Linear Algebra Review

Dave Klemish

8/29/2016

Background

This document is intended to be a primer for linear algebra in preparation of a course in linear models at the PhD level. It includes a brief summary of most of the material covered in an undergraduate linear algebra course. Basics such as matrix multiplication, transposition, inverses, etc. are not discussed, so please ask if you have any questions on these topics. All results are presented without proof.

Linear algebra is commonly thought of as the mathematics of matrices, but actually is a broader study of vector spaces which have different applications in statistics. While the majority of this document does cover matrices, there is also some discussion on more general vector spaces. Please keep in mind that this is not necessary for linear models.

1 Basic Matrix Stuff

1.1 Definitions

- The *identity matrix* $\mathbf{I}$ is a square matrix with 1's on the diagonal and 0 on all non-diagonal elements.
- A square matrix $\mathbf{A}$ is *symmetric* if $\mathbf{A} = \mathbf{A}$'.
- A square matrix $\mathbf{A}$ is *positive definite* if for any vector $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x}'\mathbf{A}\mathbf{x} > 0$.
- A square matrix $\mathbf{A}$ is *invertible* or *nonsingular* if there exists a matrix $\mathbf{A}^{-1}$ such that $\mathbf{A}\mathbf{A}^{-1} = I$. If such a matrix does not exist, $\mathbf{A}$ is said to be *singular*.
  - Not all matrices have an inverse. More on this below.
  - All matrices (even non-square) have a *generalized pseudoinverse*, which is a matrix $\mathbf{A}^-$ such that $\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}$
- A square matrix $\mathbf{A}$ is *orthogonal* if $\mathbf{A}' = \mathbf{A}^{-1}$.
- A square matrix $\mathbf{A}$ is *idempotent* if $\mathbf{A}\mathbf{A} = \mathbf{A}$.

1.2 Determinants

- The *determinant* $|\mathbf{A}|$ of a square matrix $\mathbf{A}$ is a real number that specifies whether $\mathbf{A}$ is invertible.
- $\mathbf{A}$ is invertible if and only if (iff) $|\mathbf{A}| \neq 0$.
- If $\mathbf{A}$ is invertible, $|\mathbf{A}^{-1}| = 1/|\mathbf{A}|$.
- $|\mathbf{A}' = |\mathbf{A}|$.
- If $\mathbf{A}$, $\mathbf{B}$ are both square matrices, $|\mathbf{A}\mathbf{B} = |\mathbf{A}||\mathbf{B}|$
- How do you calculate the determinant?
  - If a matrix is triangular, the determinant equals the product of the diagonal elements.
  - The determinant also equals the product of the eigenvalues of the matrix (more on this later).

  - One can also use the cofactor expansion to calculate the determinant (not discussed here).
- Note that while a singular matrix must have a determinant of zero that numerical computation of the determninant may not be exactly zero due to roundoff errors.

1.3 Trace

- The *trace* of a square matrix $\mathbf{A}$ is the sum of the diagonal elements, usually denoted as $tr(\mathbf{A})$.
- The trace is also equal to the sum of the eigenvalues of $\mathbf{A}$.
- Note that trace is a cyclic operator, so $tr(ABC) = tr(CAB) = tr(BCA)$, i.e. we can cyclically permute the matrices inside a trace operator (rotate them so the order is always the same) and the trace will remain unchanged. This is not true if you change the order, $tr(ABC) \neq tr(BAC)$.

1.4 Useful Random Things

1. $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$.
2. $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.
3. If we need to sum the elements of a vector $\mathbf{y}$, we can calculate $\mathbf{y}'\mathbf{1}$, where $\mathbf{1}$ is a vector of all 1's of the same length as $\mathbf{y}$.
$$\sum_{i=1}^{n} y_i = \mathbf{y}'\mathbf{1}$$
4. Similarly, if we need to sum the squared elements of a vector, we can calculate the vector multiplication of $\mathbf{y}$ with itself.
$$\sum_{i=1}^{n} y_i^2 = \mathbf{y}'\mathbf{y}$$

2 Vector Spaces

2.1 Definition

Let V be a set on which two operations are defined. Elements of V are referred to as vectors.

- The first operation (commonly referred to as addition) acts on two vectors $\mathbf{x}$ & $\mathbf{y} \in V$.
- The second operation (commonly referred to as scalar multiplication) acts on a vector $\mathbf{x}$ and a scalar α.

If V is closed under these two operations (i.e. if $\mathbf{x}$, $\mathbf{y} \in V$ and α a scalar, then $\mathbf{x}+\mathbf{y} \in V$ & $\alpha\mathbf{x} \in V$) **and** the following axioms are satisfied, then V is a vector space.

1. $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$ (additive commutativity)
2. $(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z})$ (additive associativity)
3. $\exists\ \mathbf{0} \in V$ such that $\mathbf{x} + \mathbf{0} = \mathbf{x}$ (additive identity)
4. $\forall\ \mathbf{x} \in V\ \exists$ -$\mathbf{x} \in V$ such that $\mathbf{x} + (\text{-}\mathbf{x}) = \mathbf{0}$ (additive inverse)
5. $\alpha(\mathbf{x} + \mathbf{y}) = \alpha\mathbf{x} + \alpha\mathbf{y}$ (multiplication is distributive over vectors)
6. $(\alpha + \beta)\mathbf{x} = \alpha\mathbf{x} + \beta\mathbf{x}$ (multiplication is distributive over vectors)
7. $(\alpha\beta)\mathbf{x} = \alpha(\beta\mathbf{x})$ (multiplicative associativity)
8. $1^*\mathbf{x} = \mathbf{x}$ (multiplicative identity)

where the above holds $\forall\ \mathbf{x}, \mathbf{y}, \mathbf{z} \in V$ and all scalars α, β.

Examples of vector spaces include:

1. $\mathbb{R}^n$: the set of all n-element column vectors of real numbers
2. $\mathbb{R}^{m\times n}$: the set of all m by n matrices of real numbers
3. $C[a, b]$: the set of all continuous functions on the closed interval $[a, b]$.

Note for the last example that a vector in this vector space is a continuous function, where the addition of two vectors f, g is defined as $(f + g)(x) = f(x) + g(x)$ and scalar multiplication is defined as $(\alpha f)(x) = \alpha f(x)$. This example is included to demonstrate the general nature of vector spaces.

A fun math fact for those who have taken abstract algebra: vector spaces are Abelian/commutative groups (over addition) with an extra component of scalar multiplication - axioms 2-4 define a group, and axiom 1 makes the group Abelian.

2.2 Subspaces

If S is a subset of V and S is closed under the operations of V (i.e. addition and scalar multiplication of elements in S result in elements in S), then S is a subspace of V.

Examples include:

1. Let $V = \mathbb{R}^2$, and let $S = \{(x_1, x_2)' | x_1 = x_2\}$. S is a subspace of V.
2. Let $V = \mathbb{R}^3$, and let $S = \{(x_1, x_2, 0)'\}$. S is a subspace of V.
3. Let $V = \mathbb{R}^2$, and let $S = \{(x_1, 1)'\}$. S is **not** a subspace of V, since S is not closed under addition or scalar multiplication.
4. Let $V = \mathbb{R}^{2\times 2}$, and let $S = \{A | a_{11} = a_{22}\}$. S is a subspace of V.
5. Let $V = C[a, b]$, and let $S = C^n[a, b]$, the set of all functions that have a continuous n^{th} derivative on $[a, b]$. S is a subspace of V.

2.3 Span, Basis and Dimension

- Let $\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n}$ be n vectors in a vector space V. A linear combination of these vectors is the sum $c_1\mathbf{x_1} + ... + c_n\mathbf{x_n}$ where $c_1, ..., c_n$ are scalars. The set of *all* linear combinations of $\mathbf{x_1}, ..., \mathbf{x_n}$ is the span of $\mathbf{x_1}, ..., \mathbf{x_n}$.
- Note that the span of a collection of vectors in a vector space V is a subspace of V.
- If every vector in V can be written as a combination of the vectors in that collection, that collection is said to be a spanning set for V.
- A collection of vectors $\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n}$ is *linearly independent* if $c_1\mathbf{x_1} + ... + c_n\mathbf{x_n} = \mathbf{0}$ implies that $c_1, ..., c_n = 0$ (i.e. the only way to additively combine the vectors to get the zero vector / additive identity is for all scalar coefficients to be zero). If you can choose $c_1, ..., c_n$ not all zero such that the sum is the zero vector, then the collection is *linearly dependent*.
- If a collection of vectors $\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n}$ are linearly independent and form a spanning set for V, that collection is said to form a *basis* for the vector space V.
- The *dimension* of a vector space V is the number of vectors n in any basis of V.
  - Note that one can choose many different bases for a vector space, but any basis you choose will have the same number of elements exactly equal to the dimension of V.

For examples 1-2 below, assume that $V = \mathbb{R}^3$.

1. Let $\mathbf{x_1} = (1, 0, 0)'$ and $\mathbf{x_2} = (0, 1, 0)'$.
   - The span of these vectors is the set of all vectors of the form $(c_1, c_2, 0)$, which is the entire xy-plane if we use x,y,z notation for $\mathbb{R}^{\not\vdash}$.
   - These vectors are linearly independent, since the only way to add these two vectors together to get the zero vector $(0, 0, 0)$ is to choose $c_1 = c_2 = 0$.
   - They do not form a basis for $\mathbb{R}^3$, since you cannot get a vector of the form (c_1, c_2, c_3) from any linear combination of these two vectors alone.
2. Let $\mathbf{x_1} = (1, 1, 1)', \mathbf{x_2} = (1, 2, 3)'$.
   - The span of these vectors is $\beta_0(1, 1, 1)' + \beta_1(1, 2, 3)'$ (note use of β's... we'll come back to this example again!) This defines a plane in 3 dimensions (i.e. is a subspace of $\mathbb{R}^3$), that goes through the origin, where the vector $(1, -2, 1)$ forms a right angle with the plane (one way to determine this is to use the cross product of the two vectors).
   - These vectors are linearly independent. There's no way to add these two vectors to get the zero vector unless we choose $\beta_0 = \beta_1 = 0$.
   - They do not form a basis for $\mathbb{R}^3$. To be a basis, the span needs to be all of $\mathbb{R}^3$, but as discussed above the span is only a two-dimensional plane. Note that since $\mathbb{R}^3$ is a 3-dimensional that any basis requires 3 vectors.
3. Let V be $C[a, b]$, with $\mathbf{x_1} = sin(y), \mathbf{x_2} = cos(y)$.
   - The span of these vectors is the set of all possible functions on $[a, b]$ that can be represented as the sum of $c_1 \sin(y) + c_2 \cos(y)$.
   - It can be shown that these vectors (i.e. functions) are linearly independent (one way to show is to calculate the Wronskian... not important for this course!). In other words, there is no way to get a function that is uniformly zero on $[a, b]$ just by choosing $c_1, c_2 \neq 0$.
   - Note that $C[a, b]$ is an infinite dimensional space, so any finite # of basis functions can not span the space and therefore these two functions cannot form a basis for the space.

2.4 Column Space & Nullspace

- For a $n \times p$ matrix $\mathbf{X}$, imagine that we separately consider the columns of $\mathbf{X}$. This would be p vectors in $\mathbb{R}^n$. The *column space* of $\mathbf{X}$, denoted $C(\mathbf{X})$ is the subspace of $\mathbb{R}^n$ spanned by these vectors.
- The *rank* of $\mathbf{X}$ is the dimension of the column space of $\mathbf{A}$, denoted $r(\mathbf{X})$.
- Note that one can similarly define the row space of a matrix, but for our purposes the column space is more important.
- The rank of a matrix is also equal to the number of linearly independent columns in that matrix.

For example, let

$$\mathbf{X} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix}$$

The column space of $\mathbf{X}$ is the subspace of $\mathbb{R}^3$ spanned by the vectors $(1,1,1)'$ and $(1,2,3)'$. This is the same subspace we saw in the previous section, which had a dimension of 2. Therefore, the $r(\mathbf{X}) = 2$.

- Let $\mathbf{X}$ be a $n \times p$ matrix. The *nullspace* of $\mathbf{X}$ is defined as the set of all vectors β in $\mathbb{R}^p$ such that $\mathbf{X}\beta = \mathbf{0}$, and is denoted $\mathrm{N}(\mathbf{X})$.
- The nullspace of a matrix is a subspace of $\mathbb{R}^p$.
- For a $n \times p$ matrix $\mathbf{X}$, the matrix rank plus the dimension of the nullspace (the nullity) equals p (the rank-nullity theorem).

Examples:

1. Let $\mathbf{X} = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{smallmatrix}\right)$ as before. One can show (using Gauss-Jordan elimination, again not very important for our purposes) that the only vector β for which $\mathbf{X}\beta = \mathbf{0}$ is the zero vector itself, which is defined to have dimension 0. Therefore the nullspace of $\mathbf{X}$ is the zero vector and the $r(\mathbf{X}) + dim(N(\mathbf{X})) = 2 + 0 = 2$, which is the number of columns of $\mathbf{X}$.
2. Let $\mathbf{X} = \left(\begin{smallmatrix} 1 & 1 & 2 & 3 \\ 1 & 2 & 3 & 4 \\ 1 & 3 & 4 & 5 \end{smallmatrix}\right)$. Note that the third and fourth columns are generated by adding the first column to the second column once and twice, respectively.
   - The column space of $\mathbf{X}$ is the same as before, since the additional columns do not expand the subspace of $\mathbb{R}^3$ spanned by the first two column vectors. Therefore, the rank of $\mathbf{X}$ still equals 2.
   - The nullspace of $\mathbf{X}$ is the span of the vectors $(-1,-1,1,0)'$ and $(-2,-1,0,1)'$. As a result the dimension of the nullspace is 2.
   - The sum of the rank and the dimension of the null space $= 2 + 2 = 4$, which again is the number of columns of $\mathbf{X}$.

2.5 Why We Need This for Linear Models

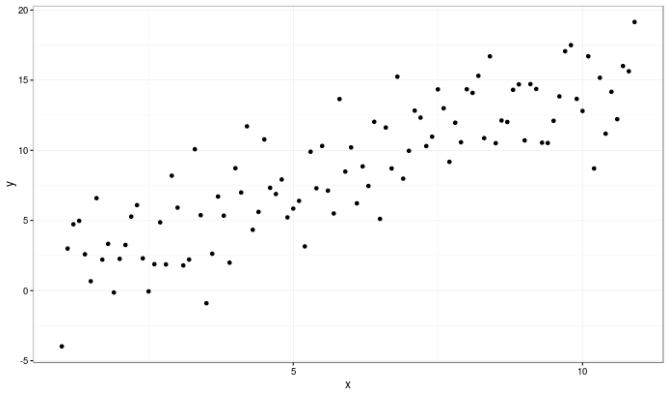
Let's start by initially considering simple linear regression, where each data observation i consists of one response value y_i and one explanatory covariate x_i. As an example, let's assume that we have 100 data points. The linear model setup is

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + \epsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_2 + \epsilon_2 \\ &\dots \\ y_{100} &= \beta_0 + \beta_1 x_{100} + \epsilon_{100} \end{aligned}$$

or equivalently,

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

where the standard assumption is that the ϵ_i are iid normally distributed with 0 mean and some variance. We generally plot the data as follows: The implicit view here is that the data consists of 100 data points in two-dimensional



space. However, one of the keys to linear models is that we can alternatively view our data consisting of *two data vectors in 100-dimensional space*! (Actually three vectors if we include an intercept β_0 in our model).

In this view, the true mean response vector (that in a sense generates the observed data by having a vector of random noise added to it) exists in the column space of $\mathbf{X}$. To fit a model and estimate the parameters β, we want to find the vector in $C(\mathbf{X})$ that is "closest" to the observed data vector $\mathbf{y}$. To do that, we need to define distance between vectors and how to find the closest vectors (see next section). Also, the concept of vector function spaces is useful for building up non-linear models

3 Inner Products, Orthogonality & Projections

3.1 Inner Products

An *inner product* on a vector space V is a function that maps two input vectors $\mathbf{x}$ & $\mathbf{y}$ in V to a real number $\langle \mathbf{x}, \mathbf{y} \rangle$ that satisfies the following conditions:

1. $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$, with equality holding only if $\mathbf{x}$ is the zero vector
2. $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$ for all vectors $\mathbf{x}$, $\mathbf{y}$ in V.
3. $\langle a\mathbf{x} + b\mathbf{y}, \mathbf{z} \rangle = a\langle \mathbf{x}, \mathbf{z} \rangle + b\langle \mathbf{y}, \mathbf{z} \rangle$ for all vectors $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$ in V and scalars a and b.

Common inner products on different vector spaces are:

- In $\mathbb{R}^n$: $\langle \mathbf{x}, \mathbf{x} \rangle = \mathbf{x}'\mathbf{y}$
- In $\mathbb{R}^{m \times n}$, the vector space of all real $n \times p$ matrices: $\langle A, B \rangle = \sum_{i=1}^{n} \sum j = 1^{p} a_{ij} b_{ij}$, i.e. multiply the two matrices element-wise and then sum over all resulting elements.
- For $C[a, b]$, the vector space of all continuous functions on $[a, b]$: $\langle f, g \rangle = \int_a^b f(x)g(x)dx$

Examples include:

1. In $\mathbb{R}^n$, the inner product of $(1, 1, 1)$ and $(1, 2, 3)$ is 6.
2. For two matrices $A = \left(\begin{smallmatrix} 1 & 2 \\ 3 & 4 \end{smallmatrix}\right)$, $B = \left(\begin{smallmatrix} 5 & 0 \\ 1 & 2 \end{smallmatrix}\right)$, the inner product is 1*5 + 2*0 + 3*1 + 4*2 = 16.
3. In $C[-\pi/2, \pi/2]$, the inner product of the sine and cosine functions is:

$$\int_{-\pi/2}^{\pi/2} sin(x)cos(x)dx = \int_{-1}^{1} udu = 0$$

One way to define a norm (length of a vector), denoted $\|\mathbf{x}\|$ is to take the square root of the inner product of a vector with itself, although one can define many other norms that are not based on inner products.

3.2 Orthogonality

- Two vectors are *orthogonal* if their inner product equals 0.
- Two subspaces $\mathbf{X}$, $\mathbf{Y}$ of a vector space are *orthogonal* if every vector in $\mathbf{X}$ is orthogonal to every vector in $\mathbf{Y}$.
- The *orthogonal complement* for a subspace $\mathbf{X}$ of a vector space V is the set of all vectors in V that are orthogonal to every vector in $\mathbf{X}$.
  - The orthogonal complement of $\mathbf{X}$ is denoted as $\mathbf{X}^{\perp}$.

Examples:

1. The vectors $\mathbf{x} = (1, 1, 0)'$ and $\mathbf{y} = (0, 0, 1)'$ are orthogonal.
2. If we consider the column spaces of these two vectors (i.e. the two one-dimensional subspaces of $\mathbb{R}^3$ spanned by each of these vectors), these two subspaces C($\mathbf{x}$) and C($\mathbf{y}$) are also orthogonal to each other.
3. In fact, any vector of the form $(a, -a, b)'$ is orthogonal to any vector in the column space of $\mathbf{x}$. Any such vector can be rewritten as $a(1, -1, 0)' + b(0, 0, 1)'$ for any scalars a, b. Therefore, the orthogonal complement of C($\mathbf{x}$) is the subspace spanned by $(1, -1, 0)$ and $(0, 0, 1)$, which is a two-dimensional plane in $\mathbb{R}^3$.

The orthogonal complement of a subspace $\mathbf{X}$ of a vector space V is effectively the largest subspace of V that has nothing "in common" with $\mathbf{X}$. The following makes that more a formal statement.

- Given two disjoint subspaces $\mathbf{X}$, $\mathbf{Y}$ of a vector space V, (i.e. $\mathbf{X} \cap \mathbf{Y} = \{\emptyset\}$), the subspace $\mathbf{Z}$ (the set of all possible linear combinations of vectors in $\mathbf{X}$ and $\mathbf{Y}$) is the *direct sum* of $\mathbf{X}$ and $\mathbf{Y}$ and is denoted $\mathbf{Z} = \mathbf{X} \oplus \mathbf{Y}$.
- It can be proven that if $\mathbf{X}$ is a subspace of a vector space V that $V = \mathbf{X} \oplus \mathbf{X}^{\perp}$, so the combination of a subspace and its orthogonal complement can generate any element of vector space.

3.3 Projections

In many instances, given a vector $\underline{y}$ we will want to find the the "closest" vector $\underline{x}$ contained in some subspace of a vector space V, where $\underline{x}$, $\underline{y} \in V$, that is "closest" (as measured by some norm). This can be done as follows:

Let S is a subspace of an inner product space V (i.e. a vector space assigned an inner product), $\mathbf{y}$ be a vector in V, and $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n$ be an orthonormal basis for S (orthonormal meaning that all basis vectors are orthogonal to each other and have length/norm 1). If

$$\mathbf{p} = \sum_{i=1}^{n} c_i \mathbf{e}_i$$

where $c_i = \langle \mathbf{y}, \mathbf{b}_i \rangle$ then

1. $\mathbf{p} \in S$
2. $\mathbf{p} - \mathbf{y} \in S^{\perp}$
3. $\mathbf{p}$ is the element of S that is closest to $\mathbf{y}$, in the sense that $\|\mathbf{x} - \mathbf{y}\| > \|\mathbf{p} - \mathbf{y}\|$ for any $\mathbf{x} \neq \mathbf{p}$ in S.

Examples:

1. In $\mathbb{R}^3$, to find the closest vector in the column space of the vector $\mathbf{x} = (1, 1, 1)'$ to the vector $\mathbf{y} = (1, 9, 5)$, note that $(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}})'$ is an orthonormal basis for $\mathbf{y}$. Therefore the projection of $\mathbf{y}$ onto $\mathbf{x}$ is

$$\begin{aligned} \mathbf{p} = \langle \mathbf{y}, \mathbf{x} \rangle \mathbf{x} &= \frac{1+9+5}{\sqrt{3}} (\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}})' \\ &= (5, 5, 5)' \end{aligned}$$

   Similarly, we can project a vector onto other subspaces of $\mathbb{R}^n$ such as planes or hyper-planes.

2. In the vector space of continuous functions on a closed interval, we can also project a function onto some subspace of all continuous functions. For example, we can use this to determine the best linear or quadratic polynomial approximation to some non-linear function on a closed interval. This is also leads into subjects such as Fourier series analysis.

If our focus is specifically on projecting vectors into $\mathbb{R}^n$ into the column space of a matrix $\mathbf{X}$ (which is often the case in linear models), it is easier to use a projection operator.

If $\mathbf{X}$ is a matrix in $\mathbb{R}^{n \times p}$, and we wish to project a vector $\mathbf{y}$ onto C($\mathbf{X}$), the column space of $\mathbf{X}$, let

$$P_X = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

be the projection operator onto the column space of $\mathbf{X}$. Then $P_X \mathbf{y}$ is the projection of $\mathbf{y}$ onto C($\mathbf{X}$).

Other notes:

- P_X is idempotent ($P_X^2 = P_X$) and symmetric.
- $I - P_X$ is a projection operator onto $\mathrm{C}(\mathbf{X})^{\cup}$).

4 Eigenvalues

4.1 Definitions

Let $\mathbf{A}$ be a square $n \times n$ matrix. If there exists a vector $\mathbf{v} \in \mathbb{R}^n$ such that $\mathbf{Av} = \lambda\mathbf{v}$ for some scalar λ, then λ and $\mathbf{v}$ are defined to be an *eigenvalue* and *eigenvector* of $\mathbf{A}$, respectively.

If we think of multiplying a vector $\mathbf{x}$ by a matrix $\mathbf{A}$ as a rotation of $\mathbf{x}$, one way to interpret the eigenvectors of $\mathbf{A}$ is the set of vectors that are invariant to the rotation induced by $\mathbf{A}$ i.e. these vectors are not rotated, but merely stretched/shrunk by some factor, specifically by a factor equal to the eigenvalue.

For small matrices, one can calculate the eigenvalues by noting that the definition is equivalent to $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = 0$. This has a nontrivial solution only if $(\mathbf{A} - \lambda\mathbf{I})$ is singular and equivalently if the determinant of $(\mathbf{A} - \lambda\mathbf{I})$ is zero. For larger matrices, other algorithms are more appropriate.

4.2 Other Notes

- For real matrices, eigenvalues and elements of the eigenvectors may be complex numbers. However, if this is the case then the complex conjugate must also be eigenvalue / eigenvector.
- The determinant of a matrix is equal to the product of its eigenvalues.
- Therefore a matrix is invertible only if it has all non-zero eigenvalues.
- The trace of a matrix is equal to the sum of the eigenvalues.
- A positive definite matrix is symmetric with all positive eigenvalues.

5 Matrix Decompositions

Matrix decompositions are a way of factorizing a given matrix into the product of two or more matrices. This section is intended solely as an introduction to some decompositions, and will not discuss how these are calculated.

5.1 Spectral Decomposition

Also known as eigendecomposition. Given a square $n \times n$ matrix $\mathbf{A}$ with n linearly independent eigenvectors, then

$$\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1}$$

where $\mathbf{Q}$ is a square matrix where the i^{th} column is an eigenvector of $\mathbf{A}$, and Λ is a diagonal matrix where the i^{th} entry consists of the corresponding eigenvalue.

Note that if $\mathbf{A}$ has n unique eigenvalues, then it has n linearly independent eigenvectors and so the above decomposition can be done. If some eigenvalues are not unique, then all eigenvectors may or may not be linearly independent and it may not be possible to spectrally decompose the matrix.

If $\mathbf{A}$ is symmetric, then the eigenvectors are orthogonal to each other and the decomposition can then be written as $\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}'$.

5.2 Cholesky Decomposition

Given a real positive definite square $n \times n$ matrix $\mathbf{A}$, then

$$\mathbf{A} = \mathbf{L}\mathbf{L}'$$

where $\mathbf{L}$ is a unique lower triangular matrix with positive diagonal elements. It is often used to decompose covariance matrices.

5.3 QR Decomposition

Given a real square $n \times n$ matrix $\mathbf{A}$, then

$$\mathbf{A} = \mathbf{QR}$$

where $\mathbf{Q}$ is an orthogonal $n \times n$ matrix, and $\mathbf{R}$ is an upper triangular $n \times n$ matrix.

- Often used in practice to numerically estimate linear regression problems rather than calculating a projection matrix due to numerical stability issues.
- Is actually more general in that one can also decompose $n \times p$ matrices. In that case, $\mathbf{R}$ is an $n \times p$ upper triangular matrix, where the bottom $n - p$ rows are all zero.

5.4 Singular Value Decomposition

Given a real square $n \times p$ matrix $\mathbf{A}$, then

$$\mathbf{A} = \mathbf{U\Sigma V'}$$

where $\mathbf{U}$, $\mathbf{V}$ are a $n \times n$ and $p \times p$ orthogonal matrices, respectively, and Σ is a $n \times p$ triangular diagonal matrix (so bottom rows are 0 if $n > p$, otherwise rightmost columns are 0).

- The columns of $\mathbf{U}$ are the orthonormal eigenvectors of $\mathbf{AA'}$.
- The columns of $\mathbf{V}$ are the orthonormal eigenvectors of $\mathbf{A'A}$.
- The non-zero diagonal elements of $\mathbf{R}$ (the singular elements) are the square roots of the eigenvalues of $\mathbf{AA'}$ and $\mathbf{A'A}$

SVD is used in many different applications in statistics and machine learning.

6 Matrix Calculus

There are times when we need to take the derivative of a function with respect to a vector or a matrix. Most commonly we'll need to take derivatives of scalar functions (i.e. functions that map a vector or matrix to a real number), so that will be the focus of this section. Derivatives of functions that map matrices to matrices requires tensors.

Much more detail can be found in the Wikipedia page on matrix calculus or in the Matrix Cookbook. You normally won't derive these, but use look them up and use them as needed.

6.1 Vector Input, Scalar Output

Let $y = f(\mathbf{x})$, where $\mathbf{x}$ is a n-dimensional vector. Then define the derivative of y with respect to $\mathbf{x}$ as

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \\ \vdots \\ \frac{\partial y}{\partial x_n} \end{pmatrix}$$

Examples:

1. Let $y = f(\mathbf{x}) = \mathbf{a'x} = a_1x_1 + a_2x_2 + \ldots + a_nx_n$. Then

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \\ \vdots \\ \frac{\partial y}{\partial x_n} \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \mathbf{a}$$

2. Let

$$\begin{aligned} y = f(\mathbf{x}) &= \mathbf{x}'\mathbf{A}\mathbf{x} \\ &= (a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n)x_1 + (a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n)x_2 + \ldots \end{aligned}$$

Then

$$\begin{aligned} \frac{\partial y}{\partial \mathbf{x}} &= \begin{pmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \\ \vdots \\ \frac{\partial y}{\partial x_n} \end{pmatrix} \\ &= \begin{pmatrix} 2a_{11}x_1 + (a_{12} + a_{21})x_2 + \ldots + (a_{1n} + a_{n1})x_n \\ (a_{12} + a_{21})x_1 + 2a_{22}x_2 + \ldots + (a_{2n} + a_{n2})x_n \\ \vdots \\ (a_{1n} + a_{n1})x_1 + (a_{2n} + a_{n2})x_2 + \ldots + 2a_{nn}x_n \end{pmatrix} \\ &= (\mathbf{A} + \mathbf{A}')\mathbf{x} \end{aligned}$$

6.2 Matrix Input, Scalar Output

Similarly, let $y = f(\mathbf{X})$, where $\mathbf{X}$ is a $n \times p$-dimensional matrix. Then define the derivative of y with respect to $\mathbf{X}$ as

$$\frac{\partial y}{\partial \mathbf{X}} = \begin{pmatrix} \frac{\partial y}{\partial x_{11}} & \frac{\partial y}{\partial x_{12}} & \cdots & \frac{\partial y}{\partial x_{1n}} \\ \frac{\partial y}{\partial x_{21}} & \frac{\partial y}{\partial x_{22}} & \cdots & \frac{\partial y}{\partial x_{2n}} \\ \vdots & & & \\ \frac{\partial y}{\partial x_{n1}} & \frac{\partial y}{\partial x_{n2}} & \cdots & \frac{\partial y}{\partial x_{nn}} \end{pmatrix}$$

Examples:

1. Let

$$\begin{aligned} y = f(\mathbf{x}) &= \mathbf{x}'\mathbf{A}\mathbf{x} \\ &= (a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n)x_1 + (a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n)x_2 + \ldots \end{aligned}$$

Then

$$\begin{aligned} \frac{\partial y}{\partial \mathbf{A}} &= \begin{pmatrix} x_1^2 & x_1x_2 & \ldots & x_1x_n \\ x_1x_2 & x_2^2 & \ldots & x_2x_n \\ \vdots & & & \\ x_1x_n & x_2x_n & \ldots & x_n^2 \end{pmatrix} \\ &= \mathbf{x}\mathbf{x}' \end{aligned}$$

2. One can show that

$$\begin{aligned} \frac{\partial tr(\mathbf{A}\mathbf{X})}{\partial \mathbf{X}} &= \mathbf{A}' \\ \frac{\partial |\mathbf{A}\mathbf{X}|}{\partial \mathbf{X}} &= |\mathbf{A}\mathbf{X}|(\mathbf{X}^{-1})' \end{aligned}$$

6.3 Vector Input, Vector Output

If a function maps a vector in $\mathbb{R}^n$ to a vector in $\mathbb{R}^p$, then we can define

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_p} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_p} \\ \vdots & & & \\ \frac{\partial y_n}{\partial x_1} & \frac{\partial y_n}{\partial x_2} & \cdots & \frac{\partial y_n}{\partial x_p} \end{pmatrix}$$

This is the Jacobian matrix.

Note that if we take the derivative of a scalar function with respect to a vector input, the result is a vector. If we take the derivative of this with respect to the same vector, we get the Hessian matrix, which can be written as:

$$\frac{\partial y}{\partial \mathbf{x} \partial \mathbf{x}'} = \begin{pmatrix} \frac{\partial y}{\partial x_1^2} & \frac{\partial y}{\partial x_1 \partial x_2} & \cdots & \frac{\partial y}{\partial x_1 \partial x_p} \\ \frac{\partial y}{\partial x_1 \partial x_2} & \frac{\partial y}{\partial x_2^2} & \cdots & \frac{\partial y}{\partial x_2 \partial x_p} \\ \vdots & & & \\ \frac{\partial y}{\partial x_1 \partial x_p} & \frac{\partial y}{\partial x_2 \partial x_p} & \cdots & \frac{\partial y}{\partial x_p^2} \end{pmatrix}$$

This is used all the time in optimization problems, as well as in determining the Fisher information matrix in statistics. It can also be written as $\frac{\partial y}{\partial \mathbf{x}^2}$.