Lecture `linalg4`: Least squares and the normal equations

October 28, 2022

Summary: Statement of least squares problem and solution via the normal equations.

References: T. Sauer's *Numerical Analysis*, 2nd edition, Section 4.1, pages 188–200.

Basic idea behind least squares

Say we want to fit a *line* to the data set $\mathcal{D}_3 = \{(x_1, y_1), (x_2, y_2), (x_3, y_3)\} = \{(-2, 4), (0, 2), (4, 10)\}$. The available modal set is $\mathcal{B}_2 = \{\phi_1(x), \phi_2(x)\} = \{1, x\}$, and the equations we would like to solve are the following:

$$\begin{pmatrix} \phi_1(x_1) & \phi_2(x_1) \\ \phi_1(x_2) & \phi_2(x_2) \\ \phi_1(x_3) & \phi_2(x_3) \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} \implies \begin{pmatrix} 1 & -2 \\ 1 & 0 \\ 1 & 4 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 4 \\ 2 \\ 10 \end{pmatrix} \tag{1}$$

This is an *overdetermined* system: there are more equations than unknowns, and there's no guarantee that a solution exists. Indeed, the middle equation says $c_1 = 2$, and then from the first we have $-2c_2 = 4 - c_1 = 2$, and so $c_2 = -1$. But $c_1 + 4c_2 = -2 \neq 10$, so the third equation is not satisfied. Of course our system has no solution because the data has been drawn from the parabola $y = \frac{1}{2}x^2 + 2$. If we change the data to $\mathcal{D}'_3 = \{(-2, -1), (0, 1), (4, 5)\}$, taken from the line $y = x + 1$, then

$$\begin{pmatrix} 1 & -2 \\ 1 & 0 \\ 1 & 4 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \\ 5 \end{pmatrix} \tag{2}$$

does have a solution, namely $(c_1, c_2) = (1, 1)$. Nevertheless, as our first attempt with the data set $\mathcal{D}_3$ shows, we can't expect to solve the equation

$$\begin{pmatrix} 1 & -2 \\ 1 & 0 \\ 1 & 4 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}. \tag{3}$$

Linear combinations of *two* length–3 vectors can never reach all conceivable length–3 vectors (that is, an arbitrary $\mathbf{b}$).

Since we can't expect to solve the above equation, let's change our goals. We instead define the *residual* vector (the "what's left over" vector),

$$\begin{pmatrix} r_1 \\ r_2 \\ r_3 \end{pmatrix} = \begin{pmatrix} 1 & -2 \\ 1 & 0 \\ 1 & 4 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} - \begin{pmatrix} 4 \\ 2 \\ 10 \end{pmatrix}. \tag{4}$$

Our goal now is NOT to make the residual vector $\mathbf{r}$ zero, for that would be tantamount to finding a solution to the equations (as seen, not possible). Rather our goal now is to make $\mathbf{r}$ as small as

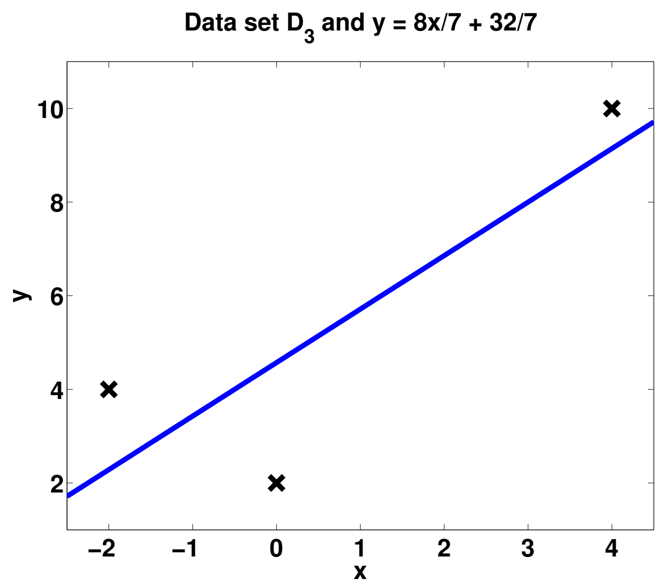

Figure 1: Linear least–squares fitting of the data set $\mathcal{D}_3$.

possible. Now, we know how to measure the size of a vector. We simply use a norm, here the 2-norm $\|\bullet\| = \|\bullet\|_2$, so our goal is to make $\|\mathbf{r}\|$ as small as possible. It proves equivalent and easier to instead work with

$$\|\mathbf{r}\|^2 = r_1^2 + r_2^2 + r_3^3. \tag{5}$$

If we make $\|\mathbf{r}\|^2$ as small as possible, then that ensures $\|\mathbf{r}\| = \sqrt{\|\mathbf{r}\|^2}$ is as small as possible (the square–root function is strictly increasing, and so preserves order relations between non-negative numbers). So we want to *minimize*

$$\|\mathbf{r}\|^2(c_1, c_2) = (c_1 - 2c_2 - 4)^2 + (c_1 - 2)^2 + (c_1 + 4c_2 - 10)^2 \tag{6}$$

over all possible choices of (c_1, c_2). Notice that the residual is a function of the two variables (c_1, c_2), and we write $\|\mathbf{r}\|^2(c_1, c_2)$ here to emphasize this dependence. From Calculus, we know that the minimum of a differentiable function $f(x)$ typically occurs at a stationary point, that is a point x where $f'(x) = 0$. To look for a stationary 2–point of $\|\mathbf{r}\|^2(c_1, c_2)$, we demand that both its partial derivatives vanish, that is

$$\begin{aligned}
\frac{\partial}{\partial c_1}\|\mathbf{r}\|^2(c_1, c_2) &= 0 = 2(c_1 - 2c_2 - 4) + 2(c_1 - 2) + 2(c_1 + 4c_2 - 10) = 6c_1 + 4c_2 - 32 \\
\frac{\partial}{\partial c_2}\|\mathbf{r}\|^2(c_1, c_2) &= 0 = -4(c_1 - 2c_2 - 4) + 8(c_1 + 4c_2 - 10) = 4c_1 + 40c_2 - 64.
\end{aligned} \tag{7}$$

But these equations can be written as a square linear system! Indeed,

$$\begin{pmatrix} 3 & 2 \\ 2 & 20 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 16 \\ 32 \end{pmatrix}, \tag{8}$$

which has solution $(c_1, c_2) = (32/7, 8/7)$. Fig. 1 depicts $y = c_1 + c_2 x = 8x/7 + 32/7$ and the data set $\mathcal{D}_3$. Here is a remarkable observation. Multiplication of (1) by the transpose A^T of the coefficient matrix yields

$$\begin{pmatrix} 1 & 1 & 1 \\ -2 & 0 & 4 \end{pmatrix}\begin{pmatrix} 1 & -2 \\ 1 & 0 \\ 1 & 4 \end{pmatrix}\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ -2 & 0 & 4 \end{pmatrix}\begin{pmatrix} 4 \\ 2 \\ 10 \end{pmatrix} \Longrightarrow \begin{pmatrix} 3 & 2 \\ 2 & 20 \end{pmatrix}\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 16 \\ 32 \end{pmatrix}, \tag{9}$$

that is precisely the square system that we derived above by setting to zero the partial derivatives of $\|\mathbf{r}\|^2(c_1, c_2)$.

Normal equations

It turns out that the above calculations go through for an essentially general overdetermined linear system $A\mathbf{x} = \mathbf{b}$, which we write as follows

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ b_m \end{pmatrix}, \tag{10}$$

here assuming that both $n < m$ and A has full column rank (that is, the columns of A are linearly independent). The system (10) is overdetermined, *and generally we expect that no solution* $\mathbf{x}$ *exists.* Of course, a solution may exist if $\mathbf{b}$ is exceptional, for example $(b_1, b_2, \cdots, b_m)^T = (a_{11}, a_{21}, \cdots, a_{m1})^T$. Here $\mathbf{b}$ is the first column of A, so just pick $x_1 = 1$, and $x_k = 0$ for $k = 2, \cdots, n$. But *in general* there will be no solution. Notice that A is m–by–n, $\mathbf{x}$ is n–by–1, and $\mathbf{b}$ is m–by–1. The transpose A^T is then n–by–m, so that A^TA is n–by–n, and $A^T\mathbf{b}$ is n–by–1. Therefore, $A^TA\mathbf{x} = A^T\mathbf{b}$ is a **square** n–by–n system, and the equations which make up this square system are collectively referred to as the *normal equations*. Provided $\det(A^TA) \neq 0$, the system of normal equations $A^TA\mathbf{x} = A^T\mathbf{b}$ does indeed have a unique solution, but this solution is generally NOT a solution to (10), but it's always as close as we can get to one.

Fact. If the columns of A are linearly independent, that is A has full column rank, then $\det(A^TA) \neq 0$. We'll not prove this fact, but note that this is the case we're most interested in. Like in the Vandermonde example (1) above, we usually have an A with linearly independent columns.

Lemma 1. Suppose A is m–by–n with $m > n$, and that A has full column rank. Then the *unique* solution $\mathbf{x}_{LS}$ (LS for "least squares") to the normal equations $A^TA\mathbf{x}_{LS} = A^T\mathbf{b}$, solves the associated least squares problem, that is minimizes

$$\|\mathbf{r}(\mathbf{x})\| = \|A\mathbf{x} - \mathbf{b}\|.$$

over all possible $\mathbf{x}$. Otherwise put, $\mathbf{x}_{LS}$ is as close as we can get to solving (10). Moreover, if (10) has a solution (in the case of an exceptional $\mathbf{b}$), then $\mathbf{x}_{LS}$ will be the solution to (10).

To prove the lemma, consider the general vector $\mathbf{x} = \mathbf{x}_{LS} + (\mathbf{x} - \mathbf{x}_{LS}) = \mathbf{x}_{LS} + \mathbf{e}$, and compute

$$\begin{aligned}
\|A\mathbf{x} - \mathbf{b}\|^2 &= (A\mathbf{x} - \mathbf{b}) \cdot (A\mathbf{x} - \mathbf{b}) \\
&= (A\mathbf{x} - \mathbf{b})^T (A\mathbf{x} - \mathbf{b}) \\
&= (A\mathbf{x}_{LS} + A\mathbf{e} - \mathbf{b})^T (A\mathbf{x}_{LS} + A\mathbf{e} - \mathbf{b}) \\
&= (A\mathbf{x}_{LS} - \mathbf{b})^T (A\mathbf{x}_{LS} - \mathbf{b}) + (A\mathbf{e})^T (A\mathbf{e}) + (A\mathbf{e})^T (A\mathbf{x}_{LS} - \mathbf{b}) + (A\mathbf{x}_{LS} - \mathbf{b})^T (A\mathbf{e}) \\
&= (A\mathbf{x}_{LS} - \mathbf{b})^T (A\mathbf{x}_{LS} - \mathbf{b}) + (A\mathbf{e})^T (A\mathbf{e}) + 2e^T (A^T A\mathbf{x}_{LS} - A^T\mathbf{b}) \\
&= \|A\mathbf{x}_{LS} - \mathbf{b}\|^2 + \|A\mathbf{e}\|^2. \qquad (11)
\end{aligned}$$

To reach the last line, we have used the fact that $\mathbf{x}_{LS}$ solves the normal equations to kill the last term in the second–to–last line. To reach the second–to–last line from the third–to–last, we have used the fact that $\mathbf{w} \cdot \mathbf{v} = \mathbf{w}^T\mathbf{v} = \mathbf{v}^T\mathbf{w}$. The result of our calculation is therefore

$$\|A\mathbf{x} - \mathbf{b}\|^2 = \|A\mathbf{x}_{LS} - \mathbf{b}\|^2 + \|A\mathbf{e}\|^2. \qquad (12)$$

Now, as we have assumed that A has full column rank, $\|A\mathbf{e}\|^2 > 0$ for $\mathbf{e} \neq \mathbf{0}$ (or else the columns of A are linearly dependent after all). This clearly shows that $\mathbf{e} = \mathbf{0}$ minimizes $\|\mathbf{r}(\mathbf{x})\|^2 = \|\mathbf{r}(\mathbf{x}_{LS} + \mathbf{e})\|^2$, and $\mathbf{e} = \mathbf{0}$ corresponds to $\mathbf{x} = \mathbf{x}_{LS}$. Therefore, $\|\mathbf{r}(\mathbf{x}_{LS})\|^2 < \|\mathbf{r}(\mathbf{x})\|^2$ for $\mathbf{x} \neq \mathbf{x}_{LS}$. We then get $\|\mathbf{r}(\mathbf{x}_{LS})\| < \|\mathbf{r}(\mathbf{x})\|$ for $\mathbf{x} \neq \mathbf{x}_{LS}$, since the square root is an increasing function, and so preserves order relations. □

Lemma 2. Suppose A is m–by–n with $m > n$, A has full column rank, and $\mathbf{x}_{LS}$ is the unique solution to the least squares problem. Then

$$\left(\nabla \|A\mathbf{x} - \mathbf{b}\|^2\right)\Big|_{\mathbf{x}=\mathbf{x}_{LS}} = \mathbf{0}. \qquad (13)$$

Roughly speaking, this statement is analogous to the following scenario from Calculus. If a function $f(x)$ has a (perhaps local) minimum value at x_* and is differentiable at x_*, then $f'(x_*) = 0$. In our case we know from the first lemma that $\|\mathbf{r}(\mathbf{x}_{LS})\|$ is actually the global minimum value of $\|\mathbf{r}(\mathbf{x})\|$.

To prove the lemma, we proceed in index notation,

$$\begin{aligned}
\|A\mathbf{x} - \mathbf{b}\|^2 &= (A\mathbf{x} - \mathbf{b}) \cdot (A\mathbf{x} - \mathbf{b}) \\
&= \sum_{k=1}^{m} (A\mathbf{x} - \mathbf{b})_k (A\mathbf{x} - \mathbf{b})_k \\
&= \sum_{k=1}^{m} \left(\sum_{j=1}^{n} A_{kj} x_j - b_k \right) \left(\sum_{p=1}^{n} A_{kp} x_p - b_k \right) \\
&= \sum_{j=1}^{n} \sum_{p=1}^{n} \sum_{k=1}^{m} x_p A_{kp} A_{kj} x_j - \sum_{j=1}^{n} \sum_{k=1}^{m} A_{kj} x_j b_k - \sum_{p=1}^{n} \sum_{k=1}^{m} A_{kp} x_p b_k + \sum_{k=1}^{m} b_k b_k \\
&= \sum_{j=1}^{n} \sum_{p=1}^{n} x_p (A^T A)_{pj} x_j - 2 \sum_{j=1}^{n} (\mathbf{b}^T A)_j x_j + \sum_{k=1}^{m} b_k b_k. \qquad (14)
\end{aligned}$$

In performing these calculations, we have used $A_{kj} = (A^T)_{jk}$. To take the gradient of the last expression, we will use the result

$$\frac{\partial x_k}{\partial x_j} = \delta_{kj} = \begin{cases} 1 & \text{for } j = k \\ 0 & \text{for } j \neq k. \end{cases} \qquad (15)$$

Proceeding, we then have

$$\begin{aligned}
\frac{\partial}{\partial x^\ell}\|A\mathbf{x}-\mathbf{b}\|^2 &= \sum_{j=1}^{n}\sum_{p=1}^{n}\delta_{\ell p}(A^TA)_{pj}x_j + \sum_{j=1}^{n}\sum_{p=1}^{n}x_p(A^TA)_{pj}\delta_{\ell j} - 2\sum_{j=1}^{n}(\mathbf{b}^TA)_j\delta_{\ell j} \\
&= \sum_{j=1}^{n}(A^TA)_{\ell j}x_j + \sum_{p=1}^{n}x_p(A^TA)_{p\ell} - 2(\mathbf{b}^TA)_\ell \\
&= \sum_{j=1}^{n}2(A^TA)_{\ell j}x_j - 2(\mathbf{b}^TA)_\ell,
\end{aligned} \tag{16}$$

where we have used the fact that sums on Kronecker δ symbols collapse to single terms, for example $\sum_{j=1}^{n}(\mathbf{b}^TA)_j\delta_{\ell j} = (\mathbf{b}^TA)_\ell$. An equivalent expression of the last equation is

$$\nabla\|A\mathbf{x}-\mathbf{b}\|^2 = 2A^TA\mathbf{x} - 2A^T\mathbf{b}, \tag{17}$$

from which the lemma immediately follows. $\square$

Example. Consider the data set $\mathcal{D}_4 = \{(-1,1),(0,1),(1,3),(2,11)\}$ taken from $y = \frac{2}{3}x^3 + x^2 + \frac{1}{3}x + 1$. Let us construct be best quadratic polynomial which fits the data. Using monomials, our modal set is then $\mathcal{B}_3 = \{\phi_1(x),\phi_2(x),\phi_3(x)\} = \{1, x, x^2\}$. The nonsquare Vandermonde matrix is then

$$V = \begin{pmatrix} \phi_1(-1) & \phi_2(-1) & \phi_3(-1) \\ \phi_1(0) & \phi_2(0) & \phi_3(0) \\ \phi_1(1) & \phi_2(1) & \phi_3(1) \\ \phi_1(2) & \phi_2(2) & \phi_3(2) \end{pmatrix} = \begin{pmatrix} 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{pmatrix}, \tag{18}$$

and we want minimize the residual

$$\begin{pmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{pmatrix} = \begin{pmatrix} 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{pmatrix}\begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} - \begin{pmatrix} 1 \\ 1 \\ 3 \\ 11 \end{pmatrix}. \tag{19}$$

Straightforward computations yield

$$V^TV = \begin{pmatrix} 4 & 2 & 6 \\ 2 & 6 & 8 \\ 6 & 8 & 18 \end{pmatrix}, \qquad V^T\mathbf{y} = \begin{pmatrix} 16 \\ 24 \\ 48 \end{pmatrix}. \tag{20}$$

Whence the normal equations $V^TV\mathbf{c} = V^T\mathbf{y}$ are

$$\begin{pmatrix} 4 & 2 & 6 \\ 2 & 6 & 8 \\ 6 & 8 & 18 \end{pmatrix}\begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} 16 \\ 24 \\ 48 \end{pmatrix}, \tag{21}$$

and the solution is $(c_1, c_2, c_3) = (2/5, 6/5, 2)$. Fig. 2 shows the fit.

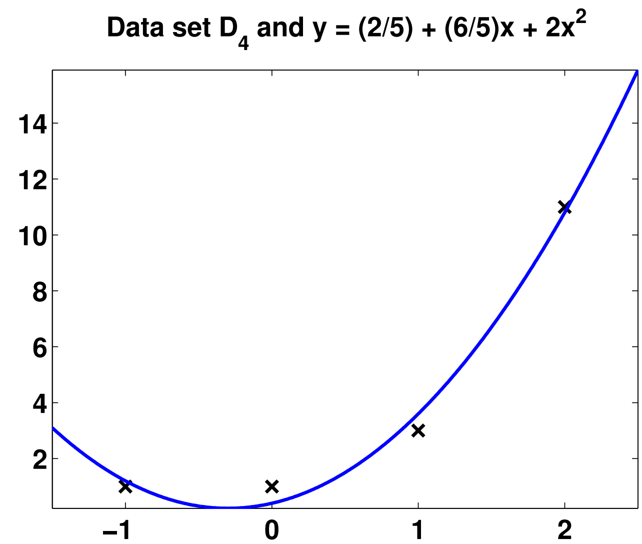

Figure 2: Quadratic least–square fitting of the data set $\mathcal{D}_4$.