Week 8

Data Science Laboratory 1

Dr John Evans
j.evans8@herts.ac.uk

University of Hertfordshire UH

Plan for today

Least-Squares Regression Generalised

Projection matrices

Goal: Explain how linear algebra is used in linear regression

Key Tool: A projection matrix

Projection matrices

Goal: Explain how linear algebra is used in linear regression

Key Tool: A projection matrix

- This can be written as

$$P = A(A^T A)^{-1} A^T,$$

  where A is some $m \times n$ matrix.

Question: What does this matrix do?

Answer: It produces a projection of a vector **b** to the nearest vector in the column space of A.

Does our definition make sense?

- If $\mathbf{b}$ is already in the column space, $P\mathbf{b}$ should equal $\mathbf{b}$, whereas if $\mathbf{b}$ is orthogonal to the column space (i.e. is at a right angle), then $P\mathbf{b}$ should be $\mathbf{0}$. We can check this is the case.

Does our definition make sense?

- If **b** is already in the column space, $P\mathbf{b}$ should equal **b**, whereas if **b** is orthogonal to the column space (i.e. is at a right angle), then $P\mathbf{b}$ should be **0**. We can check this is the case.
- Suppose **b** is in the the column space of A, i.e. $\mathbf{b} = A\mathbf{x}$, for some vector **x**. Then,

$$\begin{aligned} P\mathbf{b} &= A(A^TA)^{-1}A^TA\mathbf{x} \\ &= A\cancel{(A^TA)^{-1}}\cancel{(A^TA)}\mathbf{x} \\ &= AI_n\mathbf{x} \\ &= A\mathbf{x} \quad = \mathbf{b}. \end{aligned}$$

What if b is orthogonal to the column space?

What if b is orthogonal to the column space?

null space

- In this case, $\mathbf{b}^T(A\mathbf{x}) = 0$ for any $\mathbf{x}$.

What if b is orthogonal to the column space?

- In this case, $\mathbf{b}^T(A\mathbf{x}) = 0$ for any $\mathbf{x}$.
- Now, the right hand side is just a number (so it is equal to its transpose). This means the left hand side is a number (so it is equal to its transpose).
- Therefore, $\mathbf{x}^T A^T \mathbf{b} = 0$.

What if b is orthogonal to the column space?

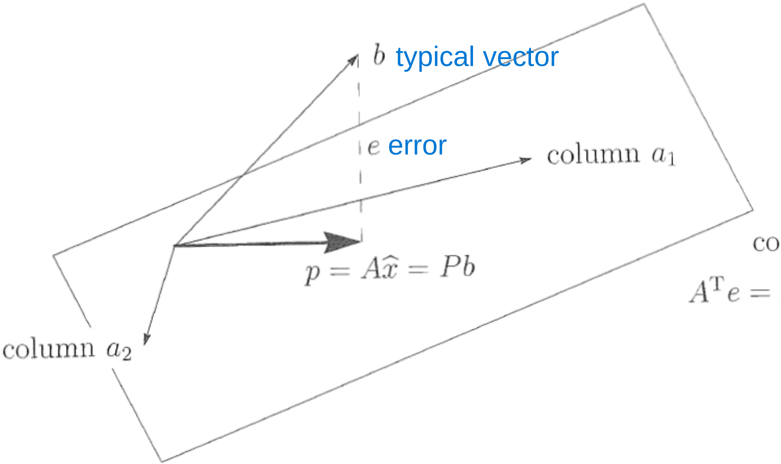
- In this case, $\mathbf{b}^T(A\mathbf{x}) = 0$ for any $\mathbf{x}$.
- ~~Now, the right hand side is just a number (so it is equal to its transpose). This means the left hand side is a number (so it is equal to its transpose)~~.
- Therefore, $\mathbf{x}^T A^T \mathbf{b} = 0$.
- This is true for any $\mathbf{x}$ and so $A^T\mathbf{b} = \mathbf{0}$.
- In fact, from general linear algebra, a big result says that the nullspace of A^T is the orthogonal complement of the column space of A i.e. $A^T\mathbf{b} = \mathbf{0}$.
- In any case, we therefore have,

$$P\mathbf{b} = A(A^TA)^{-1}A^Tb = A(A^TA)^{-1}\mathbf{0} = \mathbf{0}.$$

General picture

- We therefore know what happens to the extreme cases.
- Of course, a typical vector has some components in the column space of A, and some components orthogonal to the column space.
- What the projection does is kill the orthogonal bit and preserves the part in the column space.

Geometric Picture: Suppose A is a 3×2 matrix., i.e. two columns.



$$a_1^{\mathrm{T}} e = 0$$

$$a_2^{\mathrm{T}} e = 0$$

combine into

$$A^{\mathrm{T}} e = A^{\mathrm{T}}(b - A\widehat{x}) = 0$$

The image explained

- We have a typical vector $\mathbf{b}$ and we are projecting onto the column space.
- Of course, we also have another piece $\mathbf{e}$ which is the error (the bit which is not in the column space, that is being mapped to the nullspace of A^T).
- In particular, if $p = P\mathbf{b} = A\hat{\mathbf{x}}$ is the projected bit, then $p + e = \mathbf{b}$. In other words, the projected bit and the error together make up all of $\mathbf{b}$.

e is the projection sideways

A second projection matrix

- We have seen that **e** is in the nullspace of A^T.
- So, we can think of **b** as being projected onto **e** by some matrix, in the same way that **b** is being projected onto **p** by P.

  Question: What is this second projection matrix?

A second projection matrix

- We have seen that **e** is in the nullspace of A^T.
- So, we can think of **b** as being projected onto **e** by some matrix, in the same way that **b** is being projected onto **p** by P.

  Question: What is this second projection matrix?

  Answer: $I_m - P$, we just want to get the rest of the vector.

Im - P is the projection downwards

A second projection matrix

- We have seen that **e** is in the nullspace of A^T.
- So, we can think of **b** as being projected onto **e** by some matrix, in the same way that **b** is being projected onto **p** by P.

  Question: What is this second projection matrix?

  Answer: $I_m - P$, we just want to get the rest of the vector.
- So, P and $I_m - P$ are both projections.
- Some other really nice facts are that if P is symmetric, then $I_n - P$ is symmetric, and if $P^2 = P$, then $(I_m - P)^2 = (I_m - P)$. If we have the above picture in our mind, this should be very natural.

Summary

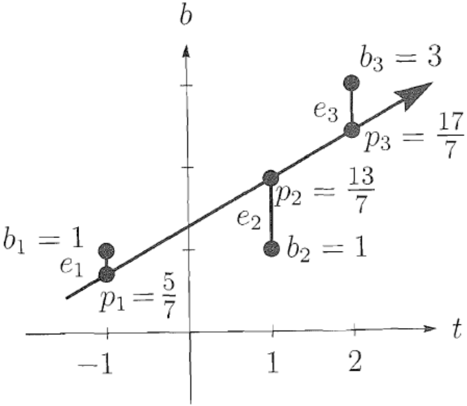
- We have a vector $\mathbf{b} \in \mathbb{R}^m$, say.
- We have an $m \times n$ matrix A and form the matrix $P = A(A^T A)^{-1} A^T$.
- What this does is project $\mathbf{b}$ down into $Col(A)$ with the rest (the error) being projected onto $Null(A^T)$ by $I_m - P$.
- We then have $\mathbf{p} = P\mathbf{b}$ and $\mathbf{e} = (I_m - P)\mathbf{b}$.
- Now it is time to perform linear regression.

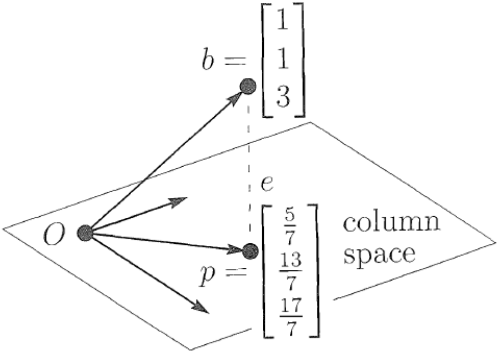
Example

We have three points $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$ given by,

$$\mathbf{b}_1 = \begin{pmatrix} -1 \\ 1 \end{pmatrix} \;,\; \mathbf{b}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \;,\; \mathbf{b}_3 = \begin{pmatrix} 2 \\ 3 \end{pmatrix} .$$

Problem: Find the straight line that passes through these which minimises the error between each point.


b
$b_3 = 3$
e_3
$p_3 = \frac{17}{7}$
$p_2 = \frac{13}{7}$
e_2
$b_2 = 1$
$b_1 = 1$
e_1
$p_1 = \frac{5}{7}$
t
-1
1
2


$b = \begin{bmatrix} 1 \\ 1 \\ 3 \end{bmatrix}$
e
O
$p = \begin{bmatrix} \frac{5}{7} \\ \frac{13}{7} \\ \frac{17}{7} \end{bmatrix}$
column space

Example continued

- We want to find a straight line $y = c + dt$ which is the best straight line fit of these three points. Obviously, no straight line will work exactly because they do not all lie on a straight line, but we can find the best straight line.

Example continued

- We want to find a straight line $y = c + dt$ which is the best straight line fit of these three points. Obviously, no straight line will work exactly because they do not all lie on a straight line, but we can find the best straight line.
- To begin, suppose we *do* have a straight line that passes through $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$. We can then solve $1 = c - d$. Likewise, we could solve $1 = c + d$ and $3 = c + 2d$. This should be recognisable. It is simply a system of three equations in two variables.

$$\underbrace{\begin{pmatrix} 1 & -1 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} c \\ d \end{pmatrix}}_{x} = \underbrace{\begin{pmatrix} 1 \\ 1 \\ 3 \end{pmatrix}}_{b}.$$

Example continued

- We want to find a straight line $y = c + dt$ which is the best straight line fit of these three points. Obviously, no straight line will work exactly because they do not all lie on a straight line, but we can find the best straight line.
- To begin, suppose we *do* have a straight line that passes through $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$. We can then solve $1 = c - d$. Likewise, we could solve $1 = c + d$ and $3 = c + 2d$. This should be recognisable. It is simply a system of three equations in two variables.

$$\begin{pmatrix} 1 & -1 \\ 1 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} c \\ d \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 3 \end{pmatrix}.$$

- Call this 3 × 2 matrix A and the right hand side $\mathbf{b}$. This has no solution (check this) but it has a best solution. By best solution, we mean we can minimise the error between this straight line and the initial points.

How do we measure the error?

Answer: The error is simply $A\mathbf{x} - \mathbf{b} = \mathbf{e}$, where $\mathbf{x} \in \mathbb{R}^2$ is our solution which we want to minimise.

How do we measure the error?

Answer: The error is simply $A\mathbf{x} - \mathbf{b} = \mathbf{e}$, where $\mathbf{x} \in \mathbb{R}^2$ is our solution which we want to minimise. To measure the error, we can use Euclidean distance, i.e. we sum the squares and take the square root. Suppose,

$$A\mathbf{x} - \mathbf{b} = \begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix}.$$

Thus, we want to minimise $e_1^2 + e - 2^2 + e_3^2$. Really, we should take the square root, but this doesn't change much and so it is actually convenient to minimise the square of the distance.

Some observations

- The error will be $\mathbf{0}$ precisely when $A\mathbf{x} = \mathbf{b}$, i.e. when a solution exists.
- Since we are squaring, this is very sensitive to outliers. As such, we do have the usual worry when dealing with such approaches. While this method is commonly used, there are alternatives.
- There are points on this line we are trying to find. We can call them p_1, p_2, p_3 and they can be found from $A\hat{\mathbf{x}} = \mathbf{p}$, where the hat signifies that this is the best fit. In the language of linear algebra, $\mathbf{p}$ is in the columns space of A.

Back to our example

- From the last observation, we want to find $\hat{\mathbf{x}}$ and hence $\mathbf{p}$.
- Since $A\hat{\mathbf{x}} = \mathbf{p}$, we know $A^TA\hat{\mathbf{x}} = A^T\mathbf{p}$ and $\mathbf{p} = \mathbf{b} - \mathbf{e}$.
- However, as we have seen, $\mathbf{e}$ is in the null space of A^T. Thus, we want to solve $A^TA\hat{\mathbf{x}} = A^T\mathbf{b}$. This we can do. First, we compute,

$$A^TA = \begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix}.$$

- This is invertible, and so $\hat{\mathbf{x}} = (A^TA)^{-1}A^T\mathbf{b}$. It therefore follows that $\mathbf{p} = A\hat{\mathbf{x}} = A(A^TA)^{-1}A\mathbf{b}$ and we are back to our projection matrix. Plugging in all the numbers, we get

Ax - b = e

$$A\hat{\mathbf{x}} = \frac{1}{14}\overset{A}{\begin{pmatrix} 1 & -1 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}}\overset{((A^T)*A)^{(-1)}}{\begin{pmatrix} 6 & -2 \\ -2 & 3 \end{pmatrix}}\overset{A^T}{\begin{pmatrix} 1 & 1 & 1 \\ -1 & 1 & 2 \end{pmatrix}}\overset{b}{\begin{pmatrix} 1 \\ 1 \\ 3 \end{pmatrix}} = \frac{1}{7}\overset{B\ (beta)}{\begin{pmatrix} 5 \\ 13 \\ 17 \end{pmatrix}}.$$

Normal equations

- Note that if we went back to $A^T A\hat{\mathbf{x}} = A^T\mathbf{b}$ then we could solve this without finding inverses and so on.
- If we multiply everything out then we are actually solving,

$$\begin{array}{ccc} 3\hat{c} & +2\hat{d} & = & 5 \\ 2\hat{c} & +6\hat{d} & = & 6 \end{array}.$$

- These are called the *normal equations*.
- Further, if we went back to the problem of minimising $e_1^2 + e_2^2 + e_3^2$, where $e_1 = (c - d - 1)$ and so on, then we could approach this from a calculus point of view.
- In other words, expand out the squares, then differentiate with respect to c, d and set to 0. This will once again give the above normal equations. We leave this as an exercise.

Summary

- We have seen how projection matrices can be used to perform linear regression.
- $P = A(A^TA)^{-1}A^T$ is a projection matrix which projects a vector **b** down to the column space.
- The rest (the error) is projected to the nullspace by $I_m - P$.
- We have $\mathbf{p} = P\mathbf{b}$ and $\mathbf{e} = (I_m - P)\mathbf{b}$.