4. Projections and the line of best fit

4.1. Motivation: the line of best fit. As motivation for what follows later in this section, consider the following situation that occurs routinely in laboratories (presumably!). Suppose that we know that two quantities of interest are related *linearly*—which means that one is a function of the other and that its graph with respect to the other is a straight line—and that we are trying to determine this straight line experimentally. We vary one of the quantities (over a finite set of values) and measure the corresponding values of the other, thereby getting a set of data points. Now if we try to find a straight line running through our set of data points, there is often no such line! This after all should not be so surprising, there being several reasons for the deviation from the ideal behaviour, not the least of which is experimental error. At any rate, our problem now is to find a line that "best fits" the data points.

One way of formulating the problem and the "best fit" criterion is as follows. Let the line of best fit be $y = mx + c$, where m is the slope and c the y-intercept.[2] Let (x_1, y_1), ..., (x_n, y_n) be the set of data points. The point (x_k, y_k) lies on $y = mx + c$ if and only if $y_k = mx_k + c$. The ideal situation (which as noted above is rarely the case) would be when we can solve the following system of linear equations for m and c (in other words, when all the data points do lie on a single line):

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix} \begin{pmatrix} m \\ c \end{pmatrix} \tag{5}$$

Note that there is a solution to the above system if and only if the column vector on the left side belongs to the column space of the $n \times 2$ matrix on the right. Confronted with the problem of being forced to "solve" this system when no solution exists, a natural thing to do would be to replace the column vector on the left side by the vector "closest" to it in the column space of the $n \times 2$ matrix and then solve. Such a closest vector is given by the "orthogonal projection" (to be defined presently) of the column vector on the left side on to the column space. This approach demands that we know how to compute orthogonal projections.

4.2. Definition of the orthogonal projection to a subspace. Suppose that we are given a subspace V of $\mathbb{R}^m$. Define $V^\perp := \{w \in \mathbb{R}^m \mid w^t v = 0 \text{ for all } v \in V\}$. Then $V \cap V^\perp = 0$, for $v^t v = 0$ implies $v = 0$ for v in $\mathbb{R}^m$. Moreover, the dimension of $V^\perp$ is such that $\dim V + \dim V^\perp = m$. Indeed, this follows from the rank-nullity theorem, given the interpretation that $V^\perp$ is the solution space of $A^t w = 0$, where A is a matrix of size $m \times \dim V$ whose column space is V.

Putting together a basis of V with a basis of $V^\perp$ therefore gives a basis of $\mathbb{R}^m$. In other words, each element x of $\mathbb{R}^m$ has a unique expression of the form $x = v + v'$ with v in V and v' in $V^\perp$. The association $x \mapsto v$ is a linear transformation from $\mathbb{R}^m$ to V (or to $\mathbb{R}^m$,

[2] This so-called slope-intercept form of the line would not be appropriate if we expect the line of best fit to be vertical. But since, following convention, we plot the "dependent variable" on the x-axis, and there are many different values of this variable, a vertical line is ruled out, and we are justified in our choice of the form of equation for the required line.

if one prefers). It is characterized by the properties that it is identity on V and vanishes on $V^\perp$. It is called the *orthogonal projection to the subspace* V.

4.3. **A formula for the orthogonal projection.** Now suppose that V is specified for us as being the column space of a $m \times n$ matrix A with linearly independent columns. The matrix A of course determines the orthogonal projection—call it P—on to V. The question now is: how do we write the matrix of P (with respect to the standard basis of $\mathbb{R}^m$) given A?

The answer is:

$$\boxed{P = A(A^tA)^{-1}A^t} \tag{6}$$

For the proof, we first observe that the A^tA is an invertible $n \times n$ matrix, so that the inverse in the formula makes sense. Suppose that $(A^tA)x = 0$ for some x in $\mathbb{R}^n$. Then, multiplying by x^t on the left, we get $(x^tA^t)(Ax) = 0$. But this means $||Ax|| = 0$, so $Ax = 0$. Since the columns of A are linearly independent, this in turn means $x = 0$. This proves that the endomorphism of $\mathbb{R}^n$ represented by A^tA is injective and so also bijective. Thus A^tA is invertible.

For an element v of V we have $v = Ax$ for some x in $\mathbb{R}^n$, so that $Pv = A(A^tA)^{-1}A^t(Ax) = A(A^tA)^{-1}(A^tA)x = Ax = v$. And for an element w of $V^\perp$ we have $A^tw = 0$ (because the m entries of A^tw are precisely the inner products of w with the columns of A which span V), and so $Pw = 0$. This proves the formula.

4.4. **Remarks.** We make various remarks about the argument in the preceding subsection.

(1) Note that if $m = n$, then we get $P = A(A^tA)^{-1}A^t = AA^{-1}(A^t)^{-1}A^t =$ identity, which makes sense.
(2) In the course of the proof we have shown the following: the map A^t restricted to the image of A is injective (where A is a real matrix). Indeed, if $A^tAx = 0$, then $x^tA^tAx = 0$ and so $||Ax|| = 0$ and $Ax = 0$.
(3) Observe directly (without recourse to formula (6)) that the matrix P representing (with respect to the standard basis of $\mathbb{R}^m$) the orthogonal projection to any subspace of $\mathbb{R}^m$ is symmetric and satisfies $P^2 = P$.
(4) Suppose that P is an $m \times m$ symmetric matrix such that $P^2 = P$. Then P represents with respect to the standard basis of $\mathbb{R}^m$ the orthogonal projection onto its column space.
(5) If in §4.3 the columns of A are orthonormal, then A^tA is the identity matrix, so formula (6) reduces to $P = AA^t$. This motivates the Gram-Schmidt orthogonalization procedure for computing an orthonormal basis for a subspace of $\mathbb{R}^m$ starting from any given basis for that subspace.

4.5. **Approximate solution to an overdetermined linear system.** Motivated by the need to find the line of best fit and armed with the formula of the previous subsection, we now proceed to give an approximate solution to an overdetermined linear system of equations. Suppose that we want to solve $Ax = b$ for x, where A is an $m \times n$ matrix with linearly independent columns (so, in particular, $m \geq n$). In general b may not be in the column space of A. We replace b by its orthogonal projection on to the column space of A, which by the formula of the previous subsection is $A(A^tA)^{-1}A^tb$. We get $Ax = A(A^tA)^{-1}A^tb$. But since A has linearly indepedent columns, we can cancel the leading A from both sides, so we get

$$\boxed{x = (A^tA)^{-1}A^tb} \tag{7}$$

4.6. **Illustration.** As an illustration of the method just described, let us work out the line that best fits the three points $(1,1)$, $(2,3)$ and $(3,3)$. The slope m and y-intercept c are obtained by an application of (7) as follows.

$$\begin{pmatrix} m \\ c \end{pmatrix} = (A^tA)^{-1}A^tb \quad \text{where } A = \begin{pmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \end{pmatrix} \text{ and } b = \begin{pmatrix} 1 \\ 3 \\ 3 \end{pmatrix}. \text{ We have } A^tA = \begin{pmatrix} 14 & 6 \\ 6 & 3 \end{pmatrix}.$$

Computation of $(A^tA)^{-1}$ (see §2.3):

$$\left(\begin{array}{cc|cc} 14 & 6 & 1 & 0 \\ 6 & 3 & 0 & 1 \end{array}\right) \longrightarrow \left(\begin{array}{cc|cc} 2 & 0 & 1 & -2 \\ 6 & 3 & 0 & 1 \end{array}\right) \longrightarrow \left(\begin{array}{cc|cc} 2 & 0 & 1 & -2 \\ 0 & 3 & -3 & 7 \end{array}\right) \longrightarrow \left(\begin{array}{cc|cc} 1 & 0 & 1/2 & -1 \\ 0 & 1 & -1 & 7/3 \end{array}\right)$$

Thus we have:

$$\begin{pmatrix} m \\ c \end{pmatrix} = \begin{pmatrix} 1/2 & -1 \\ -1 & 7/3 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 3 \\ 3 \end{pmatrix} = \begin{pmatrix} 1/2 & -1 \\ -1 & 7/3 \end{pmatrix} \begin{pmatrix} 16 \\ 7 \end{pmatrix} = \begin{pmatrix} 1 \\ \frac{1}{3} \end{pmatrix}$$

Thus the line that best fits the points $(1,1)$, $(2,3)$, and $(3,3)$ is $y = x + \frac{1}{3}$.