Some Linear Algebra for Econometrics

Frank Pinter*

January 17, 2020

Contents

1 Introduction

If probability and statistics are the foundation of econometrics, linear algebra is something closer to a toolbox. Vectors and matrices are an unavoidable part of the work of econometrics, and vector and matrix operations figure prominently in our formulas and proofs. As soon as we begin talking about vector-valued random variables, these operations grab a foothold in our work. Vectors and matrices also come up in situations like the following:

- We store data in the computer using data vectors and data matrices.
- Tools like R, Matlab, and NumPy are optimized for matrix calculations, so speaking the language of matrices makes our code run more quickly.[1]

Linear algebra also lets us think at a higher level of abstraction, and simplify our calculations and proofs. To raise the level of abstraction yet higher, we can use tools from functional analysis that conceptually unify the operations we perform on finite data (sample mean, variance, covariance, ordinary least squares) with their population analogs (expectation, variance, covariance, population regression).[2]

In these notes, I present results more generally than you probably saw in linear algebra, but more concretely than any course in functional analysis would. I do this partly to allow the jump to random variables to come naturally, but also to give a flavor of how deep these results are.

1.1 Note on references

These notes were originally written for the Harvard Economics math camp for incoming PhD students, August 2019. They are heavily based on textbooks and notes from other math and econometrics courses, including Gary Chamberlain's lecture notes, Paul Sally's *Tools of the Trade*, and the graduate Scientific Computing course (AM205) at Harvard. Their style is originally based on Ashesh Rambachan's probability and statistics notes.

These notes are not intended to replace a good course in linear algebra. In particular, they are targeted towards concepts that occur often in the econometrics course. Many standard topics are not covered, or are only covered briefly. In addition, these notes omit proofs, which are an essential part of learning the material. I recommend Hoffman and Kunze for an overview of linear algebra, Luenberger for inner product spaces and the projection theorem, and Golub and Van Loan for matrix decompositions.

2 Preliminaries

Abstract linear algebra is built on a litany of mathematical objects. Bear with me.

Definition 1. *A **commutative group** (or **abelian group**), denoted $(G, *)$, is a set G and an operation $* : G \times G \to G$ satisfying:*

[1] Computational linear algebra is a large field, and all empirical economists owe it an enormous debt. For a taste, consult Golub and Van Loan.

[2] "Some readers may look with great expectation toward functional analysis, hoping to discover new powerful techniques that will enable them to solve important problems beyond the reach of simpler mathematical analysis. Such hopes are rarely realized in practice. The primary utility of functional analysis for the purposes of this book is its role as a unifying discipline, gathering a number of apparently diverse, specialized mathematical tricks into one or a few general geometric principles." —David Luenberger

1. *Closure: for all* $a, b \in G$*, we have* $a * b \in G$*.*

2. *Associativity: for all* $a, b, c \in G$*, we have* $(a * b) * c = a * (b * c)$*.*

3. *Commutativity: for all* $a, b \in G$*, we have* $a * b = b * a$*.*

4. *Identity element: there exists* $e \in G$ *such that, for all* $a \in G$*,* $a * e = a$*.*

5. *Inverses: for each* $a \in G$*, there exists* $a' \in G$ *such that* $a * a' = e$*.*

Definition 2. *A* ***field****, denoted* $(F, +, \cdot)$*, is a set* F *and two operations* $+ : F \times F \to F$ *and* $\cdot : F \times F \to F$ *satisfying:*[3]

1. $(F, +)$ *forms a commutative group. We denote the additive identity element by* $0 \in F$ *and the additive inverse of* $a \in F$ *by* $-a$*.*

2. *Closure under* $\cdot$*: for all* $a, b \in F$*, we have* $a \cdot b \in F$*.*

3. *Associativity of* $\cdot$*: for all* $a, b, c \in F$*, we have* $(a \cdot b) \cdot c = a \cdot (b \cdot c)$*.*

4. *Commutativity of* $\cdot$*: for all* $a, b \in F$*, we have* $a \cdot b = b \cdot a$*.*

5. *Multiplicative identity element: there exists* $1 \in F$ *such that, for all* $a \in F$*,* $a \cdot 1 = a$*.*

6. *Multiplicative inverses for nonzero elements: for each* $a \in F \setminus \{0\}$*, there exists* $a^{-1} \in F$ *such that* $a \cdot a^{-1} = 1$*.*

7. *Distributivity: for all* $a, b, c \in F$*,* $a \cdot (b + c) = (a \cdot b) + (a \cdot c)$*.*

The rational numbers $\mathbb{Q}$, the real numbers $\mathbb{R}$, and the complex numbers[4] $\mathbb{C}$ are all fields. This note generally focuses on the real numbers here, though the results will generalize to the complex numbers.

2.1 Vector spaces

Definition 3. *A* ***vector space*** *over a field* F*, denoted* $(V, +, \cdot)$*, is a set* V *and two operations* $+ : V \times V \to V$ *and* $\cdot : F \times V \to V$ *satisfying:*

1. $(V, +)$ *forms a commutative group. We denote the additive identity element by* $0 \in V$ *and the additive inverse of* $v \in V$ *by* $-v$*.*

2. *Closure under* $\cdot$*: for all* $a \in F$ *and* $v \in V$*, we have* $a \cdot v \in V$*.*

3. *For all* $a \in F$ *and* $v_1, v_2 \in V$*, we have* $a \cdot (v_1 + v_2) = (a \cdot v_1) + (a \cdot v_2)$*.*

4. *For all* $a, b \in F$ *and* $v \in V$*, we have* $(a + b) \cdot v = (a \cdot v) + (b \cdot v)$*.*

5. *For all* $a, b \in F$ *and* $v \in V$*, we have* $(ab) \cdot v = a \cdot (b \cdot v)$*.*

6. *For all* $v \in V$*, we have* $1 \cdot v = v$*.*

[3] The $\cdot$ in multiplication is usually dropped.

[4] Numbers of the form $a + b\sqrt{-1}$, where a (the 'real part') and b (the 'imaginary part') are real numbers.

Elements of a vector space are called **vectors**, elements of the field are called **scalars**, and the operation $\cdot$ is called **scalar multiplication**.[5] The most familiar example is $\mathbb{R}^n$, considered as a vector space over $\mathbb{R}$; we will usually treat our data vectors as elements.[6]

Exercise 1. *Let X and Y be random variables. Show that $\{\alpha X + \beta Y \mid \alpha, \beta \in \mathbb{R}\}$ is a vector space over $\mathbb{R}$ with the usual addition and scalar multiplication operations for random variables. What is the additive identity element?*

Definition 4. *Let $(V, +, \cdot)$ be a vector space over a field F, and let $W \subseteq V$. We say W is a **subspace** of V if W is closed under addition and scalar multiplication.*

2.2 Dimension and basis

Let V be a vector space over a field F, let $v_1, \ldots, v_n \in V$, and let $\alpha_1, \ldots, \alpha_n \in F$. Then

$$\alpha_1 v_1 + \cdots + \alpha_n v_n$$

is a **linear combination** of the vectors $v_1, \ldots, v_n$. This is an enormously useful object. In particular, once we have a notion of a basis, we can uniquely represent every vector as a linear combination of basis vectors. These linear combinations are more concrete and easier to work with than the abstract vectors they represent, especially if the basis has nice properties.

Definition 5. *Let V be a vector space over a field F, and let $v_1, \ldots, v_n$ be nonzero vectors. The set $\{v_1, \ldots, v_n\}$ is **linearly independent** if, for any $\alpha_1, \ldots, \alpha_n \in F$,*

$$\alpha_1 v_1 + \cdots + \alpha_n v_n = 0 \implies \alpha_1 = \alpha_2 = \cdots = \alpha_n = 0.$$

Equivalently, the set is linearly independent if no element can be written as a linear combination of the other elements.

Definition 6. *Let V be a vector space over a field F, and let $v_1, \ldots, v_n \in V$. The set $\{v_1, \ldots, v_n\}$ **spans** V if, for any $v \in V$, v can be written as a linear combination of $\{v_1, \ldots, v_n\}$.[7] That is, there exist $\alpha_1, \ldots, \alpha_n \in F$ such that $v = \alpha_1 v_1 + \cdots + \alpha_n v_n$.*

We also sometimes refer to the **span** of $\{v_1, \ldots, v_n\}$, the set of vectors $v \in V$ that can be written as a linear combination of $\{v_1, \ldots, v_n\}$. The span of a set of vectors forms a subspace of V.

Definition 7. *Let V be a vector space over a field F and let $B \subseteq V$. We say B is a **basis** for V if B is linearly independent and spans V.*

Definition 8. *Let V be a vector space over a field F. We say V is **finite-dimensional** if there exists a finite $S \subseteq V$ that spans V.*

Theorem 1. *Let V be a nonzero finite-dimensional vector space over a field F. Then V has a finite basis, and every basis of V has the same number of elements.*

[5] The $\cdot$ in scalar multiplication is usually dropped.

[6] Some authors, such as Luenberger, distinguish between the space of points in $\mathbb{R}^n$ and the space of n-dimensional Euclidean vectors, which they label $\mathbb{E}^n$.

[7] This definition requires n to be finite. In general, a set S spans V if every $v \in V$ can be written as a linear combination of finitely many elements of S.

Definition 9. *Let V be a nonzero finite-dimensional vector space over a field F. Let B be a set of n vectors that forms a basis of V. Then the* ***dimension*** *of V, written $\dim V$, is equal to n.*

If $V = \{0\}$ then define $\dim V = 0$.

Every linearly independent set in V has at most $\dim V$ elements, and every set that spans V has at least $\dim V$ elements. Think of a basis as the (non-unique) smallest set that spans V.[8]

Example 1. *Consider $V = \mathbb{R}^n$ as a vector space over $\mathbb{R}$ with the usual addition and scalar multiplication operations. The* ***standard basis*** *$\{e_1, \ldots, e_n\}$, where e_j is has 1 in the jth coordinate and 0 everywhere else, is a basis for V, so $\dim V = n$.*

Let V be an n-dimensional vector space over a field F, and let $B = \{v_1, \ldots, v_n\}$ be a basis of V.[9] We can write $v \in V$ uniquely as a linear combination of basis vectors,

$$v = \sum_{i=1}^{n} \alpha_i v_i,$$

where $\alpha_1, \ldots, \alpha_n$ are the **coefficients** of v relative to the basis B. We conventionally write the coefficients in a **column vector** or $n \times 1$ matrix, and call this the **representation of v with respect to the basis B**:[10]

$$\begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix}.$$

In the case of $\mathbb{R}^n$, representation with respect to the standard basis coincides with the standard coordinate representation.[11]

Example 2. *Consider $\mathbb{R}^2$ as a vector space over $\mathbb{R}$. You can show that $B = \{(1,0), (1,1)\}$ is a basis of $\mathbb{R}^2$. Let $(a,b) \in \mathbb{R}^2$; its representation with respect to B is*

$$\begin{pmatrix} a-b \\ b \end{pmatrix}.$$

Vector spaces that are not finite-dimensional are called **infinite-dimensional**. The existence of a basis for every infinite-dimensional vector space is equivalent to the axiom of choice.[12]

[8] In a finite-dimensional vector space, any linearly independent set that is not a basis can be completed to form a basis.

Similarly, any set that spans V and is not a basis can drop elements until it becomes linearly independent.

[9] Throughout this note, we will write $B = \{v_1, \ldots, v_n\}$ as if B were a set. Since the order of the basis vectors matters, B is actually a sequence (or an **ordered basis**).

[10] Hoffman and Kunze call this the **coordinate matrix of v with respect to the ordered basis B** and denote it $[v]_B$.

[11] We can think of column vector representation as building an alternative coordinate system, with the basis vectors as building blocks.

[12] The set of polynomials with real coefficients is an infinite-dimensional vector space over $\mathbb{R}$. Can you think of a basis for it?

3 Linear Transformations and Matrices

Every matrix represents a function that maps one vector space to another. When thinking of the properties of matrices, I find it helpful look for a geometric intuition in terms of that function.[13]

Definition 10. *Let V and W be vector spaces over a field F. A* ***linear transformation*** *is a function $T : V \to W$ satisfying:*

1. *For all $v_1, v_2 \in V$, $T(v_1 + v_2) = T(v_1) + T(v_2)$.*
2. *For all $v \in V$ and $\alpha \in F$, $T(\alpha v) = \alpha T(v)$.*

3.1 Matrix representation

A linear transformation between two vector spaces can be represented as a matrix, by writing down what the transformation does to each basis vector. Let V and W be vector spaces over a field F, let $\{v_1, \dots, v_n\}$ be a basis for V, and let $\{w_1, \dots, w_m\}$ be a basis for W. For $k = 1, \dots, n$, consider the coefficients of $T(v_k)$ relative to the basis $\{w_1, \dots, w_m\}$:

$$T(v_k) = \sum_{j=1}^{m} a_{jk} w_j.$$

Then, since any $v \in V$ can be written as $v = \sum_{k=1}^{n} b_k v_k$, we can write $T(v)$ as a linear combination of $\{T(v_1), \dots, T(v_n)\}$:

$$T(v) = \sum_{k=1}^{n} b_k T(v_k) = \sum_{k=1}^{n} b_k \sum_{j=1}^{m} a_{jk} w_j = \sum_{j=1}^{m} \left(\sum_{k=1}^{n} b_k a_{jk} \right) w_j.$$

This gives us the coefficients of $T(v)$ relative to $\{w_1, \dots, w_m\}$. The results that follow just repeat this result using matrix multiplication.

Definition 11. *Let V and W be vector spaces over a field F, let $\{v_1, \dots, v_n\}$ be a basis for V, and let $\{w_1, \dots, w_m\}$ be a basis for W. Let $T : V \to W$ be a linear transformation. The* ***matrix representation*** *of T with respect to $\{v_1, \dots, v_n\}$ and $\{w_1, \dots, w_m\}$[14] is the $m \times n$ matrix with scalar entries*

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}$$

where the kth column, $(a_{1k}, \dots, a_{mk})$, contains the coefficients of $T(v_k)$ relative to $\{w_1, \dots, w_m\}$.[15]

The matrix representation is unique given a basis of V and a basis of W. We will say the ijth entry of a matrix is a_{ij}, the entry in row i and column j. We sometimes write the set of $m \times n$ matrices with entries in F as $F^{m \times n}$. An $n \times n$ matrix is called **square**.

[13] "You should be aware of the fact that an $m \times n$ matrix A with entries a_{ij} is more than just a static array of mn numbers. It is dynamic. It can act." —Charles Pugh

[14] When $V = W$, we will typically use the same basis for the domain and the range.

[15] Seeing the column vector of v as an $n \times 1$ matrix makes more sense if you consider F as a vector space over itself, with basis $\{1\}$, and consider a linear transformation $T : F \to V$ with $T(1) = v$.

Definition 12. *Let $A \in F^{m \times p}$ with entries a_{ij} and let $B \in F^{p \times n}$ with entries b_{jk}.* ***Matrix multiplication*** *is defined as*[16]

$$A \cdot B = C \text{ where } C \in F^{m \times n} \text{ with entries } c_{ik} = \sum_{j=1}^{p} a_{ij} b_{jk}.$$

Matrix multiplication represents function composition.

Theorem 2. *Let V and W be vector spaces over a field F, let $\{v_1, \ldots, v_n\}$ be a basis for V, and let $\{w_1, \ldots, w_m\}$ be a basis for W. Let $T : V \to W$ be a linear transformation, and let A be its matrix representation with respect to $\{v_1, \ldots, v_n\}$ and $\{w_1, \ldots, w_m\}$. For all $v \in V$, the representation of $T(v)$ with respect to $\{w_1, \ldots, w_m\}$ is $A \cdot v$.*

3.2 Properties of linear transformations

Though some results use the language of matrices, remember that linear transformations are always working in the background.[17]

Definition 13. *Let V be a vector space over a field F. The* ***identity function*** *$I : V \to V$ is the linear transformation defined by $I(v) = v$.*

Let $\dim V = n$. The matrix representation of I, with respect to any basis, is the $n \times n$ **identity matrix**, denoted I_n, with $a_{ij} = 1(i = j)$.

Definition 14. *Let V be a vector space over a field F, and let $T : V \to V$ be a linear transformation. The* ***inverse*** *of T, if it exists, is the function $T^{-1} : V \to V$ such that*

$$T^{-1}(T(v)) = I(v) = v.$$

If T^{-1} exists, then T is said to be ***invertible****.*

Proposition 1. *If T^{-1} exists, then T^{-1} is a linear transformation.*

Let $\dim V = n$ and let A be the matrix representation of T with respect to some basis. The **inverse matrix** of A, denoted A^{-1}, is the matrix representation of T^{-1} with respect to the same basis, and

$$A^{-1} A = I_n.$$

If A^{-1} exists, then A is said to be invertible.[18]

We will now give some of the (many) conditions that are equivalent to the invertibility of T.

Definition 15. *Let V and W be vector spaces over a field F and let $T : V \to W$ be a linear transformation. The* ***null space*** *(or* ***kernel****) of T is the set of vectors that T maps to the zero vector, $\{v \in V \mid T(v) = 0\}$. If V is finite-dimensional, the* ***nullity*** *of T is the dimension of the null space of T.*

[16] The $\cdot$ in matrix multiplication is usually dropped.

[17] To see these results in full generality using the language of linear transformations, see Hoffman and Kunze.

[18] Calculate the matrix inverse using `solve` in R, `inv` in Matlab, and `np.linalg.inv` in NumPy. To solve a system of linear equations $Ax = b$, instead use `solve(A, b)` in R, `A\b` in Matlab, and `np.linalg.solve(A, b)` in NumPy.

Definition 16. *If V is finite-dimensional, the **rank** of T is the dimension of the subspace $T(V) \subseteq W$.*

Theorem 3. *Rank-nullity theorem*

Let V and W be vector spaces over a field F and let $T : V \to W$ be a linear transformation. Suppose that V is finite-dimensional. Then

$$\operatorname{rank}(T) + \operatorname{nullity}(T) = \dim V.$$

The notion of rank for linear transformations corresponds to the familiar notion of rank for matrices.

Definition 17. *Let $A \in F^{m\times n}$, where F is a field. Write the jth column as $v_j = (a_{1j}, \ldots, a_{mj}) \in F^m$. The **column space** of A is the subspace of F^m spanned by the columns, $\{v_1, \ldots, v_n\}$, and the **column rank** of A is the dimension of its column space.*

*Likewise, write the ith row of A as $w_i = (a_{i1}, \ldots, a_{in}) \in F^n$. The **row space** of A is the subspace of F^n spanned by the rows, $\{w_1, \ldots, w_m\}$, and the **row rank** of A is the dimension of its row space.*

Proposition 2. *Let V and W be vector spaces over a field F and let $T : V \to W$ be a linear transformation. Let A be a matrix representation of T; then $A \in F^{m\times n}$. The rank of T is equal to the row rank of A and to the column rank of A.*

It follows that the row rank and the column rank are equal. We call this the rank of A.

Theorem 4. *Let V be a vector space over a field F with $\dim V = n$ and let $T : V \to V$ be a linear transformation. Let A be the matrix representation of T with respect to a basis. The following are equivalent:*

1. *T is invertible.*
2. *T is **full rank**; that is, $\operatorname{rank}(T) = n$.*
3. *T is **nonsingular**; that is, $\operatorname{nullity}(T) = 0$.*
4. *A is an invertible matrix.*
5. *A is full rank; that is, $\operatorname{rank}(A) = n$.*
6. *A is nonsingular; that is, $Av = 0$ implies $v = 0$.*

As a result, 'nonsingular' is sometimes used interchangeably with 'invertible'.

3.3 Transpose

The familiar notion of a matrix transpose is described below.[19]

Definition 18. *Let $A \in F^{m\times n}$ with entries a_{ij}. The **transpose** of A, denoted A' or $A^\intercal$, is the $n \times m$ matrix whose ijth entry is a_{ji}.*

A square matrix A for which $A' = A$ is called **symmetric**. You can show that, if A is $m \times p$ and B is $p \times n$, then $(AB)' = B'A'$.

[19] See Sections 3.5–3.7 of Hoffman and Kunze for the definition of the transpose of a linear transformation.

3.4 Diagonal and triangular matrices

Definition 19. *Let $A \in F^{m \times n}$ with entries a_{ij}. A is **diagonal** if $a_{ij} = 0$ when $i \neq j$.*

A square $n \times n$ diagonal matrix takes the form

$$\begin{pmatrix} c_1 & 0 & \cdots & 0 \\ 0 & c_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & c_n \end{pmatrix}.$$

It represents the linear transformation $T : V \to V$ that scales each basis vector v_i by the scalar c_i. It is clearly symmetric, and if all the $c_i \neq 0$, then the inverse is the diagonal matrix with diagonal entries $c_1^{-1}, c_2^{-1}, \ldots, c_n^{-1}$. In this case, the system of linear equations $Ax = b$ is easy to solve: $x_i = c_i^{-1} b_i$.

The computational simplicity remains if we add entries above or below the diagonal, but not both.[20]

Definition 20. *Let $A \in F^{n \times n}$ with entries a_{ij}.*

*A is **lower triangular** if all entries above the diagonal are zero, that is, $a_{ij} = 0$ when $i < j$.*

*A is **upper triangular** if all entries below the diagonal are zero, that is, $a_{ij} = 0$ when $i > j$.*

A diagonal matrix is both lower and upper triangular. The product of upper triangular matrices is upper triangular, and the product of lower triangular matrices is lower triangular.

3.5 Trace, determinant, and eigenvalues

Trace and determinant occur all the time in matrix computations and proofs, but will not recur for the rest of this note. Eigenvalues, a fascinating and rich topic in linear algebra, will only be discussed briefly.

Definition 21. *Let $A \in F^{n \times n}$ and label its entries by a_{ij}. The **trace** of A, denoted $\operatorname{trace}(A)$, is the sum of the diagonal entries, $\sum_{i=1}^n a_{ii}$.*

Definition 22. *If $B \in F^{n \times n}$ is invertible and $A = B^{-1}CB$, then A and C are **similar**.*

If two matrices are similar, they have the same trace. In addition, if A is $m \times n$ and B is $n \times m$, then $\operatorname{trace}(AB) = \operatorname{trace}(BA)$.

To get intuition for the determinant, consider $A \in \mathbb{R}^{n \times n}$. Label its columns $v_1, \ldots, v_n$ and picture the convex hull of its columns, $\{\alpha_1 v_1 + \cdots + \alpha_n v_n \mid \alpha_i \in [0, 1]\}$. This is a figure in $\mathbb{R}^n$ (called a parallelepiped; see Figure 1), and the determinant of A is its oriented volume.

Definition 23. *Let $A \in F^{n \times n}$ with entries a_{ij}. The **determinant** of A, denoted $\det A$ or $|A|$, is defined as follows.[21] If $n = 1$, then $\det A = a_{11}$. For $n > 1$, let A_{ij} be the $(n-1) \times (n-1)$ matrix generated by deleting the ith row and jth column of A. Then, for any $i \in \{1, \ldots, n\}$,*

$$\det A = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det A_{ij}.$$

[20] Lower triangular systems are easy to solve by forward substitution, and upper triangular systems are easy to solve by back substitution. See Golub and Van Loan.

[21] See Hoffman and Kunze, Chapter 5, for other equivalent characterizations.

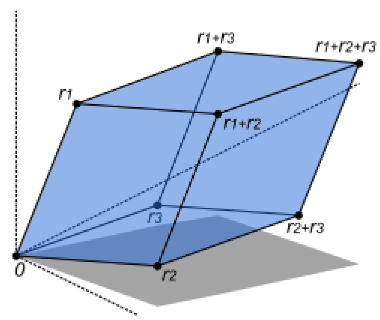

Figure 1: Parallelepiped in $\mathbb{R}^3$. (Claudio Rocchini, Wikimedia Commons)

Familiar facts about determinants can be guessed from the picture.

- $\det I_n = 1$. In general, the determinant of a diagonal matrix is the product of the diagonal entries.
- $\det A \neq 0$ if and only if A is invertible.[22]
- If A and B are square, then $\det AB = \det A \cdot \det B$. It follows that if A is invertible, then $\det A^{-1} = 1/\det A$.
- $\det A = \det A'$.

Next we discuss eigenvectors and eigenvalues, which measure the ways that a linear transformation shrinks, grows, flips, or deforms the shapes it acts on.

Definition 24. *Let $A \in F^{n\times n}$. An* ***eigenvalue*** *of A is a scalar $\lambda \in F$ for which there exists some $v \in F^n \setminus \{0\}$, called an* ***eigenvector****, such that*

$$Av = \lambda v.$$

Any nonzero scalar multiple of v is also an eigenvector corresponding to λ.

Theorem 5. *The eigenvalues of A are the roots of the* ***characteristic polynomial*** *of A,*

$$p(x) = \det(A - xI_n).$$

Suppose $F = \mathbb{C}$. Then, because p is an n-order polynomial over $\mathbb{C}$, it follows from the fundamental theorem of algebra that p has n complex roots, counted with multiplicity.[23] We can label the eigenvalues (repeating with multiplicity) by $(\lambda_1, \dots, \lambda_n)$.

The following result is surprisingly useful.

Theorem 6. *Let $A \in \mathbb{C}^{n\times n}$. Then*

$$\operatorname{trace}(A) = \sum_{i=1}^{n} \lambda_i \quad \text{and} \quad \det A = \prod_{i=1}^{n} \lambda_i.$$

[22]The dimension of the parallelepiped is the column rank of A; if A is not full rank, then the parallelepiped is a zero-measure set in $\mathbb{R}^n$.

[23]A root λ has multiplicity k if we can factor $p(x) = (x-\lambda)^k s(x)$, where $s(x)$ is a polynomial and $s(\lambda) \neq 0$. The multiplicity of the root is called the **algebraic multiplicity** of the corresponding eigenvalue.

Exercise 2. *Prove that if two matrices are similar, then they have the same characteristic polynomial, and therefore the same eigenvalues.*

The eigendecomposition anticipates the matrix decompositions we will consider later. It is sometimes helpful in computations and proofs, and gives insight into the workings of a linear transformation.[24]

Definition 25. *Let $A \in F^{n\times n}$. A is* ***diagonalizable*** *if there exists a basis of F^n whose elements are all eigenvectors of A.*

For A to be diagonalizable, it is sufficient that (1) A has n distinct eigenvalues, or (2) A is symmetric with real entries.[25]

Theorem 7. *Eigendecomposition (or spectral decomposition)*

Let $A \in \mathbb{C}^{n\times n}$ be diagonalizable. Write its eigenvalues, with multiplicity, as $(\lambda_1, \dots, \lambda_n)$. Then $A = BCB^{-1}$, where C is the diagonal matrix with diagonal elements $\lambda_1, \dots, \lambda_n$, and the columns of B are eigenvectors of A corresponding to those eigenvalues.

4 Inner Products

From now on, we will restrict ourselves to vector spaces over $\mathbb{R}$.

Definition 26. *An* ***inner product space*** *(or* ***pre-Hilbert space****) over $\mathbb{R}$ is a vector space V over $\mathbb{R}$ together with a function $\langle\cdot,\cdot\rangle : V \times V \to \mathbb{R}$, known as an* ***inner product*** *(or* ***positive definite symmetric bilinear form****), satisfying:*[26]

1. *Symmetry: for all $v, w \in V$, $\langle v, w\rangle = \langle w, v\rangle$.*
2. *Bilinearity (I): for all $v_1, v_2, w \in V$, $\langle v_1 + v_2, w\rangle = \langle v_1, w\rangle + \langle v_2, w\rangle$.*
3. *Bilinearity (II): for all $v, w \in V$ and $\alpha \in \mathbb{R}$, $\langle \alpha v, w\rangle = \alpha\langle v, w\rangle$.*
4. *Positive definiteness: for all $v \in V$, $\langle v, v\rangle \geq 0$ and $\langle v, v\rangle = 0$ if and only if $v = 0$.*

Definition 27. *A* ***normed linear vector space*** *over a field F is a vector space V over F together with a function $\|\cdot\| : V \to F$, known as a* ***norm****, satisfying:*

1. *For all $v \in V$, $\|v\| \geq 0$, with $\|v\| = 0$ if and only if $v = 0$.*
2. *Triangle inequality: for all $v, w \in V$, $\|v + w\| \leq \|v\| + \|w\|$.*
3. *For all $v \in V$ and $\alpha \in F$, $\|\alpha v\| = |\alpha|\|v\|$.*

Theorem 8. *Let V be an inner product space over $\mathbb{R}$. Then the function $\|\cdot\| : V \to \mathbb{R}$ defined by $\|v\| = \sqrt{\langle v, v\rangle}$ is a norm.*

[24] Calculate the eigendecomposition using `eigen` in R, `eig` in Matlab, and `np.linalg.eig` in NumPy.

[25] If $A \in \mathbb{C}^{n\times n}$ is symmetric and every entry has imaginary part zero, then all its eigenvalues have imaginary part zero.

[26] The notion of an inner product space over $\mathbb{C}$ is also well-defined, if we replace symmetry with conjugate symmetry. All the results below hold for inner product spaces over $\mathbb{R}$ or $\mathbb{C}$, so I will usually drop "over $\mathbb{R}$".

We make this distinction because some normed linear vector spaces have norms that do not come from an inner product. The proof of the triangle inequality uses the following result:

Theorem 9. *Cauchy-Schwarz inequality*

Let V be an inner product space. For all $v, w \in V$, we must have that $|\langle v, w\rangle| \leq \|v\|\|w\|$.

Furthermore, equality holds if and only if one of the vectors is a scalar multiple of the other.

4.1 Dot product and positive definiteness

In the geometry of $\mathbb{R}^n$, the familiar dot product is an inner product. It provides much of the geometric intuition we will lean on for inner product spaces.

Example 3. *$\mathbb{R}^n$ with the dot product is an inner product space over $\mathbb{R}$. Let $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$. The* ***dot product*** *of x and y, denoted $x \cdot y$ or $x'y$,[27] is $x_1y_1 + \cdots + x_ny_n$.*

The norm induced by the dot product has a nice interpretation as the length or Euclidean distance. The dot product itself also has a geometric interpretation. Let $x, y \in \mathbb{R}^n$, and let θ be the angle between x and y ($0° \leq \theta \leq 180°$). Then

$$\cos\theta = \frac{x \cdot y}{\|x\|\|y\|}.$$

Definition 28. *Let $A \in \mathbb{R}^{n\times n}$.*

A is ***positive semidefinite*** *if, for all $x \in \mathbb{R}^n$, $x \cdot Ax \geq 0$.*

A is ***positive definite*** *if, for all nonzero $x \in \mathbb{R}^n$, $x \cdot Ax > 0$.*

That is, the angle between x and Ax is (weakly) less than 90°. Objects of the form $x'Ax$ are called **quadratic forms**.

Proposition 3. *If A is symmetric and positive definite, then its eigenvalues are positive. If A is symmetric and positive semidefinite, then its eigenvalues are nonnegative.*

4.2 Generalizations of dot product

The dot product generalizes conveniently to inner products in infinite-dimensional vector spaces over $\mathbb{R}$. These are useful when we move to sequences, functions, and random variables.

Example 4. *Consider infinite sequences of numbers.[28] Define l_2 as the space of real-valued infinite sequences (x_n) where $\sum_{i=1}^{\infty} |x_i|^2 < \infty$, with the inner product as an infinite dot product,*

$$\langle (x_n), (y_n)\rangle = \sum_{i=1}^{\infty} x_i y_i.$$

Example 5. *Now consider well-behaved functions.[29] For any $a < b$ define $L_2[a, b]$ as the space of functions $x : [a, b] \to \mathbb{R}$ for which $|x(t)|^2$ is Lebesgue integrable.[30] The inner product is a Lebesgue*

[27] We snuck in matrix multiplication here, by representing x and y as column vectors with respect to the standard basis.

[28] This generalizes to discrete random variables with (countably) infinite support.

[29] This generalizes to well-behaved continuous random variables.

[30] Because the Lebesgue integral is not sensitive to the behavior of x on zero-measure sets, we need to consider two functions x and y the same if they are the same almost everywhere (they only differ on a set of measure zero).

integral,

$$\langle x, y\rangle = \int_a^b x(t)y(t)\,dt.$$

4.3 Orthogonality

The notion of orthogonality is useful in two ways. First, it is closely connected to linear projection. Second, an orthogonal set (and particularly an orthonormal set) makes for a convenient basis.

Definition 29. *Let V be an inner product space and let $v, w \in V$. We say v and w are* ***orthogonal*** *(or* ***perpendicular****) if $\langle v, w\rangle = 0$.*

A vector v is orthogonal to a set $W \subseteq V$ if $\langle v, w\rangle = 0$ for all $w \in W$.

Definition 30. *Let V be an inner product space. We say $\{v_1, \ldots, v_n\} \subseteq V$ is an* ***orthonormal set*** *if it is pairwise orthogonal (that is, $\langle v_i, v_j\rangle = 0$ whenever $i \neq j$) and $\langle v_i, v_i\rangle = 1$ for all i.*

If $n = \dim V$, then $\{v_1, \ldots, v_n\}$ forms an **orthonormal basis** of V. The standard basis is an example.[31]

Definition 31. *Let $Q \in \mathbb{R}^{n\times n}$. Q is an* ***orthogonal matrix*** *if its columns[32] form an orthonormal basis of $\mathbb{R}^n$ with the dot product.*

If Q is an orthogonal matrix, then

$$Q'Q = QQ' = I_n.$$

That is, $Q' = Q^{-1}$. Orthogonal matrices represent functions that perform rotation or reflection. In particular, they preserve length:

$$\|Qv\|^2 = (Qv)'Qv = v'Q'Qv = v'v = \|v\|^2.$$

5 Linear Projections

In econometrics we are often faced with problems of the form

$$\widehat{v} = \operatorname*{arg\,min}_{w\in W} \|v - w\|$$

asking us to find the best approximation to a vector.[33] The projection theorem gives us simple conditions that guarantee existence and uniqueness of solutions. Furthermore, it gives us a simple way to compute them as the solution to a system of equations.[34]

The best approximation problem asks us to find the projection of a vector on a subspace. From geometry in $\mathbb{R}^n$, we have an intuition that the line connecting the projection with the original vector should be perpendicular to the subspace. Start with the projection of one vector onto another.

[31] As is any rotation or reflection thereof.
[32] We could equivalently define it using the rows.
[33] For example, ordinary least squares.
[34] No calculus or second order conditions required.

Definition 32. *Let V be an inner product space over $\mathbb{R}$ and let $v, w \in V$. The **projection** of v onto w is defined by*

$$\operatorname{proj}_w(v) = \frac{\langle v, w \rangle}{\|w\|^2}\, w.$$

This is the closest point to v in the subspace spanned by w. Furthermore, $v - \operatorname{proj}_w(v)$ is orthogonal to w.

5.1 Gram–Schmidt orthogonalization

This is the insight that allows us to generate a basis of orthonormal vectors for any finite-dimensional inner product space. The **Gram–Schmidt orthogonalization procedure** takes any finite linearly independent set of vectors, labeled $\{v_1, \ldots, v_n\}$, and generates an orthonormal set that spans the same subspace. It works by iteratively projecting v_k on the vectors that came before, keeping only the orthogonal residual. For $k = 1, \ldots, n$:

$$\begin{aligned}
z_1 &= v_1, & e_1 &= \frac{z_1}{\|z_1\|} \\
z_2 &= v_2 - \langle v_2, e_1 \rangle e_1, & e_2 &= \frac{z_2}{\|z_2\|} \\
&\vdots \\
z_k &= v_k - \sum_{i=1}^{k-1} \langle v_k, e_j \rangle e_j, & e_k &= \frac{z_k}{\|z_k\|}.
\end{aligned}$$

Notice the similarity to residual regression.[35]

Proposition 4. *Let $\{v_1, \ldots, v_n\}$ be a set of linearly independent vectors in an inner product space V. For each $k = 1, \ldots, n$, the subspace spanned by $\{v_1, \ldots, v_k\}$ is also spanned by the orthogonal set $\{z_1, \ldots, z_k\}$ and by the orthonormal set $\{e_1, \ldots, e_k\}$.*

5.2 Uniqueness of the projection

Here we generalize from projections onto one-dimensional subspaces to multidimensional subspace W. If the projection is possible, the fact that $v - \operatorname{proj}_w(v)$ is orthogonal to w generalizes nicely.

Theorem 10. *Let V be an inner product space, let W be a subspace of V, and let $v \in V$. If there exists $\widehat{v} \in W$ such that $\|v - \widehat{v}\| \leq \|v - w\|$ for all $w \in W$, then $\widehat{v}$ is unique.*

A necessary and sufficient condition for $\widehat{v}$ to be the unique minimizing vector is that $v - \widehat{v}$ is orthogonal to W.

The orthogonality condition, $\langle v - \widehat{v}, w \rangle = 0$ for all $w \in W$, is often much easier to solve than the original minimization problem.

5.3 Hilbert spaces and the projection theorem

We need further assumptions to guarantee that $\widehat{v}$ exists. The tools we need should be familiar from real analysis.

[35] Gram–Schmidt is one way to calculate the QR decomposition, which reduces ordinary least squares to the exact solution of a triangular system of equations. R and Matlab use the QR decomposition to calculate OLS because it is more stable than the traditional formula when $X'X$ is almost singular.

Definition 33. *Let X be a space equipped with a norm $\|\cdot\|$. A sequence (x_n) of elements in X is* ***Cauchy*** *if, for each $\epsilon > 0$, there exists an N such that $\|x_n - x_m\| \leq \epsilon$ for all $n, m > N$.*

Definition 34. *We say X is* ***complete*** *if every Cauchy sequence in X converges to a point in X.*

Definition 35. *A complete inner product space is called a* ***Hilbert space****.*

$\mathbb{R}^n$, l_2, and $L_2[a, b]$ are all Hilbert spaces.

Theorem 11. *Classical projection theorem*

Let V be a Hilbert space and let W be a closed subspace[36] *of V. For each $v \in V$, there exists a unique $\widehat{v} \in W$ such that $\|v - \widehat{v}\| \leq \|v - w\|$ for all $w \in W$.*

As before, a necessary and sufficient condition for $\widehat{v}$ to be the unique minimizing vector is that $v - \widehat{v}$ is orthogonal to W.

We call $\widehat{v}$ the **orthogonal projection** of v onto W.

Exercise 3. *Let $\{e_1, \ldots, e_n\}$ be an orthonormal basis of W. Show that the orthogonal projection of v onto W is $\sum_{k=1}^{n} \langle v, e_k \rangle e_k$.*

The projection theorem can be framed in another way, which might make the connection to linear regression easier to see.

Definition 36. *Let V be an inner product space, and let W be a subspace of V. The* ***orthogonal complement*** *of W, denoted $W^{\perp}$, is the set of vectors orthogonal to W.*

Proposition 5. *From the projection theorem: for any Hilbert space V and any closed subspace W of V, any $v \in V$ can be written uniquely in the form $v = \widehat{v} + \epsilon$, where $\widehat{v} \in W$ and $\epsilon \in W^{\perp}$.*

6 Matrix Decompositions

6.1 Singular value decomposition

The singular value decomposition is used in proofs, and can also be helpful in computations.[37] The idea is that every linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$ does the following three things, in order:

1. A rotation and/or reflection in $\mathbb{R}^n$.
2. A scaling along each basis vector in $\mathbb{R}^n$, with the output in $\mathbb{R}^m$.
3. A rotation and/or reflection, now in $\mathbb{R}^m$.

Consider the unit sphere in $\mathbb{R}^n$. Its image under any linear transformation is a hyperellipse. The lengths of the ellipse's axes capture the scaling in the second step, and are called singular values.

Theorem 12. *Singular value decomposition (SVD)*

Let $A \in \mathbb{R}^{m \times n}$. Then there exist an $m \times m$ orthogonal matrix U and an $n \times n$ orthogonal matrix V such that

$$A = U \Sigma V'$$

where Σ is a diagonal matrix with $p = \min\{n, m\}$ diagonal entries.

[36] In the sense that W contains all its limit points.

[37] Calculate it using `svd` in R and Matlab, and `np.linalg.svd` in NumPy.

The diagonal entries of Σ are called **singular values** and labeled $(\sigma_1, \ldots, \sigma_p)$. They are uniquely determined,[38] all nonnegative, and conventionally sorted so that $\sigma_1 \geq \cdots \geq \sigma_p \geq 0$. The columns of U are called **left singular vectors** and the columns of V are called **right singular vectors**.

Theorem 13. *Reduced SVD*

Suppose that $m \geq n$. Then there exist an $m \times n$ matrix U_1 with orthonormal columns and an $n \times n$ orthogonal matrix V such that

$$A = U_1 \widehat{\Sigma} V'$$

where $\widehat{\Sigma}$ is an $n \times n$ diagonal matrix.

Note that U_1 just contains the first n columns of U; we could partition $U = \begin{pmatrix} U_1 & U_2 \end{pmatrix}$. Intuitively, we go from the reduced SVD to the full SVD by stuffing U_1 with $m - n$ orthonormal vectors and adding $m - n$ zero rows to the bottom of $\widehat{\Sigma}$.

It follows in the $m \geq n$ case that, letting $\{v_1, \ldots, v_n\}$ be the columns of V and letting $\{u_1, \ldots, u_n\}$ be the columns of U_1,

$$Av_i = \sigma_i u_i, \quad A' u_i = \sigma_i v_i.$$

An interesting consequence is that $(\sigma_1^2, \ldots, \sigma_p^2)$ are the eigenvalues of $A'A$ and AA'.

6.2 Cholesky factorization

The Cholesky factorization can be helpful for sampling from multivariate normal distributions.[39]

Theorem 14. *Cholesky factorization*

Let $A \in \mathbb{R}^{n \times n}$ be symmetric and positive definite. Then there exists a unique lower triangular matrix $L \in \mathbb{R}^{n \times n}$ with positive diagonal entries such that $A = LL'$.

Remark 1. *Matrix square root*

Suppose A is a symmetric positive definite matrix and we are interested in finding B such that $B^2 = A$.[40] Construct it as follows. Let $A = LL'$ using the Cholesky factorization. Let $L = U\Sigma V'$ using the singular value decomposition. Then take

$$A^{1/2} = U \Sigma U'.$$

You can check that $A^{1/2}$ is symmetric positive definite, and $A^{1/2} A^{1/2} = A$.

7 Matrix Stacking and the Kronecker Product

When working with matrices as data objects, it is often convenient to move around entries of the matrix to support calculations. The vec operator and the Kronecker product can be useful toward this end.[41]

[38] That is, every SVD of A has the same singular values, though perhaps in a different order.
[39] Calculate it using `chol` in R and Matlab, and `np.linalg.cholesky` in NumPy.
[40] This is easy if A is diagonal.
[41] Calculate the Kronecker product using `kronecker` in R, `kron` in Matlab, and `np.kron` in NumPy.

Definition 37. *Let $B \in \mathbb{R}^{m_1 \times n_1}$ with entries b_{ij} and $C \in \mathbb{R}^{m_2 \times n_2}$. Their* ***Kronecker product*** *(or* ***Kronecker tensor product****) $B \otimes C$ is the $m_1 m_2 \times n_1 n_2$ matrix given in the following block matrix form:*

$$B \otimes C = \begin{pmatrix} b_{11}C & \cdots & b_{1n_1}C \\ \vdots & \ddots & \vdots \\ b_{m_1 1}C & \cdots & b_{m_1 n_1}C \end{pmatrix}.$$

The Kronecker product has the following useful properties:

- $(B \otimes C)' = B' \otimes C'$.
- $(B \otimes C)(D \otimes F) = BD \otimes CF$, if the matrices are conformable.[42]
- $(B \otimes C)^{-1} = B^{-1} \otimes C^{-1}$.

When is this useful? The operation of matrix stacking gives a hint.

Definition 38. *Let $A \in \mathbb{R}^{m \times n}$ and denote its columns by $v_1, \ldots, v_n$. The* ***vec operator*** $\text{vec}(A)$ *maps A to the $nm \times 1$ column vector generated by stacking its columns:*

$$\text{vec}(A) = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}.$$

Proposition 6. *If $B \in \mathbb{R}^{m_1 \times n_1}$, $C \in \mathbb{R}^{m_2 \times n_2}$, and $X \in \mathbb{R}^{n_1 \times m_2}$, then*

$$Y = CXB' \iff \text{vec}(Y) = (B \otimes C)\ \text{vec}(X).$$

References

Golub, G. and C. Van Loan. (2013). *Matrix Computations*, 4th ed.

Hoffman, K. and R. Kunze. (1971). *Linear Algebra*, 2nd ed.

Luenberger, D. (1969). *Optimization by Vector Space Methods.*

[42] That is, if matrix multiplication makes sense here.