THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

Department of Biostatistics
Department of Stat. and OR

Refresher course, Summer 2015

Linear Algebra

Original Author:
Oleg Mayba
(UC Berkeley, 2006)

Modified By:
Eric Lock (UNC, 2010 & 2011)
Gen Li (UNC, 2012)
Michael Lamm (UNC, 2013)
Wen Jenny Shi (UNC, 2014)
Meilei Jiang (UNC, 2015)

Instructor:
Meilei Jiang
(UNC at Chapel Hill)

Based on the NSF sponsored (DMS Grant No 0130526) VIGRE Boot camp lecture notes in the Department of Statistics, University of California, Berkeley

July 28, 2015

Contents

1 Introduction

These notes are intended for use in the warm-up camp for incoming UNC STOR and Biostatistics 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 9 algebraic classes in total (one class per weekday for two weeks, excluding a day for the university orientation). Each class will last two hours with a short break in between.

Go Tar Heels!

2 Vector Spaces

A set V is a **vector space** over $\mathbb{R}$ (field), and its elements are called **vectors**, if there are 2 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 $rv = w$)

that satisfy the following 8 conditions $\forall\ v_1, v_2, v_3 \in V$ and $\forall\ r_1, r_2 \in \mathbb{R}$ (filed):

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'll be needing 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 $0_{n\times 1}$.

Vector space is an elementary object considered in the linear algebra. Here are some concrete examples:

1. Vector space $\mathbb{R}^n$ with usual operations of element-wise addition and scalar multiplication. An example of these operations in $\mathbb{R}^2$ is illustrated above.
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 $(rf)(x) = rf(x)$.

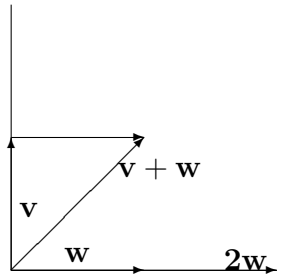

Figure 1: Vector Addition and Scalar Multiplication

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. A 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 $w = r_1 v_1 + r_2 v_2 + \ldots + r_n v_n$. 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.

Some properties of subspaces:

1. Subspaces are closed under linear combinations.
2. A nonempty set S is a subspace if and only if every linear combination of (finitely many) elements of S also belongs to S.

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 $r_1 v_1 + r_2 v_2 + \ldots + r_n v_n = 0 \Longrightarrow r_1 = r_2 = \ldots = r_n = 0$, 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.

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: This proof will be by contradiction, and use two facts that can be easily verified from the definitions above. First, if $I \subset V$ is linearly independent, then $I \cup \{x\}$ is linearly dependent if and only if (iff) $x \in \text{span}(I)$. Second, if $S, T \subset V$ with $T \subset \text{span}(S)$ then $\text{span}(T) \subset \text{span}(S)$.

If the theorem's claim does not hold. Then there exists a $x \in I$ such that for all $y \in S$ $\{y\} \cup I \backslash \{x\}$ is linearly dependent. Let $I' = I \backslash \{x\}$. By I linearly independent it follows that I' is also linearly independent. Then by the first fact above, $\{y\} \cup I'$ 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 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 by convention, and if $I \subset S$ the claim holds directly. So assume $|S| = m < \infty$, and $I \not\subset S$.

Consider now 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, else 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. $\square$

Basis and dimension Now suppose that $v_1, v_2, \ldots, v_n$ span V and that, moreover, they are linearly independent. Then we say that the set $\{v_1, v_2, \ldots, v_n\}$ is a **basis** for V.

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 call the unique number of vectors in a basis for V the **dimension** of V (denoted $\dim(V)$).

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.

2.2 Special Spaces

Inner product space An **inner product** is 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 $\langle v, v \rangle = 0$ iff $v = 0$ (positive-definiteness)

We note here that not all vector spaces have inner products defined on them. We call the vector spaces where the inner products are defined the **inner product space**.

Examples:

1. Given 2 vectors $x = [x_1, x_2, \cdots, x_n]'$ and $y = [y_1, y_2, \cdots, y_n]'$ in $\mathbb{R}^n$, we define their inner product $x'y = \langle x, y\rangle = \sum_{i=1}^{n} x_i y_i$. You can check yourself that the 3 properties above are satisfied, and the meaning of notation $x'y$ will become clear from the next section.
2. Given $f, g \in C_{[-1,1]}$, 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.

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: 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

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

□

we can show the result with equality if and only if $v = \lambda w$. Namely, the inequality holds and it's equality if and only if v and w are linearly dependent.

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

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

The angle is in $[0, \pi)$. This generates nice geometry for the inner product space.

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** $x := \|x\| = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2} = \sqrt{x'x}$, or $\|x\|^2 = x'x$. 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 $f \in C_{[-1,1]}$, we define $\|f\|_p = \left(\int_{-1}^{1} |f(x)|^p dx\right)^{\frac{1}{p}}$, which is also a norm. (see Minkowski Inequality)

4. For any inner product space V, $\|x\|^2 = \langle x, x\rangle$ 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, in most cases, the opposite (i.e. to obtain inner products from norms) is not obvious.

Metric Space A more general definition on the vector space is the **metric**. The 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 vector space equipped with a metric is called **metric space**. Many analytic definitions (e.g. completeness, compactness, continuity, etc) can be defined under metric space. Please refer to the analysis material for more information.

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. For example, assuming we define the discrete metric on the space V, where $d(x, y) = 0$ if $x = y$ and $d(x, y) = 1$ if $x \neq y$; it is not obvious what kind of norm should be defined in this space.

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

1. $d(x, y) = d(x + r, y + r)$ (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. Given a vector space V, if we define an inner product in it, we can naturally derived a norm in it; if we have a norm in it, we can naturally derived a metric in it. The opposite is not true.

2.3 Orthogonality

We say that vectors v, w in V are **orthogonal** if $\langle v, w\rangle = 0$, or equivalently, $angle(v, w) = \pi/2$. It is denoted as $v \perp w$.

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:

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

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

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 V, i.e., $\langle q_i, q_j\rangle = 0\ \forall i \neq j$, and suppose that $r_1q_1 + r_2q_2 + \ldots + r_nq_n = 0$. 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 $x = [x_1, x_2, \cdots, x_n]'$. Then if we let $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, we can see that vector $e = [x_1 - \bar{x}, x_2 - \bar{x}, \cdots x_n - \bar{x}]'$ is orthogonal to vector $\hat{x} = [\bar{x}, \bar{x}, \cdots, \bar{x}]'$, since $\sum_{i=1}^{n} \bar{x}(x_i - \bar{x}) = \bar{x}\sum_{i=1}^{n} x_i - \bar{x}\sum_{i=1}^{n} \bar{x} = n\bar{x}^2 - n\bar{x}^2 = 0$.

Orthogonal subspace and complement Suppose S, T are subspaces of 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 $[0, 0, 1]'$, 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 $\dim(S^{\perp}) = n - k$.

2.4 Gram-Schmidt Process

Suppose we're given linearly independent vectors $v_1, v_2, \ldots, v_n$ in V, and there's an inner product defined on V. Then we know that $v_1, v_2, \ldots, v_n$ form a basis for the subspace which they span (why?). 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 $q_i = v_i - \left(\frac{\langle q_1, v_i\rangle}{\langle q_1, q_1\rangle} q_1 + \ldots + \frac{\langle q_{i-1}, v_i\rangle}{\langle q_{i-1}, q_{i-1}\rangle} q_{i-1}\right)$, 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.

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.

Exercises

2.1 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.

2.2 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.

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

2.4 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.

2.5 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 $F_{[-1,1]}$.

2.6 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 $h(x) = x^2$.

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

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 $x = [x_1, x_2, \cdots x_n]' \in \mathbb{R}^n$ 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 $A + B = B + A$.

Scalar multiplication Scalar multiplication is also defined elementwise. 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 $m \times q$. 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.

Below we list some very useful ways to think about matrix multiplication:

1. Suppose A is $m \times n$ matrix, and x is a $n \times 1$ column vector. Then if we let $a_1, a_2, \dots, a_n$ denote the respective columns of A, and $x_1, x_2, \dots, x_n$ denote the components of x, we get a $m \times 1$ vector $Ax = x_1 a_1 + x_2 a_2 + \dots + x_n a_n$, a linear combination of the columns of A. Thus applying matrix A to a vector always returns a vector in the column space of A (see below for definition of column space).

2. Now, let A be $m \times n$, and let x be a $1 \times m$ row vector. Let $a_1, a_2, \ldots, a_m$ denote rows of A, and $x_1, x_2, \ldots, x_m$ denote the components of x. Then multiplying A on the left by x, we obtain a $1 \times n$ row vector $xA = x_1a_1 + x_2a_2 + \ldots + x_ma_m$, a linear combination of the rows of A. Thus multiplying matrix on the right by a row vector always returns a vector in the row space of A (see below for definition of row space)

3. Now let A be $m \times n$, and let B be $n \times k$, and let $a_1, a_2, \ldots, a_n$ denote columns of A and $b_1, b_2, \ldots, b_k$ denote the columns of B, and let c_j denote the j-th column of $m \times k$ $C = AB$. Then $c_j = Ab_j = b_{1j}a_1 + b_{2j}a_2 + \ldots + b_{nj}a_n$, i.e. we get the columns of product matrix by applying A to the columns of B. Notice that it also implies that every column of product matrix is a linear combination of columns of A.

4. Once again, consider $m \times n$ A and $n \times k$ B, and let $a_1, a_2, \ldots a_n$ denote rows of A (they are, of course, just $1 \times n$ row vectors). Then letting c_i denote the i-th row of $C = AB$, we have $c_i = a_iB$, i.e. we get the rows of the product matrix by applying rows of A to B. Notice, that it means that every row of C is a linear combination of rows of B.

5. Finally, let A be $m \times n$ and B be $n \times k$. Then if we let $a_1, a_2, \ldots, a_n$ denote the columns of A and $b_1, b_2, \ldots, b_n$ denote the rows of B, then $AB = a_1b_1 + a_2b_2 + \ldots + a_nb_n$, the sum of n matrices, each of which is a product of a row and a column (check this for yourself!).

Transpose Let A be $m \times n$, then the **transpose** of A is the $n \times m$ matrix A', s.t. $a_{ij} = a'_{ji}$. 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 $x'y$ 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. $(AB)' = B'A'$

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

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 $(A^{-1})' = (A')^{-1}$

   Proof: Exercise 3.1(b).

4. A is invertible iff $Ax = 0 \Longrightarrow x = 0$ (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 $x = 0$. Then the columns $a_1, ..., a_n$ of A are linearly independent and therefore form a basis for $\mathbb{R}^n$ (Exercise 2.4). So, if $e_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$, $e_2 = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}$, ..., $e_n = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$, we can write

$$c_{1i}a_1 + c_{2i}a_2 + ... + c_{ni}a_n = A \begin{bmatrix} c_{1,i} \\ c_{2,i} \\ \vdots \\ c_{n,i} \end{bmatrix} = e_i$$

   for all $i = 1, ..., n$. Hence, if C is given by $C_{ij} = c_{ij}$, then

$$AC = [e_1\, e_2 \,...\, e_n] = I_n.$$

   To see that $CA = I$, note ACA=IA=A, and therefore A(CA-I)=0. Let $Z = CA - I$. Then $Az_i = 0$ for each column of Z. By assumption $Az_i = 0 \Rightarrow z_i = 0$, so $Z = 0$, $CA = I$. Hence, $C = A^{-1}$ and A is invertible. □

3.2 Special Matrices

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

$$< Qx, Qy >= x'Q'Qy = x'I_{n\times n}y = x'y.$$

In particular $\|Qx\| = \sqrt{x'Q'Qx} = \|x\|$. Also, if A, and B are orthogonal, then so are A^{-1} and AB. We say that a square matrix A is **idempotent** if $A^2 = A$.

We say that a square matrix A is **positive definite** if A is symmetric and if $\forall\ n \times 1$ vectors $x \neq 0_{n\times 1}$, we have $x'Ax > 0$. 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 $x'Ax \geq 0$. 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' (more generally, if A is positive semi-definite, then so is A').

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.
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 $tr(A + B) = tr(A) + tr(B)$.

3.3 Fundamental Spaces

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 $col(A')$.

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 $\{x \in \mathbb{R}^m : x'A = 0\}$, and we notice that this is precisely $N(A')$.

The fundamental theorem of linear algebra states:

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

2. $col(A) = (N(A'))^{\perp}$ and $N(A) = (col(A'))^{\perp}$. 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 $\dim(N(A'))$ $= m - r$.

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 iff the nullspace doesn't contain any vector other than zero. Similarly, rows are linearly independent iff the left nullspace doesn't contain any vector other than zero.

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 $\mathbb{R}^n$, 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 iff it is invertible.

We now give the following results:

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

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

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

   Proof: Similar to (3).

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 $(A^{-1})' = (A')^{-1}$

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

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

3.4 *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.

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

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. $< b^*, b - b^* >= 0$. 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 s^* in S s.t.

$$s^* = \arg\min_{s \in S} \|v - s\|$$

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

Proof (sketch):

First construct a sequence $y_n \in S$ such that

$$\|y_n - v\| \to \inf_{s \in S} \|v - s\|.$$

Then use the Parallelogram Law ($\frac{1}{2}\|x-y\|^2 + \frac{1}{2}\|x+y\|^2 = \|x\|^2 + \|y\|^2$) with $x = y_n - v$ and $y = v - y_m$. Rearranging terms, using convexity and appropriate bounds, take the lim inf of each side to show that the sequence $\{y_n\}_{n=1}^{\infty}$ is Cauchy. This combined with V complete gives the existence of

$$s^* = \arg\min_{s \in S} \|v - s\|.$$

To obtain uniqueness use the Parallelogram Law and the convexity of S. $\square$

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 iff $< v - s^*, s >= 0 \quad \forall s \in S$.

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 $s^* - ts \in S$, 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

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

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

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 $s \in S$

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

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 $x, y \in V$, $\|x - y\| \geq \|P_Sx - P_Sy\|$.

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

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

Taking $a = x - y$ and $b = P_Sx - P_Sy$, Fact 1 and the Corollary imply that $< a - b, b >= 0$ and thus

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

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.

Fact 4: Let $\{a_1, \ldots, a_n\}$ and S be as above, and $A = [a_1 \ldots a_n]$. Then the **projection matrix** $P_S = P = A(A'A)^{-1}A'$.

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 $col(A)^\perp = N(A')$ and thus

$$A'(b - Ax) = 0 \Rightarrow A'Ax = A'b$$

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

We follow up with some properties of projection matrices:

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 $A_S = [a_1\, a_2 \ldots a_q]$ and $A_{S^{\perp}} = [a_{q+1} \ldots a_m]$ be such that $a_1, \ldots, a_q$ form a basis for S and $a_{q+1}, \ldots, a_m$ form a basis for $S_{\perp}$. By 3, if P_S is the projection onto $col(A_S)$ and $P_{S^{\perp}}$ is the projection onto $col(A_{S^{\perp}})$, $\forall b \in \mathbb{R}^m$

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

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

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

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

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, $P'(b - Pb) = P'(I - P)b = P(I - P)b = (P - P^2)b = 0b = 0$. $\square$

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

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 $q_1 q_1' b + q_2 q_2' b + \ldots + q_m q_m' b$ 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

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 $p \times 1$ 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 leaste 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 equivilient to minimizing $\|Y - \tilde{Y}\|^2 = \|Y - X\tilde{\beta}\|^2$. It follows that the **fitted values** associated with the lse satisfy

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

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

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

The matrix $H = X(X'X)^{-1}X'$ 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 $V = C_{[-1,1]}$ with $\|f\|^2 =< f, f >= \int_{-1}^{1} f(x)f(x)dx$. 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 $< f - P_S f, s >= 0$ for all linear functions $s \in S$. By previous (in class) exercise finding $P_S f$ requires finding constans a and b such that

$$< x^2 - (ax + b), 1 >= 0 =< x^2 - (ax + b), x >$$

. First we solve

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

Next,

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

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 $col(A)$, $tr(P) = r$.

5 Differentiation

5.1 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 $x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix}$ 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 $\frac{\partial f}{\partial x} = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_k} \end{bmatrix}$.

Below are the extensions

1. Let $a \in \mathbb{R}^k$, and let $y = a'x = a_1x_1 + a_2x_2 + \ldots + a_kx_k$. Then $\frac{\partial y}{\partial x} = a$

   Proof: Follows immediately from definition.

2. Let $y = x'x$, 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 $y = a'Ax$. Then $\frac{\partial y}{\partial x} = A'a$

   Proof: Note that $a'A$ is $1 \times k$. Writing $y = a'Ax = (A'a)'x$ it's then clear from 1 that $\frac{\partial y}{\partial x} = A'a$. $\square$

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

   Proof: Exercise 5.1(b).

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

Exercises

5.1 Prove the following properties of vector derivatives:

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

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

5.2 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 $\frac{\partial \theta}{\partial y}$.

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

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 3.

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

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

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

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

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 instead of row i we now have row $i-rj$. Then splitting the determinant of the new matrix along this row we have det(original) + det(original matrix with row rj in place of row i. That last determinant is just r times determinant of original matrxi with row j in place of row i, and since the matrix has two equal rows, the determinant is 0. So the determinant of the new matrix has to be equal to the determinant of the original. □

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. $\det(A) = 0$ iff A is not invertible (proof involves ideas of elimination)

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

   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 $A_* = [a_{11}^* \, a_{12}^* \ldots a_{1n}^*] \in \mathbb{R}^n$ 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. $\det(A') = \det(A)$. This is true since expanding along the row of A' is the same as expanding along the corresponding column of A.

6.2 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 λ. For small n, we find eigenvalues by noticing that

$$Ax = \lambda x \iff (A - \lambda I)x = 0 \iff A - \lambda I$$

is not invertible $\iff \det(A - \lambda I) = 0$. We then write out the formula for the determinant (which will be a polynomial of degree n in λ) and solve it. Every $n \times n$ A then has n eigenvalues (possibly repeated and/or complex), since every polynomial of degree n has n roots. Eigenvectors for a specific value of λ are found by calculating the basis for nullspace of $A - \lambda I$ via standard elimination techniques. If $n \geq 5$, there's a theorem in algebra that states that no formulaic expression for the roots of the polynomial of degree n exists, so other techniques are used, which we will not be covering. Also, you should be able to see that the eigenvalues of A and A' 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.

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 $r_1Ax_1 + r_2Ax_2 = A0 = 0 \implies \lambda_1 r_1 x_1 + \lambda_2 r_2 x_2 = 0$. Multiplying first equation by λ_1 and subtracting it from the second, we get $\lambda_1 r_1 x_1 + \lambda_2 r_2 x_2 - (\lambda_1 r_1 x_1 + \lambda_1 r_2 x_2) = 0 - 0 = 0 \implies r_2(\lambda_2 - \lambda_1)x_2 = 0$ and since $x_1 \neq 0$, 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. $\square$

We say that $n \times n$ A is **diagonalizable** if it has n linearly independent eigenvectors. Certainly, every matrix that has n DISTINCT eigenvalues is diagonalizable (by the proof above), but some matrices that fail to have n distinct eigenvalues may still be diagonalizable, as we'll see in a moment. The reasoning behind the term is as follows: 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 corresponding eigenvalues (note that they need not be distinct), and let S be $n \times n$ matrix the j-th column of which is s_j. Then if we let Λ be $n \times n$ diagonal matrix s.t. the ii-th entry on the main diagonal is λ_i, then from familiar rules of matrix multiplication we can see that $AS = S\Lambda$, and since S is invertible (why?) we have $S^{-1}AS = \Lambda$ (Exercise 6.2). 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 and Basic Results

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 $\mathbb{C}^n$?) We also note that we need to tweak a bit the earlier definition of transpose to account for the fact that if $x = \begin{bmatrix} 1 \\ i \end{bmatrix} \in \mathbb{C}^2$, then

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

We note that in the complex case $\|x\|^2 = (\bar{x})'x$, where $\bar{x}$ is the complex conjugate of x, and we introduce the notation x^H to denote the transpose-conjugate $\bar{x}'$ (thus we have $x^H x = \|x\|^2$).

You can easily see for yourself that if $x \in \mathbb{R}^n$, then $x^H = x'$. $A^H = (\bar{A})'$ for $n \times n$ matrix A is defined similarly and we call A^H **Hermitian transpose** of A. 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 $x^H y = y^H x$).

We say that $n \times n$ matrix A is **Hermitian** if $A = A^H$. We say that $n \times n$ A is **unitary** if $A^H A = AA^H = I (A^H = A^{-1})$. You should check for yourself that every symmetric real matrix is Hermitian, and every orthogonal real matrix is unitary. We say that a square matrix A is **normal** if it commutes with its Hermitian transpose: $A^H A = AA^H$. You should check for yourself that Hermitian (and therefore symmetric) and unitary (and therefore orthogonal) matrices are normal. We next present some very important results about Hermitian and unitary matrices (which also include as special cases symmetric and orthogonal matrices respectively):

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. 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 $x^H Ax > 0 \Longrightarrow x^H(\lambda x) > 0 \Longrightarrow \lambda \|x\|^2 > 0 \Longrightarrow \lambda > 0$. $\square$

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 mutally 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 $A = Q \Lambda Q'$. 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 $|\lambda_1| = 1 \Longrightarrow \bar{\lambda_1} \lambda_1 = 1$, and we conclude that $\lambda_1 = \lambda_2$, a contradiction. Therefore, $x^H y = 0$. $\square$

9. For EVERY square matrix A, $\exists$ some unitary matrix U s.t. $U^{-1} A U = U^H A U = T$, where T is upper triangular. We will not prove this result, but the proof can be found, for example, in section 5.6 of G.Strang's 'Linear Algebra and Its Applications' (3rd ed.) This is a very important result which we're going to use in just a moment to prove the so-called Spectral Theorem.

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

    Proof: $BB^H = (U^H A U)(U^H A U)^H = U^H A U U^H A^H U = U^H A A^H U = U^H A^H A U$ (since A is normal) $= U^H A^H U U^H A U = (U^H A U)^H (U^H A U) = B^H B$. $\square$

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

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

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 $\|A^H x\|^2 = |a_{11}|^2 + |a_{12}|^2 + \ldots + |a_{1n}|^2$. 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. $\square$

13. We have just succeded in proving the Spectral Theorem: If A is $n \times n$ symmetric matrix, then we can write it as $A = Q \Lambda Q'$. We know that if A is symmetric, then it's normal, and we know that we can find some unitary U s.t. $U^{-1} A U = T$, where T is upper triangular. But we know that T is also normal, and being upper triangular, it is then diagonal. So A is diagonalizable and by discussion above, the entries of $T = \Lambda$ are eigenvalues of A (and therefore real) and the columns of U are corresponding unit eigenvectors of A (and therefore real), so U is a real orthogonal matrix.

14. More generally, we have shown that every normal matrix is diagonalizable.

15. If A is positive definite, it has a square root B, s.t. $B^2 = A$.

    Proof: We know that we can write $A = Q\Lambda Q'$, where all diagonal entries of Λ are positive. Let $B = Q\Lambda^{1/2}Q'$, 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. $\square$

16. 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 $Ax = Q\Lambda Q'x$. Then (do it as an exercise!) carrying out the matrix multiplication on $Q\Lambda Q'$ 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: $Q\Lambda Q' = \lambda_1 q_1 q_1' + \lambda_2 q_2 q_2' + \ldots + \lambda_n q_n q_n'$ and so $Ax = \lambda_1 q_1 q_1' x + \lambda_2 q_2 q_2' x + \ldots + \lambda_n q_n q_n' x$. We recognize $q_i q_i'$ 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.

6.4 SVD and Pseudo-inverse

Theorem: Every $m \times n$ matrix A can be written as $A = Q_1 \Sigma Q_2'$, where Q_1 is $m \times m$ orthogonal, Σ is $m \times n$ pseudo-diagonal (meaning that that the first r diagonal entries σ_{ii} are non-zero and the rest of the matrix entries are zero, where $r = rank(A)$), and Q_2 is $n \times n$ orthogonal. Moreover, the first r columns of Q_1 form an orthonormal basis for $col(A)$, the last $m - r$ columns of Q_1 form an orthonormal basis for $N(A')$, the first r columns of Q_2 form an orthonormal basis for $col(A')$, last $n - r$ columns of Q_2 form an orthonormal basis for $N(A)$, and the non-zero entries of Σ are the square roots of non-zero eigenvalues of both AA' and $A'A$. (It is a good exersise at this point for you to prove that AA' and $A'A$ do in fact have same eigenvalues. What is the relationship between eigenvectors?). This is known as the **Singular Value Decomposition** or SVD.

Proof: $A'A$ is $n \times n$ symmetric and therefore has a set of n real orthonormal eigenvectors. Since $rank(A'A) = rank(A) = r$, we can see that $A'A$ has r non-zero (possibly-repeated) eigenvalues (Exercise 6.3). Arrange the eigenvectors $x_1, x_2, \ldots, x_n$ in such a way that the first $x_1, x_2, \ldots, x_r$ correspond to non-zero $\lambda_1, \lambda_2, \ldots, \lambda_r$ and put $x_1, x_2, \ldots, x_n$ as columns of Q_2. Note that as $x_{r+1}, x_{r+2}, \ldots, x_n$ form a basis for $N(A)$ by Exercise 2.4 as they are linearly independent, $dim(N(A)) = n - r$ and

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

Therefore $x_1, x_2, \ldots, x_r$ form a basis for the row space of A. Now set $\sigma_{ii} = \sqrt{\lambda_i}$ for $1 \leq i \leq r$, and let the rest of the entries of $m \times n$ Σ be 0. Finally, for $1 \leq i \leq r$, let $q_i = \frac{Ax_i}{\sigma_{ii}}$.

You should verify for yourself that q_i's are orthonormal ($q_i'q_j = 0$ if $i \neq j$, and $q_i'q_i = 1$). By Gram-Schmidt, we can extend the set $q_1, q_2, \ldots, q_r$ to a complete orthonormal basis for $\mathbb{R}^m$, $q_1, q_2, \ldots, q_r, q_{r+1}, \ldots, q_n$. As $q_1, q_2, \ldots, q_r$ are each in the column space of A and linearly independent, they form an orthonormal basis for column space of A and therefore $q_{r+1}, q_{r+2}, \ldots, q_n$ form an orthonormal basis for the left nullspace of A. We now verify that $A = Q_1 \Sigma Q_2'$ by checking that $Q_1' A Q_2 = \Sigma$. Consider ij-th entry of $Q_1' A Q_2$. It is equal to $q_i' A x_j$. For $j > r$, $Ax_j = 0$ (why?), and for $j \leq r$ the expresison becomes $q_i' \sigma_{jj} q_j = \sigma_{jj} q_i' q_j = 0$(if $i \neq j$) or 1 (if $i = j$). And therefore $Q_1' A Q_2 = \Sigma$, as claimed. $\square$

One important application of this decomposition is in estimating β in the system we had before when the columns of X are linearly dependent. Then $X'X$ is not invertible, and more than one value of $\hat{\beta}$ will result in $X'(Y - X\hat{\beta}) = 0$. By convention, in cases like this, we choose $\hat{\beta}$ that has the smallest length. For example, if both $\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$ satisfy the normal equations, then we'll choose the latter and not the former. This optimal value of $\hat{\beta}$ is given by $\hat{\beta} = X^+ Y$, where X^+ is a $p \times n$ matrix defined as follows: suppose X has rank $r < p$ and it has S.V.D. $Q_1 \Sigma Q_2'$. Then $X^+ = Q_2 \Sigma^+ Q_1'$, where Σ^+ is $p \times n$ matrix s.t. for $1 \leq i \leq r$ we let ${\sigma^+}_{ii} = 1/\sigma_{ii}$ and ${\sigma^+}_{ij} = 0$ otherwise. We will not prove this fact, but the proof can be found (among other places) in appendix 1 of Strang's book. The matrix X^+ is called the **pseudo-inverse** of the matrix X. The pseudo-inverse is defined and unique for all matrices whose entries are real or complex numbers.

Exercises

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

6.2 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 $S^{-1}AS = \Lambda$.

6.3 Show that if $rank(A) = r$, then $A'A$ has r non-zero eigenvalues.

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

7 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.

7.1 Expectation, Variance and Covariance

The **expected value** of a continuous random variable X, with probability density function f, is defined by

$$EX = \int_{-\infty}^{\infty} xf(x)dx.$$

The expected value of a discrete random variable X, with probability mass function p, is defined by

$$EX = \sum_{x \in \mathbb{R}, p(x) \neq 0} xp(x).$$

The expected value is **well-defined** if $E|X| < \infty$.

We now list some basic properties of $E(\cdot)$:

1. $X \leq Y$ imples $EX \leq EY$

   Proof: Follows directly from properties of $\int$ and $\sum$.

2. For $a, b \in \mathbb{R}$, $E(aX + bY) = aEX + bEY$.

   Proof: Follows directly from properties of $\int$ and $\sum$.

3. $|EX| \leq E|X|$

   Proof: Note that $X, -X \leq |X|$. Hence, $EX, -EX \leq E|X|$ and therefore $|EX| \leq E|X|$. $\square$

4. If X and Y are independent ($X \coprod Y$), then $E(XY) = EX \cdot EY$.

   Proof: See Theorem 4.2.10 in Casella and Berger.

5. If X is a non-negative continuous random variable, then

$$EX = \int_0^{\infty} P(X \geq t)dt = \int_0^{\infty} (1 - F(t))dt.$$

Proof: Suppose $X \sim f$. Then,

$$\begin{aligned}
\int_0^\infty P(X > t)dt &= \int_0^\infty [\int_t^\infty f(x)dx]dt \\
&= \int_0^\infty [\int_0^\infty f(x)I(x > t)dx]dt \\
&= \int_0^\infty \int_0^\infty f(x)I(x > t)dtdx \quad \text{(Fubini)} \\
&= \int_0^\infty f(x)[\int_0^\infty I(x > t)dt]dx \\
&= \int_0^\infty xf(x)dx = EX \qquad \square
\end{aligned}$$

6. If $X \sim f$ then $Eg(X) = \int g(x)f(x)dx$.
   If $X \sim p$ then $Eg(x) = \sum_x g(x)p(x)$.

   Proof: Follows from definition of $Eg(X)$.

The **variance** of a random variable X is defined by

$$\begin{aligned}
\text{Var}(X) &= E(X - EX)^2 \\
&= EX^2 - (EX)^2.
\end{aligned}$$

Note that $\text{Var}(X)$ is finite (and therefore well-defined) if $EX^2 < \infty$. The **covariance** of two random variables X and Y is defined by

$$\begin{aligned}
\text{Cov}(X, Y) &= E[(X - EX)(Y - EY)] \\
&= E(XY) - EXEY.
\end{aligned}$$

Note that $\text{Cov}(X, Y)$ is finite if $EX^2, EY^2 < \infty$.

We now list some general properties, that follow from the definition of variance and covariance:

1. $\text{Var}(X) \geq 0$, with "=" iff X is constant with probability 1.
2. For $a, b \in \mathbb{R}$, $\text{Var}(aX + b) = a^2\text{Var}(X)$.
3. If $X \coprod Y$, 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 $EX_i^2 < \infty$, then

$$\text{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \text{Var}(X_i) + 2\sum_{i<j} \text{Cov}(X_i, X_j).$$

7.2 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

$$\begin{aligned}P(X+Y \leq v) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_X(x)f_Y(y)I\{(x,y) : x+y \leq v\}dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{v-y} f_X(x)f_Y(y)dxdy \\ &= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{v-y} f_X(x)dx\right]f_Y(y)dy \\ &= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{v} f_X(u-y)du\right]f_Y(y)dy \qquad (u = y+x) \\ &= \int_{-\infty}^{v}\left[\int_{-\infty}^{\infty} f_X(u-y)f_Y(y)dy\right]du \\ &= \int_{-\infty}^{v}(f_X * f_Y)(u)du. \qquad \square\end{aligned}$$

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 $1 \leq i, j \leq k$, 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**[1]) 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

$$f_{Y_1,Y_2}(y_1, y_2) = \frac{1}{2} f_{X_1,X_2}(\frac{y_1 + y_2}{2}, \frac{y_1 - y_2}{2}).$$

[1]The partial derivative matrix D is sometimes called the **Jacobain matrix** (see Section 5.2).

As a special case, assume X_1, X_2 are $\mathbb{N}(0,1)$ and independent. Then,

$$\begin{aligned} f_{Y_1,Y_2}(y_1,y_2) &= \frac{1}{2}\phi(\frac{y_1+y_2}{2})\phi(\frac{y_1-y_2}{2}) \\ &= \frac{1}{4\pi}\exp\left\{-\frac{(y_1+y_2)^2}{8}-\frac{(y_1-y_2)^2}{8}\right\} \\ &= \frac{1}{4\pi}\exp\left\{-\frac{2y_1^2+2y_2^2}{8}\right\} \\ &= \frac{1}{4\pi}\exp\left\{-\frac{y_1^2}{4}\right\}\exp\left\{-\frac{y_2^2}{4}\right\}. \end{aligned}$$

So, both Y_1 and Y_2 are $\mathbb{N}(0,2)$, and they are independent!

7.3 Derivation of Common Univariate Distributions

Double Exponential

If $X_1, X_2 \sim \text{Exp}(\lambda)$ and $X_1 \coprod X_2$, 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 $EX = \frac{\alpha}{\beta}$, $\mathrm{Var}(X) = \frac{\alpha}{\beta^2}$.

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 $x > 0$,

$$\begin{aligned} f(x) &= (g_{\alpha_1,\beta} * g_{\alpha_2,\beta})(x) \\ &= \int_0^x g_{\alpha_1,\beta}(x-u) g_{\alpha_2,\beta}(u) du \\ &= \frac{\beta^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} e^{-\beta x} \int_0^x (x-u)^{\alpha_1-1} u^{\alpha_2-1} du \qquad (1) \\ &= \text{const} \cdot e^{-\beta x} x^{\alpha_1+\alpha_2-1} \end{aligned}$$

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. $\square$

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

$$f_Y(y) = \frac{\lambda^2 y^{n-1} e^{-\lambda y}}{(n-1)!}.$$

This is also known as an **Erlang** distribution with parameters n and λ.

It follow from equation (1), with $x = 1$ that

$$\frac{\beta^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} e^{-\beta} \int_0^1 (1-u)^{\alpha_1-1} u^{\alpha_2-1} du$$
$$= g_{\alpha_1+\alpha_2,\beta}(1) = \frac{\beta^{\alpha_1+\alpha_2} e^{-\beta}}{\Gamma(\alpha_1+\alpha_2)}.$$

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

$$EX = \frac{r}{r+s}, \qquad \mathrm{Var}(X) = \frac{rs}{(r+s)^2(r+s+1)}.$$

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 $Y = X_1^2 + \ldots + X_k^2 \sim \chi_k^2$, then

$$EY = E(X_1^2 + \ldots + X_k^2) = kEX_1^2 = k.$$

$$\begin{aligned}\mathrm{Var}(Y) &= k\mathrm{Var}(X_1^2) = k(EX_1^4 - (EX_1^2)^2) \\ &= k(3-1) = 2k.\end{aligned}$$

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 $X \sim \chi_m^2$, $Y \sim \chi_n^2$, and $X \coprod Y$.

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 $X, Y \sim \mathbb{N}(0,1)$ 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 $EX = 0$ for $n \geq 2$, undefined for $n = 1$; $\mathrm{Var}(X) = \frac{n}{n-2}$ 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 $n \to \infty$.

7.4 Random Vectors: Expectation and Variance

A **random vector** is a vector $X = [X_1\, X_2\, ...X_k]'$ 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 $E(X)$, if it exists, is given by the expected value of each component:

$$E(X) = [EX_1\, EX_2\, ...\, EX_k]'.$$

The covariance matrix of a random vector $Cov(X)$ is given by

$$Cov(X) = E[(X - EX)(X - EX)'] = E(XX') - EXEX'.$$

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 $k \times 1$ $\mu = E(X)$ and $k \times k$ $V = Cov(X)$ ($v_{ij} = Cov(X_i, X_j)$):

1. $E(AX) = AE(X)$

   Proof: Exercise 7.5(a).

2. $Var(a'X) = a'Va$.

   Proof: Note that

$$\begin{aligned} var(a'X) &= var(a_1X_1 + a_2X_2 + \ldots + a_kX_k) \\ &= \sum_{i=1}^{k}\sum_{j=1}^{k} a_ia_jCov(X_iX_j) \\ &= \sum_{i=1}^{k}\sum_{j=1}^{k} v_{ij}a_ia_j = a'Va \quad \square \end{aligned}$$

3. $Cov(AX) = AVA'$

   Proof Exercise 7.5(b).

4. $E(X'AX) = tr(AV) + \mu'A\mu$

   Proof: Let A_i be the ith row of A and a_{ij} be the ijth entry of A.
   Note that $tr(AV) = tr(A(E(XX') - EXEX')) = tr(AE(XX')) - tr(AEXEX')$.

$$\begin{aligned} tr(AE(XX')) &= tr\left(\begin{pmatrix} A_1E(XX') \\ \vdots \\ A_kE(XX') \end{pmatrix}\right) \\ &= \sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij}E(X_jX_i) \\ &= E\left(\sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij}X_jX_i\right) \\ &= E\left(\sum_{i=1}^{k} A_iXX_i\right) \\ &= E\left(\left(\sum_{i=1}^{k} X_iA_i\right)X\right) \\ &= E(X'AX) \end{aligned}$$

Meanwhile,

$$tr(AEXEX') = tr(EX'AEX) = EX'AEX = \mu' A\mu.$$

So we have $E(X'AX) = tr(AV) + \mu' A\mu$. $\square$

5. Covariance matrix V is positive semi-definite.

   Proof: $y'Vy = Var(y'X) \geq 0 \ \forall y \neq 0$. Since V is symmetric (why?), it follows that $V^{1/2} = (V^{1/2})'$. $\square$

6. $Cov(a'X, b'X) = a'Vb$

   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 $Cov(AX, BX) = AVB'$.

   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,

   $$c_{ij} = A_i V B_j = ij\text{th entry of } AVB'. \ \square$$

Exercises

7.1 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 $X \geq 0$.]

7.2 Let X be a continuous random variable with density f_X. Find the density of $Y = |X|$ in terms of f_X.

7.3 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 $Y_2 \sim Beta(\alpha_1, \alpha_2)$.

7.4 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 $t > 0$.

7.5 Prove the following results about vector expectations and variance:

(a) $E(Ax) = AE(x)$

(b) $Cov(Ax) = AVA'$

(c) $Cov(a'x, b'x) = a'Vb$

8 Further Applications to Statistics: Normal Theory and F-test

8.1 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 $X_1 \sim N(\mu_1, \sigma_1)$ and $X_2 \sim N(\mu_2, \sigma_2)$, 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

$$\Sigma = AA' = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

and $\rho = \frac{Cov(X_1, X_2)}{\sigma_1\sigma_2}$ (you should verify that the entries of $\Sigma = AA'$ 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

$$f_X(x) = \frac{1}{2\pi\sqrt{det(\Sigma)}} \exp\left\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right\}.$$

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,

$$\sqrt{det(\Sigma)} = \sqrt{det(AA')} = \sqrt{det(A)^2} = |det(A)| = \sigma_1\sigma_2\sqrt{1-\rho^2}$$

(you should verify the last step). We know that $f_Z(z) = \frac{1}{2\pi}\exp\left(-\frac{1}{2}z'z\right)$ and since $X = AZ + \mu$ we have by the result above:

$$\begin{aligned} f_X(x) &= \frac{1}{2\pi|det(A)|}\exp\left(-\frac{1}{2}(A^{-1}(x-\mu))'(A^{-1}(x-\mu))\right) \\ &= \frac{1}{2\pi|det(A)|}\exp\left(-\frac{1}{2}(x-\mu)'(A^{-1})'(A^{-1})(x-\mu)\right) \\ &= \frac{1}{2\pi|det(A)|}\exp\left(-\frac{1}{2}(x-\mu)'(AA')^{-1}(x-\mu)\right) \\ &= \frac{1}{2\pi\sqrt{det(\Sigma)}}\exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right) \end{aligned}$$

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 $N(B\mu + d, BVB')$.

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 $N(0, \sigma^2 AA')$ random vector.

8.2 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 $\frac{x'Ax}{\sigma^2} \sim {\chi_r}^2$, where $r = rank(A)$.

   Proof: We write $\frac{x'Ax}{\sigma^2} = \frac{x'Q}{\sigma}\Lambda\frac{Q'x}{\sigma}$ and we note that $\frac{Q'x}{\sigma} \sim N(0, \frac{1}{\sigma^2} \times \sigma^2 Q'Q) = N(0, I)$, i.e.$\frac{Q'x}{\sigma}$ 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 $r = rank(B) = rank(V^{1/2}AV^{1/2})$. 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 $\epsilon \sim N(0, \sigma^2)$, by one of the above results we have $\frac{\epsilon'(I-H)\epsilon}{\sigma^2} \sim {\chi_r}^2$, where $r = rank(I-H)$. But we know (why?) that $rank(I - H) = tr(I - H) = tr(I - X(X'X)^{-1}X') = tr(I) - tr(X(X'X)^{-1}X') = n - tr(X'X(X'X)^{-1}) = n - tr(I_{p\times p}) = n - p$. 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 $X_1' = X_1'X(X'X)^{-1}X'$ and $X_2' = X_2'X(X'X)^{-1}X'$. 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

$$\frac{\|\hat{Y}_f\|^2 - \|\hat{Y}_r\|^2}{\sigma^2} = \frac{Y'(X(X'X)^{-1}X' - X_1(X_1'X_1)^{-1}X_1')Y}{\sigma^2},$$

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

$$Y'(X(X'X)^{-1}X' - X_1(X_1'X_1)^{-1}X_1')Y$$

$$= \epsilon'(X(X'X)^{-1}X' - X_1(X_1'X_1)^{-1}X_1')\epsilon + \beta_1'X_1'(X(X'X)^{-1}X' - X_1(X_1'X_1)^{-1}X_1')X_1\beta_1.$$

We recognize the second summand as

$$(\beta_1'X_1'X(X'X)^{-1}X' - \beta_1'X_1'X_1(X_1'X_1)^{-1}X_1')X_1\beta_1 = (\beta_1'X_1' - \beta_1'X_1')X_1\beta_1 = 0.$$

So, letting $A = X(X'X)^{-1}X' - X_1(X_1'X_1)^{-1}X_1'$, 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 $\|e_f\|^2 = \epsilon'(I-H)\epsilon$. We proceed to show now that the two quadratic forms $\epsilon'(I-H)\epsilon$ and $\epsilon' A \epsilon$ are independent, by showing that $(I-H)\sigma^2 IA = \sigma^2(I-H)A = 0$. 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

$$\frac{\epsilon' A \epsilon}{p_2 \sigma^2} \Big/ \frac{\epsilon'(I-H)\epsilon}{(n-p)\sigma^2} = \frac{\|\hat{Y}_f\|^2 - \|\hat{Y}_r\|^2}{p_2 MS_{Res}} \sim F_{p_2, n-p}.$$

We can now test our null hypothesis $\beta_2 = 0$, using this statistic, and we would reject for large values of F.

Exercises

8.1 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 $N(B\mu + d, BVB')$.

8.2 Assume V is positive definite, AV, and $B = V^{\frac{1}{2}} A V^{\frac{1}{2}}$ is idempotent. Show that $rank(B) = rank(AV)$ (hint: consider the nullspaces, and invertible transformation $v = V^{1/2} w$).

8.3 Let $X = [X_1\, X_2]'$ for $n \times p_1$ X_1 and $n \times p_2$ X_2, and $A = X(X'X)^{-1}X' - X_1(X_1'X_1)^{-1}X_1'$. Show that A is symmetric and idempotent of rank p_2 (use trace to determine rank of A).

9 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. Montgomery, Douglas C. et al, 'Introduction to Linear Regression Analysis', 3rd ed., 2001, John Wiley & Sons

5. Horn, Roger A; Johnson, Charles R., 'Matrix Analysis', 1985, Cambridge University Press

6. Strang, G, 'Linear Algebra and Its Applications', 3rd ed., 1988, Saunders College Publishing