Mathematics/Statistics Bootcamp

Part I: Linear Algebra

Steven Winter   Christine Shen

Department of Statistical Science
Duke University

MSS Orientation, August 2022

Motivation

The real world is non-linear. Non-linear objects are usually:

- Difficult to study mathematically.
- Difficult to solve numerically.

We can get quite far with linear approximations. Linear objects are usually:

- Easy to study mathematically.
- Easy[1] to solve numerically.

Linear algebra studies linear/vector spaces and linear transformations.

[1] Nothing is easy in 1 million dimensions.

Outline

Basic Linear Algebra

Vector Spaces

A **real vector space** V is a set equipped with two functions $+ : V \times V \to V$ and $\cdot : \mathbb{R} \times V \to V$ satisfying

1. $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$,
2. $\mathbf{u} + (\mathbf{v} + \mathbf{w}) = (\mathbf{u} + \mathbf{v}) + \mathbf{w}$,
3. There exists $\mathbf{0} \in V$ such that $\mathbf{0} + \mathbf{v} = \mathbf{v}$,
4. For any $\mathbf{v} \in V$, there exists $-\mathbf{v} \in V$ such that $\mathbf{v} + (-\mathbf{v}) = \mathbf{0}$,
5. $a \cdot (b \cdot \mathbf{v}) = (ab) \cdot \mathbf{v}$,
6. $1 \cdot \mathbf{v} = \mathbf{v}$,
7. $a \cdot (\mathbf{u} + \mathbf{v}) = a \cdot \mathbf{u} + a \cdot \mathbf{v}$,
8. $(a + b) \cdot \mathbf{v} = a \cdot \mathbf{v} + b \cdot \mathbf{v}$.

Usually $V = \mathbb{R}^n$, and $+, \cdot$ are defined coordinate-wise. Random variables with pth moments also form a vector space called L^p.

Linear Transformations

Let V, W be real vector spaces. A **linear transformation** is a function $T : V \to W$ such that

1. $T(\mathbf{u} + \mathbf{v}) = T(\mathbf{u}) + T(\mathbf{v})$,
2. $T(c \cdot \mathbf{v}) = c \cdot T(\mathbf{v})$.

Examples: $V = W = \mathbb{R}^3$. Which are linear?

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} \mapsto \begin{pmatrix} x \\ y \\ 0 \end{pmatrix}, \quad \begin{pmatrix} x \\ y \\ z \end{pmatrix} \mapsto \begin{pmatrix} 3x - y/2 \\ y + z \\ x - y \end{pmatrix}, \quad \begin{pmatrix} x \\ y \\ z \end{pmatrix} \mapsto \begin{pmatrix} \log(x) \\ \log(y) \\ \log(z) \end{pmatrix}$$

Linear Independence

A set of vectors $\{\mathbf{v}_1, ..., \mathbf{v}_N\}$ is **linearly dependent** if there exist scalars $c_1, c_2, \ldots, c_N$, not all equal to zero, such that

$$\sum_{i=1}^{N} c_i \mathbf{v}_i = 0$$

For example,

$$\begin{pmatrix} -1 \\ -1 \end{pmatrix} \text{ and } \begin{pmatrix} \pi \\ \pi \end{pmatrix}$$

are linearly dependent as elements of $V = \mathbb{R}^2$.

If no such scalars exist, the set is said to be **linearly independent**.

Basis

Recall

$$\operatorname{span}_{\mathbb{R}}(\{\mathbf{v}_1, \dots, \mathbf{v}_N\}) = \left\{ \sum_{i=1}^{N} c_i \mathbf{v}_i \mid c_1, \dots, c_N \in \mathbb{R} \right\}.$$

A set of vectors $\{\mathbf{v}_1, \dots, \mathbf{v}_N\}$ forms a **basis** for a vector space V if it is linearly independent and spans V.

The number of basis vectors, $\dim(V)$, is the **dimension** of V.

Almost always, $V = \mathbb{R}^n$ with standard basis

$$\begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \dots, \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}.$$

Dot Products

The **dot product** is a function $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ defined by

$$\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^T \mathbf{v} = \sum_{i=1}^{n} u_i v_i.$$

Vectors are **orthogonal** if $\mathbf{u}^T \mathbf{v} = 0$.

The average of a vector can be written as

$$\bar{\mathbf{v}} = (\mathbf{1}^T \mathbf{1})^{-1} \mathbf{1}^T \mathbf{v} = \frac{1}{n} \sum_{i=1}^{n} v_i.$$

Inner products (specifically kernels) are *very* useful in statistics: covariances, feature expansion (Mercer's theorem), building Gaussian processes, etc.

Norms

The dot product induces the Euclidean **norm**,

$$||\mathbf{v}||_2 = \sqrt{\mathbf{v}^T\mathbf{v}} = \sqrt{\sum_{i=1}^{n} v_i^2}$$

Recall

1. $||c\mathbf{v}||_2 = c||\mathbf{v}||_2$,
2. $||\mathbf{v}||_2 = 0$ if and only if $\mathbf{v} = \mathbf{0}$,
3. $||\mathbf{u} + \mathbf{v}||_2 \leq ||\mathbf{u}||_2 + ||\mathbf{v}||_2$,
4. $\Big| ||\mathbf{u}||_2 - ||\mathbf{v}||_2 \Big| \leq ||\mathbf{u} - \mathbf{v}||_2$.
5. $|\langle \mathbf{u}, \mathbf{v} \rangle| \leq ||\mathbf{u}||_2 ||\mathbf{v}||_2$. (Cauchy-Schwarz)

A vector is a **unit vector** if $||\mathbf{v}||_2 = 1$.

Basic Matrix Theory

Notation

A matrix represents a linear transformation $T : V \to W$ in a fixed basis. Always assume $V = \mathbb{R}^n$, $W = \mathbb{R}^m$ with the standard basis.

Write

$$\mathbf{A} = (a_{ij}) = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{pmatrix} \in \mathbb{R}^{m \times n}$$

Matrix operations follow naturally from properties of linear transformations.

Fundamental Subspaces

A subset S of a vector space V is a **subspace** if it is also a vector space. E.g., $\mathbb{R} \subseteq \mathbb{R}^2$.

Fix $\mathbf{A} \in \mathbb{R}^{m \times n}$. The **column space**, $C(\mathbf{A})$ is the subspace of $\mathbb{R}^m$ spanned by the columns of $\mathbf{A}$. By definition,

$$C(\mathbf{A}) = \{\mathbf{A}\mathbf{v} \mid \mathbf{v} \in \mathbb{R}^n\}.$$

The **row space**, $C(\mathbf{A}^T)$, is defined similarly.

The **rank** of $\mathbf{A}$ is the dimension of the column space (*equivalently the row space*). An $n \times n$ matrix $\mathbf{A}$ is **full rank** if $\text{rank}(\mathbf{A}) = n$. This is equivalent to being invertible.

Rank Nullity

The **null space**, $N(\mathbf{A})$, is the vector subspace of $\mathbb{R}^n$ defined by

$$N(\mathbf{A}) = \{\mathbf{v} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{v} = \mathbf{0}\}.$$

The null space is orthogonal to the row space: if $\mathbf{A}\mathbf{v} = \mathbf{0}$ and $\mathbf{u} = \mathbf{A}^T\mathbf{w} \in C(\mathbf{A}^T)$, then

$$\mathbf{v}^T\mathbf{u} = \mathbf{v}^T\mathbf{A}^T\mathbf{w} = (\mathbf{A}\mathbf{v})^T\mathbf{w} = \mathbf{0}.$$

The **rank-nullity theorem** says

$$\dim(C(\mathbf{A})) + \dim(N(\mathbf{A})) = n$$

Example

Consider

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

Then $\dim(C(\mathbf{A})) \leq 4$, $\dim(C(\mathbf{A}^T)) \leq 3$, so the rank is at most 3.

The column space includes

$$\begin{pmatrix} 4 \\ -2 \\ -3 \end{pmatrix} = 2\begin{pmatrix} 2 \\ -1 \\ 0 \end{pmatrix} + 0\begin{pmatrix} -4 \\ 2 \\ 0 \end{pmatrix} - 3\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} + 0\begin{pmatrix} 0 \\ 0 \\ 2 \end{pmatrix}$$

but not $(2,0,0)^T$. What is the dimension of the column space? Basis? Rank? Dimension of null space? Basis?

Matrix Addition

Corresponds to adding linear transformations. Find sums element-wise:

$$\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n} \implies (A + B)_{ij} = a_{ij} + b_{ij}.$$

Associative and commutative:

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

Typically $O(n^2)$.

Matrix Multiplication

Corresponds to composing linear transformations. Multiply by dotting rows and columns:

$$\mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{B} \in \mathbb{R}^{n \times q} \implies (\mathbf{AB})_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}.$$

Equivalently: $\mathbf{AB} = \sum_{i=1}^{n} \mathbf{a}_i \mathbf{b}^i$. Get $\mathbf{AB} \in \mathbb{R}^{m \times q}$.

For example:

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 5 & 6 & 7 \\ 8 & 9 & 10 \end{pmatrix} = \begin{pmatrix} 1(5)+2(8) & 1(6)+2(9) & 1(7)+2(10) \\ 3(5)+4(8) & 3(6)+4(9) & 3(7)+4(10) \end{pmatrix}$$

Naively $O(n^3)$.

Matrix Multiplication Properties

Associative, but generally not commutative:

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

$$\mathbf{A}\mathbf{B} \neq \mathbf{B}\mathbf{A} \quad \text{(usually)}$$

Respects addition

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

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

and scalar multiplication

$$c(\mathbf{A} + \mathbf{B}) = c\mathbf{A} + c\mathbf{B}$$

$$c\mathbf{A}\mathbf{B} = (c\mathbf{A})\mathbf{B} = \mathbf{A}(c\mathbf{B})$$

The **identity matrix**, $\mathbf{I} = \mathrm{diag}(1, ..., 1)$, satisfies $\mathbf{I}\mathbf{A} = \mathbf{A}\mathbf{I} = \mathbf{A}$.

Matrix Inversion

Corresponds to inverting linear transformations. A matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is **invertible** (or **nonsingular**) if and only if $\exists \mathbf{A}^{-1} \in \mathbb{R}^{n \times n}$ such that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$.

For example:

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \implies \mathbf{A}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -c \\ -b & a \end{pmatrix}$$

Naively $O(n^3)$. Try to avoid it entirely if you're solving $\mathbf{Ax} = \mathbf{b}$.[2]

[2] `http://gregorygundersen.com/blog/2020/12/09/matrix-inversion/`

Matrix Inversion Properties

Let $\mathbf{A}, \mathbf{B}$ be nonsingular and $c \neq 0$. Then

$$
\begin{aligned}
(\mathbf{A}^{-1})^{-1} &= \mathbf{A} \\
(c\mathbf{A})^{-1} &= \frac{1}{c}\mathbf{A}^{-1} \\
(\mathbf{A}\mathbf{B})^{-1} &= \mathbf{B}^{-1}\mathbf{A}^{-1} \\
(\mathbf{A}_1\mathbf{A}_2\cdots\mathbf{A}_N)^{-1} &= \mathbf{A}_N^{-1}\mathbf{A}_{N-1}^{-1}\cdots\mathbf{A}_1^{-1}
\end{aligned}
$$

Transposes

Corresponds to the adjoint/dual linear transformation. Swap rows and columns: if $\mathbf{A} \in \mathbb{R}^{m \times n}$, then $\mathbf{A}^T \in \mathbb{R}^{n \times m}$ and $(\mathbf{A}^T)_{ij} = a_{ji}$.
Useful properties:

$$
\begin{aligned}
(\mathbf{A}^T)^T &= \mathbf{A} \\
(\mathbf{A} + \mathbf{B})^T &= \mathbf{A}^T + \mathbf{B}^T \\
(\mathbf{A}\mathbf{B})^T &= \mathbf{B}^T \mathbf{A}^T \\
(\mathbf{A}_1 \mathbf{A}_2 \cdots \mathbf{A}_N)^T &= \mathbf{A}_N^T \mathbf{A}_{N-1}^T \cdots \mathbf{A}_1^T \\
(\mathbf{A}^{-1})^T &= (\mathbf{A}^T)^{-1}
\end{aligned}
$$

Exercises

1. a) Fix $\gamma > 0$ and let $V = \mathbb{R}^3$. Find a set of linearly independent vectors $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ such that $\langle \mathbf{v}_i, \mathbf{v}_j \rangle = \gamma$ for all $i \neq j$.

   b) (Bonus) Let V be the vector space of smooth functions:

   $$V = \{f : \mathbb{R} \to \mathbb{R} \mid \text{all derivatives of } f \text{ exist and are continuous}\}$$

   equipped with pointwise addition and the usual scalar multiplication. Are $f(t) = e^t$ and $g(t) = -3e^{2t}$ linearly independent? Prove/disprove.

2. Verify a special case of the Sherman–Morrison–Woodbury formula: $(\mathbf{I} + \mathbf{UV})^{-1} = \mathbf{I} - \mathbf{U}(\mathbf{I} + \mathbf{VU})^{-1}\mathbf{V}$.

3. Prove $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$ by definition.

Special Matrices

Special Matrices

Some common structures:

- A matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$ is **square** if $n = m$. Write $\mathbf{A}^k$ for $\mathbf{A}\mathbf{A} \cdots \mathbf{A}$.
- A square matrix $\mathbf{A}$ is **diagonal** if $i \neq j \implies a_{ij} = 0$.
- The **identity** matrix $\mathbf{I}$ is diagonal with all diagonal elements equal to 1. Recall $\mathbf{A}\mathbf{I} = \mathbf{I}\mathbf{A} = \mathbf{A}$.
- A square matrix $\mathbf{A}$ is **symmetric** if $\mathbf{A}^T = \mathbf{A}$. E.g., covariances.
- A square matrix $\mathbf{A}$ is **idempotent** if $\mathbf{A}^2 = \mathbf{A}$.
- An invertible matrix $\mathbf{A}$ is **orthogonal** (or **orthonormal**) if $\mathbf{A}^T = \mathbf{A}^{-1}$. E.g., rotations, reflections, permutations.
- Triangular matrices, partitioned matrices, quadratic forms, projection matrices, etc.

Triangular Matrices

A square matrix $\mathbf{U}$ is **upper triangular** if $i > j \implies u_{ij} = 0$. For example:

$$\mathbf{U} = \begin{pmatrix} u_{11} & u_{12} & u_{13} & \dots & u_{1n} \\ 0 & u_{22} & u_{23} & \dots & u_{2n} \\ 0 & 0 & u_{33} & \dots & u_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & u_{nn} \end{pmatrix}$$

Inversion and solving $\mathbf{Ux} = \mathbf{b}$ is $O(n^2)$. **Lower triangular** matrices defined analogously.

Partitioned Matrices

Obtain a **submatrix** of $\mathbf{A}$ by deleting rows and/or columns. A **partitioned matrix** has the following decomposition:

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \dots & \mathbf{A}_{1c} \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \dots & \mathbf{A}_{2c} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}_{r1} & \mathbf{A}_{r2} & \dots & \mathbf{A}_{rc} \end{pmatrix}$$

where the submatrix $\mathbf{A}_{ij}$ is referred to as the ijth block of $\mathbf{A}$. All operations (e.g., multiplication) pass to submatrices.

Quadratic Forms

Let $\mathbf{A}$ be a square symmetric matrix. A **quadratic form** is a function mapping vectors to scalars:

$$\mathbf{x} \mapsto \mathbf{x}^T \mathbf{A} \mathbf{x}.$$

If $\mathbf{x}^T \mathbf{A} \mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$, then $\mathbf{A}$ is **positive definite** (PD). If instead $\mathbf{x}^T \mathbf{A} \mathbf{x} \geq 0$, then $\mathbf{A}$ is **positive semi-definite** (PSD).

Covariance matrices must be PSD.[3]

[3] Exercise: prove this after the probability session.

Projection Matrices

A **projection** matrix $\mathbf{P}$ is an idempotent matrix: $\mathbf{P}^2 = \mathbf{P}$.

An orthogonal projection is a projection that is symmetric: $\mathbf{P}^T = \mathbf{P}$. Can show an orthogonal projection $\mathbf{P}$ sends a vector to the closest point in $C(\mathbf{P})$ (see board).

Extremely important in statistics - e.g., linear regression.

Quick Exercises

Let $\mathbf{P}$ be an orthogonal projection.

1. Show $\mathbf{I} - \mathbf{P}$ is also an orthogonal projection.
2. Show $(\mathbf{I} - \mathbf{P})^T \mathbf{P} = \mathbf{0}$.
3. Show $\mathbf{P}\mathbf{v} = \mathbf{v}$ for $\mathbf{v} \in C(\mathbf{P})$.

Key Example: Linear Models

Linear Models

We have a response y_i (e.g., lifespan) and covariates $\mathbf{x}_i \in \mathbb{R}^p$ (e.g., heart rate, blood pressure, etc) for individuals $i = 1, \ldots n$.

Try modeling y_i as a *linear combination* of the $\mathbf{x}_i$, plus noise:

$$y_i = \boldsymbol{\beta}^T \mathbf{x}_i + \varepsilon_i$$

Here $\boldsymbol{\beta} \in \mathbb{R}^p$ are unknown regression **coefficients** and the ε_i are unobserved mean zero **errors**.

Goal: understand how changing $\mathbf{x}$ influences $\mathbf{y}$.

Ordinary Least Squares

Let $\mathbf{Y} = (y_1, ..., y_n)^T$, and $\mathbf{X} \in \mathbb{R}^{n \times p}$ have rows $\mathbf{x}_1^T, ..., \mathbf{x}_n^T$. We can write the linear model as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$$

or equivalently $E[\mathbf{Y}] \in C(\mathbf{X})$.

How to estimate $\boldsymbol{\beta}$? Often we minimize the **residual sum of squares**,

$$\text{RSS}(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \boldsymbol{\beta}^T \mathbf{x}_i)^2 = ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||_2^2$$

Calculus approach: compute $d\text{RSS}(\boldsymbol{\beta})/d\boldsymbol{\beta}$, set to zero, etc.

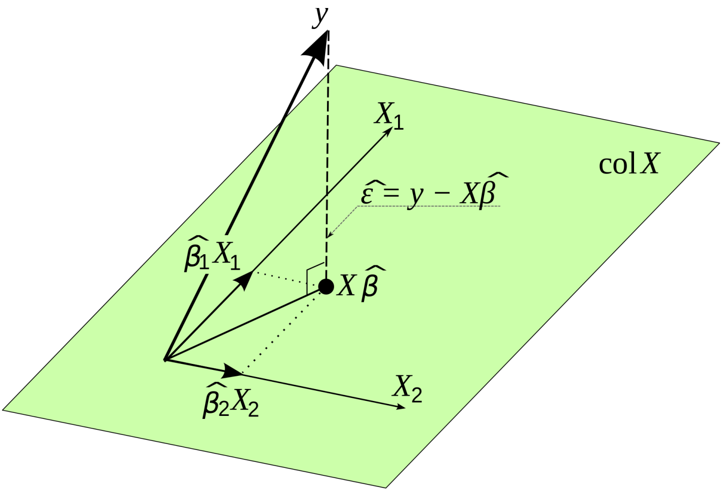
OLS via Projections

Let P be an orthogonal projection with $C(\mathbf{P}) = C(\mathbf{X})$.Then

$$
\begin{aligned}
\mathrm{RSS}(\beta) &= ||\mathbf{Y} - \mathbf{X}\beta||_2^2 \\
&= ||(\mathbf{Y} - \mathbf{PY}) + (\mathbf{PY} - \mathbf{X}\beta)||_2^2 \\
&= ||(\mathbf{I} - \mathbf{P})\mathbf{Y} + \mathbf{P}(\mathbf{Y} - \mathbf{X}\beta)||_2^2 \\
&= \mathbf{Y}^T(\mathbf{I} - \mathbf{P})^T(\mathbf{I} - \mathbf{P})\mathbf{Y} + 2\mathbf{Y}^T(\mathbf{I} - \mathbf{P})^T\mathbf{P}(\mathbf{Y} - \mathbf{X}\beta) \\
&\quad + (\mathbf{Y} - \mathbf{X}\beta)^T\mathbf{P}^T\mathbf{P}(\mathbf{Y} - \mathbf{X}\beta) \\
&= ||(\mathbf{I} - \mathbf{P})\mathbf{Y}||_2^2 + 0 + ||\mathbf{P}(\mathbf{Y} - \mathbf{X}\beta)||_2^2 \\
&= ||(\mathbf{I} - \mathbf{P})\mathbf{Y}||_2^2 + ||\mathbf{PY} - \mathbf{X}\beta||_2^2 \\
&\geq ||(\mathbf{I} - \mathbf{P})\mathbf{Y}||_2^2
\end{aligned}
$$

The minimizer $\hat{\beta}$ satisfies $\mathbf{PY} = \mathbf{X}\hat{\beta}$. No calculus!

OLS Geometry



From
`https://en.wikipedia.org/wiki/Ordinary_least_squares`.

Exercises

Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ have rank $p \leq n$ (so $\mathbf{X}^T\mathbf{X}$ is invertible). Consider the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$.

1. Show $\mathbf{P}_\mathbf{X} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is an orthogonal projection matrix and $\mathbf{P}_\mathbf{X}\mathbf{X} = \mathbf{X}$. Guess $C(\mathbf{P}_\mathbf{X})$ and $C(\mathbf{I} - \mathbf{P}_\mathbf{X})$ but don't worry about proving it.
2. Assume $\mathbf{P}_\mathbf{X}\mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\beta}}$. Does this imply $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$? Why or why not?
3. Now assume $\mathbf{X} = [\mathbf{1} \quad \mathbf{z}] \in \mathbb{R}^{n \times 2}$ for some $\mathbf{z} \in \mathbb{R}^n$. Describe the model in words. Calculate $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \in \mathbb{R}^2$ and interpret these values. How do things simplify if $\mathbf{z}$ has mean zero?

Intermediate Matrix Theory

Trace

The **trace** is a function $\mathrm{Tr} : \mathbb{R}^{n \times n} \to \mathbb{R}$ defined by summing the diagonal elements:

$$\mathrm{Tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$$

Some properties of the trace are

$$\begin{aligned}
\mathrm{Tr}(c\mathbf{A}) &= c\,\mathrm{Tr}(\mathbf{A}) \\
\mathrm{Tr}(\mathbf{A} + \mathbf{B}) &= \mathrm{Tr}(\mathbf{A}) + \mathrm{Tr}(\mathbf{B}) \\
\mathrm{Tr}(\mathbf{A}^T) &= \mathrm{Tr}(\mathbf{A}) \\
\mathrm{Tr}(\mathbf{AB}) &= \mathrm{Tr}(\mathbf{BA}) \\
\mathrm{Tr}(\mathbf{A}_1\mathbf{A}_2 \cdots \mathbf{A}_N) &= \mathrm{Tr}(\mathbf{A}_N\mathbf{A}_1\mathbf{A}_2 \cdots \mathbf{A}_{N-1})
\end{aligned}$$

Defining Determinants

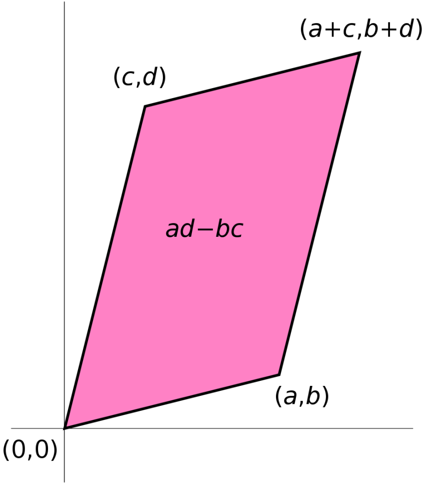
Let

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

The **determinant** is defined as

$$|\mathbf{A}| = \det(\mathbf{A}) = ad - bc$$

Determinant Geometry



From `https://en.wikipedia.org/wiki/Determinant`.

Extending to Square Matrices

The **minor** $\mathbf{M}_{ij}$ of a_{ij} is the $n-1 \times n-1$ matrix that is formed by removing the ith row and jth column from $\mathbf{A}$. Determinants for $n \times n$ matrices are found with cofactor expansion:

$$|\mathbf{A}| = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} |\mathbf{M}_{ij}|$$

Properties:

$$|\mathbf{A}^T| = |\mathbf{A}|$$
$$|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}| = |\mathbf{B}||\mathbf{A}| = |\mathbf{BA}|$$
$$|c\mathbf{A}| = c^n |\mathbf{A}|$$
$$\mathbf{A} \text{ singular} \iff |\mathbf{A}| = 0$$
$$|\mathbf{A}^{-1}| = \frac{1}{|\mathbf{A}|}$$

Eigenvalues and Eigenvectors

Let $\mathbf{A}$ be a square matrix. If there is a vector $\mathbf{v} \neq \mathbf{0}$ such that

$$\mathbf{A}\mathbf{v} = \lambda \mathbf{v}.$$

for some scalar λ, then λ is called an eigenvalue with eigenvector $\mathbf{v}$. Eigenvectors are special vectors that are stretched, but not rotated.

The rank of $\mathbf{A}$ is the number of nonzero eigenvalues.

The set of eigenvalues is called the **spectrum** of $\mathbf{A}$.

Spectral Theorem (Eigendecomposition)

Let $\mathbf{A}$ be an invertible $n \times n$ symmetric square matrix. We can always choose orthonormal eigenvectors $\mathbf{v}_1, ..., \mathbf{v}_n$ for eigenvalues $\lambda_1 \geq ... \geq \lambda_n$. This gives the unique decomposition

$$\mathbf{A} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i^T$$

where $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, ..., \lambda_n)$ and $\mathbf{V}$ has columns $\mathbf{v}_1, ..., \mathbf{v}_n$. Still works if $\mathbf{A}$ is not symmetric, but then $\mathbf{A} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{-1}$.

Computational complexity? Geometric interpretations?.

Note $\mathbf{V}^T\mathbf{V} = \mathbf{I}$. *Very* important in statistics - e.g., if $\mathbf{Y} \sim N(\mathbf{0}, \mathbf{A})$ then $\mathbf{V}^T\mathbf{Y} \sim N(\mathbf{0}, \boldsymbol{\Lambda})$. Entries become independent! Also PCA.

Application: Pseudoinverses and Square Roots

A **pseudoinverse** of $\mathbf{A}$ is a matrix $\mathbf{G}$ satisfying

$$\mathbf{AGA} = \mathbf{A}.$$

If $\mathbf{A}$ is invertible then $\mathbf{G} = \mathbf{A}^{-1}$ is the unique pseudoinverse. Otherwise there are infinitely many $\mathbf{G}$.

Most common is the **Moore-Penrose inverse** for a symmetric[4] matrix $\mathbf{A}$:

$$\mathbf{G} = \mathbf{V}\boldsymbol{\Lambda}^{-}\mathbf{V}^T$$

where $\boldsymbol{\Lambda}^{-} = \mathrm{diag}(1/\lambda_1, ..., 1/\lambda_k, 0, ..., 0)$. Useful when $\mathbf{X}^T\mathbf{X}$ is singular (e.g., OLS).

Ideas for defining $\mathbf{A}^{1/2}$?

[4] General case via SVD.

SVD

The **singular value decomposition** generalizes the eigendecomposition. Factor $\mathbf{A} \in \mathbb{R}^{m \times n}$ as

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T$$

where $\mathbf{U} \in \mathbb{R}^{m \times m}, \mathbf{V} \in \mathbb{R}^{n \times n}$ are such that $\mathbf{U}^T\mathbf{U} = \mathbf{I_m}$, $\mathbf{V}^T\mathbf{V} = \mathbf{I_n}$, and $\mathbf{D} \in \mathbb{R}^{m \times n}$ is a nonnegative rectangular diagonal matrix of **singular values** $d_1 \geq \cdots \geq d_n$.

How are $\mathbf{U}, \mathbf{D}, \mathbf{V}$ related to the eigendecompositions of $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}\mathbf{A}^T$?

Compact SVD (Optional)

The **compact singular value decomposition** factors a rank r matrix as $\mathbf{A} \in \mathbb{R}^{m \times n}$ as

$$\mathbf{A} = \mathbf{U}_r \mathbf{D}_r \mathbf{V}_r^T$$

where $\mathbf{U}_r \in \mathbb{R}^{m \times r}, \mathbf{V} \in \mathbb{R}^{n \times r}$ are such that $\mathbf{U}^T \mathbf{U} = \mathbf{I_r}$, $\mathbf{V}^T \mathbf{V} = \mathbf{I_r}$, and $\mathbf{D} \in \mathbb{R}^{r \times r}$ is a nonnegative square diagonal matrix of nonzero singular values $d_1 \geq \cdots \geq d_r$.

Can write

$$\mathbf{A} = \sum_{i=1}^{r} d_i \mathbf{u}_i \mathbf{v}_i^T.$$

Cholesky Decomposition

We can write any symmetric PSD matrix (e.g., covariances) as

$$\mathbf{A} = \mathbf{L}\mathbf{L}^T$$

where $\mathbf{L}$ is lower triangular. Naively $O(n^3)$

Can efficiently simulate multivariate normals after you have $\mathbf{L}$: if $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{I})$, then $\boldsymbol{\mu} + \mathbf{L}\mathbf{Z} \sim N(\boldsymbol{\mu}, \mathbf{L}\mathbf{L}^T)$.

If you have $\mathbf{A} = \mathbf{L}\mathbf{L}^T$, then you can find the Cholesky of

$$a\mathbf{A} + b\mathbf{v}\mathbf{v}^T$$

in $O(n^2)$.[5] *Order of magnitude faster* for adaptive Metropolis, approximating Gaussian processes, etc.

[5] "A More Efficient Rank-one Covariance Matrix Update for Evolution Strategies" by Oswin Krause and Christian Igel.

Other Decompositions

Many other ways to decompose a matrix:

1. LU decomposition for a square matrix: $\mathbf{A} = \mathbf{LU}$ with $\mathbf{L}$ lower triangular and $\mathbf{U}$ upper triangular. Good for solving equations.
2. QR decomposition for a general $m \times n$ matrix: $\mathbf{A} = \mathbf{QR}$, where $\mathbf{Q}$ is an orthogonal $m \times m$ matrix and $\mathbf{R}$ is an upper triangular $m \times n$ matrix. Useful for least squares.
3. Polar decomposition for a general $m \times n$ matrix: $\mathbf{A} = \mathbf{QS}^{1/2}$ where $\mathbf{Q}$ is an orthogonal $m \times n$ matrix and $\mathbf{S}$ is a symmetric square root of $\mathbf{A}^T\mathbf{A}$. Good for sampling orthogonal matrices.

Warning: matrix decomposition functions will often pad or transpose the things you want. For example: `np.linalg.svd` both pads the singular vectors and returns $\mathbf{V}^T$.

Exercises

1. Prove a symmetric matrix is PSD if and only if all eigenvalues are non-negative.
2. Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be symmetric with eigenvalues $\lambda_1, ..., \lambda_n$. Prove

$$\mathrm{Tr}(\mathbf{A}) = \sum_{i=1}^{n} \lambda_i \quad \text{and} \quad |\mathbf{A}| = \prod_{i=1}^{n} \lambda_i.$$

   Bonus: prove it without assuming symmetry.
3. Let $\mathbf{P} \in \mathbb{R}^{n\times n}$ be a singular[6] projection matrix of rank $k < n$. Find all eigenvalues of $\mathbf{P}$. Use this to find $|\mathbf{I} + c\mathbf{P}|$.

[6] Bonus exercise: prove $\mathbf{I}$ is the only full rank projection matrix.

Useful References

- *Mathematics for Machine Learning* - Garrett Thomas[7]
- *Matrix Algebra from a Statistician's Perspective* - Harville
- *The Matrix Cookbook* - Petersen and Pedersen

[7] Professor recommended for 521!

Acknowledgements

Past contributors:

- Jordan Bryan, PhD student
- Brian Cozzi, MSS alumni
- Michael Valancius, MSS alumni
- Graham Tierney, PhD student
- Becky Tang, PhD student