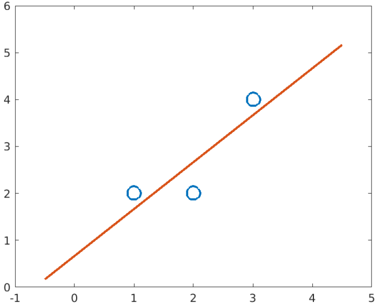
Week 11: Best Approximation and Least Squares

Dr Niall Madden

15 and 18 November, 2022



These slides are adapted (slightly) from ones by Tobias Rossmann.

Outline

For more details,

- Section 6.3 (Best Approximation) and 6.6 (Least Squares) in Lay et al:
  https://nuigalway-primo.hosted.exlibrisgroup.com/permalink/f/1pmb9lf/353GAL_ALMA_DS5192067630003626
- Chapters 10 and 11 of *Linear Algebra for Data Science*
  https://shainarace.github.io/LinearAlgebra/leastsquares.html

MA313
Week 11: Best Approximation and Least Squares

Start of ...

PART 1: Announcements and Preview of Week 11

Assignment 5

Assignment 5 opened on Thursday 10 Nov). Deadline is 5pm, Friday, 25th of November.

Communication Skills : Next steps...

- Instructions at `https://www.niallmadden.ie/2223-MA313/22_23_Communication_Skills.pdf` have been updated.
- Deadline is 5pm Friday, 18 November.
- Presentations will be during the week 21–25 November:
  - Monday at 12.00 in AC204 (i.e., MA335 class time)
  - Tuesday at 13.00 in AC202 (i.e., MA313 class time)
  - Thursday at 12.00 in IT206 (i.e., MA313 tutorial time)

Next week

- Tuesday's and Thursdays classes will be used for presentations.
- Friday's class will be used to review the module and preview the exam.
- I'll also provide some sample exam-type questions, with solutions, and video.

The big ideas from this week will be solving Least Squares Problems.

- Why it is that the orthogonal projection is the best solution.
- How to find it.

These are the essential ideas from recent lectures that you need for this week.

- The **INNER PRODUCT** of vectors u and v in $\mathbb{R}^n$ is the real number given by

$$u \cdot v = u^T v = u_1 v_1 + u_2 v_2 + \cdots + u_n v_n.$$

- The **LENGTH** (or "Euclidian norm") of a vector $v \in \mathbb{R}^n$ is $\|v\| := \sqrt{v \cdot v} = \sqrt{v_1^2 + \cdots + v_n^2}$. If $\|u\| = 0$ that means all the entries in u are zero.
- The **distance** between vectors $u, v \in \mathbb{R}^n$ is $\|u - v\| = u \cdot v = u^\top v$
- $u, v \in \mathbb{R}^n$ are **orthogonal** if $u \cdot v = 0$. We may write this as $u \perp v$.
- **Pythagorean Theorem:** If $u \perp v$, then $\|u + v\|^2 = \|u\|^2 + \|v\|^2$.

- Given a subspace, W, of $\mathbb{R}^n$, the vector $z \in \mathbb{R}^n$ is **orthogonal** to W if $z \perp w$ for all $w \in W$.
- In particular, given a matrix A, if z is orthogonal to $\mathrm{Col}\, A$, then $a_j \perp z$ for a_j is column j of A. That is $a_j^\top z = 0$. Since this is true of any j, in fact $A^\top z = 0$.
- Every vector $v \in \mathbb{R}^n$ has a **unique representation**

$$v = \hat{v} + z \qquad \text{for } \hat{v} \in W, \qquad \text{and } z \perp W.$$

MA313
Week 11: Best Approximation and Least Squares

Start of ...

PART 2: Orthogonal Matrices

This is actually left over from last week... I'll skim through it.

Definition: ORTHONORMAL

The vectors $u_1, \ldots, u_p \in \mathbb{R}^n$ are **orthonormal** if they are orthogonal unit vectors. That is:

- $u_i \perp u_j$ for all $i \neq j$. Equivalently, $u_i \cdot u_j = 0$ for all $i \neq j$.
- $\|u_i\| = 1$ for all i.

Note: If $u_1, \ldots, u_p$ are orthogonal and all non-zero, then $\frac{1}{\|u_1\|}u_1, \ldots, \frac{1}{\|u_p\|}u_p$ are orthonormal.

Definition: ORTHONORMAL BASIS

An **orthonormal basis** of a subspace W of $\mathbb{R}^n$ is a basis of W that consists of orthonormal vectors.

Example: The standard basis of $\mathbb{R}^n$ is orthonormal.

Theorem

Let A be an $n \times n$ matrix. Then the following are equivalent:

(a) The columns of A form an orthonormal basis of $\mathbb{R}^n$.

(b) $A^\top A = I_n = AA^\top$. (That is, A is invertible and $A^{-1} = A^\top$.)

(c) $Ax \cdot Ay = x \cdot y$ for all $x, y \in \mathbb{R}^n$.

(d) $\|Ax\| = \|x\|$ for all $x \in \mathbb{R}^n$

Definition: ORTHOGONAL MATRIX

An $n \times n$ matrix A is **orthogonal** if $A^\top A = I_n$ (in which case also $AA^\top = I_n$).

Example

- Reflections: e.g. $\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$.
- Rotations: $\begin{bmatrix} \cos\vartheta & -\sin\vartheta \\ \sin\vartheta & \cos\vartheta \end{bmatrix}$ for $\vartheta \in \mathbb{R}$, e.g. $\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$ or $\begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$.

MA313
Week 11: Best Approximation and Least Squares

Start of ...

PART 3: Best Approximation

What are orthogonal projections used for?

Part 3: Best Approximation

Let W by a subspace of $\mathbb{R}^n$.

The **orthogonal projection** of a vector $v \in \mathbb{R}^n$ is denoted $\hat{v} = \mathrm{proj}_W\, v$.

It has the property that $(v - \hat{v}) \perp W$.

Question

What are orthogonal projections good for?

Hint: take W to be a one-dimensional subspace of R^2.

Best Approximation Theorem

Let W be a subspace of $\mathbb{R}^n$. Let

$$\mathrm{proj}_W \colon \mathbb{R}^n \to W, \quad v \mapsto \hat{v}$$

be the orthogonal projection onto W. Then for any $v \in \mathbb{R}^n$,

$$\|v - \hat{v}\| \leqslant \|v - w\| \qquad \text{for any } w \in W,$$

with equality if and only if $w = \hat{v}$.
Hence: $\hat{v}$ is the unique vector in W which minimises the distance from v.

Part 3: Best Approximation

Proof

- We want to show that $\|v - \hat{v}\| \leqslant \|v - w\|$. for any $w \in W$.
- Since $\hat{v}$ is the orthogonal projection of v onto W, we know that $(v - \hat{v}) \perp W$.
- Also, both $\hat{v}$ and w are in W, so $\hat{v} - w \in W$.
- It follows that $(v - \hat{v}) \perp (\hat{v} - w)$.
- So we can apply Pythagoras' Theorem:

$$\|(v - \hat{v}) + (\hat{v} - w)\|^2 = \|v - \hat{v}\|^2 + \|\hat{v} - w\|^2.$$

- That gives $\|v - w\|^2 = \|v - \hat{v}\|^2 + \|\hat{v} - w\|^2$.
- But $\|\hat{v} - w\| \geq 0$, so we can conclude $\|v - w\|^2 \geq \|v - \hat{v}\|^2$.

(Note: see diagram on next slide).

Part 3: Best Approximation

MA313
Week 11: Best Approximation and Least Squares

Start of ...

PART 4: Least Squares Problems

Part 4: Least Squares Problems

Motivation

Suppose that some *mathematical model* of a phenomenon predicts that it can be described by the equation of a line: $y = u_1 + u_2 x$ for some unknown coefficients $u_1, u_2 \in \mathbb{R}$.

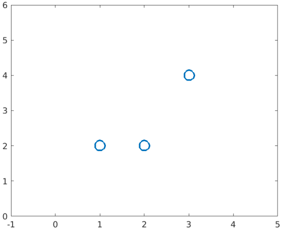
By taking **measurements**, we find points $\begin{bmatrix} x_1 \\ y_1 \end{bmatrix}, \ldots, \begin{bmatrix} x_N \\ y_N \end{bmatrix}$ that the line should fit. Due to measurement errors, there might not be any pair u_1, u_2 with $u_1 + u_2 x_i = y_i$ for all $i = 1, \ldots, N$. The system is usually **over determined**, meaning there are too many equations to be satisfied at the same time.

There is no solution for works for all equations, so we try to find the *best approximation*.

Part 4: Least Squares Problems

Example

Suppose we wanted to find the line, $y = u_1 + u_2 x$ that best fits the data $(1, 2)$, $(2, 2)$, and $(3, 4)$. Write down a matrix-vector equation for this problem.



Part 4: Least Squares Problems

We would like to solve a problem $Ax = b$, meaning that we try to find x. If we could, then $\|b - Ax\| = 0$.

But there is no solution. So we try to find the $\hat{x}$ that makes $A\hat{x} - b$ as small as possible.

Defn: LINEAR LEAST-SQUARES PROBLEM

Given an $m \times n$ matrix A and a vector $b \in \mathbb{R}^m$, the associated **linear least-squares problem** is to minimise the length of the **residual** (also called "approximation error"), $\|A\hat{x} - b\|$, to an exact solution "$Ax = b$" among all vectors $x \in \mathbb{R}^n$.

More formally: A **least-squares solution** of the system "$Ax = b$" is any $\hat{x} \in \mathbb{R}^n$ such that $\|A\hat{x} - b\| \leqslant \|Ax - b\|$ for all $x \in \mathbb{R}^n$.

Part 4: Least Squares Problems

Questions

1. Is there always a choice of $\hat{x}$ which minimises $\|A\hat{x} - b\|$?
2. If so, how can we find it?
3. Why is this called "least-squares"?

Part 4: Least Squares Problems

Let's explain this problem in terms of the terminology we've developed recently:

1. Solving $Ax = b$ means we are trying to find the coefficients in x that allow us to express b as a linear combination of the columns of A.
2. If $b \notin \text{Col}\, A$, then there is no solution.
3. But $\text{Col}\, A$ is a subspace of $\mathbb{R}^n$, so we can look at the orthogonal projection of b onto $\text{Col}\, A$. That is...

We wish to find $\hat{x} \in \mathbb{R}^n$ such that $A\hat{x}$ is the closest point to b within all of $\mathrm{Col}\, A$. By the Best Approximation Theorem, this means that

$$A\hat{x} = \mathrm{proj}_{\mathrm{Col}\, A}(b).$$

It follows that such an $\hat{x}$ exists. But how can we find it?

1. - Suppose $\hat{b} = \mathrm{proj}_{\mathrm{Col}\,A}(b)$, and then that $\hat{x}$ solves $A\hat{x} = \hat{b}$.
   - Since $\hat{b}$ is the orthogonal projection of b onto $\mathrm{Col}\,A$, we know that $(b - \hat{b}) \perp \mathrm{Col}\,A$
   - That gives $A^\top(b - \hat{b}) = 0$.
   - So now, $A^\top b - A^\top \hat{b} = 0$
   - But $\hat{b} = A\hat{x}$, so $A^\top b - A^\top(A\hat{x}) = 0$.
   - Thus $\hat{x}$ is the solution to $(A^\top A)\hat{x} = A^\top b$.

2. - If $\hat{x}$ solves $A^\top A\hat{x} = A^\top b$, then $A^\top(A\hat{x} - b) = 0$.
   - So $(A\hat{x} - b) \perp \mathrm{Col}\,A$.
   - So b can be decomposed as $b = (A\hat{x}) + \big(b - A\hat{x}\big)$.

   - We see that $A\hat{x}$ is the best approximation for b in $\mathrm{Col}\,A$.

Algorithm: NORMAL EQUATIONS

Given $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$, to solve the Least-Squares Problem,

- Form the $n \times n$ matrix $A^T A$, and the vector $A^\top b \in \mathbb{R}^m$.
- Solve the **normal equation**

$$(A^\top A)\hat{x} = A^\top b. \tag{1}$$

(e.g., using Gaussian elimination.)

Theorem

The always exists a least-squares solution of "$Ax = b$".
The least-squares solutions of "$Ax = b$" are precisely the exact solutions of the **normal equation** (1)

Example

Find a least-squares solution of the system $Ax = b$ when

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \quad \text{and} \quad b = \begin{bmatrix} 2 \\ 2 \\ 4 \end{bmatrix}.$$

Also compute the length of the residual (=approximation error).

In case we don't get to it in class, you should find that

$$A^T A = \begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix}, \quad A^T b = \begin{bmatrix} 8 \\ 18 \end{bmatrix}.$$

Solving $A^\top A\hat{x} = A^T b$ should give $\hat{x} = \begin{bmatrix} 2/3 \\ 1 \end{bmatrix}$. Then the length of the residual is

$$\|A\hat{x} - b\| = \left\| \begin{bmatrix} -1/3 \\ 2/3 \\ -1/3 \end{bmatrix} \right\| \quad A^T b = \begin{bmatrix} 8 \\ 18 \end{bmatrix} = \sqrt{2/3} = 0.8165.$$

For simplicity, the example above, and exercises below, focus on solving Least Squares Problems with three equations and two unknowns. However, the approach is much more general, and exactly the same approach works with more equations and unknowns. The only significant difference is in how one would understand what the unknowns represent – for example one could seek a function of the form $u_1 + u_2 x + u_3 x^2$ that fits the data.

Also:

- How do we know that $A^\top A$ is invertible?
- Normal matrices are so important, there are special algorithms for solving the associated systems.
- The history is fascinating too!

We can also derive the normal equations using calculus. This is because we are solving a minimization problem. So solving where the derivative of the objective function, $\|A\hat{x} - b\|$ is zero gives the equation. See the end of Chapter 10 of Linear Algebra for Data Science for more details:
https://shainarace.github.io/LinearAlgebra/leastsquares.html

At this point we switched to slides with R code in a Jupyter Notebook.

Exercises

Least-Squares Problems

These exercises are taken from Section 6.5 of the textbook.

1. *6.5.1* Find a least-squares solution of $Ax = b$, where

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

2. *6.5.2* Find a least-squares solution of $Ax = b$, where

$$A = \begin{bmatrix} 2 & 1 \\ -2 & 0 \\ 2 & 3 \end{bmatrix}, \quad b = \begin{bmatrix} -5 \\ 8 \\ 1 \end{bmatrix}.$$