Lightning Review of Linear Algebra

36-462/662, Fall 2022

22 March 2022 (Lecture 17)

Contents

> What follows is both brief and incomplete. It is *not* a complete guide to linear algebra — that's a *pre-requisite* for this class. (If you got through linear regression without using these ideas, I'm afraid you didn't really understand what you were doing.) The point of the lecture was merely to jog your memory, and to help you see where things fit together.

You need to know what vectors and matrices are, and how to do arithmetic with them.

1 Linear Combinations, Linear Dependence and Independence, Dimension, Subspace

Definition 1 *A* **linear combination** *of vectors $\vec{v}_1, \ldots \vec{v}_r$ is a weighted sum of the vectors, i.e., any vector which can be written*

$$\sum_{i=1}^{r} c_i \vec{v}_i \tag{1}$$

If all the $c_i = 0$, the linear combination is **trivial**, *otherwise it is* **non-trivial**.

The set of all vectors which are linear combinations of $\vec{v}_1, \ldots \vec{v}_r$ is the **span** *of that set of vectors.*

Definition 2 *The non-zero vectors $\vec{v}_1, \vec{v}_2, \ldots \vec{v}_r$ are* **linearly dependent** *when a non-trivial linear combination of them is zero:*

$$\sum_{i=1}^{r} c_i \vec{v}_i = 0 \tag{2}$$

with some (or all) of the $c_i \neq 0$. If a collection of vectors aren't linearly dependent, they are **linearly independent**, *i.e.,*

$$\sum_{i=1}^{r} c_i \vec{v}_i = 0 \tag{3}$$

if and only if all the $c_i = 0$.

"Linearly" is often dropped when it's implied by context.

Definition 3 *A vector space has* **dimension** *p when it has at least one set of p linearly independent vectors, but every set of $p+1$ vectors is linearly dependent.*

(For this definition to make sense, it can't be the case that there are linearly independent sets of p vectors, but that only *some* sets of $p+1$ vectors are linearly dependent.)

Proposition 1 *Ordinary vectors which are ordered lists of p numbers, i.e., $\vec{v} \in \mathbb{R}^p$, do indeed have dimension p.*

Definition 4 *A* **subspace** *of a vector space is a subset S which is closed under vector addition and scalar multiplication: if $\vec{u}, \vec{v} \in S$, then so is $a\vec{u} + b\vec{v}$.*

People sometimes say "linear subspace".

- For any vector $\vec{u}$, its multiples $a\vec{u}$ form a one-dimensional subspce.
- If $r < p$, the dimension of the space, then the linear combinations of $\vec{v}_1, \ldots \vec{v}_r$ form a subspace of dimension at most r.

2 Inner Product, Norm, Orthogonal Vectors

Definition 5 (Inner or dot product) *The* **inner product** *of two vectors $\vec{v}, \vec{u} \in \mathbb{R}^p$ is a scalar,*

$$\vec{v} \cdot \vec{u} = \sum_{i=1}^{p} v_i u_i \tag{4}$$

This is also called the **dot product** *because it's written with the $\cdot$ sign.*

Proposition 2 *The inner product has the following basic properties:*

1. *Linearity: $(a\vec{v} + b\vec{w}) \cdot \vec{u} = a(\vec{v} \cdot \vec{u}) + b(\vec{w} \cdot \vec{u})$*
2. *Symmetry: $\vec{v} \cdot \vec{u} = \vec{u} \cdot \vec{v}$*
3. *Positive-definiteness: $\vec{v} \cdot \vec{v} \geq 0$, and $\vec{v} \cdot \vec{v} = 0$ iff $\vec{v} = \vec{0}$.*

— In more advanced linear algebra, we might consider other sorts of vector spaces, and we are willing to regard something as an "inner product" if it obeys those three properties. In such contexts, we sometimes write the inner product as $\langle v, u \rangle$, or (especially in physics) $\langle v|u \rangle$.

Example 1 *If we insist on writing p-dimensional vectors as $p \times 1$ column matrices, the usual inner product is*

$$\vec{v} \cdot \vec{u} = v^T u \tag{5}$$

Definition 6 *The* **norm** *of a vector is the square root of its inner product with itself:*

$$\|\vec{v}\| = \sqrt{\vec{v} \cdot \vec{v}} \tag{6}$$

Example 2 *For p-dimensional vectors, the ordinary or Euclidean norm is*

$$\|\vec{v}\| = \sqrt{\sum_{i=1}^{p} v_i^2} \tag{7}$$

This is also called the **length** *of the vector (by the Pythagorean theorem).*

— Note:

1. In R, `length(v)` will give the number of entries in the vector `v`, i.e., its dimension.
2. We can actually define multiple norms just on finite-dimensional vector spaces, which all obey some very natural properties, which make it sensible to call them all "norms". When we looked at penalties, for instance, we say the "L_2" norm (which is what we defined here), but also the "L_1" norm, $\|\vec{v}\|_1 = \sum_{i=1}^{p} |v_i|$. (You can work out from those what the L_p norm might be.) We won't get into the axioms which define the term "norm" here, because you can find them easily and we won't go any further in this direction.

Definition 7 *Any vector of length 1 is a* **unit vector**.

Proposition 3 *For any vector $\vec{v}$, the vector $\frac{1}{\|\vec{v}\|}\vec{v}$ is a unit vector. We often write this as $\frac{\vec{v}}{\|\vec{v}\|}$. A common convention is to write this as $\hat{v}$, but, since that conflicts with using hats to indicate estimates, we'll use $\tilde{v}$.*

A common convention is to write the unit vectors along the coordinate axes as $\tilde{e}_1, \ldots \tilde{e}_p$. (Physicists would write $\hat{i}, \hat{j}, \hat{k}$, or $\tilde{i}, \tilde{j}, \tilde{k}$, but we'll often need more than three dimensions.)

2.1 Angles and Inner Products

Proposition 4 *For any two finite-dimensional vectors $\vec{u}$ and $\vec{v}$,*

$$\vec{v} \cdot \vec{u} = \|\vec{v}\|\|\vec{u}\| \cos\theta \tag{8}$$

where θ is the angle between the vectors.

(For infinite-dimensional vector spaces, we sometimes use this formula to *define* the angle between vectors.)

To understand what's going on here, consider a very special case first, where instead of $\vec{v} \cdot \vec{u}$ in full generality, we consider $\vec{v} \cdot \tilde{e}_1$, the inner product of $\vec{v}$ with the unit vector along the first axis.

By definition,

$$\vec{v} \cdot \tilde{e}_1 = v_1 \tag{9}$$

the first coordinate of $\vec{v}$. In the diagram, this is shown by dropping a perpendicular line (in blue) from the end of $\vec{v}$ on to the horizontal axis, which is the line defined by $\tilde{e}_1$. The length of the segment in purple is therefore $v_1 = \vec{v} \cdot \tilde{e}_1$. (It is offset slightly below the axis for clarity.)

The black, blue and purple line segments define a right triangle, so we can apply trigonometry. $\cos\theta$ is the length of the side of the triangle *adjacent* to the angle, divided by the length of the hypotenuse. But here that's

$$\cos\theta = \frac{\vec{v} \cdot \tilde{e}_1}{\|\vec{v}\|} \tag{10}$$

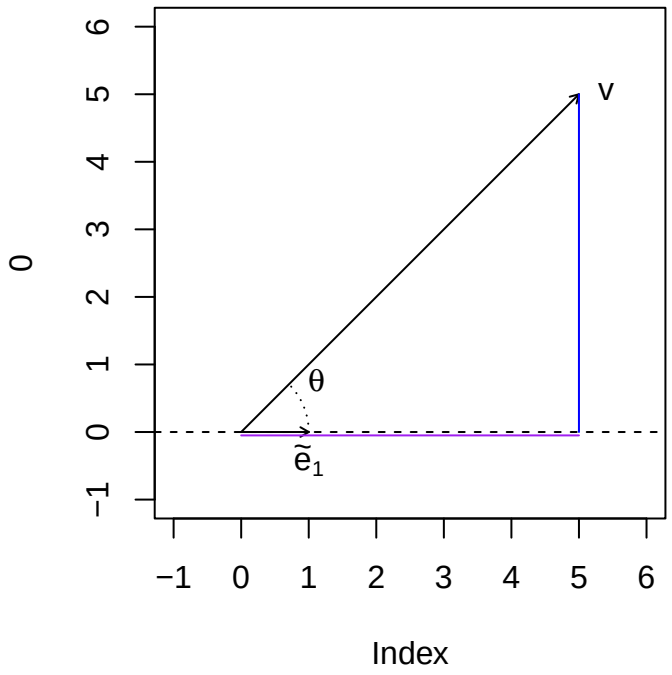

Figure 1: Hand-waving about the relationship between inner products and cosines; see text.

Turned around,

$$\vec{v} \cdot \tilde{e}_1 = \|\vec{v}\| \cos\theta \tag{11}$$

Since inner product is linear in each argument,

$$\vec{v} \cdot (a\tilde{e}_1) = a\|\vec{v}\| \cos\theta = \|a\tilde{e}_1\|\|\vec{v}\| \cos\theta \tag{12}$$

Clearly, nothing depended here on taking the inner product with $\tilde{e}_1$ as opposed to $\tilde{e}_2$ or $\tilde{e}_{137}$, so this argument works for every coordinate unit vector, and every multiple of those unit vectors. That it works for vectors pointing in an *arbitrary* direction is a bit less obvious, but I didn't promise you any proofs.

2.2 Orthogonal Vectors

Definition 8 *Vectors are* **orthogonal** *(to each other) when their inner product is zero:*

$$\vec{v} \cdot \vec{u} = 0 \tag{13}$$

We sometimes write this $\vec{v} \perp \vec{u}$.

This is because $\cos\theta = 0$ if, and only if, θ is a right angle ("rtho-" being "right" in Greek).

2.2.1 Orthogonalizing a collection of vectors

Given any collection of vectors, $\vec{v}_1, \vec{v}_2, \ldots \vec{v}_r$, here is a procedure for constructing a new set of vectors which are all orthogonal to each other.

1. $\vec{u}_1 = \vec{v}_1$. (Leave the first vector alone.)
2. Set

$$\vec{u}_2 = \vec{v}_2 - \frac{\vec{v}_2 \cdot \vec{u}_1}{\|\vec{u}_1\|^2}\vec{u}_1 \tag{14}$$

Exercise: Show that $\vec{u}_2 \cdot \vec{u}_1 = 0$.

3. Set
$$\vec{u}_3 = \vec{v}_3 - \frac{\vec{v}_3 \cdot \vec{u}_1}{\|\vec{u}_1\|^2}\vec{u}_1 - \frac{\vec{v}_3 \cdot \vec{u}_2}{\|\vec{u}_2\|^2}\vec{u}_2 \tag{15}$$

Exercise: Show that $\vec{u}_3 \cdot \vec{u}_1 = \vec{u}_3 \cdot \vec{u}_2 = 0$.

etc.

For the t^{th} vector,
$$\vec{u}_t = \vec{v}_t - \sum_{s=1}^{t-1} \frac{\vec{v}_t \cdot \vec{u}_s}{\|\vec{u}_s\|^2}\vec{u}_s \tag{16}$$

$\vec{u}_t$ will be orthogonal to all of the earlier $\vec{u}_s$.

Notes:

0. It's common to normalize the $\vec{u}$s at the end to have length 1, and doing so as you go can simplify calculations, but it's not necessary.
1. The span of the $\vec{v}$s is always the same as the span of the $\vec{u}$s. (Can you show this?)
2. The results of this procedure will be different, depending on the initial ordering of the vectors. But, again, the span is always the same.
3. If the initial vectors $\vec{v}_i$ are linearly dependent, some of the later vectors $\vec{u}_i$ will be zero. (Why?)

3 Bases, Orthonormal Bases

Definition 9 (Basis, orthogonal basis, normal basis, orthonormal basis) *The vectors $\vec{b}_1, \ldots \vec{b}_p$ are a* **basis** *for their vector space, when, given any $\vec{v}$*
$$\vec{v} = \sum_{j=1}^{p} v_j \vec{b}_j \tag{17}$$
for some coefficients $v_1, \ldots v_p$. The basis is **orthogonal** *when $\vec{b}_i \cdot \vec{b}_j = 0$ (unless $i = j$), and* **normal** *when each $\vec{b}_i$ has norm 1. A basis which is both orthogonal and normal is* **orthonormal**.

In a finite-dimensional space, the unit vectors along the different coordinate axes form an orthonormal basis, so no orthonormal basis has to have more vectors than the space has dimensions. (A non-orthogonal basis might need more.)

The orthogonalization procedure of the previous section can be used to construct an orthogonal basis from any non-orthogonal basis, and any basis can be normalized, so, given any non-orthogonal basis, we can construct an orthonormal basis.

Proposition 5 *If a basis is orthonormal, then the coefficients v_j in the expansion are just given by inner products:*
$$\vec{v} = \sum_{j=1}^{p} (\vec{v} \cdot \vec{b}_j)\vec{b}_j \tag{18}$$

Proposition 6 *In any p-dimensional vector space, once we choose an orthonormal basis, we can represent any vector by a list of p numbers, and the inner product between vectors is the sum of the pairwise products of their coordinates,*
$$\vec{v} \cdot \vec{u} = \sum_{i=1}^{p} v_i u_i \tag{19}$$

The choice of basis doesn't matter:

Proposition 7 *The inner product between two vectors, from Eq. 19, is the same in all bases.*

4 Matrices and Operators

Definition 10 *A* **(linear) operator** $\mathcal{A}$ *is a vector-valued linear function of a vector, i.e., a function where, for any two vectors* $\vec{v}, \vec{u} \in \mathcal{V}$*, and any two scalars* c, d*,*

$$\mathcal{A}(c\vec{v} + d\vec{u}) = c\mathcal{A}(\vec{v}) + d\mathcal{A}(\vec{u}) \tag{20}$$

Proposition 8 *In the vector space of* $p \times 1$ *column matrices, left-multiplying by a* $p \times p$ *matrix* $\mathbf{a}$ *defines a linear operator,* $\vec{v} \mapsto \mathbf{a}\vec{v}$*.*

Proposition 9 *In any p-dimensional vector space, once we choose an orthonormal basis, any linear operator can be represented by a* $p \times p$ *matrix.*

Proof: Say the basis vectors are $\vec{b}_1, \ldots \vec{b}_p$. Define the matrix through

$$a_{ij} = \vec{b}_i \cdot (\mathcal{A}(\vec{b}_j) \tag{21}$$

Now pick any arbitrary vector $\vec{v}$. In this basis, it is represented by a $p \times 1$ matrix with entries $v_j = \vec{b}_j \cdot \vec{v}$. By linearity, $\mathcal{A}(\vec{v}) = \sum_{j=1}^{p} v_j \mathcal{A}(\vec{b_j})$, which will be represented by a column matrix whose entries are $\sum_j \vec{b}_i \cdot (\mathcal{A}(\vec{b}_j) v_j$. But this is the same as $\mathbf{a}\vec{v}$. □

4.1 Rank

Definition 11 (Rank, full rank, rank deficient, range space) *The* **column rank** *of a* $p \times q$ *matrix* $\mathbf{a}$ *is the number of linearly independent columns of* $\mathbf{a}$*. If all the columns are linearly independent, then* $\mathbf{a}$ *has* **full** *(column) rank, otherwise it is (column) rank* **deficient***. Column rank equals the dimension of the (column)* **range (space)** *of* $\mathbf{a}$*, the linear subspace of vectors p-vectors of the form* $\mathbf{a}\vec{v}$*. The* **row rank***, similarly, is the number of linearly independent rows of* $\mathbf{a}$*, and the definitions of full row rank and row rank deficiency are parallel.*

Proposition 10 *The row rank and column rank of a matrix are always equal.*

Definition 12 *The (column)* **null space** *of* $\mathbf{a}$ *is the set of all vectors such that* $\mathbf{a}\vec{v} = 0$*.*

Exercise 3 proves that the null space is, in fact, a linear subspace.

Proposition 11 *For a q-column matrix, the sum of the column rank and the dimension of the null space is always q; for a p-row matrix, the sum of the row rank and the dimension of the row null space is p.*

5 Eigenvalues and Eigenvectors of Matrices

Definition 13 (Eigenvalue, eigenvector, spectrum, leading eigen) *A non-zero vector* $\vec{v}$ *is an* **eigenvector** *of an operator* $\mathbf{a}$*, with* **eigenvalue** λ*, when*

$$\mathbf{a}\vec{v} = \lambda\vec{v} \tag{22}$$

An operator's **spectrum** *is its set of eigenvalues. If we put the eigenvalues in order, the largest one is the* **leading** *eigenvalue, and the corresponding eigenvector is also called the leading eigenvector.*

Since matrices represent operators on finite-dimensional spaces, we uses these same definitions for the eigenvectors, eigenvalues, and spectra of matrices.

Note: in general, eigenvalues may be complex numbers, not just real numbers. In this class, the places where we need eigenvalues are ones where the eigenvalues will be real, but complex eigenvalues are important in many other applications.

Proposition 12 *If* $\vec{v}$ *is an eigenvector with eigenvalue* λ*, then so is* $b\vec{v}$*, for any b.*

Because of this, it's usual to normalize the eigenvector to have norm 1. (Sometimes we normalize to have its entries sum to 1.)

Proposition 13 (Number of eigenvalues) *An $p \times p$ matrix $\mathbf{a}$ has at most p distinct eigenvalues. When it has fewer than p distinct eigenvalues, it is still conventional to write $\lambda_1, \ldots \lambda_p$, with some eigenvalues repeated.*

The reason for this is that the eigenvalues are roots of a polynomial equation, and for a $p \times p$ matrix the polynomial has degree p.

If a matrix has fewer than p distinct eigenvalues, the "repeated" ones are **degenerate**. If λ is k-**fold** degenerate, or **has multiplicity** k, then there is $k-$-dimensional linear subspace of eigenvectors with eigenvalue λ (Exercise 9). A matrix with no repeated eigenvalues is **non-degenerate**.

Example 3 *The $p \times p$ identity matrix has the unique eigenvalue 1, which is p-fold degenerate; all vectors are eigenvectors of the identity matrix.*

5.1 Trace and determinant in terms of eigenvalues

Definition 14 *The* **trace** *of a square matrix, $\mathrm{tr}\,\mathbf{a}$, is the sum of its diagonal entries:*

$$\mathrm{tr}\,\mathbf{a} \equiv \sum_{i=1}^{p} a_{ii} \tag{23}$$

Proposition 14 *1. Trace is a linear operator on matrices, $\mathrm{tr}\,(c\mathbf{a} + d\mathbf{b}) = c\,\mathrm{tr}\,\mathbf{a} + d\,\mathrm{tr}\,\mathbf{b}$.*

2. The trace of matrix equals the sum of its eigenvalues, "counted with mutliplicity":

$$\mathrm{tr}\,\mathbf{a} = \sum_{i=1}^{p} \lambda_i \tag{24}$$

3. The trace of a matrix product doesn't care about the order of multiplication:

$$\mathrm{tr}\,(\mathbf{ab}) = \mathrm{tr}\,(\mathbf{ba}) \tag{25}$$

Definition 15 *The* **determinant** *of $\mathbf{a}$, $\det \mathbf{a}$ or $|\mathbf{a}|$, is the product of its eigenvalues (again, "counted with multiplicity"):*

$$\det(\mathbf{a}) \equiv |\mathbf{a}| \equiv \prod_{i=1}^{p} \lambda_i \tag{26}$$

(There are, of course, other, equivalent, ways to define the determinant.)

5.2 Matrix inversion and eigenvalues

Proposition 15 *A square matrix is invertible if and only if $|\mathbf{a}| \neq 0$, i.e., if and only if all of its eigenvalues are non-zero.*

Proposition 16 *The rank of an $p \times p$ matrix is p minus the eigenvalue multiplicity of 0.*

Proof: Exercise 10.

Proposition 17 *A square matrix is invertible if and only if it is of full rank.*

Proof: Follows from the previous two propositions. □

5.2.1 Multicollinearity and eigenvalues

You will recall that when we do ordinary least squares for linear models, the coefficient vector estimate is given by

$$\hat{\beta} = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{y} \tag{27}$$

where $\mathbf{x}$ is the $n \times p$ matrix of regressor values, and $\mathbf{y}$ is the $n \times 1$ vector of response values. This expression of course makes no sense when $\mathbf{x}^T\mathbf{x}$ can't be inverted. We say that this happens when the regressors are **collinear**, or **multicollinear**. What does this mean? Two equivalent things:

- One column (or more) of $\mathbf{x}$ can be written as a linear combination of the other columns.
- $\mathbf{x}^T\mathbf{x}$ has at least one zero eigenvalue.

If $\mathbf{x}^T\mathbf{x}$ has a zero eigenvalue, there must be an eigenvector which goes with it. Because $\mathbf{x}^T\mathbf{x}$ is a $p \times p$ matrix, that eigenvector will be a vector of length p. That vector gives the coefficients in a linear combination of the regressor variables whose variance (in the sample) is exactly 0. (You will prove this in HW 9.)

One way to check for multicollinearity is therefore to find the eigenvalues and eigenvectors of $\mathbf{x}^T\mathbf{x}$. Zero or near-zero eigenvalues indicate exact or approximate multicollinearit. The corresponding eigenvectors tell us which combinations of regressor variables are causing the trouble. Sometimes this suggests useful simplifications of the data set; sometimes it tells us we're up the creek. For more, see Shalizi (2015), ch. 15.

5.2.2 Matrix powers and inverses, and eigenvalues and eigenvectors

Proposition 18 *The eigenvalues of $\mathbf{a}^k$, for positive integer k, are the powers of the eigenvalues of $\mathbf{a}$; the eigenvectors of $\mathbf{a}^k$ are the eigenvectors of $\mathbf{a}$. The eigenvalues of $\mathbf{a}^{-1}$ (when that matrix exists) are the reciprocals of the eigenvalues of $\mathbf{a}$.*

Proof (partial): If $\vec{v}$ is an eigenvector $\mathbf{a}$ with eigenvalue λ, then it's also an eigenvector of $\mathbf{a}^2$ with eigenvalue λ^2. This is because $\mathbf{a}^2\vec{v} = \mathbf{a}\mathbf{a}\vec{v} = \mathbf{a}\lambda\vec{v} = \lambda\mathbf{a}\vec{v} = \lambda^2\vec{v}$; iterating, $\vec{v}$ is an eigenvalue of $\mathbf{a}^k$ with eigenvalue λ^k. To see that $\mathbf{a}^k$ has no *other* eigenvalues, and to show the result for the inverse matrix $\mathbf{a}^{-1}$, is a little more involved, and left as an exercise.

5.3 Special Kinds of Matrix and Their Eigenvalues

Definition 16 *A matrix $\mathbf{a}$ is* **normal** *when $\mathbf{a}^T\mathbf{a} = \mathbf{a}\mathbf{a}^T$.*

Proposition 19 *All of the eigenvectors of a normal matrix are (or can be chosen to be) orthogonal to each other.*

Definition 17 *A square matrix is* **symmetric** *when it is equal to its own transpose, $\mathbf{a}^T = \mathbf{a}$.*

Proposition 20 *Every symmetric matrix is normal (in the sense of Definition 16).*

Proof: For a symmetric matrix, $\mathbf{a}^T\mathbf{a} = \mathbf{a}\mathbf{a}^T = \mathbf{a}^2$. □

Proposition 21 *For a symmetric matrix, all of the eigenvalues are real (not complex) numbers, and all of the eigenvectors can be chosen to be orthogonal to each other.*

Definition 18 *A square matrix is* **positive-semi-definite**, *or* **non-negative-definite**, *when, for all non-zero vectors $\vec{v}$, $\vec{v} \cdot \mathbf{a}\vec{v} \geq 0$. This is often abbreviated $\mathbf{a} \succeq 0$. If $\vec{v} \cdot \mathbf{a}\vec{v} > 0$, then the matrix is* **positive-definite**, *$\mathbf{a} \succ 0$.*

Proposition 22 *All the eigenvalues of a positive-semi-definite matrix are ≥ 0; all the eigenvalues of a positive definite matrix are > 0.*

Negative definite matrices are defined similarly, and have negative eigenvalues, but have fewer uses in statistics.

Definition 19 *A matrix is* **orthogonal** *when it is inverse is the same as its transpose, $\mathbf{o}^T = \mathbf{o}^{-1}$.*

Proposition 23 $\mathbf{o}$ *is orthogonal if and only if its rows are a set of orthonormal vectors, which happens if and only if its columns are orthonormal as well.*

On the basis of this proposition, it might have been better to call these matrices "orthonormal" rather than "orthogonal"; I haven't been able to work out where the name comes from.

Proposition 24 *1. All eigenvalues of an orthogonal matrix have magnitude 1.*

2. Since $\mathbf{o}^T\mathbf{o} = \mathbf{I}$*, if* $\mathbf{b}^T\mathbf{b} = \mathbf{a}$*, then* $(\mathbf{ob})^T(\mathbf{ob}) = \mathbf{a}$*.*

6 Eigendecomposition

Proposition 25 (Eigendecomposition or spectral decomposition) *Suppose* $\mathbf{a}$ *is a square,* $p \times p$ *matrix, and its eigenvectors are linearly independent. Then*

$$\mathbf{a} = \mathbf{vdv}^{-1} \tag{28}$$

where $\mathbf{v}$ *is the* $p \times p$ *matrix whose columns are the eigenvectors of* $\mathbf{a}$*, and* $\mathbf{d}$ *is the* $p \times p$ *diagonal matrix of* $\mathbf{a}$*'s eigenvalues.*

PROOF: Exercise 25. □

Corollary 1 *If the eigenvectors of* $\mathbf{a}$ *are linearly independent, then*

$$\mathbf{a}^{-1} = \mathbf{vd}^{-1}\mathbf{v}^{-1} \tag{29}$$

PROOF: $\mathbf{a}^{-1} = \left(\mathbf{vdv}^{-1}\right)^{-1}$, by the proposition. But this is just $(\mathbf{v}^{-1})^{-1}\mathbf{d}^{-1}\mathbf{v}^{-1}$, and the corollary follows. □

Corollary 2 *If the eigenvectors of* $\mathbf{a}$ *are linearly independent, and, in addition,* $\mathbf{a}$ *is normal (Definition 16), then* $\mathbf{v}$ *is orthogonal (Definition 19), and*

$$\mathbf{a} = \mathbf{vdv}^T \tag{30}$$

Remember that every symmetric matrix is normal, so the corollary applies to symmetric matrices.

6.1 Singular Value Decomposition

The spectral decomposition shows how to break up nice, square matrices into products of their eigenvalues and eigenvectors. It turns out that *any* matrix can be broken up into a product of eigenvalues and eigenvectors of related matrices.

Proposition 26 (Singular value decomposition) *An* $n \times m$ *matrix* $\mathbf{b}$ *has as its* **singular value decomposition**

$$\mathbf{b} = \mathbf{udv}^T \tag{31}$$

where $\mathbf{u}$*, the* $n \times n$ *matrix of* **left singular vectors***, contains the eigenvectors of* $\mathbf{bb}^T$*;* $\mathbf{v}$*, the* $m \times m$ *matrix of* **right singular vectors***, contains the eigenvectors of* $\mathbf{b}^T\mathbf{b}$*; and* $\mathbf{d}$ *is an* $n \times m$ *diagonal matrix of* **singular values***, containing the square roots of the non-zero eigenvalues of* $\mathbf{bb}^T$ *(which are also the eigenvalues of* $\mathbf{b}^T\mathbf{b}$*).*

In the special case of a square, $n \times n$ matrix, all of the matrices in the SVD are also $n \times n$, but don't much simplify further. However, one can interpret the terms more easily: multiplying by $\mathbf{v}^T$ rotates a vector to a new set of coordinate axes, multiplying by $\mathbf{d}$ stretches the vector along the axes (and possibly reflects along some of them), and then multiplying by $\mathbf{u}$ does a final rotation to new coordinates.

Proposition 27 *If* $\mathbf{b}$ *is symmetric as well as square, then* $\mathbf{bb}^T = \mathbf{b}^T\mathbf{b}$*, and the matrices* $\mathbf{u}$ *and* $\mathbf{v}$ *are identical. Moreover,* $\mathbf{d}$ *is simply the diagonal matrix of eigenvalues of* $\mathbf{b}$*.*

6.2 Square Root of a Matrix

Any matrix $\mathbf{b}$ such that $\mathbf{b}^T\mathbf{b} = \mathbf{a}$ is a **square root** of a symmetric matrix $\mathbf{a}$. There are infinitely many of them[1], but a fairly straightforward one can be defined through the eigendecomposition:

$$\mathbf{a}^{1/2} = \mathbf{d}^{1/2}\mathbf{v} \tag{32}$$

where $\mathbf{d}^{1/2}$ takes the square root of each element of $\mathbf{d}$, the diagonal matrix of eigenvalues of $\mathbf{a}$.

7 Orthogonal Projections, Idempotent Matrices

Definition 20 *The* **projection** *of a vector $\vec{v}$ on to a set s is the point in the set which comes closest to $\vec{v}$.*

Definition 21 *If s is a $q < p$ dimensional linear sub-space, we can find the projection by using an orthonormal basis for that subspace:*

$$\vec{v}^{\parallel} = \sum_{j=1}^{q} (\vec{v} \cdot \vec{s}_j)\vec{s}_j \tag{33}$$

This is the **orthogonal projection** *of $\vec{v}$ on to s.*

Proposition 28 *Any vector has a unique decomposition into its projection on to the subspace and a residual $\vec{v} = \vec{v}^{\parallel} + \vec{v}^{\perp}$. The residual $\vec{v}^{\perp}$ is orthogonal to the subspace, i.e., orthogonal to every vector in the subspace.*

Definition 22 *An operator (or matrix) $\mathbf{p}$ is* **idempotent** *when its powers are equal to itself, i.e., when $\mathbf{p}^k = \mathbf{p}$ for all integer $k > 1$.*

Proposition 29 *Every orthogonal project is idempotent.*

Geometrically, this means that once a vector has been projected into a subspace, projecting into the same space again does nothing.

Proposition 30 *All eigenvalues of an idempotent matrix are either 0 or 1, hence its rank is equal to its trace.*

Proof: Exercise 11. □

This fact is surprisingly useful for linear regression, and for linear smoothers in general. (See the section "Some General Theory for Linear Smoothers" in Chapter 1 of *ADAfaEPoV*.)

8 R Commands for Linear Algebra

For matrix multiplication, use the `%*%` operator; if one of its arguments is an R vector, it will (sometimes) try to convert it to an appropriate matrix, but you're probably better off doing that explicitly yourself.

`t()` transposes, `det()` and `determinant` calculate the determinant.

If `a` is an existing matrix, `diag(a)` can be used to extract or set the entries on its diagonal. Similarly, `lower.tri()` and `upper.tri` can extract or set the lower or upper triangular parts of a matrix.

If `v` is a vector, `diag(v)` creates the corresponding diagonal matrix. `diag(k)` for integer `k` creates a $k \times k$ identity matrix.

There is, oddly, no built-in function for calculating a trace, but it's easy to write one.

For a matrix `a` and a vector `b`, `solve(a,b)` solves the linear system $\mathbf{a}\vec{x} = \vec{b}$. With just a matrix, `solve(a)` finds $\mathbf{a}^{-1}$.

[1] Because, for any orthogonal matrix $\mathbf{c}$, $\mathbf{b}^T\mathbf{b} = \mathbf{b}^T\mathbf{o}^T\mathbf{o}\mathbf{b} = (\mathbf{o}\mathbf{b})^T(\mathbf{o}\mathbf{b})$.

`eigen(a)` returns a list containing the eigenvalues and (normalized) eigenvectors of `a`. Similarly, `svd(a)` finds the singular value decomposition.

9 Vector Calculus

Definition 23 *The* **gradient** *of a scalar-valued function is the vector-valued function of its first partial derivatives:*

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_p} \end{bmatrix} \tag{34}$$

∇f is a vector-valued function; $\nabla f(\vec{x})$ is a vector.

∇ is pronounced "grad" or, more rarely, "del". ("Nabla" is by far the least common way of pronouncing it, and apparently originated as a joke; it's a bit unfortunate it's how the makers of LaTeXchose to represent the symbol.)

Definition 24 *The derivative of f in direction $\tilde{u}$ at $\vec{x}$, or* **directional derivative***, is the rate of change in f as we move away from $\vec{x}$ in the direction $\tilde{u}$*

$$D_{\tilde{u}} f(x) \equiv \frac{df(\vec{x} + h\tilde{u})}{dh}(0) \tag{35}$$

Proposition 31 *The derivative in direction $\tilde{u}$ is the inner product of $\tilde{u}$ and the gradient:*

$$D_{\tilde{u}} f(\vec{x} = \tilde{u} \cdot \nabla f(\vec{x}) \tag{36}$$

Proof: Use the chain rule. □

At any point x, $\nabla f(\vec{x})$ is the direction in which the function f increases most rapidly. More exactly, the unit vector $\nabla f(\vec{x})/\|\nabla f(\vec{x})\|$ gives the direction of "steepest ascent", and $\|\nabla f(\vec{x})\|$ is the slope in that direction. $-\nabla f(\vec{x})$ likewise tells us the slope and direction of steepest descent.

The gradient itself changes; these changes show up in the second derivatives of f.

Definition 25 *The* **Hessian (matrix) H**f *of a scalar-valued function f is the matrix of its second partial derivatives,*

$$H_{ij} f = \frac{\partial^2 f}{\partial x_i \partial x_j} \tag{37}$$

Because $\partial^2 f/\partial x_i \partial x_j = \partial^2 f/\partial x_j \partial x_i$, the Hessian is always a symmetric matrix.

Some alternative notations are $\nabla^2 f$, $\nabla\nabla f$, and $\nabla \otimes \nabla f$, which all indicate the idea of taking derivatives of the gradient ∇f.

Some people use "the Hessian" to refer to the *determinant* of this matrix. But calling the matrix itself "the Hessian" is standard in statistical, machine learning, and much of optimization. If there's any chance of confusion, "the Hessian matrix" is a safe phrase.

In many situations, we do not need all of the second derivatives, but just the sum of the second derivatives along the coordinates.

Definition 26 *Let f be a scalar function of a p-dimensional vector space. Its* **Laplacian** *is the sum of its second partial derivatives along each coordinate axis:*

$$\Delta f = \sum_{i=1}^{p} \frac{\partial^2 f}{\partial x_i \partial x_i} \tag{38}$$

Alternate notations are $\nabla^2 f$ and $\nabla \cdot \nabla f$.

Notice that $\Delta f = \operatorname{tr} Hf$. While the Hessian matrix Hf is obviously coordinate-dependent, one can show that Δf is not. The Laplacian is centrally important to diffusion, and used extensively in nonlinear dimensionality reduction.

When we transform scalars into scalars, we often need to keep track of how lengths get altered. If we apply a function f, a small region of length dx around x gets transformed into a region of length $|df/dx|dx$ around $f(x)$. When we transform vectors into vectors, we need to keep track of how volumes transform, and this job is done a determinant.

Definition 27 *Let f be a mapping from a p-dimensional vector space to a q-dimensional vector space. The* **Jacobian** *of f is the determinant of its matrix of partial derivatives:*

$$Jf = \begin{vmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_p} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_q}{\partial x_1} & \frac{\partial f_q}{\partial x_2} & \cdots & \frac{\partial f_q}{\partial x_p} \end{vmatrix} \tag{39}$$

9.0.0.1 Scalar, linear functions of $\vec{x}$

$$\begin{aligned} \nabla \vec{b} \cdot \vec{x} &= \vec{b} && (40) \\ \text{because } \frac{\partial}{\partial x_i} \sum_{j=1}^{p} b_j x_j &= \sum_{j=1}^{p} b_j \frac{\partial x_j}{\partial x_i} && (41) \\ &= b_i && (42) \end{aligned}$$

So $\nabla \vec{b}^T \vec{x} = \nabla \vec{x}^T \vec{b}$.

9.0.0.2 Quadratic forms in $\vec{x}$

A **quadratic form** in a vector is an expression like $\vec{x} \cdot \mathbf{a}\vec{x} = \vec{x}^T \mathbf{a} \vec{x}$. This is a scalar, but it's quadratic in the coordinates of $\vec{x}$.

$$\begin{aligned} \nabla \vec{x} \cdot \mathbf{a}\vec{x} &= (\mathbf{a} + \mathbf{a}^T)\vec{x} && (43) \\ \text{because} \frac{\partial}{\partial x_i} \left(\sum_{j=1}^{p} x_j \sum_{k=1}^{p} a_{jk} x_k \right) &= \frac{\partial}{\partial x_i} \left(\sum_{j=1}^{p} \sum_{k=1}^{p} a_{jk} x_j x_k \right) && (44) \\ &= \frac{\partial}{\partial x_i} \left(\sum_{j=1}^{p} a_{jj} x_j^2 + \sum_{j=1}^{p} \sum_{k \neq j} a_{jk} x_j x_k \right) && (45) \\ &= 2a_{ii} x_i + \sum_{j \neq i} a_{ij} x_j + \sum_{k \neq i} a_{ki} x_k && (46) \\ &= \sum_{j=1}^{p} a_{ij} x_j + \sum_{k=1}^{p} a_{ki} x_k && (47) \\ &= (\mathbf{a}\vec{x})_i + (\mathbf{a}^T \vec{x})_i && (48) \end{aligned}$$

If $\mathbf{a}$ is symmetric so $\mathbf{a} = \mathbf{a}^T$, this simplifies:

$$\nabla \vec{x} \cdot \mathbf{a}\vec{x} = 2\mathbf{a}\vec{x} \tag{49}$$

10 Function Spaces

Spaces of functions can often be treated as linear vector spaces. The usual definition of inner product is

$$\langle f, g \rangle = \int f(x)g(x)dx \tag{50}$$

This leads to a norm for functions,

$$\|f\|_2^2 = \int f^2(x)dx \tag{51}$$

Functions where $\|f\|_2 < \infty$ are **square-integrable**, and the space of square-integrable functions is called L_2 (or sometimes L^2). (Exercise: Suggest a definition of L_p, without looking it up.) Sometimes it is convenient to introduce a "base" or "reference" density μ, and to modify the definition of inner product to

$$\langle f, g \rangle = \int f(x)g(x)\mu(x)dx \tag{52}$$

and

$$\|f\|_2^2 = \int f^2(x)\mu(x)dx \tag{53}$$

The space of function is then called $L_2(\mu)$. The inner product is linear in both f and g, with or without the factor of μ.

Similarly, one can define inner products and square-integrability for functions on a restricted domain, e.g., $\int_0^1 f^2(x)dx$, or combine a domain restriction with a non-uniform reference distribution.

10.1 Bases

Because $\langle f, g \rangle$ acts like an ordinary inner product, lots of what's familiar from vector spaces carries over to L_2. In particular, it makes sense to speak of a sequence of functions $\psi_1, \psi_2, \ldots$ being a basis, and even of its being an orthonormal basis, in which case

$$f = \sum_{j=1}^{\infty} \langle f, \psi_j \rangle \psi_j \tag{54}$$

Notice that the mononomials $1, x, x^2, \ldots$ are a basis for L_2 on $[0, 1]$ and on $\mathbb{R}$, but they are neither orthogonal nor normalized. Many families of orthogonal polynomials exist; the most straightforward begin by taking $\psi_1(x) = 1$, and then making ψ_k the $(k+1)$-degree polynomial which is orthogonal to all previous functions in the series.

Another basis for L_2 on $[0, 1]$ are the sines and cosines, which in this notation we may write as $\psi_1 = 1$, $\psi_{2k}(x) = \sin 2k\pi x$, $\psi_{2k+1} = \cos 2k\pi x$. This is the **Fourier basis** and the coefficients in this expansion are the Fourier coefficients or Fourier transform of the original function.

10.2 Eigenvalues and Eigenfunctions of Operators

In function spaces, an **operator** $\mathcal{O}$ is a a higher-order function, something which maps functions to other functions. You know two operators from calculus: Taking a derivative transforms one function, f, into another, df/dx. Likewise integration transforms f into $\int_{u=-\infty}^{x} f(u)du$, a function of x.

Operators can have eigenvalues and eigenfunctions, just as matrices have eigenvalues and eigenvectors:

$$\mathcal{O}f = \lambda f \tag{55}$$

You know examples of this, too, from calculus: e^{ax} is an eigenfunction of both differentiation and integration. (What are the eigenvalues?)

As the last example suggests, in general operators on function spaces have infinitely many eigenvalues and eigenvectors.

11 Further Reading

There are many, many decent references on linear algebra, often as part of a more general reference on mathematical methods, e.g., Boas (1983). Axler (1996) is notable for presenting the whole subject "from the ground up", but at the same time from the abstract perspective characteristic of modern mathematics (as opposed to sheer calculational procedures).

12 Exercises

To think through, in your copious free time, rather than to hand in, unless explicitly assigned.

1. Prove that for any $\vec{v}$, $0\vec{v} = \vec{0}$.
2. Prove that the span of of any set of vectors is a linear subspace.
3. Prove that the null space of a matrix $\mathbf{a}$, i.e., the set of all vectors where $\mathbf{a}\vec{v} = 0$, is in fact a linear subspace.
4. Suppose that $\mathbf{a}$ is symmetric, that $\mathbf{a}\vec{u} = \lambda_1\vec{u}$, $\mathbf{a}\vec{v} = \lambda_2\vec{v}$, and $\lambda_1 \neq \lambda_2$. Show that $\vec{u} \cdot \vec{v} = 0$. Why does your proof need $\mathbf{a}$ to be symmetric?
5. Assume the notation of Proposition 25
   a. Show that, for any square matrix $\mathbf{a}$,
   $$\mathbf{av} = \mathbf{vd} \tag{56}$$
   b. Show that if the eigenvectors of $\mathbf{a}$ are all linearly independent, then $\mathbf{v}^{-1}$ exists.
   c. Show that, when the eigenvectors are linearly independent, $\mathbf{a} = \mathbf{vdv}^{-1}$.
   d. Can $\mathbf{v}^{-1}$ exist if the eigenvectors of $\mathbf{a}$ are linearly dependent?
6. Show that $\vec{v}^{\parallel}$, as defined in Eq. 33, is also
   $$\underset{c_1, c_2, \ldots c_q}{\operatorname{argmin}} \|\vec{v} - \sum_{j=1}^{q} c_j \vec{s}_j\| \tag{57}$$
7. Show that $\vec{v}^{\perp} = \vec{v} - \vec{v}^{\parallel}$ is orthogonal to (i.e., has inner product zero with) any vector of the form $\sum_{j=1}^{q} b_j \vec{s}_j$.
8. Suppose that $\vec{v}_1$ and $\vec{v}_2$ are both eigenvectors of $\mathbf{a}$ with eigenvalue λ. Show that $c_1\vec{v}_1 + c_2\vec{v}_2$ is also an eigenvector of $\mathbf{a}$ with eigenvalue λ, for all c_1, c_2.
9. Let λ be an eigenvector of a matrix $\mathbf{a}$ with multiplicity k. Show that the set of eigenvectors of $\mathbf{a}$ with eigenvalue λ forms a linear subspace of dimension k. *Hints*: start with the $k = 1$ case; use Exercise 8.
10. Let $\mathbf{a}$ be an arbitrary $p \times p$ matrix. Show that its rank equal to p minus the eigenvalue multiplicity of 0. *Hint*: First, show that the dimension of $\mathbf{a}$'s null space must be $\geq$ the multiplicity of 0; then (slightly harder) that the dimension of the null space must be $\leq$ the number of zero eigenvalues.
11. Show the following:
    a. All eigenvalues of an idempotent matrix are either 0 or 1. *Hint*: try proof by contradiction: suppose there were some other eigenvalue λ, and show that this conflicts with idempotency.
    b. The rank of an idempotent matrix equals its trace.
    c. (More challenging) Is it true that if each eigenvalue of a matrix is either 0 or 1, the matrix must be idempotent? (Does it matter whether the eigenvectors form a basis?)
12. Let $\mathbf{a}$ be a symmetric, positive-definite matrix. Use the spectral decomposition to show that all its eigenvalues are > 0.

13. Show that $\langle f, g \rangle$, as defined in §10, is linear in both f and g.
14. Prove that if $\vec{v}$ is an eigenvector of $\mathbf{a}^k$, for positive integer $k > 1$, then it is also an eigenvector of $\mathbf{a}$. *Hint*: Assume it isn't, and derive a contradiction.
15. Prove that if $\vec{v}$ is an eigenvector of $\mathbf{a}$, then it is also and eigenvector of $\mathbf{a}^{-1}$, and vice versa. *Hint*: Assume it isn't, and derive a contradiction.

References

Axler, Sheldon. 1996. *Linear Algebra Done Right*. Berlin: Springer-Verlag.

Boas, Mary L. 1983. *Mathematical Methods in the Physical Sciences*. Second. New York: Wiley.

Shalizi, Cosma Rohilla. 2015. "The Truth About Linear Regression." Online Manuscript. http:///www.stat.cmu.edu/~cshalizi/TALR.