

Least Squares

1. Frustration

Suppose we want to solve the system of simultaneous equations

$$x + 2y = 3$$

$$2x + y = 3$$

$$x - 2y = 1$$

$$2x - y = 2$$

The frustration is that there is no solution. The first two equations together, for example, have only the solution $x = 1, y = 1$, and the last two together have only the solution $x = 1, y = 0$. The system has more equations than variables. It is “overdetermined”. The best we can hope for is to make the equations all approximately true, in some sense:

$$x + 2y \approx 3$$

$$2x + y \approx 3$$

$$x - 2y \approx 1$$

$$2x - y \approx 2$$

2. The sum-of-squares combined error

To be more concrete, we should ask for x and y that make the equations as nearly true as possible. In other words, we should try to find x and y that minimize the *errors*

$$\epsilon_1 = x + 2y$$

$$\epsilon_2 = 2x + y$$

$$\epsilon_3 = x - 2y$$

$$\epsilon_4 = 2x - y$$

There is still a problem, though: It is not clear what it means to minimize several errors simultaneously. We must choose some single combined measure of the errors and then minimize that.

To combine the errors we should *not* use $\epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_4$, because negative errors could cancel out positive ones. It is more reasonable, but still not good, to use $|\epsilon_1| + |\epsilon_2| + |\epsilon_3| + |\epsilon_4|$. The use of absolute values is undesirable both because the absolute value function is not differentiable and because there might be many (x, y) pairs giving the same minimum value (as it turns out).

Instead, a nice measure of the combined error is the *sum of squares* of the individual errors. Because the combined error depends on the choice of x and y , let's write

$$E(x, y) = \epsilon_1^2 + \epsilon_2^2 + \epsilon_3^2 + \epsilon_4^2.$$

Just trying some different choices for x and y gives

$$E(1, 1) = 5 \text{ (since } 0^2 + 0^2 + 2^2 + 1^2 = 5),$$

$$E(1, 0) = 5,$$

$E(0, 1) = 23$ (worse); but the least-squares method will reveal

$$E(1.4, 0.5) = 0.9 \text{ (fantastic, and the lowest possible error).}$$

It is worth noting that the least squares method will allow several small errors in preference to one large one; for example, $1^2 + 1^2 + 1^2 + 1^2$ is smaller than $4^2 + 0^2 + 0^2 + 0^2$.

3. The matrix formulation

You know how to write a square system of linear equations with matrices; an overdetermined system can be written the same way. In our example:

$$\begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 1 & -2 \\ 2 & -1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \approx \begin{bmatrix} 3 \\ 3 \\ 1 \\ 2 \end{bmatrix}. \text{ In general, an overdetermined system is of the form } \mathbf{Ax} \approx \mathbf{b}.$$

Here A is $m \times n$ with $m > n$, \mathbf{x} is an n -vector, and \mathbf{b} is an m -vector. The errors can also be expressed as a single vector $\epsilon = \langle \epsilon_1, \dots, \epsilon_m \rangle$. If ϵ is written as a column vector, then $\epsilon = \mathbf{Ax} - \mathbf{b}$.

4. Solution using normal equations

The least-squares solution can be found by a simple method that almost seems like magic before you know its derivation:

Method: To solve $\mathbf{Ax} \approx \mathbf{b}$ in the least squares sense, instead solve the *square* system of equations $N\mathbf{x} = \mathbf{d}$, where $N = A^t A$ and $\mathbf{d} = A^t \mathbf{b}$.

This system of equations is usually called the system of *normal equations*. For short, they can be written as $A^t \mathbf{Ax} = A^t \mathbf{b}$, which you can think of as the result of multiplying both sides of $\mathbf{Ax} \approx \mathbf{b}$ on the left by A^t and putting $=$ for \approx . The matrix N will always be symmetric, i.e., $N^t = N$.

Example (continued from above).

$$N = A^t A = \begin{bmatrix} 1 & 2 & 1 & 2 \\ 2 & 1 & -2 & -1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 1 & -2 \\ 2 & -1 \end{bmatrix} = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix},$$

$$\mathbf{d} = A^t \mathbf{b} = \begin{bmatrix} 1 & 2 & 1 & 2 \\ 2 & 1 & -2 & -1 \end{bmatrix} \begin{bmatrix} 3 \\ 3 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 14 \\ 5 \end{bmatrix}, \text{ so the normal equations are}$$

$$\begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 14 \\ 5 \end{bmatrix}.$$

Thus the solution is $x = 1.4, y = 0.5$, as mentioned above. (As you see, this example was specially chosen for easy arithmetic; in general N is symmetric but not diagonal.)

Notice that N needs to be nonsingular. In practice this is very likely to be the case. The reason why, as well as the derivation of the normal equations will be explained below.

5. Applications.

Least-squares problems arise in a number of ways. The simplest one is this:

Application 1. “Linear regression”. Find the best-fitting straight line $y = c_0 + c_1x$ for data points $(1, 3), (2, 3), (-2, -1), (-1, 1), (2, 2)$ (Figure 1).

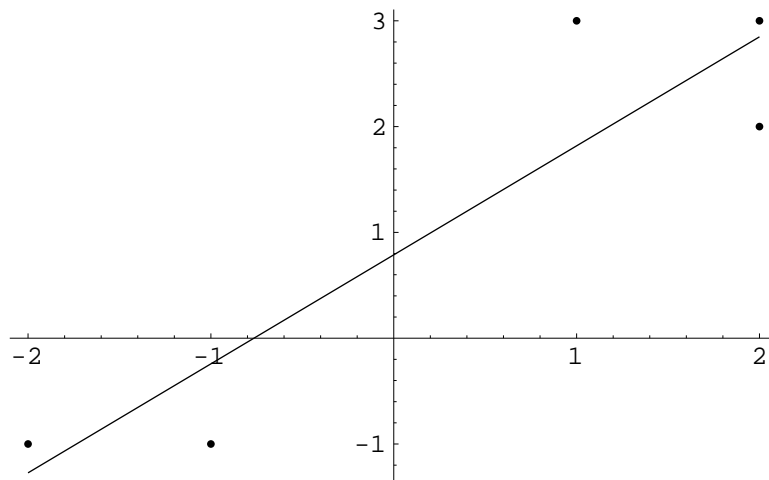


Figure 1: A linear fit

Solution. Here we need to find c_0 and c_1 , not x and y . If a data point $(2, 5)$ (say) were to be *on* the line that would mean $5 = c_0 + c_1 2$. But there are too many data points to hope that they would all be on one line. Thus we have to settle for finding c_0 and c_1 so that

$$\begin{aligned}
c_0 + 1c_1 &\approx 3 \\
c_0 + 2c_1 &\approx 3 \\
c_0 - 2c_1 &\approx -1 \\
c_0 - 1c_1 &\approx 1 \\
c_0 + 2c_1 &\approx 2
\end{aligned}$$

or in other words,
$$\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & -2 \\ 1 & -1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \end{bmatrix} \approx \begin{bmatrix} 3 \\ 3 \\ -1 \\ 1 \\ 2 \end{bmatrix}, \text{ so the normal equations are}$$

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & -2 & -1 & 2 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & -2 \\ 1 & -1 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & -2 & -1 & 2 \end{bmatrix} \begin{bmatrix} 3 \\ 3 \\ -1 \\ 1 \\ 2 \end{bmatrix}, \text{ or}$$

$$\begin{bmatrix} 5 & 2 \\ 2 & 14 \end{bmatrix} = \begin{bmatrix} 8 \\ 14 \end{bmatrix}. \text{ A computer solution gives } c_0 = 1.27273, c_1 = 0.81818, \text{ to five decimal places. Thus the best-fitting straight line in the least-squares sense is } y = 1.27273 + 0.81818x.$$

In general, for data points $(r_1, s_1), \dots, (r_m, s_m)$ you get the normal equations

$$\begin{bmatrix} m & \sum_i r_i \\ \sum_i r_i & \sum_i r_i^2 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \end{bmatrix} = \begin{bmatrix} \sum_i s_i \\ \sum_i r_i s_i \end{bmatrix}.$$

It is all right to have some of the r_i be the same. These normal equations are guaranteed to be nonsingular as long as *not all* the r_i are the same. (If they were all the same, the data points would lie on a vertical line and there would be no one best-fitting linear function.)

Application 2. Finding a best-fitting quadratic. Given data points $(r_1, s_1), \dots, (r_m, s_m)$, find the parabola $y = c_0 + c_1x + c_2x^2$ that fits them best (Figure 2). In other words, solve

$$\begin{aligned}
c_0 + c_1r_1 + c_2r_1^2 &\approx s_1 \\
c_0 + c_1r_2 + c_2r_2^2 &\approx s_2 \\
&\dots \quad \dots \quad \dots \\
c_0 + c_1r_m + c_2r_m^2 &\approx s_m
\end{aligned}$$

At first this might not look like a linear problem. Notice, though, that the r_i and s_i are data numbers, and their powers are also just certain numbers.

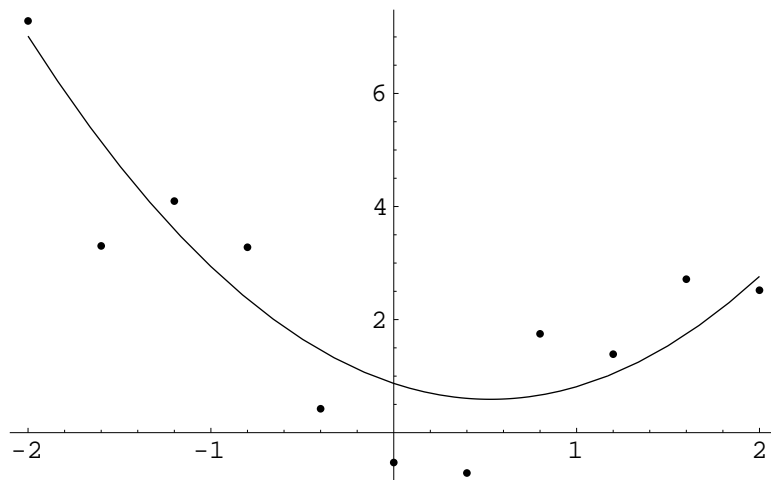


Figure 2: A quadratic fit

The unknowns are c_0, c_1, c_2 , which appear *linearly*. In fact, a polynomial *is* a linear combination—of powers. In matrix form, this problem is

$$\begin{bmatrix} 1 & r_1 & r_1^2 \\ 1 & r_2 & r_2^2 \\ \dots & \dots & \dots \\ 1 & r_m & r_m^2 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} s_1 \\ s_2 \\ \dots \\ s_m \end{bmatrix}, \text{ or for short, } \mathbf{A}\mathbf{c} \approx \mathbf{s}.$$

Solving the normal equations gives us the best c_0, c_1, c_2 .

Note. Although one can talk loosely of a “best-fitting parabola”, if c_2 happened to come out to be zero the curve would actually be a straight line.

Application 3. Finding a best-fitting linear combination of polynomials of given degree d .

This is an obvious generalization of the quadratic case. For polynomials of degree d , the only requirement to guarantee nonsingularity of the normal equations is that r_1, \dots, r_m should include at least $d + 1$ different numbers.

Application 4. Finding a best-fitting linear combination of functions. Given data points and a specified finite family of functions, find the linear combination of those functions that fits the data best.

In other words, suppose we are given functions f_1, \dots, f_n and data points $(r_1, s_1), \dots, (r_m, s_m)$. We want to find coefficients c_1, \dots, c_n so that

$$\begin{array}{rcl} c_1 f_1(r_1) + \dots + c_n f_n(r_1) & \approx & s_1 \\ c_1 f_1(r_2) + \dots + c_n f_n(r_2) & \approx & s_2 \\ & \dots & \dots \\ c_1 f_1(r_m) + \dots + c_n f_n(r_m) & \approx & s_m \end{array}$$

This is just a least-squares problem $A\mathbf{c} \approx \mathbf{b}$, where $A = [f_i(r_j)]$ and $\mathbf{b} = [s_i]$. We'll discuss the particular case where the f_i are Chebyshev polynomials.

Application 5. Finding a best-fitting polynomial parametric curve near given data points.

Let's consider a cubic curve in two dimensions. We have times t_0, \dots, t_n and data points P_1, \dots, P_n , and we want a cubic curve $P(t)$ so that $P(t_i)$ is close to P_i for each i . As a measure of closeness we can use the Euclidean distance. In other words, we want to minimize the sum of squares of the errors $\text{dist}(P(t_i), P_i)$. This situation at first sounds different from the errors previously discussed, which were vertical distances on graphs, but see what happens: Write $P(t) = \begin{bmatrix} x(t) \\ y(t) \end{bmatrix}$, $P_i = \begin{bmatrix} x_i \\ y_i \end{bmatrix}$. Then $\text{dist}(P(t_i), P_i) = \sqrt{(x(t_i) - x_i)^2 + (y(t_i) - y_i)^2}$, and the sum-of-squares error is the sum of squares of square roots: $E = \sum_i (x(t_i) - x_i)^2 + (y(t_i) - y_i)^2$. So if we write $x(t) = c_0 + c_1t + c_2t^2 + c_3t^3$ and $y(t) = d_0 + d_1t + d_2t^2 + d_3t^3$, with coefficients to be determined, the sum-of-squares error is exactly the same as for the least-squares problem

$$\begin{array}{rcl} c_0 + c_1t_1 + c_2t_1^2 + c_3t_1^3 & \approx & x_1 \\ c_0 + c_1t_2 + c_2t_2^2 + c_3t_2^3 & \approx & x_2 \\ & \dots & \dots \\ c_0 + c_1t_n + c_2t_n^2 + c_3t_n^3 & \approx & x_n \\ d_0 + d_1t_1 + d_2t_1^2 + d_3t_1^3 & \approx & y_1 \\ d_0 + d_1t_2 + d_2t_2^2 + d_3t_2^3 & \approx & y_2 \\ & \dots & \dots \\ d_0 + d_1t_n + d_2t_n^2 + d_3t_n^3 & \approx & y_n \end{array}$$

Although squared errors from both x - and y -coordinates are added together, the errors involve separate sets of variables, so this problem is really equivalent to two separate least-squares problems $A\mathbf{c} \approx \mathbf{x}$ and $A\mathbf{d} \approx \mathbf{y}$, where

$$A = \begin{bmatrix} 1 & t_1 & t_1^2 & t_1^3 \\ 1 & t_2 & t_2^2 & t_2^3 \\ \dots & \dots & \dots & \dots \\ 1 & t_n & t_n^2 & t_n^3 \end{bmatrix},$$

$$\mathbf{c} = \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix}, \mathbf{d} = \begin{bmatrix} d_0 \\ d_1 \\ d_2 \\ d_3 \end{bmatrix}, \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix}, \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix}.$$

Since both problems share the same A , we can solve them together with normal equations

$$A^t AC = A^t P, \text{ where } C = \begin{bmatrix} c_0 & d_0 \\ c_1 & d_1 \\ c_2 & d_2 \\ c_3 & d_3 \end{bmatrix} \text{ and}$$

$$P = \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \dots & \dots \\ x_n & y_n \end{bmatrix} = \text{the matrix of data points as row vectors.}$$

Other applications. Large overdetermined arrays arise in many engineering design applications. The method of normal equations works well even if A is 1000×50 . (On modern workstations it takes less than a second to solve a 50×50 set of linear equations.)

6. A geometrical interpretation

Recall that the distance between two vectors is the length of their difference. Therefore $|\epsilon|$ is the distance from $A\mathbf{x}$ to \mathbf{b} . Moreover, $E(\mathbf{x}) = \epsilon_1^2 + \dots + \epsilon_m^2 = |\epsilon|^2$. For a nonnegative function, minimizing the square of the function is the same as minimizing the function. Therefore

These statements are equivalent:
 $A\mathbf{x} \approx \mathbf{b}$ has least-squares solution \mathbf{x} ;
 $A\mathbf{x}$ is as close as possible to \mathbf{b} in \mathbf{R}^m .

Moreover, the vectors of the form $A\mathbf{x}$, for all possible column vectors \mathbf{x} , form a *subspace* W of \mathbf{R}^m . Therefore the problem is really one of finding the point in a subspace that is closest to a given point; the error vector ϵ is the vector from the given point to the point in the subspace. In Figure 3, the parallelogram represents part of the subspace W . The error vector “err” goes from b to $A\mathbf{x}$.

To see why W is a subspace, notice that W is just the image of the linear transformation given by $T(\mathbf{x}) = A\mathbf{x}$. Or, notice that $A\mathbf{x}$ can be expanded as $A\mathbf{x} = A_1x_1 + \dots + A_nx_n$, where A_j means the j -th column of A ; therefore W is the *column space* of A , i.e., the subspace spanned by the columns of A . In Figure 3, the columns of A are the sides of the parallelogram that touch the origin.

You might think that the way to get from \mathbf{b} to the closest point of W is to go along a normal to W —i.e., a line perpendicular to W —and that’s exactly right, as indicated by the right angle in Figure 3. First let’s discuss the computational method that produces the answer, and then then derivation of the method.

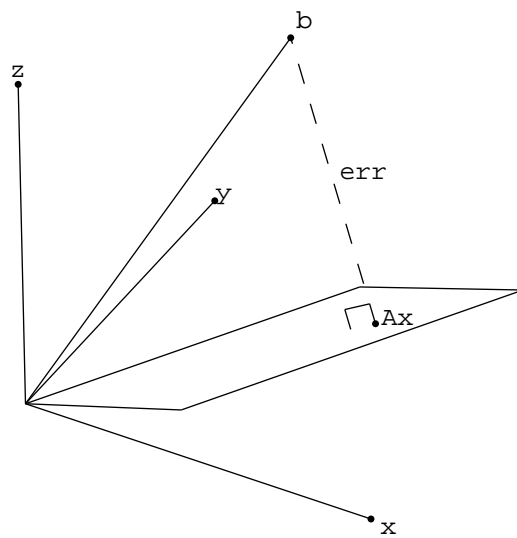


Figure 3: Geometrical interpretation

7. Explanation of the method

One way to invent the method of normal equations is to use calculus. The error is $E(x_1, \dots, x_n)$. Take the partial derivative with respect to each x_i and set it equal to zero. You get n simultaneous equations in x_1, \dots, x_n , which turn out to be the same as the normal equations.

Here is a better way, based on the geometrical interpretation discussed above. Recall that

- (i) the set W of all vectors of the form $A\mathbf{x}$ is a subspace, namely the column space of A (the subspace of \mathbf{R}^n spanned by the columns of A);
- (ii) we are looking for the particular \mathbf{x} for which $A\mathbf{x}$ is closest to \mathbf{b} . To avoid confusion, call this vector \mathbf{x}_0 ;
- (iii) the error vector is $\epsilon = A\mathbf{x}_0 - \mathbf{b}$.

We need only two more facts:

- (iv) The shortest distance from a vector \mathbf{b} to a subspace is along a perpendicular. (See Problems for a verification.)
- (v) In a matrix product AB , the entries are dot products of rows of A with columns of B .

Now, by (iv), ϵ is perpendicular to all vectors in W . By (i), the columns of A are in W , so ϵ is perpendicular to each column of A . Thus the dot product of ϵ with every column of A is zero. Write the dot products by putting A on its side and writing the matrix product $A^t\epsilon = \mathbf{0}$.

This says $A^t(A\mathbf{x}_0 - \mathbf{b}) = \mathbf{0}$, so $A^tA\mathbf{x}_0 = A^t\mathbf{b}$, the normal equations!

8. Some comments.

(a) When you first study linear algebra, you might think that \mathbf{R}^n is useful only for $n = 2$ and $n = 3$. However, if you solve a 1000×50 least-squares problem, you are really relying on geometrical properties involving the 50-dimensional subspace W of the 1000-dimensional space \mathbf{R}^{1000} !

(b) What about the nonsingularity of the normal equations in general? Here are two facts that are closely related to each other:

Fact (I). The rank of A^tA equals the rank of A .

For example, consider a 20×4 least-squares problem $A\mathbf{x} \approx \mathbf{b}$. The normal equations are 4×4 . They are nonsingular if their rank is 4. By Fact I, this happens if the rank of A is 4, i.e., if the four columns of A are linearly independent. But this is extremely likely if there is any random quality to the entries of A . In fact, since row rank = column rank, the rank of A will be 4 if any four *rows* are linearly independent.

Fact (II). If A is $m \times n$ with $m > n$, the determinant of $A^t A$ equals the sum of the squares of the determinants of all $n \times n$ submatrices of A .

Continuing the 20×4 example, if you take all the 4×4 submatrices of A (almost 5000 of them), square their determinants, and add, you get the determinant of $A^t A$. As you can see, it's quite likely to be nonzero.

9. Problems

Problem W-1. Find the best fitting straight line for data points $(-1, -1)$, $(-1, 0)$, $(0, 1)$, $(1, 2)$, $(1, 3)$. (See Figure 4.)

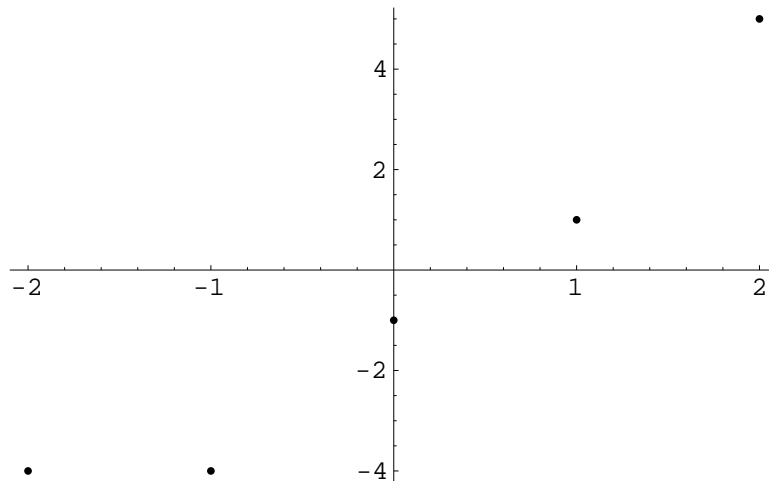


Figure 4: Data points

Problem W-2. Find the best-fitting parabola $y = f(x)$ for data points $(-2, -4)$, $(-1, -4)$, $(0, -1)$, $(1, 1)$, $(2, 5)$. (See Figure 5.)

Problem W-3. For the case of finding the best-fitting quadratic $c_0 + c_1x + c_2x^2$ for data $(r_1, s_1), \dots, (r_m, s_m)$, express the normal equations $N\mathbf{c} = \mathbf{d}$ by giving the entries of N and \mathbf{d} as summations involving the data coordinates.

Problem W-4. In the explanation of the method of normal equations, we used the fact that the subspace W of all vectors of the form $A\mathbf{x}$ contains the columns of A . Which particular \mathbf{x} give the various columns of A ?

Problem W-5. Carry out the derivation of the normal equations using calculus, as suggested in Section 7, but just for the case $n = 2$ and arbitrary m .

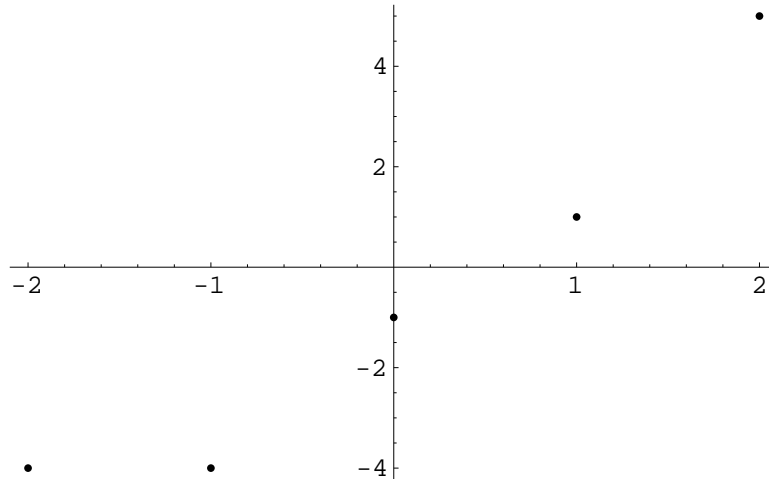


Figure 5: Data points

Problem W-6. Show that if W is a subspace of \mathbf{R}^n and \mathbf{b} is any vector in \mathbf{R}^n , then there is a vector \mathbf{w}_0 of W so that $\mathbf{w}_0 - \mathbf{b}$ is perpendicular to W .

(Method: Let k be the dimension of W and let $\mathbf{w}_1, \dots, \mathbf{w}_k$ be an orthonormal basis of W . Write $\mathbf{w}_0 = r_1\mathbf{w}_1 + \dots + r_k\mathbf{w}_k$ with unknown coefficients r_i . See if you can find a condition on the r_i so that $\mathbf{w}_0 - \mathbf{b}$ is perpendicular to \mathbf{w}_1 . Do the same for $\mathbf{w}_2, \dots, \mathbf{w}_k$. This is really the same idea as for Gram-Schmidt orthogonalization. Of course, in terms of the geometrical derivation of the least-squares method, $\mathbf{w}_0 = A\mathbf{x}_0$.)

Problem W-7. Show that in the preceding problem, of all vectors in W the vector \mathbf{w}_0 is the closest vector to \mathbf{b} , by using both a geometrical explanation and an algebraic proof.

Geometrical method: To the diagram in Section 6, add an arbitrary-looking vector \mathbf{w} in W . Consider the triangle formed by (the ends of) \mathbf{b} , \mathbf{w}_0 , and \mathbf{w} . Three points always lie in a plane, so they form an ordinary planar triangle, even if they are in a higher-dimensional space.

Algebraic method: Consider squares of lengths instead of lengths. Thus you want to show that \mathbf{w}_0 is the \mathbf{w} in W for which $(\mathbf{w} - \mathbf{b}) \cdot (\mathbf{w} - \mathbf{b})$ is least. Write arbitrary \mathbf{w} as $\mathbf{w} = \mathbf{w}_0 + \Delta\mathbf{w}$ and write $\epsilon = \mathbf{w}_0 - \mathbf{b}$. Rewrite the expression you are minimizing using this notation and expand it using simple algebraic properties of the dot product. You should get something algebraic that corresponds to the geometrical method.

Problem W-8. Consider the overdetermined system in one variable

$$\begin{aligned} x &\approx 1 \\ x &\approx 0 \end{aligned}$$

- (a) Find all values of \mathbf{x} that minimize $|\epsilon_1| + |\epsilon_2|$. Is there more than one such value?
- (b) Find the least-squares solution using calculus.
- (c) Find the least-squares solution using the normal equations.

Problem W-9. Suppose you have a least-squares problem with $m = n$ and a nonsingular coefficient matrix A . Is the least-squares solution of $A\mathbf{x} \approx \mathbf{b}$ the same as the ordinary solution of linear equations $A\mathbf{x} = \mathbf{b}$? (Explain why or why not.)

Problem W-10. (a) Suppose you wanted to find the point of the surface $P(t, u) = (3t - u - 5, u + 2t + 2, 4t + 2u - 7)$ that is closest to the origin. Show how to rephrase this as a least-squares problem. (You are not asked to finish the solution. Notice that this particular P is linear, or more precisely, affine.)

(b) Sometimes you need to invent a new method from ingredients that you know. Suppose you were given a *nonlinear* parametric surface $P(t, u)$ and were asked which point of it is closest to the origin. Invent a method to solve such a problem. (Combine ideas of linear approximation, Newtonian iteration, and least squares.)

Problem W-11. Find a parametric line $P(t)$ so that $P(0)$ is close to $\begin{bmatrix} 0 \\ 2 \end{bmatrix}$, $P(1)$ is close to $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$, $P(2)$ is close to $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$, and $P(3)$ is close to $\begin{bmatrix} 2 \\ 0 \end{bmatrix}$.

(Method: Similar to the example of a cubic parametric curve in §5.)

Problem W-12. Show how to compute a parametric curve $P(t)$ of degree at most 2 so that $P(0)$ is close to $\begin{bmatrix} 0 \\ 2 \end{bmatrix}$, $P(1)$ is close to $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$, $P(2)$ is close to $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$, and $P(3)$ is close to $\begin{bmatrix} 2 \\ 0 \end{bmatrix}$.

Rather than compute the coefficients numerically, you may leave them as a matrix expression involving matrices with explicit entries. (Matrix inverses can be left uncomputed.)

(Method: Similar to the example of a cubic parametric curve in §5.)