Stat 159/259: Linear Algebra Notes

Jarrod Millman

November 16, 2015

Abstract

These notes assume you've taken a semester of undergraduate linear algebra. In particular, I assume you are familiar with the following:

- solving systems of linear equations using Gaussian elimination
- linear combinations of vectors to produce a space
- the rank of a matrix
- the Gram–Schmidt process for orthonormalising a set of vectors
- eigenvectors and eigenvalues of a square matrix

I will briefly review some of the above topics, but if you haven't seen them before the presentation may be difficult to follow. If so, please review your introductory linear algebra textbook or use Wikipedia.

Background

Introductory linear algebra texts often start with an examination of simultaneous systems of linear equations. For example, consider the following system of linear equations

$$\begin{aligned} 2x_1 + x_2 &= 3 \\ x_1 - 3x_2 &= -2. \end{aligned}$$

For convenience, we abbreviate this system

$$\underbrace{\begin{bmatrix} 2 & 1 \\ 1 & -3 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} 3 \\ -2 \end{bmatrix}}_{b}$$

where A is the coefficient matrix, x is a column vector of unknowns, and b is a column vector of constants.

You should verify that

$$x = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

solves the given system.

More generally, a real-valued matrix A is an ordered, $m \times n$ rectangular array of real numbers, which when multiplied on the right by an n-dimensional column vector x produces an m-dimensional column vector b. Having introduced the $Ax = b$ notation, we can view the $m \times n$ matrix A as a linear function from $\mathbf{R}^n$ to $\mathbf{R}^m$. A linear function is one that respects proportions and for which the effect of a sum is the sum of individual effects. That is for any real numbers a and b and any n-dimensional real vectors x and y, a function $A : \mathbf{R}^n \to \mathbf{R}^m$ is called a *linear function* if the following identity holds

$$A(ax + by) = aA(x) + bA(y).$$

Linear functions preserve linear combinations.

Before returning our attention to real-valued matrices, let's remove some detail to see what the basic objects of linear algebra are in an abstract setting.

Vector spaces and linear transforms

Linear algebra is the study of linear transformations of vector spaces over some field. To make this precise, we need to define a few algebraic structures.

Definition 1. A *group* $(G, \cdot)$ is a set G and a binary operation $\cdot : G \times G \to G$ called (group) *multiplication* such that

1. for all a and b in G, ab is in G (i.e., G is *closed* under multiplication),
2. for all a, b, and c in G, $(ab)c = a(bc)$ (i.e., multiplication is *associative*),
3. there exists an *identity* element e in G such that $eg = ge = g$ for all g in G, and
4. for each g in G there exists an *inverse* element g^{-1} in G such that $g^{-1}g = gg^{-1} = e$.

If the group multiplication is *commutative* (i.e., $ab = ba$) then we say the group is a *commutative group*. Given a group $(G, \cdot)$, a subset H of G is called a *subgroup* if $(H, \cdot)$ is a group under the multiplication of G.

Example. The integers under addition $(\mathbf{Z}, +)$ is a group with identity element 0 and where the inverse of z is $-z$. The even integers are a subgroup of $(\mathbf{Z}, +)$. The rational numbers excluding 0 under multiplication $(\mathbf{Q} \setminus \{0\}, \cdot)$ is a group with identity element 1 and where the inverse of p/q is q/p. In both cases, the reader should verify that the group properties are met.

Definition 2. A *field* $(F, \times, +)$ is a set F and two binary operations called *field multiplication* $\times : F \times F \to F$ and *field addition* $+ : F \times F \to F$ such that $(F, +)$ is a commutative group with identity 0 and $(F \setminus \{0\}, \times)$ is a commutative group with identity denoted 1 such that multiplication distributes over addition $a(b + c) = ab + ac$.

We will mostly be interested in the field of real numbers $\mathbf{R}$ and the field of complex numbers $\mathbf{C}$. However, note that the rationals $\mathbf{Q}$ equipped with the standard operations of addition and multiplication form a field, while the integers $\mathbf{Z}$ do not (Why?).

Definition 3. A *vector space* $(V, F, \oplus, \otimes)$ is set V of *vectors* over a field F of *scalars* with the usual field operations equipped with *vector addition* $\oplus$ and *scalar multiplication* $\otimes$ such that $(V, \oplus)$ is a commutative group and scalar multiplication $\otimes : F \times V \to V$ satisfies the following identities:

$$\begin{aligned} a \otimes (v \oplus w) &= (a \otimes v) \oplus (a \otimes w), & 1 \otimes v &= v, \\ (a + b) \otimes v &= (a \otimes v) \oplus (b \otimes v), & (ab) \otimes v &= a \otimes (b \otimes v) \end{aligned}$$

for all vectors v and w in V and all scalars a and b in F.

While I used the symbols $\oplus$ and $\otimes$ in the definition above to distinguish vector addition from field addition and scalar multiplication from field multiplication, in the sequel I will follow standard practice and overload the multiplication and addition symbols. The type of multiplication and addition will be made clear from the context.

The two operations of vector addition and scalar multiplication lead to the more general notion of *linear combinations* such as

$$a_1 v_1 + a_2 v_2 + \cdots + a_n v_n = \sum a_i v_i$$

of n vectors v_i in V and n scalars a_i in F. All linear combinations of a set of n vectors $v_1, v_2, \ldots, v_n$ of V form a *subspace* of V *spanned* by $v_1, v_2, \ldots, v_n$. A set of n vectors $v_1, v_2, \ldots, v_n$ of V are *linearly independent* if $\sum a_i v_i = 0$ implies that each a_i is 0. A set of n vectors $v_1, v_2, \ldots, v_n$ of V form a *basis* of a (finite-dimensional) vector space when every vector in V is uniquely expressed as a linear combination of $v_1, v_2, \ldots, v_n$.

Definition 4. Given two vector spaces V and W over the same field F, a function $T : V \to W$ is called a *linear transform* if for all vectors x and y in V and all scalars a and b in F

$$T(ax + bv) = aT(x) + bT(y).$$

While the focus of linear algebra concerns linear transformations of vector spaces, we will often want to impose additional structure of our vector spaces. In particular, we would like to equip our spaces with some notion of distance and angle.

Definition 5. Let V be a vector space over the field F with the usual operations. Then an *inner product* is any function

$$\langle \cdot, \cdot \rangle : V \times V \to F$$

which for all x, y, and z in V and all a and b in F the following hold

- **(Symmetry)** $\langle x, y \rangle = \langle y, x \rangle$,
- **(Positive definite)** $\langle x, x \rangle \geq 0$ and $\langle x, x \rangle = 0$ implies $x = 0$, and
- **(Left linearity)** $\langle ax + by, z \rangle = a\langle x, z \rangle + b\langle y, z \rangle$.

Example. If x is in $\mathbf{R}^n$, then the dot product is an inner product. That is,

$$\langle x, y \rangle = x^\top y = \sum x_i y_i = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n.$$

You should verify that the dot product satisfies the properties for an inner product.

A vector space equipped with an inner product is called an *inner product space.* Once we have an inner product we can define what we mean by distance and angle. Thus equipping the purely algebraic structure of a vector space with an inner product enables us to rigorously define several geometric ideas such as the length of a vector or the angle between two vectors. Once we can define the angle between vectors we will be able to establish a notion of orthogonality (i.e., when are two vectors "perpendicular"). Inner product spaces are an abstraction of Euclidean space, as inner products are an abstraction of dot products.

Definition 6. Let V be a vector space over a field F equipped with an inner product. Then any two vectors x and y are said to be *orthogonal*, denoted $x \perp y$, if their inner product is 0.

Two subspaces U and W of V are called orthogonal if every vector in U is orthogonal to every vector in W.

Definition 7. Let V be a vector space over a field F equipped with an inner product and let W be any subspace of V. The *orthogonal complement* of W in V, denoted $W^\perp$, is the set of all vectors in V orthogonal to all vectors in w; that is,

$$W^\perp \equiv \{x \in V : x \perp w \text{ for all } w \in W\}.$$

Theorem 1 (Orthogonal decomposition)**.** *Let V be a vector space over a field F equipped with an inner product and let W be any subspace of V. Then any vector x in V can be uniquely written as the sum of one element from W and one element from $W^\perp$.*

Proof. Clearly $W \cap W^\perp = \{0\}$. Put $U = W + W^\perp$ (the space of vectors obtained by adding all possible pairs chosen one from W and one from $W^\perp$). (Check that this is a subspace of V.) Now take an orthonormal basis for U. To show that $U = V$, assume not for a contradiction. If $U \neq V$, then extend the orthonormal basis for U to an orthonormal basis for V. Let z be one of the new basis vectors in the extension. Then $z \perp U$ by construction. Since $W \subset U$, $z \perp W$ as well. But then z must be an element of $W^\perp$, a contradiction. Hence $U = W + W^\perp = V$. To show that there is a unique representation, assume not for a contradiction. If not, then there must be an x in V such that there are two distinct, non-zero vectors $w_1 \neq w_2$ in W and two distinct, non-zero vectors $u_1 \neq u_2$ in $W^\perp$ such that $x = w_1 + u_1$ and $x = w_2 + u_2$. This implies that $(w_1 - w_2) + (u_1 - u_2) = 0$, a contradiction. □

Corollary. *Let V be a vector space over a field F equipped with an inner product and let W be any subspace of V. Then*

$$\dim V = \dim W + \dim W^\perp.$$

Definition 8. Let V be a vector space over a subfield F of the complex numbers with the usual operations. Then a function

$$\|\cdot\| : V \to \mathbf{R}$$

is called a *norm* on V if for every x and y in V and all a in F the following properties hold

- **(Positivity)** $\|x\| \geq 0$; and $\|x\| = 0$ implies $x = 0$,
- **(Absolute homogeneity)** $\|ax\| = |a|\|x\|$, and
- **(Triangle inequality)** $\|x + y\| \leq \|x\| + \|y\|$.

Note. This definition is less general than the previous ones. This is due to the desire for the range of a norm to be a totally ordered field (e.g., **R** is totally ordered, but **C** is not).

Given an inner product $\langle \cdot, \cdot \rangle$, there is a natural norm for the space defined by that inner product

$$\|x\| \equiv \sqrt{\langle x, x \rangle}.$$

If the inner product is the dot product, this is the usual Euclidean norm

$$\|x\| \equiv \sqrt{x^\top x}.$$

Matrices

There are four fundamental subspaces associated with any real-valued $m \times n$ matrix A. Its *nullspace* $N(A)$ contains all the vectors x that satisfy $Ax = 0$. Its *column space* $R(A)$ is the set of all linear combinations of its columns. The *row space* of A is the column space $R(A^\top)$. Similarly, the *left nullspace* is the nullspace $N(A^\top)$.

The properties of the four fundamental subspaces are collected in the fundamental theorem of the field.

Theorem 2 (The fundamental theorem of linear algebra)**.** *Let A be a $m \times n$ real-valued matrix with rank r. Then*

- $\dim R(A^\top) = \dim R(A) = r$
- $\dim N(A^\top) = m - r$
- $\dim N(A) = n - r$

and

- $R(A^\top) \perp N(A), \quad R(A^\top)^\perp = N(A), \quad N(A)^\perp = R(A^\top)$
- $R(A) \perp N(A^\top), \quad R(A)^\perp = N(A^\top), \quad N(A^\top)^\perp = R(A)$

Proof. That the row and column ranks are equal is usual shown in introductory courses as a result of reducing A to its row echelon form. Here is another approach. Put $r \equiv \dim R(A)$ and let $u_1, u_2, \ldots, u_r$ be a basis for the column space of A. Necessarily each column of A is a unique linear combination of this basis. Hence we can decompose A as the product of two matrices U and W where the columns of U are the basis vectors for the column space of A and the columns of W are the unique coefficients (or weights) needed to combine the basis vectors to produce the corresponding columns of A.

$$A = \underbrace{\begin{bmatrix} | & & | \\ u_1 & \cdots & u_r \\ | & & | \end{bmatrix}}_{U_{m\times r}} \underbrace{\begin{bmatrix} | & & | \\ w_1 & \cdots & w_n \\ | & & | \end{bmatrix}}_{W_{r\times n}} = \begin{bmatrix} | & & | \\ Uw_1 & \cdots & Uw_n \\ | & & | \end{bmatrix}$$

Furthermore, the rows of A can be written as linear combinations of the r rows of W. Thus a basis for the row space must not exceed the number of rows in W. Hence, I conclude $\dim R(A) \geq \dim R(A^\top)$. Applying the same reasoning to $A^\top$ yields $\dim R(A) \leq \dim R(A^\top)$. Combining these two inequalities, I conclude the two dimensions must be equal.

The remaining dimensionality results follow from the earlier orthogonal decomposition theorem once we've proven the orthogonality properties.

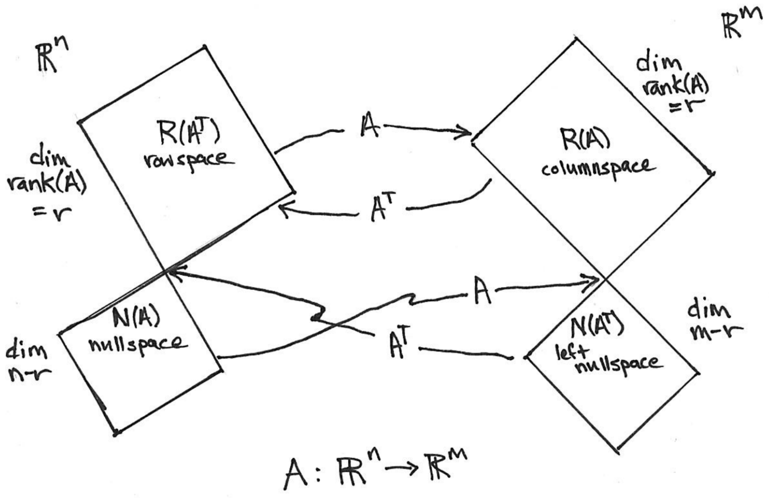

Figure 1: The four fundamental subspaces.

To prove the first orthogonality property, we first note that for any n-dimensional vector x in the nullspace of A, it must be that $Ax = 0$. Since $Ax = 0$, x is orthogonal to each row of A. Thus x is orthogonal to every linear combination of the rows of A. The second orthogonality property follows from the same argument applied to $A^\top$. □

The theorem is illustrated schematically in Figure 1 and will be made concrete during our discussion of the singular value decomposition.

The fundamental theorem of linear algebra and the singular value decomposition (SVD) are intimately related. However, before discussing the SVD of an arbitrary real-valued matrix, we will first consider certain nice matrices. In particular, we will first look at (1) square matrices, then (2) square, symmetric matrices, and finally (3) square, symmetric, positive (semi-)definite matrices. As we increasingly restrict our attention to more and more special classes of matrices, our matrix decompositions will become nicer. Once we derive the properties of the matrix decompositions of these restricted matrices, we will use them to great effect when we return to the general case.

Square matrices

A square matrix A in $\mathbf{R}^{n\times n}$ when multiplied on the right (or left) by a vector x in $\mathbf{R}^n$ will produce another vector y also in $\mathbf{R}^n$. So it is possible that there may be directions in which A behaves like a scalar. That is there may be special vectors v_i such that Av_i is just a scaled version of v_i.

Definition 9. Given A in $\mathbf{R}^{n\times n}$ a scalar λ is called an *eigenvalue* and a vector v its associated *eigenvector* if $Av = \lambda v$.

If we had n such eigenvalue-eigenvector pairs $(\lambda_1, v_1), (\lambda_2, v_2), \ldots, (\lambda_n, v_n)$, then putting the v_i in the columns of S we can perform the following calculation

$$\begin{aligned}
AS &= A \begin{bmatrix} | & & | \\ v_1 & \cdots & v_n \\ | & & | \end{bmatrix} \\
&= \begin{bmatrix} | & & | \\ Av_1 & \cdots & Av_n \\ | & & | \end{bmatrix} \\
&= \begin{bmatrix} | & & | \\ \lambda_1 v_1 & \cdots & \lambda_n v_n \\ | & & | \end{bmatrix} \\
&= \begin{bmatrix} | & & | \\ v_1 & \cdots & v_n \\ | & & | \end{bmatrix} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} \\
&= S\Lambda.
\end{aligned}$$

Furthermore, if the eigenvectors are linearly independent then S is full rank, so its inverse S^{-1} exists. Thus, we can *diagonalize* A by

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

or *decompose* A as

$$A = S\Lambda S^{-1}.$$

But how could we find them? We can use Gaussian elimination to solve $Ax = b$ for x when A and b are given (that is, if b is in the column space of A). But that will not work for solving $Ax = \lambda x$, since we now have two unknowns x and λ. However, notice that if we move the RHS over to the LHS of the equation and pull x outside the sum, then

$$(A - \lambda I)x = 0.$$

So if we hope to find a non-zero eigenvector x, then $(A - \lambda I)$ will have to have a non-trivial null space. In other words, $(A - \lambda I)$ must be singular. Hence,

$$\det(A - \lambda I) = 0.$$

Thus the roots of the $\det(A - \lambda I)$ are the eigenvalues of A. This function $\det(A - \lambda I)$ is called the *characteristic polynomial* of A. It is a nth degree polynomial in λ and by the fundamental theorem of algebra and the factor theorem it has n not necessarily distinct roots in the complex plane.

Symmetric matrices

A square matrix A in $\mathbf{R}^{n \times n}$ is *symmetric* if $A = A^\top$.

Property 1 ($S^{-1} = S^\top$)**.** If a symmetric matrix A is full rank, then it is diagonalized by $S^\top A S = \Lambda$ and decomposed as $A = S \Lambda S^\top$.

Proof. By virtue of being a square matrix,

$$A = S \Lambda S^{-1}.$$

Transposing both sides yields

$$A^\top = (S^{-1})^\top \Lambda S^\top.$$

Since $A = A^\top$,

$$S \Lambda S^{-1} = (S^{-1})^\top \Lambda S^\top.$$

I conclude $S^{-1} = S^\top$; and, since A is square, **Property 1** follows. □

Property 2 (Real eigenvalues)**.** A symmetric, real-valued matrix A has real eigenvalues.

Proof. Let λ be an eigenvalue of A and v its associated eigenvector. Then

$$Av = \lambda v \tag{1}$$

implies

$$A\overline{v} = \overline{\lambda}\overline{v} \tag{2}$$

where $\overline{v}$ is the complex conjugate of v. Multiplying equation 1 on the left by $\overline{v}^\top$ yields

$$\overline{v}^\top A v = \lambda \overline{v}^\top v. \tag{3}$$

Transposing both sides of equation 2 and replacing $A^\top$ with A gives

$$\overline{v}^\top A = \overline{\lambda}\overline{v}^\top. \tag{4}$$

Finally multiplying both sides of equation 4 on the right by v yields

$$\overline{v}^\top A v = \overline{\lambda}\overline{v}^\top v. \tag{5}$$

Since equation 3 and equation 5 have the same LHS,

$$\lambda \overline{v}^\top v = \overline{\lambda}\overline{v}^\top v.$$

Canceling $\overline{v}^\top v$ from both sides, I conclude $\lambda = \overline{\lambda}$. □

Property 3 (Orthogonal eigenvectors)**.** For a symmetric, real-valued matrix A with distinct eigenvalues $\lambda_i \neq \lambda_j$, the associated eigenvectors are orthogonal $v_i \perp v_j$.

Proof. Since $(Av_i)^\top = \lambda_i v_i^\top$, $A^\top = A$, and $Av_j = \lambda_j v_j$,

$$\begin{aligned}\lambda_i v_i^\top v_j &= v_i^\top A^\top v_j \\ &= v_i^\top A v_j \\ &= \lambda_j v_i^\top v_j.\end{aligned}$$

Subtracting the LHS from both sides of the above equation yields

$$(\lambda_i - \lambda_j) v_i^\top v_j = 0.$$

Since $(\lambda_i - \lambda_j) \neq 0$, I conclude $v_i^\top v_j = 0$. □

Let's summarize these nice properties in a theorem.

Theorem 3 (Spectral Theorem)**.** *Let A be a symmetric, real-valued $n \times n$ matrix. Then A can be decomposed as*

$$A = Q \Lambda Q^\top$$

where Q is an orthogonal matrix (i.e., $Q^\top Q = I$) and Λ is a diagonal matrix containing the n real eigenvalues, with multiplicity, of A sorted in non-increasing order $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$.

I leave it as an exercise to show that for real symmetric matrices, you can get orthogonal eigenvectors for repeated eigenvalues. The argument is straightforward, but a bit cumbersome.

Positive (semi-)definite matrices

Definition 10. A symmetric, real-valued $n \times n$ matrix A is *positive definite* if for all nonzero x in $\mathbf{R}^n$,

$$x^\top A x > 0.$$

If we replace the strict inequality $>$ with $\geq$ above, then we say A is *positive semi-definite.*

Theorem 4. *A symmetric, real-valued, positive (semi-)definite matrix A has (non-negative) positive eigenvalues.*

Proof. Let $A = Q\Lambda Q^\top$ be the spectral decomposition of A and let $u_1, u_2, \ldots, u_n$ be the eigenvectors of A (i.e., the columns of Q. Since A is symmetric, the eigenvectors are unitary $Q^\top Q = I$. Hence, if we take x to be u_i for i in $\{1, 2, \ldots, n\}$,

$$\begin{aligned} u_i^\top A u_i &= u_i^\top Q \Lambda Q^\top u_i \\ &= \lambda_i. \end{aligned}$$

Since A is positive definite, I conclude $\lambda_i > 0$. If A is positive semi-definite, then $\lambda_i \geq 0$. □

Definition 11. Let A be a $m \times n$ matrix. Then the $n \times n$ matrix $A^\top A$ is called the *Gram matrix* of A.

Theorem 5. *Let A be any $m \times n$ matrix. Then its Gram matrix $A^\top A$ is positive semi-definite.*

Proof. For any x in $\mathbf{R}^n$,

$$\begin{aligned} x^\top A^\top A x &= (Ax)^\top (Ax) \\ &= \|Ax\|^2 \\ &\geq 0. \end{aligned}$$

□

Corollary. *Let A be any $m \times n$ matrix of rank n. Then its Gram matrix $A^\top A$ is positive definite.*

Note that we could have taken $A^\top$ instead of A above, which would lead to the Gram matrix $AA^\top$.

Singular Value Decomposition

We have seen that if A has a particularly nice form (i.e., square and symmetric), we can decompose it into the product of an orthonormal basis, a diagonal matrix, and the transpose of the first matrix. Further, if A is real-valued, then all the entries of the matrix decomposition will also be real. This decomposition is particularly nice. We can directly read off several properties of A by simply looking at its spectral decomposition. For example, the number of non-zero eigenvalues is the rank of A.

It would be extremely useful to be able to similarly decompose an arbitrary real-valued $m \times n$ matrix A. However, since we are now dealing with a rectangular matrix, so the row space is a subspace of $\mathbf{R}^m$ and the column space is a subspace of $\mathbf{R}^n$. So rather than needing just one orthonormal basis, we would like to

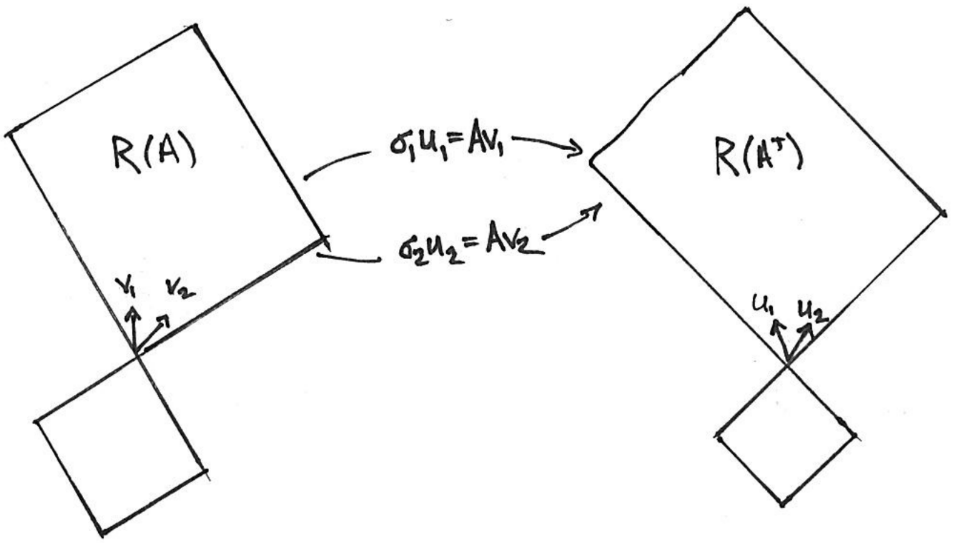

Figure 2: The goal.

decompose A into an orthonormal basis V in the row space which goes to a scaled version of an orthonormal basis U for the column space

$$AV = U\Sigma$$

where $V^\top V = I$, $U^\top U = I$, and Σ is a diagonal matrix of non-negative real numbers.

See Figure 2 and Figure 3.

Theorem 6 (Singular Value Decomposition)**.** *Let A be a real-valued $m \times n$ matrix. Then its singular value decomposition is $A = U\Sigma V^\top$.*

Proof. If $A = U\Sigma V^\top$, then

$$\begin{aligned} AA^\top &= (U\Sigma V^\top)(U\Sigma V^\top)^\top \\ &= (U\Sigma V^\top)(V\Sigma U^\top) \\ &= U\Sigma^2 U^\top \end{aligned}$$

and

$$\begin{aligned} A^\top A &= (U\Sigma V^\top)^\top(U\Sigma V^\top) \\ &= (V\Sigma U^\top)(U\Sigma V^\top) \\ &= V\Sigma^2 V^\top. \end{aligned}$$

Hence if we could decompose $AA^\top = U\Sigma^2 U^\top$ and $A^\top A = V\Sigma^2 V^\top$. □

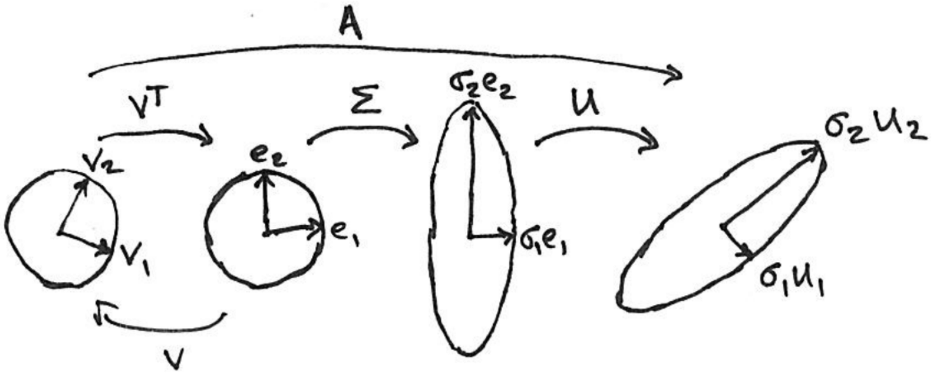

Figure 3: The action of A via its SVD.

Variational characterization of eigenvalues

Definition 12. For any square $n \times n$ matrix A and any n-dimensional vector x, the Rayleigh Quotient is

$$\frac{x^\top Ax}{x^\top x}.$$

Note that if $x^\top x = 1$, this is equivalent to $x^\top Ax$.

Theorem 7 (Rayleigh Quotient Theorem)**.** *Let A be a symmetric matrix in $\mathbf{R}^{n\times n}$ and let $A = Q\Lambda Q^\top$ be its spectral decomposition. Since A has real eigenvalues, let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Then*

$$\lambda_1 = \max_{\|x\|=1} x^\top Ax$$
$$\lambda_n = \min_{\|x\|=1} x^\top Ax$$

Proof. For any $\|x\| = 1$ in $\mathbf{R}^n$, put $y = Q^\top x$. Note that there is a one-to-one correspondence between x and y. By the rotational invariance of the Euclidean norm, we have $\|y\| = \|Q^\top x\| = \|x\| = 1$. Hence

$$\begin{aligned} x^\top Ax &= x^\top Q\Lambda Q^\top x \\ &= y^\top \Lambda y \\ &= \sum \lambda_i y_i^2. \end{aligned}$$

By assumption, $\lambda_1 \geq \lambda_i \geq \lambda_n$ for i in $\{1, 2, \ldots, n\}$, so

$$\lambda_1 = \lambda_1 \sum y_i^2 \geq \sum \lambda_i y_i^2 \geq \lambda_n \sum y_i^2 = \lambda_n.$$

Since the inequalities are obtained with v_1 and v_n respectively, we are done. □

Theorem 8 (Poincairé Inequality). *Let A be a symmetric matrix in $\mathbf{R}^{n\times n}$ and let $A = Q\Lambda Q^\top$ be its spectral decomposition. Since A has real eigenvalues, let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Fix k in $\{1, 2, \ldots, n\}$. Let V denote any k-dimensional subspace of $\mathbf{R}^n$. Then there exists x and y in V such that*

$$\begin{aligned} \lambda_k &\geq x^\top A x \\ \lambda_{n-k+1} &\leq y^\top A y \end{aligned}$$

Proof. Put

$$Q_k = \begin{bmatrix} | & & | \\ v_k & \cdots & v_n \\ | & & | \end{bmatrix}.$$

Let W denote the column space of Q_k. Clearly, the $\dim W = n - k + 1$. So by a counting argument V and W have a non-trivial intersection. For any $\|x\| = 1$ in $V \cap W$, there exists a y in $\mathbf{R}^{n-k+1}$ such that $x = Q_k y$. Hence

$$\begin{aligned} x^\top A x &= y^\top Q_k^\top Q\Lambda Q^\top Q_k y \\ &= \sum_{i=k}^{n} \lambda_i y_i^2 \\ &\leq \lambda_k. \end{aligned}$$

To show $\lambda_{n-k+1} \leq y^\top A y$, apply the previous result to $-A$. □

Theorem 9 (The Minimax Principle). *Let A be a symmetric matrix in $\mathbf{R}^{n\times n}$ and let $A = Q\Lambda Q^\top$ be its spectral decomposition. Since A has real eigenvalues, let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Fix k in $\{1, 2, \ldots, n\}$. Let V denote any subspace of $\mathbf{R}^n$ (the dimension of which will be specified below). Then*

$$\begin{aligned} \lambda_k &= \max_{\dim V = k} \; \min_{\substack{x\in V \\ \|x\|=1}} \; x^\top A x \\ &= \min_{\dim V = n-k+1} \; \max_{\substack{x\in V \\ \|x\|=1}} \; x^\top A x. \end{aligned}$$

Proof. By the Poincairé inequality, in any k-dimensional subspace V of $\mathbf{R}^n$ there is an unit length vector x such that $\lambda_k \geq x^\top A x$. Hence it is true that for any k-dimensional subspace V

$$\lambda_k \geq \min_{\substack{x\in V \\ \|x\|=1}} x^\top A x.$$

Since this bound is obtained by letting $x = v_k$, I conclude

$$\lambda_k = \max_{\dim V = k} \; \min_{\substack{x\in V \\ \|x\|=1}} \; x^\top A x.$$

For the second result, apply the first result to $-A$. □

This is also known as the *Variational theorem* or the *Courant-Fisher-Weyl minimax theorem.*

Orthogonal projections

Definition 13. An $n \times n$ matrix P is an *orthogonal projection* matrix if $P = P^2$ and $P = P^\top$.

Theorem 10. *Let A be a real-valued matrix in $\mathbf{R}^{m\times n}$ of rank n. The matrix $P_A = A(A^\top A)^{-1}A^\top$ when multiplied on the right by a vector x in $\mathbf{R}^n$ will return*

Proof. The orthogonal projection p of v onto the column space of A, which has full column rank, will be a unique combination of the columns of A; that is, $p = Ax$ for some unknown x in $\mathbf{R}^n$. Since it is an orthogonal projection, the difference $e = v - p$ will be orthogonal to the column space of A. In particular, $e \perp A$; so

$$A^\top e = A^\top(v - p) = A^\top(v - Ax) = A^\top v - A^\top Ax = 0.$$

Hence

$$A^\top Ax = A^\top v.$$

Since A has full column rank, $(A^\top A)^{-1}$ exists. So

$$x = (A^\top A)^{-1}A^\top v$$

and

$$p = Ax = \underbrace{A(A^\top A)^{-1}A^\top}_{P_A} v.$$

See Figure 4 for a schematic illustration. □

Applications

We are now ready to look at three important applications: pseudoinverses, principal component analysis, and the least square solution.

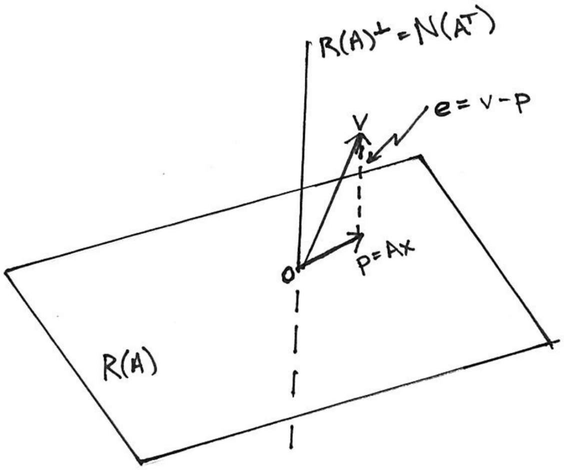

Figure 4: Schematic illustration of orthogonal projection.

Pseudoinverse

We begin by reviewing some facts about matrix inverses.

Definition 14. A matrix A has a *left inverse* A_{L}^{-1} if $A_{\mathrm{L}}^{-1}A = I$.

If A is full column rank, then it has a trivial null space and $A^\top A$ is full rank. Hence $(A^\top A)^{-1}$ exists and $(A^\top A)^{-1}A^\top A = I$. In this case, we say $A_{\mathrm{L}}^{-1} = (A^\top A)^{-1}A^\top$ is a left inverse of A (this doesn't mean that it is unique).

Definition 15. A matrix A has a *right inverse* A_{R}^{-1} if $AA_{\mathrm{R}}^{-1} = I$.

Similarly, if A is full row rank, then $(AA^\top)^{-1}$ exists and $AA^\top(AA^\top)^{-1} = I$. Then we say $A_{\mathrm{R}}^{-1} = A^\top(AA^\top)^{-1}$ is a (not necessarily unique) right inverse of A.

Definition 16. If a matrix A^{-1} is both a left and a right inverse of a matrix A, it is called the *inverse* of A.

Theorem 11. *If a square matrix A has a left inverse A_L^{-1} and a right inverse A_R^{-1}, then $A_L^{-1} = A_R^{-1}$.*

Proof. Since all the dimensions match,

$$\begin{aligned} A_{\mathrm{L}}^{-1} &= A_{\mathrm{L}}^{-1}AA_{\mathrm{R}}^{-1} \\ &= A_{\mathrm{R}}^{-1}. \end{aligned}$$

□

Definition 17. A real-valued matrix A has a *pseudoinverse* A^+ if

- $AA^+A = A$,
- $A^+AA^+ = A^+$,
- $(A^+A)^\top = A^+A$, and
- $(AA^+)^\top = AA^+$.

Theorem 12 (Existence and uniquness of A^+)**.** *For any real-valued $m \times n$ matrix A there is a unique pseudoinverse A^+.*

Proof. By computing A^+ via SVD below, I demonstrate the existence of A^+ for any real-valued matrix A. To show uniqueness, suppose both A_1^+ and A_2^+ are pseudoinverses of A. Then, by repeated application of the properties of the pseudoinverse, I have

$$\begin{aligned}
AA_1^+ &= (AA_1^+)^\top \\
&= (A_1^+)^\top A^\top \\
&= (A_1^+)^\top (AA_2^+A)^\top \\
&= (A_1^+)^\top A^\top (A_2^+)^\top A^\top \\
&= (AA_1^+)^\top (AA_2^+)^\top \\
&= AA_1^+AA_2^+ \\
&= AA_2^+.
\end{aligned}$$

Similarly, $A_1^+A = A_2^+A$. Hence

$$A_1^+ = A_1^+AA_1^+ = A_2^+AA_1^+ = A_2^+AA_2^+ = A_2^+.$$

□

Theorem 13 (Computing A^+ via SVD)**.** *For any real-valued matrix A, let $U\Sigma V^\top$ be its SVD. Then the pseudoinverse of A is*

$$A^+ = V\Sigma^+U^\top,$$

where Σ^+ is formed by transposing Σ and replacing each non-zero σ_i with its reciprocal $1/\sigma_i$.

Proof. Since $V^\top V = I$, $U^\top U = I$, and $\Sigma\Sigma^+\Sigma = \Sigma$,

$$\begin{aligned}
AA^+A &= (U\Sigma V^\top)(V\Sigma^+U^\top)(U\Sigma V^\top) \\
&= U\Sigma V^\top \\
&= A.
\end{aligned}$$

Similarly,

$$\begin{aligned} A^+AA^+ &= (V\Sigma^+U^\top)(U\Sigma V^\top)(V\Sigma^+U^\top) \\ &= A^+. \end{aligned}$$

Since $\Sigma\Sigma^+ = \Sigma^+\Sigma$,

$$\begin{aligned} (A^+A)^\top &= ((V\Sigma^+U^\top)(U\Sigma V^\top))^\top \\ &= V\Sigma\Sigma^+V^\top \\ &= V\Sigma^+\Sigma V^\top \\ &= V\Sigma^+U^\top U\Sigma V^\top \\ &= A^+A. \end{aligned}$$

Finally,

$$\begin{aligned} (AA^+)^\top &= ((U\Sigma V^\top)(V\Sigma^+U^\top))^\top \\ &= AA^+. \end{aligned}$$

I conclude that A^+ satisfies all the requirements of a pseudoinverse. □

Principal Component Analysis

Definition 18. The *matrix of ones* is a matrix where every entry is 1. We will denote an $n \times n$ matrix of ones as J_n.

Definition 19. Given n samples of an m-dimensional random vector,

$$X = \begin{bmatrix} | & & | \\ x_1 & \cdots & x_n \\ | & & | \end{bmatrix}$$

the *sample mean vector* is

$$\bar{x} = \frac{1}{n}\sum x_i$$

and the *sample covariance matrix* is

$$\hat{\Sigma} = \frac{1}{n-1}\sum (x_i - \bar{x})(x_i - \bar{x})^\top.$$

Definition 20. The *centering matrix* of size n is defined as

$$C_n = I_n - \frac{1}{n}J_n,$$

where J_n is the $n \times n$ matrix of all 1s.

You should verify that the centering matrix C_n is symmetric $C_n = C_n^\top$ and idempotent $C_n^2 = C_n$. Given an $m \times n$ matrix X, multiplying it with centering matrix on the left $C_m X$ removes the column mean from each column. Similarly, multiplying X with centering matrix on the right XC_n removes the row mean from each row. While this is not computationally efficient, it is convenient for algebraic manipulation (as we will see).

Definition 21. Given an $m \times n$ data matrix X, the *scatter matrix* is the $m \times m$ matrix

$$\begin{aligned} S &\equiv \sum (x_i - \bar{x})(x_i - \bar{x})^\top \\ &= (X - \bar{x}\mathbf{1}^\top)(X - \bar{x}\mathbf{1}^\top)^\top \\ &= (XC_n)(XC_n)^\top \\ &= XC_n C_n^\top X^\top \\ &= XC_n X^\top. \end{aligned}$$

Scaling the scatter matrix by $\frac{1}{n-1}$ yields the standard sample covariance matrix, while scaling it by $\frac{1}{n}$ yields the maximum likelihood estimate of the covariance matrix of the multivariate Gaussian distribution.

SVD, revisited

Since the SVD is so important, it is worth looking at it from a different perspective. Consider the following greedy strategy of search first for the unit direction which A increases most

$$\begin{aligned} v_1 &= \operatorname*{argmax}_{\|v\|=1} \|Av\| \\ \lambda_1 &= \|Av_1\|. \end{aligned}$$

Then the unit direction perpendicular to the first which A increases most

$$\begin{aligned} v_2 &= \operatorname*{argmax}_{\substack{\|v\|=1 \\ v \perp v_1}} \|Av\| \\ \lambda_2 &= \|Av_2\|. \end{aligned}$$

And then the unit direction perpendicular to the space spanned by the first two principle directions which A increases most

$$\begin{aligned} v_3 &= \operatorname*{argmax}_{\substack{\|v\|=1 \\ v\perp v_1,v_2}} \|Av\| \\ \lambda_3 &= \|Av_3\|. \end{aligned}$$

And continuing in this way until there are no more unit direction perpendicular to all the previous ones.

Theorem 14 (Best-fit linear subspace)**.** *Let A be a $m \times n$ matrix with right singular vectors $v_1, v_2, \ldots, v_r$. For $1 \leq k \leq r$, let $V_k = \operatorname{span}(v_1, v_2, \ldots, v_k)$ be the space spanned by the first k right singular values of A. Then V_k is the best-fit k-dimensional subspace for A.*

Definition 22. For any real matrix A with the SVD given as $U\Sigma V^\top$, the *truncated SVD* of A is $A_k = U\Sigma_k V^\top$ where Σ_k is obtained from Σ by setting all but the first $k < \operatorname{rank}(A)$ elements of the diagonal to 0.

Note that all but the first k columns of U and all but the first k rows of $V^\top$ will be multiplied by 0, so $U\Sigma_k V^\top$ is equivalent to $U_k\Sigma_k V_k^\top$ where U_k is just the first t columns of U and $V_k^\top$ is just the first k rows of $V^\top$.

Corollary (Eckart–Young theorem)**.** *The truncated SVD A_k is the best rank k approximation to A in the sense that if B is any compatible rank k, then*

$$\|A - A_k\|_F \leq \|A - B\|_F .$$

Least squares solution

Consider the least-squares problem

$$p^* = \min_x \|Ax - y\|_2$$

for A in $\mathbf{R}^{m\times n}$ and y in $\mathbf{R}^m$.

If y is in $R(A)$, then $p^* = 0$. However, if y is not in $R(A)$, then $p^* > 0$ and, at optimum, the residual vector $r = y - Ax$ is such that $r^\top y > 0, A^\top r = 0$. I will assume that $m \geq n$, and that A is full column rank.

Using the SVD of A and the rotational invariance of the L_2 norm,

$$\begin{aligned} \min_x \|Ax - y\|_2 &= \min_x \|Ax - y\|_2^2 \\ &= \min_x \|U^\top(AVV^\top x - y)\|_2^2 \\ &= \min_x \|\Sigma V^\top x - U^\top y)\|_2^2. \end{aligned}$$

Let

$$\tilde{x} \doteq V^\top x \quad \text{and} \quad \tilde{y} \doteq U^\top y = \begin{bmatrix} u_1^\top y \\ \vdots \\ u_m^\top y \\ u_{m+1}^\top y \\ \vdots \\ u_n^\top y \end{bmatrix} = \begin{bmatrix} \tilde{y}_{\mathcal{R}(A)} \\ \tilde{y}_{\mathcal{R}(A)^\perp} \end{bmatrix}$$

where I've partitioned $\tilde{y}$ into the first m elements $\tilde{y}_{\mathcal{R}(A)}$ and the remaining $n-m$ elements $\tilde{y}_{\mathcal{R}(A)^\perp}$.

Let $\tilde{\Sigma}$ denote the first m rows of Σ, which by assumption is the diagonal matrix of the m singular values. Substituting these symbols into the above equation and expanding the sum into two terms (one containing $\tilde{x}$ and the other not), I have

$$\begin{aligned} \min_x \|\Sigma V^\top x - U^\top y)\|_2^2 &= \min_{\tilde{x}} \left\| \begin{bmatrix} \tilde{\Sigma} \\ 0 \end{bmatrix} \tilde{x} - \begin{bmatrix} \tilde{y}_{\mathcal{R}(A)} \\ \tilde{y}_{\mathcal{R}(A)^\perp} \end{bmatrix} \right\|_2^2 \\ &= \min_{\tilde{x}} \left\| \begin{matrix} \tilde{\Sigma}\tilde{x} - \tilde{y}_{\mathcal{R}(A)} \\ \tilde{y}_{\mathcal{R}(A)^\perp} \end{matrix} \right\|_2^2 \\ &= \min_{\tilde{x}} \|\tilde{\Sigma}\tilde{x} - \tilde{y}_{\mathcal{R}(A)}\|_2^2 + \|\tilde{y}_{\mathcal{R}(A)^\perp}\|_2^2. \end{aligned}$$

Since $\tilde{\Sigma}$ is invertible (just replace the non-zero elements with their reciprocals), the first term can be made 0 by setting $\tilde{x} = \tilde{\Sigma}^{-1}\tilde{y}_{\mathcal{R}(A)}$. I conclude $p^* = \|\tilde{y}_{\mathcal{R}(A)^\perp}\|_2^2$ and the least square solution is given by $\hat{x} = V^\top \tilde{\Sigma}^{-1}\tilde{y}_{\mathcal{R}(A)}$.

By construction $\tilde{y}_{\mathcal{R}(A)^\perp}$ is colinear with a component of $\tilde{y}$ and thus with y. Hence $r^\top y > 0$. Also by construction, $\tilde{y}_{\mathcal{R}(A)^\perp}$ is perpendicular to A. Hence $A^\top r = 0$. The geometry is schematically illustrated in Figure 5.

References

[1] Rajendra Bhatia. *Matrix Analysis.* Graduate Text in Mathematics. Springer-Verlag, New York, 1997.

[2] Kenneth Hoffman and Ray Kunze. *Linear Algebra.* Prentice Hall, Englewood Cliffs, New Jersey, 1971.

[3] Gilbert Strang. *Introduction to Linear Algebra.* Wellesley-Cambridge Press, Massachusetts, 2013.

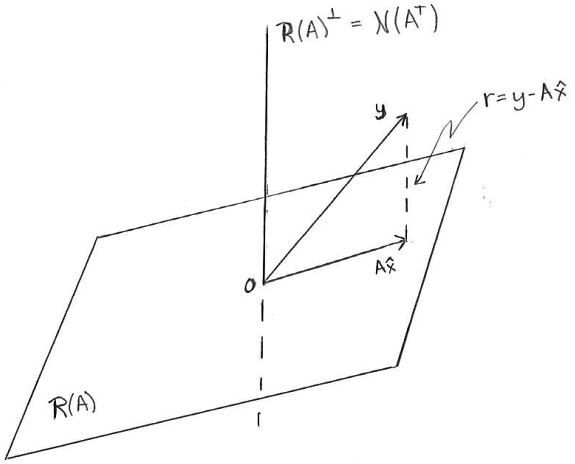

Figure 5: This diagram represents the geometry of least squares.