Linear Algebra

July 28, 2006

1 Introduction

These notes are intended for use in the warm-up camp for incoming Berkeley Statistics graduate students. Welcome to Cal! 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 already know. If you have never taken such a course before, you are strongly encouraged to do so by taking math 110 (or the honors version of it), or by covering material presented and/or mentioned here on your own. 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). Go Bears!

2 Vector Spaces

A set V is a **vector space** over $\mathbb{R}$ 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 $r \in \mathbb{R}$ 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 $\forall\ r_1, r_2 \in \mathbb{R}$:

1. $v_1 + v_2 = v_2 + v_1$
2. $(v_1 + v_2) + v_3 = v_1 + (v_2 + v_3)$
3. $\exists$ vector $0 \in V$, s.t. $v + 0 = v$, $\forall\ v \in V$
4. $\forall\ v \in V\ \exists\ -v = w \in V$ s.t. $v + w = 0$
5. $r_1(r_2 v) = (r_1 r_2)v$, $\forall\ v \in V$
6. $(r_1 + r_2)v = r_1 v + r_2 v$, $\forall\ v \in V$
7. $r(v_1 + v_2) = rv_1 + rv_2$, $\forall\ r \in \mathbb{R}$
8. $1v = v$, $\forall\ v \in V$

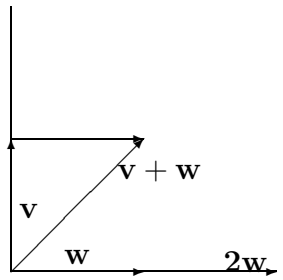

Figure 1: Vector Addition and Scalar Multiplication

Vector spaces over fields other than $\mathbb{R}$ are defined similarly, with the multiplicative identity of the field taking place of 1 in last property. We won't concern ourselves with those spaces, except for when we'll be needing complex numbers later on. Also, we'll be using symbol 0 to designate both 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 $0_{n\times 1}$.

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 $(rf)(x) = rf(x)$.

2.1 Basic Concepts

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, empty set is not allowed as a subspace.

Examples:

1. A subset $\{0\}$ is always a subspace of a vectors space V
2. Given vectors $v_1, v_2, \ldots, v_n \in V$, the set of all their linear combinations (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. You can also verify for yourself that the 2 conditions are indpendent 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.

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 $w = r_1v_1 + r_2v_2 + \ldots + r_nv_n$. If every vector in V is a linear combination of $v_1, v_2, \ldots, v_n$, then we say that $v_1, v_2, \ldots, v_n$ **span** V.

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 $r_1v_1 + r_2v_2 + \ldots + r_nv_n = 0 \Longrightarrow r_1 = r_2 = \ldots = r_n = 0$, 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.

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 $\{v_1, v_2, \ldots v_n\}$ be a basis for V, and let $\{w_1, w_2, \ldots, w_m\}$ be another basis for V. Then $n = m$.

Proof: Omitted, but can be found in any book on linear algebra.

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 Orthogonality

An inner product is a function $f : V \times V \to \mathbb{R}$ (which we denote by $f(v_1, v_2) =< v_1, v_2 >$), s.t. $\forall\ v, w, z \in V$, and $\forall r \in \mathbb{R}$:

1. $< v, w + rz > =< v, w > +r < v, z >$
2. $< v, w >=< w, v >$
3. $< v, v > \geq 0$ and $< v, v >= 0$ iff $v = 0$

We note here that not all vector spaces have inner products defined on them, but we will only be dealing with the ones that do. Examples:

1. Given 2 vectors $x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$ and $y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$ in $\mathbb{R}^n$, we define their inner product $x'y$ $=< x, y >= \sum_{i=1}^{n} x_i y_i$. You can check yourself that the 3 properties above are satisfied, and the meaning of notation $x'y$ will become clear from the next section.

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

We point out here the relationship in $\mathbb{R}^n$ between inner products and the length (or norm) of a vector. The length of a vector $x = \|x\| = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2} = \sqrt{x'x'}$, or $\|x\|^2 = x'x$.

We say that vectors v, w in V are **orthogonal** if $< v, w >= 0$. Notice that the zero vector is the only vector orthogonal to itself (why?).

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 $\|x\|^2 + \|y\|^2 = \|x - y\|^2$. Expending 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 $< x, y >= x'y = 0$.

2. Nonzero orthogonal vectors are linearly independent. Suppose we have $q_1, q_2, \ldots, q_n$, a set of nonzero mutually orthogonal ($< q_i, q_j >= 0 \ \forall i \neq j$) vectors in V, and suppose that $r_1 q_1 + r_2 q_2 + \ldots + r_n q_n = 0$. Then taking inner product of q_1 with both sides, we have $r_1 < q_1, q_1 > + r_2 < q_1, q_2 > + \ldots + r_n < q_1 q_n >=< q_1, 0 >= 0$. 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. We leave it as an exersise to show that $f(x) = 1$ and $g(x) = x$ are orthogonal in $F_{[-1,1]}$, and that if $n \neq m$ $f(x) = sin(nx)$ and $g(x) = sin(mx)$ are orthogonal in $F_{[0,2\pi]}$, and $sin(nx), cos(mx)$ are orthogonal in $F_{[0,2\pi]}$ for all values of n, m.

4. Suppose we have a $n \times 1$ vector of observations $x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$. Then if we let $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$, we can see that vector $e = \begin{bmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{bmatrix}$ is orthogonal to vector $\hat{x} = \begin{bmatrix} \bar{x} \\ \bar{x} \\ \vdots \\ \bar{x} \end{bmatrix}$, since $\sum_{i=1}^{n} \bar{x}(x_i - \bar{x}) = \bar{x} \sum_{i=1}^{n} x_i - \bar{x} \sum_{i=1}^{n} \bar{x} = n\bar{x}^2 - n\bar{x}^2 = 0.$

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 $\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$, 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 exersise that if $\dim(V) = n$, and $\dim(S) = k$, then $\dim(S^{\perp}) = n - k$.

2.3 Gram-Schmidt Process

Suppose we're given linearly independent vectors $v_1, v_2, \ldots, v_n$ in V, and there's 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?), and we can find 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 $< q_1, (v_2 - rq_1) >= 0$, and we get $r = \frac{<q_1, v_2>}{<q_1, q_1>}$. 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 $q_i = v_i - \left(\left(\frac{<q_1, v_i>}{<q_1, q_1>} \right) q_1 + \ldots + \left(\frac{<q_{i-1}, v_i>}{<q_{i-1}, q_{i-1}>} \right) q_{i-1} \right)$, 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. $< q_i, q_j >= 0 \ \forall i \neq j$ and $< q_i, q_i >= 1 \ \forall i$ (why?). We call these vectors that have length 1

unit vectors.

As an exersise, you can now construct an orthonormal basis for the subspace of $F_{[-1,1]}$ spanned by $f(x) = 1, g(x) = x$, and $h(x) = x^2$. 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.

3 Matrices and Matrix Alegbra

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

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

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 $x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n$ 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 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 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 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 $a_1, a_2, \ldots, a_n$ denote the respective columns of A, and $x_1, x_2, \ldots, x_n$ denote the components of x, we get a $m \times 1$ vector $Ax = x_1 a_1 + x_2 a_2 + \ldots + x_n a_n$, 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 $c_j = a_jB$, 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!).

Let A be $m \times n$, then we say that **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 $< x, y >$ is just $x'y$ according to matrix multiplication.

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

You should prove all of these identities as an exersise.

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 (prove this). 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 (think about nullspaces). 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 should 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 **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 **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 110).

We can see from the theorem, that the columns of A are linearly independent iff the nullspace doesnt' 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 $\mathbb{R}^m$, and therefore a solution exists for every b (though it may not be unique).

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

You should be able now to prove the following results:

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

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

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

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

4 Least Squares Estimation

4.1 Projections

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. $\|b\| = \|b'\| + \|b - b'\|$. 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).

We can now re-write the above situation in matrix terms. Let a_i be now the i-th column of the $m \times n$ matrix A. Then we want to find $x \in \mathbb{R}^n$ s.t. $(b - Ax) \perp col(A)$, or in other words $A'(b - Ax) = 0_{n\times 1}$. We now have $A'b = A'Ax$ or $x = (A'A)^{-1}A'b$ (why is $A'A$ invertible?). Then for every vector b in $\mathbb{R}^m$, its projection onto the column space of A is $Ax = A(A'A)^{-1}A'b$. We call the matrix $P = A(A'A)^{-1}A'$ that takes a vector in $\mathbb{R}^m$ and returns its projection onto $col(A)$ the **projection** matrix. We follow up with some properties of projection matrices that you should prove for yourself (unless the proof is supplied):

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

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

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 $S^\perp$

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

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

6. Every symmetric and idempotent matrix P is a projection. All we need to show if 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, $P'(b - Pb) = P'(I - P)b = P(I - P)b = (P - P^2)b = 0b = 0$.

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. $tr(P) = r$.

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 in statistics lectures of this camp) is $\hat{\beta}$ that minimizes $\sum_{i=1}^{n} (y_i - x_i\beta)^2$, where x_i is the i-th row of X. We recognize that in matrix notation $\hat{\beta}$ minimizes $(Y - X\beta)'(Y - X\beta)$. From this point there are two (or more ways) that we can arrive at the same conclusion.

First, we could recognize that the same $\hat{\beta}$ has to minimize $\|Y - X\beta\|$ (since the expresison above was just $\|Y - X\beta\|^2$, and $\|Y - X\beta\|$ is always non-negative), and therefore we pick

$\hat{\beta}$ is such a way that $X\beta$ is the projection of Y onto columns of X. Alternatively, we could differenitate $(Y - X\beta)'(Y - X\beta)$ with respect to β (see section on vector derivatives to find out how to carry out such a differentiation) and set it to zero (since if $\hat{\beta}$ minimizes the expression, the derivative at $\hat{\beta}$ should be 0) to get: $-X'Y - X'Y + 2X'X\hat{\beta} = 0$, or once again $\hat{\beta} = (X'X)^{-1}X'Y$. The projected values $\hat{Y} = X(X'X)^{-1}X'Y$ are known as **fitted values**, and the portion $e = Y - \hat{Y}$ of Y (which is orthogonal to the column space of X) is known as **residuals**.

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.

4.3 Matrix Derivatives and Other Identities

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 (which you should verify for yourself) together with 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. Let $a \in \mathbb{R}^k$, and let $y = a'x = a_1x_1 + a_2x_2 + \ldots + a_kx_k$. Then $\frac{\partial y}{\partial x} = a$
2. Let $y = x'x$, then $\frac{\partial y}{\partial x} = 2x$
3. Let A be $k \times k$, and a be $k \times 1$, and $y = a'Ax$. Then $\frac{\partial y}{\partial x} = A'a$
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.
5. $E(Ax) = AE(x)$

6. $Var(a'x) = var(a_1x_1 + a_2x_2 + \ldots + a_kx_k) = \sum_{i=1}^{k}\sum_{j=1}^{k} a_ia_jcov(x_ix_j) = \sum_{i=1}^{k}\sum_{j=1}^{k} v_{ij}a_ia_j = a'Va$

7. $Var(Ax) = AVA'$

8. $E(x'Ax) = tr(AV) + \mu'A\mu$

9. Covariance matrix V is positive semi-definite. Proof: $y'Vy = Var(y'x) >\geq 0 \ \forall y \neq 0$. Since V is symmetric (why?), prove for yourself that $V^{1/2} = (V^{1/2})'$ (this requires knowing the results on diagonalization of symmetric matrices to be presented later).

10. $Cov(a'x, b'x) = a'Vb$

11. If x, y are two $k \times 1$ vectors of random varialbes, 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 $cov(Ax, Bx) = AVB'$.

5 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.

5.1 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 we have that the determinant of the new matrix is 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 (why?)

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

9. $\det(AB) = \det(A)\det(B)$. Proof: Suppose $\det(B) = 0$. Then B is not invertible, and AB is not invertible, therefore $\det(AB) = 0$. If $\det(B) \neq 0$, show that $d(A) = \frac{det(AB)}{det(B)}$ satisfies the first 3 properties, and therefore $d(A) = \det(A)$. In particular $\det(A^{-1}) = \frac{1}{det(A)}$.

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.

5.2 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 eigenvecors 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.

We can show that eigenvectors corresponding to distinct eigenvalues are linearly independent. 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 and is left as exersise.

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 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 multiple multiplication we can see that $AS = S\Lambda$, and since S is invertible (why?) we have $S^{-1}AS = \Lambda$. 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 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 repated 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. 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).

5.3 Complex Matrices and Basic Results

We know allow complex entries in vectors and matrices. Scalar multiplicaiton now also allows multiplication by complex numbers, so we're gonna 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 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 $ntimesn$ 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 (which also include as special cases symmetric and orthogonal matrices respectively):

1. If A is Hermitian, then $\forall x \in \mathbb{C}^n$, $y = x^H Ax \in \mathbb{R}$. Proof: taking the hermitian transpose we have $y^H = x^H A^H x = x^H Ax = y$, and the only scalars in $\mathbb{C}$ that are equal to their own conjugates are the reals.

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 $\lambda \in \mathbb{R}$.

3. If A is positive definite, and λ is an eigenvalue of A, then $\lambda > 0$ (note that since A is symmetric, we know that $\lambda \in \mathbb{R}$). Proof: Let nonzero x be the eigenvector corresponding to λ. Then since A is positive definite, we have $x'Ax > 0 \Longrightarrow x'(\lambda x) > 0$ $\Longrightarrow \lambda \|x\|^2 > 0 \Longrightarrow \lambda > 0$.

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 $x^H y = 0$.

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 $\|Ux\| = \|x\|$.

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 $|\lambda| = 1$.

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, $x^H y = 0$.

9. For EVERY square matrix A, $\exists$ some unitary matrix U s.t. $U^{-1}AU = U^H AU = T$, 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 $B = U^{-1}AU$ is normal. Proof: $BB^H = (U^HAU)(U^HAU)^H = U^HAUU^HA^HU = U^HAA^HU = U^HA^HAU$ (since A is normal) $= U^HA^HUU^HAU = (U^HAU)^H(U^HAU) = B^HB$.

11. If $n \times n$ A is normal, then $\forall x \in \mathbb{C}^n$ we have $\|Ax\| = \|A^Hx\|$. Proof: $\|Ax\|^2 = (Ax)^HAx = x^HA^HAx = x^HAA^Hx = (A^Hx)^H(A^Hx) = \|A^Hx\|^2$. And since $\|Ax\| > 0 < \|A^Hx\|$, we have $\|Ax\| = \|A^H\|$.

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

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. $U^{-1}AU = T$, 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. $B^2 = A$. 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.

16. If A is symmetric and idempotent, and λ is an eigenvalue of A, then $\lambda = 1$ or $\lambda = 0$. Proof: we know that $A = Q\Lambda Q'$ and $A^2 = A$, therefore $Q\Lambda Q'Q\Lambda Q' = Q\Lambda^2 Q' = Q\Lambda Q'$, and we conclude that $\Lambda = \Lambda^2$. You should prove for yourself that this implies that diagonal entries of Λ are either 0 or 1, and that the number of 1's along the main diagonal of $\Lambda = rank(A)$. Why is this another proof that $rank(A) = tr(A)$ for symmetric and idempotent A?

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 exersise!) 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 Further Applications to Statistics: Normal Theory and F-test

6.1 Bivariate Normal Distribution

Suppose X is a vector of continuous random variables and $Y = AX + c$, where A is an invertible matrix. Then if X has probability density funciton p_X, then the probability density function of Y is given by $p_Y(y) = |det(A)|^{-1} p_X(A^{-1}(Y - c))$. 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 $X \sim N(0_{2\times 1}, \Sigma)$, 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

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

where *exp* denotes the exponential function.

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 $p_Z(z) = \frac{1}{2\pi}exp\left(-\frac{1}{2}z'z\right)$ and since $X = AZ + \mu$ we have by the result above:

$$p_X(x) = \frac{1}{2\pi|det(A)|} exp\left(-\frac{1}{2}(A^{-1}(x-\mu))'(A^{-1}(x-\mu))\right)$$

$$= \frac{1}{2\pi|det(A)|} exp\left(-\frac{1}{2}(x-\mu)'(A^{-1})'(A^{-1})(x-\mu)\right) = \frac{1}{2\pi|det(A)|} exp\left(-\frac{1}{2}(x-\mu)'(AA')^{-1}(x-\mu)\right)$$

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

which proves the theorem. The symmetric matrix Σ is the covariance matrix of X.

You should prove for yourself now 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.

6.2 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 $\frac{x'Ax}{\sigma^2} \sim {\chi_r}^2$.

2. The above result generalizes further: suppose $k \times 1$ $x \sim N(0, V)$, and $k \times k$ symmetric A is s.t. AV or VA is idempotent. Then $x'Ax \sim {\chi_r}^2$, where $r = rank(AV)$ or $rank(VA)$, respectively. 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 exersise now to show that $rank(B) = rank(AV)$ (hint: consider the nullspaces, and invertible transformation $v = V^{1/2}w$).

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.

Now, let's go back to our linear system from the first lecture. Recall 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 $\hat{Y} = X\hat{\beta} = PY$, where $P = X(X'X)^{-1}X'$ is the projection matrix onto columns of X, and $e = Y - \hat{Y} = (I - P)Y$ 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. $\hat{Y} \sim N(X\beta, \sigma^2 P)$. Proof: Clearly, $Y \sim N(X\beta, \sigma^2 I)$, and $\hat{Y} = PY \Longrightarrow \hat{Y} \sim N(PX\beta, P\sigma^2 IP') = N(X(X'X)^{-1}X'X\beta, \sigma^2 PP') = N(X\beta, \sigma^2 P)$.

2. $e \sim N(0, \sigma^2(I - P))$. Proof is analgous and is left as an exersise.

3. $\hat{Y}$ and e are independent (in probabilistic sense of the word). Proof: $cov(\hat{Y}, e) = cov(PY, (I - P)Y) = P(var(Y))(I - P) = P\sigma^2 I(I - P) = \sigma^2 P(I - P) = 0$. And since both vectors were normally distributed, zero correlation implies independence. Notice that *cov* above referred to the cross-covariance matrix.

4. $\frac{\|e\|^2}{\sigma^2} \sim {\chi^2}_{n-p}$. Proof: First notice that $e = (I - P)Y = (I - P)(X\beta + \epsilon) = (I - P)\epsilon$ (why?). Now, $\frac{\|e\|^2}{\sigma^2} = \frac{e'e}{\sigma^2} = \frac{\epsilon'(I-P)'(I-P)\epsilon}{\sigma^2} = \frac{\epsilon'(I-P)\epsilon}{\sigma^2}$. Since $(I - P)$ is symmetric and idempotent, and $\epsilon \sim N(0, \sigma^2)$, by one of the above results we have $\frac{\epsilon'(I-P)\epsilon}{\sigma^2} \sim {\chi_r}^2$, where $r = rank(I - P)$. But we know (why?) that $rank(I - P) = tr(I - P) = tr(I - X(X'X)^{-1}X') = tr(I) - tr(X(X'X)^{-1}X') = n - tr(X'X(X'X)^{-1}) = n - tr(I_{p\times p}) = n - p$. So we have $\frac{\|e\|^2}{\sigma^2} \sim {\chi^2}_{n-p}$, and in particular $E(\frac{\|e\|^2}{n-p}) = \sigma^2$.

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 $[X_1 X_2] = X(X'X)^{-1}X'[X_1, X_2] \Longrightarrow X_1 = X(X'X)^{-1}X'X_1$ and $X_2 = X(X'X)X'X_2$ (by straightforward matrix multiplicaiton) or $PX_1 = X_1$ and $PX_2 = X_2$. 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 that A is symmetric and idempotent of rank p_2, and therefore $\frac{\epsilon'A\epsilon}{\sigma^2} \sim {\chi^2}_{p2}$ (use trace to determine rank of A). 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 $\|e_f\|^2 = \epsilon'(I-P)\epsilon$. We proceed to show now that the two quadratic forms $\epsilon'(I-P)\epsilon$ and $\epsilon'A\epsilon$ are independent, by showing that $(I-P)\sigma^2IA = \sigma^2(I-P)A = 0$. The proof is left as an exersise for you. We will now denote $\frac{\|e_f\|^2}{n-p}$ by MS_{Res}, and we conclude that $\frac{\epsilon'A\epsilon}{p_2\sigma^2} / \frac{\epsilon'(I-P)\epsilon}{(n-p)\sigma^2} = \frac{\|\hat{Y}_f\|^2 - \|\hat{Y}_r\|^2}{p_2 MS_{Res}} \sim F_{p_2,n-p}$. We can

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

6.3 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 know as **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 (why?). Arrange the eigenvectors $x_1, x_2, \ldots, x_n$ is 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. You should easily verify for yourself that $x_{r+1}, x_{r+2}, \ldots, x_n$ form a basis for $N(A)$ and there-

fore $x_1, x_2, \ldots, x_r$ form a basis for 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 prove 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 $\mathbb{R}^m$, $q_1, q_2, \ldots, q_r, q_{r+1}, \ldots, q_n$. You should verify for yourself that $q_1, q_2, \ldots, q_r$ form orthonormal basis for column space of A and that therefore $q_{r+1}, q_{r+2}, \ldots, q_n$ form an orthonormal basis for left nullspace of A. We now verify that $A = Q_1 \Sigma Q_2'$ by checking that $Q_1'AQ_2 = \Sigma$. Consider ij-th entry of $Q_1'AQ_2$. It is equal to $q_i'Ax_j$. 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 $Q_1'AQ_2 = \Sigma$, as claimed.

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.

7 References

1. Bickel, Peter J; Doksum, Kjell A., 'Mathematical Statistics: Basic Ideas and Selected Topics', 2nd ed., 2001, Prentice Hall

2. Freedman, David A., 'Statistical Models: Theory and Applications', 2005, Cambridge University Press

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

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