definite it is Hermitian. This is **not true in the case of real matrices**.

Define the bilinear function $B(x, y)$ by $\langle Ax, y \rangle = y^H Ax$. Note that 'bilinear' in this case means $y \mapsto B(x, y)$ is anti-linear in that $B(x, \alpha y + \beta z) = \bar{\alpha} B(x, y) + \bar{\beta} B(x, z)$.

1. By direct calculation, show the following identity holds for all complex numbers α, β and vectors x, y:

$$\alpha\bar{\beta} B(x, y) + \bar{\alpha}\beta B(y, x) = B(\alpha x + \beta y, \alpha x + \beta y) - |\alpha|^2 B(x, x) - |\beta|^2 B(y, y)$$

2. Show that $B(x,x) = 0$ for all x implies $A = 0$, the matrix of all zeros. Hint: Apply the identity above twice, first with $\alpha = \beta = 1$ then with $\alpha = i = \sqrt{-1}$ and $\beta = 1$. Recall that for any vector v, $\langle v, x\rangle = 0$ for all x implies v is zero.

3. Show that if $B(x,x)$ is a real number for all x, then $A^H = A$.

   This last statement proves the result, since non-negative definiteness implies $B(x,x) \geq 0$ is real.

   NOTE: In fact you have shown that if $\langle Ax, x\rangle$ is *real* for all $x \in \mathbb{C}^n$, then A is hermitian.

4. Find an example of $A \in \mathbb{R}^2$ non-negative definite that is not symmetric.

6.2 Show that if a matrix is triangular, its determinant is the product of the entries on the main diagonal.

6.3 Let $s_1, s_2, \ldots, s_n \in \mathbb{R}^n$ be the set of linearly independent eigenvectors of A, let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the corresponding eigenvalues (note that they need not be distinct), and let S be the $n \times n$ matrix such that the j-th column of which is s_j. Show that if Λ is the $n \times n$ diagonal matrix s.t. the ii-th entry on the main diagonal is λ_i, then $AS = S\Lambda$, and since S is invertible (why?) we have $S^{-1}AS = \Lambda$.

6.4 Show that if $rank(A) = r$, then A^TA has r non-zero eigenvalues.

6.5 Show that if A is idempotent, and λ is an eigenvalue of A, then $\lambda = 1$ or $\lambda = 0$.

6.6 Show that if A is a real matrix that is orthogonally diagonizable, then it must be symmetric. Show that a real matrix that is normal but is not symmetric must have at least one complex eigenvalue.

Hint: Use the spectral decomposition for (complex) normal matrices and the exercise above, which says if $\langle Ax, x\rangle$ is *real* for all $x \in \mathbb{C}^n$, then A is Hermitian. But if $A \in \mathbb{R}^n$ then being Hermitian is the same as being symmetric.

6.7 Recall the Fibonacci diagonalization example above. If you had to compute the nth (say $n = 100$) Fibonacci number by hand using the formula $F_{n+2} = F_n + F_{n+1}$ how many operations would it take? How about using the diagonalization procedure above?

Make some reasonable assumptions about how long it takes a grad student to add two numbers and look up the nth power of a number on wolfram alpha; based on these assumptions, how much time would the grad student save by knowing the diagonalization formula if his advisor asked him for the $10^7 th$ Fibonacci number?

6.8

Let $X \in \mathbb{R}^{n \times d}$. Show the following two optimization problems are equivalent (see section 6.5 for context)

$$\min_{w \in \mathbb{R}^d \text{ s.t. } ||w||=1} ||X - Xww^T||_F^2$$

$$\max_{w \in \mathbb{R}^d \text{ s.t. } ||w||=1} w^T S w$$

Where $||A||_F^2 := tr(AT^A)$ is the Frobenius norm. All you need to do to solve this problem is expand the objective function in the first problem and use properties of the trace and the fact that $||w|| = 1$. You should eventually get to something that looks like the negative of the second problem (plus some constants).

7 Tensors

In the orthogonal decomposition of a Hilbert space V, we showed that if K is a closed subspace then each $v \in V$ can be written $v = k_1 + k_2$ where $k_1 \in K$ and $k_2 \in K^\perp$. More succinctly, V is given as a direct sum of subspaces written $V = K \oplus K^\perp$. The dimensions of these subspaces add to the dimension of V.

In this section we consider a way of putting together two vector spaces V, H such that the dimension of the new space is $\dim V \times \dim H$ rather than a sum. In addition, we will look at operations on pairs of vector spaces that produce such structure.

The word 'tensor' recently has become popular in machine learning and statistics. See this StackExchange post for a discussion of whether that is justified. Tensors have long-standing use in mathematics and physics. An example from statistics comes from multi-dimensional smoothing splines (see Hastie ch. 5).

Here **all vector spaces considered are finite-dimensional**, though the concept of a tensor can apply to very general spaces. See Lang or Roman, 'Advanced Linear Algebra.'

Definition: Tensor product Suppose V, U are finite-dimensional vector spaces over the same field (e.g. real or complex numbers). Suppose $\{v_i\}_1^n$ and $\{u_i\}_1^m$ are bases for V, U. Suppose $\otimes$ is a bilinear map on pairs $(v_i, u_j) \in V \times U = \{(v, u) \,:\, u \in U, v \in V\}$, that is

$$v_i \mapsto v_i \otimes u_j \quad \text{and} \quad u_j \mapsto v_i \otimes u_j \quad \text{are linear maps}$$

We extend this map to $V \times U$ by linearity, such that if $v = \sum \alpha_i v_i$ and $u = \sum \beta_j u_j$ then

$$v \otimes u = \sum \alpha_i \beta_j v_i \otimes u_j$$

Check $\{v_i \otimes u_j\}_{i=1\ldots n, j=1\ldots m}$ are linearly independent. Thus if we define the **tensor product of** V, U as

$$V \otimes U = \operatorname{span}\{v_i \otimes u_j\}_{i=1\ldots n, j=1\ldots m}$$

then $V \otimes U$ is a vector space of dimension nm whose elements are of the form $v \otimes u$.

We note that a tensor product defined in this way is a 'universal' bilinear map, in the sense that if $T : V \times U \mapsto W$ is another bilinear linear map to a vector space W, then there is a unique linear map $S : V \otimes U$ such that $T(v, u) = S(v \otimes u)$. This is done by defining $S(v_i \otimes u_j) = T(v_i, u_j)$ on basis elements and extending it to the entire space.

An alternate definition comes from Halmos, 'Finite Dimensional Vector Spaces.' First, note that we may define a direct sum $V \oplus U$ in cases where U, V are disjoint (except for the zero element) and every vector $y \in V \oplus U$ can be written uniquely as $v + u$ for $v \in V, u \in U$. This space can thus be identified with the pairs (v, u). The space of bilinear maps $(v, u) \mapsto \ell(u, v)$ is a finite-dimensional vector space and thus has a dual space of continuous linear functionals.

Alternate definition Let V, U be as above. Then $V \otimes U$ is defined as the dual space B^* of B, which is the space of bilinear maps on $V \oplus U$.

Elements of $V \otimes U$ are denoted $y = v \otimes u$, where y is such that $y(\ell) = \ell(v, u)$ for all $\ell \in B$. Note that this shows $\otimes$ is a bilinear map.

If $\{v_i\}_1^n$ and $\{u_i\}_1^m$ are bases for V, U, then $\{v_i \otimes u_j\}_{i=1\ldots n, j=1\ldots m}$ is a basis for $V \otimes U$.

Examples:

1. Say $V = \mathbb{R}^n$ and $U = \mathbb{R}^m$. $v \otimes u = vu^T$ is called the **outer product** and is a tensor product. $V \otimes U$ is the space of $n \times m$ matrices.

   Note $\langle v, u \rangle = tr(v \otimes u)$.

2. $V = \mathbb{R}^{n \times p}$ and $U = \mathbb{R}^{m \times q}$ the spaces of $n \times p$ and $m \times q$ real matrices. Define the **Kronecker product** of $A = \{a_{ij}\}_{i=1\ldots n, j=1\ldots p}$ and $B \in U$ as

$$A \otimes B = \begin{bmatrix} a_{11}B & \ldots & a_{1p}B \\ \ldots & & \\ a_{n1}B & \ldots & a_{np}B \end{bmatrix}$$

   Then $V \otimes U$ is the space of $nm \times pq$ matrices.

3. V is an n-dimensional complex vector space and $U = V^*$ its dual. Define the functional

$$v \otimes \ell = \ell(v)$$

   Then $\otimes$ is a tensor product.

4. Suppose V, U are n, m dimensional vector spaces over the same field. Suppose $\{v_i\}$ and $\{\ell_j\}$ are bases for V, U^*. We may define $V \otimes U^*$ and $\otimes$ as in the alternate definition above.

   Now consider the following map on $V \otimes U^*$:

$$L(v \otimes \ell)(u) = v\ell(u) \qquad \forall\, u \in U$$

   Thus L maps $V \otimes U^*$ to the space of linear transformations from U to V, $\hom(U, V)$, also called homomorphisms as they preserve the vector space structure. In the exercises, you are asked to show this map is in fact one-to-one and onto. In other words, $V \otimes U^*$ is isomorphic to the set of linear transformations from U to V.

7.1 Exercises

7.1 V, U are n, m dimensional vector spaces over $\mathbb{C}$. Show that if $\{v_i\}_1^n$ are linearly independent, then for $u_1 \dots u_n \in U$ we have

$$\sum v_i \otimes u_i = 0 \quad \implies u_i = 0, \ \forall i$$

and conclude that $v \otimes u = 0$ if and only if one of the arguments is zero.

7.2 V, U are n, m dimensional vector spaces over $\mathbb{C}$. Show that the map defined above $L : V \otimes U^* \mapsto \hom(U, V)$ is one-to-one and onto. Hint: Recall that there is a bijection between two finite-dimensional linear spaces if and only if they have the same dimension. You can characterize the bijection using their respective bases.

7.3 If $\otimes$ is the Kronecker product given above, show that for real matrices A, B, C, D of appropriate dimensions

1. $(A \otimes B)^T = A^T \otimes B^T$
2. $(A \otimes B) \otimes C = A \otimes (B \otimes C)$
3. $(A \otimes B)(C \otimes D) = (AC \otimes BD)$
4. A, B invertible then $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$
5. $tr(A \otimes B) = tr(A)tr(B)$
6. $rank A \otimes B = rank(A)rank(B)$

8 Singular Value Decomposition

Eigenvalues/vectors tell you a lot about a matrix[4], however, the definition only makes sense for square matrices. The Spectral theorem says every symmetric matrix has an eigenbasis which turns out to be very useful in many applied and theoretical problems. The *singular value decomposition* (SVD) generalizes the notion of eigenvectors/values to any (possibly non square, symmetric) matrix. For a matrix $A \in \mathbb{R}^{n \times d}$ the SVD is essentially the eigenvalues/vectors of the square, symmetric matrices $A^T A \in \mathbb{R}^{d \times d}$ and $AA^T \in \mathbb{R}^{n \times n}$.

The following notes give a nice summary of SVD http://www4.ncsu.edu/~ipsen/REU09/chapter4.pdf.

8.1 Definition

We first give the definition of a singular value decomposition of a matrix. It looks like there are a lot of parts to the definition, but the important quantities are the matrices U, S, and V. We then state an existence theorem (every matrix has an SVD) then state a uniqueness

[4]The prefix *eigen* means "proper" in German which should give some indication of how important mathematicians thinks eigenvalues/vectors are.

theorem (an SVD of a matrix is *unique-ish*). In this chapter we focus on real matrices, however, the story is similar for complex matrices.

Definition: Let $A \in \mathbb{R}^{n\times d}$ with $m := \min(n, d)$ and $r = \text{rank}(A)$. The SVD of A is given by

$$\begin{aligned} A :=& \begin{bmatrix} U & U_0 \end{bmatrix} \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V \\ V_0 \end{bmatrix} := U_{full} S_{full} V_{full}^T \\ =& USV^T \\ =& \sum_{k=1}^{r} s_{kk} u_k v_k^T \end{aligned} \tag{4}$$

The matrices $U_{full} \in \mathbb{R}^{n\times n}$, $S_{full} \in \mathbb{R}^{n\times d}$, $V_{full} \in \mathbb{R}^{d\times d}$ are known as the full SVD. The matrices $U \in \mathbb{R}^{n\times r}$, $S \in \mathbb{R}^{R\times R}$, $V \in \mathbb{R}^{d\times R}$ are known as the reduced SVD. Most of the time we focus on the reduced SVD.

The matrices $U \in \mathbb{R}^{n\times r}$, $U_0 \in \mathbb{R}^{n\times(m-r)}$ satisfy

- $U_{full} = \begin{bmatrix} U & U_0 \end{bmatrix} \in \mathbb{R}^{n\times n}$ is an orthogonal matrix
- The columns of U, given by $u_k \in \mathbb{R}^n$ give an orthonormal basis of the the column space of A i.e. $\text{col}(U) = \text{col}(A)$ and are called the *left singular vectors*
- The columns of U_0 span the the left kernel of A i.e. $\text{col}(U_0) = N(A^T)$

The matrices $V \in \mathbb{R}^{d\times r}$, $V_0 \in \mathbb{R}^{d\times(m-r)}$ satisfy

- $V_{full} = \begin{bmatrix} V & V_0 \end{bmatrix} \in \mathbb{R}^{d\times d}$ is an orthogonal matrix
- The columns of V, given by $v_k \in \mathbb{R}^d$ give an orthonormal basis of the the row space of A i.e. $\text{col}(V) = \text{col}(A^T)$ and are called the *right singular vectors*
- The columns of V_0 span the the right kernel of A i.e. $\text{col}(V_0) = N(A)$

The matrix $S \in \mathbb{R}^{r\times r}$ is a diagonal matrix with strictly positive entries given by $s_{kk} > 0$ which are called the *singluar values*.

Exercise 1 asks you to verify the three different versions of the SVD decomposition are in fact equal. The SVD of A is related to the eigen-decomposition of A^TA and AA^T as follows

1. The left singular vectors, $u_k \in \mathbb{R}^n$, are the eigenvectors of A^TA.
2. The right singular vectors, $v_k \in \mathbb{R}^d$, are the eigenvectors of AA^T.
3. The singular values, s_{kk}, are the square roots of the eigenvalues of A^TA and AA^T (recall these have the same eigenvalues).

Theorem (SVD existence): Every matrix has a singular value decomposition.

Proof: First we construct the matrix V_{full} by an eigen-decomposition of A^TA. Note $A^TA \in \mathbb{R}^{d\times d}$ symmetric and therefore has a set of d real orthonormal eigenvectors. Since $rank(A^TA) = rank(A) = r$, we can see that A^TA has r non-zero (possibly-repeated) eigenvalues (Exercise 6.3). Arrange the eigenvectors $v_1, v_2, \ldots, v_d$ in such a way that the first $v_1, v_2, \ldots, v_r$ correspond to non-zero $\lambda_1, \lambda_2, \ldots, \lambda_r$ and put $v_1, v_2, \ldots, v_d$ as columns of V_{full}. Note that as $v_{r+1}, v_{r+2}, \ldots, x_d$ form a basis for $N(A)$ by Exercise 2.4 as they are linearly independent, $dim(N(A)) = d - r$ and

$$v_i \in N(A) \qquad for \quad i = r+1, ..., d.$$

Therefore $v_1, v_2, \ldots, v_r$ form a basis for the row space of A.

We construct S_{full} using the eigenvalues of A^TA. Now set $s_{ii} = \sqrt{\lambda_i}$ for $1 \leq i \leq r$. Let $S_{full} \in \mathbb{R}^{n\times d}$ be the middle matrix with S on the top left and the remaining entries zero.

Finally we construct the matrix $U_{full} \in \mathbb{R}^{n\times n}$. For $1 \leq i \leq r$, let

$$u_i = \frac{Av_i}{s_{ii}} \in \mathbb{R}^n$$

be the first r columns of U_{full}. You should verify for yourself that u_i's are orthonormal ($u_i^Tu_j = 0$ if $i \neq j$, and $u_i^Tu_i = 1$). By Gram-Schmidt, we can extend the set $u_1, u_2, \ldots, u_r$ to a complete orthonormal basis for $\mathbb{R}^n$, $u_1, u_2, \ldots, u_r, u_{r+1}, \ldots, u_n$. As $u_1, u_2, \ldots, u_r$ are each in the column space of A and linearly independent, they form an orthonormal basis for column space of A and therefore $u_{r+1}, u_{r+2}, \ldots, u_n$ form an orthonormal basis for the left nullspace of A.

We now verify that $A = U_{full}S_{full}V_{full}^T$ by checking that $U_{full}^TAV_{full} = S_{full}$. Consider ij-th entry of $U_{full}^TAV_{full}$. It is equal to $u_i^TAv_j$. For $j > r$, $Av_j = 0$ (why?), and for $j \leq r$ the expression becomes $u_i^Ts_{jj}u_j = s_{jj}u_i^Tu_j = 0$(if $i \neq j$) or s_{ii} (if $i = j$). And therefore $U_{full}^TAV_{full} = S_{full}$, as claimed. $\square$

Theorem (SVD uniqueness-ish): Suppose U, S, V is the reduced SVD of a matrix A and $Q \in \mathbb{R}^{r\times}$ is an orthogonal matrix. Then UQ, S, UQ is another SVD of A. We can make a similar statement for the matrices U_0 and V_0. This set of orthogonal transofmations gives every SVD of A.

Remark: The singular values of a matrix are unique i.e. they are the square root of the eigenvalues of A^TA and AA^T.

Remark: The singular vectors of a matrix are not unique. However, the subspaces spanned by the singular vectors (e.g. span(u_1, u_u)) are unique.

One of the main applications of SVD is finding a lower rank approximation of A. Note that we can write A in terms of outer products of its singular vectors by

$$A = \sum_{i=1}^{r} s_iu_iv_i^T$$

Note we ordered the singular values in non-increasing order i.e. $s_1 \geq s_2 \ldots$. We might approximate A by retaining the first $k < r$ singular vectors i.e.

$$\tilde{A} = \sum_{i=1}^{k} s_i u_i v_i^T$$

We refer to this approximation as the rank k SVD of A. Exercise 7.2 asks you to justify the fact that $\tilde{A}$ is rank k. This approximation is at the heart of methods like PCA and is discussed in more depth in Section 7.2 below.

8.2 Low rank approximation

Suppose we are given a matrix $A \in \mathbb{R}^{n \times d}$ and we want to approximate it with a low rank matrix. Consider the following optimization problem: given an integer K we find a rank K approximation of A by

$$\begin{array}{ll} \text{maximize} & ||A - X||_F^2 \\ \text{subject to} & \text{rank}(X) \leq K. \end{array} \tag{5}$$

where $||M||_F^2 = \text{tr}(M^T M)$ is the Frobenius norm of a matrix. In other words, we find the matrix of rank less than or equal to K that is closes to A in the Frobenius norm.
Eckart-Young theorem: The solution to Problem 5 is given by the rank K SVD of A.

Remark: The rank K SVD approximation of A is given by $\sum_{i=1}^{K} s_i u_i v_i^T$ where s_i, u_i, v_i are as in the definition of the SVD (e.g. u_i is the ith left singular vector of A.)

Remark: The discussion in Section 6.5 above on the first Principal Components direction proves the Eckart-Young theorem in the case $k = 1$ (why?)

For a matrix $A \in \mathbb{R}^{n \times d}$ denote *2-norm* or *spectral norm* by

$$||A||_2 = \sqrt{s_1} \text{ (i.e. square root the largest singular value of } A)$$

Warning: the 2-norm is **not** equal to the Frobenius norm. For a discussion of matrix norms (which will also clarify the naming conventions) see wikipedia https://en.wikipedia.org/wiki/Matrix_norm. Exercise 7.3 asks you to show that the 2-norm is a norm.

For the proof we point of Eckart-Young we point the reader to: https://en.wikipedia.org/wiki/Low-rank_approximation#Proof_of_Eckart.E2.80.93Young.E2.80.93Mirsky_theorem_.28for_Frobenius_norm.29. This proof uses some facts about the matrix 2-norm and the Frobenius norm which are left as exercises at the end of this chapter.

8.3 A few Applications

Pseudoinverse and least squares Suppose we have a matrix $A \in \mathbb{R}^{n \times d}$ and we are interested in finding some kind of inverse. If $n = d$ and A is full rank then we have an actual

inverse, however, if A is not full rank or. Even worse, if A is not square, it's not exactly clear what an inverse would even mean.

To make this problem a little more concrete consider solving a linear system, given A and $b \in \mathbb{R}^n$ find $x \in \mathbb{R}^d$ such that

$$Ax = b.$$

If $b \notin \text{col}(A)$ then there is no solution. If $b \in \text{col}(A)$, but the columns of A a linearly dependent then there is an infinite number of solutions. To rectify both of these situations let's relax the problem a little and consider finding x to minimize

$$||Ax - b||_2$$

We can show the unique solution, $x*$, is given using the full SVD of A

$$x* = V_{full} S^+_{full} U^T_{full} b$$

where $S^+_{full} = \begin{bmatrix} S^{-1} & 0 \\ 0 & 0 \end{bmatrix}$ where S^{-1} always exists since S is a diagonal matrix with strictly positive entries. It's a good exercise to show that if A is square an invertible the above equation gives the normal least squares solution.

The *Moore-Penrose* pseudoinverse of a matrix $A \in \mathbb{R}^{n \times d}$, denoted $A^+ \in \mathbb{R}^{d \times n}$, is given by $A^+ = V_{full} S^+_{full} U^T_{full}$. This pseudo inverse always exists since every matrix has an SVD. For a discussion of it's properties see https://en.wikipedia.org/wiki/Moore-Penrose_pseudoinverse.

We can now write the minimum norm solution to a linear equation as

$$x* = A^+ b$$

which makes A^+ suggestive of being an inverse. More suggestively, one can show that A^+ solves both

$$\min_{X \in \mathbb{R}^{d \times n}} ||AX - I_n||_F$$

$$\min_{X \in \mathbb{R}^{d \times n}} ||XA - I_d||_F$$

Remark: The pseudoinverse is an example of the following problem solving principle: **if you can't solve the problem you are given then try solving a related, easier problem.**

Applications with Data We briefly mention a few applications statistics/machine learning applications of SVD. For a more detailed discussion of these methods see Eigenproblems in Pattern Recognition (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.128.2674&rep=rep1&type=pdf), standard statistics references, or Google.

1. The rank k PCA decomposition is equal to the rank k SVD of the data matrix after centering the variables. If $X \in \mathbb{R}^{n \times d}$ whose columns have been centered then the rank k PCA approximation is given by the rank k SVD approximation of X. The resulting data can be represented by the *unnormalizes scores*, $U_k S_k$ where $U_k \in \mathbb{R}^{n \times k}$ is the first k left singular vectors (i.e. first k columns of U) and $S_k \in \mathbb{R}^{k \times k}$. The *loadings* $V_k \in \mathbb{R}^{d \times k}$ (first k right singular vectors) represent the k direction in $\mathbb{R}^d$ of maximal variance.

2. An $n \times d$ pixel, black and white image is given by a matrix $X \in \mathbb{R}^{n \times d}$. The rank k SVD of an image matrix X can be used to *compress* the image: store an approximate image that looks almost exactly like the original image, but takes a lot less memory. For example, the original image requires saving nd numbers (i.e. one number for each pixel). Suppose we compute a rank k SVD of X and store the resulting singular values/vectors; the SVD approximation requires $k(n+d)+k$ numbers. Suppose $n = d = 1000$; the original image takes 1,000,000 numbers while the rank $k = 20$ SVD approximation takes 2020 numbers. Low rank image approximations often give remarkably good approximations of the original image (e.g. google "SVD image compression" to see some examples).

3. Suppose we have two data sets $X \in \mathbb{R}^{n \times d}$ and $Y \in \mathbb{R}^{n \times p}$ with the same rows but different columns. For example, consider a study where we have $n = 200$ patients, $d = 100$ clinical variables (e.g. height, weight, etc) and $p = 10000$ gene expression measurements. One way of investigating associations between two data matrices is Partial Least Squares (PLS). Let $S_{xy} \in \mathbb{R}^{d \times p}$ be the "cross-covariance matrix" given by $S_{xy} := X^T Y$ (assuming we have mean centered the columns of both data matrices). One version of PLS[5] amounts to computing the rank k SVD of S_{xy}.

4. *Spectral clustering* is a way of incorporating non-linearity into a clustering algorithm. Suppose we ave given a data set $X \in \mathbb{R}^{n \times d}$ and we would like to cluster there observations into K clusters. Spectral clustering amounts the following three steps

   (a) Fix a measure of distance between observations, $k : \mathbb{R}^d \times \mathbb{R}^d$, where k might be the standard inner product or more generally any kernel (e.g. radial basis kernel). Let $A \in \mathbb{R}^{n \times n}$ be given by $A_{ij} = k(x_i, x_j)$ where x_i is the ith observation (ith row of X). Let $D \in \mathbb{R}^{n \times n}$ be the diagonal matrix whose ith element, $D_{ii} = \sum_{j=1}^{n} A_{ij}$. Now let $L \in \mathbb{R}^{n \times n}$ be the *laplacian* matrix given by $L = D - A$.

   (b) Computing there rank K SVD of the laplacian matrix L resulting in a new matrix $Y \in \mathbb{R}^{n \times K}$ which we use to represent the original data.

   (c) Apply a standard clustering algorithm such as K-means to the new matrix Y.

   Spectral clustering is motivated by community detection for networks which explains the above naming/notation (Laplacian, A, etc). For a survey on spectral clustering see A Tutorial on Spectral Clustering (found at https://arxiv.org/pdf/0711.0189.pdf)

8.4 Principal Components Analysis

For an introduction to PCA see pages 373-385 from Introduction to Statistical Learning (PDF available at http://www-bcf.usc.edu/~gareth/ISL/). For some interesting PCA visualizations see http://projector.tensorflow.org/. This section discusses a number of different perspectives on PCA and connects them using linear algebra.

[5] Unfortunately there are a number of slightly different algorithms that are all called PLS. The one we are referring to is called EZ-PLS or PLS-SVD.

Suppose we have a data matrix $X \in \mathbb{R}^{n \times d}$ with n observations and d variables. Let $x_1, \ldots, x_n \in \mathbb{R}^n$ be the observations (rows of X). PCA can be used for dimensionality reduction (i.e. find a matrix in $\mathbb{R}^{n \times k}, k < d$ that represents the data with fewer variables) or directly for data analysis. The rank k-PCA returns

1. A *scores* matrix $U \in \mathbb{R}^{n \times k}$ whose columns are $u_1, \ldots, u_k \in \mathbb{R}^n$
2. A *loadings* matrix $V \in \mathbb{R}^{d \times k}$ whose columns are $v_1, \ldots, v_k \in \mathbb{R}^d$
3. A diagonal matrix $S \in \mathbb{R}^{k \times k}$ whose diagonal entries are $s_1, \ldots, s_k > 0$.

The value of k can range between 1 and $\min(n, d)$. PCA should remind you of SVD because it is an SVD.

Fact: The rank k PCA is the rank k SVD of the data matrix after centering the column means. In other words, let $X_c \in \mathbb{R}^{n \times d}$ be the matrix whose ith row is given by $x_i - \bar{x}$ (where $\bar{x} \in \mathbb{R}^d$ is the sample mean). Then k-PCA(X) = k-SVD(X_c).[6] The scores (loadings) are the left (right) singular vectors of X_c.

PCA is computed by an SVD, however, this fact alone does not provide much intuition into what PCA is doing. We present two more geometric perspectives on the first PCA component.

Two geometric perspectives of the first PCA component Consider a candidate loading vector $v \in \mathbb{R}^d$. Assume $||v|| = 1$.

One quantity of interest is the variance of data points after being projected onto the line spanned by v (i.e. $v^T x_1, \ldots, v^T x_n$ which turn out to be the scores). This quantity is var $\left(\{x_i^T v\}_{i=1}^n\right) = \frac{1}{n-1}\left(\sum_{i=1}^n x_i^T v - (\frac{1}{n}\sum_{j=1}^n x_i^T v)\right)$. To make life easier for ourselves **let's assume for the rest of this section that have first centered the data** i.e. $x_i \to x_i - \bar{x}$. In this case the variance is now given by

$$\text{var}\left(\{x_i^T v\}_{i=1}^n\right) = \frac{1}{n-1}\sum_{i=1}^n (x_i^T v)^2$$

We might decide that a "good" direction maximizes this variance.

Another quantity we are interested in is the residuals of the projections. Recall that $vv^T x_i \in \mathbb{R}^d$ gives the projection of the data point x_i onto the line spanned by v. We might decide a "good" direction minimizes the distance been the original data points x_i and their projections $vv^T x_i$. In other words, we might try to minimize the sum of the squared residuals $r_i = x_i - vv^T x_i \in \mathbb{R}^d$

$$\sum_{i=1}^n ||x_i - vv^T x_i||_2^2$$

Figure 2, borrowed from this discussion https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues, visualizes the two

[6] One can of course compute k-SVD(X). PCA(X) is an affine approximation of a data cloud while SVD(X) is a subspace approximation of a data cloud. Think the difference between linear regression with an intercept (affine) vs. with no intercept (subspace).

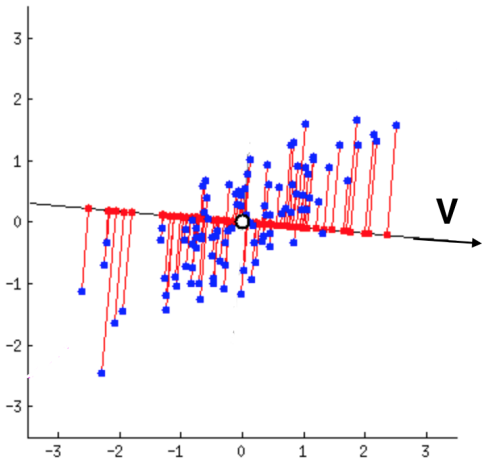

Figure 2: Two geometric perspectives of first PCA direction.

geometric quantities of interest. This image shows a number of two dimensional data points (in blue). The line spanned by the vector v is shown in black. The projections of the $vv^T x_i^T$ are the red points lying on the black line. The residuals are the red lines connecting the blue data points to their red projections.

To summarize, we consider two objectives. The first component $v \in \mathbb{R}^d$, $||v|| = 1$ that

1. maximizes the variance of the projects

$$\max_{||v||=1} \sum_{i=1}^{n} (x_i^T v)^2 \left(= \text{var}(\{x_i^T v\}_{i=1}^n)\right)$$

2. minimizes the squared residuals of the projected data

$$\min_{||v||=1} \sum_{i=1}^{n} ||x_i - vv^T x_i||_2^2$$

You can show that both of these formulations are related to the quadratic form $v^T S v$ where $S \in \mathbb{R}^{d \times d}$ is the sample covariance matrix ($S = \sum_{i=1}^n x_i x_i^T$ e.g. see section 3.4). Exercise 7.8 asks you to show that these two perspectives are give the same problem which is solved by SVD.

Other PCA components The higher order loadings, $v_2, v_3, \ldots$ can be understood using the above geometric perspectives plus an additional orthogonality constraint. In other words, we let v_2 be the direction in $\mathbb{R}^d$ orthogonal to v_1 that minimizes the residuals of the projected data points (or equivalently maximizes the variance of the projected data). Similarly for $v_3, v_4, \ldots$.

The scores are u_i are given by $u_i = X_c v_i$. You can show $u_i \perp u_j$ which means the new variables derived from PCA are uncorrelated.

8.5 The $25,000,000,000 Eigenvector: the Linear Algebra Behind Google

See this paper for a linear algebra perspective on PageRank: https://www.rose-hulman.edu/~bryan/googleFinalVersionFixed.pdf

Exercises

8.1 Verify the three different versions of the SVD decomposition are equivalent (e.g. $U_{full}S_{full}V_{full}^T = USV^T$).

8.2 Let $u_i \in \mathbb{R}^n$ and $u_i \in \mathbb{R}^d$ for $i = 1, \ldots, r$. Assume the u_i and the v_i are orthonormal (e.g. $u_i^T u_j = 0$ if $i \neq j$). Let $s_1, \ldots, s_r \in \mathbb{R}$ be scalars. Show that

$$A := \sum_{i=1}^{n} s_i u_i v_i^T$$

is a has rank R.

8.3 Show the matrix 2-norm $||A||_2 = \sqrt{s_1}$ is a norm on the set of $\mathbb{R}^{n \times d}$ matrices.

8.4 Show the matrix 2-norm is given by the following optimization problem

$$||A||_2^2 = \max_{u \in \mathbb{R}^d \text{ s.t. } ||u||_2 = 1} ||Au||_2$$

where $||u||_2$ above is the usual Euclidean vector norm. This property is where the matrix 2-norm get its (possibly confusing) name. In general, we can define a matrix p-norm for $p \geq 1$ by $||A||_p := \max_{u \in \mathbb{R}^d \text{ s.t. } ||u||_p = 1} ||Au||_p$ where the p norm of a vector is given by $||u||_p = \left(\sum_{i=1}^d u_i^p \right)^{\frac{1}{p}}$.

8.5 Prove the left inequality and right equality of the following

$$||A||_2 \leq \left(\sum_{i=1}^{n} \sum_{j=1}^{d} |A_{ij}|^2 \right)^{\frac{1}{2}} = ||A||_F$$

8.6 Prove the Frobenius norm squared is equal to the of the sum of the squared singular values i.e.

$$||A||_F^2 = \sqrt{\sum_{i=1}^{r} s_i^2}$$

This this should make the inequality $||A||_2 \leq ||A||_F$ immediate.

8.7 Show that $x* = A^x b$ is the solution to

$$\min_x ||Ax - b||_2$$

where A^+ is the psueudoinverse (defined in section 7.3).

8.8 Show the two perspectives of the first PCA component give the same result i.e. show

$$\min_{||v||=1} \sum_{i=1}^{n} ||x_i - vv^T x_i||_2^2$$

$$\max_{||v||=1} \sum_{i=1}^{n} (x_i^T v)^2$$

have the same solution.

8.9 Show the two the above two problems are both solved by the first right eigenvector of X_c (*Hint*: think about the Rayleigh quotient of the sample covariance matrix).

8.10 Let $A \in \mathbb{R}^{n \times d}$. Show that the solution to the following problem is given by the leading left and right singular vectors. Minimize $x^T Ay$ over $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^d$ with $||x|| = ||y|| = 1$. *Hint*: start with the SVD of A.

9 Matrix functions and differentiation

9.1 Basics

Here we just list the results on taking derivatives of expressions with respect to a vector of variables (as opposed to a single variable). We start out by defining what that actually means: Let $x = [x_1, x_2, \cdots, x_k]^T$ be a vector of variables, and let f be some real-valued function of x (for example $f(x) = sin(x_2) + x_4$ or $f(x) = {x_1}^{x_7} + x_{11} log(x_3)$). Then we define $\frac{\partial f}{\partial x} = \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \cdots, \frac{\partial f}{\partial x_k}\right]^T$. Below are the extensions

1. Let $a \in \mathbb{R}^k$, and let $y = a^T x = a_1 x_1 + a_2 x_2 + \ldots + a_k x_k$. Then $\frac{\partial y}{\partial x} = a$

   Proof: Follows immediately from definition.

2. Let $y = x^T x$, then $\frac{\partial y}{\partial x} = 2x$

   Proof: Exercise 5.1(a).

3. Let A be $k \times k$, and a be $k \times 1$, and $y = a^T A x$. Then $\frac{\partial y}{\partial x} = A^T a$

   Proof: Note that $a^T A$ is $1 \times k$. Writing $y = a^T A x = (A^T a)^T x$ it's then clear from 1 that $\frac{\partial y}{\partial x} = A^T a$. $\square$

4. Let $y = x^T A x$, then $\frac{\partial y}{\partial x} = Ax + A^T x$ and if A is symmetric $\frac{\partial y}{\partial x} = 2Ax$. We call the expression $x^T A x = \sum_{i=1}^{k} \sum_{j=1}^{k} a_{ij} x_i x_j$, a **quadratic form** with corresponding matrix A.

   Proof: Exercise 5.1(b).

9.2 Jacobian and Chain Rule

A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is said to be **differentiable at** x if there exists a linear function $L : \mathbb{R}^n \to \mathbb{R}^m$ such that

$$\lim_{x' \to x, x' \neq x} \frac{f(x') - f(x) - L(x' - x)}{\|x' - x\|} = 0.$$

It is not hard to see that such a linear function L, if any, is uniquely defined by the above equation. It is called the differential of f at x. Moreover, if f is differentiable at x, then all of its partial derivatives exist, and we write the Jacobian matrix of f at x by arranging its partial derivatives into a $m \times n$ matrix,

$$Df(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x) & \cdots & \frac{\partial f_1}{\partial x_n}(x) \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1}(x) & \cdots & \frac{\partial f_m}{\partial x_n}(x) \end{bmatrix}.$$

It is not hard to see that the differential L is exactly represented by the Jacobian matrix $Df(x)$. Hence,

$$\lim_{x' \to x, x' \neq x} \frac{f(x') - f(x) - Df(x)(x' - x)}{\|x' - x\|} = 0$$

whenever f is differentiable at x.

In particular, if f is of the form $f(x) = Mx + b$, then $Df(x) \equiv M$.

Now consider the case where f is a function from $\mathbb{R}^n$ to $\mathbb{R}$. The Jacobian matrix $Df(x)$ is a n-dimensional row vector, whose transpose is the gradient. That is, $Df(x) = \nabla f(x)^T$. Moreover, if f is twice differentiable and we define $g(x) = \nabla f(x)$, then Jacobian matrix of g is the Hessian matrix of f. That is,

$$Dg(x) = \nabla^2 f(x).$$

Suppose that $f : \mathbb{R}^n \to \mathbb{R}^m$ and $h : \mathbb{R}^k \to \mathbb{R}^n$ are two differentiable functions. The **chain rule of differentiability** says that the function g defined by $g(x) = f(h(x))$ is also differentiable, with

$$Dg(x) = Df(h(x))Dh(x).$$

For the case $k = m = 1$, where h is from $\mathbb{R}$ to $\mathbb{R}^n$ and f is from $\mathbb{R}^n$ to $\mathbb{R}$, the equation above becomes

$$g'(x) = Df(h(x))Dh(x) = \langle \nabla f(h(x)), Dh(x) \rangle = \sum_{i=1}^{n} \partial_i f(h(x)) h_i'(x)$$

where $\partial_i f(h(x))$ is the ith partial derivative of f at $h(x)$ and $h_i'(x)$ is the derivative of the ith component of h at x.

Finally, suppose that $f : \mathbb{R}^n \to \mathbb{R}^m$ and $h : \mathbb{R}^n \to \mathbb{R}^m$ are two differentiable functions, then the function g defined by $g(x) = \langle f(x), h(x) \rangle$ is also differentiable, with

$$Dg(x) = f(x)^T Dh(x) + h(x)^T Df(x).$$

Taking transposes on both sides, we get

$$\nabla g(x) = Dh(x)^T f(x) + Df(x)^T h(x).$$

Example 1. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function. Let $x^* \in \mathbb{R}^n$ and $d \in \mathbb{R}^n$ be fixed. Define a function $g : \mathbb{R} \to \mathbb{R}$ by $g(t) = f(x^* + td)$. If we write $h(t) = x^* + td$, then $g(t) = f(h(t))$. We have

$$g'(t) = \langle \nabla f(x^* + td), Dh(t) \rangle = \langle \nabla f(x^* + td), d \rangle.$$

In particular,

$$g'(0) = \langle \nabla f(x^*), d \rangle.$$

Suppose in addition that f is twice differentiable. Write $F(x) = \nabla f(x)$. Then $g'(t) = \langle d, F(x^* + td) \rangle = \langle d, F(h(t)) \rangle = d^T F(h(t))$. We have

$$g''(t) = d^T DF(h(t))Dh(t) = d^T \nabla^2 f(h(t))d = \langle d, \nabla^2 f(x^* + td)d \rangle.$$

In particular,

$$g''(0) = \langle d, \nabla^2 f(x^*) d \rangle.$$

Example 2. Let M be an $n \times n$ matrix and let $b \in \mathbb{R}^n$, and define a function $f : \mathbb{R}^n \to \mathbb{R}$ by $f(x) = x^T M x + b^T x$. Because $f(x) = \langle x, Mx + b \rangle$, we have

$$\nabla f(x) = M^T x + Mx + b = (M^T + M)x + b,$$

and

$$\nabla^2 f(x) = M^T + M.$$

In particular, if M is symmetric then $\nabla f(x) = 2Mx + b$ and $\nabla^2 f(x) = 2M$.

9.3 Matrix functions

We may define matrix-valued functions for sufficiently smooth functions f using Taylor expansions. We will focus on $f : \mathbb{R} \mapsto \mathbb{R}$ such that f is analytic, i.e. it has Taylor series expansion

$$f(x) = \sum_0^\infty \alpha_k x^k / k!, \qquad \alpha_k = f^{(k)}(0)$$

Then for a matrix $A \in \mathbb{R}^{n \times n}$, we define

$$f(A) = \sum_0^\infty \alpha_k A^k / k!$$

Whether this function exists depends on the radius of convergence of the Taylor expansion of f relative to $\|A\|$, where $\|\cdot\|$ is some matrix norm satisfying $\|AB\| \leq \|A\|\|B\|$, such as the Frobenius norm defined elsewhere in these notes.

If A is diagonizable and the series above exists, we have the following much simpler characterization

$$A = U \Lambda U^T \quad \Longrightarrow \quad f(A) = U f(\Lambda) U^T$$

Since Λ is diagonal with entries $\lambda_1 \dots \lambda_n$, we see from the power series expansion that $f(\Lambda)$ is diagonal with entries $f(\lambda_1) \dots f(\lambda_n)$.

Generalizations to infinite-dimensional settings also exist and rely on the spectral theorem. See Lax.

Example: Solution to linear ODE Solutions $x : [0, \infty) \mapsto \mathbb{R}^n$ to the ODE

$$\frac{d}{dt} x(t) = Ax(t) \qquad x(0) = x_0$$

are given by

$$x(t) = e^{tA} x_0$$

where the matrix exponential is defined as above for the function $f(x) = e^{tx}$.

Example: Markov processes The matrix function $e^{tA} = P_t$ is important in the theory of Markov processes on finite state spaces, where the ij term of P_t gives the probability of the process being in state j after t time given that it started in state i. The matrix A is, in this case, is called the infinitessimal generator of the process.

Exercises

9.1 Prove the following properties of vector derivatives:

(a) Let $y = x^T x$, then $\frac{\partial y}{\partial x} = 2x$

(b) Let $y = x^T Ax$, then $\frac{\partial y}{\partial x} = Ax + A^T x$ and if A is symmetric $\frac{\partial y}{\partial x} = 2Ax$.

9.2 The **inverse function theorem** states that for a function $f : \mathbb{R}^n \to \mathbb{R}^n$, the inverse of the Jacobian matrix for f is the Jacobian of f^{-1}:

$$(Df)^{-1} = D(f^{-1}).$$

Now consider the function $f : \mathbb{R}^2 \to \mathbb{R}^2$ that maps from polar (r, θ) to cartesian coordinates (x, y):

$$f(r, \theta) = \begin{bmatrix} r\,cos(\theta) \\ r\,sin(\theta) \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix}.$$

Find Df, then invert the two-by-two matrix to find $\frac{\partial r}{\partial x}$, $\frac{\partial r}{\partial y}$, $\frac{\partial \theta}{\partial x}$, and $\frac{\partial \theta}{\partial y}$.

9.3 Show that if $P_t = e^{tA}$ for $A \in \mathbb{R}^{n \times n}$, then $\{P_t\}_{t \geq 0}$ forms a semigroup. That is, check

- $P_t P_s = P_{t+s}$ for all $t, s \geq 0$
- $P_0 = I$ the identity

Such objects are important in the theory of linear evolution equations and in Markov processes.

10 Computation

This chapter discussions some computational algorithms that are relevant to problems we often encounter in statistics (e.g eigenvector decomposition).

10.1 Power method

Let $A \in \mathbb{R}^{n \times n}$ be a matrix with eigenvalues $|\lambda_1| > |\lambda_2| \geq \ldots, \geq |\lambda_n|$ and corresponding eigenvectors $v_1, \ldots, v_n$. In other words A has a *leading eigenvalue/vector*. The goal of this section is to compute the leading eigenvector $v_1 \in \mathbb{R}^n$.

One way to do this would be to solve for the eigenvalues of A by computing the roots of the polynomial $p(\lambda) = \det(A - \lambda I)$ then finding the eigenvectors. From the discussion in Section 6.2 we know this a bad idea since finding the roots of a polynomial is a very hard problem to solve in general.

One of the most straightforward methods find the leading eigenvector of a matrix is called the *power method*. The power method essentially consists of repeatedly multiplying a random vector w_0 by the matrix A i.e. computing $A^k w_0$. It turns out that as $k \to \infty$ the resulting vector will converge to v_1.

Let $w_0 \in \mathbb{R}^d$ be our starting point (typically selected randomly). Iterate the following for a while

$$\begin{aligned} &\text{for } k = 1, 2, \ldots \\ &\quad w_{k+1} = A w_k \\ &\quad w_{k+1} = \frac{w_{k+1}}{||w_{k+1}||} \end{aligned}$$

We stop after some finite (hopefully small) number of iterations and the let the resulting w_k be our estimate of v_1. Most of the action is in the first line, Aw_k. The second line just normalizes the magnitude of w_k (for example, if $w_k = v_1$ and $\lambda_1 > 1$ then $A^k w_0 = \lambda_1^k v_1$ has a really large magnitude which our computer will not be happy about).

Theorem: Suppose A has a strictly largest eigenvalue and is diagonalizable. Additionally, suppose w_0 is not orthogonal to v_1. Then as $k \to \infty$, $w_k \to \pm v_1$ (i.e. we get either v_1 or $-v_1$). Additionally, $||w_k - v_1|| = O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right)$.

Remark: This theorem not only says the power method converges (under some assumptions) it also gives a convergence rate. We prove this theorem under the assumption that A is diagonalizable (e.g. if A is symmetric we can apply the spectral theorem). The same theorem holds without the diagonalizable assumption; this proof is similar, but uses *Jordan canonical form*.

Proof: Since A is diagonalizable there exists an orthonormal basis of eigenvectors $v_1, \ldots, v_n$. Let $w_0 = \sum_{i=1}^n a_i v_i$ (note $a_1 \neq 0$). The core of the proof is the following

calculation

$$\begin{aligned} A^k w_0 &= \sum_{i=1}^{n} a_i \lambda_i^k v_i \\ &= \lambda_1^k a_1 v_1 + \lambda_1^k \sum_{i=2}^{n} a_i \left(\frac{\lambda_i}{\lambda_1}\right)^k v_i \end{aligned}$$

We are essentially done since the the second term in the above expression dies since $|\frac{\lambda_i}{\lambda_1}| < 1$ by assumption. Note that the power method re-normalizes w_k each time so we only need to show convergence up to a constant times $A^k w_0$.

$$\begin{aligned} ||a_1 v_1 - \frac{1}{\lambda_1^k} A^k w_0||_2 &= ||a_1 v_1 - a_1 v_1 - \sum_{i=2}^{n} a_i \left(\frac{\lambda_i}{\lambda_1}\right)^k v_i|| \\ &= ||\sum_{i=2}^{n} a_i \left(\frac{\lambda_i}{\lambda_1}\right)^k v_i|| \\ &\leq \sum_{i=2}^{n} |a_i| \left|\frac{\lambda_i}{\lambda_1}\right|^k \\ &= \sum_{i=2}^{n} \tilde{a} \left|\frac{\lambda_2}{\lambda_1}\right|^k \\ &= (n-1)\tilde{a} \left|\frac{\lambda_2}{\lambda_1}\right|^k \end{aligned}$$

where $\tilde{a} = \max_{i=2,\ldots,n} |a_i|$. $\square|$.

The power method is the basis of more sophisticated *Krylov subspace* methods for computing eigenvectors. In many cases the power method converges very quickly. Notice, however, the convergence rate is related to the quantity $\frac{\lambda_2}{\lambda_1}$ i.e. if $\lambda_1 >> \lambda_2$ the power method will converge faster.

Squeezing the power method to get more eigenvectors We can use the power method to compute larger eigenvectors. Let $f(A)$ be any algorithm (e.g. the power method) that returns the leading eigenvector of a matrix. Suppose we have a matrix A such that the eigenvalues are strictly decreasing i.e. $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|$. We can find the eigenvectors by repeatedly computing $f(A)$ then projecting the columns A onto the orthogonal complement of each new eigenvector.

Let $A_1 := A$ then compute $v_1, \ldots, v_n$ by

$$\begin{aligned} &\text{for } k = 1, 2, \ldots, n \\ &\quad v_k = f(A_k) \\ &\quad A_{k+1} = A(I - v_k v_k^T) \end{aligned}$$

This is not the state of the art way to compute SVD, but it works. For a recent analysis of this particular method see LazySVD: Even Faster SVD Decomposition Yet Without Agonizing Pain (pdf can be found at https://arxiv.org/abs/1607.03463). For a more general

treatment of the power method see Chapter 7 from http://people.inf.ethz.ch/arbenz/ewp/Lnotes/lsevp.pdf.

10.2 Gradient Descent

Let $F : \mathbb{R}^d \to \mathbb{R}$ be a function and consider the problem of *unconstrained minimization* i.e. find $x^*\mathbb{R}^d$ that minimizes F

$$\min_{x\mathbb{R}^d} F(x)$$

For example, if $F(x) = x^2+bx+c$ you can solve this problem easily. If $F(\beta) = ||X\beta-y||_2$ where now $\beta\mathbb{R}^d$ is the variable and $X \in \mathbb{R}^{n\times d}$, $y \in \mathbb{R}^n$ then $\beta^* = (X^TX)^{-1}X^Ty$ is a solution if X^TX is invertible (this is linear regression).

In general, if we make no assumptions about F then minimizing it is really hard.

11 Statistics: Random Variables

This sections covers some basic properties of random variables. While this material is not necessarily tied directly to linear algebra, it is essential background for graduate level Statistics, O.R., and Biostatistics. For further review of these concepts, see Casella and Berger, sections 2.1, 2.2, 2.3, 3.1, 3.2, 3.3, 4.1, 4.2, 4.5, and 4.6.

Much of this section is gratefully adapted from Andrew Nobel's lecture notes.

11.1 Expectation, Variance and Covariance

Expectation The **expected value** of a continuous random variable X, with probability density function f, is defined by

$$\mathbb{E}X = \int_{-\infty}^{\infty} xf(x)dx.$$

The expected value of a discrete random variable X, with probability mass function p, is defined by

$$\mathbb{E}X = \sum_{x\in\mathbb{R}, p(x)\neq 0} xp(x).$$

The expected value is **well-defined** if $\mathbb{E}|X| < \infty$.

We now list some basic properties of $E(\cdot)$:

1. $X \leq Y$ imples $\mathbb{E}X \leq \mathbb{E}Y$

   Proof: Follows directly from properties of $\int$ and $\sum$.

2. For $a, b \in \mathbb{R}$, $\mathbb{E}(aX + bY) = a\mathbb{E}X + b\mathbb{E}Y$.

   Proof: Follows directly from properties of $\int$ and $\sum$.

3. $|\mathbb{E}X| \leq \mathbb{E}|X|$

   Proof: Note that $X, -X \leq |X|$. Hence, $\mathbb{E}X, -\mathbb{E}X \leq \mathbb{E}|X|$ and therefore $|\mathbb{E}X| \leq \mathbb{E}|X|$. □

4. If X and Y are independent ($X \perp Y$), then $\mathbb{E}(XY) = \mathbb{E}X \cdot \mathbb{E}Y$.

   Proof: See Theorem 4.2.10 in Casella and Berger.

5. If X is a non-negative continuous random variable, then

$$\mathbb{E}X = \int_0^{\infty} P(X \geq t)dt = \int_0^{\infty} (1 - F(t))dt.$$

Proof: Suppose $X \sim f$. Then,

$$\begin{aligned}
\int_0^\infty P(X > t)dt &= \int_0^\infty [\int_t^\infty f(x)dx]dt \\
&= \int_0^\infty [\int_0^\infty f(x)I(x > t)dx]dt \\
&= \int_0^\infty \int_0^\infty f(x)I(x > t)dtdx \quad \text{(Fubini)} \\
&= \int_0^\infty f(x)[\int_0^\infty I(x > t)dt]dx \\
&= \int_0^\infty xf(x)dx = \mathbb{E}X \qquad \square
\end{aligned}$$

6. If $X \sim f$, then $\mathbb{E}g(X) = \int g(x)f(x)dx$.
   If $X \sim p$, then $\mathbb{E}g(X) = \sum_x g(x)p(x)$.

   Proof: Follows from definition of $Eg(X)$.

Variance and Corvariance The **variance** of a random variable X is defined by

$$\begin{aligned}
\text{Var}(X) &= \mathbb{E}(X - \mathbb{E}X)^2 \\
&= \mathbb{E}X^2 - (\mathbb{E}X)^2.
\end{aligned}$$

Note that $\text{Var}(X)$ is finite (and therefore well-defined) if $\mathbb{E}X^2 < \infty$. The **covariance** of two random variables X and Y is defined by

$$\begin{aligned}
\text{Cov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)] \\
&= \mathbb{E}(XY) - \mathbb{E}X\mathbb{E}Y.
\end{aligned}$$

Note that $\text{Cov}(X, Y)$ is finite if $\mathbb{E}X^2, \mathbb{E}Y^2 < \infty$.

We now list some general properties, that follow from the definition of variance and covariance:

1. $\text{Var}(X) \geq 0$, with "=" if and only if X is constant with probability 1.
2. For $a, b \in \mathbb{R}$, $\text{Var}(aX + b) = a^2\text{Var}(X)$.
3. If $X \perp Y$, then $\text{Cov}(X, Y) = 0$. The converse, however, is not true in general.
4. $\text{Cov}(aX + b, cY + d) = ac\text{Cov}(X, Y)$.
5. If $X_1, \ldots, X_n$ satisfy $\mathbb{E}X_i^2 < \infty$, then

$$\text{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \text{Var}(X_i) + 2\sum_{i<j} \text{Cov}(X_i, X_j).$$

11.2 Distribution of Functions of Random Variables

Here we describe various methods to calculate the distribution of a function of one or more random variables.

CDF method For the single variable case, given $X \sim f_X$ and $g : \mathbb{R} \to \mathbb{R}$ we would like to find the density of $Y = g(X)$, if it exists. A straightforward approach is the **CDF method**:

- Find F_Y in terms of F_X
- Differentiate F_Y to get f_Y

Example 1: Location and scale. Let $X \sim f_X$ and $Y = aX + b$, with $a > 0$. Then,

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(aX + b \leq y) = P(X \leq \frac{y-b}{a}) \\ &= F_X(\frac{y-b}{a}). \end{aligned}$$

Thus, $f_Y(y) = F_Y'(y) = a^{-1} f_X(\frac{y-b}{a})$.

If $a < 0$, a similar argument shows $f_Y(y) = |a|^{-1} f(\frac{y-b}{a})$.

Example 2 If $X \sim \mathbb{N}(0,1)$ and Y = aX + b, then

$$\begin{aligned} f_Y(y) &= |a|^{-1} \phi(\frac{y-b}{a}) \\ &= \frac{1}{\sqrt{2\pi a^2}} \exp\left\{ -\frac{(y-b)^2}{2a^2} \right\} \\ &= \mathbb{N}(b, a^2). \end{aligned}$$

Example 3 Suppose $X \sim \mathbb{N}(0,1)$. Let $Z = X^2$. Then,

$$\begin{aligned} F_Z(z) &= P(Z \leq z) = P(X^2 \leq z) = P(-\sqrt{z} \leq X \leq \sqrt{z}) \\ &= \Phi(\sqrt{z}) - \Phi(-\sqrt{z}) = 1 - 2\Phi(-\sqrt{z}). \end{aligned}$$

Thus, $f_Z(z) = z^{-1/2} \phi(-\sqrt{z}) = \frac{1}{\sqrt{2\pi}} z^{-1/2} e^{-z/2}$.

Convolutions The **convolution** $f = f_1 * f_2$ of two densities f_1 and f_2 is defined by

$$f(x) = \int_{-\infty}^{\infty} f_1(x-y) f_2(y) dy.$$

Note that $f(x) \geq 0$, and

$$\begin{aligned} \int_{-\infty}^{\infty} f(x) dx &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} f_1(x-y) f_2(y) dy \right] dx \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_1(x-y) f_2(y) d_x d_y \\ &= \int_{-\infty}^{\infty} f_2(y) \left[\int_{-\infty}^{\infty} f_1(x-y) dx \right] dy = \int_{-\infty}^{\infty} f_2(y) dy = 1. \end{aligned}$$

So, $f = f_1 * f_2$ is a density.

Theorem: If $X \sim f_X$, and $Y \sim f_Y$ and X and Y are independent, then $X + Y \sim f_X * f_Y$.

Proof: Note that

$$\begin{aligned}
P(X+Y \le v) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_X(x)f_Y(y)I\{(x,y) : x+y \le v\}dxdy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{v-y} f_X(x)f_Y(y)dxdy \\
&= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{v-y} f_X(x)dx\right] f_Y(y)dy \\
&= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{v} f_X(u-y)du\right] f_Y(y)dy \qquad (u = y+x) \\
&= \int_{-\infty}^{v}\left[\int_{-\infty}^{\infty} f_X(u-y)f_Y(y)dy\right] du \\
&= \int_{-\infty}^{v} (f_X * f_Y)(u)du. \qquad \square
\end{aligned}$$

Corollary: Convolutions are commutative and associative. If f_1, f_2, f_3 are densities, then

$$\begin{aligned}
f_1 * f_2 &= f_2 * f_1 \\
(f_1 * f_2) * f_3 &= f_1 * (f_2 * f_3).
\end{aligned}$$

Change of Variables We now consider functions of more than one random variable. In particular, let U, V be open subsets in $\mathbb{R}^k$, and $H : U \to V$. Then, if $\vec{x}$ is a vector in U,

$$H(\vec{x}) = (h_1(\vec{x}), \ldots, h_k(\vec{x}))^t.$$

is a vector in V. The functions $h_1(\cdot), \ldots, h_k(\cdot)$ are the **coordinate functions** of H. If $\vec{X}$ is a continuous random vector, we would like to find the density of $H(\vec{X})$. First, some further assumptions:

(A1) $H : U \to V$ is one-to-one and onto.

(A2) H is continuous.

(A3) For every $1 \le i, j \le k$, the partial derivatives

$$h'_{ij} \equiv \frac{\partial h_i}{\partial x_j}$$

exist and are continuous.

Let $D_H(\vec{x})$ be the matrix of partial derivatives of H:

$$D_H(\vec{x}) = [h'_{ij}(\vec{x}) : 1 \leq i, j \leq k].$$

Then, the **Jacobian** (or **Jacobian determinant**[7]) of H at $\vec{x}$ is the determinant of $D_H(\vec{x})$:

$$J_H(\vec{x}) = \det(D_H(\vec{x})).$$

The assumptions **A1-3** imply that $H^{-1} : V \to U$ exists and is differentiable on V with

$$J_{H^{-1}}(\vec{y}) = (J_H(H^{-1}(y)))^{-1}.$$

Theorem: Suppose $J_H(\vec{x}) \neq 0$ on U. If $\vec{X} \sim f_{\vec{X}}$ is a k-dimensional random vector such that $P(\vec{X} \in U) = 1$, then $\vec{Y} = H(\vec{X})$ has density

$$\begin{aligned} f_{\vec{Y}}(\vec{y}) &= f_{\vec{X}}(H^{-1}(\vec{y})) \cdot |J_{H^{-1}}(\vec{y})| \\ &= f_{\vec{X}}(H^{-1}(\vec{y})) \cdot |J_H(H^{-1}(\vec{y}))|^{-1}. \end{aligned}$$

Example: Suppose X_1, X_2 are jointly continuous with density f_{X_1,X_2}. Let $Y_1 = X_1 + X_2$, $Y_2 = X_1 - X_2$, and find f_{Y_1,Y_2}.

Here

$$\begin{aligned} y_1 = h_1(x_1, x_2) &= x_1 + x_2 \\ y_2 = h_2(x_1, x_2) &= x_1 - x_2 \\ x_1 = g_1(y_1, y_2) &= \frac{1}{2}(y_1 + y_2) \\ x_2 = g_2(y_1, y_2) &= \frac{1}{2}(y_1 - y_2), \end{aligned}$$

and

$$J_H(x_1, x_2) = \begin{vmatrix} \frac{\partial h_1}{\partial x_1} & \frac{\partial h_1}{\partial x_2} \\ \frac{\partial h_2}{\partial x_1} & \frac{\partial h_2}{\partial x_2} \end{vmatrix} = \begin{vmatrix} 1 & 1 \\ 1 & -1 \end{vmatrix} = -2 \neq 0.$$

So, applying the theorem, we get

$$f_{Y_1,Y_2}(y_1, y_2) = \frac{1}{2} f_{X_1,X_2}(\frac{y_1 + y_2}{2}, \frac{y_1 - y_2}{2}).$$

As a special case, assume X_1, X_2 are $\mathbb{N}(0, 1)$ and independent. Then,

$$\begin{aligned} f_{Y_1,Y_2}(y_1, y_2) &= \frac{1}{2}\phi(\frac{y_1 + y_2}{2})\phi(\frac{y_1 - y_2}{2}) \\ &= \frac{1}{4\pi}\exp\left\{-\frac{(y_1 + y_2)^2}{8} - \frac{(y_1 - y_2)^2}{8}\right\} \\ &= \frac{1}{4\pi}\exp\left\{-\frac{2y_1^2 + 2y_2^2}{8}\right\} \\ &= \frac{1}{4\pi}\exp\left\{-\frac{y_1^2}{4}\right\}\exp\left\{-\frac{y_2^2}{4}\right\}. \end{aligned}$$

[7]The partial derivative matrix D is sometimes called the **Jacobain matrix** (see Section 9.2).

So, both Y_1 and Y_2 are $\mathbb{N}(0, 2)$, and they are independent!

11.3 Derivation of Common Univariate Distributions

Double Exponential If $X_1, X_2 \sim \text{Exp}(\lambda)$ and $X_1 \perp X_2$, then $X_1 - X_2$ has a **double exponential** (or **Laplace**) distribution: $X_1 - X_2 \sim \text{DE}(\lambda)$. The density of $\text{DE}(\lambda)$,

$$f(x) = \frac{\lambda}{2} e^{-\lambda|x|} \qquad -\infty < x < \infty,$$

can be derived through the convolution formula.

Gamma and Beta Distributions The **gamma function**, a component in several probability distributions, is defined by

$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx, \qquad t > 0.$$

Here are some basic properties of $\Gamma(\cdot)$:

1. $\Gamma(t)$ is well-defined for $t > 0$.
   Proof: For $t > 0$,
   $$0 \le \Gamma(t) \le \int_0^1 x^{t-1} dx + \int_1^\infty x^{t-1} e^{-x} dx < \infty. \qquad \square$$
2. $\Gamma(1) = 1$.
   Proof: Clear.
3. $\forall x > 0$, $\Gamma(x + 1) = x\Gamma(x)$.
   Proof: Exercise 7.4.
4. $\Gamma(n + 1) = n!$ for $n = 0, 1, 2, ...$
   Proof: Follows from $2, 3$.
5. $\log \Gamma(\cdot)$ is convex on $[0, \infty)$.

The **gamma distribution** with parameters $\alpha, \beta > 0$, $\Gamma(\alpha, \beta)$, has density

$$g_{\alpha,\beta}(x) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}, \qquad x > 0.$$

Note: A basic change of variables shows that for $s > 0$,

$$X \sim \Gamma(\alpha, \beta) \iff sX \sim \Gamma\left(\alpha, \frac{\beta}{s}\right).$$

So, β acts as a scale parameter of the $\Gamma(\alpha, \cdot)$ family. The parameter α controls shape:

- If $0 < \alpha < 1$, then $g_{\alpha,\beta}(\cdot)$ is convex and $g_{\alpha,\beta} \uparrow \infty$ as $x \to 0$.
- If $\alpha > 1$, then $g_{\alpha,\beta}(\cdot)$ is unimodal, with maximum at $x = \frac{\alpha-1}{\beta}$.

If $X \sim \Gamma(\alpha, \beta)$, then $\mathbb{E}X = \frac{\alpha}{\beta}$, $\mathrm{Var}(X) = \frac{\alpha}{\beta^2}$.

We now use convolutions to show that if $X \sim \Gamma(\alpha_1, \beta), Y \sim \Gamma(\alpha_2, \beta)$ are independent then $X + Y \sim \Gamma(\alpha_1 + \alpha_2, \beta)$:

Theorem: The family of distributions $\{\Gamma(\cdot, \beta)\}$ is closed under convolutions. In particular

$$\Gamma(\alpha_1, \beta) * \Gamma(\alpha_2, \beta) = \Gamma(\alpha_1 + \alpha_2, \beta).$$

Proof: For $x > 0$,

$$\begin{aligned} f(x) &= (g_{\alpha_1,\beta} * g_{\alpha_2,\beta})(x) \\ &= \int_0^x g_{\alpha_1,\beta}(x-u) g_{\alpha_2,\beta}(u) du \\ &= \frac{\beta^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} e^{-\beta x} \int_0^x (x-u)^{\alpha_1-1} u^{\alpha_2-1} du \qquad (6) \\ &= \text{const} \cdot e^{-\beta x} x^{\alpha_1+\alpha_2-1} \end{aligned}$$

Thus, $f(x)$ and $g_{\alpha_1+\alpha_2,\beta}(x)$ agree up to constants. As both integrate to 1, they are the same function. □

Corollary: Note if $\alpha = 1$, then $\Gamma(1, \beta) = \mathrm{Exp}(\beta)$. Hence, If $X_1, \ldots, X_n$ are iid $\sim \mathrm{Exp}(\lambda)$, then

$$Y = X_1 + \ldots + X_n \sim \Gamma(n, \lambda),$$

with density

$$f_Y(y) = \frac{\lambda^n y^{n-1} e^{-\lambda y}}{(n-1)!}.$$

This is also known as an **Erlang** distribution with parameters n and λ.

It follow from equation (6), with $x = 1$ that

$$\begin{aligned} \frac{\beta^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} e^{-\beta} \int_0^1 (1-u)^{\alpha_1-1} u^{\alpha_2-1} du \\ = g_{\alpha_1+\alpha_2,\beta}(1) = \frac{\beta^{\alpha_1+\alpha_2} e^{-\beta}}{\Gamma(\alpha_1+\alpha_2)}. \end{aligned}$$

Rearranging terms shows that for $r, s > 0$,

$$B(r, s) = \frac{\Gamma(r)\Gamma(s)}{\Gamma(r+s)} = \int_0^1 (1-u)^{r-1} u^{s-1} du.$$

Here $B(\cdot, \cdot)$ is known as the **beta function** with parameters r, s. The **beta distribution** $Beta(r, s)$ has density

$$b_{r,s}(x) = B(r, s)^{-1} \cdot x^{r-1}(1-x)^{s-1}, \qquad 0 < x < 1.$$

The parameters r, s play symmetric roles. If $r = s$ then $Beta(r, s)$ is symmetric about $1/2$. $Beta(r, r)$ is u-shaped if $r < 1$, uniform if $r = 1$, and unimodal (bell shaped) if $r > 1$. If $r > s > 0$ then $Beta(r, s)$ is skewed to the right, if $0 < s < r$ then $Beta(r, s)$ is skewed left. The random variable $X \sim Beta(r, s)$ has expection and variance

$$\mathbb{E}X = \frac{r}{r+s}, \qquad \mathrm{Var}(X) = \frac{rs}{(r+s)^2(r+s+1)}.$$

Chi-square distributions Fix an integer $k \geq 1$. Then, the chi-square distribution with k degrees of freedom, written χ^2_k, is $\Gamma(k/2, 1/2)$. Thus, χ^2_k has density

$$f_k(x) = \frac{1}{2^{k/2}\Gamma(\frac{k}{2})} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}, \qquad x > 0.$$

Theorem: If $X_1, \ldots, X_k$ are iid $\mathbb{N}(0, 1)$, then $X_1^2 + \ldots + X_k^2 \sim \chi^2_k$.

Proof: Recall that if $X \sim \mathbb{N}(0, 1)$ then $X^2 \sim f(x) = \frac{1}{2\sqrt{\pi}} e^{-\frac{x}{2}} = \Gamma(\frac{1}{2}, \frac{1}{2})$. Thus, $X^2 \sim \chi^2_1$. Furthermore,

$$X_1^2 + \ldots + X_k^2 \sim \Gamma\left(\frac{k}{2}, \frac{1}{2}\right) = \chi^2_k. \qquad \square$$

If $Y = X_1^2 + \ldots + X_k^2 \sim \chi^2_k$, then

$$\mathbb{E}Y = \mathbb{E}(X_1^2 + \ldots + X_k^2) = k\mathbb{E}X_1^2 = k.$$

$$\begin{aligned} \mathrm{Var}(Y) &= k\mathrm{Var}(X_1^2) = k(\mathbb{E}X_1^4 - (\mathbb{E}X_1^2)^2) \\ &= k(3-1) = 2k. \end{aligned}$$

F and t-distributions The **F-distribution** with with m,n degrees of freedom, $F(m, n)$, is the distribution of the ratio

$$\frac{X/m}{Y/n},$$

where $X \sim \chi^2_m$, $Y \sim \chi^2_n$, and $X \perp Y$.

The density of $F(m, n)$ is

$$f_{m,n}(x) = B^{-1}\left(\frac{m}{2}, \frac{n}{2}\right)\left(\frac{m}{n}\right)^{\frac{m}{2}} x^{\frac{m}{2}-1}\left(1 + \frac{m}{n}x\right)^{-\frac{1}{2}(m+n)}.$$

The **t-distribution** with n degrees of freedom, t_n, is the distribution of the ratio

$$\frac{X}{\sqrt{Y^2/n}},$$

where $X \sim \mathbb{N}(0,1), Y^2 \sim \chi^2_n$ are independent. Equivalently, t_n is the distribution of $\sqrt{Z}$ where $Z \sim F(1,n)$. The density of t_n is

$$f_n(t) = \frac{1}{nB\left(\frac{1}{2}, \frac{n}{2}\right)} \cdot \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}.$$

Some other properties of the t-distribution:

1. t_1 is the **Cauchy distribution**.
2. If $X \sim t_n$ then $\mathbb{E}X = 0$ for $n \geq 2$, undefined for $n = 1$; $\text{Var}(X) = \frac{n}{n-2}$ for $n \geq 3$, undefined for $n = 1, 2$.
3. The density $f_n(t)$ converges to the density of a standard normal, $\phi(t)$, as $n \to \infty$.

11.4 Random Vectors: Expectation and Variance

A **random vector** is a vector $X = [X_1\, X_2 \ldots X_k]^T$ whose components $X_1, X_2, ..., X_k$ are real-valued random variables defined on the same probability space. The expectation of a random vector $\mathbb{E}(X)$, if it exists, is given by the expected value of each component:

$$\mathbb{E}(X) = [\mathbb{E}X_1\, \mathbb{E}X_2 \ldots \mathbb{E}X_k]^T.$$

The covariance matrix of a random vector $Cov(X)$ is given by

$$Cov(X) = \mathbb{E}[(X - \mathbb{E}X)(X - \mathbb{E}X)^T] = \mathbb{E}(XX^T) - \mathbb{E}X\mathbb{E}X^T.$$

We now give some general results on expectations and variances. We supply reasonings for some of them, and you should verify the rest (usually by the method of entry-by-entry comparison). We assume in what follows that $k \times k$ A and $k \times 1$ a are constant, and we let $k \times 1$ $\mu = \mathbb{E}(X)$ and $k \times k$ $V = Cov(X)$ ($v_{ij} = Cov(X_i, X_j)$):

1. $\mathbb{E}(AX) = A\mathbb{E}(X)$

   Proof: Exercise 7.5(a).

2. $Var(a^T X) = a^T V a$.

   Proof: Note that

$$\begin{aligned} var(a^T X) &= var(a_1 X_1 + a_2 X_2 + \ldots + a_k X_k) \\ &= \sum_{i=1}^{k} \sum_{j=1}^{k} a_i a_j Cov(X_i X_j) \\ &= \sum_{i=1}^{k} \sum_{j=1}^{k} v_{ij} a_i a_j = a^T V a \quad \square \end{aligned}$$

3. $Cov(AX) = AVA^T$

   Proof Exercise 7.5(b).

4. $\mathbb{E}(X^TAX) = tr(AV) + \mu^TA\mu$

   Proof: Let A_i be the ith row of A and a_{ij} be the ijth entry of A.
   Note that $tr(AV) = tr(A(\mathbb{E}(XX^T) - \mathbb{E}X\mathbb{E}X^T)) = tr(A\mathbb{E}(XX^T)) - tr(A\mathbb{E}X\mathbb{E}X^T)$.

$$\begin{aligned} tr(A\mathbb{E}(XX^T)) &= tr\left(\begin{pmatrix} A_1\mathbb{E}(XX^T) \\ \vdots \\ A_k\mathbb{E}(XX^T) \end{pmatrix}\right) \\ &= \sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij}\mathbb{E}(X_jX_i) \\ &= \mathbb{E}\left(\sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij}X_jX_i\right) \\ &= \mathbb{E}\left(\sum_{i=1}^{k} A_iXX_i\right) \\ &= \mathbb{E}\left(\left(\sum_{i=1}^{k} X_iA_i\right)X\right) \\ &= \mathbb{E}(X^TAX) \end{aligned}$$

   Meanwhile,

$$tr(A\mathbb{E}X\mathbb{E}X^T) = tr(\mathbb{E}X^TA\mathbb{E}X) = \mathbb{E}X^TA\mathbb{E}X = \mu^TA\mu.$$

   So we have $\mathbb{E}(X^TAX) = tr(AV) + \mu^TA\mu$. $\square$

5. Covariance matrix V is positive semi-definite.

   Proof: $y^TVy = Var(y^TX) \geq 0\ \forall y \neq 0$. Since V is symmetric (why?), it follows that $V^{1/2} = (V^{1/2})^T$. $\square$

6. $Cov(a^TX, b^TX) = a^TVb$

   Proof: Exercise 7.5(c).

7. If X, Y are two $k \times 1$ vectors of random variables, we define their **cross-covariance** matrix C as follows : $c_{ij} = Cov(X_i, Y_j)$. Notice that unlike usual covariance matrices, a cross-covariance matrix is not (usually) symmetric. We still use the notation $Cov(X, Y)$ and the meaning should be clear from the context. Now, suppose A, B are $k \times k$. Then $Cov(AX, BX) = AVB^T$.

   Proof: Let c_{ij} be the ijth entry of $Cov(AX, BX)$. Denote the ith row vectors of A and B as A_i and B_i, respectively. By the result above,

$$c_{ij} = A_iVB_j = ij\text{th entry of } AVB^T. \quad \square$$

Exercises

11.1 Show that if $X \sim f$ and $g(\cdot)$ is non-negative, then $Eg(X) = \int_{-\infty}^{\infty} g(x)f(x)dx$. [Hint: Recall that $EX = \int_0^{\infty} P(X > t)dt$ if $X \geq 0$.]

11.2 Let X be a continuous random variable with density f_X. Find the density of $Y = |X|$ in terms of f_X.

11.3 Let $X_1 \sim \Gamma(\alpha_1, 1)$ and $X_2 \sim \Gamma(\alpha_2, 1)$ be indepedent. Use the two-dimensional change of variables formula to show that $Y_1 = X_1 + X_2$ and $Y_2 = X_1/(X_1 + X_2)$ are independent with $Y_1 \sim \Gamma(\alpha_1 + \alpha_2, 1)$ and $Y_2 \sim Beta(\alpha_1, \alpha_2)$.

11.4 Using integration by parts, show that the gamma function $\Gamma(t) = \int_0^{\infty} x^{t-1}e^{-x}dx$ satisfies the relation $\Gamma(t + 1) = t\Gamma(t)$ for $t > 0$.

11.5 Prove the following results about vector expectations and variance:

(a) $E(Ax) = AE(x)$

(b) $Cov(Ax) = AVA^T$

(c) $Cov(a^T x, b^T x) = a^T V b$

12 Further Applications to Statistics: Normal Theory and F-test

12.1 Bivariate Normal Distribution

Suppose X is a vector of continuous random variables and $Y = AX + c$, where A is an invertible matrix and c is a constant vector. If X has probability density function f_X, then the probability density function of Y is given by

$$f_Y(y) = |det(A)|^{-1} f_X(A^{-1}(Y - c)).$$

The proof of this result can be found in appendix B.2.1 of Bickel and Doksum.

We say that 2×1 vector $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ has a **bivariate normal** distribution if $\exists$ Z_1, Z_2 I.I.D $N(0,1)$, s.t. $X = AZ + \mu$. In what follows we will moreover assume that A is invertible. You should check at this point for yourself that $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$, where $\sigma_1 = \sqrt{{a_{11}}^2 + {a_{12}}^2}$ and $\sigma_2 = \sqrt{{a_{21}}^2 + {a_{22}}^2}$, and that $Cov(X_1, X_2) = a_{11}a_{21} + a_{12}a_{22}$. We then say that $X \sim N(\mu, \Sigma)$, where

$$\Sigma = AA^T = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

and $\rho = \frac{Cov(X_1, X_2)}{\sigma_1\sigma_2}$ (you should verify that the entries of $\Sigma = AA^T$ are as we claim). The meaning behind this definition is made explicit by the following theorem:

Theorem: Suppose $\sigma_1 \neq 0 \neq \sigma_2$ and $|\rho| < 1$. Then

$$f_X(x) = \frac{1}{2\pi\sqrt{det(\Sigma)}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}.$$

Proof Note first of all that if A is invertible, then it follows directly that $\sigma_1 \neq 0 \neq \sigma_2$ and $|\rho| < 1$ (why?). Also,

$$\sqrt{det(\Sigma)} = \sqrt{det(AA^T)} = \sqrt{det(A)^2} = |det(A)| = \sigma_1\sigma_2\sqrt{1-\rho^2}$$

(you should verify the last step). We know that $f_Z(z) = \frac{1}{2\pi}\exp\left(-\frac{1}{2}z^Tz\right)$ and since $X = AZ + \mu$ we have by the result above:

$$\begin{aligned} f_X(x) &= \frac{1}{2\pi|det(A)|}\exp\left(-\frac{1}{2}(A^{-1}(x-\mu))^T(A^{-1}(x-\mu))\right) \\ &= \frac{1}{2\pi|det(A)|}\exp\left(-\frac{1}{2}(x-\mu)^T(A^{-1})^T(A^{-1})(x-\mu)\right) \\ &= \frac{1}{2\pi|det(A)|}\exp\left(-\frac{1}{2}(x-\mu)^T(AA^T)^{-1}(x-\mu)\right) \\ &= \frac{1}{2\pi\sqrt{det(\Sigma)}}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \end{aligned}$$

which proves the theorem. The symmetric matrix Σ is the covariance matrix of X. $\square$

You should prove for yourself (Exercise 8.1) that if X has a bivariate normal distribution $N(\mu, V)$, and B is invertible, then $Y = BX + d$ has a bivariate normal distribution $N(B\mu + d, BVB^T)$.

These results generalize to more than two variables and lead to multivariate normal distributions. You can familiarize yourself with some of the extensions in appendix B.6 of Bickel and Doksum. In particular,we note here that if x is a $k \times 1$ vector of IID $N(0, \sigma^2)$ random variables, then Ax is distributed as a multivariate $N(0, \sigma^2 AA^T)$ random vector.

12.2 F-test

We will need a couple more results about quadratic forms:

1. Suppose $k \times k$ A is symmetric and idempotent and $k \times 1$ $x \sim N(0_{k\times 1}, \sigma^2 I_{k\times k})$. Then $\frac{x^T Ax}{\sigma^2} \sim {\chi_r}^2$, where $r = rank(A)$.

   Proof: We write $\frac{x^T Ax}{\sigma^2} = \frac{x^T Q}{\sigma}\Lambda\frac{Q^T x}{\sigma}$ and we note that $\frac{Q^T x}{\sigma} \sim N(0, \frac{1}{\sigma^2} \times \sigma^2 Q^T Q) = N(0, I)$, i.e.$\frac{Q^T x}{\sigma}$ is a vector of IID $N(0, 1)$ random variables. We also know that the Λ is diagonal and its main diagonal consist of r 1's and $k - r$ 0's, where $r = rank(A)$. You can then easily see from matrix multiplication that $\frac{x'Q}{\sigma}\Lambda\frac{Q'x}{\sigma} = {z_1}^2 + {z_2}^2 + \ldots + {z_r}^2$, where the z_i's are IID $N(0, 1)$. Therefore $\frac{x'Ax}{\sigma^2} \sim {\chi_r}^2$. $\square$

2. The above result generalizes further: suppose $k \times 1$ $x \sim N(0, V)$, and $k \times k$ symmetric A is s.t. V is positive definite and either AV or VA is idempotent. Then $x'Ax \sim {\chi_r}^2$, where $r = rank(AV)$ or $rank(VA)$, respectively.

   Proof: We will prove it for the case of idempotent AV and the proof for idempotent VA is essentially the same. We know that $x \sim V^{1/2}z$, where $z \sim N(0, I_{k\times k})$, and we know that $V^{1/2} = (V^{1/2})'$, so we have: $x'Ax = z'(V^{1/2})'AV^{1/2}z = z'V^{1/2}AV^{1/2}z$. Consider $B = V^{1/2}AV^{1/2}$. B is symmetric, and $B^2 = V^{1/2}AV^{1/2}V^{1/2}AV^{1/2} = V^{1/2}AVAVV^{-1/2} = V^{1/2}AVV^{-1/2} = V^{1/2}AV^{1/2} = B$, so B is also idempotent. Then from the previous result (with $\sigma = 1$), we have $z'Bz \sim {\chi_r}^2$, and therefore $x'Ax \sim {\chi_r}^2$, where $rrank(B) = rank(V^{1/2}AV^{1/2})$. It is a good exercise (Exercise 8.2) to show that $rank(B) = rank(AV)$. $\square$

3. Let $U = x'Ax$ and $V = x'Bx$. Then the two quadratic forms are independent (in the probabilistic sense of the word) if $AVB = 0$. We will not prove this result, but we will use it.

Recall (Section 4.2) that we had a model $Y = X\beta + \epsilon$, where Y is $n \times 1$ vector of observations, X is $n \times p$ matrix of explanatory variables (with linearly independent columns), β is $p \times 1$ vector of coefficients that we're interested in estimating, and ϵ is $n \times 1$ vector of error terms with $E(\epsilon) = 0$. Recall that we estimate $\hat{\beta} = (X'X)^{-1}X'Y$, and we denote fitted values $\hat{Y} = X\hat{\beta} = HY$, where the hat matrix $H = X(X'X)^{-1}X'$ is the projection matrix onto columns of X, and $e = Y - \hat{Y} = (I - H)Y$ is the vector of residuals. Recall also that $X'e = 0$. Suppose now that $\epsilon \sim N(0, \sigma^2 I)$, i.e. the errors are IID $N(0, \sigma^2)$ random variables. Then we can derive some very useful distributional results:

1. $\hat{Y} \sim N(X\beta, \sigma^2 H)$.

   Proof: Clearly, $Y \sim N(X\beta, \sigma^2 I)$, and $\hat{Y} = HY \implies \hat{Y} \sim N(HX\beta, H\sigma^2 IH') = N(X(X'X)^{-1}X'X\beta, \sigma^2 HH') = N(X\beta, \sigma^2 H)$. $\square$

2. $e \sim N(0, \sigma^2(I - H))$.

   Proof: Analagous to 1.

3. $\hat{Y}$ and e are independent (in probabilistic sense of the word).

   Proof: $Cov(\hat{Y}, e) = Cov(HY, (I - H)Y) = H(var(Y))(I - H) = H\sigma^2 I(I - H) = \sigma^2 H(I - H) = 0$. And since both vectors were normally distributed, zero correlation implies independence. Notice that Cov above referred to the cross-covariance matrix. $\square$

4. $\frac{\|e\|^2}{\sigma^2} \sim {\chi^2}_{n-p}$.

   Proof: First notice that $e = (I - H)Y = (I - H)(X\beta + \epsilon) = (I - H)\epsilon$ (why?). Now,

$$\frac{\|e\|^2}{\sigma^2} = \frac{e'e}{\sigma^2} = \frac{\epsilon'(I - H)'(I - H)\epsilon}{\sigma^2} = \frac{\epsilon'(I - H)\epsilon}{\sigma^2}$$

   Since $(I - H)$ is symmetric and idempotent, and $\epsilon \sim N(0, \sigma^2 I)$, by one of the above results we have $\frac{\epsilon'(I-H)\epsilon}{\sigma^2} \sim {\chi_r}^2$, where $r = \text{rank}(I - H)$. But we know (why?) that

$$\begin{aligned}\text{rank}(I - H) &= \text{tr}(I - H) = \text{tr}(I - X(X'X)^{-1}X') \\ &= \text{tr}(I) - \text{tr}(X(X'X)^{-1}X') = n - \text{tr}(X'X(X'X)^{-1}) \\ &= n - tr(I_{p\times p}) = n - p\end{aligned}$$

   So we have $\frac{\|e\|^2}{\sigma^2} \sim {\chi^2}_{n-p}$, and in particular $E(\frac{\|e\|^2}{n-p}) = \sigma^2$. $\square$

Before we introduce the F-test, we are going to establish one fact about partitioned matrices. Suppose we partition $X = [X_1 \ X_2]$. Then $[X_1 \ X_2] = X(X'X)^{-1}X'[X_1 \ X_2] \implies X_1 = X(X'X)^{-1}X'X_1$ and $X_2 = X(X'X)X'X_2$ (by straightforward matrix multiplicaiton) or $HX_1 = X_1$ and $HX_2 = X_2$. Taking transposes we also obtain $X_1^T = X_1^T X(X^T X)^{-1}X^T$ and $X_2^T = X_2^T X(X^T X)^{-1}X'$. Now suppose we want to test a theory that the last p_2 coefficients of β are actually zero (note that if we're interested in coefficients scattered throught β, we can just re-arrange the columns of X). In other words, splitting our system into $Y = X_1\beta_1 + X_2\beta_2 + \epsilon$, with $n \times p_1$ X_1 and $n \times p_2$ X_2 ($p_1 + p_2 = p$), we want to see if $\beta_2 = 0$.

We consider the test statistic

$$\frac{\|\hat{Y}_f\|^2 - \|\hat{Y}_r\|^2}{\sigma^2} = \frac{Y^T(X(X^T X)^{-1}X^T - X_1(X_1^T X_1)^{-1}X_1^T)Y}{\sigma^2},$$

where $\hat{Y}_f$ is the vector of fitted values when we regress with respect to all columns of X (full system), and $\hat{Y}_r$ is the vector of fited values when we regress with respect to only first p_1

columns of X (restricted system). Under null hypothesis ($\beta_2 = 0$), we have $Y = X_1\beta_1 + \epsilon$, and expanding the numerator of the expression above, we get

$$Y^T(X(X^TX)^{-1}X^T - X_1(X_1^TX_1)^{-1}X_1^T)Y$$

$$= \epsilon^T(X(X^TX)^{-1}X^T - X_1(X_1^TX_1)^{-1}X_1^T)\epsilon + \beta_1^TX_1^T(X(X^TX)^{-1}X^T - X_1(X_1^TX_1)^{-1}X_1^T)X_1\beta_1.$$

We recognize the second summand as

$$(\beta_1^TX_1^TX(X^TX)^{-1}X^T - \beta_1^TX_1^TX_1(X_1^TX_1)^{-1}X_1^T)X_1\beta_1 = (\beta_1^TX_1^T - \beta_1^TX_1^T)X_1\beta_1 = 0.$$

So, letting $A = X(X^TX)^{-1}X^T - X_1(X_1^TX_1)^{-1}X_1^T$, under null hypothesis our test statistic is $\frac{\epsilon' A\epsilon}{\sigma^2}$. You should prove for yourself (Exercise 8.3) that A is symmetric and idempotent of rank p_2, and therefore $\frac{\epsilon' A\epsilon}{\sigma^2} \sim {\chi^2}_{p_2}$. That doesn't help us all that much yet since we don't know the value of σ^2.

We have already established above that $\frac{\|e_f\|^2}{\sigma^2} \sim {\chi^2}_{n-p}$, where $\|e_f\|^2 = \epsilon^T(I-H)\epsilon$. We proceed to show now that the two quadratic forms $\epsilon^T(I-H)\epsilon$ and $\epsilon^TA\epsilon$ are independent, by showing that $(I-H)\sigma^2IA = \sigma^2(I-H)A = 0$. The proof is left as an exercise for you. We will now denote $\frac{\|e_f\|^2}{n-p}$ by MS_{Res}, and we conclude that under the null hypothesis

$$\frac{\epsilon^TA\epsilon}{p_2\sigma^2} \Big/ \frac{\epsilon^T(I-H)\epsilon}{(n-p)\sigma^2} = \frac{\|\hat{Y}_f\|^2 - \|\hat{Y}_r\|^2}{p_2MS_{Res}} \sim F_{p_2,n-p}.$$

We can now test our null hypothesis $\beta_2 = 0$, using this statistic, and we would reject for large values of F.

Exercises

12.1 Show that if X has a bivariate normal distribution $N(\mu, V)$, and B is invertible, then $Y = BX + d$ has a bivariate normal distribution $N(B\mu + d, BVB^T)$.

12.2 Assume V is positive definite, AV, and $B = V^{\frac{1}{2}}AV^{\frac{1}{2}}$ is idempotent. Show that $rank(B) = rank(AV)$ (hint: consider the nullspaces, and invertible transformation $v = V^{1/2}w$).

12.3 Let $X = [X_1\, X_2]^T$ for $n \times p_1$ X_1 and $n \times p_2$ X_2, and $A = X(X^TX)^{-1}X^T - X_1(X_1^TX_1)^{-1}X_1^T$. Show that A is symmetric and idempotent of rank p_2 (use trace to determine rank of A).

13 References

1. Bickel, Peter J and Doksum, Kjell A., 'Mathematical Statistics: Basic Ideas and Selected Topics', 2nd ed., 2001, Prentice Hall

2. Casella, George and Berger, Roger L, 'Statistical Inference', 2nd ed., 2001, Duxbury Press

3. Freedman, David A., 'Statistical Models: Theory and Applications', 2005, Cambridge University Press

4. Garcia, S. R., & Horn, R. A., 'A Second Course in Linear Algebra', 2017, Cambridge University Press.

5. Montgomery, Douglas C. et al, 'Introduction to Linear Regression Analysis', 3rd ed., 2001, John Wiley & Sons

6. Horn, Roger A; Johnson, Charles R., 'Matrix Analysis', 1985, Cambridge University Press

7. Strang, G, 'Linear Algebra and Its Applications', 3rd ed., 1988, Saunders College Publishing

8. Trefethen, Lloyd N., and David Bau III. Numerical linear algebra. Vol. 50. Siam, 1997. APA

9. Lax, Peter. Functional Analysis.

10. Hastie, Trevor, et al. The Elements of Statistical Learning.