BIOS 7345: Advanced Regression for Independent Data

Andrew J. Spieker, Ph.D.

Associate Professor of Biostatistics
Vanderbilt University

Set 18 (Supplement 1): Linear algebra

Version: 08/26/2025

Supplementary notes:

- Note: Linear algebra is *shockingly* self-referential. These distilled notes are not a substitute for a good first course in linear algebra; rather, they are intended to serve as a refresher on key concepts, ideas, and results—and as a source of reference material for the course.
- We will not go through everything in these notes, but you should review everything independently (you may skim over the material on generalized inverses).

Table of Contents

Notation and key definitions

Vectors:

- Typically given by bold, lower-case letters (e.g, $\mathbf{x}$).
  - On the board, I may or may not write $\vec{x}$.
- Exclusively *column* vectors:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}$$

- Absolute (L_1) norm, denoted by $||\mathbf{x}||_1$, given by $\sum_{i=1}^{N} |x_i|$.
- Euclidean (L_2) norm, denoted by $||\mathbf{x}||_2$, given by $\sqrt{\sum_{i=1}^{N} x_i^2}$.
- The zero-vector is denoted by $\mathbf{0}$.
- Similarly, $\mathbf{1}$ denotes a vector of ones.

Matrices:

- Typically given by bold, upper-case letters (e.g, $\mathbf{A}$).

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1K} \\ a_{21} & a_{22} & \cdots & a_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NK} \end{bmatrix} = \begin{bmatrix} \mathbf{a}_{\cdot 1} & \mathbf{a}_{\cdot 2} & \cdots & \mathbf{a}_{\cdot K} \end{bmatrix} = \begin{bmatrix} \mathbf{a}_{1\cdot}^{\mathsf{T}} \\ \mathbf{a}_{2\cdot}^{\mathsf{T}} \\ \vdots \\ \mathbf{a}_{N\cdot}^{\mathsf{T}} \end{bmatrix}.$$

- a_{ij} denotes the entry in the i^{th} row and j^{th} column of $\mathbf{A}$.
- $\mathbf{a}_{i\cdot}$ and $\mathbf{a}_{\cdot j}$ denote vectors of the i^{th} row and the j^{th} column of $\mathbf{A}$.
  - $\mathbf{a}_{i\cdot}^{\mathsf{T}}$ is "horizontal."
- We say $\mathbf{A}$ is an $N \times K$ matrix (read "N by K," meaning it has N rows and K columns).

Vectors and matrices: Definitions

- A **matrix** is a rectangular array of numbers.
- A **vector** is a matrix that consists of only one column.
- A **scalar** is a single constant.
- Note: matrix/vector elements and scalars will generally be assumed to be finite and real-valued. I'm don't believe complex numbers come up in this course at all.

Notation and key definitions

Square matrices:

- $\mathbf{A}$ is said to be *square* if $N = K$.
  - That is, if it has the same number of rows and columns.
- For instance:

$$\mathbf{A} = \begin{bmatrix} 1 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 1 & 2 \\ -1 & 0 \end{bmatrix}, \text{ and } \mathbf{C} = \begin{bmatrix} 1 & 2 & -3 \\ -1 & 0 & 0 \\ 0 & -3 & 1 \end{bmatrix}$$

are all examples of square matrices.

Notation and key definitions

Diagonal matrices:

- A square matrix, $\mathbf{A}$, is said to be *diagonal* if $a_{ij} = 0$ when $i \neq j$.
  - That is, if its non-diagonal entries are zero.
- For instance:

$$\mathbf{A} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 4 \end{bmatrix}$$

  is an example of a diagonal matrix. Importantly, the entries on the diagonals needn't be non-zero for the matrix to qualify as diagonal.
- Notation (context is important; if you don't know, ask :)):
  - $\text{diag}(\mathbf{x})$: a diagonal matrix with $\mathbf{x}$ along the diagonal (R notation).
  - $\text{diag}(K)$: a $K \times K$ identity matrix (R notation).
  - $\text{diag}(x_i)$: a diagonal matrix with $x_1, \ldots, x_N$ along the diagonal.
  - $\text{vec}(x_i)$: a vector with entries $x_1, \ldots, x_N$.

Identity matrices:

- $\mathbf{A}$ is an *identity* matrix if $a_{ij} = 1$ if $i = j$ and 0 if $i \neq j$.
  - It is a diagonal matrix with ones along the diagonal entries.
- For instance:

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

  is a 3×3 identity matrix.
- Note: We use notation $\mathbf{I}_K$ to denote a $K \times K$ identity matrix.
- Note: If context makes the value of K evident, we simplify the notation by dropping the subscript (i.e., $\mathbf{I}$ denotes the identity).

Matrices of ones:

- We use the notation $\mathbf{J}_N$ to denote an $N \times N$ matrix of all ones.
- For instance:

$$\mathbf{J}_3 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}.$$

- Similarly, $\mathbf{J}$ is a matrix of ones, such that the context will dictate the dimensions.

Notation and key definitions

Upper-triangular matrices:

- $\mathbf{A}$ is an *upper-triangular* if $a_{ij} = 0$ for $i > j$.
  - The entries "below" the diagonal are zero.
- For instance:

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 3 \\ 0 & 3 & -1 \\ 0 & 0 & 1 \end{bmatrix}$$

is an example of an upper triangular matrix. Importantly, it is not required for the entries in the upper triangle to be non-zero for the matrix to qualify as an upper-triangular matrix.

Notation and key definitions

Lower-triangular matrices:

- $\mathbf{A}$ is a *lower-triangular* matrix if $a_{ij} = 0$ for $i < j$.
  - The entries "above" the diagonal are zero.
- For instance:

$$\mathbf{A} = \begin{bmatrix} 2 & 0 & 0 \\ 2 & -2 & 0 \\ -1 & 3 & 2 \end{bmatrix}$$

  is an example of a lower-triangular matrix. Importantly, it is not required for the entries in the lower triangle to be non-zero for the matrix to qualify as an lower-triangular matrix.

Matrix addition:

- If **A** and **B** are of the same dimension, they may be added.
  - To add two matrices, you add element-wise.
- For instance:

$$\mathbf{A} = \begin{bmatrix} 1 & 4 & 1 \\ 3 & 0 & -1 \end{bmatrix} \text{ and } \mathbf{B} = \begin{bmatrix} 5 & -4 & 0 \\ -2 & 6 & -2 \end{bmatrix}.$$

Therefore,

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} 1+5 & 4+(-4) & 1+0 \\ 3+(-2) & 0+6 & -1+(-2) \end{bmatrix} = \begin{bmatrix} 6 & 0 & 1 \\ 1 & 6 & -3 \end{bmatrix}.$$

Matrix operations

Matrix addition: Properties

- Matrix addition is associative:

$$(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C}).$$

- Matrix addition is commutative:

$$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}.$$

Scalar multiplication:

- If **A** is any matrix, it can be multiplied by a scalar.
  - To multiply a matrix by a scalar, multiply element-wise.
- For instance:

$$\mathbf{A} = \begin{bmatrix} 1 & 4 & 1 \\ 3 & 0 & -1 \end{bmatrix} \text{ and } c = 2.$$

Therefore,

$$c\mathbf{A} = \begin{bmatrix} 2 \times 1 & 2 \times 4 & 2 \times 1 \\ 2 \times 3 & 2 \times 0 & 2 \times (-1) \end{bmatrix} = \begin{bmatrix} 2 & 8 & 2 \\ 6 & 0 & -2 \end{bmatrix}.$$

The dot product:

- If $\mathbf{x}$ and $\mathbf{y}$ are of length K, then $\mathbf{x} \cdot \mathbf{y} = \sum_{k=1}^{K} x_k y_k$.
  - The dot product is the sum of the element-wise products (well defined if vectors are of the same length).
- For instance:

$$\mathbf{x} = \begin{bmatrix} 1 \\ 4 \\ 1 \end{bmatrix} \text{ and } \mathbf{y} = \begin{bmatrix} 5 \\ -4 \\ 0 \end{bmatrix}.$$

  Therefore,

$$\mathbf{x} \cdot \mathbf{y} = (1 \times 5) + (4 \times -4) + (1 \times 0) = -11.$$

- Note: $\mathbf{x} \cdot \mathbf{y} = ||\mathbf{x}|| \, ||\mathbf{y}|| \cos\theta$, where θ marks the angle between $\mathbf{x}$ and $\mathbf{y}$.
- If $\mathbf{x}$ and $\mathbf{y}$ are orthogonal vectors, then $\mathbf{x} \cdot \mathbf{y} = 0$. Try proving this: both using the formula above and *not* using the formula above!

Matrix transposition:

- If $\mathbf{A}$ is an $N \times K$ matrix, its *transpose*, $\mathbf{A}^\intercal$ is a $K \times N$ matrix that reverses the role of the rows and columns of $\mathbf{A}$.
  - The (i, j) entry of $\mathbf{A}$ is the (j, i) entry of $\mathbf{A}^\intercal$.
- For instance:

$$\mathbf{A} = \begin{bmatrix} 1 & 4 & 1 \\ 3 & 0 & -1 \end{bmatrix} \Longrightarrow \mathbf{A}^\intercal = \begin{bmatrix} 1 & 3 \\ 4 & 0 \\ 1 & -1 \end{bmatrix}.$$

- Some sources will denote the transpose as $\mathbf{A}'$. I use the "$\intercal$"-notation just out of habit.

Matrix transposition: Properties

- $(\mathbf{A}^\intercal)^\intercal = \mathbf{A}$.
- $(\mathbf{A} + \mathbf{B})^\intercal = \mathbf{A}^\intercal + \mathbf{B}^\intercal$.
- $(\mathbf{AB})^\intercal = \mathbf{B}^\intercal \mathbf{A}^\intercal$.
- $(c\mathbf{A})^\intercal = c\mathbf{A}^\intercal$.
- $\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^\intercal \mathbf{y} = \mathbf{y}^\intercal \mathbf{x}$.
- $\mathbf{x} \cdot \mathbf{x} = \mathbf{x}^\intercal \mathbf{x} = \sum_{k=1}^{K} x_i^2 = ||\mathbf{x}||_2^2$.

A matrix times a vector:

- If $\mathbf{A}$ is an $N \times K$ and $\mathbf{x}$ is a vector of length K, then the product $\mathbf{Ax}$ is well defined.
  - The product $\mathbf{Ax}$ produces a vector of length N.
  - Elements of $\mathbf{x}$ inform you how to combine columns of $\mathbf{A}$.
- For instance:

$$\mathbf{A} = \begin{bmatrix} 1 & 4 & 1 \\ 3 & 0 & -1 \end{bmatrix} \text{ and } \mathbf{x} = \begin{bmatrix} 5 \\ -4 \\ 0 \end{bmatrix}.$$

Therefore,

$$\mathbf{Ax} = \left(5 \times \begin{bmatrix} 1 \\ 3 \end{bmatrix}\right) + \left(-4 \times \begin{bmatrix} 4 \\ 0 \end{bmatrix}\right) + \left(0 \times \begin{bmatrix} 1 \\ -1 \end{bmatrix}\right) = \begin{bmatrix} -11 \\ 15 \end{bmatrix}.$$

A matrix times a vector:

- In practice, we combine steps: the i^{th} element of $\mathbf{Ax}$ is given by the dot product of the i^{th} row of $\mathbf{A}$ with $\mathbf{x}$.
  - In other words, $[\mathbf{Ax}]_i = \mathbf{a}_{i\cdot}^\intercal \cdot \mathbf{x} = \mathbf{a}_{i\cdot}\mathbf{x}$.
- Mental image associated with first step in this operation:

$$\begin{bmatrix} \boxed{1 \quad 4 \quad 1} \\ 3 \quad 0 \quad -1 \end{bmatrix} \begin{bmatrix} 5 \\ -4 \\ 0 \end{bmatrix} = \begin{bmatrix} \boxed{-11} \\ ? \end{bmatrix}.$$

- Mental image associated with second step in this operation:

$$\begin{bmatrix} 1 \quad 4 \quad 1 \\ \boxed{3 \quad 0 \quad -1} \end{bmatrix} \begin{bmatrix} 5 \\ -4 \\ 0 \end{bmatrix} = \begin{bmatrix} -11 \\ \boxed{15} \end{bmatrix}.$$

A matrix times a matrix:

- Matrix product **AB** well defined if number of columns of **A** matches number of rows of **B**.
  - Each column obtained by multiplying **A** by columns of **B**.
- For instance:

$$\mathbf{A} = \begin{bmatrix} 1 & 4 & 1 \\ 3 & 0 & -1 \end{bmatrix} \text{ and } \mathbf{B} = \begin{bmatrix} 5 & -2 \\ -4 & 6 \\ 0 & -2 \end{bmatrix}.$$

Therefore,

$$[\mathbf{AB}]_{\cdot 1} = 5 \times \begin{bmatrix} 1 \\ 3 \end{bmatrix} + (-4) \times \begin{bmatrix} 4 \\ 0 \end{bmatrix} + (0) \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} -11 \\ 15 \end{bmatrix}, \text{ and}$$

$$[\mathbf{AB}]_{\cdot 2} = (-2) \times \begin{bmatrix} 1 \\ 3 \end{bmatrix} + 6 \times \begin{bmatrix} 4 \\ 0 \end{bmatrix} + (-2) \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 20 \\ -4 \end{bmatrix}.$$

A matrix times a matrix:

- We use the visual tool to obtain elements of **AB**.
- Mental image associated with first step in this operation:

$$\begin{bmatrix} \boxed{1 \quad 4 \quad 1} \\ 3 \quad 0 \quad -1 \end{bmatrix} \begin{bmatrix} \boxed{\begin{matrix} 5 \\ -4 \\ 0 \end{matrix}} & \begin{matrix} -2 \\ 6 \\ -2 \end{matrix} \end{bmatrix} = \begin{bmatrix} \boxed{-11} & ? \\ ? & ? \end{bmatrix}.$$

- . . . and so on . . .
- Mental image associated with fourth step in this operation:

$$\begin{bmatrix} 1 \quad 4 \quad 1 \\ \boxed{3 \quad 0 \quad -1} \end{bmatrix} \begin{bmatrix} \begin{matrix} 5 \\ -4 \\ 0 \end{matrix} & \boxed{\begin{matrix} -2 \\ 6 \\ -2 \end{matrix}} \end{bmatrix} = \begin{bmatrix} -11 & 20 \\ 15 & \boxed{-4} \end{bmatrix}.$$

A matrix times a matrix: Properties

- Matrix multiplication is associative:

$$(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC}).$$

- Matrix multiplication is distributive over matrix addition:

$$\mathbf{A}(\mathbf{B}+\mathbf{C}) = \mathbf{AB}+\mathbf{AC} \text{ and } (\mathbf{A}+\mathbf{B})\mathbf{C} = \mathbf{AC}+\mathbf{BC}.$$

- Matrix multiplication is not generally commutative:

$$\mathbf{AB} \neq \mathbf{BA} \text{ (generally)}.$$

A matrix's trace:

- For a square matrix, $\mathbf{A}$, its trace is given by $\text{tr}(\mathbf{A}) = \sum_{n=1}^{N} A_{nn}$.
  - That is, the trace is the sum of the entries on the main diagonal.
- For instance:

$$\mathbf{A} = \begin{bmatrix} 2 & -4 \\ 3 & -1 \end{bmatrix}.$$

Therefore,

$$\text{tr}(\mathbf{A}) = 2 + (-1) = 1.$$

A matrix's trace: Properties

- $\text{tr}(\mathbf{A}) = \text{tr}(\mathbf{A}^\intercal)$.
- $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$.
- $\text{tr}(\mathbf{A} + \mathbf{B}) = \text{tr}(\mathbf{A}) + \text{tr}(\mathbf{B})$.
- $\text{tr}(\mathbf{ABCD}) = \text{tr}(\mathbf{BCDA}) = \text{tr}(\mathbf{CDAB}) = \text{tr}(\mathbf{DABC})$.
  - In general, trace is invariant under "cyclic permutations."
    - Note that I didn't include, for instance, $\text{tr}(\mathbf{ADCB})$ in the equality above.
  - However, if all matrices are symmetric, we can strengthen this result to say that the trace is invariant to *any* permutation of the matrix product—not just the cyclic ones.
- $\text{tr}(\mathbf{x}\mathbf{x}^\intercal) = \mathbf{x}^\intercal\mathbf{x}$.
  - $\mathbf{x}^\intercal\mathbf{x} = \langle \mathbf{x}, \mathbf{x} \rangle$ is referred to as an "inner" product, producing a scalar.
  - $\mathbf{x}\mathbf{x}^\intercal = \mathbf{x} \otimes \mathbf{x}$ is referred to as an "outer" product, producing a matrix.

Table of Contents

Gaussian elimination:

- Suppose we seek to solve $\mathbf{Ax} = \mathbf{c}$ for $\mathbf{x}$ (a *very* convenient matrix notation to represent a system of N equations with K unknowns).
  - The proper way to think about this from the standpoint of linear algebra is that $\mathbf{x}$ is telling you the correct linear combination (if one exists) of the *columns* of $\mathbf{A}$ that produce the vector $\mathbf{c}$.
- The typical way we learn how to do this is through a process called Gaussian elimination, sequentially scaling, switching, and adding multiples of rows from the augmented matrix $\begin{bmatrix}\mathbf{A} & \mathbf{c}\end{bmatrix}$ until $\mathbf{A}$ has been transformed into an upper-triangular matrix.
- We may then back-substitute to learn the values of $\mathbf{x}$.

Gaussian elimination: Example

- Equations:

$$\begin{aligned} x + 2y + z &= 2 \\ 3x + 8y + z &= 12 \\ 4y + z &= 2 \end{aligned}$$

- From "good" to "better": keep variables aligned and refer to them as x_1, x_2, and x_3:

$$\begin{aligned} x_1 + 2x_2 + x_3 &= 2 \\ 3x_1 + 8x_2 + x_3 &= 12 \\ 0x_1 + 4x_2 + x_3 &= 2 \end{aligned}$$

Elimination with matrices

Gaussian elimination: Example

- From "good" to "better": keep variables aligned.

$$\begin{aligned} x_1 + 2x_2 + x_3 &= 2 \\ 3x_1 + 8x_2 + x_3 &= 12 \\ 0x_1 + 4x_2 + x_3 &= 2 \end{aligned}$$

- From "better" to "best": matrix notation ($\mathbf{Ax} = \mathbf{c}$).

$$\begin{bmatrix} 1 & 2 & 1 \\ 3 & 8 & 1 \\ 0 & 4 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 12 \\ 2 \end{bmatrix}$$

Elimination with matrices

Gaussian elimination: Example

- Augmented matrix $\begin{bmatrix}\mathbf{A} & \mathbf{c}\end{bmatrix}$ includes a vertical slash for clarity:

$$\left[\begin{array}{ccc|c} 1 & 2 & 1 & 2 \\ 3 & 8 & 1 & 12 \\ 0 & 4 & 1 & 2 \end{array}\right]$$

- Next step: want a 0 in the (2,1) position.
- Natural choice: subtract three of row 1 from row 2.
- Proper way to think about this from a linear algebra standpoint: multiply both sides by 3×3 matrix $\mathbf{E}_{21}$ that leaves rows 1 and 3 unchanged but subtracts three of row 1 from row 2:

$$\mathbf{E}_{21} = \begin{bmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Gaussian elimination: Example

- Applying the desired operation $\mathbf{E}_{21}\begin{bmatrix}\mathbf{A} & \mathbf{c}\end{bmatrix}$:

$$\left[\begin{array}{ccc|c} 1 & 2 & 1 & 2 \\ 0 & 2 & -2 & 6 \\ 0 & 4 & 1 & 2 \end{array}\right]$$

- Next step: want 0 in the (3,2) position.
- Natural choice: subtract two of row 2 from row 3.
- Proper way to think about this from a linear algebra standpoint: multiply both sides by 3×3 matrix $\mathbf{E}_{32}$ that leaves rows 1 and 2 unchanged but subtracts two of row 2 from row 3:

$$\mathbf{E}_{32} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -2 & 1 \end{bmatrix}.$$

Gaussian elimination: Example

- Applying the desired operation $\mathbf{E}_{32}\mathbf{E}_{21}\begin{bmatrix}\mathbf{A} & \mathbf{c}\end{bmatrix}$:

$$\left[\begin{array}{ccc|c} 1 & 2 & 1 & 2 \\ 0 & 2 & -2 & 6 \\ 0 & 0 & 5 & -10 \end{array}\right]$$

- The algorithm is complete and we use back-substitution to find:

$$\mathbf{x} = \begin{bmatrix} 2 \\ 1 \\ -2 \end{bmatrix}$$

- If the rows (equations) aren't in the "right" order, may need to switch rows (equivalent to multiplying by permutation matrix) to get to the right place.

Gaussian elimination: Properties in this nice case

- In this special case, the matrix was placed in row echelon form.
- $\mathbf{A}$ had three "pivots" (and is thus characterized as *rank three*).
- Can be read as the first non-zero column of each row in echelon form.
- Though we haven't yet discussed the determinant, it is equal to the product of the pivots ($\det(\mathbf{A}) = 1 \times 2 \times 5 = 10$).
- Note that the echelon form is upper-triangular.
- Though we haven't yet discussed inverses, note that $(\mathbf{E}_{32}\mathbf{E}_{21})^{-1}$ is lower-triangular:

$$(\mathbf{E}_{32}\mathbf{E}_{21})^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 6 & 2 & 1 \end{bmatrix}$$

- This is the idea behind the well-known decomposition $\mathbf{A} = \mathbf{LU}$.

More about the rank:

- $\text{rank}(\mathbf{A}) = \#$ of pivots in row echelon form (zeros can be in a pivot position, but a pivot cannot be zero).
- A square ($N \times N$) matrix is said to be of full rank if it has N pivots.
- For any matrix $\mathbf{A}$,
  - $\text{rank}(\mathbf{A}) \leq N$ (number of rows).
  - $\text{rank}(\mathbf{A}) \leq K$ (number of columns).
  - $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}^\intercal) = \text{rank}(\mathbf{A}\mathbf{A}^\intercal) = \text{rank}(\mathbf{A}^\intercal\mathbf{A})$.
- If the matrix product $\mathbf{AB}$ is well defined, then the following are true:
  - $\text{rank}(\mathbf{AB}) \leq \text{rank}(\mathbf{A})$.
  - $\text{rank}(\mathbf{AB}) \leq \text{rank}(\mathbf{B})$.
  - $\text{rank}(\mathbf{AB}) = \text{rank}(\mathbf{A})$ if $\mathbf{B}$ is full-rank.
- If the matrix sum $\mathbf{A} + \mathbf{B}$ is well defined, then the following is true:
  - $\text{rank}(\mathbf{A} + \mathbf{B}) \leq \text{rank}(\mathbf{A}) + \text{rank}(\mathbf{B})$.

More about the rank:

- To put some of these ideas more concretely/elegantly, define the augmented matrix $\mathbf{A}_{\text{aug}} = \begin{bmatrix} \mathbf{A} & \mathbf{c} \end{bmatrix}$.
- Then,
  1. If $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}_{\text{aug}}) = K$, there is a unique solution to $\mathbf{Ax} = \mathbf{c}$.
  2. If $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}_{\text{aug}}) < K$, there are many solutions to $\mathbf{Ax} = \mathbf{c}$.
  3. If $\text{rank}(\mathbf{A}) < \text{rank}(\mathbf{A}_{\text{aug}})$, there are no solutions to $\mathbf{Ax} = \mathbf{c}$.

Table of Contents

Key ideas: Let $\mathbf{v}_1, \ldots, \mathbf{v}_K \in \mathbb{R}^N$

- $\text{span}(\mathbf{v}_1, \ldots, \mathbf{v}_K) = \sum_{k=1}^{K} a_k \mathbf{v}_k$ (all possible linear combinations).
- Vectors said to be linearly dependent if there exist $c_1, \ldots, c_K$ (not all zero) such that

$$\sum_{k=1}^{K} c_k \mathbf{v}_k = \mathbf{0},$$

  and are otherwise said to be linearly independent.

Key ideas: Let $\mathbf{v}_1, \ldots, \mathbf{v}_K \in \mathbb{R}^N$

- If $\mathbf{v}_1, \ldots, \mathbf{v}_K$ are linearly independent, each $\mathbf{y} \in \mathbb{R}^N$ can be written in the form for some values $a_1, \ldots, a_K$:

$$\mathbf{y} \; = \; \sum_{k=1}^{K} a_k \mathbf{v}_k,$$

- Linearly independent vectors that *span* a vector space in this way are said to form a *basis* for that vector space.

Notation: Direct sum

- If $\mathcal{B}_1$ and $\mathcal{B}_2$ are vector spaces, then the direct sum, given by $\mathcal{B}_3 = \mathcal{B}_1 \oplus \mathcal{B}_2$, comprises all vectors of the form

$$\mathbf{b}_3 = c_1\mathbf{b}_1 + c_2\mathbf{b}_2,$$

where $\mathbf{b}_1 \in \mathcal{B}_1$ and $\mathbf{b}_2 \in \mathcal{B}_2$, and $c_1, c_2 \in \mathbb{R}$.

Understanding the key linear subspaces

Motivation:

- Square matrices represent the setting in which the number of equations and number of unknown variables match, which is what gives us the hope of finding a unique solution to $\mathbf{Ax} = \mathbf{c}$ for any $\mathbf{c}$.
  - $N > K$: number of equations exceeds number of unknowns.
  - $N < K$: number of unknowns exceeds number of equations.
- Of course, sometimes the equations in a system can represent either redundant or inconsistent information.

Example 18.1: Redundancy even in the case of $N = K$

- Despite the fact that there are three equations and three unknowns below, only two of the equations contribute unique information:

$$\begin{aligned} 2x_1 + 3x_2 + 5x_3 &= 0 \\ 2x_1 + 2x_2 + 4x_3 &= 0 \\ 2x_1 + 4x_2 + 6x_3 &= 0 \end{aligned}$$

- The "redundancy" in this example is clearer by looking at the columns of the associated matrix:

$$\mathbf{A} = \begin{bmatrix} 2 & 3 & 5 \\ 2 & 2 & 4 \\ 2 & 4 & 6 \end{bmatrix}$$

Example 18.1: Redundancy even in the case of $N = K$

- Matrix form:

$$\mathbf{A} = \begin{bmatrix} 2 & 3 & 5 \\ 2 & 2 & 4 \\ 2 & 4 & 6 \end{bmatrix}$$

- The third column is the sum of the other two.
- Very importantly: rank($\mathbf{A}$) marks *both* the number of linearly independent rows *and* the number of linearly independent columns.
- Therefore, I know that any row can be written as a linear combination of the other two without having to do the calculation.

Example 18.1: Redundancy even in the case of $N = K$

- The linear dependence of the columns (or equivalently, rows) will make itself apparent when reducing to row echelon form:

$$\begin{bmatrix} 2 & 3 & 5 \\ 2 & 2 & 4 \\ 2 & 4 & 6 \end{bmatrix} \xrightarrow{\mathbf{E}_{21}} \begin{bmatrix} 2 & 3 & 5 \\ 0 & -1 & -1 \\ 0 & 1 & 1 \end{bmatrix} \xrightarrow{\mathbf{E}_{32}} \begin{bmatrix} 2 & 3 & 5 \\ 0 & -1 & -1 \\ 0 & 0 & 0 \end{bmatrix}$$

- We learn ALL of the following:
  - The pivots of $\mathbf{A}$ are 2 and -1 ($\text{rank}(\mathbf{A}) = 2$).
  - There are two linearly independent columns.
  - There are two linearly independent rows.
  - There are infinitely many solutions to the equation $\mathbf{Ax} = \mathbf{0}$.
  - The values $\mathbf{c}$ for which $\mathbf{Ax} = \mathbf{c}$ can be expressed as a linear combination of any two linearly independent columns of $\mathbf{A}$ (that is, any two linearly independent columns of $\mathbf{A}$ form a *basis* for its column space).

The row space:

- The *row space* of $\mathbf{A}$, denoted by $\mathcal{R}(\mathbf{A})$, is defined as:

$$\mathcal{R}(\mathbf{A}) = \left\{ \mathbf{y} \in \mathbb{R}^K : \exists \mathbf{x} \in \mathbb{R}^N \text{ with } \mathbf{y} = \mathbf{A}^\intercal \mathbf{x} \right\}.$$

- Characterizations:
  - The space *spanned* by the rows of $\mathbf{A}$.
  - Note that in my notation above, I'm avoiding row vectors by noting that the row space of $\mathbf{A}$ is simply the "column space" of $\mathbf{A}^\intercal$.
- Properties:
  - $\mathbf{0}_K \in \mathcal{R}(\mathbf{A})$.
  - If $\mathbf{v}_1 \in \mathcal{R}(\mathbf{A})$ and $\mathbf{v}_2 \in \mathcal{R}(\mathbf{A})$, then $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 \in \mathcal{R}(\mathbf{A})$.
  - $\dim(\mathcal{R}(\mathbf{A})) = \text{rank}(\mathbf{A}) = \#$ pivots in echelon form.

The null space:

- The *null space* of $\mathbf{A}$, denoted by $\mathcal{N}(\mathbf{A})$, is defined as:

$$\mathcal{N}(\mathbf{A}) = \left\{ \mathbf{x} \in \mathbb{R}^K : \mathbf{A}\mathbf{x} = \mathbf{0}_N \right\}.$$

- Characterizations:
  - The set of vectors that vanish when $\mathbf{A}$ is "applied" to them.
- Properties:
  - $\mathbf{0}_K \in \mathcal{N}(\mathbf{A})$.
  - If $\mathbf{v}_1 \in \mathcal{N}(\mathbf{A})$ and $\mathbf{v}_2 \in \mathcal{N}(\mathbf{A})$, then $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 \in \mathcal{N}(\mathbf{A})$.
  - If $\mathcal{N}(\mathbf{A})$ *only* includes $\mathbf{0}_K$, then $\mathbf{A}$ is invertible (meaning that $\mathbf{A}\mathbf{x} = \mathbf{c}$ has a solution for every $\mathbf{c}$).
  - Every vector $\mathbf{v}_1 \in \mathcal{N}(\mathbf{A})$ is orthogonal to every vector $\mathbf{v}_2 \in \mathcal{R}(\mathbf{A})$. This, together with the fact that $\mathcal{R}(\mathbf{A}) \oplus \mathcal{N}(\mathbf{A}) = \mathbb{R}^K$, is the reason we say that the null space and row space are *orthogonal complements* ($\mathcal{N}(\mathbf{A}) = [\mathcal{R}(\mathbf{A})]^\perp$).

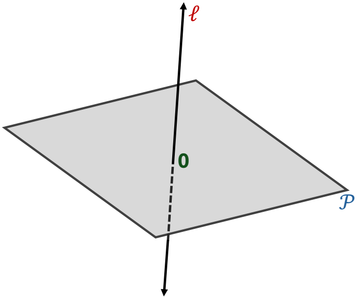
The null space and the row space: Elaboration

- Key statement on the previous slide: "Every vector in a matrix's null space is orthogonal to every vector in its row space."
- To show this, suppose $\mathbf{v}_1 \in \mathcal{N}(\mathbf{A})$. Then, by definition, $\mathbf{A}\mathbf{v}_1 = \mathbf{0}$, which is to say:

$$\mathbf{a}_{1\cdot}^{\mathsf{T}}\mathbf{v}_1 = 0 \text{ and } \cdots \text{ and } \mathbf{a}_{N\cdot}^{\mathsf{T}}\mathbf{v}_1 = 0.$$

- If $\mathbf{v}_2 \in \mathcal{R}(\mathbf{A})$, then $\mathbf{v}_2 = \mathbf{c}^{\mathsf{T}}\mathbf{A} = c_1\mathbf{a}_{1\cdot} + \cdots + c_N\mathbf{a}_{N\cdot}$ for some $\mathbf{c}$.
- Then, $\mathbf{v}_2^{\mathsf{T}}\mathbf{v}_1 = c_1\mathbf{a}_{1\cdot}^{\mathsf{T}}\mathbf{v}_1 + \cdots + c_N\mathbf{a}_{N\cdot}^{\mathsf{T}}\mathbf{v}_1 = 0$, which is to say that $\mathbf{v}_1$ and $\mathbf{v}_2$ are orthogonal.

The null space and row space: Orthogonal complements (in $\mathbb{R}^3$)



- If $\mathcal{N}(\mathbf{A})$ is the plane, $\mathcal{P}$, then $\mathcal{R}(\mathbf{A})$ is the line, ℓ—and vice versa.
- If $\mathcal{N}(\mathbf{A})$ is the origin, $\mathbf{0}$, then $\mathcal{R}(\mathbf{A})$ is $\mathbb{R}^3$—and vice versa.

The left null space:

- The *left null space* of $\mathbf{A}$, denoted by $\mathcal{N}(\mathbf{A}^\intercal)$, is defined as:

$$\mathcal{N}(\mathbf{A}^\intercal) = \left\{ \mathbf{x} \in \mathbb{R}^N : \mathbf{x}^\intercal \mathbf{A} = \mathbf{0}_K \right\}.$$

- Characterizations:
  - The set of vectors that vanish when "applied to" $\mathbf{A}$.
  - The term "left null space" comes from the fact that the $\mathbf{x}$ is on the left rather than on the right.
  - The notation $\mathcal{N}(\mathbf{A}^\intercal)$ come from the idea that the left null space of $\mathbf{A}$ is simply the null space of $\mathbf{A}^\intercal$.
- Properties:
  - $\mathbf{0}_N \in \mathcal{N}(\mathbf{A}^\intercal)$.
  - If $\mathbf{v}_1 \in \mathcal{N}(\mathbf{A}^\intercal)$ and $\mathbf{v}_2 \in \mathcal{N}(\mathbf{A}^\intercal)$, then $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 \in \mathcal{N}(\mathbf{A}^\intercal)$.

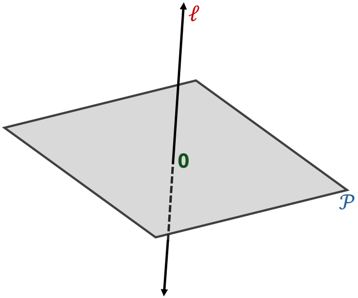
The column space:

- The *column space* of $\mathbf{A}$, denoted by $\mathcal{C}(\mathbf{A})$, is defined as:

$$\mathcal{C}(\mathbf{A}) = \left\{ \mathbf{y} \in \mathbb{R}^N : \exists \mathbf{x} \in \mathbb{R}^K \text{ with } \mathbf{y} = \mathbf{A}\mathbf{x} \right\}.$$

- Characterizations:
  - The space *spanned* by the columns of $\mathbf{A}$.
  - Characterizes all vectors $\mathbf{c}$ for which $\mathbf{A}\mathbf{x} = \mathbf{c}$ has a solution.
- Properties:
  - $\mathbf{0}_N \in \mathcal{C}(\mathbf{A})$
  - If $\mathbf{v}_1 \in \mathcal{C}(\mathbf{A})$ and $\mathbf{v}_2 \in \mathcal{C}(\mathbf{A})$, then $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 \in \mathcal{C}(\mathbf{A})$.
  - $\dim(\mathcal{C}(\mathbf{A})) = \text{rank}(\mathbf{A}) = \#$ pivots in echelon form.
  - $\dim(\mathcal{N}(\mathbf{A})) + \dim(\mathcal{C}(\mathbf{A})) = K = \#$ columns in $\mathbf{A}$.
  - Every vector $\mathbf{v}_1 \in \mathcal{C}(\mathbf{A})$ is orthogonal to every vector $\mathbf{v}_2 \in \mathcal{N}(\mathbf{A}^\intercal)$. This, together with the fact that $\mathcal{C}(\mathbf{A}) \oplus \mathcal{N}(\mathbf{A}^\intercal) = \mathbb{R}^N$, is the reason we say that the left null space and column space are *orthogonal complements* ($\mathcal{N}(\mathbf{A}^\intercal) = [\mathcal{C}(\mathbf{A})]^\perp$).

The left null space and column space: Orthogonal complements (in $\mathbb{R}^3$)



- If $\mathcal{N}(\mathbf{A}^\intercal)$ is the plane, $\mathcal{P}$, then $\mathcal{C}(\mathbf{A})$ is the line, ℓ—and vice versa.
- If $\mathcal{N}(\mathbf{A}^\intercal)$ is the origin, $\mathbf{0}$, then $\mathcal{C}(\mathbf{A})$ is $\mathbb{R}^3$—and vice versa.

Return to prior example:

- Equations:

$$\begin{bmatrix} 2 & 3 & 5 \\ 2 & 2 & 4 \\ 2 & 4 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \mathbf{0}.$$

- We found that $\text{rank}(\mathbf{A}) = 2$. From the prior properties, we know:
  - $\mathcal{C}(\mathbf{A})$ is a plane in $\mathbb{R}^3$ through the origin.
    - Specifically, the plane is characterized by the span of any two of the columns of **A** and has equation $4x_1 - 2x_2 - 2x_3 = 0$, which I learned by taking the cross product of the first two linearly independent column vectors.
  - $\mathcal{N}(\mathbf{A})$ is a line in $\mathbb{R}^3$ through the origin.
    - Specifically, the line is spanned by the vector $(1, 1, -1)^\intercal$, which I learned from the rank-nullity theorem and by recognizing this as a vector in the null space.

Example 18.2: Linear models

- We often find ourselves in the case where $N > K$, with more equations than unknowns:

$$\begin{aligned} x_1 + 2x_2 &= 4 \\ x_1 + 3x_2 &= 3 \\ x_1 + 4x_2 &= 6 \\ x_1 + 3x_2 &= 5 \\ x_1 + 1x_2 &= 2 \end{aligned}$$

- In this case, the equations are inconsistent and the most we can hope for is a "best" solution for x_1 and x_2 (which are playing the role of the intercept and a slope in a simple linear regression model).
- It is extraordinarily important to unify these ideas with properties regarding matrix-related subspaces.

Table of Contents

A matrix's inverse:

- A square matrix $\mathbf{A}$ is invertible if there is exists a matrix $\mathbf{A}^{-1}$ such that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$.
- An invertible matrix is also referred to as a non-singular matrix.
- For instance, if

$$\mathbf{A} = \begin{bmatrix} 2 & -2 \\ 1 & 4 \end{bmatrix}.$$

you can verify that

$$\mathbf{A}^{-1} = \begin{bmatrix} 0.4 & 0.2 \\ -0.1 & 0.2 \end{bmatrix}.$$

Algorithm:

- You can attempt to find a matrix's inverse via the Gauss-Jordan algorithm, which extends the Gaussian elimination algorithm.
- Idea: Take augmented matrix $\begin{bmatrix}\mathbf{A} & \mathbf{I}\end{bmatrix}$ and apply the steps of elimination–but keep going until you can get $\mathbf{A}$ into the form $\mathbf{I}$.
- If this is not possible, $\mathbf{A}$ is not invertible. But if it is, then the matrix to which the identity transforms is nothing other than $\mathbf{A}^{-1}$.

Matrix inverses

Gauss-Jordan: Example

- Let's take one of our prior examples:

$$\left[\begin{array}{ccc|ccc} 1 & 2 & 1 & 1 & 0 & 0 \\ 3 & 8 & 1 & 0 & 1 & 0 \\ 0 & 4 & 1 & 0 & 0 & 1 \end{array}\right]$$

- Step 1 from elimination: subtract three of row 1 from row 2.
- Step 2 from elimination: subtract two of row 2 from row 3.
- Result:

$$\left[\begin{array}{ccc|ccc} 1 & 2 & 1 & 1 & 0 & 0 \\ 0 & 2 & -2 & -3 & 1 & 0 \\ 0 & 0 & 5 & 6 & -2 & 1 \end{array}\right]$$

- Gauss says stop, but Jordan says "keep going."

Gauss-Jordan: Example

- New starting point:

$$\left[\begin{array}{ccc|ccc} 1 & 2 & 1 & 1 & 0 & 0 \\ 0 & 2 & -2 & -3 & 1 & 0 \\ 0 & 0 & 5 & 6 & -2 & 1 \end{array}\right]$$

- Step 3: Take one of row 2 away from row 1.

$$\left[\begin{array}{ccc|ccc} 1 & 0 & 3 & 4 & -1 & 0 \\ 0 & 2 & -2 & -3 & 1 & 0 \\ 0 & 0 & 5 & 6 & -2 & 1 \end{array}\right]$$

Gauss-Jordan: Example

- Not done yet; keep going!

$$\left[\begin{array}{ccc|ccc} 1 & 0 & 3 & 4 & -1 & 0 \\ 0 & 2 & -2 & -3 & 1 & 0 \\ 0 & 0 & 5 & 6 & -2 & 1 \end{array}\right]$$

- Step 4: Take 3/5 of row 3 away from row 1.

$$\left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 4-18/5 & -1+6/5 & -3/5 \\ 0 & 2 & -2 & -3 & 1 & 0 \\ 0 & 0 & 5 & 6 & -2 & 1 \end{array}\right]$$

Gauss-Jordan: Example

- Not done yet; keep going!

$$\left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 4-18/5 & -1+6/5 & -3/5 \\ 0 & 2 & -2 & -3 & 1 & 0 \\ 0 & 0 & 5 & 6 & -2 & 1 \end{array}\right]$$

- Step 5: Add 2/5 of row 3 to row 2.

$$\left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 4-18/5 & -1+6/5 & -3/5 \\ 0 & 2 & 0 & -3+12/5 & 1-4/5 & 2/5 \\ 0 & 0 & 5 & 6 & -2 & 1 \end{array}\right]$$

Gauss-Jordan: Example

- Not done yet; keep going!

$$\left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 4-18/5 & -1+6/5 & -3/5 \\ 0 & 2 & 0 & -3+12/5 & 1-4/5 & 2/5 \\ 0 & 0 & 5 & 6 & -2 & 1 \end{array}\right]$$

- Steps 6 and 7: Divide row 2 by 2 and row 5 by 5

$$\left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 4-18/5 & -1+6/5 & -3/5 \\ 0 & 1 & 0 & -3/2+12/10 & 1/2-4/10 & 2/10 \\ 0 & 0 & 1 & 6/5 & -2/5 & 1/5 \end{array}\right]$$

- Cleaning up:

$$\mathbf{A}^{-1} = \begin{bmatrix} 2/5 & 1/5 & -3/5 \\ -3/10 & 1/10 & 1/5 \\ 6/5 & -2/5 & 1/5 \end{bmatrix}.$$

Gauss-Jordan: Example

- We can use this to verify that our prior solution to $\mathbf{Ax} = \mathbf{c}$ was correct when $\mathbf{c} = (2, 12, 2)$:

$$\mathbf{x} \;=\; \mathbf{A}^{-1}\mathbf{c} = \begin{bmatrix} 2/5 & 1/5 & -3/5 \\ -3/10 & 1/10 & 1/5 \\ 6/5 & -2/5 & 1/5 \end{bmatrix} \begin{bmatrix} 2 \\ 12 \\ 2 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ -2 \end{bmatrix}.$$

- We can also use this to find solutions involving other values of $\mathbf{c}$ without having to cycle through the elimination steps again.

A matrix's inverse: Meaning and properties

- Meaning: If $\mathbf{A}$ $(N \times N)$ is non-singular (invertible), then...
  - The equation $\mathbf{A}\mathbf{x} = \mathbf{c}$ can be solved for $\mathbf{x}$ for each $\mathbf{c} \in \mathbb{R}^N$ and those solutions are unique ($\mathbf{x} = \mathbf{A}^{-1}\mathbf{c}$).
  - $\mathcal{C}(\mathbf{A}) = \mathbb{R}^N$.
  - $\mathcal{N}(\mathbf{A}) = \{\mathbf{0}\}$.
  - $\mathbf{A}$ has N pivots.
  - $\mathbf{A}$ is of full rank (that is, $\text{rank}(\mathbf{A}) = N$).
- Properties:
  - If $\mathbf{A}$ and $\mathbf{B}$ are invertible $N \times N$ matrices, then $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$
  - $(c\mathbf{A})^{-1} = c^{-1}(\mathbf{A}^{-1})$.
  - $(\mathbf{A}^\intercal)^{-1} = (\mathbf{A}^{-1})^\intercal$.
- Notation:
  - $\mathbf{A}^{-\intercal} = (\mathbf{A}^{-1})^\intercal = (\mathbf{A}^\intercal)^{-1}$ (combining notation for convenience).

Table of Contents

Symmetric matrices:

- A square matrix, $\mathbf{A}$ is said to be *symmetric* if $a_{ij} = a_{ji}$.
  - Or, to put it another way, $\mathbf{A}^\intercal = \mathbf{A}$, meaning that reversing the roles of the rows and columns does not alter a symmetric matrix.
- For instance:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 5 & -1 \\ 3 & -1 & 0 \end{bmatrix}$$

  is an example of a symmetric matrix.
- Example: $\mathbf{A}^\intercal\mathbf{A}$ is a symmetric matrix.
- Example: For $\mathbf{v} \in \mathbb{R}^N$, $\mathbf{v}\mathbf{v}^\intercal$ is a symmetric $N \times N$ matrix (referred to as the outer product, just as $\mathbf{v}^\intercal\mathbf{v}$ is referred to as the inner product).

Very special matrices

Symmetric matrices:

- Matrices of the form $\mathbf{A}^\top \mathbf{A}$ (or $\mathbf{A}\mathbf{A}^\top$) are generally very special in linear algebra and you'll see this form all the time in this course.

Lemma 18.1: Property of $\mathbf{A}^\intercal \mathbf{A}$

$\mathbf{A}^\intercal \mathbf{A} = \mathbf{0}$ if and only if $\mathbf{A} = \mathbf{0}$.

Lemma 18.1: Proof

- If $\mathbf{A}^\intercal\mathbf{A} = \mathbf{0}$, then $\text{tr}(\mathbf{A}^\intercal\mathbf{A}) = 0$, which can only happen if $\mathbf{A} = \mathbf{0}$.
- The proof in the other direction in trivial.

Idempotent matrices:

- An *idempotent* matrix is one for which $\mathbf{A}^2 = \mathbf{AA} = \mathbf{A}$.
  - Multiplying the matrix by itself gives back the original matrix.
- For instance:

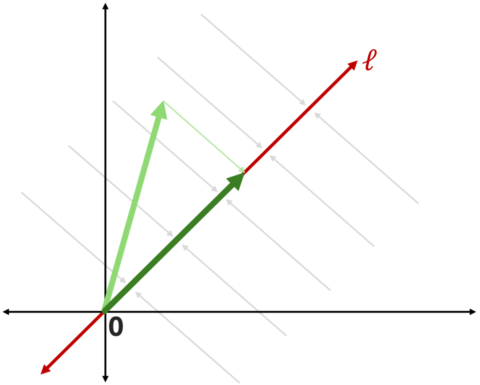
$$\mathbf{A} = \begin{bmatrix} 3 & -6 \\ 1 & -2 \end{bmatrix}$$

  is an example of an idempotent matrix.
- Note: Idempotent matrices must be square.
- Note: the rank of an idempotent matrix equals its trace.
- Note: an idempotent matrix can only have eigenvalues of zero and one (though we haven't yet discussed eigenvalues).

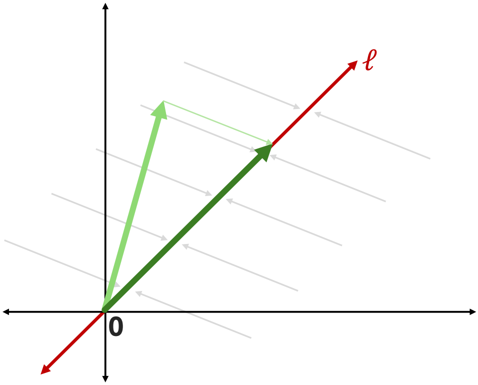
Projection matrices:

- An orthogonal projection matrix, $\mathbf{P}$, is symmetric and idempotent.
    - Non-symmetric idempotent matrices are *oblique* projection matrices.
    - Unless otherwise specified, any reference to projection matrices refers to orthogonal projection matrices in this course.
- If $\mathbf{P}$ is a projection matrix, then by idempotence $\text{rank}(\mathbf{P}) = \text{tr}(\mathbf{P})$.
- If $\mathbf{P}$ is a projection matrix, then by Lemma 18.1 $\mathbf{P}^\intercal\mathbf{P} = \mathbf{0}$ if and only if $\mathbf{P} = \mathbf{0}$.
- Matrices of the form $\mathbf{P} = \mathbf{A}(\mathbf{A}^\intercal\mathbf{A})^{-1}\mathbf{A}^\intercal$ are very important examples of a projection matrices in which $\mathbf{v} = \mathbf{P}\mathbf{x}$ marks the projection of vector $\mathbf{x}$ onto the subspace spanned by the columns of $\mathbf{A}$ (presumes $\mathbf{A}$ is of full column rank, in which case $\mathbf{A}^\intercal\mathbf{A}$ is invertible).

Orthogonal projections: Visualized in $\mathbb{R}^2$



Oblique projections: Visualized in $\mathbb{R}^2$



Projection matrices:

- For the time being, let me use $\mathbf{X}$ to denote my "matrix of interest" to follow familiar notation from regression modeling, and let $\mathbf{P} = \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal$ denote the projection matrix.
- Let's verify that $\mathbf{P}$ is a projection matrix. To do this, we need to check whether $\mathbf{P}$ is symmetric:

$$\begin{aligned}\mathbf{P}^\intercal &= (\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal)^\intercal \\ &= ((\mathbf{X}^\intercal)^\intercal(\mathbf{X}^\intercal\mathbf{X})^{-\intercal}(\mathbf{X})^\intercal) \\ &= (\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal) = \mathbf{P}.\end{aligned}$$

- I also need to check whether $\mathbf{P}$ is idempotent:

$$\begin{aligned}\mathbf{P}^2 &= \mathbf{P}\mathbf{P} = (\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal)(\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal) \\ &= \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}(\mathbf{X}^\intercal\mathbf{X})(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal \\ &= \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal = \mathbf{P}.\end{aligned}$$

Projection matrices:

- Let's further show that $\mathbf{I} - \mathbf{P} = \mathbf{I} - \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal$ is a projection matrix by verifying that it too is symmetric and idempotent:

$$\begin{aligned}(\mathbf{I} - \mathbf{P})^\intercal &= \mathbf{I}^\intercal - \mathbf{P}^\intercal \\ &= \mathbf{I} - \mathbf{P}.\end{aligned}$$

- Further,

$$\begin{aligned}(\mathbf{I} - \mathbf{P})^2 &= (\mathbf{I} - \mathbf{P})(\mathbf{I} - \mathbf{P}) \\ &= \mathbf{I}^2 - \mathbf{I}\mathbf{P} - \mathbf{P}\mathbf{I} + \mathbf{P}\mathbf{P} \\ &= \mathbf{I} - 2\mathbf{P} + \mathbf{P} \\ &= \mathbf{I} - \mathbf{P}.\end{aligned}$$

Projection matrices:

- Continuing our example, you should be able to show that $\mathbf{PX} = \mathbf{X}$ easily. What about $(\mathbf{I} - \mathbf{P})\mathbf{X}$?

$$\begin{aligned}(\mathbf{I} - \mathbf{P})\mathbf{X} &= \mathbf{X} - \mathbf{PX} \\ &= \mathbf{X} - \mathbf{X} = \mathbf{0}\end{aligned}$$

- How does this square with our intuition about what $\mathbf{P}$ is doing?
  - $\mathbf{P}$ projects vectors onto the space spanned by the columns of $\mathbf{X}$, while $\mathbf{I} - \mathbf{P}$ projects vectors onto the orthogonal complement of $\mathbf{X}$.
  - Projecting $\mathbf{X}$ onto its own column space should leave $\mathbf{X}$ unchanged. On the other hand, projecting the columns of $\mathbf{X}$ onto a space that is orthogonal to the columns of $\mathbf{X}$ should indeed give the zero vector.

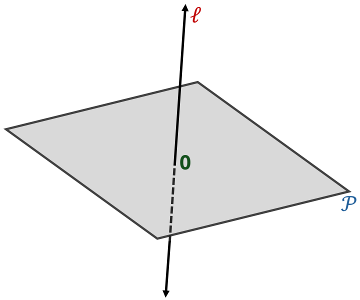
Projection matrices:

- As an example, let $\mathbf{P} = \frac{1}{2}\begin{bmatrix}1 & 1\\1 & 1\end{bmatrix}$, a projection matrix.
- $\mathbf{Px} = \frac{1}{2}\begin{bmatrix}x_1 + x_2\\x_1 + x_2\end{bmatrix}$. **P** projects vectors onto the line spanned by $(1, 1)^\intercal$.
- On the other hand,

$$\mathbf{I} - \mathbf{P} = \frac{1}{2}\begin{bmatrix}1 & -1\\-1 & 1\end{bmatrix}$$

- $(\mathbf{I} - \mathbf{P})\mathbf{x} = \frac{1}{2}\begin{bmatrix}x_1 - x_2\\x_2 - x_1\end{bmatrix}$. $\mathbf{I} - \mathbf{P}$ projects vectors onto the line spanned by $(1, -1)^\intercal$.

Orthogonal projections: $\mathbf{P}$ and $\mathbf{I} - \mathbf{P}$ (visualized in $\mathbb{R}^3$)



- If $\mathbf{P}$ projects onto the plane, $\mathcal{P}$, then $\mathbf{I} - \mathbf{P}$ projects onto the line, ℓ—and vice versa.
- If $\mathbf{P}$ projects onto $\mathbb{R}^3$ (implying $\mathbf{P} = \mathbf{I}$), then $\mathbf{P} - \mathbf{I} = \mathbf{0}$ sends everything to zero—and vice versa.

Projection matrices:

- More generally, $\mathbf{J}_N/N$, an $N \times N$ matrix with all entries given by $1/N$, is a projection matrix.
- Clearly, $\mathbf{J}_N$ is symmetric.
- To show idempotence, it is helpful to write $\mathbf{J} = \mathbf{1}_N\mathbf{1}_N^\intercal$, where $\mathbf{1}_N$ is a length-N vector of ones.
- Then,

$$
\begin{aligned}
\left(\frac{1}{N}\mathbf{J}_N\right)^2 &= \left(\frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\intercal\right)\left(\frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\intercal\right) \\
&= \frac{1}{N^2}\mathbf{1}_N(\mathbf{1}_N^\intercal\mathbf{1}_N)\mathbf{1}_N^\intercal = \frac{1}{N^2}\mathbf{1}_N(N)\mathbf{1}_N^\intercal \\
&= \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\intercal = \frac{1}{N}\mathbf{J}_N.
\end{aligned}
$$

Orthogonal matrices:

- A square matrix, $\mathbf{A}$, is said to be *orthogonal* if its columns are mutually orthogonal and each unit length (orthonormal).
- For instance:

$$\mathbf{A} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

is an example of an orthogonal matrix.

- Note: If $\mathbf{A}$ is orthogonal, then $\mathbf{A}^\intercal\mathbf{A} = \mathbf{I}$, which is another way of saying that $\mathbf{A}^\intercal = \mathbf{A}^{-1}$.
  - Diagonal entry k of $\mathbf{A}^\intercal\mathbf{A}$ is $||\mathbf{a}_{\cdot k}||^2 = 1$ and off-diagonal entry ij is a dot product of orthogonal vectors, $\mathbf{a}_{\cdot i}^\intercal\mathbf{a}_{\cdot j}$, so this is intuitive.
- Geometrically, these matrices are marked by rotations, reflections, rotoinversions, permutations (and their products).

Table of Contents

Determinants

A (square) matrix's determinant:

- The determinant of $\mathbf{A}$ is typically denoted by $\det(\mathbf{A})$ or $|\mathbf{A}|$.
- It is a quantity that shows up *repeatedly* in linear algebra.
- The value of a determinant is uniquely defined/identified through its possession of the following properties:
  1. The identity matrix has a determinant of one.
  2. Row exchanges reverse the sign of the determinant.
  3. The determinant has a "row-wise linearity," which is easiest to describe in the case of a 2×2 matrix:

$$\begin{vmatrix} ta & tb \\ c & d \end{vmatrix} = t\begin{vmatrix} a & b \\ c & d \end{vmatrix} \text{ and } \begin{vmatrix} a+a' & b+b' \\ c & d \end{vmatrix} = \begin{vmatrix} a & b \\ c & d \end{vmatrix} + \begin{vmatrix} a' & b' \\ c & d \end{vmatrix}$$

A matrix's determinant:

- You should be able to verify most of the below properties of the determinant of an $N \times N$ matrix $\mathbf{A}$:
  1. If $\mathbf{A}$ has two equal rows, $|\mathbf{A}| = 0$.
  2. If $\mathbf{A}$ has a row of zeros, $|\mathbf{A}| = 0$.
  3. Subtracting a multiple of row i from row k does not change $|\mathbf{A}|$.
  4. The determinant of an upper-triangular matrix is the product of the diagonal entries.
  5. $|\mathbf{A}| = 0$ if and only if $\mathbf{A}$ is singular.
  6. $|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}|$.
  7. If $\mathbf{A}$ is non-singular, $|\mathbf{A}^{-1}| = 1/|\mathbf{A}|$.
  8. $|c\mathbf{A}| = c^N|\mathbf{A}|$.
  9. $|\mathbf{A}^\mathsf{T}| = |\mathbf{A}|$.
  10. $|\mathbf{I} + \mathbf{x}\mathbf{y}^\mathsf{T}| = 1 + \mathbf{x}^\mathsf{T}\mathbf{y}$.
  11. An orthogonal matrix has a determinant of ± 1.

A matrix's determinant:

- The absolute value of the determinant can be conceptualized geometrically as vol(P), where

$$P = \left\{ c_1 \mathbf{A}_{\cdot 1} + \cdots + c_K \mathbf{A}_{\cdot K} \middle| 0 \leq c_i \leq 1 \; \forall i \right\}.$$

- This represents the volume of the parallelepiped with each of its sides being represented/implied by the columns of **A**.

DETERMINANTS

A matrix's determinant:

- There are many formulas and procedures to determine $|\mathbf{A}|$.
- In the 2×2 case, this is given by:

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc,$$

  which is absolutely necessary to commit to memory.

- In the general 3×3 case, we often break down:

$$\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = a \begin{vmatrix} e & f \\ h & i \end{vmatrix} - b \begin{vmatrix} d & f \\ g & i \end{vmatrix} + c \begin{vmatrix} d & e \\ g & h \end{vmatrix}$$

- Note that we could have gone "down a column" of **A** rather than "across a row" and we would still get the same answer.

Connection to inverses:

- If $\mathbf{A}$ is invertible, then:

$$\mathbf{A}^{-1} = \frac{1}{|\mathbf{A}|}\mathbf{C}^\intercal,$$

where $\mathbf{C}$ is the matrix of cofactors, and $\mathbf{C}^\intercal = \text{adj}(\mathbf{A})$ is referred to as the adjugate matrix.

- In the 2×2 case, we have that

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Connection to inverses:

- In the 3×3 case, we have that

$$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}^{-1} = \frac{1}{|\mathbf{A}|} \begin{bmatrix} +\begin{vmatrix} e & f \\ h & i \end{vmatrix} & -\begin{vmatrix} d & f \\ g & i \end{vmatrix} & +\begin{vmatrix} d & e \\ g & h \end{vmatrix} \\ -\begin{vmatrix} b & c \\ h & i \end{vmatrix} & +\begin{vmatrix} a & c \\ g & i \end{vmatrix} & -\begin{vmatrix} a & b \\ g & h \end{vmatrix} \\ +\begin{vmatrix} b & c \\ e & f \end{vmatrix} & -\begin{vmatrix} a & c \\ d & f \end{vmatrix} & +\begin{vmatrix} a & b \\ d & e \end{vmatrix} \end{bmatrix}$$

Table of Contents

Motivation:

- Often useful to motivate linear algebra problems geometrically.
- Suppose you want to find vectors $\mathbf{x}$ for which $\mathbf{y} = \mathbf{A}\mathbf{x} \in \mathcal{C}(\mathbf{A})$ is "in the same direction" as $\mathbf{x}$ (or in the reverse direction).
- Put algebraically, can we figure out which vectors $\mathbf{x}$ solve $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$?
- These vectors are called the *eigenvectors* of $\mathbf{A}$.
- The values of λ for which the above equation is satisfied are called the eigenvalues.
- Naturally, this concept only makes sense if $\mathbf{A}$ is square.

Ideas:

- If in fact $\mathbf{Ax} = \lambda\mathbf{x}$, then $\mathbf{Ax} - \lambda\mathbf{x} = (\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$.
- By convention, we don't consider $\mathbf{x} = \mathbf{0}$ to be an eigenvector of $\mathbf{A}$ (although λ can be zero—in fact, we already have a name for the eigenvectors for which $\lambda = 0$).
- Now, there is a solution $\mathbf{x} \neq \mathbf{0}$ exactly when the columns of $\mathbf{A} - \lambda\mathbf{I}$ are linearly dependent (or, $\mathbf{A} - \lambda\mathbf{I}$ is singular and has determinant zero).
- This gives us a clue on how to find the eigenvalues, which is to solve the *characteristic equation*, $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$, for λ. The left-hand side is called the characteristic polynomial (with degree given by the number of columns/rows).
- To then find the eigenvectors (i.e., the vectors $\mathbf{x} \in \mathcal{N}(\mathbf{A} - \lambda\mathbf{I})$), we plug in the eigenvalues into $\mathbf{Ax} = \lambda\mathbf{x}$ and solve for $\mathbf{x}$.

Example 18.3: Finding eigenvalues and eigenvectors

- Take the matrix $\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$ as an example.
- We have that $\mathbf{A} - \lambda\mathbf{I} = \begin{bmatrix} 3-\lambda & 1 \\ 1 & 3-\lambda \end{bmatrix}$.
- Now, $\det(\mathbf{A} - \lambda\mathbf{I}) = (3-\lambda)^2 - 1 = \lambda^2 - 6\lambda + 8 = (\lambda - 4)(\lambda - 2)$.
- This tells us that $\lambda = 2$ and $\lambda = 4$ are the eigenvalues of $\mathbf{A}$.
- To characterize the eigenvectors, we want to solve $(\mathbf{A} - 2\mathbf{I})\mathbf{x} = \mathbf{0}$ and $(\mathbf{A} - 4\mathbf{I})\mathbf{x} = \mathbf{0}$ for $\mathbf{x}$.

Example 18.3: Finding eigenvalues and eigenvectors

- Now, $\mathbf{A} - 2\mathbf{I} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$.
- We should be able to identify that this matrix will map any vector of the form $(c, -c)^\intercal$ to zero. We typically take $c = 1/\sqrt{2}$ so that the eigenvector is of length one.
- Next, $\mathbf{A} - 4\mathbf{I} = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}$.
- We should be able to identify that this matrix will map any vector of the form $(c, c)^\intercal$ to zero. We typically take $c = 1/\sqrt{2}$ so that the eigenvector is of length one.
- A lot of really nice things happened in this example that were, of course, no accident.

Properties: All of which check out with the prior example

- The sum of the eigenvalues of $\mathbf{A}$ is equal to $\text{tr}(\mathbf{A})$.
- The product of the eigenvalues of $\mathbf{A}$ is equal to $\det(\mathbf{A})$.
- Symmetric matrices in particular have real eigenvalues.
- Symmetric matrices in particular have orthogonal eigenvectors.
- Positive definite matrices have positive eigenvalues (although we have not talked about positive definite matrices yet).
- If λ is an eigenvalue of $\mathbf{A}$, then $\lambda + c$ is an eigenvalue of $\mathbf{A} + c\mathbf{I}$.

Lemma 18.2: Property of matrix products

If $\mathbf{A}$ and $\mathbf{B}$ are square matrices of the same dimension, then $\mathbf{AB}$ and $\mathbf{BA}$ have the same eigenvalues (that is, even if the matrix product does not commute).

Lemma 18.2: Proof

- Suppose λ is an eigenvalue of **AB**, with eigenvector **v**. Then,

$$\begin{aligned} (\mathbf{AB})\mathbf{v} &= \lambda\mathbf{v} \\ \Rightarrow \mathbf{B}(\mathbf{AB})\mathbf{v} &= \mathbf{B}(\lambda\mathbf{v}) \\ \Rightarrow (\mathbf{BA})(\mathbf{Bv}) &= \lambda(\mathbf{Bv}) \\ \Rightarrow (\mathbf{BA})(\mathbf{w}) &= \lambda\mathbf{w}, \end{aligned}$$

where $\mathbf{w} = \mathbf{Bv}$ is itself a vector, and in particular an eigenvector of **BA** with corresponding eigenvalue λ. The proof in the other direction is analogous.

More properties:

- If $\mathbf{A}$ is triangular, then the diagonal elements of $\mathbf{A}$ are the eigenvalues.
- If λ is an eigenvalue of an invertible matrix $\mathbf{A}$ with eigenvector $\mathbf{v}$, then λ^{-1} is an eigenvalue of $\mathbf{A}^{-1}$ with eigenvector $\mathbf{x}$.
- A matrix is invertible if and only if it has *only* nonzero eigenvalues.
- The eigenvalues of a matrix *can be complex*, but they come in pairs when they are (rest assured that considerations regarding complex eigenvalues are not the focus of this course).
- If an $N \times N$ matrix has N distinct eigenvalues, then there are N linearly independent eigenvectors.
- The eigenvalues of a matrix *can be repeated*. If the N eigenvalues of an $N \times N$ matrix are not distinct, you *may not* be able to find N linearly independent eigenvectors.

Example 18.4: Repeated eigenvalues

- Take the matrix $\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ as an example.
- We have that $\mathbf{A} - \lambda \mathbf{I} = \begin{bmatrix} 1-\lambda & 0 \\ 0 & 1-\lambda \end{bmatrix}$.
- Now, $\det(\mathbf{A} - \lambda \mathbf{I}) = (1-\lambda)^2$.
- This tells us that $\lambda = 1$ is the eigenvalue of $\mathbf{A}$, but it's been repeated.
- To characterize the eigenvectors, we want to solve $(\mathbf{A} - \mathbf{I})\mathbf{x} = \mathbf{0}$.
- Every vector is an eigenvector in this example.

Example 18.4: Repeated eigenvalues

- Take the matrix $\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ as an example.
- We have that $\mathbf{A} - \lambda \mathbf{I} = \begin{bmatrix} 1-\lambda & 1 \\ 0 & 1-\lambda \end{bmatrix}$.
- Now, $\det(\mathbf{A} - \lambda \mathbf{I}) = (1-\lambda)^2$.
- This tells us that $\lambda = 1$ is the eigenvalue of $\mathbf{A}$, but it's been repeated.
- To characterize the eigenvectors, we want to solve $(\mathbf{A} - \mathbf{I})\mathbf{x} = \mathbf{0}$.
- You should verify that the eigenvectors are all of the form $(c, 0)^\intercal$.
- In this case, the *algebraic* multiplicity of the eigenvalue does not match the *geometric* multiplicity, in which case we call $\mathbf{A}$ a defective matrix (which seems kind of mean).
- $\mathbf{A}$ is an example of a shear matrix, representing the addition of a multiple of one row/column to another. Shear matrices are defective.

Tying into projection matrices:

- If $\mathbf{P}$ is an $N \times N$ projection matrix of rank R, it has R eigenvalues of one and $N - R$ eigenvalues of zero.
- Further, there must be an orthogonal matrix $\mathbf{Q}$ such that

$$\mathbf{Q}^\intercal \mathbf{P} \mathbf{Q} = \mathbf{\Lambda},$$

  where $\mathbf{\Lambda} = \mathrm{diag}([\mathbf{1}_R, \mathbf{0}_{N-R}])$.
- This property has to do with the eigendecomposition, which we will discuss later.

Table of Contents

Reminder about symmetric matrices:

- Recall that all eigenvalues of a symmetric matrix are real and its eigenvectors are orthogonal.
  - In the case where every vector is an eigenvector, we can follow the convention of *choosing* the vectors to be orthogonal.
- When $\mathbf{A}$ is symmetric, it also happens that the signs of its pivots match the signs of its eigenvalues, which is computationally useful!

Positive definite matrices

Definition:

- A symmetric matrix $\mathbf{A}$ is said to be positive definite if for every $\mathbf{x} \neq \mathbf{0}$, we have that $\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} > 0$. Put another way, $\mathbf{A}$ is positive definite if $\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} = 0$ only when $\mathbf{x} = \mathbf{0}$.
- The quantity $\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}$ is known as a *quadratic form*.
- To see why, consider the 2×2 case:

$$\begin{aligned}\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} &= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} a_1 & a_2 \\ a_2 & a_3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} a_1 x_1 + a_2 x_2 \\ a_2 x_1 + a_3 x_2 \end{bmatrix} \\ &= a_1 x_1^2 + (2a_2) x_1 x_2 + a_3 x_2^2.\end{aligned}$$

- When $\mathbf{A}$ is positive definite, the graph of $y = \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}$ has minimum at the origin and "faces upward" as an elliptical paraboloid.
  - When $\mathbf{A}$ is symmetric but not positive definite, there are other geometries associated with the quadratic form (including the hyperbolic paraboloid and the parabolic cylinder).

Properties of positive definite matrices:

- If $\mathbf{A}$ is a positive definite matrix, then the following properties hold:
  1. $\mathbf{A}$ has all positive eigenvalues (you can use this as a test for positive definiteness).
  2. $\mathbf{A}$ has all positive pivots.
  3. $\mathbf{A}$ has a positive determinant (because the eigenvalues are all positive).
  4. $\mathbf{A}$ will rotate input vectors less than than 90 degrees.
     - ⋆ To see this, note that $\mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{x} = \mathbf{x} \cdot (\mathbf{A}\mathbf{x}) = ||\mathbf{x}|| \, ||\mathbf{A}\mathbf{x}|| \cos\theta > 0$.

Positive semi-definite matrices: Definition and properties

- A symmetric matrix $\mathbf{A}$ is said to be positive semi-definite if for every $\mathbf{x}$, we have that $\mathbf{x}^\intercal \mathbf{A} \mathbf{x} \geq 0$.
- Sometimes referred to as "non-negative definite."
- If $\mathbf{A}$ is positive semi-definite, then $\mathbf{A}$ has all non-negative eigenvalues.
- If $\mathbf{A}$ is positive definite, then $\mathbf{A}$ is positive semi-definite.
- If $\mathbf{A}$ is positive semi-definite but not positive definite, then $\mathbf{A}$ is singular.
- Projection matrices are positive semi-definite because they are symmetric and have non-negative eigenvalues.

Shorthand notation:

- We sometimes write $\mathbf{A} \succ 0$ to communicate that $\mathbf{A}$ is positive definite.
  - Likewise, $\mathbf{A} \succcurlyeq 0$ means that $\mathbf{A}$ is positive semi-definite.
- As you might imagine, $\mathbf{A} \prec 0$ mean that $\mathbf{A}$ is negative definite (all negative eigenvalues), and $\mathbf{A} \preccurlyeq 0$ means that $\mathbf{A}$ is negative semi-definite.
- To say that $\mathbf{A} \succ \mathbf{B}$ is to say that $\mathbf{A} - \mathbf{B} \succ 0$.
  - Likewise for $\succcurlyeq$, $\prec$, and $\preccurlyeq$.
- Property: If $\mathbf{A}_{N \times N} \succ 0$ and $\mathbf{C}_{Q \times N}$ is of rank Q, then $\mathbf{CAC}^\intercal \succ 0$.
  - See how much easier it is to state the theorem under the shorthand notation? Try proving this!

Table of Contents

Decomposition 1: Lower times upper triangular

- We have already learned how to decompose the matrix $\mathbf{A}$ into the form $\mathbf{A} = \mathbf{LU}$ via Gaussian elimination when there were no row exchanges.
- In fact, we already have an example of this from an earlier section. The $\mathbf{LU}$ decomposition will not directly come into play so much in this class, but the next two decompositions we learn certainly will.

Decomposition 2: Eigendecomposition

- An $N \times N$ matrix $\mathbf{A}$ with N linearly independent eigenvectors can be expressed in the form $\mathbf{A} = \mathbf{S}\boldsymbol{\Lambda}\mathbf{S}^{-1}$, where:
    1. $\mathbf{S}$ is an $N \times N$ matrix of eigenvectors of $\mathbf{A}$.
    2. $\boldsymbol{\Lambda}$ is an $N \times N$ diagonal matrix with the eigenvalues of $\mathbf{A}$ along the diagonal (in descending order, by convention).
- Interestingly, $\mathbf{A}^r = \mathbf{S}\boldsymbol{\Lambda}^r\mathbf{S}^{-1}$ (easy to verify when $r \in \mathbb{Z}^+$).
- When $\mathbf{A}$ is symmetric, the eigenvectors are orthogonal and so, if we scale their lengths to unity, the eigendecomposition can be expressed as $\mathbf{A} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^\intercal$.
- Matrices with this factorization are called *diagonalizable*. Defective matrices are not diagonalizable because they do not have a complete basis of eigenvectors.

Decomposition 2: Eigendecomposition

- The eigendecomposition is useful for quickly computing powers of $\mathbf{A}$, particularly large ones.

```
set.seed(7345)
K <- 100
A <- matrix(sample(1:(K^2)), nrow = K)
A <- A/rep(rowSums(A), each = K)
```

- An $\mathbf{A}$ generated in this way will almost certainly be diagonalizable.

Decomposition 2: Eigendecomposition

- See below:

```
eA <- eigen(A)
Lambda <- diag(eA$values)
S <- eA$vectors

Acb.1 <- S %*% Lambda^3 %*% solve(S)
Acb.2 <- A %*% A %*% A

> Acb.1[1:3,1:3]
                [,1]            [,2]          [,3]
[1,] 0.010315443-0i 0.009203797-0i 0.01187848+0i
[2,] 0.010653311-0i 0.009514142-0i 0.01230300+0i
[3,] 0.009243248-0i 0.008278312-0i 0.01058362+0i

> Acb.2[1:3,1:3]
            [,1]        [,2]       [,3]
[1,] 0.010315443 0.009203797 0.01187848
[2,] 0.010653311 0.009514142 0.01230300
[3,] 0.009243248 0.008278312 0.01058362

> sum(abs(Acb.2 - Acb.1))
[1] 1.211255e-12
```

- You can imagine trying to do $\mathbf{A}^{100}$ (for loop?).

Theorem 18.1: Symmetric matrices and projection matrices

A symmetric $N \times N$ matrix, $\mathbf{A}$, can be realized as a linear combination of N mutually orthogonal projection matrices.

Theorem 18.1: Argument

- Let $\mathbf{A} = \mathbf{S}\mathbf{\Lambda}\mathbf{S}^{-1} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{\intercal}$ denote the eigendecomposition of $\mathbf{A}$ (recall that since $\mathbf{A}$ is symmetric, $\mathbf{S}$ is orthogonal).
- Let $\mathbf{q}_i$ denote the i^{th} column of $\mathbf{Q}$ and let λ_i the i^{th} diagonal entry of $\mathbf{\Lambda}$. Then we can rewrite $\mathbf{A}$ as $\lambda_1 \mathbf{q}_1 \mathbf{q}_1^{\intercal} + \cdots + \lambda_N \mathbf{q}_N \mathbf{q}_N^{\intercal}$.
- Now, note that because $\mathbf{q}_i$ is of unit length, $\mathbf{q}_i^{\intercal}\mathbf{q}_i = 1$, and

$$\mathbf{q}_i \mathbf{q}_i^{\intercal} \quad = \quad \mathbf{q}_i (\mathbf{q}_i^{\intercal} \mathbf{q}_i)^{-1} \mathbf{q}_i.$$

- That is to say that $\mathbf{q}_i \mathbf{q}_i^{\intercal}$ is an $N \times N$ orthogonal projection matrix that projects any vector onto the linear subspace spanned by the eigenvector $\mathbf{q}_i$ (all of which are mutually orthogonal).
- This theorem is known in mathematics by a famous name. What is it?

Matrix decompositions

Decomposition 3: Motivation for singular value decomposition (SVD)

- Consider an $N \times K$ matrix $\mathbf{A}$. Want to find orthonormal vectors in $\mathbb{R}^K$ that, after applying $\mathbf{A}$, come out as orthonormal in $\mathbb{R}^N$:

$$
\begin{bmatrix} \mathbf{A}\mathbf{v}_1 & \cdots & \mathbf{A}\mathbf{v}_K \end{bmatrix} = \begin{bmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_N \end{bmatrix}
\begin{bmatrix}
D_{11} & 0 & \cdots & 0 & \color{red}{0} & \color{red}{\cdots} & \color{red}{0} \\
0 & D_{22} & \ddots & \vdots & \color{red}{0} & \color{red}{\cdots} & \color{red}{0} \\
\vdots & \ddots & \ddots & 0 & \color{red}{\vdots} & \color{red}{\ddots} & \color{red}{\vdots} \\
0 & \cdots & 0 & D_{PP} & \color{red}{0} & \color{red}{\cdots} & \color{red}{0} \\
\color{blue}{0} & \color{blue}{0} & \color{blue}{\cdots} & \color{blue}{0} & & & \\
\color{blue}{\vdots} & \color{blue}{\vdots} & \color{blue}{\ddots} & \color{blue}{\vdots} & & - & \\
\color{blue}{0} & \color{blue}{0} & \color{blue}{\cdots} & \color{blue}{0} & & &
\end{bmatrix}
$$

- If $N > K$, need the blue zeros to make dimensions match ($P = K$).
- If $N < K$, need the red zeros to make dimensions match ($P = N$).
- If $N = K$, no extra zeros needed ($P = K = N$).

Decomposition 3: Motivation for singular value decomposition (SVD)

- Consider an $N \times K$ matrix $\mathbf{A}$. Want to find orthonormal vectors in $\mathbb{R}^K$ that, after applying $\mathbf{A}$, come out as orthonormal in $\mathbb{R}^N$:

$$\begin{bmatrix}\mathbf{A}\mathbf{v}_1 & \cdots & \mathbf{A}\mathbf{v}_K\end{bmatrix} = \begin{bmatrix}\mathbf{u}_1 & \cdots & \mathbf{u}_N\end{bmatrix} \begin{bmatrix} D_{11} & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & D_{22} & \ddots & \vdots & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & 0 & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & D_{PP} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & & & \\ \vdots & \vdots & \ddots & \vdots & & - & \\ 0 & 0 & \cdots & 0 & & & \end{bmatrix}$$

- By convention, take $D_{11} > \cdots > D_{PP} \geq 0$ (call them $d_1, \ldots, d_P$, the *singular values*). The first $R = \text{rank}(\mathbf{A})$ are positive.

Matrix decompositions

Decomposition 3: Motivation for singular value decomposition (SVD)

- Consider an $N \times K$ matrix $\mathbf{A}$. Want to find orthonormal vectors in $\mathbb{R}^K$ that, after applying $\mathbf{A}$, come out as orthonormal in $\mathbb{R}^N$:

$$\begin{bmatrix}\mathbf{A}\mathbf{v}_1 & \cdots & \mathbf{A}\mathbf{v}_K\end{bmatrix} = \begin{bmatrix}\mathbf{u}_1 & \cdots & \mathbf{u}_N\end{bmatrix} \begin{bmatrix} D_{11} & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & D_{22} & \ddots & \vdots & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & 0 & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & D_{PP} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & & & \\ \vdots & \vdots & \ddots & \vdots & & - & \\ 0 & 0 & \cdots & 0 & & & \end{bmatrix}$$

- More compactly, find $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{D}$ such that $\mathbf{AV} = \mathbf{UD}$.
- Since $\mathbf{V}$ is orthogonal, $\mathbf{V}^\top\mathbf{V} = \mathbf{I}$, and we can write $\mathbf{A} = \mathbf{UDV}^\top$.

Decomposition 3: Clues to determine $\mathbf{U}$ and $\mathbf{D}$

- Note first that, assuming such a decomposition is possible, that:

$$\begin{aligned} \mathbf{A}\mathbf{A}^\intercal &= \mathbf{U}\mathbf{D}\mathbf{V}^\intercal(\mathbf{U}\mathbf{D}\mathbf{V}^\intercal)^\intercal \\ &= \mathbf{U}\mathbf{D}\mathbf{V}^\intercal\mathbf{V}\mathbf{D}^\intercal\mathbf{U}^\intercal \\ &= \mathbf{U}\mathbf{D}\mathbf{D}^\intercal\mathbf{U}^\intercal \end{aligned}$$

- This is the eigendecomposition of $\mathbf{A}\mathbf{A}^\intercal$, so $\mathbf{U}$ is based on the eigenvectors of $\mathbf{A}\mathbf{A}^\intercal$ and $\mathbf{D}$ on the eigenvalues of $\mathbf{A}\mathbf{A}^\intercal$.
- The eigenvalues of $\mathbf{A}\mathbf{A}^\intercal$ will be real. Why?
- The eigenvectors of $\mathbf{A}\mathbf{A}^\intercal$ will be orthogonal. Why?

Decomposition 3: Clues to determine $\mathbf{V}$

- Note further that:

$$\begin{aligned} \mathbf{A}^\intercal\mathbf{A} &= (\mathbf{UDV}^\intercal)^\intercal\mathbf{UDV}^\intercal \\ &= \mathbf{VD}^\intercal\mathbf{U}^\intercal\mathbf{UDV}^\intercal \\ &= \mathbf{VD}^\intercal\mathbf{DV}^\intercal \end{aligned}$$

- This is the eigendecomposition of $\mathbf{A}^\intercal\mathbf{A}$, so $\mathbf{V}$ is based on the eigenvectors of $\mathbf{A}^\intercal\mathbf{A}$ and $\mathbf{D}$ on the eigenvalues of $\mathbf{A}^\intercal\mathbf{A}$.
- The eigenvalues of $\mathbf{A}^\intercal\mathbf{A}$ will be real. Why?
- The eigenvectors of $\mathbf{A}^\intercal\mathbf{A}$ will be orthogonal. Why?
- Note that $\mathbf{A}^\intercal\mathbf{A}$ and $\mathbf{AA}^\intercal$ share eigenvalues; one may have extra eigenvalues of zero(s) depending upon the shape of $\mathbf{A}$.

Decomposition 3: Singular value decomposition (SVD)

- Although we didn't prove it, it turns out every $N \times K$ matrix $\mathbf{A}$ can be expressed in the form $\mathbf{A} = \mathbf{UDV}^\intercal$, and the previous slides more or less show you how to do it:
    1. $\mathbf{U}$ is an $N \times N$ matrix of orthonormal eigenvectors of $\mathbf{AA}^\intercal$.
    2. $\mathbf{D}$ is an $N \times K$ "diagonal" matrix with the positive square-rooted eigenvalues of $\mathbf{AA}^\intercal$ along the diagonal and are arranged in descending order (called the singular values).
    3. $\mathbf{V}$ is a $K \times K$ matrix of orthonormal eigenvectors of $\mathbf{A}^\intercal\mathbf{A}$.
- Must be *careful* when setting up the eigenvectors, because if $\mathbf{v}$ is an eigenvector, so is $-\mathbf{v}$.

Decomposition 3: Singular value decomposition (SVD)

- As an example, suppose I want to find the SVD of the following matrix:

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix}.$$

- We could just cycle through this by brute force using the steps on the prior slide.

Decomposition 3: Singular value decomposition (SVD)

- Step 1: Find the eigenvectors of $\mathbf{A}\mathbf{A}^\intercal$ (because $\mathbf{A}$ is symmetric, they should exist and they should be orthogonal).

$$\mathbf{A}\mathbf{A}^\intercal = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}.$$

- Can identify $\mathbf{u}_1 = (1,1)^\intercal$ and $\mathbf{u}_2 = (1,-1)^\intercal$ as eigenvectors with corresponding eigenvalues of $\lambda_1 = 3$ and $\lambda_2 = 1$.
  - Careful!! Choosing eigenvectors this way (and not $(-1,-1)^\intercal$ and/or $(-1,1)^\intercal$, for instance) affects choice of eigenvectors that form $\mathbf{V}$.
- Suitably scaling, this suggests that:

$$\mathbf{U} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \text{ and } \mathbf{D} = \begin{bmatrix} \sqrt{3} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

Decomposition 3: Singular value decomposition (SVD)

- Step 2: Find the eigenvectors of $\mathbf{A}^\intercal\mathbf{A}$ (because $\mathbf{A}$ is symmetric, they should exist and they should be orthogonal).

$$\mathbf{A}^\intercal\mathbf{A} \;=\; \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

- Can identify eigenvalues $\lambda_1 = 3$, $\lambda_2 = 1$, and $\lambda_3 = 0$ (no accident).
- An eigenvector corresponding to λ_1 could be chosen as, say, $\mathbf{v}_1 = (1, 2, 1)^\intercal$ or $\mathbf{v}_1 = (-1, -2, -1)^\intercal$. Which one is correct?
- Recall that the whole point of this we need $\mathbf{A}\mathbf{v}_1 = \mathbf{u}_1 d_1$ for non-negative d_1. This only checks out with for the first choice.

Decomposition 3: Singular value decomposition (SVD)

- Step 2: Find the eigenvectors of $\mathbf{A}^\intercal\mathbf{A}$ (because $\mathbf{A}$ is symmetric, they should exist and they should be orthogonal).

$$\mathbf{A}^\intercal\mathbf{A} \quad = \quad \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

- Can identify eigenvalues $\lambda_1 = 3$, $\lambda_2 = 1$, and $\lambda_3 = 0$ (no accident).
- An eigenvector corresponding to λ_2 could be chosen as, say, $\mathbf{v}_2 = (1, 0, -1)^\intercal$ or $\mathbf{v}_2 = (-1, 0, 1)^\intercal$. Which one is correct?
- Recall that the whole point of this we need $\mathbf{A}\mathbf{v}_2 = \mathbf{u}_2 d_2$ for non-negative d_2. This only checks out with for the second choice.

Decomposition 3: Singular value decomposition (SVD)

- Step 2: Find the eigenvectors of $\mathbf{A}^\intercal\mathbf{A}$ (because $\mathbf{A}$ is symmetric, they should exist and they should be orthogonal).

$$\mathbf{A}^\intercal\mathbf{A} \quad = \quad \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

- Can identify eigenvalues $\lambda_1 = 3$, $\lambda_2 = 1$, and $\lambda_3 = 0$ (no accident).
- An eigenvector corresponding to λ_3 could be chosen as, say, $\mathbf{v}_3 = (1, -1, 1)^\intercal$ or $\mathbf{v}_3 = (-1, 1, -1)^\intercal$. Which one is correct?
- If I'm not mistaken, it shouldn't make a difference because there is no $\mathbf{u}_3$ with which I'm trying to synergize the direction of $\mathbf{v}_3$. Let's try it both ways and pray?

Decomposition 3: Singular value decomposition (SVD)

- Step 2: Find the eigenvectors of $\mathbf{A}^{\intercal}\mathbf{A}$ (because $\mathbf{A}$ is symmetric, they should exist and they should be orthogonal).
- Suitably scaling, *one* choice of $\mathbf{V}$ is given by:

$$\mathbf{V} = \begin{bmatrix} 1/\sqrt{6} & -1/\sqrt{2} & 1/\sqrt{3} \\ 2/\sqrt{6} & 0 & -1/\sqrt{3} \\ 1/\sqrt{6} & 1/\sqrt{2} & 1/\sqrt{3} \end{bmatrix}.$$

- The other is given by:

$$\widetilde{\mathbf{V}} = \begin{bmatrix} 1/\sqrt{6} & -1/\sqrt{2} & -1/\sqrt{3} \\ 2/\sqrt{6} & 0 & 1/\sqrt{3} \\ 1/\sqrt{6} & 1/\sqrt{2} & -1/\sqrt{3} \end{bmatrix}.$$

Decomposition 3: Singular value decomposition (SVD)

- Step 3: Does this check out when we use **V**?

$$
\begin{aligned}
\mathbf{UDV}^{\mathsf{T}} &= \frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}\begin{bmatrix}\sqrt{3} & 0 & 0\\ 0 & 1 & 0\end{bmatrix}\begin{bmatrix}1/\sqrt{6} & 2/\sqrt{6} & 1/\sqrt{6}\\ -1/\sqrt{2} & 0 & 1/\sqrt{2}\\ 1/\sqrt{3} & -1/\sqrt{3} & 1/\sqrt{3}\end{bmatrix}\\
&= \frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}\begin{bmatrix}\sqrt{3}/\sqrt{6} & 2\sqrt{3}/\sqrt{6} & \sqrt{3}/\sqrt{6}\\ -1/\sqrt{2} & 0 & 1/\sqrt{2}\end{bmatrix}\\
&= \begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}\begin{bmatrix}1/2 & 1 & 1/2\\ -1/2 & 0 & 1/2\end{bmatrix}\\
&= \begin{bmatrix}0 & 1 & 1\\ 1 & 1 & 0\end{bmatrix}
\end{aligned}
$$

- And we breathe a huge sigh of relief!

Decomposition 3: Singular value decomposition (SVD)

- Step 3: Does this check out when we use $\widetilde{\mathbf{V}}$?

$$\mathbf{U}\mathbf{D}\widetilde{\mathbf{V}}^{\intercal} = \frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\\1 & -1\end{bmatrix}\begin{bmatrix}\sqrt{3} & 0 & 0\\0 & 1 & 0\end{bmatrix}\begin{bmatrix}1/\sqrt{6} & 2/\sqrt{6} & 1/\sqrt{6}\\-1/\sqrt{2} & 0 & 1/\sqrt{2}\\-1/\sqrt{3} & 1/\sqrt{3} & -1/\sqrt{3}\end{bmatrix}$$

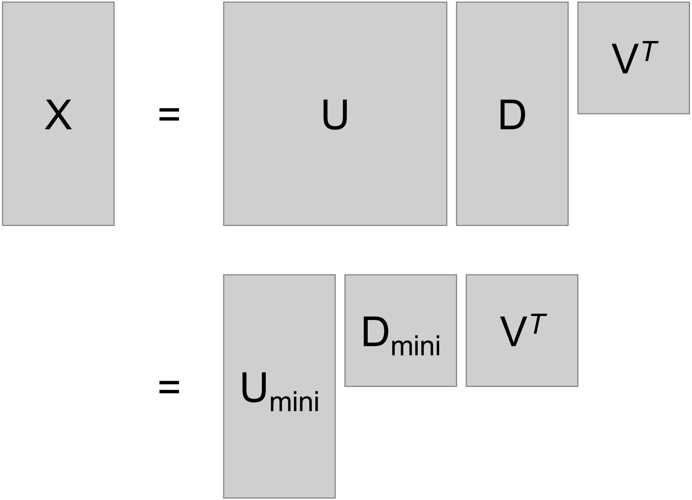
$$= \frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\\1 & -1\end{bmatrix}\begin{bmatrix}\sqrt{3}/\sqrt{6} & 2\sqrt{3}/\sqrt{6} & \sqrt{3}/\sqrt{6}\\-1/\sqrt{2} & 0 & 1/\sqrt{2}\end{bmatrix}\ldots\text{umm}\ldots$$

- We can stop here. Do you see what happened?
- Because $\mathbf{D}$ has a right column of zeros, the third *row* of $\widetilde{\mathbf{V}}^{\intercal}$ never got picked up. We could have chosen $\mathbf{v}_3 = (\log(\pi), e^2, -2\sinh(7.3))^{\intercal}$ and the math would have worked just fine. But of course $\mathbf{V}$ would then not be orthogonal and wouldn't qualify as a valid SVD.

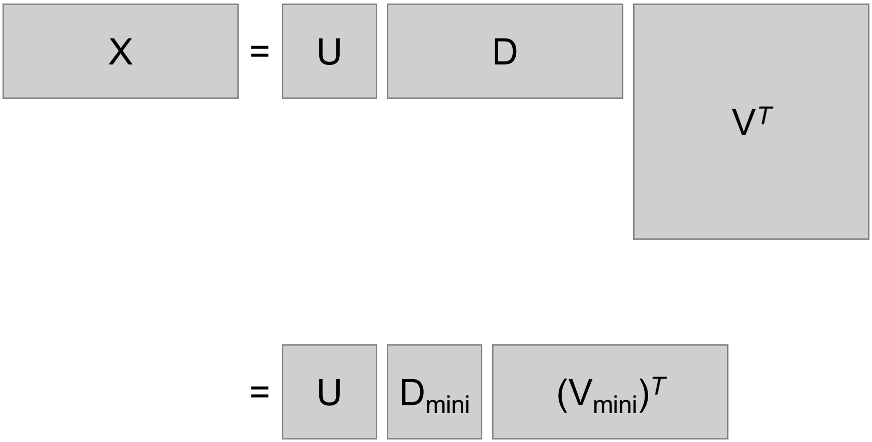
Decomposition 3: Singular value decomposition (SVD)

- Given the convention of ordering the singular values from largest to smallest, the SVD is unique up to sign changes in the eigenvectors (and, for non-square matrices, re-ordering of the "extraneous" eigenvectors).
- As we just saw, when $N \neq K$, the SVD contains some "extra" information that we can theoretically toss and come up with a decomposition (it's just not the SVD in particular).

Decomposition 3: SVD for narrow matrices (*only* the top is an SVD)



Decomposition 3: SVD for wide matrices (*only* the top is an SVD)



Decomposition 3: Singular value decomposition (SVD)

- If $\mathbf{A}$ is positive definite, then the SVD is none other than the eigendecomposition.
  - Because $\mathbf{A}$ is symmetric, it has real eigenvalues and orthogonal eigenvectors, so it can be written as $\mathbf{A} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^\intercal$.
  - Because $\mathbf{A}$ is positive definite in particular, it has only positive eigenvalues and so $\boldsymbol{\Lambda}$ is a diagonal matrix of only positive values.
  - When $\mathbf{A}$ is symmetric but *not* positive definite, the eigendecomposition and SVD *almost* coincide, but $\mathbf{D}$ contains the absolute values of the eigenvalues and the orientation of the vectors in $\mathbf{U}$ may need to be switched.

Theorem 18.2: A condition for $\mathbf{A}^\intercal\mathbf{A}$ to be positive definite

If $\mathbf{A}$ is an $N \times K$ matrix of rank K, then $\mathbf{A}^\intercal\mathbf{A}$ is positive definite.

Theorem 18.2: Proof

- We will use the SVD. Let $\mathbf{A} = \mathbf{UDV}^\intercal$.
- $\mathbf{A}^\intercal\mathbf{A} = (\mathbf{UDV}^\intercal)^\intercal(\mathbf{UDV}^\intercal) = \mathbf{VD}^\intercal\mathbf{U}^\intercal\mathbf{UDV}^\intercal = \mathbf{VD}^2\mathbf{V}^\intercal$.
- This must be the eigendecomposition of $\mathbf{A}^\intercal\mathbf{A}$ with $\mathbf{\Lambda} = \mathbf{D}^2$; since the rank of $\mathbf{A}$ is K, the diagonal entries of $\mathbf{D}$ are nonzero and so the diagonal entires of $\mathbf{\Lambda} = \mathbf{D}^2$ are positive.
- Therefore, $\mathbf{A}^\intercal\mathbf{A}$ has only positive eigenvalues, which concludes the proof.

Lemma 18.3: Property of matrix products

Suppose $\mathbf{A}$ and $\mathbf{B}$ are positive definite of the same dimension and that $\mathbf{A} - \mathbf{B} \succ 0$. Then, $\mathbf{B}^{-1} - \mathbf{A}^{-1} \succ 0$.

The proof of this lemma implicitly relies on knowledge of a key matrix decomposition. Can you spot it on the following slide?

Lemma 18.3: Proof

- Knowing $\mathbf{A} - \mathbf{B} \succ 0$, multiply through by $\mathbf{B}^{-1/2}$ on left and right.
  - To see why this is a valid step, note that $\mathbf{z}^\mathsf{T} = \mathbf{x}^\mathsf{T}\mathbf{B}^{-1/2}$ is a vector and $\mathbf{z}^\mathsf{T}(\mathbf{A} - \mathbf{B})\mathbf{z} > 0$ by the fact that $\mathbf{A} - \mathbf{B} \succ 0$.
- So, we have:

$$\begin{aligned}\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2} - \mathbf{B}^{-1/2}\mathbf{B}\mathbf{B}^{-1/2} &\succ 0\\ \mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2} - \mathbf{I} &\succ 0\end{aligned}$$

- If λ is an eigenvalue of $\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}$, $\lambda > 1$ (Slide 828). Therefore,

$$\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2} = (\mathbf{B}^{-1/2}\mathbf{A}^{1/2})(\mathbf{A}^{1/2}\mathbf{B}^{-1/2})$$

- $(\mathbf{B}^{-1/2}\mathbf{A}^{1/2})(\mathbf{A}^{1/2}\mathbf{B}^{-1/2})$ has the same eigenvalues as $(\mathbf{A}^{1/2}\mathbf{B}^{-1/2})(\mathbf{B}^{-1/2}\mathbf{A}^{1/2}) = \mathbf{A}^{1/2}\mathbf{B}^{-1}\mathbf{A}^{1/2}$ (Lemma 18.2); all > 1.

$$\begin{aligned}\mathbf{A}^{1/2}\mathbf{B}^{-1}\mathbf{A}^{1/2} - \mathbf{I} &\succ 0\\ \mathbf{A}^{-1/2}\mathbf{A}^{1/2}\mathbf{B}^{-1}\mathbf{A}^{1/2}\mathbf{A}^{-1/2} - \mathbf{A}^{-1/2}\mathbf{I}\mathbf{A}^{-1/2} &\succ 0\\ \mathbf{B}^{-1} - \mathbf{A}^{-1} &\succ 0.\end{aligned}$$

Table of Contents

Basic ideas:

- Let $f(\mathbf{x}) \in \mathbb{R}$ for $\mathbf{x} \in \mathbb{R}^N$, then:

$$\frac{\partial f}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_N} \end{bmatrix}.$$

- It is very convenient to be able to write down derivatives of (multivariable) functions in terms of matrices and vectors.
- I am following the "denominator layout" convention for vector/matrix calculus in this course. Since I am treating vectors as column vectors in this course, using denominator layout helps me avoid the need to transpose all of my resulting derivatives to get column vectors back.

Example 18.5: Real-valued function of a vector

- If $f(\mathbf{x}) = \mathbf{c}^\intercal \mathbf{x} = \mathbf{x}^\intercal \mathbf{c} = c_1 x_1 + \cdots + c_N x_N$, then:

$$\frac{\partial f}{\partial \mathbf{x}} = \begin{pmatrix} c_1 \\ \vdots \\ c_N \end{pmatrix} = \mathbf{c}.$$

- This should give you insights into why I want to use denominator layout; I'd really like the derivative of a real-valued function of a vector to be a column. Our derivative rules will flow and follow from this convention.
- By numerator layout, $\partial f / \partial \mathbf{x} = \mathbf{c}^\intercal$; we are not going to use numerator layout, but don't be confused if when you look it up you see a discrepancy of this sort.

Example 18.6: Vector-valued function of a vector

- If $f(\mathbf{x}) = \mathbf{A}\mathbf{x}$, then $\partial f / \partial \mathbf{x} = \mathbf{A}^\intercal$.
- If $f(\mathbf{x}) = \mathbf{x}^\intercal \mathbf{A}$, then $\partial f / \partial \mathbf{x} = \mathbf{A}$.
- This is reminiscent of the rule you already know from single-variable calculus: if $f(x) = ax$, then $f'(x) = a$.
- Not to be a broken record, but this example also underscores the point that choosing numerator vs. denominator layout doesn't fundamentally avoid having to keep track of transposes altogether.
- Try showing these on your own!

Rule: Vector elements as a function of a linear form

- Suppose $f(\cdot)$ is differentiable, and let $\mathbf{w} = \text{vec}(f(\mathbf{v}_i^\intercal \mathbf{u}))$. Then,

$$\frac{\partial \mathbf{w}}{\partial \mathbf{u}} = \mathbf{V}^\intercal \text{diag}(f'(\mathbf{v}_i^\intercal \mathbf{u})),$$

  where $\mathbf{v}_i^\intercal$ denotes the i^{th} (length-K) row of the $N \times K$ matrix $\mathbf{V}$.
- Note:
  - $\mathbf{w}$ is a length-N vector.
  - $\mathbf{u}$ is a length-K vector.
  - $\text{diag}(f'(\mathbf{v}_i^\intercal \mathbf{u}))$ is an $N \times N$ matrix.
  - $\partial \mathbf{w}/\partial \mathbf{u}$ is a $K \times N$ matrix.
- This is reminiscent of a chain rule of sorts.

Example 18.7: Vector elements as a function of a linear form

- Determine the derivative of $\text{vec}((\mathbf{x}_i^\intercal \boldsymbol{\beta})^{-2})$ with respect to $\boldsymbol{\beta}$:

$$\frac{\partial \text{vec}((\mathbf{x}_i^\intercal \boldsymbol{\beta})^{-2})}{\partial \boldsymbol{\beta}} = \mathbf{X}^\intercal \text{diag}(-2(\mathbf{x}_i^\intercal \boldsymbol{\beta})^{-3}).$$

- If you're wondering how this rule will get applied in this course, notice that I am tipping my hand in the notation I'm using for this example.

Rule: Quadratic forms

- Let $\mathbf{A}$ denote a symmetric $K \times K$ matrix.
- If $f(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}$, then $\partial f / \partial \mathbf{x} = 2\mathbf{A}\mathbf{x}$.
- To prove this, write in summation notation:

$$
\begin{aligned}
f(\mathbf{x}) &= \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} = \sum_{i=1}^{K}\sum_{j=1}^{K} a_{ij}x_i x_j \\
\frac{\partial f}{\partial x_k} &= \sum_{i=1}^{K}\sum_{j=1}^{K} \frac{\partial}{\partial x_k} a_{ij}x_i x_j = \sum_{i=1}^{K} a_{ki}x_i + \sum_{j=1}^{K} a_{jk}x_j \\
&= 2\sum_{i=1}^{K} a_{ki}x_i = 2\mathbf{a}_{k\cdot}^\mathsf{T}\mathbf{x}
\end{aligned}
$$

Example 18.8: Quadratic forms

- If $f(\mathbf{x}) = ||\mathbf{y} - \mathbf{Ax}||^2$, then $\partial f / \partial \mathbf{x} = -2\mathbf{A}^\mathsf{T}(\mathbf{y} - \mathbf{Ax})$.
- To see this, note that:

$$
\begin{aligned}
||\mathbf{y} - \mathbf{Ax}||^2 &= (\mathbf{y} - \mathbf{Ax})^\mathsf{T}(\mathbf{y} - \mathbf{Ax}) \\
&= \mathbf{y}^\mathsf{T}\mathbf{y} - \mathbf{y}^\mathsf{T}(\mathbf{Ax}) - (\mathbf{Ax})^\mathsf{T}\mathbf{y} + (\mathbf{Ax})^\mathsf{T}(\mathbf{Ax}) \\
&= \mathbf{y}^\mathsf{T}\mathbf{y} - \mathbf{y}^\mathsf{T}(\mathbf{Ax}) - (\mathbf{Ax})^\mathsf{T}\mathbf{y} + \mathbf{x}^\mathsf{T}(\mathbf{A}^\mathsf{T}\mathbf{A})\mathbf{x}.
\end{aligned}
$$

- Apply rules on previous slides and rearrange to arrive at result.
- Notice that this sort of behaves like the power rule and the chain rule.
- I expect that this example gives you a clue as to why we want to spend a little time on matrix calculus.

Rule: Derivatives of inverses

- If $\mathbf{A}$ is non-singular with each entry a real-valued function of x, then:

$$\frac{\partial}{\partial x}\mathbf{A}^{-1} = -\mathbf{A}^{-1}\left(\frac{\partial \mathbf{A}}{\partial x}\right)\mathbf{A}^{-1}.$$

- To see this, note that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ and differentiate both sides:

$$\begin{aligned}
\frac{\partial}{\partial x}\mathbf{A}^{-1}\mathbf{A} &= \frac{\partial}{\partial x}\mathbf{I} = \mathbf{0}. \\
\Rightarrow \left(\frac{\partial \mathbf{A}^{-1}}{\partial x}\right)\mathbf{A} + \mathbf{A}^{-1}\left(\frac{\partial \mathbf{A}}{\partial x}\right) &= \mathbf{0} \\
\Rightarrow \left(\frac{\partial \mathbf{A}^{-1}}{\partial x}\right)\mathbf{A} &= -\mathbf{A}^{-1}\left(\frac{\partial \mathbf{A}}{\partial x}\right) \\
\Rightarrow \frac{\partial \mathbf{A}^{-1}}{\partial x} &= -\mathbf{A}^{-1}\left(\frac{\partial \mathbf{A}}{\partial x}\right)\mathbf{A}^{-1}.
\end{aligned}$$

Basic ideas: Derivatives with respect to matrices

- Let $f(\mathbf{X}) \in \mathbb{R}$ for an $N \times K$ matrix $\mathbf{X}$, then:

$$\frac{\partial f}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial f}{\partial x_{11}} & \cdots & \frac{\partial f}{\partial x_{1K}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_{N1}} & \cdots & \frac{\partial f}{\partial x_{NK}} \end{bmatrix}.$$

Example 18.9: Log-determinant and trace

- For a $K \times K$ matrix $\mathbf{X}$, an $N \times K$ matrix $\mathbf{A}_1$, and a $K \times N$ matrix $\mathbf{A}_2$:

$$\frac{\partial}{\partial \mathbf{X}} \text{tr}(\mathbf{A}_1 \mathbf{X} \mathbf{A}_2) = \mathbf{A}_1^\intercal \mathbf{A}_2^\intercal.$$

- For an invertible matrix $\mathbf{X}$:

$$\frac{\partial}{\partial \mathbf{X}} \log(|\mathbf{X}|) = (\mathbf{X}^{-1})^\intercal.$$

- Try verifying these for a 2×2 or a 3×3 matrix!

Rule: Matrix chain rule

- If $f = f(\mathbf{x}) : \mathbb{R}^K \longrightarrow \mathbb{R}$ is a function of $\mathbf{x}$ and $\mathbf{x}(\mathbf{y}) : \mathbb{R}^M \longrightarrow \mathbb{R}^K$ is a function of $\mathbf{y}$, then:

$$\frac{\partial f}{\partial \mathbf{y}} = \frac{\partial \mathbf{x}}{\partial \mathbf{y}} \frac{\partial f}{\partial \mathbf{x}}.$$

- Notice, as a heuristic, the "cancellation of $\partial \mathbf{x}$."
- Notice the reverse-ordering. The dimensions need to match!
  - $\partial f / \partial \mathbf{y}$ is a length-M vector.
  - $\partial \mathbf{x} / \partial \mathbf{y}$ is an $M \times K$ matrix.
  - $\partial f / \partial \mathbf{x}$ is a length-K vector.
- This extends in the way you might expect:

$$\frac{\partial f}{\partial \mathbf{z}} = \frac{\partial \mathbf{y}}{\partial \mathbf{z}} \frac{\partial \mathbf{x}}{\partial \mathbf{y}} \frac{\partial f}{\partial \mathbf{x}}.$$

Rule: Constant matrix times a vector

- Suppose $\mathbf{A}$ is an $N \times K$ matrix of constants, $\mathbf{z} = \mathbf{z}(\mathbf{x})$ is a length-K vector, and $\mathbf{x}$ is a length-M vector. Then,

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{A}\mathbf{z}) = \frac{\partial \mathbf{z}}{\partial \mathbf{x}} \mathbf{A}^{\intercal}.$$

- Note: $\partial \mathbf{z}/\partial \mathbf{x}$ is an $M \times K$ matrix.
- We'll invoke this implicitly many times.

Rule: Diagonal matrix times a vector

- Suppose $\mathbf{u} = \mathbf{u}(\mathbf{x})$ and $\mathbf{v} = \mathbf{v}(\mathbf{x})$ are length-N vectors, both functions of a length-K vector, $\mathbf{x}$. Then,

$$\frac{\partial}{\partial \mathbf{x}}(\text{diag}(\mathbf{u})\mathbf{v}) = \frac{\partial \mathbf{v}}{\partial \mathbf{x}}\text{diag}(\mathbf{u}) + \frac{\partial \mathbf{u}}{\partial \mathbf{x}}\text{diag}(\mathbf{v}).$$

- Note:
  - diag($\mathbf{u}$) and diag($\mathbf{v}$) are $N \times N$ diagonal matrices with the entries of $\mathbf{u}$ and $\mathbf{v}$ along the respective diagonals.
  - diag($\mathbf{u}$)$\mathbf{v}$ is a length-N vector (with a $K \times N$ derivative).
  - $\partial \mathbf{u}/\partial \mathbf{x}$ and $\partial \mathbf{v}/\partial \mathbf{x}$ are $K \times N$ matrices.
- This result seems obscure, but rest assured that this formula will come up when we cover generalized linear models (you do **not** need to memorize the formula :)).

Example 18.10: Diagonal matrix times a vector

- Suppose $\mathbf{u}(\boldsymbol{\beta}) = \text{vec}((\mathbf{x}_i^\intercal \boldsymbol{\beta})^{-2})$ and $\mathbf{v}(\boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta}$.
- Application of the rule on the previous slide should reveal that:

$$\frac{\partial}{\partial \boldsymbol{\beta}}(\text{diag}(\mathbf{u})\mathbf{v}) = -\mathbf{X}^\intercal \text{diag}((\mathbf{x}_i^\intercal \boldsymbol{\beta})^{-2}).$$

- I recommend trying this. I also recommend trying to show this by first simplifying the multiplication and applying the rule on Slide 872—it should give you the same answer.
- More generally, please note that I do not expect you to derive matrix calculus rules so much as I expect you to apply existing ones.

Table of Contents

Motivation:

- As we know, not all matrices are invertible.
- A generalized inverse (or pseudoinverse) of $\mathbf{A}$, denoted by $\mathbf{A}^{-}$, is one that satisfies $\mathbf{A}\mathbf{A}^{-}\mathbf{A} = \mathbf{A}$.
  - Such a matrix always exists.
  - If $\mathbf{A}$ is non-singular, then $\mathbf{A}^{-}$ is unique and $\mathbf{A}^{-} = \mathbf{A}^{-1}$.
  - Otherwise, $\mathbf{A}^{-}$ is not unique.

Example 18.11: Square matrix

- Let $\mathbf{A} = \begin{bmatrix} 2 & 2 & 3 \\ 1 & 0 & 1 \\ 3 & 2 & 4 \end{bmatrix}$, which is singular.
- Both of the following are generalized inverses (g-inverses, henceforth):

$$\mathbf{A}_1^- = \begin{bmatrix} 0 & 1 & 0 \\ 1/2 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \text{ and } \mathbf{A}_2^- = \begin{bmatrix} 0 & 1 & 0 \\ 0 & -3/2 & 1/2 \\ 0 & 0 & 0 \end{bmatrix}$$

- You can verify that $\mathbf{A}\mathbf{A}_1^-\mathbf{A} = \mathbf{A}\mathbf{A}_2^-\mathbf{A} = \mathbf{A}$.

Example 18.12: Non-square matrix

- Let $\mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}$, which is not even square!
- Both of the following are g-inverses:

$$\mathbf{x}_1^- = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix} \text{ and } \mathbf{x}_2^- = \begin{bmatrix} 0 & 1/2 & 0 & 0 \end{bmatrix}$$

- You can verify that $\mathbf{x}\mathbf{x}_1^-\mathbf{x} = \mathbf{x}\mathbf{x}_2^-\mathbf{x} = \mathbf{x}$.

Properties of g-inverses:

- Let $\mathbf{A}$ have rank R, and let $\mathbf{A}^-$ be a g-inverse for $\mathbf{A}$. Then, $\text{rank}(\mathbf{A}^-\mathbf{A}) = \text{rank}(\mathbf{A}\mathbf{A}^-) = \text{rank}(\mathbf{A}) = R$.
- If the system $\mathbf{Ax} = \mathbf{c}$ is consistent, then for any g-inverse $\mathbf{A}^-$ of $\mathbf{A}$, $\mathbf{x} = \mathbf{A}^-\mathbf{c}$ is a solution (different choices for $\mathbf{A}^-$ yield different solutions to the equation).
- The system $\mathbf{Ax} = \mathbf{c}$ has a solution if and only if $\mathbf{A}\mathbf{A}^-\mathbf{c} = \mathbf{c}$ for any g-inverse $\mathbf{A}^-$ of $\mathbf{A}$.
- The following six slides comprise Lemmas 18.4 through 18.9, which are extremely useful in proving key results relevant to ANOVA methods for design matrices that are not of full rank.

Lemma 18.4: Another property involving $\mathbf{A}^\intercal\mathbf{A}$

If $\mathbf{B}\mathbf{A}^\intercal\mathbf{A} = \mathbf{C}\mathbf{A}^\intercal\mathbf{A}$, then $\mathbf{B}\mathbf{A}^\intercal = \mathbf{C}\mathbf{A}^\intercal$.

Rough idea of proof: Show that $(\mathbf{B}\mathbf{A}^\intercal - \mathbf{C}\mathbf{A}^\intercal)(\mathbf{B}\mathbf{A}^\intercal - \mathbf{C}\mathbf{A}^\intercal)^\intercal = \mathbf{0}$ using Lemma 18.1.

Lemma 18.5

If $(\mathbf{X}^\intercal\mathbf{X})^-$ is a g-inverse of $\mathbf{X}^\intercal\mathbf{X}$, then $[(\mathbf{X}^\intercal\mathbf{X})^-]^\intercal$ is also a g-inverse of $\mathbf{X}^\intercal\mathbf{X}$.

Rough idea of proof: Since $(\mathbf{X}^\intercal\mathbf{X})^-$ is a g-inverse of $\mathbf{X}^\intercal\mathbf{X}$, we have that $(\mathbf{X}^\intercal\mathbf{X})(\mathbf{X}^\intercal\mathbf{X})^-(\mathbf{X}^\intercal\mathbf{X}) = \mathbf{X}^\intercal\mathbf{X}$. Transpose both sides.

Lemma 18.6

If $(\mathbf{X}^\intercal\mathbf{X})^-$ is a g-inverse of $\mathbf{X}^\intercal\mathbf{X}$, then $(\mathbf{X}^\intercal\mathbf{X})^-\mathbf{X}^\intercal\mathbf{X}[(\mathbf{X}^\intercal\mathbf{X})^-]^\intercal$ is a symmetric, reflexive g-inverse of $\mathbf{X}^\intercal\mathbf{X}$.

Rough idea of proof: Symmetry is clear. You will want to apply Lemma 18.5 twice to show this.

Lemma 18.7

If $(\mathbf{X}^\intercal\mathbf{X})^-$ is a g-inverse of $\mathbf{X}^\intercal\mathbf{X}$, then $\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^-$ is a g-inverse of $\mathbf{X}^\intercal$ and $(\mathbf{X}^\intercal\mathbf{X})^-\mathbf{X}^\intercal$ is a g-inverse of $\mathbf{X}$.

Rough idea of proof: You know that $(\mathbf{X}^\intercal\mathbf{X})(\mathbf{X}^\intercal\mathbf{X})^-(\mathbf{X}^\intercal\mathbf{X}) = \mathbf{X}^\intercal\mathbf{X}$; you will want to apply Lemma 18.4.

Lemma 18.8

If $\mathbf{G}$ and $\widetilde{\mathbf{G}}$ are both g-inverses of $\mathbf{X}^\intercal\mathbf{X}$, then $\mathbf{X}\mathbf{G}\mathbf{X}^\intercal = \mathbf{X}\widetilde{\mathbf{G}}\mathbf{X}^\intercal$.

Rough idea of proof: You will want to apply Lemma 18.7 and then Lemma 18.4.

Lemma 18.9

If $(\mathbf{X}^\intercal\mathbf{X})^-$ is a g-inverse of $\mathbf{X}^\intercal\mathbf{X}$, then $\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^-\mathbf{X}^\intercal$ is symmetric.

Rough idea of proof: You will want to apply Lemma 18.5 and then Lemma 18.8.

The Moore-Penrose pseudoinverse:

- This is a g-inverse for $\mathbf{A}$ that satisfies the following criteria:
  1. $\mathbf{A}\mathbf{A}^{-}\mathbf{A} = \mathbf{A}$.
  2. $\mathbf{A}^{-}\mathbf{A}\mathbf{A}^{-} = \mathbf{A}^{-}$.
  3. $\mathbf{A}^{-}\mathbf{A}$ is symmetric.
  4. $\mathbf{A}\mathbf{A}^{-}$ is symmetric.
- A g-inverse that that satisfies the first two properties is *reflexive*.
- The Moore-Penrose pseudoinverse is unique.

The Moore-Penrose pseudoinverse: What is it?

- Let $\mathbf{A} = \mathbf{UDV}^\intercal$ denote the SVD. The Moore-Penrose pseudoinverse is given by $\mathbf{A}^- = \mathbf{V}(\mathbf{D}^+)^\intercal\mathbf{U}^\intercal$, where $\mathbf{D}^+$ is obtained by replacing each nonzero diagonal element of $\mathbf{D}$ with its reciprocal.
- The properties of the Moore-Penrose pseudoinverse are easy to verify:
  1. $\mathbf{AA}^-\mathbf{A} = \mathbf{UDV}^\intercal\mathbf{V}(\mathbf{D}^+)^\intercal\mathbf{U}^\intercal\mathbf{UDV}^\intercal = \mathbf{UDV}^\intercal = \mathbf{A}$.
  2. $\mathbf{A}^-\mathbf{AA}^- = \mathbf{V}(\mathbf{D}^+)^\intercal\mathbf{U}^\intercal\mathbf{UDV}^\intercal\mathbf{V}(\mathbf{D}^+)^\intercal\mathbf{U}^\intercal = (\mathbf{VDU}^\intercal)^\intercal = \mathbf{UDV}^\intercal = \mathbf{A}$.
  3. $\mathbf{A}^-\mathbf{A} = \mathbf{V}(\mathbf{D}^+)^\intercal\mathbf{U}^\intercal\mathbf{UDV}^\intercal = \mathbf{VV}^\intercal$ (symmetric).
  4. $\mathbf{AA}^- = \mathbf{UDV}^\intercal\mathbf{V}(\mathbf{D}^+)^\intercal\mathbf{U}^\intercal = \mathbf{UU}^\intercal$ (symmetric).
- Honestly, the SVD is very often the best possible way to think about what a matrix is and what it does.

Example 18.13: The Moore-Penrose pseudoinverse

- Let's determine the Moore-Penrose pseudoinverse for

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix}.$$

- We did the SVD for this matrix already, with:

$$\begin{aligned} \mathbf{V} &= \begin{bmatrix} 1/\sqrt{6} & -1/\sqrt{2} & 1/\sqrt{3} \\ 2/\sqrt{6} & 0 & -1/\sqrt{3} \\ 1/\sqrt{6} & 1/\sqrt{2} & 1/\sqrt{3} \end{bmatrix}. \\ (\mathbf{D}^{+})^{\intercal} &= \begin{bmatrix} 1/\sqrt{3} & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \\ \mathbf{U}^{\intercal} &= \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \end{aligned}$$

Example 18.13: The Moore-Penrose pseudoinverse

- Leveraging our knowledge of how this will play out, let's convert to "mini" matrices by chopping out the extraneous information.

$$\begin{aligned}
\mathbf{V}_{\text{mini}} &= \begin{bmatrix} 1/\sqrt{6} & -1/\sqrt{2} \\ 2/\sqrt{6} & 0 \\ 1/\sqrt{6} & 1/\sqrt{2} \end{bmatrix}. \\
(\mathbf{D}^{+}_{\text{mini}})^{\mathsf{T}} &= \begin{bmatrix} 1/\sqrt{3} & 0 \\ 0 & 1 \end{bmatrix} \\
\mathbf{U}^{\mathsf{T}} &= \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}
\end{aligned}$$

Example 18.13: The Moore-Penrose pseudoinverse

- Continuing on,

$$
\begin{aligned}
\mathbf{A}^{-} &= \mathbf{V}_{\text{mini}}(\mathbf{D}^{+}_{\text{mini}})^{\mathsf{T}}\mathbf{U}^{\mathsf{T}} \\
&= \frac{1}{\sqrt{2}} \begin{bmatrix} 1/\sqrt{6} & -1/\sqrt{2} \\ 2/\sqrt{6} & 0 \\ 1/\sqrt{6} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 1/\sqrt{3} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \\
&= \begin{bmatrix} 1/2\sqrt{3} & -1/2 \\ 1/\sqrt{3} & 0 \\ 1/2\sqrt{3} & 1/2 \end{bmatrix} \begin{bmatrix} 1/\sqrt{3} & 1/\sqrt{3} \\ 1 & -1 \end{bmatrix} \\
&= \begin{bmatrix} 1/6 - 1/2 & 1/6 + 1/2 \\ 1/3 & 1/3 \\ 1/6 + 1/2 & 1/6 - 1/2 \end{bmatrix} = \frac{1}{3} \begin{bmatrix} -1 & 2 \\ 1 & 1 \\ 2 & -1 \end{bmatrix}
\end{aligned}
$$

Example 18.13: The Moore-Penrose pseudoinverse

- We should be able to verify at the very least that it is a g-inverse.
- We also should be able to verify that it satisfies the additional properties of the Moore-Penrose pseudoinverse.