An Application: Linear Regression

This is a particular use of the concept of projections, to statistics. The setup is somewhat cumbersome. We'll see what we can make of it.

The idea is to find a best fit for a set of data. We'll start with finding a linear best fit, a line that is as close to as many of the data points as possible. So, if we have the points

$$(x, y) = \quad (1, 2), \quad (2, 4), \quad (3, 3)$$

our objective is to find a line $y = a_0 + a_1 x$ that is as close as possible to those points. If we were to try to solve for a SINGLE line that matched those points, the system would have

$$a_0 + a_1(1) = 2 \qquad a_0 + a_1(2) = 4 \qquad a_0 + a_2(3) = 3$$

which forms the system

$$[D|\mathbf{y}] = \left[\begin{array}{cc|c} 1 & 1 & 2 \\ 1 & 2 & 4 \\ 1 & 3 & 3 \end{array}\right] \longrightarrow \text{Linear Reduction} \longrightarrow \left[\begin{array}{cc|c} 1 & 1 & 2 \\ 0 & 1 & 2 \\ 0 & 0 & -1 \end{array}\right]$$

so no single solution. What we want then, is to get the $\mathbf{a} = \begin{bmatrix} a_0 \\ a_1 \end{bmatrix}$ such that we minimize the error from the fit:

$$D\mathbf{a} - \mathbf{y} \quad = \quad \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} - \begin{bmatrix} 2 \\ 4 \\ 3 \end{bmatrix}.$$

So, how do we do that? Look at the error's expression. It is the difference between two terms. The first is a multiplication of a matrix with a column vector, so an element of the column space of the matrix. The second is just a single vector. So, what we're looking for is *minimizing the distance between a subspace of* $\mathbb{R}^3$ *and a vector*. The point in the column space of D closest to $\mathbf{y}$ will be $\text{Proj}_{\text{Col}(D)}(\mathbf{y})$. So, $D\mathbf{a}$ will be equal to the projection of $\mathbf{y}$ onto the column space of D. Unfortunately, the column vectors of A are not orthogonal. So, Gram-Schmidt:

$$C = \left\{ \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right\} \longrightarrow C = \left\{ \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} - \frac{6}{3} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right\} = \left\{ \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} \right\}.$$

$$\text{Proj}_{\text{Col}(D)}(\mathbf{y}) = \frac{9}{3} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + \frac{1}{2} \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{5}{2} \\ 3 \\ \frac{7}{2} \end{bmatrix}.$$

Those coefficients, unfortunately, are not those for the actual system. Why don't we just solve the new system?

$$\left[\begin{array}{cc|c} 1 & 1 & \frac{5}{2} \\ 1 & 2 & 3 \\ 1 & 3 & \frac{7}{2} \end{array}\right] \qquad \begin{array}{c} \\ -R_1 \\ -R1 \end{array} \left[\begin{array}{cc|c} 1 & 1 & \frac{5}{2} \\ 0 & 1 & \frac{1}{2} \\ 0 & 2 & 1 \end{array}\right] \qquad \begin{array}{c} -R_2 \\ \\ -2R_2 \end{array} \left[\begin{array}{cc|c} 1 & 0 & 2 \\ 0 & 1 & \frac{1}{2} \\ 0 & 0 & 0 \end{array}\right].$$

This makes our best fit the line $2 + \frac{1}{2}x$.

We could have also used the fact that the orthogonal set had vectors $\mathbf{y}_1 = \mathbf{v}_1$ and $\mathbf{y}_2 = \mathbf{v}_2 - 2\mathbf{v}_1$ (with the $\mathbf{y}$ the vectors from D and the $\mathbf{v}$ the new ones). We can use that with the projection coefficients $b_1 = 3$ and $b_2 = \frac{1}{2}$ for

$$\begin{aligned}\mathbf{y} = 3\mathbf{y}_1 + \frac{1}{2}\mathbf{y}_2 &= 3\mathbf{v}_1 + \frac{1}{2}(\mathbf{v}_2 - 2\mathbf{v}_1) \\ &= 3\mathbf{v}_1 + \frac{1}{2}\mathbf{v}_2 - \mathbf{v}_1 \\ &= 2\mathbf{v}_1 + \frac{1}{2}\mathbf{v}_2.\end{aligned}$$

Either way, final results are:

$$\text{approximation:} \begin{bmatrix} \frac{5}{2} \\ 3 \\ \frac{7}{2} \end{bmatrix}, \qquad \text{true values:} \begin{bmatrix} 2 \\ 4 \\ 3 \end{bmatrix} \qquad \text{error:} \begin{bmatrix} \frac{1}{2} \\ -1 \\ \frac{1}{2} \end{bmatrix}.$$

Those errors are orthogonal to the columns of the data matrix. Check, if you like.

Matrix Inverses

Definition: The inverse the matrix A, written A^{-1}, has the property

$$A\,A^{-1} = A^{-1}\,A = I$$

where I is the identity matrix.

This only applies to square matrices, and NOT ALL have an inverse. A matrix that has an inverse can be called 'invertible,' and those that do NOT have an inverse are called 'singular.' You'll notice that the term for a non-invertible matrix is rather more impressive sounding, well, that's because a random square matrix will almost always be invertible. Remember, as usual, that we don't use random matrices. Keep alert.

There are slightly weaker versions of inverses we will see very little of:

Definition: The *Left* Inverse of A is a matrix B such that $BA = I$.

Definition: The *Right* Inverse of A is a matrix B such that $AB = I$.

We won't use they often, since for a square matrix they are the same thing:

Property: if A and B are $n \times n$ matrices and $AB = I$ then $B = A^{-1}$.

Proof:
This is harder than it looks. Notice that A is not stated to be invertible. First:

$$A(BA - I) = ABA - A = A - A = 0.$$

Note that this, alone, does not prove that $BA - I = 0$, it does, however, prove that every column of $BA - I$ is orthogonal to every row of A:

$$\mathrm{Col}(BA - I) \perp \mathrm{Row}(A).$$

Recall that $AB = I$, so $\mathrm{Col}(AB) = \mathrm{Col}(I) = \mathbb{R}^n$. $\mathrm{Col}(AB)$ is a subset of $\mathrm{Col}(A)$:

$$\mathrm{Col}(AB) = \{(AB)\mathbf{x}|\mathbf{x} \in \mathbb{R}^n\} = \{A(B\mathbf{x})|\mathbf{x} \in \mathbb{R}^n\} \subseteq \{A\mathbf{x}|\mathbf{x} \in \mathbb{R}^n\} = \mathrm{Col}(A).$$

So $\mathrm{Col}(A) = \mathbb{R}^n$ meaning $\mathrm{Rank}(A) = n$. This means that the rows of A span $\mathbb{R}^n$, meaning that

$$(BA - I) \perp \mathbb{R}^n \quad \Longrightarrow \quad (BA - I) = 0,$$

meaning $BA = I$.

So, if we find a right inverse for a square matrix, it is THE inverse and the matrix is invertible. Also, this takes advantage (slightly) of another property of invertible matrices: you can almost treat them like non-zero numbers. If you have $AB = AC$, with A invertible, then $A^{-1}AB = A^{-1}AC \quad \Longrightarrow \quad B = C$.

Properties:

- A^{-1} exists implies $(A^T)^{-1} = (A^{-1})^T$.
- A, B, both invertible, implies $(AB)^{-1} = B^{-1}A^{-1}$.
- A^{-1} exists means $(A^{-1})^{-1} = A$.
- A invertible then cA is too, for $c \in \mathbb{R} \neq 0$. $(cA)^{-1} = \frac{A^{-1}}{c}$.
- I is invertible. $I^{-1} = I$.
- A invertible then A^k is invertible, $(A^k)^{-1} = (A^{-1})^k = A^{-k}$ (use $A^0 = I$).
- If A is invertible, its inverse is unique.

Now take a look at what this means for a linear system:

$$A\mathbf{x} = \mathbf{b}, \quad A \text{ invertible} \quad \Longrightarrow \quad A^{-1}Ax = A^{-1}b \quad \Longrightarrow \quad x = A^{-1}b.$$

Having an invertible A on the left hand side means you can get the solution that way (usually not worth the effort, just solve it), so

- You ALWAYS get a solution (always consistent).
- The solution is unique.

This is primarily of value when one has to calculate many solutions of the form $A\mathbf{x} = \mathbf{b}$. Each can then be solved with a mere multiplication, nothing else.

Calculating Inverses: Not as hard as you may expect. First, it's really easy for any matrix smaller than 3×3.

For a 1×1, $[a]^{-1} = \left[\frac{1}{a}\right]$. Sorry, couldn't resist.

For a 2×2: $\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$.

Example:

$$\begin{bmatrix} 2 & -1 \\ 4 & -3 \end{bmatrix}^{-1} = \frac{1}{-6+4}\begin{bmatrix} -3 & 1 \\ -4 & 2 \end{bmatrix} = \begin{bmatrix} \frac{3}{2} & -\frac{1}{2} \\ 2 & -1 \end{bmatrix}.$$

Once you're dealing with a bigger matrix than that, you have to row reduce. For A, an $n \times n$ matrix, simply solve the linear system $\left[A \middle| I\right]$, row reduce the left down to I and the right will be A^{-1}.

Example:

$$\left[\begin{array}{rrr|rrr} 0 & 2 & 1 & 1 & 0 & 0 \\ 1 & 0 & -1 & 0 & 1 & 0 \\ -1 & 1 & 1 & 0 & 0 & 1 \end{array}\right]$$

First, step, put that leading one on top then clear below it:

$$\begin{array}{c} R_2 \\ R_1 \\ \end{array} \left[\begin{array}{rrr|rrr} 1 & 0 & -1 & 0 & 1 & 0 \\ 0 & 2 & 1 & 1 & 0 & 0 \\ -1 & 1 & 1 & 0 & 0 & 1 \end{array}\right] \qquad \begin{array}{c} \\ \\ R_3 + R_1 \end{array} \left[\begin{array}{rrr|rrr} 1 & 0 & -1 & 0 & 1 & 0 \\ 0 & 2 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 \end{array}\right]$$

$$\begin{array}{c} \\ R_3 \\ R_2 \end{array} \left[\begin{array}{rrr|rrr} 1 & 0 & -1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 2 & 1 & 1 & 0 & 0 \end{array}\right] \qquad \begin{array}{c} \\ \\ R_3 - 2R_2 \end{array} \left[\begin{array}{rrr|rrr} 1 & 0 & -1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & -2 & -2 \end{array}\right]$$

$$\begin{array}{c} R_1 + R_3 \\ \\ \end{array} \left[\begin{array}{rrr|rrr} 1 & 0 & 0 & 1 & -1 & -2 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & -2 & -2 \end{array}\right] = \left[I \middle| A^{-1}\right].$$

Now to check:

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

$$\begin{bmatrix} 0 & 2 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1 & -2 \\ 0 & 1 & 1 \\ 1 & -2 & -2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Only check one side, though. Your choice.

Notice that, using the earlier property, this means that A ($n \times n$) is invertible if and only if we can reduce it down to I. This means that A (square, $n \times n$) is invertible if

- A is row equivalent to I
- A has rank n
- $\text{Row}(A) = \mathbb{R}^n$
- $\text{Col}(A) = \mathbb{R}^n$
- $\text{Null}(A) = \{\mathbf{0}\}$

and a whole lot more...

Solving/Simplifying Matrix Equations

These come up every so often. The idea is to see if you can follow the rules of matrix multiplication/addition.
Example: Solve for X (as much as possible) in

$$AXB^TA - C + DA = \mathbf{0} \qquad \text{given } A, B \text{ are invertible.}$$

This is much like the usual algebra, except there are fewer rules to take advantage of. First step, get the constants onto the right:

$$AXB^TA = C - DA$$

then multiply each side by A^{-1} on the right and left (remember, there's a difference)

$$XB^T = A^{-1}CA^{-1} - A^{-1}D$$

then multiply by the inverse of B^T on the right for

$$X = A^{-1}CA^{-1}(B^T)^{-1} - A^{-1}D(B^{-1})^T.$$

Determinants

The determinant of square matrix A is simply a real number written $\det(A)$ or $|A|$. For a 1×1 matrix it is simply the only entry (not absolute valued, positive or negative). For a 2×2:

$$\det\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\right) = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc,$$

is that familiar?

The definition for a larger one is rather complicated. It uses a concept called a *Cofactor.*
Definition: The Cofactor of A is defined

$$C_{i,j}(A) = (-1)^{i,j} \det(A_{i,j}),$$

where $A_{i,j}$ here means the matrix A with the ith row and jth columns removed.

That power of (-1) means that the value is constantly multiplied by a value from the following checkerboard pattern:

$$\begin{bmatrix} + & - & + & - & \cdots \\ - & + & - & + & \\ + & - & + & - & \\ - & + & - & + & \\ \vdots & & & & \ddots \end{bmatrix}$$

How to Calculate it:

$$\det(A) = a_{i,1}C_{i,1}(A) + a_{i,2}C_{i,2}(A) + \cdots + a_{i,n}C_{i,n}(A)$$
$$\det(A) = a_{1,i}C_{1,i}(A) + a_{2,i}C_{2,i}(A) + \cdots + a_{n,i}C_{n,i}(A).$$

This actually works, since we know how to calculate anything for $n = 1$ or 2, we can therefore do it for $n = 3$, $n = 4$, etc. For any size.

Example: Calculate $\begin{vmatrix} 1 & -2 & 0 \\ 3 & 1 & 4 \\ 0 & -2 & 2 \end{vmatrix}$.

$$\begin{vmatrix} 1 & -2 & 0 \\ 3 & 1 & 4 \\ 0 & -2 & 2 \end{vmatrix} = (1)\begin{vmatrix} 1 & 4 \\ -2 & 2 \end{vmatrix} - (-2)\begin{vmatrix} 3 & 4 \\ 0 & 2 \end{vmatrix} + 0\begin{vmatrix} 3 & 1 \\ 0 & -2 \end{vmatrix}$$
$$= 10 + 12 + 0 \qquad = 22.$$

If we had taken the middle column we'd have

$$\begin{vmatrix} 1 & -2 & 0 \\ 3 & 1 & 4 \\ 0 & -2 & 2 \end{vmatrix} = -(-2)\begin{vmatrix} 3 & 4 \\ 0 & 2 \end{vmatrix} + (1)\begin{vmatrix} 1 & 0 \\ 0 & 2 \end{vmatrix} - (-2)\begin{vmatrix} 1 & 0 \\ 3 & 4 \end{vmatrix}$$
$$= 2(6) + 2 + 8 \qquad = 22.$$

It makes no difference. It's generally easier to use the row/column with the most zeros in it.

There are several properties of a determinant that make it worth discussing. Here's the main one:

Theorem: $\det(AB) = \det(A)\det(B)$ for all $n \times n$ matrices A and B.

Example Questions:

Section 1.5: 2.bdfh), 5.bdf), 7.bd), 9.b), 19.b)
Section 2.1: 3.bdfhjl), 13.bd)