Matrix Algebra Refresher

Diógenes Cruz Figueroa García

November, 2023

Contents

1 Notation and Definitions **2**
1.1 Types of matrices . . . 3
1.2 Transpose and Conjugate Transpose . . . 4

2 Vectors **5**
2.1 Linear Combinations . . . 5
2.2 Dot Product and euclidean Norm . . . 7
2.3 Orthogonality and Orthonormality . . . 8
2.4 Cauchy-Schwarz and Triangle Inequalities . . . 9

3 Matrix algebra **9**
3.1 Matrix-vector multiplication . . . 9
3.1.1 Vector-vector mutliplication, dot product and orthogonality (again) . . . 11
3.2 Matrix operators and properties . . . 13
3.2.1 The trace operator . . . 13
3.2.2 Span, Rank, Kernel and Nullity . . . 13
3.2.3 The Determinant . . . 15
3.2.4 Eigenvalues and Eigenvectors . . . 17
3.2.5 Signing a matrix . . . 18
3.2.6 Norm of a matrix, and submultiplicativity (Cauchy-Schwartz) . . . 19
3.2.7 Matrix-matrix multiplication . . . 19
3.2.8 The Inverse Matrix . . . 21
3.2.9 Orthogonal and Unitary matrices . . . 22
3.3 The Rank-Nullity Theorem . . . 23
3.4 Matrix decomposition . . . 23
3.4.1 Spectral Decomposition . . . 24
3.4.2 Eigendecomposition . . . 24
3.4.3 Singular Value Decomposition . . . 24
3.4.4 Rank Decomposition . . . 24
3.5 Inverse generalizations . . . 25
3.5.1 Right and left inverses . . . 25
3.5.2 Moore-Penrose inverse or Pseudoinverse . . . 25
3.6 System of equations . . . 26

3.6.1 Linear equations . . . 26
3.6.2 Affine transformation . . . 27
3.7 Idempotent matrix . . . 27
3.7.1 Application: Distribution of the variance estimator . . . 28
3.8 Demeaning and first-differencing matrices . . . 29

4 Matrix Calculus **30**
4.1 Gradient, Hessian and Jacobian . . . 30
4.2 Matrix derivatives . . . 31
4.3 Chain rule . . . 32
4.4 Examples . . . 33
4.4.1 OLS . . . 33
4.4.2 Logit . . . 34
4.4.3 GMM criterion function . . . 34

1 Notation and Definitions

First off, denote a some scalar. For now, we will assume that $a \in \mathbb{R}$.

Though vectors refer to a broader class of objects, we focus our attention on geometric vectors, i.e., an object with magnitude and direction. We also presupose the notions of a k-dimensional euclidean space, and a the cartesian coordinate system as the method of uniquely determining the position of an object in said space, by means of displacement on each of the (muutually orthogonal) axis.

We denote a vector with a boldface lowercase letter, e.g., $\boldsymbol{a}$. A vector $\boldsymbol{a} \in \mathbb{R}^k$ is a $k \times 1$ list of numbers with a geometrical interpretation of a shift from the origin to the point in a k-dimensional space given by the coordinates of vector $\boldsymbol{a}$.

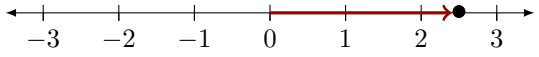
$$\boldsymbol{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{bmatrix}$$

Though vectors can be rows or columns, for economics and econometrics, when referring to a vector, we usually mean a column vector. We can represent a row vector with the notation $\boldsymbol{a}'$ or $\boldsymbol{a}^T$.

$$\boldsymbol{a}' = \begin{bmatrix} a_1 & a_2 & \cdots & a_k \end{bmatrix}$$

We call $\boldsymbol{a}'$ $\boldsymbol{a}$ transposed, and it just means that we flipped the vector $\boldsymbol{a}$ on its side.

Graphically, a vector is a shift (think of a walk) from the origin to some point in a k-dimensional (euclidean) space, and it's depicted as an arrow starting from the origin of the euclidean space in the direction and the magnitude given by the vector (the coordinates being represented by the vector). For example, a 1×1 vector is an arrow on the number line pointing from 0 to $\boldsymbol{a}$ (e.g., $\boldsymbol{a} = (2.5)$):



A 2×1 vector is a point on a plane ($\boldsymbol{a}' = (2.5, 3.5)$)[1], and a 3×1 vector is a point in space ($\boldsymbol{a}' = (2, 3, 1)$).

And so on. All the lines get messy, so onwards consider only the point as the representation of the vector in space. The $l-th$ element number of the vector $\boldsymbol{a}$, a_l tells us how far the point is along the $l-th$ axis. For example, the vector $\boldsymbol{a} = (2.5, 3.5)'$ is 2.5 units along the x_1 axis, and 3.5 units from the x_2 axis.

[1] Though notationally vectors are not sepparated by commas, in-text row vectors are separated by commas to avoid ambiguity.

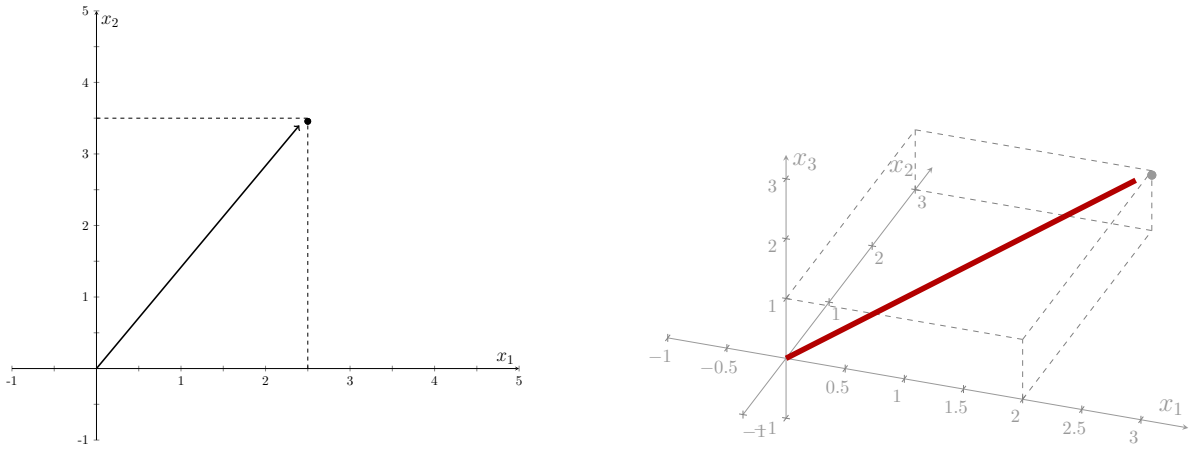


Figure 1: A 2×1 vector and a 3×1 vector

A matrix $\boldsymbol{A}$ is a $k \times r$ array of numbers, where k denotes the number of rows, and r the number of columns. We denote a matrix with a boldface uppercase letter, e.g., $\boldsymbol{A}$.

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1r} \\ a_{21} & a_{22} & \cdots & a_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kr} \end{bmatrix}$$

Matrices are used to represent linear transformations. Graphically, a matrix is a transformation of the space, and hence, of every possible vector on that space. I'll come back to this once we cover matrix operations.

We can consider vectors to be a matrix of multiple rows but only one column (column vector), or only one row and multiple columns (row vector). A 1×1 matrix is just a scalar, and scalars enter any operation as would any regular multiplication.

1.1 Types of matrices

A **square matrix** is a matrix with the same number of rows and columns:

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kk} \end{bmatrix}$$

Otherwise, it is called a **rectangular matrix**.

A **triangular matrix** is a square matrix where all the elements above or below the diagonal are zero. If all the elements above the diagonal are zero, it's called a lower triangular matrix, denoted $\boldsymbol{L}$, and if all the elements below the diagonal are zero, it's called an upper triangular matrix, denoted $\boldsymbol{U}$.

$$\boldsymbol{U} = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1r} \\ 0 & u_{22} & \cdots & u_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & u_{rr} \end{bmatrix} \qquad \boldsymbol{L} = \begin{bmatrix} l_{11} & 0 & \cdots & 0 \\ l_{21} & l_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{r1} & l_{r2} & \cdots & l_{rr} \end{bmatrix}$$

A **diagonal matrix** is a matrix with elements only on the diagonal, and zeroes everywhere else. Hence, it's

considered a matrix that is both upper triangular and lower triangular. It is denoted usually $\boldsymbol{D}$. E.g.:

$$\boldsymbol{D} = \begin{bmatrix} d_{11} & 0 & \cdots & 0 \\ 0 & d_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_{kk} \end{bmatrix}$$

We can also define a **rectangular diagonal matrix** as a matrix that has elements only on the diagonal starting from the $1,1$ element, and zeroes everywhere else. E.g., for a $r \times k$ matrix, $r \neq k$, we could have either of:

$$\boldsymbol{D} = \begin{bmatrix} d_{11} & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & d_{22} & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_{rr} & 0 & \cdots & 0 \end{bmatrix} \qquad \boldsymbol{D} = \begin{bmatrix} d_{11} & 0 & \cdots & 0 \\ 0 & d_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_{kk} \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

The **identity matrix** is a special case of a square diagonal matrix with 1's on the main diagonal. It is denoted as $\boldsymbol{I}_k$, where k denotes the dimension of the matrix. When all the elements of the diagonal are constant, it is denoted a scalar matrix $\boldsymbol{C} = c\boldsymbol{I}_k$

$$\boldsymbol{I}_k = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \qquad \boldsymbol{C} = \begin{bmatrix} c & 0 & \cdots & 0 \\ 0 & c & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & c \end{bmatrix}$$

The **leading principal submatrix of order** r of a $k \times k$ square matrix $\boldsymbol{A}$, denoted $\boldsymbol{A}_r$, is the $r \times r$ matrix formed by the first r rows and columns of $\boldsymbol{A}$:

$$\boldsymbol{A} = \begin{bmatrix} a_1 1 & a_{12} & \cdots & a_{1r} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2r} & \cdots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{r1} & a_{r2} & \cdots & a_{rr} & \cdots & a_{rk} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kr} & \cdots & a_{kk} \end{bmatrix} \qquad \boldsymbol{A}_r = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1r} \\ a_{21} & a_{22} & \cdots & a_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ a_{r1} & a_{r2} & \cdots & a_{rr} \end{bmatrix}$$

1.2 Transpose and Conjugate Transpose

The transpose of a matrix $\boldsymbol{A}$, denoted $\boldsymbol{A}^{'}$ is defined as:

$$A^{'} = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{k1} \\ a_{12} & a_{22} & \cdots & a_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1r} & a_{2r} & \cdots & a_{kr} \end{bmatrix}$$

We rotated the matrix on its diagonal, and the dimensions changed from $k \times r$ to $r \times k$.

Properties of the transpose:[a]

1. $(\boldsymbol{A}^{'})^{'} = \boldsymbol{A}$
2. $(\boldsymbol{A} + \boldsymbol{B})^{'} = \boldsymbol{A}^{'} + \boldsymbol{B}^{'}$

3. $(c\boldsymbol{A})' = c\boldsymbol{A}'$
4. $(AB)' = B'A'$
5. $\det(\boldsymbol{A}) = \det(\boldsymbol{A}')$
6. $\boldsymbol{A}'\boldsymbol{A}$ is a symmetric positive semidefinite matrix.
7. If the matrix is invertible, $(A')^{-1} = (A^{-1})'$

[a] Some definitions will show up later

If $\boldsymbol{A}$ is a square $k \times k$ matrix, and $a_{ij} = a_{ji}$ for all $i \neq j$, we say that the matrix is **symmetric**. Symmetric matrices have the property that $\boldsymbol{A} = \boldsymbol{A}'$.

Though we'll be dealing almost exclusively with matrices in a real domain, it's worth introducing matrices in the complex domain as those whose entries have a real part and an imaginary part. I.e., matrix $\boldsymbol{A}$ is complex if it's of the form:

$$\boldsymbol{A} = \begin{bmatrix} a_{11} + ic_{11} & a_{12} + ic_{12} & \cdots & a_{1r} + ic_{1r} \\ a_{21} + ic_{21} & a_{22} + ic_{22} & \cdots & a_{2r} + ic_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1} + ic_{k1} & a_{k2} + ic_{k2} & \cdots & a_{kr} + ic_{kr} \end{bmatrix}$$

The **complex conjugate** of a complex number is the number with the same real part, but with an opposite imaginary part. E.g., the complex conjugate of $a + ic$ is $a - ic$. The **conjugate transpose** of matrix $\boldsymbol{A}$, denoted $\boldsymbol{A}^*$ is the matrix obtained by taking the complex conjugate of each of its entries, and then transposing the matrix:

$$\boldsymbol{A}^* = \begin{bmatrix} a_{11} - ic_{ii} & a_{21} - ic_{21} & \cdots & a_{k1} - ic_{k1} \\ a_{12} - ic_{12} & a_{22} - ic_{22} & \cdots & a_{k2} - ic_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1r} - ic_{1r} & a_{2r} - ic_{2r} & \cdots & a_{kr} - ic_{kr} \end{bmatrix}$$

Matrix $\boldsymbol{A}$ is said to be **Hermitian** if it is equal to its conjugate transpose: $\boldsymbol{A} = \boldsymbol{A}^*$. From this, it should be obvious that Hermitian matrices are square, and with only real entries on its diagonal. Further, if a real matrix is symmetric, it is Hermitian.. Though we don't deal much with complex matrices, several theorems that are proven for Hermitian matrices are applicable to any symmetric matrix by extension.

2 Vectors

2.1 Linear Combinations

Take any two vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^k$ and a scalar $c \in \mathbb{R}$. Vector addition $\boldsymbol{a} + \boldsymbol{b}$ is defined as the sum of the elements of the vectors, whilst scalar multiplication $c\boldsymbol{a}$ is the product of the scalar and each element of the vector:

$$\boldsymbol{a} + \boldsymbol{b} = \begin{bmatrix} a_1 + b_1 \\ a_2 + b_2 \\ \vdots \\ a_k + b_k \end{bmatrix} \qquad c\boldsymbol{a} = \begin{bmatrix} ca_1 \\ ca_2 \\ \vdots \\ ca_k \end{bmatrix}$$

We can represent vector addition as a succession of shifts starting from the origin, and then moving on from each successive point reached by the shift of the respective added vector. Scalar multiplication is analogous, considering addition of identical vectors. Consider $\boldsymbol{a} = (1.5, 2.5)'$ and $\boldsymbol{b} = (2, 1)'$, and $c = 2$. Graphically, $\boldsymbol{a} + \boldsymbol{b} = (3.5, 3.5)'$ (left) and $c\boldsymbol{a} = (3, 5)'$ (right):

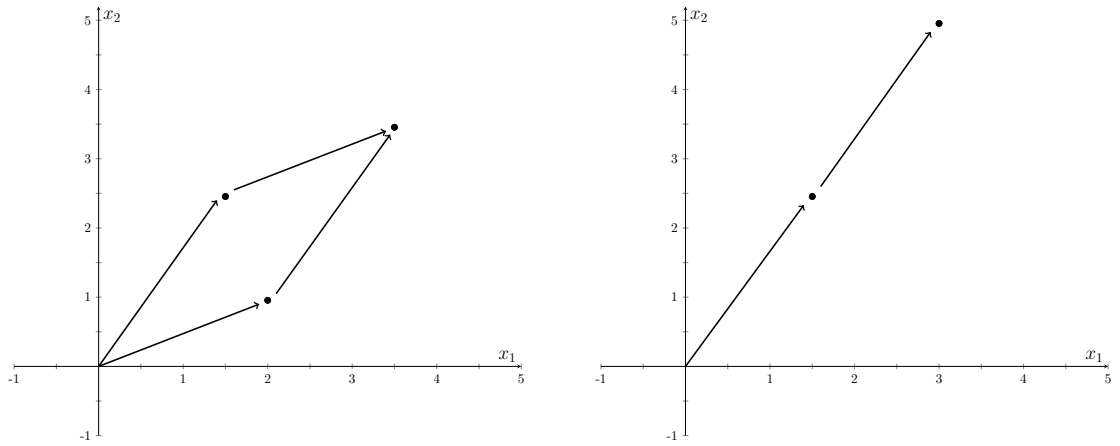


Figure 2: Vector addition and scalar multiplication

Any vector operation using vector addition and scalar multiplication is called a linear combination. Now think of vectors $\hat{\mathbf{i}} = (1,0)' \in \mathbb{R}^2$ and $\hat{\mathbf{j}} = (0,1)' \in \mathbb{R}^2$. These vectors are called the canonical basis of the coordinate system on $\mathbb{R}^2$. Any vector $\boldsymbol{a} = (a_1, a_2)' \in \mathbb{R}^2$ can be written as a linear combination of $\hat{\mathbf{i}}$ and $\hat{\mathbf{j}}$:

$$\boldsymbol{a} = a_1\hat{\mathbf{i}} + a_2\hat{\mathbf{j}} = \begin{bmatrix} a_1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ a_2 \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$$

Any space $\mathbb{R}^k$ has a basis of k linearly independent vectors $\hat{\mathbf{i}}_1, \hat{\mathbf{i}}_2, \ldots, \hat{\mathbf{i}}_k$, and thus any $k \times 1$ vector can be represented as a linear combination of these vectors. Naturally, using the canonical basis vectors, we can represent any $k \times 1$ vector as a linear combination of the base vectors. However, we may chose a different set of basis vectors, and define coordinates in relationship to these new vectors. Consider using $\hat{\mathbf{i}}_1 = (1.5, 2.5)'$ and $\hat{\mathbf{i}}_2 = (2,1)'$ as the basis of $\mathbb{R}^2$. Then, any vector in $\mathbb{R}^2$ can be represented as a linear combination of $\hat{\mathbf{i}}_1$ and $\hat{\mathbf{i}}_2$. E.g., consider vector $(1,4)' = 2\hat{\mathbf{i}}_1 - 1\hat{\mathbf{i}}_2$.

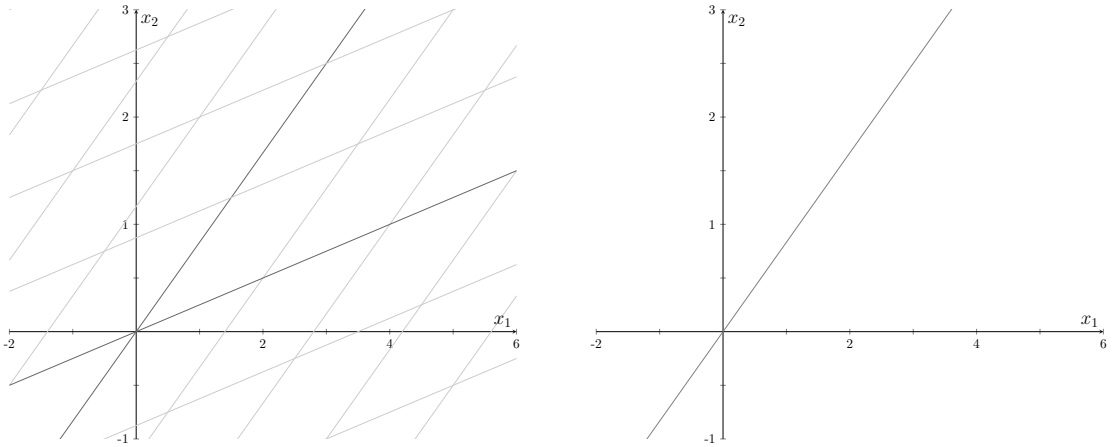


Figure 3: Span of two vectors

Graphically, go back to Figure 2, and note we selected $\hat{\mathbf{i}} = \boldsymbol{a}$ and $\hat{\mathbf{j}} = \boldsymbol{b}$. Go back to Figure 2 on the left panel and consider the rhombus formed by the vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ and their addition. Extend the line representing $\boldsymbol{a}$. Any point on that line is a scalar multiplication of $\boldsymbol{a}$. From any one of those points, start moving in the direction prescribed by $\boldsymbol{b}$, and any point reached is a linear combination of $\boldsymbol{a}$ and $\boldsymbol{b}$. On the left of Figure 3, we depict this by a new grid that is formed by the linear combination of $\hat{\mathbf{i}} = \boldsymbol{a}$ and $\hat{\mathbf{j}} = \boldsymbol{b}$. Since this grid actually covers the totality of $\mathbb{R}^2$, we say that the span of these two vectors equals $\mathbb{R}^2$. Since the span of these two vectors is equal to the dimension in which both of them live in, we say that these vectors are linearly independent.

In this case, the space is squished into a grid where the x-axis is tilted upwards and strecthed, and the y-axis is tilted to the right and stretched. Different vectors used as basis vectors will result in different transformation of

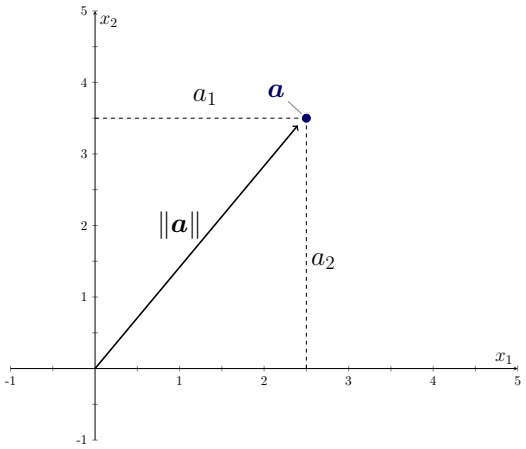


Figure 4: Euclidean norm of a vector

the vector space.

Now consider $\hat{\mathbf{i}} = (1.5, 2.5)$, and $\hat{\mathbf{j}} = (3, 5)$. Note that if we follow the same procedure as above, we up only with a straight line, instead of a grid (area) as depicted on the right of Figure 3. This is because one vector can be obtained through a linera combination of the other. In this case, the span of these two vectors is only $\mathbb{R}$, and we say that these vectors are linearly dependent.

2.2 Dot Product and euclidean Norm

Take two products in the same dimension, $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^k$. The **dot product** of these two vectors, $\boldsymbol{a} \cdot \boldsymbol{b}$ is defined as the sum of the products of the elements of the vectors:

$$\boldsymbol{a} \cdot \boldsymbol{b} = a_1 b_1 + a_2 b_2 + \cdots + a_k b_k$$

Note that the dot product is always a scalar.

The **euclidean norm** of a vector $\boldsymbol{a}$, denoted $\|\boldsymbol{a}\|$, is a measure of its magnitude, and it's defined as the squared root of the sum of the squares of the elements of the vector:

$$\|\boldsymbol{a}\| = \sqrt{a_1^2 + a_2^2 + \cdots + a_k^2} = \sqrt{\boldsymbol{a} \cdot \boldsymbol{a}}$$

Note that when $k = 1$, the euclidean magnitude of a vector is the absolute value of the element of the vector. For $k = 2$, the euclidean magnitude of a vector takes the form of the Pythagorean theorem, and thus, the euclidean magnitude of a vector is the has the interpretation of the length of the vector in the geometric space. For $k > 2$, the euclidean magnitude of a vector takes exactly the same interpretation: the distance from the origin to the point in space represented by the vector.

I'll come back to the norms on Matrices.

Take any vectors $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^k$, and any scalar $c \in \mathbb{R}$, then the following properties hold:
Properties of the dot product

1. $\boldsymbol{x} \cdot \boldsymbol{y} = \boldsymbol{y} \cdot \boldsymbol{x}$ (commutative property).
2. $\boldsymbol{x} \cdot (\boldsymbol{y} + \boldsymbol{z}) = \boldsymbol{x} \cdot \boldsymbol{y} + \boldsymbol{x} \cdot \boldsymbol{z}$ (distributive property).
3. $c(\boldsymbol{x} \cdot \boldsymbol{y}) = (c\boldsymbol{x}) \cdot \boldsymbol{y}$ (associative property).
4. $\boldsymbol{x} \cdot \boldsymbol{x} = \|\boldsymbol{x}\|^2$

Properties of the euclidean norm

1. $\|\boldsymbol{x}\| \geq 0$, and $\|\boldsymbol{x}\| = 0$ if and only if $\boldsymbol{x} = \mathbf{0}$.
2. $\|c\boldsymbol{x}\| = |c|\|\boldsymbol{x}\|$.
3. **Triangle Inequality**: $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$.

2.3 Orthogonality and Orthonormality

Vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ are said to be orthogonal, denoted $\boldsymbol{a} \perp \boldsymbol{b}$, if the angle between the lines that they form is 90° (or 270°), which occurs if $\boldsymbol{a} \cdot \boldsymbol{b} = 0$. The dot product is, interestingly so, associated to the angle between the respective vectors. Consider vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^k$, as shown in Figure 5

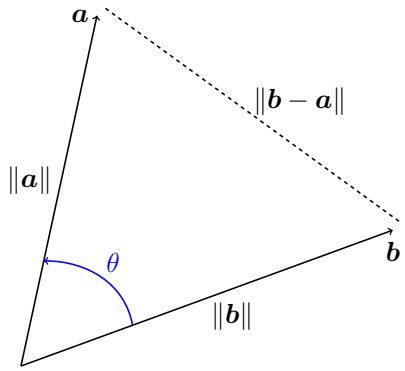


Figure 5: Dot Product and angle between two vectors

The law of cosines states that, for a triangle, the square of the length of the opposite side of an angle is equal to the sum of the squares of the adjacent sides' length minus twice their product times the cosine of the angle. In Figure 5, the law of cosines translates to:

$$\|\boldsymbol{b} - \boldsymbol{a}\|^2 = \|\boldsymbol{b}\|^2 + \|\boldsymbol{a}\|^2 - 2\|\boldsymbol{a}\|\|\boldsymbol{b}\| \cos(\theta)$$

Now, from the properties of the dot product and the euclidean norm:

$$\|\boldsymbol{b} - \boldsymbol{a}\|^2 = (\boldsymbol{b} - \boldsymbol{a}) \cdot (\boldsymbol{b} - \boldsymbol{a}) = \boldsymbol{b} \cdot \boldsymbol{b} - 2\boldsymbol{a} \cdot \boldsymbol{b} + \boldsymbol{a} \cdot \boldsymbol{a} = \|\boldsymbol{b}\|^2 - 2\boldsymbol{b} \cdot \boldsymbol{a} + \|\boldsymbol{a}\|^2$$

Back into the law of cosines:

$$\begin{aligned} \|\boldsymbol{b}\|^2 - 2\boldsymbol{b} \cdot \boldsymbol{a} + \|\boldsymbol{a}\|^2 &= \|\boldsymbol{b}\|^2 + \|\boldsymbol{a}\|^2 - 2\|\boldsymbol{a}\|\|\boldsymbol{b}\| \cos(\theta) \\ \implies \qquad \boldsymbol{b} \cdot \boldsymbol{a} &= \|\boldsymbol{b}\|\|\boldsymbol{a}\| \cos(\theta) \end{aligned}$$

Now, from trigonometry, $\cos(90°) = \cos(90°) = 0$. So, by the equation above, two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ are orthogonal if and only if $\boldsymbol{a} \cdot \boldsymbol{b} = 0$. As a corollary, the dot product $\boldsymbol{a} \cdot 0 = 0$, so the zero vector is always orthogonal to any vector. If the angle formed by two vectors is 0° or 180°, then $\cos(\theta) = 1$ or $\cos(\theta) = -1$, respectively. Hence, the absolute value of the dot product of two vectors equals the product of their respective euclidean norms. In this case, the vectors are said to be **colinear**.

Lastly, two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ are said to be orthonormal if they are orthogonal and have unit length, i.e., $\boldsymbol{a} \cdot \boldsymbol{b} = 0$ and $\|\boldsymbol{a}\| = \|\boldsymbol{b}\| = 1$. In several instances, vectors are normalized to have unit length, and thus, be orthonormal. This is particularly useful when it comes to matrix decompositions.

2.4 Cauchy-Schwarz and Triangle Inequalities

Above we showed that $\boldsymbol{a} \cdot \boldsymbol{b} = \|\boldsymbol{a}\|\|\boldsymbol{b}\| \cos(\theta)$. Take the absolute value:

$$|\boldsymbol{a} \cdot \boldsymbol{b}| = \|\boldsymbol{a}\|\|\boldsymbol{b}\||\cos(\theta)| \leq \|\boldsymbol{a}\|\|\boldsymbol{b}\|$$

where the inequality comes from the fact that $|\cos(\theta)| \in [0, 1]$. This is known as the **Cauchy-Schwarz inequality**.

$$|\boldsymbol{a} \cdot \boldsymbol{b}| \leq \|\boldsymbol{a}\|\|\boldsymbol{b}\|$$

This inequality can be expressed using sums:

$$\sum_{i=1}^{k}(a_i b_i) \leq \left(\sum_{i=1}^{k} a_i^2\right)^{1/2} \left(\sum_{i=1}^{k} b_i^2\right)^{1/2}$$

And sometimes is also found in its squared form. Since $\|\boldsymbol{a}\|^2 = \boldsymbol{a} \cdot \boldsymbol{a}$, we have:

$$\begin{aligned}(\boldsymbol{a} \cdot \boldsymbol{b})^2 \leq &(\boldsymbol{a} \cdot \boldsymbol{a})(\boldsymbol{b} \cdot \boldsymbol{b}) \\ \left(\sum_{i=1}^{k}(a_i b_i)\right)^2 \leq &\left(\sum_{i=1}^{k} a_i^2\right)\left(\sum_{i=1}^{k} b_i^2\right)\end{aligned}$$

A consequence of the Cauchy-Schwarz inequality is the **triangle inequality**:

$$\begin{aligned}\|\boldsymbol{a} + \boldsymbol{b}\|^2 =&(\boldsymbol{a} + \boldsymbol{b}) \cdot (\boldsymbol{a} + \boldsymbol{b}) = (\boldsymbol{a} \cdot \boldsymbol{a}) + 2(\boldsymbol{a} \cdot \boldsymbol{b}) + (\boldsymbol{b} \cdot \boldsymbol{b}) \\ \leq&(\boldsymbol{a} \cdot \boldsymbol{a}) + 2|\boldsymbol{a} \cdot \boldsymbol{b}| + (\boldsymbol{b} \cdot \boldsymbol{b}) \leq (\boldsymbol{a} \cdot \boldsymbol{a}) + 2\|\boldsymbol{a}\|\|\boldsymbol{b}\| + (\boldsymbol{b} \cdot \boldsymbol{b}) \\ =&\|\boldsymbol{a}\|^2 + 2\|\boldsymbol{a}\|\|\boldsymbol{b}\| + \|\boldsymbol{b}\|^2 = (\|\boldsymbol{a}\| + \|\boldsymbol{b}\|)^2\end{aligned}$$

where the first inequality comes from a scalar always being less than or equal to its absolute value, and the second results from the Cauchy-Schwartz inequality. Taking the square root from both sides, we arrive to the triangle inequality:

$$\|\boldsymbol{a} + \boldsymbol{b}\| \leq \|\boldsymbol{a}\| + \|\boldsymbol{b}\|$$

3 Matrix algebra

3.1 Matrix-vector multiplication

A linear transformation is a function that maps a vector to another vector. These take the form of:

$$\boldsymbol{A}\boldsymbol{x}$$

where $\boldsymbol{A}$ is $r \times k$ matrix and $\boldsymbol{x}$ is $k \times 1$. Matrix-vector multiplication requires a match between the dimension of the vector, and the number of columns of the matrix. Take $\boldsymbol{A}$ as defined above, and some vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_k)'$. Then, $\boldsymbol{A}\boldsymbol{x}$ is defined as:

$$\boldsymbol{A}'x = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{r1} & a_{r2} & \cdots & a_{rk} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1k}x_k \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{k2}x_k \\ \vdots \\ a_{r1}x_1 + a_{r2}x_2 + \cdots + a_{rk}x_k \end{bmatrix}$$

Another way to look at this is defining:

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{r1} & a_{r2} & \cdots & a_{rk} \end{bmatrix} \equiv \begin{bmatrix} a_1^{'} \\ a_2^{'} \\ \vdots \\ a_r^{'} \end{bmatrix}$$

Then,

$$\boldsymbol{Ax} = \begin{bmatrix} a_1^{'} \cdot x \\ a_2^{'} \cdot x \\ \vdots \\ a_r^{'} \cdot x \end{bmatrix}$$

in other words, the vector that results from the dot product of x and each row of $\boldsymbol{A}^{'}$.

Note that we are left with a $r \times 1$ vector. For any matrix-vector multiplication, the dimension match is:

$$\underset{r \times k}{\boldsymbol{A}^{'}} \underset{k \times 1}{x}$$

In multiplication, the inner dimensions must be equal (in this case, k and k), and the outer dimensions indicate the dimension of the resulting vector (in this case, $r \times 1$). A geometric interpretation of a linear transformation is a transformation of the space, and hence, of every possible vector on that space. To see this, consider the Identity matrix, denoted $\boldsymbol{I_k}$, as defined previously. Also, take any vector $\boldsymbol{x}$. The result from vector multiplication is:

$$\boldsymbol{I_k x} \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix} = \begin{bmatrix} x_1 + 0 + \cdots + 0 \\ 0 + x_2 + \cdot + 0 \\ \vdots \\ 0 + 0 + \cdots + x_k \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix} = \boldsymbol{x}$$

The identity matrix is akin to multiplying any scalar times 1: it leaves the multiplicand unchanged. Consider the base vectors of a k-dimensional space to be $\hat{\mathbf{i}}_1 = (1, 0, \cdots, 0)^{'}$, $\hat{\mathbf{i}}_2 = (0, 1, \cdots, 0)^{'}$, and so on, until $\hat{\mathbf{i}}_k = (0, 0, \cdots, 1)^{'}$. These are known as the canonical basis vectors of the coordinate system. Rewrite the identity matrix as: the identity matrix as:

$$\boldsymbol{I_k} = \begin{bmatrix} \hat{\mathbf{i}}_1 & \hat{\mathbf{i}}_2 & \cdots & \hat{\mathbf{i}}_k \end{bmatrix}$$

Then $\boldsymbol{I_k x}$ **projects** a vector x into the system of coordinates defined by $\boldsymbol{I_k}$, i.e., projects a vector into the canonical system of cooordinates. However, as we saw above, we can take any set of vectors as the basis of our coordinate system. Consider the basis of the coordinate system to be given by the vectors $\hat{\boldsymbol{a}}_1 = (a_{11}, a_{12}, \cdots, a_{1r})^{'}$, $\hat{\boldsymbol{a}}_2 = (a_{21}, a_{22}, \cdots, a_{2r})^{'}$, and so on, until $\hat{\boldsymbol{a}}_k = (a_{k1}, a_{k2}, \cdots, a_{kr})^{'}$. and consider:

$$\boldsymbol{A} = \begin{bmatrix} \hat{\boldsymbol{a}}_1 & \hat{\boldsymbol{a}}_2 & \cdots & \hat{\boldsymbol{a}}_k \end{bmatrix}$$

Then matrix $\boldsymbol{A}$ projects the vector $\boldsymbol{x}$ into the system of coordinates defined by the basis $\hat{\boldsymbol{a}}_1, \hat{\boldsymbol{a}}_2, \cdots, \hat{\boldsymbol{a}}_k$. Consider the example we gave above where the basis vectors were $\hat{\mathbf{i}}_1 = (1.5, 2.5)^{'}$ and $\hat{\mathbf{i}}_2 = (2, 1)^{'}$. Then, the matrix $\boldsymbol{A}^{'}$ is:

$$\boldsymbol{A} = \begin{bmatrix} 1.5 & 2 \\ 2.5 & 1 \end{bmatrix}$$

Now take vector $\boldsymbol{x} = (2, -1)^{'}$. Then,

$$\boldsymbol{Ax} = \begin{bmatrix} 1.5 & 2 \\ 2.5 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ -1 \end{bmatrix} = \begin{bmatrix} 1.5 \cdot 2 + 2 \cdot (-1) \\ 2.5 \cdot 2 + 1 \cdot (-1) \end{bmatrix} = \begin{bmatrix} 1 \\ 4 \end{bmatrix}$$

So, point $(2, -1)^{'}$ in the standard coordinate system is projected into point $(1, 4)^{'}$ in the space transformed by matrix $\boldsymbol{A}$. To see this graphically, also consider the points of reference, the unit point on the x-axis $(1, 0)^{'}$ and the unit point on the y-axis $(0, 1)^{'}$. When we apply the transformation from $\boldsymbol{A}$, these points end up in:

$$\boldsymbol{A}\begin{bmatrix}1\\0\end{bmatrix} = \begin{bmatrix}1.5\\2.5\end{bmatrix} = \hat{\mathbf{i}}_1 \qquad \boldsymbol{A}\begin{bmatrix}0\\1\end{bmatrix} = \begin{bmatrix}2\\1\end{bmatrix} = \hat{\mathbf{i}}_2$$

Now look to Figure 6, which shows how these three points are shifted. First, the unit vector along the x-axis goes way up and to the right, whilst the unit vector along the y-axis just gets shifted to the right (so spacially, there is kind of a rotation of the space). As you can see, vector $y = \boldsymbol{A}\boldsymbol{x}$ results from going two units along $\hat{\mathbf{i}}_1$, and minus one unit along $\hat{\mathbf{i}}_2$.

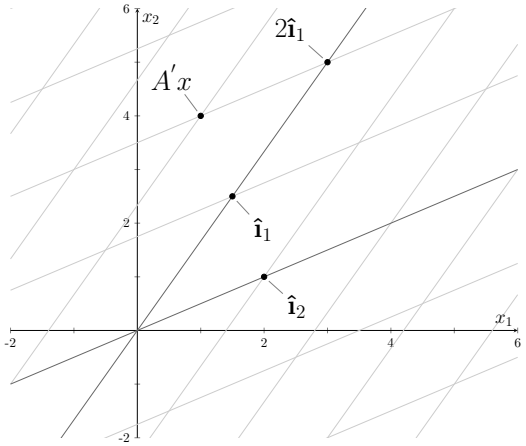


Figure 6: Transformation by Matrix A

The previous examples worked very well with square matrices. Little changes with matrix-vector multiplication with rectangular matrices. The main difference is that the matrix transforms the vector into a space with a different number of dimensions, which might be higher, or lower. E.g.:

$$\begin{bmatrix}1.5 & 2\\2.5 & 1\\1 & 1\end{bmatrix}\begin{bmatrix}2\\-1\end{bmatrix} = \begin{bmatrix}1\\4\\1\end{bmatrix} \qquad \begin{bmatrix}3.5 & 2\end{bmatrix}\begin{bmatrix}2\\-1\end{bmatrix} = 5.$$

The geometric interpretation changes. Whereas matrices with less columns than k rows result in a collapse of dimensions, matrices with more than k columns move the vectors to a higher dimensional space, though will not span the totality of that new space.

3.1.1 Vector-vector mutliplication, dot product and orthogonality (again)

Vectors can be multiplied between other vectors as a special case of matrix multiplication (a $1 \times k$ matrix multiplying a vector).

Take two $k \times 1$ vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. Note that the dot product between them can be written as matrix multiplication:

$$\boldsymbol{x} \cdot \boldsymbol{y} = \boldsymbol{x}^{'}\boldsymbol{y} = \boldsymbol{y}^{'}\boldsymbol{x}$$

This implies that the dot product has the interpretation of a maping of a vector onto a 1-dimensional space. In particular, for $\boldsymbol{x}^{'}\boldsymbol{y}$, think of $\boldsymbol{y} = (1, 1, \cdots, 1)^{'} = \hat{\mathbf{i}}_1 + \hat{\mathbf{i}}_2 + \cdots + \hat{\mathbf{i}}_k$, where $\hat{\mathbf{i}}_i$ is the i-th unit vector (canonical basis vectors for the standard coordinate system). Then, $\boldsymbol{x}$ just states where each of the canonical vectors of the $\mathbb{R}^k$ space end up in the number line. For example, if $\boldsymbol{x} = (3, 2, -1)^{'}$, then:

Thus, the dot product, as a projection, tells us overall, where a vector $\boldsymbol{y}$ lands on the unit line when projecting onto the vector $\boldsymbol{x}$. Interestingly enough, however, is that it is the same spot where $\boldsymbol{x}$ lands when projecting onto vector $\boldsymbol{y}$ giving some nice symmetry that may seem unintuitive at first. This is depicted in Figure 7 below.

Take for instance, vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^2$ as on the left of Figure 8 and project vector b onto vector a (i.e., $\boldsymbol{a}^{'}\boldsymbol{b}$). This is represented as if casting a shadow of vector $\boldsymbol{b}$ onto vector $\boldsymbol{a}$ (note the dashed line is orthogonal to vector $\boldsymbol{a}$), denoted $\boldsymbol{p}$. The dot product is the result of multiplying the norm of the shadow, $\|\boldsymbol{p}\|$ times the norm of $\boldsymbol{a}$, $\|\boldsymbol{a}\|$.

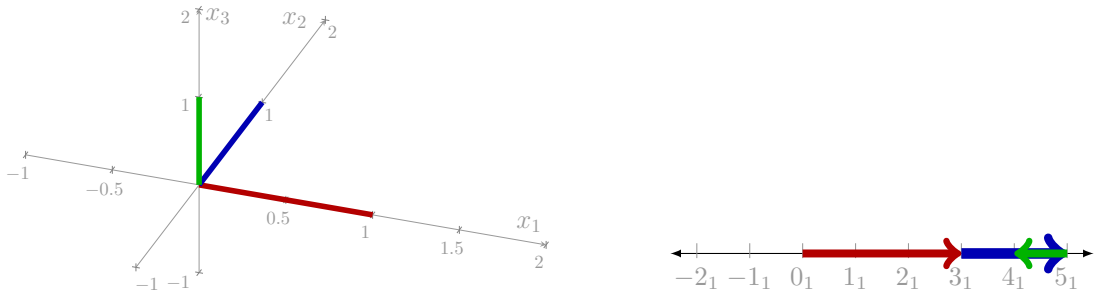


Figure 7: Dot product as a projection onto the unit line

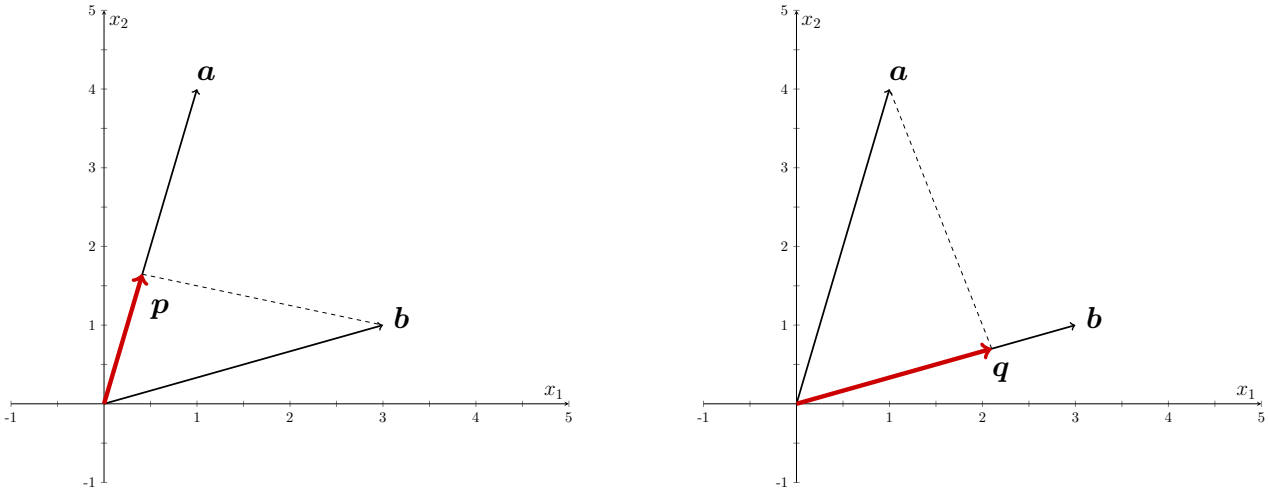


Figure 8: Conmutativity of projecting one vector onto another.

What is interesting, is that the resulting number is the same as projecting $\boldsymbol{a}$ onto $\boldsymbol{b}$ ($\boldsymbol{b}'\boldsymbol{a}$), call this vector $\boldsymbol{q}$, and then multiplying the lengths of both vectors, i.e., $\|\boldsymbol{a}\|\|\boldsymbol{p}\| = \|\boldsymbol{b}\|\|\boldsymbol{q}\|$. This is depicted on the right of Figuer 8.

A special case is the multiplication of a vector times itself. Take a $k \times 1$ vector $\boldsymbol{x}$. Then:

$$\boldsymbol{x}'\boldsymbol{x} = \sum_{i=1}^{k} x_i^2 = \boldsymbol{x} \cdot \boldsymbol{x} = \|\boldsymbol{x}\|^2$$

Note that:

$$\boldsymbol{x}'\boldsymbol{x} = \boldsymbol{x} \cdot \boldsymbol{x} = \|x\|^2$$

$x'x$ comes up often, e.g., if x is a vector of errors, then $x'x$ is the sum of squared errors, or for the Herfindahl Index of market concentration if x is a vector of the market share of the k firms. Now consider any vector $\boldsymbol{y} \neq 0$ and normalize it by the norm of $\boldsymbol{x}$, $\|\boldsymbol{x}\|$:

$$\tilde{\boldsymbol{y}} = \frac{\boldsymbol{y}}{\|\boldsymbol{x}\|}$$

It can be easily shown, using the Kuhn-Tucker method, that from all the possible vectors $\boldsymbol{y} \neq 0$, $\boldsymbol{x}$ is the one that maximizes $\boldsymbol{x}'\tilde{\boldsymbol{y}}$:

$$\boldsymbol{x} = \arg\max_{\boldsymbol{y} \neq 0} \quad \boldsymbol{x}'\tilde{\boldsymbol{y}} \qquad \text{s.t. } \|\boldsymbol{y}\| = \|\boldsymbol{x}\|$$

$$\implies \boldsymbol{x} = \arg\max_{\boldsymbol{y} \neq 0} \quad \frac{\boldsymbol{x}'\boldsymbol{y}}{\|\boldsymbol{x}\|} \qquad \text{s.t. } \|\boldsymbol{y}\| = \|\boldsymbol{x}\|$$

Since $\boldsymbol{x}'\boldsymbol{x} = \|\boldsymbol{x}\|^2$:

$$\max_{\boldsymbol{y}\neq 0} \frac{\boldsymbol{x}'\boldsymbol{y}}{\|\boldsymbol{x}\|} = \frac{\boldsymbol{x}'\boldsymbol{x}}{\|\boldsymbol{x}\|} = \|\boldsymbol{x}\|$$

Hence, the euclidean norm has a geometric interpretation as the maximum gain that a vector can apply to another vector. This idea is behind the definition we use for matrix norms in section 3.2.6. In a similar manner, the minimal gain that a vector can apply to another vector is zero, when the vector is either the zero vector, or when the vectors are orthogonal (zero dot product). This gives some insight into why orthogonality is important. Mapping a vector into another vector gives an idea of how much information is shared between two vectors. When two vectors are orthogonal, they encode completely different information. This lack of redundancy is desirable in many applications. E.g., in regression analysis, if the regressors are orthogonal, they encode different information, which results in reduced standard errors (recall the dot prouduct comes up in the formula for standard errors), and allows for better inference.

On a more general note, we say that the quadratic form associated to a $k \times k$ matrix $\boldsymbol{A}$ is given by $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$. Note that the quadratic form is a scalar, and often serves to summarize the information of a vector. This form is used to sign a matrix, as will be shown later. In statistics, the quadratic form of a vector of normal random is used to create test statistics which have a chi-squared distribution, such as the Hausman test, or the Wald test.

3.2 Matrix operators and properties

3.2.1 The trace operator

For a square $k \times k$ matrix $\boldsymbol{A}$, the trace operator is defined as the sum of the elements on the main diagonal of $\boldsymbol{A}$:

$$\text{tr}(\boldsymbol{A}) = \sum_{i=1}^{k} a_{ii}$$

Properties of the Trace operator

- If both $\boldsymbol{A}$ and $\boldsymbol{B}$ are $k \times r$ matrices, $\text{tr}(\boldsymbol{A} + \boldsymbol{B}) = \text{tr}(\boldsymbol{A}) + \text{tr}(\boldsymbol{B})$
- For any scalar c, $\text{tr}(c\boldsymbol{A}) = c\text{tr}(\boldsymbol{A})$
- $\text{tr}(\boldsymbol{A}) = \text{tr}(\boldsymbol{A}')$
- If both $\boldsymbol{A}$ and $\boldsymbol{B}$ are $k \times k$ matrices, $\text{tr}(\boldsymbol{A}\boldsymbol{B}) = \text{tr}(\boldsymbol{B}\boldsymbol{A})$.
- More generally, for $k \times r$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, $\text{tr}(\boldsymbol{A}'\boldsymbol{B}) = \text{tr}(\boldsymbol{B}'\boldsymbol{A}) = \text{tr}(\boldsymbol{A}\boldsymbol{B}') = \text{tr}(\boldsymbol{B}\boldsymbol{A}')$.
- For $k \times 1$ vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, $\text{tr}(\boldsymbol{a}'\boldsymbol{b}) = \boldsymbol{a}'\boldsymbol{b}$

3.2.2 Span, Rank, Kernel and Nullity

First, the **span** of a matrix $\boldsymbol{A}$, denoted $\text{Sp}(\boldsymbol{A})$ is the span of the vectors that make it up, i.e., the set of all points that may result from any linear combination of the column vectors that make it up. If a function $f(x) = \boldsymbol{A}\boldsymbol{x}$, then the range of f is the span of $\boldsymbol{A}$.

Consider a square matrix $\boldsymbol{A}$ given by:

$$\boldsymbol{A} = \begin{bmatrix} 3 & 6 & 5 \\ 7 & 1 & 3 \\ 5 & 8 & 7 \end{bmatrix}$$

The span of this matrix is a $2-$dimansional plane. Note that the first 2 columns are linearly independent, but the third column can be obtained as a linear combination of the first two (multiplying the first one times $\frac{1}{3}$, and the

second column by $\frac{2}{3}$, we get the third column). Figure 9 shows partially the space created by the first two columns of matrix $\boldsymbol{A}$, with the two vectors being the thick black lines on the edges of the plane. The third line in the middle shows the third column, which also lies on the same plane, since it's a linear combination of the first two columns.

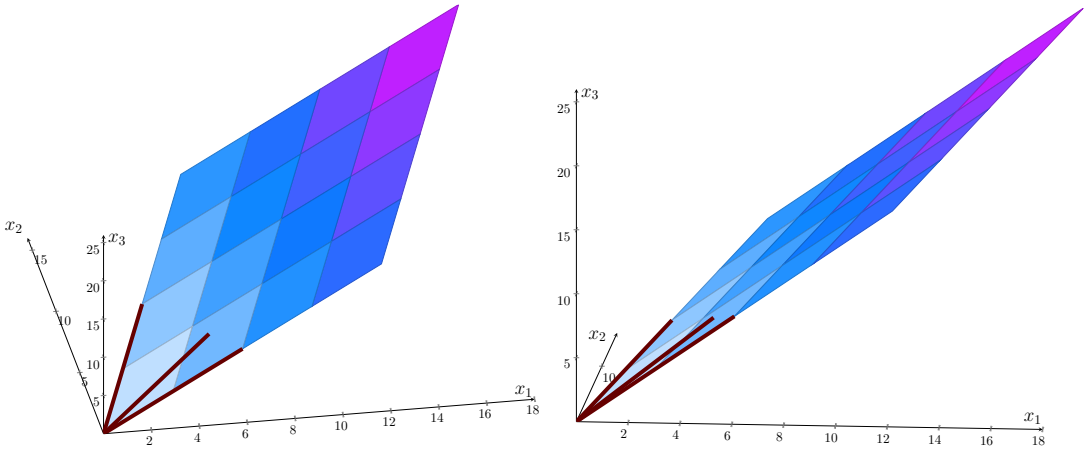


Figure 9: Span of a linearly dependent matrix

Note that when we rotate the point of view, the third line remains on the plane. If we instead consider:

$$\boldsymbol{B} = \begin{bmatrix} 3 & 6 & 5 \\ 7 & 1 & 3 \\ 5 & 8 & 2 \end{bmatrix}$$

the last column now is linearly independent. Since we can create infinite linear combinations of the vectors using this linearly independent third column, moving it around results in a span that's equal to the entire $\mathbb{R}^3$ space. This is shown in Figure 10. The plane is the same as in figure 9, since both matrices share the same first two columns. However, the third column of $\boldsymbol{B}$ is now linearly independent. The vector of this third column now lies outside the plane. We can use this third vector to move about a third dimension, allowing for the span of $\boldsymbol{B}$ to be the entire $\mathbb{R}^3$.

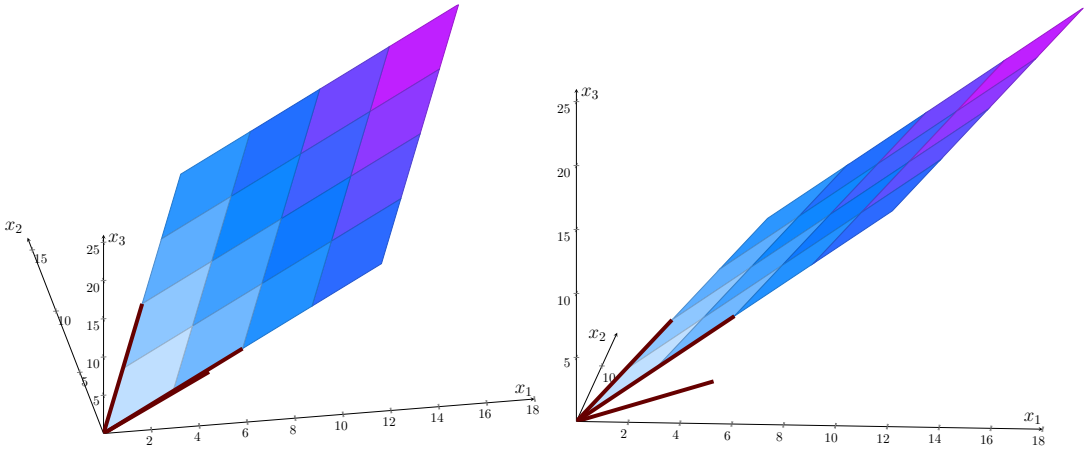


Figure 10: Span of a linearly independent matrix

The rank of a matrix $\boldsymbol{A}$, denoted $\text{rank}(\boldsymbol{A})$, is the dimension of it's span. If the span of $\boldsymbol{A}$ is $\mathbb{R}^k$, then $\text{rank}(\boldsymbol{A}) = k$.

The kernel of a matrix A, $\ker(\boldsymbol{A})$, is the set of all vectors that are mapped to the zero vector by the matrix $\boldsymbol{A}$. I.e.:

$$\ker(\boldsymbol{A}) = \{\boldsymbol{x} \in \mathbb{R}^k : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}\}$$

The nullity of a matrix $\boldsymbol{A}$, denoted $\text{nul}(\boldsymbol{A})$, is the dimension of the kernel of $\boldsymbol{A}$.

Think back to the example we gave above, where we had a matrix $\boldsymbol{A}$ with a linearly dependent third column. Intuitively, what happens is that several vectors in $\mathbb{R}^3$ get squished into the plane on figure 9. The kernel of $\boldsymbol{A}$ is the set of all vectors that get squished into the origin. Take, for example, vector $\boldsymbol{x} = (2, 1, 3)$. When transformed by matrix A, this vector ends up in:

$$\boldsymbol{A} = \begin{bmatrix} 3 & 6 & 5 \\ 7 & 1 & 3 \\ 5 & 8 & 7 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 27 \\ 24 \\ 39 \end{bmatrix}$$

Vector $\boldsymbol{x}$ is not in the kernel of $\boldsymbol{A}$, since it doesn't get mapped to the origin. However, vector $\boldsymbol{y} = (1, 2, -3)$ is in the kernel of $\boldsymbol{A}$, since:

$$\boldsymbol{A} = \begin{bmatrix} 3 & 6 & 5 \\ 7 & 1 & 3 \\ 5 & 8 & 7 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ -3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

So transforming vector $\boldsymbol{y}$ by matrix $\boldsymbol{A}$ results in the zero vector. How about matrix $\boldsymbol{B}$ from the example above? Note that this matrix is linearly independent, and thus maps every single vector in $\mathbb{R}^3$ into another unique vector in $\mathbb{R}^3$. Thus the only vector that gets mapped to the zero vector is the zero vector. In general, if a matrix is linearly independent, then $\ker(\boldsymbol{A}) = \{\boldsymbol{0}\}$, and $\text{nul}(\boldsymbol{A}) = 0$.

3.2.3 The Determinant

The determinant of a matrix is a scalar that can be computed for any square matrix. The deteriminant of a 2×2 matrix is given by:

$$\det(\boldsymbol{A}) = \det\left(\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}\right) = a_{11}a_{22} - a_{12}a_{21}$$

i.e., the product of the diagonal elements minus the product of the off-diagonal elements. Now define $\boldsymbol{A}^{(\neg j, \neg l)}$ as the matrix resulting from removing the $j-$th row and the $l-$th column. The determinant of a $k \times k$ matrix is defined recursively by:

$$\det(\boldsymbol{A}) = \sum_{i=1}^{k} (-1)^{1+i} a_{1i} \det(\boldsymbol{A}^{(\neg 1, \neg i)})$$

As an example for a 3×3 matrix:

$$\det\left(\begin{bmatrix} 3 & 6 & 5 \\ 7 & 1 & 3 \\ 5 & 8 & 2 \end{bmatrix}\right) = 3\det\left(\begin{bmatrix} 1 & 3 \\ 8 & 2 \end{bmatrix}\right) - 6\det\left(\begin{bmatrix} 7 & 3 \\ 5 & 2 \end{bmatrix}\right) + 5\det\left(\begin{bmatrix} 7 & 1 \\ 5 & 8 \end{bmatrix}\right) = 3(-22) - 6(-1) + 5(51) = 195$$

Though the process is rather simple for 2×2 and 3×3 matrices, it gets increasingly complicated to estimate by hand for larger matrices. Just let a computer do it, in case you need it. Manually, the algorithm actually allows you to take any row or any column, and expand the determinant by that row or column, which might make things easier, but then again, if it's complicated enough, just let a computer handle it.

Geometrically, the determinant of of a 2×2 matrix is the area of the parallelogram defined by the two vectors that make up the columns of the matrix. The determinant of a 3×3 matrix is the volume of the parallelepiped defined by the three vectors that make up the columns of the matrix. The determinant of matrices of larger dimensions have a similar interpretation. For now, take the 2×2 matrix and draw the vectors that make up the columns of the matrix, and draw the parallelogram by joining the endopoints to the sum of the vectors. The area of this parallelogram is the determinant of the matrix. This is represented in Figure 11. Note, first we can enclose the parallelogram in a rectangle with base $a_{11} + a_{12}$ and height $a_{21} + a_{22}$, so the total area is

$$(a_{11} + a_{12}) \times (a_{21} + a_{22}) = a_{11}a_{21} + a_{11}a_{22} + a_{12}a_{21} + a_{12}a_{22}$$

We want to substract what's not in the parallelogram, which we can break into 4 triangles, 2 pairs with equal areas, which should be relatively easy to calculate as shown in Figure 11.

Hence, the total area is given by:

$$a_{11}a_{21} + a_{11}a_{22} + a_{12}a_{21} + a_{12}a_{22} - (a_{11} + a_{12})a_{21} - (a_{21} + a_{22})a_{12} = a_{11}a_{22} - a_{12}a_{21} = \det(\boldsymbol{A})$$

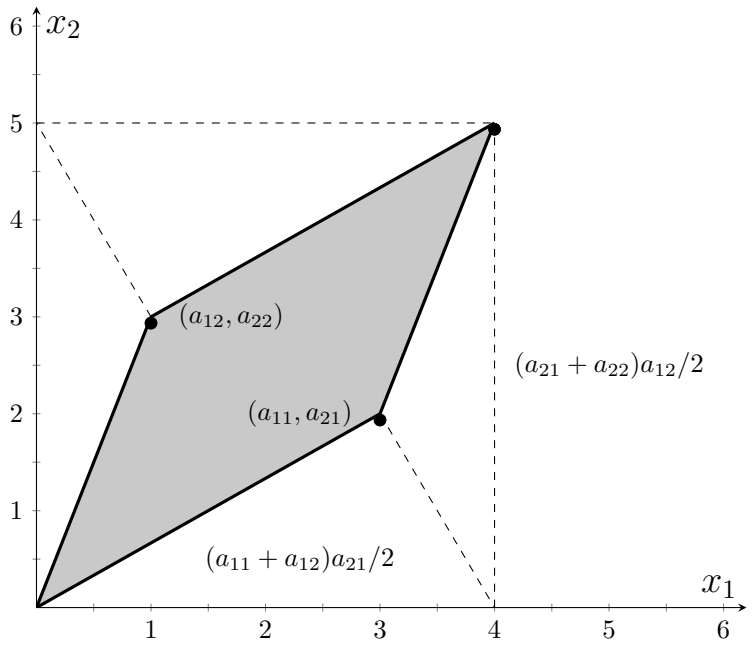


Figure 11: Determinant of a 2×2 matrix

A similar process can be done for higher dimensions to build intuition. When $\text{rank}(\boldsymbol{A}) < k$ for a $k \times k$ matrix, then $\det(\boldsymbol{A}) = 0$. The determinant may take on negative values. This does not alter the geometrical interpretation, but rather just says that the parallelogram or parallelepiped is oriented in the opposite direction (e.g., if there's a 90° rotation).

When we have a linearly dependent matrix, we concluded this results in squuishing a space into a lower dimension. This means that the area of a parallelogram collapses into a line or a point, or the volume of a parallelepiped collapses into a surface, a line or a point. This intuitively should result in a determinant equal to 0:

$$\det\left(\begin{bmatrix}3 & 6 & 5\\ 7 & 1 & 3\\ 5 & 8 & 7\end{bmatrix}\right) = 3\det\left(\begin{bmatrix}1 & 3\\ 8 & 7\end{bmatrix}\right) - 6\det\left(\begin{bmatrix}7 & 3\\ 5 & 7\end{bmatrix}\right) + 5\det\left(\begin{bmatrix}7 & 1\\ 5 & 8\end{bmatrix}\right) = 3(-17) - 6(34) + 5(51) = 0$$

Properties of the Determinant For $k \times k$ matrices $\boldsymbol{A}$, the following properties hold:

1. Homogeneity: $\det(c\boldsymbol{A}) = c^k \det(\boldsymbol{A})$
2. If $\tilde{A}$ results from interchanging two rows or two columns from $\boldsymbol{A}$, then $\det(\tilde{\boldsymbol{A}}) = -\det(\boldsymbol{A})$.
3. If any column or row of $\boldsymbol{A}$ can be represented as a linear combination of other columns or rows, then $\det(\boldsymbol{A}) = 0$.
4. If $\boldsymbol{A}$ is a triangular matrix (i.e., $a_{ij} = 0 \; \forall \; i > j$ or $i < j$), then
$$\det \boldsymbol{A} = \prod_{i=1}^{k} a_{ii}$$
5. If $\boldsymbol{A}$ is invertible, then $\det(\boldsymbol{A}^{-1}) = \det(\boldsymbol{A})^{-1}$.
6. $\det(\boldsymbol{AB}) = \det(\boldsymbol{A})\det(\boldsymbol{B})$.

3.2.4 Eigenvalues and Eigenvectors

Geometrically, we've seen that matrices may be interpreted as transformations of space. Usually, vectors change direction when transformed by a matrix. However, there are some vectors that don't change direction, but only change their magnitude. These vectors are called eigenvectors, and the magnitude by which they change is called their eigenvalue. Formally, a $k \times k$ matrix $\boldsymbol{A}$ is said to have eigenvalue λ and eigenvector $\boldsymbol{v}$ if:

$$\boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{v}, \ \boldsymbol{v} \neq 0$$

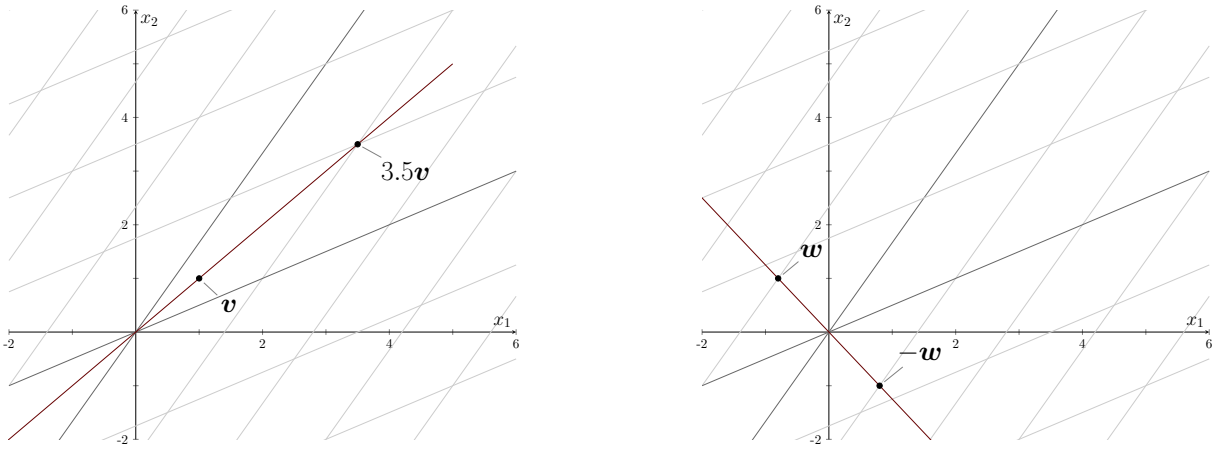


Figure 12: eigenvectors and eigenvalues A

Go back to the example we gave on matrix-vector multiplication, and multiply that matrix times $\boldsymbol{v} = (1, 1)'$, and times $\boldsymbol{w} = (-0.8, 1)'$:

$$\boldsymbol{A}\boldsymbol{v} = \begin{bmatrix} 1.5 & 2 \\ 2.5 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3.5 \\ 3.5 \end{bmatrix} = 3.5\boldsymbol{v} \qquad \boldsymbol{A}\boldsymbol{w} = \begin{bmatrix} 1.5 & 2 \\ 2.5 & 1 \end{bmatrix} \begin{bmatrix} -0.8 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.8 \\ -1 \end{bmatrix} = -\boldsymbol{w}$$

Figure 12 shows the transformation of the space by matrix $\boldsymbol{A}$. Note that though several points are squished and shifted in space, vector $\boldsymbol{v}$ remains in the same direction, but is scaled by a factor of 3.5.

Recall the formula that characterizes the eigenvalues and eigenvectors: $\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v}$. This is equivalent to $(\boldsymbol{A} - \lambda\boldsymbol{I}_k)\boldsymbol{v} = 0$. Since $\boldsymbol{v}$ is not the zero vector, then to find the eigenvalues of $\boldsymbol{A}$, we need to find the values of λ such that

$$\det(\boldsymbol{A} - \lambda\boldsymbol{I}_k) = 0.$$

This is known as the characteristic equation of the square $k \times k$ matrix $\boldsymbol{A}$. $\det(\boldsymbol{A} - \lambda\boldsymbol{I}_k)$ is a polynomial of degree k in λ, and thus has k roots, some of which may be complex (this is a direct result of the fundamental theorem of algebra), and some of which may be repeated roots (multiplicites). These roots are the eigenvalues of $\boldsymbol{A}$. In case there are more than 1 eigenvalues with the same value (multiplicities), we say that the matrix is defective.

Properties of eigenvalues and eigenvectors
Denote $\{\lambda_i, \boldsymbol{v}_i\}_{i=1}^n$ the set of n eigenvalues with corresponding eigenvectors of a matrix $\boldsymbol{A}$. Then:

1. Eigenvectors are defined up to a multiplicative constant, i.e., for $c \neq 0$, if $\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v}$, then $\boldsymbol{A}(c\boldsymbol{v}) = \lambda(c\boldsymbol{v})$. In general, we prefer to work with normalizations of the eigenvectors such that they have a unit length, i.e., $\|\boldsymbol{v}\| = 1$.
2. $\{1/\lambda_i\}_{i=1}^n$ is the set of eigenvalues of $\boldsymbol{A}^{-1}$.
3. $\{\lambda^k\}_{i=1}^n$ is the set of eigenvalues of $\boldsymbol{A}^k$.

4. The determinant of a matrix is the product of its eigenvalues:

$$\det(\boldsymbol{A}) = \prod_{i=1}^{n} \lambda_i$$

5. A matrix is invertible if all of its eigenvalues are non-zero, $\lambda_i \neq 0 \; \forall i$.

6. The trace of a matrix is the sum of its eigenvalues:

$$\text{tr}(\boldsymbol{A}) = \sum_{i=1}^{n} \lambda_i$$

7. If $\boldsymbol{A}$ is symmetric, then all its eigenvectors are orthogonal and all its eigenvalues are real.

8. If $\boldsymbol{A}$ is a square $k \times k$ matrix, then the eigenvalues of $I_k + \boldsymbol{A}$ are $\{1 + \lambda_i\}_{i=1}^{k}$.

As a sidenote, we've been defining the domain of matrix elements over the real plane $\mathbb{R}$. However, nothing stops us from using complex numbers $\mathbb{C}$ instead. When dealing with complex matrices, we can extend a property: a $k \times k$ matrix $\boldsymbol{A}$ has k eigenvalues, counting multiplicities, and k eigenvectors, some of which may be complex. This, again, is a result of applying the fundamental theorem of algebra to the characteristic polynomial.

3.2.5 Signing a matrix

We are interested in extending the concept of positive and negative numbers to matrices. For any real number $a \in \mathbb{R}$, and $x \in \mathbb{R}$, $x \neq 0$, we know that a is weakly positive if $ax^2 \geq 0$ and strictly positive if $ax^2 > 0$. An analogous concept follows for negative numbers. We can extend this concept to square matrices using the quadratic form associated to such a matrix. If for every $\boldsymbol{x} \in \mathbb{R}^k$, $\boldsymbol{x} \neq 0$

1. $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} \geq 0$, we say that $\boldsymbol{A}$ is **positive semidefinite**.
2. $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} > 0$, we say that $\boldsymbol{A}$ is **positive definite**.
3. $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} \leq 0$, we say that $\boldsymbol{A}$ is **negative semidefinite**.
4. $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} < 0$, we say that $\boldsymbol{A}$ is **negative definite**.
5. If neither of the above, we say that $\boldsymbol{A}$ is **indefinite**.

This definition might be hard to apply in practice, so we can use the following theorems:
Theorem: A square matrix $\boldsymbol{A}$ is

- **Positive semidefinite** if and only if all of its eigenvalues are non-negative, $\lambda \geq 0$.
- **Positive definite** if and only if all of its eigenvalues are positive, $\lambda > 0$
- **Negative semidefinite** if and only if all of its eigenvalues are non-positive, $\lambda \leq 0$
- **Negative definite** if and only if all of its eigenvalues are negative, $\lambda < 0$
- **Indefinite** if some of it's eigenvalues are positive, and some are negative.

Now, take a $k \times k$ matrix $\boldsymbol{A}$, and define as its leading principal submatrix of order r, $\boldsymbol{A}_r$, the matrix resulting from keeping the first r rows and columns of $\boldsymbol{A}$. Define the principal minor of order r as $\det(\boldsymbol{A}_r)$.
Theorem: A square matrix $\boldsymbol{A}$ is:

- **Positive semidefinite** if and only if all of its principal minors are non-negative, $\det(\boldsymbol{A}_r) \geq 0 \; \forall \; r$.
- **Positive definite** if and only if all of its principal minors are non-negative, $\det(\boldsymbol{A}_r) > 0 \; \forall \; r$.

- **Negative semidefinite** if and only if the principal minors of order r have the sign $(-1)^r$ or are zero.
- **Negative definite** if and only if the principal minors of order r have the sign $(-1)^r$
- **Indefinite** if it is neither of the above.

As a collorary, a positive definite matrix requires all the elements on its diagonal to be positive, whereas a negative definite matrix requires all the elements on its diagonal to be negative. Similarly, for positive semidefiniteness and negative semidefiniteness, the diagonal elements must be non-negative and non-positive, respectively.

3.2.6 Norm of a matrix, and submultiplicativity (Cauchy-Schwartz)

The norm of a matrix is, just as in the case of the norm of a vector, a measure of its magnitude. Though there are several measurements, we stick with the spectral norm, which is a scalar that measures the maximum gain that the matrix can apply to a vector:

$$\|\boldsymbol{A}\| = \max_{\boldsymbol{x}\neq 0} \frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|}$$

where $\|\boldsymbol{A}\boldsymbol{x}\|$ and $\|\boldsymbol{x}\|$ are the euclidean norms of the vectors $\boldsymbol{A}\boldsymbol{x}$ and $\boldsymbol{x}$. We stick to this measure of magnitude, because it holds a similar interpretation to a vector norm. Though a vector norm is defined as its distance from the origin (multiple applications of the Pythagorean Theorem), in section 3.1.1 we saw that that the euclidean norm of a vector also measures the maximum gain that a vector can apply to any other vector that it projects.

Define λ_{max} as the largest eigenvalue of $\boldsymbol{A}'\boldsymbol{A}$, it can be shown that:

$$\|\boldsymbol{A}\| = \sqrt{\lambda_{max}}$$

For matrix norms, we can also show something similar to the Cauchy-Schwartz inequality for vector dot products:

$$\|\boldsymbol{A}\boldsymbol{B}\| \leq \|\boldsymbol{A}\|\|\boldsymbol{B}\|$$

A simple way to show this is:

$$\begin{aligned}\|\boldsymbol{A}\boldsymbol{B}\| &= \max_{\boldsymbol{x}\neq 0} \frac{\|\boldsymbol{A}\boldsymbol{B}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \max_{\boldsymbol{B}\boldsymbol{x}\neq 0} \frac{\|\boldsymbol{A}\boldsymbol{B}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \max_{\boldsymbol{B}\boldsymbol{x}\neq 0} \frac{\|\boldsymbol{A}\boldsymbol{B}\boldsymbol{x}\|}{\|\boldsymbol{B}\boldsymbol{x}\|}\frac{\|\boldsymbol{B}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \\ &\leq \left(\max_{\boldsymbol{y}\neq 0} \frac{\|\boldsymbol{A}\boldsymbol{y}\|}{\|\boldsymbol{y}\|}\right)\left(\max_{\boldsymbol{x}\neq 0} \frac{\|\boldsymbol{B}\boldsymbol{y}\|}{\|\boldsymbol{y}\|}\right) = \|A\|\|B\|\end{aligned}$$

3.2.7 Matrix-matrix multiplication

Before juping to matrix-matrix multiplication, it should be straightforward that matrices may always be multiplied by a scalar. If c is a scalar, then $c\boldsymbol{A}$ is the matrix that results from multiplying all of its elements time the scalar c.

Going back, a matrix can be interpreted as transformation of space. In the example above, we warped the x-axis by moving scaling it across the x-axis by a factor of 1.5, and across thn y-axis by a factor of 2.5. Similarly, the y-axis was scaled by a factor of 2 across the x-axis, and kept as was across the y-axis. Matrix multiplication may be seen as a way to realize multiple subsequent transformations of the space. Denote $\boldsymbol{A}$ a $k \times r$ matrix, and denote $\boldsymbol{B}$ a $r \times s$ matrix. Matrix multiplication $\boldsymbol{A}\boldsymbol{B}$ is defined as:

$$AB = \begin{bmatrix} \sum_{i=1}^{r} a_{1i}b_{i1} & \sum_{i=1}^{r} a_{1i}b_{i2} & \cdots & \sum_{i=1}^{r} a_{1i}b_{is} \\ \sum_{i=1}^{r} a_{2i}b_{i1} & \sum_{i=1}^{r} a_{2i}b_{i2} & \cdots & \sum_{i=1}^{r} a_{2i}b_{is} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{r} a_{ki}b_{i1} & \sum_{i=1}^{r} a_{ki}b_{i2} & \cdots & \sum_{i=1}^{r} a_{ki}b_{is} \end{bmatrix}$$

As a general algorithm, the $ij-$eth element of the matrix AB is constructed by taking the $i-$th row from matrix $\boldsymbol{A}$ and the the $j-$eth column of matrix $\boldsymbol{B}$, and obtaining the dot product between these two vectors. Define $\boldsymbol{a_i}$ the

vector resulting from extracting the $i-$eth **row** from $\boldsymbol{A}$ (there's k such vectors), and define $\boldsymbol{b_j}$ the vector resulting from extracting the $j-$eth **column** from B (there's s such vectors):

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1r} \\ a_{21} & a_{22} & \cdots & a_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kr} \end{bmatrix} = \begin{bmatrix} \boldsymbol{a}_1^{'} \\ \boldsymbol{a}_2^{'} \\ \vdots \\ \boldsymbol{a}_k^{'} \end{bmatrix} \qquad \boldsymbol{B} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1s} \\ b_{21} & b_{22} & \cdots & b_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ b_{r1} & b_{r2} & \cdots & b_{rs} \end{bmatrix} = \begin{bmatrix} \boldsymbol{b}_1 & \boldsymbol{b}_2 & \cdots \boldsymbol{b}_s \end{bmatrix}$$

Matrix multiplication is then defined as:

$$\underset{k\times r}{\boldsymbol{A}}\ \underset{r\times s}{\boldsymbol{B}} = \begin{bmatrix} \boldsymbol{a}_1^{'} \cdot \boldsymbol{b_1} & \boldsymbol{a}_1^{'} \cdot \boldsymbol{b_2} & \cdots & \boldsymbol{a}_1^{'} \cdot \boldsymbol{b_s} \\ \boldsymbol{a}_2^{'} \cdot \boldsymbol{b_1} & \boldsymbol{a}_2^{'} \cdot \boldsymbol{b_2} & \cdots & \boldsymbol{a}_2^{'} \cdot \boldsymbol{b_s} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{a}_k^{'} \cdot \boldsymbol{b_1} & \boldsymbol{a}_k^{'} \cdot \boldsymbol{b_2} & \cdots & \boldsymbol{a}_k^{'} \cdot \boldsymbol{b_s} \end{bmatrix}$$

Matrix multiplication is ordered. For instance, note that $\boldsymbol{BA}$ is not defined, since the dimmensions don't match. Just for a second assume both $\boldsymbol{A}$ and $\boldsymbol{B}$ are $k \times k$ matrices. Then, **as a general rule**:

$$AB = \begin{bmatrix} \sum_{i=1}^r a_{1i}b_{i1} & \sum_{i=1}^r a_{1i}b_{i2} & \cdots & \sum_{i=1}^r a_{1i}b_{is} \\ \sum_{i=1}^r a_{2i}b_{i1} & \sum_{i=1}^r a_{2i}b_{i2} & \cdots & \sum_{i=1}^r a_{2i}b_{is} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^r a_{ki}b_{i1} & \sum_{i=1}^r a_{ki}b_{i2} & \cdots & \sum_{i=1}^r a_{ki}b_{is} \end{bmatrix} \neq \begin{bmatrix} \sum_{i=1}^r a_{i1}b_{1i} & \sum_{i=1}^r a_{i2}b_{1i} & \cdots & \sum_{i=1}^r a_{ik}b_{1i} \\ \sum_{i=1}^r a_{i1}b_{2i} & \sum_{i=1}^r a_{i2}b_{2i} & \cdots & \sum_{i=1}^r a_{ik}b_{2i} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^r a_{i1}b_{ki} & \sum_{i=1}^r a_{i2}b_{ki} & \cdots & \sum_{i=1}^r a_{ik}b_{ki} \end{bmatrix} = BA$$

There may be exceptions, but generally it's a no go. So matrix multiplication does not maintain the property of commutativity.

Consider again $\boldsymbol{A}$ a $k \times r$ matrix, and denote $\boldsymbol{B}$ a $r \times s$ matrix, and denote $\boldsymbol{x}$ a $s \times 1$ vector. Now consider $\boldsymbol{ABx}$ a matrix-matrix-vector multiplication. By general rule, matrix multiplication is conducted from right to left. So $\boldsymbol{ABx}$ can be decomposed as:

$$\boldsymbol{ABx} = \underbrace{\boldsymbol{A}}_{\text{second transformation}} \underbrace{(\boldsymbol{Bx})}_{\text{first trannsformation}}$$

However, matrix multiplication does maitain the property of associativity:

$$\boldsymbol{ABx} = \boldsymbol{A}(\boldsymbol{Bx}) = (\boldsymbol{AB})x$$

So we can interpret $\boldsymbol{ABx}$ either as transforming $\boldsymbol{x}$ using $\boldsymbol{B}$ and then the resulting vector using $\boldsymbol{A}$, or first obtaining a total transformation matrix $\boldsymbol{AB}$, and then using it to transform $\boldsymbol{x}$.

Properties of Matrix Multiplication

1. Non-commutativity: $\boldsymbol{AB} \neq \boldsymbol{BA}$.
2. Distributivity: $\boldsymbol{A}(\boldsymbol{B} + \boldsymbol{C}) = \boldsymbol{AB} + \boldsymbol{AC}$.
3. Product with a scalar: $c(\boldsymbol{AB}) = (c\boldsymbol{A})\boldsymbol{B} = \boldsymbol{A}(c\boldsymbol{B})$.
4. Transpose: $(\boldsymbol{AB})^{'} = \boldsymbol{B}^{'}\boldsymbol{A}^{'}$.
5. Associativity: $\boldsymbol{ABC} = (\boldsymbol{AB})\boldsymbol{C} = \boldsymbol{A}(\boldsymbol{BC})$.

Vector-vector multiplication can also be seen as a special case of matrix-matrix multiplication. Consider:

$$\boldsymbol{xx}^{'} = \begin{bmatrix} x_1^2 & x_1x_2 & \cdots & x_1x_k \\ x_1x_2 & x_2^2 & \cdots & x_2x_k \\ \vdots & \vdots & \ddots & \vdots \\ x_1x_k & x_2x_k & \cdots & x_kx_k \end{bmatrix}$$

$\boldsymbol{xx}'$, on the other hand, shows how individual elements of the vector move in tandem with other vectors. So if x is a vector of random variables with expected value of zero, then the expected value, $E[xx']$ would be a vector of variances and covariances (recall the expected value is a linear operation, so the expected value of a matrix is the matrix of expected values of each of its components).

The example above is a special case of the outer product of a vector.

3.2.8 The Inverse Matrix

We are interested in a matrix $\boldsymbol{B}$ that reverts the changes in space caused by matrix $\boldsymbol{A}$. I.e., $\boldsymbol{BA} = \boldsymbol{I_k}$. Note this causes vector $\boldsymbol{x}$ to be projected back into the canonical coordinate system if these transformations are applied. If such a matrix exists, it is known as the inverse matrix of $\boldsymbol{A}$, and is denoted as $\boldsymbol{A}^{-1}$. In this case, we also say that $\boldsymbol{A}$ is invertible. For the inverse matrix to exist, we need $\boldsymbol{A}$ to be a square matrix (though some generalizations exist). In the case of a 2×2 matrix, if

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

then matrix A^{-1}

$$A^{-1} = \begin{bmatrix} w & x \\ y & z \end{bmatrix}$$

is such that:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} w & x \\ y & z \end{bmatrix} = \begin{bmatrix} aw + by & ax + bz \\ cw + dy & cx + dz \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

This is just a system of equations with 4 equations and 4 unknowns (w, x, y, z). The solution to this problem, if it exists, looks as:

$$\boldsymbol{A}^{-1} = \begin{bmatrix} \frac{d}{ad-cb} & \frac{-b}{ad-cb} \\ \frac{-c}{ad-cb} & \frac{a}{ad-cb} \end{bmatrix} = \frac{1}{det(A)} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Note this is well defined (and thus the inverse exists) whenever $det(A) \neq 0$.

Properties of the inverse matrix. If $\boldsymbol{A}$ and $\boldsymbol{B}$ are $k \times k$ invertible matrices, then:

1. $\boldsymbol{AB}$ is invertible, and $\boldsymbol{(AB)^{-1} = B^{-1}A^{-1}}$
2. $(\boldsymbol{A}^{-1})^{-1} = \boldsymbol{A}$, and hence $\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I_k}$
3. $(n\boldsymbol{A})^{-1} = \frac{1}{n}A^{-1}$
4. The inverse of a diagonal matrix $\boldsymbol{D}$, $\boldsymbol{D}^{-1}$ is a diagonal matrix with entries $\frac{1}{d_{ii}}$ in the diagonal.

As we said before, matrices are a way of projecting vectors into a different coordinate system. When this projection results in each vector being mapped to a unique new vector in each new coordinate system, (i.e., vector $\boldsymbol{x}$ and only vector $\boldsymbol{x}$ gets projected into vector $\boldsymbol{y} = \boldsymbol{Ax}$), we can inverse the operation (i.e. if we are at $\boldsymbol{y}$, we can go back to $\boldsymbol{x}$). However, if two or more vectors get projected into some other vector (both $\boldsymbol{x}$ and $\boldsymbol{w}$ get projected into vector $\boldsymbol{y} = \boldsymbol{Ax} = \boldsymbol{Aw}$), then we cannot reverse the operation (i.e. if we are at $\boldsymbol{y}$, we don't know whether we should get back to $\boldsymbol{x}$ or $\boldsymbol{w}$).

Take, for example, the next very simple linearly dependent matrix:

$$\boldsymbol{A} = \begin{bmatrix} 0.5 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

This matrix halves the x_1 coordinate, keeps the x_2 coordinate as is, and gets rid of the x_3 dimension. So, it squishes the space into a plane as seen in Figure 13. Since several vectors in the cube to the left land on the plane to the right, it is impossible to go back to the original space. I.e., there is no transformation that takes a vector in the plane to the right, uniquely projects it into the $3 - D$ space, and allows us to retrieve the original space.

To continue with this easy example, now take matrix

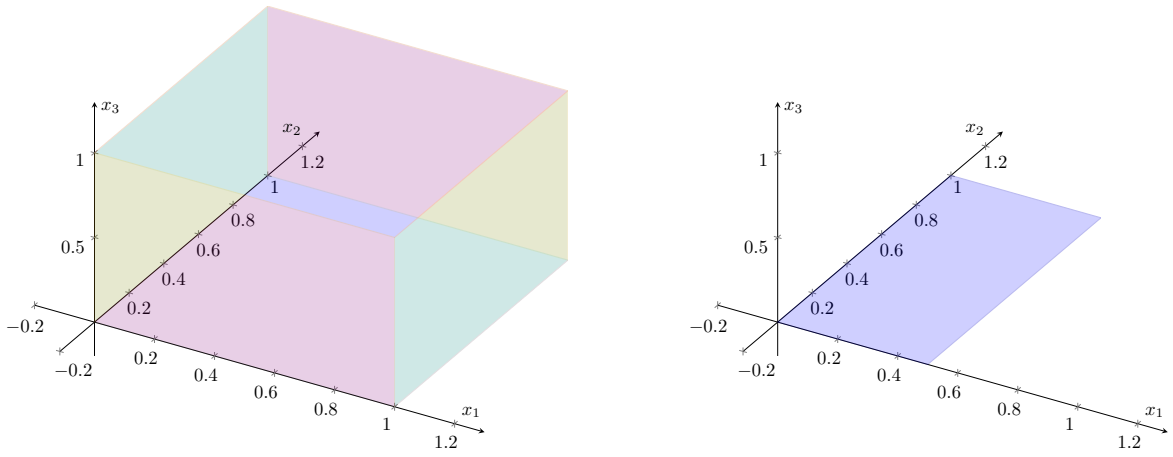


Figure 13: When matrix $\boldsymbol{A}$ is linearly dependent, it has no inverse

$$\boldsymbol{B} = \begin{bmatrix} 0.5 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

which again, halves the x_1 coordinates, keeps the x_2 coordinates as is, and doubles the x_3 coordinates, so it transforms a cube into a rectangular prism. This transformation is invertible, since each vector in the rectangular prisme, a 3D space, can be uniquely mapped back into the cube (see Figure 14)

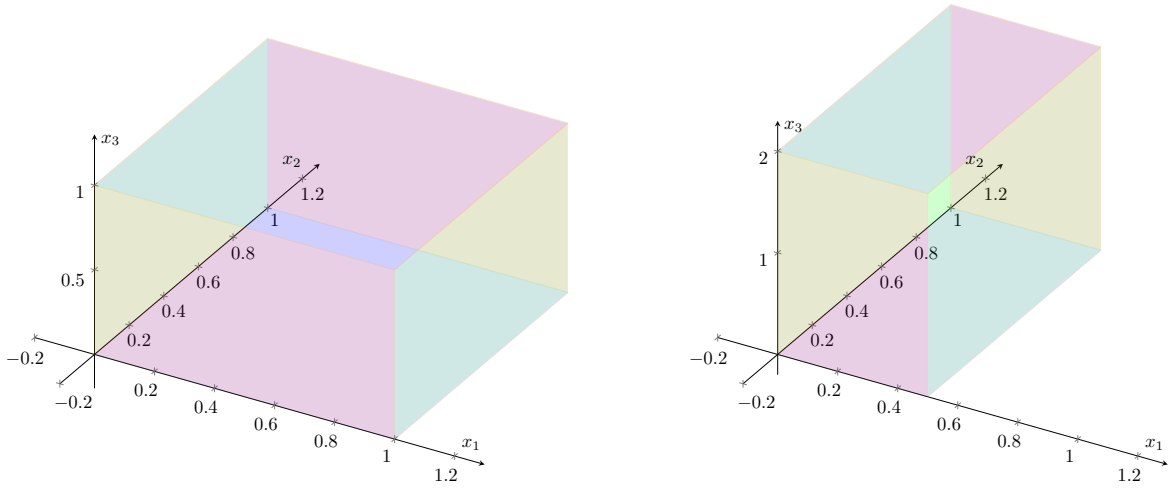


Figure 14: When matrix $\boldsymbol{A}$ uniquely maps vectors, it has an inverse.

In this case:

$$\boldsymbol{B}^{-} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0.5 \end{bmatrix}$$

i.e., a matrix that doubles the x_1 coordinate, leaves the x_2 coordinate as is, and halves the x_3 coordinate. (the exact inverse procedure). Note this would be impossible for matrix $\boldsymbol{A}$ since it would require to divide by 0.

3.2.9 Orthogonal and Unitary matrices

A real square $k \times k$ matrix, $\boldsymbol{Q}$ is said to be **orthogonal or orthonormal** if all it's columns and rows are orthonormal. This implies that $\boldsymbol{Q}$ is invertible, and that:

$$\boldsymbol{Q}\boldsymbol{Q}^{'} = \boldsymbol{Q}^{'}\boldsymbol{Q} = \boldsymbol{I}_k \qquad \Longrightarrow \qquad \boldsymbol{Q}^{-1} = \boldsymbol{Q}^{'}.$$

A generalization for complex $k \times k$ matrices is the concept of **unitary** matrices. A complex matrix $\boldsymbol{U}$ is said to be unitary if its conjugate transpose is also its inverse:

$$\boldsymbol{U}^* = \boldsymbol{U}^{-1} \qquad \Longrightarrow \qquad \boldsymbol{U}^*\boldsymbol{U} = \boldsymbol{U}\boldsymbol{U}^* = \boldsymbol{I}_k$$

Properties of Orthogonal matrices

1. The determinant of an orthogonal matrix is either 1 or -1.
2. For real $k \times 1$ vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, $(\boldsymbol{Q}\boldsymbol{x}) \cdot (\boldsymbol{Q}\boldsymbol{y}) = \boldsymbol{x} \cdot \boldsymbol{y}$.

Properties of Unitary matrices

1. The determinant of a unitary matrix is a complex number with absolute value 1.
2. For complex $k \times 1$ vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, $(\boldsymbol{U}\boldsymbol{x}) \cdot (\boldsymbol{U}\boldsymbol{y}) = \boldsymbol{x} \cdot \boldsymbol{y}$.
3. $\boldsymbol{U}$ is diagonizable, as a consequence of the spectral theorem.
4. The eigenvalues of $\boldsymbol{U}$ are complex numbers with absolute value 1.

Now that all necessary concepts have been defined, I introduce the fundamental theorem of linear algebra:

3.3 The Rank-Nullity Theorem

(Also known as the fundamental theorem of linear algebra)
Given an $r \times k$ matrix $\boldsymbol{A}$, then:

1. $\text{rank}(\boldsymbol{A}) + \text{nul}(\boldsymbol{A}) = k$
2. $\left\{\boldsymbol{x} \in \mathbb{R}^k | \boldsymbol{y} \in Sp(\boldsymbol{A}^{'}), \boldsymbol{x} \cdot \boldsymbol{y}\right\} = \ker(\boldsymbol{A})$

Further, the following statements are equivalent:

1. $\boldsymbol{A}$ is invertible
2. $\text{rank}(\boldsymbol{A}) = k$ and $\text{nul}(\boldsymbol{A}) = 0$.
3. $\det(\boldsymbol{A}) \neq 0$.
4. The columns of $\boldsymbol{A}$ are linearly independent.

No proof is provided.

3.4 Matrix decomposition

Decoposition is an operation that is similar to factorization. Decompositions come in handy specially when trying to save memory when doing calculations on the computer. Some help us understand the properties of a matrix. Among decompositions methods, those who rely on diagonalization are of particular importance, among which we have the spectral decomposition and eigendecomposition. We also have the the Cholesky and the singular value decompositions are also of great importance. However, they far exceed a mere refreshser.

Another decomposition I intruduce is the rank decomposition, since I'll be using it later to prove a property of idempotent matrices.

3.4.1 Spectral Decomposition

This decomposition comes from the **spectral theorem** that states that for any symmetric $k \times k$ matrix, there are exactly k eigenvalues, including multiplicities, the eigenvectors are orthogonal, and can be chosen from an orthonormal basis.

So now take any symmetric $k \times k$ matrix $\boldsymbol{A}$. Then, we can decompose $\boldsymbol{A}$ as:

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{'}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix with $\boldsymbol{A}$'s eigenvalues (all real, since the matrix is symmetric), and $\boldsymbol{Q}$ is a matrix with the respective eigenvectors, which are orthogonal (again, since the matrix is symmetric). Since all eigenvalues are real, all eigenvectors orthogonal, and the eigenvectors are defined up to a multiplicative constant, we can choose $\boldsymbol{Q}$ to be an orthonormal matrix, i.e., $\boldsymbol{Q}^{'}\boldsymbol{Q} = \boldsymbol{I_k}$.

To construct this decomposition, consider any eigenvector $\boldsymbol{v}$ chosen from an orthonormal basis, and let $\boldsymbol{Q}$ be the matrix of these orthonormal eigenvectors:

$$\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v} \quad \Longrightarrow \quad \boldsymbol{A}\boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{\Lambda} \quad \Longrightarrow \quad \boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{'}$$

3.4.2 Eigendecomposition

Take any square $k \times k$ matrix $\boldsymbol{A}$. Denote $\boldsymbol{Q}$ the matrix of it's eigenvectors, and Λ a diagonal matrix with its associated eigenvalues. The **eigendecomposition theorem** tells us that if $\boldsymbol{Q}$ is a square matrix, then we can decompose $\boldsymbol{A}$ as:

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

This decomposition is obtained as follows, starting from the definition of eigenvectors:

$$\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v} \quad \Longrightarrow \quad \boldsymbol{A}\boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{\Lambda} \quad \Longrightarrow \quad \boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{-1}$$

Where the last step requires $\boldsymbol{Q}$ to be invertible (linearly independent eigenvectors). The inverse of a matrix can be obtained using eigendecomposition:

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

3.4.3 Singular Value Decomposition

The **SVD** generalizes the eigendecomosition to rectangular complex matrices. In particular, for any $k \times r$ complex matrix $\boldsymbol{A}$, there exists a $k \times k$ unitary matrix $\boldsymbol{U}$, a $r \times r$ unitary matrix $\boldsymbol{V}$, and a $k \times r$ diagonal matrix $\boldsymbol{\Sigma}$ such that:

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{*}$$

In addition, if $\boldsymbol{A}$ is a real matrix, then $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal basis (see eigendecomposition), and the decomposition is denoted as $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{'}$.

3.4.4 Rank Decomposition

Rank decomposition Take any $\boldsymbol{k \times s}$ matrix $\boldsymbol{A}$ with rank r. Then there exists a $k \times r$ matrix $\boldsymbol{C}$, and an $r \times s$ matrix $\boldsymbol{D}$ such that:

$$\boldsymbol{A} = \boldsymbol{C}\boldsymbol{D}$$

Though a theorem guarrantees this decomposition to exist, is not unique, as alternative matrices $\tilde{\boldsymbol{C}}$ and $\tilde{\boldsymbol{D}}$ may also yield $\boldsymbol{A}$ when multiplied.

3.5 Inverse generalizations

3.5.1 Right and left inverses

Now consider $\boldsymbol{A}$ a $k \times r$ matrix. We say that $\boldsymbol{A}$ is left invertible if ther exists an $r \times k$ matrix $\boldsymbol{B}$ such that $\boldsymbol{BA} = \boldsymbol{I_r}$. $\boldsymbol{B}$ is called the left inverse of $\boldsymbol{A}$. Similarly, we say that $\boldsymbol{A}$ is right invertible if there exists an $r \times k$ matrix $\boldsymbol{C}$ such that $\boldsymbol{AC} = \boldsymbol{I_k}$. In this case, $\boldsymbol{C}$ is called the right inverse of $\boldsymbol{A}$.

If a matrix is both right and left invertible, then it is invertible.

For any $k \times r$ matrix $\boldsymbol{A}$:

1. if $k > r$ (more rows than columns), then $\boldsymbol{A}$ is not right invertible, and it is left invertible if its row echelon form[a] has a pivot in every column.
2. if $k < r$ (less rows than columns), then $\boldsymbol{A}$ is not left invertible, and it is right invertible if its row echelon form has a pivot in every row.

Note that if a matrix is both right and left invertible, then it must be a square matrix. Further, it's left and right inverse are the same: $\boldsymbol{A}^{-1} = \boldsymbol{B} = \boldsymbol{C}$.

[a] the form resulting from Gauss-Jordan elimination on the rows

3.5.2 Moore-Penrose inverse or Pseudoinverse

For a real matrix $\boldsymbol{A} \in \mathbb{R}^{k \times r}$ or a complex matrix $\boldsymbol{A} \in \mathbb{C}^{k \times r}$, the **pseudoinverse**, denoted $\boldsymbol{A}^+$ or $\boldsymbol{A}^\dagger$, is an $r \times k$ matrix defined overa similar domain (real or complex) that satisfies:

1. $\boldsymbol{AA}^+\boldsymbol{A} = \boldsymbol{A}$
2. $\boldsymbol{A}^+\boldsymbol{AA}^+ = \boldsymbol{A}^+$
3. Hermitian: $(\boldsymbol{AA}^+)^* = \boldsymbol{AA}^+$ and $(\boldsymbol{A}^+\boldsymbol{A})^* = \boldsymbol{A}^+\boldsymbol{A}$.

The pseudoinverse is unique and is guarrantieed to exist for any matrix.

Properties of the pseudoinverse

1. If $\boldsymbol{A}$ is real, so is $\boldsymbol{A}^+$.
2. If $\boldsymbol{A}$ is invertible, then $\boldsymbol{A}^+ = \boldsymbol{A}^{-1}$.
3. $(\boldsymbol{A}^+)^+ = \boldsymbol{A}$
4. $(c\boldsymbol{A})^+ = c^{-1}\boldsymbol{A}^+$.
5. The pseudoinverse of a $k \times r$ rectangular diagonal matrix $\boldsymbol{D}$, with elements entries d_{ii} on the diagonal, is a $r \times k$ rectangular diagonal matrix with elements d_{ii}^{-1} on the diagonal.
6. When $\boldsymbol{A}$ has full rank (i.e., $\text{rank}(\boldsymbol{A}) = \min\{r, k\}$), then
   - If $\boldsymbol{A}$ has linearly independent columns, $\boldsymbol{A}^+ = (\boldsymbol{A}^*\boldsymbol{A})^{-1}\boldsymbol{A}^*$. In particular, if $\boldsymbol{A}$ is real, $\boldsymbol{A}^+ = (\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'$.
   - If $\boldsymbol{A}$ has linearly independent rows, $\boldsymbol{A}^+ = \boldsymbol{A}^*(\boldsymbol{AA}^*)^{-1}$. In particular, if $\boldsymbol{A}$ is real, $\boldsymbol{A}^+ = \boldsymbol{A}'(\boldsymbol{A}'\boldsymbol{A})^{-1}$.

Point 6. is rather easy to prove. Just remember that if $\boldsymbol{A}$ has full rank, then $\boldsymbol{A}^*\boldsymbol{A}$ is a positive definite matrix, and as such, has an inverse. Then it's straightforward to show that these forms satisfy the definition of the Moore-Penrose inverse.

3.6 System of equations

3.6.1 Linear equations

Consider a system of r linear equations with k unknowns given by:

$$\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1k}x_k \\ y_2 &= a_{21}x_1 + a_{22}x_2 + \cdots + a_{2k}x_k \\ \vdots &= \qquad\qquad \vdots \\ y_r &= a_{r1}x_1 + a_{r2}x_2 + \cdots + a_{rk}x_k \end{aligned}$$

We can rewrite this in matrix form as:

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$$

Where

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_r \end{bmatrix} \qquad \boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{r1} & a_{r2} & \cdots & a_{rk} \end{bmatrix} \qquad \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix}$$

We can interpret it as $\boldsymbol{y}$ being a transformation applied on $\boldsymbol{x}$ by matrix $\boldsymbol{A}$, where vector $\boldsymbol{y}$ and matrix $\boldsymbol{A}$ are known. Focus on the case where $\boldsymbol{A}$ is a square invertible matrix. This case is known as the **just identified** case, since we have the same number of linearly independent equations as the number of unknowns. We can also solve this problem by applying the inverse of $\boldsymbol{A}$ to solve for $\boldsymbol{x}$.

$$\boldsymbol{A^{-1}y} = \boldsymbol{A^{-1}A}\boldsymbol{x} = x$$

However, when $\boldsymbol{A}$ is not invertible (e.g., it is a rectangular matrix), there is either no solution, or the solution is not unique. However, we can use the SVD to find a solution.

The first case is when the model is **underidentified**, meaning there are less equations than unknowns, i.e., for $r < k$. In general, this model allows for an infinite number of solutions, though we could engineer some examples for which this is not the case. The second case is when the model is **overidentified**, meaining there are more equations than unknowns, i.e., for $r > k$. In this case, and as a general rule, there is no solution that satisfies all equations, though again, we could engineer some examples for which there is a unique solution, or even an infinite number of solutions.

The SVD allows to use the pseudoinverse of $\boldsymbol{A}$ to find a "close enough" solution for the overidentified case (which incidentally coincides with the minimization of the sum of squared residuals in the linear regression model), or find the solution of the underidentified case with the smallest norm. In particular, focus on the a real matrix $\boldsymbol{A}$, and using the SVD, $\boldsymbol{A} = \boldsymbol{U\Sigma V}'$. Using the properties of the pseudoinverse:

$$\boldsymbol{A}^+ = \boldsymbol{V\Sigma}^+\boldsymbol{U}'$$

Where $\boldsymbol{\Sigma}^+$ is a rectangular diagonal matrix with the reciprocals of the diagonal elements of Σ, and the transpose of $\boldsymbol{U}$ and $\boldsymbol{V}$ follow from them being unitary matrices. Then, we can provide a "close" solution to the underidentified and overidentified cases as:

$$\tilde{\boldsymbol{x}} = \boldsymbol{A}^+\boldsymbol{y} = \boldsymbol{V\Sigma}^+\boldsymbol{U}'\boldsymbol{y}$$

For the underidentified case, from all the possible vectors $\boldsymbol{x}$ that satisfy the condition, the solution above yields the one with the smallest norm. For the overidentified case, the solution chooses $\boldsymbol{x}$ so as to minimize the norm of $\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|$, which is known as the least squares solution.

Application: linear regression Consider the linear regression model where an outcome y_i is determined by a set of k regressors $\boldsymbol{X}_i = (x_1, \cdots, x_k)$ and their coefficients $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_k)$. We have n observations. Denote $\boldsymbol{y} = (y_1, \cdots, y_n)^{'}$, $\boldsymbol{X} = (X_1, \cdots, X_n)^{'}$, thus the model can be rewritten as:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}$$

Assume that $\boldsymbol{X}$ has full rank, and $n > k$. A solution for this model is given by:

$$\hat{\boldsymbol{\beta}} = \boldsymbol{X}^{+}\boldsymbol{y}$$

Where $\boldsymbol{X}^{+}$ is the pseudoinverse of $\boldsymbol{X}$. Since we assumed $\boldsymbol{X}$ has full rank, and $n > k$, then $\boldsymbol{X}^{+} = (\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}$. Thus, the least square solution for this model is:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\boldsymbol{y}$$

3.6.2 Affine transformation

An affine transformation is a linear transformation with a displacement, of the form:

$$\underset{r\times 1}{y} = \underset{r\times k}{\boldsymbol{A}^{'}}\ \underset{k\times 1}{x} + \underset{r\times 1}{\boldsymbol{b}}$$

where $b \neq 0$. This form can be used to solve for more general system of equations with linearity on the unknowns. Take

$$\boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_r \end{bmatrix}$$

Then we can solve for x by following a similar procedure as before:

$$\boldsymbol{y} - \boldsymbol{b} = \boldsymbol{A}\boldsymbol{x} \quad \Longrightarrow \quad \boldsymbol{x} = \boldsymbol{A}^{+}(\boldsymbol{y} - \boldsymbol{b})$$

3.7 Idempotent matrix

We say that a matrix is idempotent if its square is equal to itself. Formally, a matrix $\boldsymbol{P}$ is idempotent if $\boldsymbol{P}^2 = \boldsymbol{P} = \boldsymbol{P}$. For this product to be defined, is is obvious that $\boldsymbol{P}$ must be a square. More generally, an idempotent matrix satisfies $\boldsymbol{P} = \boldsymbol{P}^n$ for any $n \geq 2$. The canonical example of a square matrix is the identity matrix. Take a 2×2 example:

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Take this other example from Wikipedia:

$$\begin{bmatrix} 3 & -6 \\ 1 & -2 \end{bmatrix}\begin{bmatrix} 3 & -6 \\ 1 & -2 \end{bmatrix} = \begin{bmatrix} 3 & -6 \\ 1 & -2 \end{bmatrix}$$

Note, however, that for this second matrix, its determminant is equal to zero (and thus, it is not invertible, as stated by the rank-nullity theorem). In general we can prove that if a matrix is idempotent and invertible, it must be the identity matrix. Assume matrix $\boldsymbol{P}$ is idempotent and invertible, then:

$$\boldsymbol{P}^2 = \boldsymbol{P} \implies \boldsymbol{P}^{-1}\boldsymbol{P}^2 = \boldsymbol{P}^{-1}\boldsymbol{P} \implies \boldsymbol{P} = \boldsymbol{I}$$

So, if a $k \times k$ matrix $\boldsymbol{P}$ is idempotent and it's not the identity matrix, then it must not be invertible. By the rank-nullity theorem, this means that such matrix must have a rank of less than k.

Two important results. First, An idempotent matrix has only eigenvalues of 0 and 1, specifically, the number of eigenvalues equal to 1 is rank($\boldsymbol{P}$), and thus, the number of eigenvalues equal to 0 is $k - \text{rank}(\boldsymbol{P}) = \text{nul}(\boldsymbol{P})$. Second, the trace of an idempotent matrix is equal to its rank.

We prove the first statement. For any $n \geq 1$, we have that:

$$\lambda \boldsymbol{x} = \boldsymbol{P}\boldsymbol{x} = \boldsymbol{P}^n \boldsymbol{x} = \lambda^n \boldsymbol{x} \qquad \Longrightarrow \lambda = \lambda^n$$

Which is only possible if $\lambda = 0$ or $\lambda = 1$, hence, all eigenvalues are either 0 or 1.

If $\boldsymbol{P}$ is symmetric, then we can use the spectral decomposition to prove the first statement. Since $\boldsymbol{P} = \boldsymbol{P}^n$,

$$\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{-1} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{-1}\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{-1} \cdots \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{-1} = \boldsymbol{Q}\boldsymbol{\Lambda}^{\boldsymbol{n}}\boldsymbol{Q}^{-1} \qquad \Longrightarrow \qquad \Lambda = \Lambda^n$$

Which, again, is only possible if Λ is a diagonal matrix with only 1's and 0's in its diagonal.

To prove that the rank of an idempotent matrix $\boldsymbol{P}$ is equal to its trace, we'll make use of rank decomposition. Especifically, for a $k \times k$ idempotent matrix $\boldsymbol{P}$ with rank $r < k$, we can decompose it as: $\boldsymbol{P} = \boldsymbol{C}\boldsymbol{D}$, where $\boldsymbol{C}$ is a $k \times r$ matrix, and $\boldsymbol{D}$ is a $r \times k$ matrix. Note that since $\boldsymbol{C}$ and $\boldsymbol{D}$ have full column and row rank, respectively, their reduced row echelon forms have a pivot in every column and row, respectively. Thus, $\boldsymbol{C}$ is left invertible, and $\boldsymbol{D}$ is right invertible. Denote the left inverse of $\boldsymbol{C}$, $\boldsymbol{C}^{-1}$, and the right inverse of $\boldsymbol{D}$, $\boldsymbol{D}^{-1}$. Then, we can write:

$$\begin{aligned} \boldsymbol{P}^2 = \boldsymbol{C}\boldsymbol{D}\boldsymbol{C}\boldsymbol{D} &= \boldsymbol{C}\boldsymbol{D} = \boldsymbol{P} \\ \boldsymbol{C}^{-1}\boldsymbol{C}\boldsymbol{D}\boldsymbol{C}\boldsymbol{D}\boldsymbol{D}^{-1} &= \boldsymbol{C}^{-1}\boldsymbol{C}\boldsymbol{D}\boldsymbol{D}^{-\boldsymbol{1}} \\ \boldsymbol{D}\boldsymbol{C} &= \boldsymbol{I}_r \end{aligned}$$

So:

$$\text{tr}(\boldsymbol{P}) = \text{tr}(\boldsymbol{C}\boldsymbol{D}) = \text{tr}(\boldsymbol{D}\boldsymbol{C}) = \text{tr}(\boldsymbol{I}_r) = r = \text{rank}(\boldsymbol{P})$$

Finally, since the trace of a matrix is equal to the sum of its eigenvalues, all of which are either 1 or 0 for an idempotent matrix, we have that the number of 1's in the diagonal of $\boldsymbol{P}$ is equal to its rank, and the number of 0's is equal to its nullity.

3.7.1 Application: Distribution of the variance estimator

Take a linear regression model, $Y_i = \boldsymbol{X}_i^{'}\boldsymbol{\beta} + e_i$, $i \in \{1, \cdots, n\}$, where the error term is normally distributed with mean zero and variance σ^2, $e_i | \boldsymbol{X}_i \sim N(0, \sigma^2)$, so the errors are homoskedastic and uncorrelated. Here, $\boldsymbol{X}_i$ is a $k \times 1$ vector of regressors, and $\boldsymbol{\beta}$ is a $k \times 1$ vector of coefficients.

We can obtain the OLS estimator $\hat{\boldsymbol{\beta}}$ as in section 3.6.1. Denote $\boldsymbol{e} = (e_1, \cdots, e_n)^{'}$ the vector of errors. Since they are homoskedastic and uncorrelated, their joint distribution is denoted as $\boldsymbol{e} \sim N(0, \sigma^2 \boldsymbol{I}_n)$

We defined the fitted values as $\hat{Y}_i = \boldsymbol{X}_i^{'}\hat{\boldsymbol{\beta}}$, and the regression residuals as $\hat{e}_i = Y_i - \hat{Y}_i$. An unbiased estimator of the variance of the error term is given by:

$$s^2 = \frac{1}{n-k}\sum_{i=1}^{n} \hat{e}_i^2 = \frac{1}{n-k}\hat{\boldsymbol{e}}^{'}\hat{\boldsymbol{e}}$$

where $\hat{\boldsymbol{e}} = (\hat{e}_1, \cdots, \hat{e}_n)$. Define $\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}$, called the Annihilation Matrix. It is easy to show that it is idempotent, symmetric, and has rank $n - k$. It is also straightforward to show that $\hat{\boldsymbol{e}} = \boldsymbol{M}\boldsymbol{e}$. We're interested in the distribution of the sum of squared residuals s^2. Construct the test statistic as:

$$(n-k)\frac{s^2}{\sigma^2} = \frac{1}{\sigma^2}\hat{\boldsymbol{e}}^{'}\hat{\boldsymbol{e}} = \frac{1}{\sigma^2}\left(\boldsymbol{M}\boldsymbol{e}\right)^{'}\boldsymbol{M}\boldsymbol{e} = \frac{1}{\sigma^2}\boldsymbol{e}^{'}\boldsymbol{M}^{'}\boldsymbol{M}\boldsymbol{e} = \left(\frac{1}{\sigma}\boldsymbol{e}^{'}\right)\boldsymbol{M}\left(\frac{1}{\sigma}\boldsymbol{e}\right) \equiv \boldsymbol{z}^{'}\boldsymbol{M}\boldsymbol{z}$$

where the last steps follow from $\boldsymbol{M} = \boldsymbol{M}^{'}$ and $\boldsymbol{M}^2 = \boldsymbol{M}$ (symmetric and idempotent), and then we defined $z \equiv (1/\sigma)\boldsymbol{e} \sim N(0, \boldsymbol{I}_n)$. Note this leaves us with a quadratic form. Since $\boldsymbol{M}$ is symmetric, we use a spectral decomposition $\boldsymbol{M} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{'}$, where $\boldsymbol{Q}$ is orthonormal. Since $\boldsymbol{M}$ is idempotent with rank $n-k$, $\boldsymbol{\Lambda}$ is a diagonal $n \times n$

matrix with $n-k$ 1's and k 0's in its diagonal. We can rearrange the matrix of eigenvalues and of eigenvectors so as to leave the 1's in $\boldsymbol{\Lambda}$ in the top left corner:

$$(n-k)\frac{s^2}{\sigma^2} = \boldsymbol{z}^{'}\boldsymbol{M}\boldsymbol{z} = \boldsymbol{z}^{'}\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{'}\boldsymbol{z} = \boldsymbol{z}^{'}\boldsymbol{Q}\begin{bmatrix}\boldsymbol{I}_{n-k} & 0\\ 0 & 0\end{bmatrix}\boldsymbol{Q}^{'}\boldsymbol{z}$$

Since $\boldsymbol{z}$ follows a multivariate standard normal distribution, and $\boldsymbol{Q}$ is orthonormal ($\boldsymbol{Q}\boldsymbol{Q}^{'} = \boldsymbol{I}_n$)

$$\boldsymbol{Q}^{'}\boldsymbol{z} \sim N(0, \boldsymbol{Q}\boldsymbol{I}_n\boldsymbol{Q}^{'}) = N(0, \boldsymbol{I}_n)$$

Define $\tilde{\boldsymbol{z}} \equiv \boldsymbol{Q}^{'}\boldsymbol{z}$. Then:

$$(n-k)\frac{s^2}{\sigma^2} = \tilde{\boldsymbol{z}}^{'}\begin{bmatrix}\boldsymbol{I}_{n-k} & 0\\ 0 & 0\end{bmatrix}\tilde{\boldsymbol{z}} = \sum_{i=1}^{n-k}\tilde{z}_i^2$$

Note this is the sum of $n-k$ squared standard normal random variables, and thus, it follows a chi-squared distribution with $n-k$ degrees of freedom. Thus:

$$(n-k)\frac{s^2}{\sigma^2} \sim \chi^2_{n-k}$$

3.8 Demeaning and first-differencing matrices

Demeaning and first differencing matrices are widely used in econometrics as a way to deal with time invariant regressors in panel data. Namely, they are used to obtain the fixed-effects estimator and the first-differencing estimator. A **demeaning matrix** substracts the mean of a vector from each element of the vector. It is defined as:

$$\boldsymbol{M} = \boldsymbol{I_k} - \mathbf{1}(\mathbf{1}^{'}\mathbf{1})^{-1}\mathbf{1}^{'} = \boldsymbol{I_k} - \frac{1}{k}\mathbf{1}\mathbf{1}^{'}$$

where $\mathbf{1}$ is a $k \times 1$ vector of ones, $\mathbf{1} = (1, 1, \cdots, 1)$.

To see why it's called a demeaning matrix, consider a $k \times 1$ vector $\boldsymbol{x}$, and let $\bar{x}$ be its mean. Then, $\boldsymbol{M}\boldsymbol{x}$ is a $k \times 1$ vector with elements:

$$\boldsymbol{M}\boldsymbol{x} = \left(\begin{bmatrix}1 & 0 & \cdots & 0\\ 0 & 1 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 1\end{bmatrix} - \frac{1}{k}\begin{bmatrix}1 & 1 & \cdots & 1\\ 1 & 1 & \cdots & 1\\ \vdots & \vdots & \ddots & \vdots\\ 1 & 1 & \cdots & 1\end{bmatrix}\right)\begin{bmatrix}x_1\\ x_2\\ \vdots\\ x_k\end{bmatrix} = \begin{bmatrix}x_1\\ x_2\\ \vdots\\ x_k\end{bmatrix} - \begin{bmatrix}\frac{1}{k}\sum_{i=1}^{k}x_i\\ \frac{1}{k}\sum_{i=1}^{k}x_i\\ \vdots\\ \frac{1}{k}\sum_{i=1}^{k}x_i\end{bmatrix} = \begin{bmatrix}x_1 - \bar{x}\\ x_2 - \bar{x}\\ \vdots\\ x_k - \bar{x}\end{bmatrix}$$

The **first-differencing** matrix is a $k-1 \times k$ matrix that substracts the previous element of a vector from the current one (for obvious reasons, it starts from $a_2 - a_1$):

$$D = \begin{bmatrix}-1 & 1 & 0 & \cdots & 0 & 0\\ 0 & -1 & 1 & \cdots & 0 & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & 0 & \cdots & -1 & 1\end{bmatrix}$$

To see this first differencing:

$$\boldsymbol{D}\boldsymbol{x} = \begin{bmatrix}-1 & 1 & 0 & \cdots & 0 & 0\\ 0 & -1 & 1 & \cdots & 0 & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & 0 & \cdots & -1 & 1\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ \vdots\\ x_k\end{bmatrix} = \begin{bmatrix}-x_1 + x_2\\ -x_2 + x_3\\ \vdots\\ -x_{k-1} + x_k\end{bmatrix} = \begin{bmatrix}x_2 - x_1\\ x_3 - x_2\\ \vdots\\ x_k - x_{k-1}\end{bmatrix}$$

4 Matrix Calculus

4.1 Gradient, Hessian and Jacobian

Let $\boldsymbol{x}$ be a $k \times 1$ vector, and $f : \mathbb{R}^k \to \mathbb{R}$ a function of class C^1. $\boldsymbol{x} \mapsto f(\boldsymbol{x})$. The derivative of f with respect to $\boldsymbol{x}$ and with respect to $\boldsymbol{x}^{'}$ are defined as:[2]

$$\frac{\partial f}{\partial \boldsymbol{x}} f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_k} \end{bmatrix} \qquad \frac{\partial f}{\partial \boldsymbol{x}^{'}} f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_k} \end{bmatrix}$$

The derivative with rexpect to $\boldsymbol{x}$ is a $k \times 1$ commonly referred to as the **gradient** of f with respect to $\boldsymbol{x}$, and is denoted ∇f or $Df(\boldsymbol{x})$. Among its many properties, it is a vector that points in the direction of the steepest ascent (fastest growth) of f at $\boldsymbol{x}$. For instance,

$$f(x_1, x_2) = -\frac{1}{2}(x_1^2 + x_2^2) \qquad \frac{\partial f}{\partial \boldsymbol{x}} f(\boldsymbol{x}) = \begin{bmatrix} -x_1 \\ -x_2 \end{bmatrix}$$

The plots on Figure 15 shows the contour lines of f, and the gradient at the point $(1, 1)$, which is $(-1, -1)^{'}$:

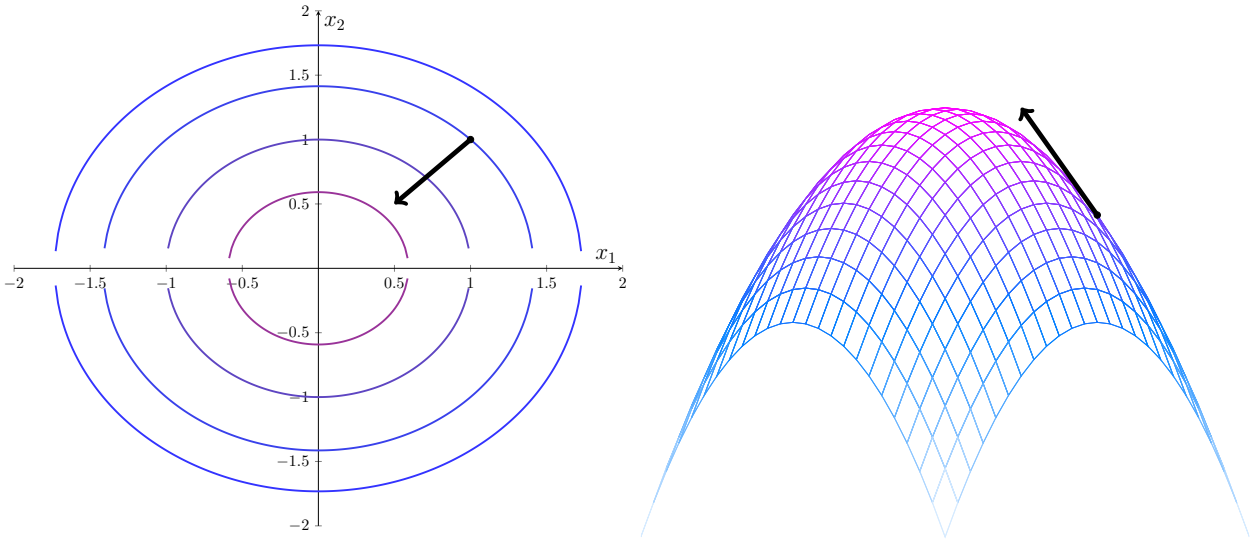


Figure 15: Gradient of $f(\boldsymbol{x})$

Note on the plot of the right that the gradient also allows to obtain the plane tangent to the surface of f, similar to how the derivative is the slope of the tangent line in the case of a function of one variable. Gradiennts are useful in optimization problems. Critical points are obtained when $\nabla f = 0$, and they defined global or local maxima and minima, as well as saddle points. Intuitively, when $\nabla f = 0$, then there is no direction for growth (or decline, for that matter) for infinitesimal changes in $\boldsymbol{x}$. In Figure 15, this occurrs at the top of the paraboloid.

If the function is of the class C^2, we can obtain the second derivative (first, w.r.t $\boldsymbol{x}$, and then, w.r.t. its transpose,

[2] Mathematicians are more used to using row vectors, and thus several of the definitions used by the mathematicians are the transpose of what we'll be seing here. Fear not, is mere convention. Economists and Ecoonometricians are more used to using column vectors, due to the structure of the data we usually work with, and the convention stuck. So if you find elsewere that the definitions are the transpose of what we have, well, there's nothing wrong.

$\boldsymbol{x}^{'}$):

$$\frac{\partial^2}{\partial x \partial x^{'}} f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial^2}{\partial x_1^2} f(\boldsymbol{x}) & \frac{\partial^2}{\partial x_1 \partial x_2} f(\boldsymbol{x}) & \cdots & \frac{\partial^2}{\partial x_1 \partial x_k} f(\boldsymbol{x}) \\ \frac{\partial^2}{\partial x_2 \partial x_1} f(\boldsymbol{x}) & \frac{\partial^2}{\partial x_2^2} f(\boldsymbol{x}) & \cdots & \frac{\partial^2}{\partial x_2 \partial x_k} f(\boldsymbol{x}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial x_k \partial x_1} f(\boldsymbol{x}) & \frac{\partial^2}{\partial x_k \partial x_2} f(\boldsymbol{x}) & \cdots & \frac{\partial^2}{\partial x_k^2} f(\boldsymbol{x}) \end{bmatrix}$$

This matrix is also known as the **Hessian**, denoted $\mathcal{H}$ or $D^2 f(\boldsymbol{x})$ and it is a square matrix. Further, it is symmetric since, by Young's theorem,

$$\frac{\partial^2}{\partial x_i \partial x_j} f(\boldsymbol{x}) = \frac{\partial^2}{\partial x_j \partial x_i} f(\boldsymbol{x}) \quad \forall i, j$$

Just as in the univariate case, a function f is concave at $\boldsymbol{x}$ if its Hessian is negative semidefinite at $\boldsymbol{x}$, and convex at $\boldsymbol{x}$ if its positive semidefinite at $\boldsymbol{x}$. Further, a function is strictly concave or strictly convex at $\boldsymbol{x}$ if its Hessian is negative or positive definite at $\boldsymbol{x}$ respectively. A function is concave if it is concave at every $\boldsymbol{x}$. Analogous definitions apply for the function being convex, strictly concave and strictly convex.

Take the function $f(x_1, x_2) = -\frac{1}{2}(x_1^2 + x_2^2)$. Its Hessian is:

$$\mathcal{H}(\boldsymbol{x}) = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$$

Since the first element of the diagonal is negative, and the determinant of the matrix is positive (alternating signs of the principal minors), the Hessian is negative definite, and thus, the function is strictly concave (Figure 15, right plot).

Now, consider a vector valued function $g : \mathbb{R}^k \to \mathbb{R}^r$, $\boldsymbol{x} \mapsto g(\boldsymbol{x})$ i.e., a function that maps from $\mathbb{R}^k$ to $\mathbb{R}^r$ for $r \geq 2$. Such an example is a system of demand functions, that maps for every vector of prices $\boldsymbol{p}$ to a vector of quantities demanded $\boldsymbol{q}(\boldsymbol{p})$. We can rewrite this function as:

$$g(\boldsymbol{x}) = \begin{bmatrix} g_1(\boldsymbol{x}) \\ g_2(\boldsymbol{x}) \\ \vdots \\ g_r(\boldsymbol{x}) \end{bmatrix}$$

In this case, we define the **Jacobian** matrix as:

$$J_x(x) = \begin{bmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} & \cdots & \frac{\partial g_1}{\partial x_k} \\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_2}{\partial x_2} & \cdots & \frac{\partial g_2}{\partial x_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_r}{\partial x_1} & \frac{\partial g_r}{\partial x_2} & \cdots & \frac{\partial g_r}{\partial x_k} \end{bmatrix}$$

were the subscript x indicates that the Jacobian is with respect to $\boldsymbol{x}$. Note that if g is a scalar function, then the Jacobian is the gradient of g with respect to $\boldsymbol{x}$. The Jacobiann is also denoted as $Dg(\boldsymbol{x})$, similar to the gradient. I'll stick to this notation going forward ($Df(\boldsymbol{x})$ for both, the Jacobian and the gradient in case of a scalar valued function, and $D^2 f(\boldsymbol{x})$ for the Hessian of a scalar-valued function).

4.2 Matrix derivatives

Consider $\boldsymbol{A}$ a $r \times k$ matrix, and $\boldsymbol{a}$ and $\boldsymbol{x}$, both $k \times 1$ vectors. Then:

And:

$$\frac{\partial^2}{\partial x \partial x^{'}} (x^{'} A x) = A + A^{'}$$

In particular, if we define a symmetric, square matrix $\boldsymbol{W}$, then:

$$\frac{\partial}{\partial x} (x^{'} W x) = 2Wx \qquad \frac{\partial^2}{\partial x \partial x^{'}} (x^{'} W x) = 2W$$

$f(\boldsymbol{x})$	$\frac{\partial f}{\partial \boldsymbol{x}}$	$\frac{\partial f}{\partial \boldsymbol{x}'}$
$\boldsymbol{a}'\boldsymbol{x}$	$\boldsymbol{a}$	$\boldsymbol{a}'$
$\boldsymbol{A}\boldsymbol{x}$	$\boldsymbol{A}'$	$\boldsymbol{A}$
$\boldsymbol{x}'\boldsymbol{A}$	$\boldsymbol{A}$	$\boldsymbol{A}'$
$\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$	$(\boldsymbol{A}+\boldsymbol{A}')\boldsymbol{x}$	$\boldsymbol{x}'(\boldsymbol{A}+\boldsymbol{A}')$

4.3 Chain rule

Consider the following functions:

- $f : \mathbb{R}^k \to \mathbb{R}^s$, $\boldsymbol{x} \mapsto f(\boldsymbol{x})$, where $\boldsymbol{x} \in \mathbb{R}^k$.
- $g : \mathbb{R}^r \to \mathbb{R}^k$, $\boldsymbol{y} \mapsto g(\boldsymbol{y})$, where $\boldsymbol{y} \in \mathbb{R}^r$.
- $f \circ g : \mathbb{R}^r \to \mathbb{R}^s$, $\boldsymbol{y} \mapsto f(g(\boldsymbol{y}))$.

Then the Jacobian of $f \circ g$ is given by:

$$\underset{r\times s}{D_y f \circ g(\boldsymbol{y})} = \underset{r\times k}{D_y g(y)} \underset{k\times s}{D_x f(g(\boldsymbol{y}))}$$

To see this, expand a bit the Jacobian of g and f:

$$Dg(\boldsymbol{y}) = \begin{bmatrix} \frac{\partial g_1}{\partial y_1} & \frac{\partial g_1}{\partial y_2} & \cdots & \frac{\partial g_1}{\partial y_r} \\ \frac{\partial g_2}{\partial y_1} & \frac{\partial g_2}{\partial y_2} & \cdots & \frac{\partial g_2}{\partial y_r} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_k}{\partial y_1} & \frac{\partial g_k}{\partial y_2} & \cdots & \frac{\partial g_k}{\partial y_r} \end{bmatrix} = \begin{bmatrix} Dg_1(\boldsymbol{y}) \\ Dg_2(\boldsymbol{y}) \\ \vdots \\ Dg_k(\boldsymbol{y}) \end{bmatrix}$$

$$Df(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_k} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_s}{\partial x_1} & \frac{\partial f_s}{\partial x_2} & \cdots & \frac{\partial f_s}{\partial x_k} \end{bmatrix} = \begin{bmatrix} Df_1(\boldsymbol{x}) & Df_2(\boldsymbol{x}) & \cdots & Df_k(\boldsymbol{x}) \end{bmatrix}$$

Now, note that the element on the $i-$th row and $j-$th column of the Jacobian of $f \circ g$ should denote the change of function $f_j(g(\boldsymbol{y}))$ when y_i changes. So we must consider how each function $g_1(\boldsymbol{y}), g_2(\boldsymbol{y}), \cdots, g_k(\boldsymbol{y})$ changes when y_i changes. This is given by the dot product between the $i-$th row of $D_y g(y)$, $D_y g_i(\boldsymbol{y})$ and the $j-$th column of $D_x f(g(\boldsymbol{y}))$, $D_x f_j(g(\boldsymbol{y}))$:

$$D_y f \circ g(\boldsymbol{y})_{ij} = \frac{\partial}{\partial y_i} f_j(g_1(\boldsymbol{y}), g_2(\boldsymbol{y}), \cdots, g_k(\boldsymbol{y})) = \sum_{m=1}^{k} \frac{\partial f_j}{\partial x_m}(g(\boldsymbol{y})) \frac{\partial g_m}{\partial y_i}(\boldsymbol{y}) = D_y g_i(\boldsymbol{y}) \cdot D_x f_j(g(\boldsymbol{y}))$$

where the subscript ij denotes the element on the $i-$th row and $j-$th column of the Jacobian of $f \circ g$. On the end, we obtain that this is equal to the dot product between the $i-$eth row of $D_y g(y)$ and the $j-$th column of $D_x f(g(\boldsymbol{y}))$, which satisfies the definition of matrix multiplication.

We can nest several chain rules as before. Now, consider functions f and g as defined above, and consider a third function, $h : \mathbb{R}^t \to \mathbb{R}^r$, $\boldsymbol{z} \mapsto h(\boldsymbol{z})$, where $\boldsymbol{z} \in \mathbb{R}^t$. Consider the composite function $f \circ g \circ h : \mathbb{R}^t \to \mathbb{R}^s$, $\boldsymbol{z} \mapsto f(g(h(\boldsymbol{z})))$. Then:

$$\underset{l\times s}{D_z f \circ g \circ h(z)} = \underset{l\times r}{D_z h(z)} \underset{r\times k}{D_y g(h(z))} \underset{k\times s}{D_x f(g(h(y)))}$$

4.4 Examples

4.4.1 OLS

Consider an outcome variable $Y_i = \boldsymbol{X}_i^{'}\boldsymbol{\beta} + e_i$, where X_i is a $k \times 1$ vector of regressors, β is a $k \times 1$ vector of coefficients, and e_i is a scalar error term. The OLS estimator minimizes the sum of squared errors, $e_i^2 = (Y_i - \boldsymbol{X}_i^{'}\boldsymbol{\beta})^2$ across observations:

$$\min_{\beta} \sum_{i=1}^{n} (Y_i - \boldsymbol{X}_i^{'}\boldsymbol{\beta})^2$$

Denote

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad \boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1^{'} \\ \boldsymbol{X}_2^{'} \\ \vdots \\ \boldsymbol{X}_n^{'} \end{bmatrix} \qquad \boldsymbol{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

Rewrite the OLS estiatmor:

$$\hat{\beta} = \quad \arg\min_{\beta} \quad (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^{'}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$$

Take the functions:

$$\underset{1\times 1}{f(\boldsymbol{z})} = \underset{1\times n}{\boldsymbol{z}^{'}}\,\underset{n\times 1}{\boldsymbol{z}} \qquad\qquad \underset{n\times 1}{g(\boldsymbol{\beta})} = \underset{n\times 1}{\boldsymbol{Y}} - \underset{n\times k}{\boldsymbol{X}}\,\underset{k\times 1}{\boldsymbol{\beta}}$$

To take the first derivative, we use the chain rule for the Jacobian above. First,

$$D_{\beta} g(\boldsymbol{\beta}) = -\boldsymbol{X}^{'}$$

and

$$D_z f(\boldsymbol{z}) = 2\boldsymbol{z} \qquad\qquad \implies D_z f(g(\boldsymbol{\beta})) = 2(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$$

Putting it all together by the chain rule:

$$\text{FOC}: \qquad \frac{\partial ...}{\partial \hat{\boldsymbol{\beta}}} = (-\boldsymbol{X}^{'})(2(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})) = -2\boldsymbol{X}^{'}\boldsymbol{Y} + 2\boldsymbol{X}^{'}\boldsymbol{X}\hat{\boldsymbol{\beta}} = 0$$

We get rid of the 2 and the negative sign. $\boldsymbol{X}^{'}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = 0$. Solving for $\hat{\boldsymbol{\beta}}$:

$$\boldsymbol{X}^{'}\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^{'}\boldsymbol{Y} \qquad\qquad \implies \qquad\qquad \hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\boldsymbol{Y}$$

Note this coincides with the least squares solution given by the Moore-Penrose inverse approach. A necessary and sufficient condition for the existence of the inverse is that $\boldsymbol{X}^{'}\boldsymbol{X}$ has full column rank (for which it suffices that $\boldsymbol{X}_i\boldsymbol{X}_i^{'}$ has full rank for all i). In particular, since we know that $\boldsymbol{X}_i\boldsymbol{X}_i^{'}$ is a quadratic form, we need it to be positive definite for it to be invertible. As for the second order condition, we now take the derivative of the FOC with respect to the transpose, $\boldsymbol{\beta}^{'}$:

$$\text{SOC}: \qquad \frac{\partial^2 ...}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{'}} = 2\boldsymbol{X}^{'}\boldsymbol{X}$$

A condition for the solution given by $\hat{\boldsymbol{\beta}}$ to be a unique minimizer is for the Hessian to be positive definite, (strict convexity). This occurs if $\boldsymbol{X}^{'}\boldsymbol{X}$ is positive definite, again, for which it suffices for $\boldsymbol{X}_i\boldsymbol{X}_i^{'}$ to be positive definite for all i.

4.4.2 Logit

The logit considers that the probability of a binary outcome Y_i is given by:

$$\Pr(Y_i = 1|\boldsymbol{X}_i) = \frac{e^{\boldsymbol{X}_i'\boldsymbol{\beta}}}{1 + e^{\boldsymbol{X}_i'\boldsymbol{\beta}}} \equiv \Lambda(\boldsymbol{X}_i'\boldsymbol{\beta})$$

Consider function $f(w) = \log(w)$, $g(z) = \Lambda(z)$, where the logistic distribution c.d.f. is:

$$\Lambda(z) = \frac{e^z}{1 + e^z} \qquad\qquad \frac{\partial \Lambda}{\partial z}(z) = \Lambda(z)(1 - \Lambda(z))$$

and function $h(\boldsymbol{\beta}) = \boldsymbol{X_i'}\boldsymbol{\beta}$. So the complete function would be $f(g(h(\boldsymbol{\beta}))) = \log(\Lambda(\boldsymbol{X}_i'\boldsymbol{\beta}))$.
The log-likelihood function is:

$$\hat{\boldsymbol{\beta}} = \quad \arg\max_{\beta} \quad l(\beta) = \sum_{i=1}^{n} \left[Y_i \log \Lambda(\boldsymbol{X}_i'\boldsymbol{\beta}) + (1 - Y_i) \log(1 - \Lambda(\boldsymbol{X}_i'\boldsymbol{\beta})) \right]$$

Again, we can use the chain rule to obtain the FOC:

$$\begin{aligned}
\frac{\partial l}{\partial \boldsymbol{\beta}} =& \sum_{i=1}^{n} \left[Y_i \left(\boldsymbol{X}_i \Lambda'(\boldsymbol{X}_i'\hat{\boldsymbol{\beta}})(1 - \Lambda(\boldsymbol{X}_i'\hat{\boldsymbol{\beta}})) \right) \frac{1}{\Lambda(\boldsymbol{X}_i'\hat{\boldsymbol{\beta}})} - (1 - Y_i) \left(\boldsymbol{X}_i \Lambda'(\boldsymbol{X}_i'\hat{\boldsymbol{\beta}})(1 - \Lambda(\boldsymbol{X}_i'\hat{\boldsymbol{\beta}})) \right) \frac{1}{1 - \Lambda(\boldsymbol{X}_i'\hat{\boldsymbol{\beta}})} \right] \\
=& \sum_{i=1}^{n} \left[\boldsymbol{X}_i Y_i (1 - \Lambda(\boldsymbol{X}_i'\hat{\boldsymbol{\beta}})) - \boldsymbol{X}_i (1 - Y_i) \Lambda(\boldsymbol{X}_i'\hat{\boldsymbol{\beta}}) \right] \\
=& \sum_{i=1}^{n} \left[\boldsymbol{X}_i Y_i - \boldsymbol{X}_i \Lambda(\boldsymbol{X}_i'\hat{\boldsymbol{\beta}}) \right] = 0
\end{aligned}$$

We can't solve for $\hat{\boldsymbol{\beta}}$ analytically, so we use numerical methods. As for the SOC, since we are looking for a maximum, we need the function to be strictly concave. Obtaining the Hessian:

$$\frac{\partial l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = \sum_{i=1}^{n} (-\boldsymbol{X}_i \boldsymbol{X}_i' \Lambda(\boldsymbol{X}_i'\boldsymbol{\beta})(1 - \Lambda(\boldsymbol{X}_i'\hat{\boldsymbol{\beta}})))$$

Since $\Lambda(z) \in (0, 1)$, for the function above to be negative definite, and thus strictly concave for the unique maximizer, a sufficient condition is for $\boldsymbol{X}_i \boldsymbol{X}_i'$ to be positive definite for all i.

4.4.3 GMM criterion function

Finally consider the General Methods of Moments estimator for a linear model with instrumental variable. Definte the function $g_i(\beta)$:

$$g_i(\boldsymbol{\beta}) = \boldsymbol{Z}_i (Y_i - \boldsymbol{X}_i'\boldsymbol{\beta})$$

where $\boldsymbol{Z}_i$ is a $r \times 1$ vector of instrumental variables. Define:

$$\bar{g}_n(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} g_i(\boldsymbol{\beta})$$

and for some diagonal (symmetric) positive definite weight matrix W, define the criterion function as

$$J(\boldsymbol{\beta}) = n \bar{g}_n(\boldsymbol{\beta})' \boldsymbol{W} \bar{g}_n(\boldsymbol{\beta})$$

The GMM estimator $\hat{\boldsymbol{\beta}}$ is given by the coefficients that minimize the criterion function:

$$\hat{\boldsymbol{\beta}}_{gmm} = \quad \arg\min_{\beta} \quad n \left(\sum_{i=1}^{n} \boldsymbol{Z}_i (Y_i - \boldsymbol{X}_i'\beta) \right)' \boldsymbol{W} \left(\sum_{i=1}^{n} \boldsymbol{Z}_i (Y_i - \boldsymbol{X}_i'\beta) \right)$$

Deriving the First order conditions:

$$\frac{\partial J}{\partial \boldsymbol{\beta}} = 2n \left(\sum_{i=1}^{n} -\boldsymbol{X}_i \boldsymbol{Z}_i^{'} \right) \boldsymbol{W} \left(\sum_{i=1}^{n} \boldsymbol{Z}_i (Y_i - \boldsymbol{X}_i^{'} \hat{\boldsymbol{\beta}}) \right) = 0$$

Again we can get rid of $2n$, and solve for $\hat{\boldsymbol{\beta}}$:

$$\left(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{Z}_i^{'} \right) \boldsymbol{W} \left(\sum_{i=1}^{n} \boldsymbol{Z}_i \boldsymbol{X}_i^{'} \hat{\boldsymbol{\beta}} \right) = \left(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{Z}_i^{'} \right) \boldsymbol{W} \left(\sum_{i=1}^{n} \boldsymbol{Z}_i Y_i \right)$$

$$\hat{\boldsymbol{\beta}}_{gmm} = \left[\left(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{Z}_i^{'} \right) \boldsymbol{W} \left(\sum_{i=1}^{n} \boldsymbol{Z}_i \boldsymbol{X}_i^{'} \right) \right]^{-1} \left(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{Z}_i^{'} \right) \boldsymbol{W} \left(\sum_{i=1}^{n} \boldsymbol{Z}_i Y_i \right)$$

As for the SOC:

$$\frac{\partial^2 J}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = 2n \left(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{Z}_i^{'} \right) \boldsymbol{W} \left(\sum_{i=1}^{n} \boldsymbol{Z}_i \boldsymbol{X}_i^{'} \right)$$

There is a unique minimizer if the Hessian is positive definite. Since this is a quadratic form, and W is positive definite by assumption, a sufficient condition for a positive definite Hessian is for $\boldsymbol{X}_i \boldsymbol{Z}_i^{'}$ to have a full column rank for all i.