Least-squares solutions

Now we turn to the general problem of finding our best approximation to

$$A\mathbf{x} = \mathbf{b} \qquad (\star)$$

If $\mathbf{b}$ is in the column space, $C(A)$, of A, then $(\star)$ has a solution. Then we can find an $\mathbf{x}$ so that the residual

$$\mathbf{r} = \mathbf{b} - A\mathbf{x}$$

is zero. But if $\mathbf{b}$ is not in $C(A)$, then there is no solution. We will look for an approximation. We try to find a vector in $C(A)$ that is as close to $\mathbf{b}$ as possible. That is, we try to minimize the length

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

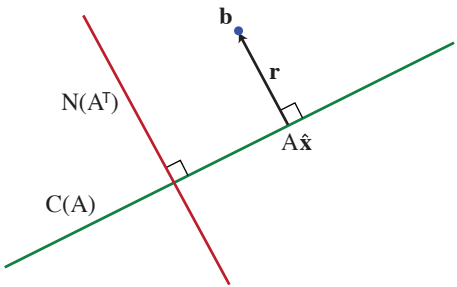
of the residual. Since $A\mathbf{x}$ is in $C(A)$ for any $\mathbf{x}$,

this length will be minimized when $\mathbf{r}$ *is orthogonal to* $C(A)$.

Recall the fundamental theorem of linear algebra. The column space is the orthogonal complement of the left nullspace:

$$C(A) = N(A^T)^\perp$$

Since $\|\mathbf{r}\|$ will be minimized when $\mathbf{r}$ is orthogonal to $C(A)$, the minimum occurs when $\mathbf{r} \in N(A^T)$.

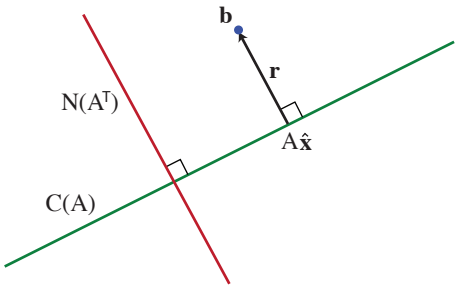


That is, the residual is minimized when

$$A^T\mathbf{r} = \mathbf{0} \quad \text{or} \quad A^T(\mathbf{b} - A\mathbf{x}) = \mathbf{0}$$

We call the solution of this equation $\hat{\mathbf{x}}$, and the equation for $\hat{\mathbf{x}}$ can be written in the following way:

$$A^T A\hat{\mathbf{x}} = A^T\mathbf{b}$$



$$A^T A\hat{\mathbf{x}} = A^T\mathbf{b} \qquad (\dagger)$$

An $\hat{\mathbf{x}}$ that satisfies this equation is called a *least-squares solution*. The equations in the system (†) are referred to as the *normal equations*. So least-squares solutions are solutions of normal equations.

It is an amazing fact that, regardless of what A is, even though $A\mathbf{x} = \mathbf{b}$ might not have a solution, when we multiply both sides by A^T to get the normal equations (†), there is always a solution, and moreover it is, in a very deep sense, the "best" solution we can find!

The least squares solution is unique, as long as the columns of A are LI. This is because...

Theorem

The nullspaces of A and A^TA are the same.

Proof. Suppose $A\mathbf{x} = \mathbf{0}$. Then $A^TA\mathbf{x} = 0$. Likewise, suppose $A^TA\mathbf{x} = 0$. Then multiply on the left by $\mathbf{x}^T$. So

$$0 = \mathbf{x}^TA^TA\mathbf{x} = (A\mathbf{x})^T(A\mathbf{x}) = \|A\mathbf{x}\|^2$$

which implies that $A\mathbf{x} = \mathbf{0}$.

The above theorem implies that if A has LI columns, and hence $A\mathbf{x} = 0$ has only the zero solution, then $A^TA\mathbf{x} = \mathbf{0}$ has only the zero solution, so A^TA is invertible. So we have proven the following.

Corollary

A^TA is invertible if and only if A has LI columns.

Linear fits to data

Suppose we are given n data points

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

and we want to fit a line through these points the best we can. We write our line as

$$y = mx + b$$

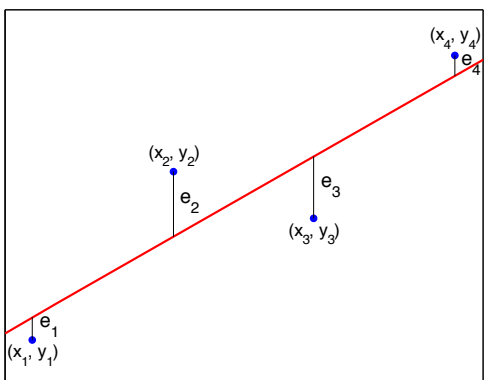
and try to find the best slope m and $y-$intercept b. **If** all the points lie on the line, we would then be able to solve the system of equations

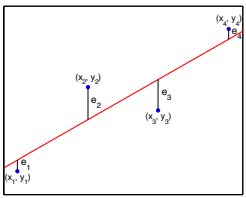
$$\begin{aligned} mx_1 + b &= y_1 \\ mx_2 + b &= y_2 \\ &\vdots \\ mx_n + b &= y_n \end{aligned} \qquad (\star)$$

For the line $\boxed{y = mx + b}$ m and b would satisfy

$$mx_1 + b = y_1$$
$$\vdots$$
$$mx_n + b = y_n$$

Generally, the points will not all lie on a line, so the above system is inconsistent. We will find a least-squares fit of the line to the data. The error at each point is the distance from y_i to the point on the line, or $e_i = y_i - (mx_i + b)$. The figure below shows a least squares line through 4 points, and the errors at each point.





The error vector $\mathbf{e} = (e_1, e_2, \ldots, e_m)$ contains all of these errors. The least squares line is the line that minimizes the sum of the squares of the errors, which is the same as minimizing the length of the error vector. That is, the least squares line minimizes

$$E = \|\mathbf{e}\|^2 = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - mx_i - b)^2$$

To find the least squares line we find the least squares solution to the inconsistent system $(\star)$. This inconsistent system can be written in matrix-vector form as

$$A\mathbf{v} = \mathbf{y}$$

$$mx_1 + b = y_1$$
$$mx_2 + b = y_2$$
$$\vdots$$
$$mx_n + b = y_n$$

This is

$$A\mathbf{v} = \mathbf{y}$$

where

$$A = \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix}, \qquad \mathbf{v} = \begin{pmatrix} m \\ b \end{pmatrix}, \qquad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \tag{0.1}$$

The vector $\mathbf{y}$ is given, and has the $y-$coordinates. The matrix A is given, and has the $x-$coordinates. We want to solve for the vector $\mathbf{v}$ to give us the slope and $y-$intercept of the least-squares line. So we find a least-squares solution by solving the normal equations

$$\boxed{A^T A\mathbf{v} = A^T\mathbf{y}} \tag{0.2}$$

The system

$$A^T A\mathbf{v} = A^T\mathbf{y}$$

for the least squares line is the 2×2 system

$$\begin{pmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & n \end{pmatrix} \begin{pmatrix} m \\ b \end{pmatrix} = \begin{pmatrix} \sum x_i y_i \\ \sum y_i \end{pmatrix} \tag{0.3}$$

Example

Find the least squares line through the points $(0, 0)$, $(2, 1)$, $(3, 2)$.

In the above example, the matrix

$$A = \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix}$$

is called the *design matrix* (written as X usually in your text). In that example, we fit the linear combination of the basis functions $\phi_1(x) = x,\ \phi_2(x) = 1$ as

$$y(x) = c_1\phi_1(x) + c_2\phi_2(x)$$

to the data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

Generally, to fit a linear combination of basis functions ϕ_i

$$y(t) = c_1\phi_1(t) + \cdots + c_m\phi_m(t)$$

to data $(t_1, y_1), \cdots, (t_n, y_n)$, we form the design matrix

$$X = \begin{pmatrix} \phi_1(t_1) & \cdots & \phi_m(t_1) \\ \vdots & \vdots & \vdots \\ \phi_1(t_n) & \cdots & \phi_m(t_n) \end{pmatrix}$$

Then, for $\beta = (c_1, \ldots, c_m)^T$, $y = (y_1, \ldots, y_n)^T$, we find the least squares solution of

$$X\beta = y$$

Pro-tip

This is implemented in Matlab as

$$\beta = X\backslash y$$

Example

Fit

$$y(t) = c_1 t + c_2 e^t$$

through the three points $(0, 1), (1, 1), (2, 2)$.

Curve fitting

In most situations a line is not the best fit for data. Usually we will try to fit some kind of curve to the data. This can be done to estimate some trend, or uncover some law underlying the observations. For example, one might try to fit population data with an exponential function to try to predict the future population. The method of least-squares can be used in these cases as well.

A classic example of curve fitting to discover a natural law was performed by the German astronomer Johannes Kepler in the early 1600's. Kepler tried to make sense of the incredibly accurate (naked eye!) observations of the Danish astronomer Tycho Brahe on the motion of the planets. Kepler found that the data did not fit the prevailing geocentric models nor the heliocentric model of Copernicus. These models were based on a 2000+ year old tradition that the motions of the planets were made up of circles. Kepler proposed the revolutionary idea that the planets not only orbited the sun, instead of the other way around, but that they did so in *ellipses*! This model fit the data better, although not by that much! Kepler formulated three laws of planetary motion based on this model. His *third law* states that the square of the orbital period of a planet is proportional to the cube of the semi-major axis of its orbit. In other words,

$$T^2 \propto x^3 \tag{0.4}$$

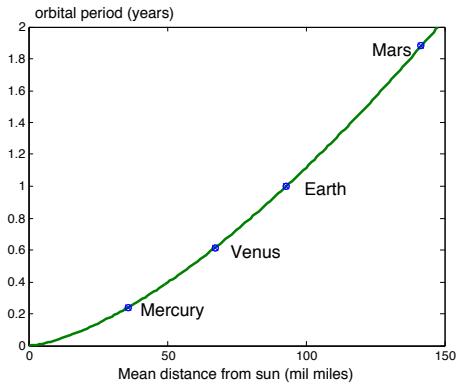
where T is the orbital period and x is the semi-major axis.

Kepler's Third Law: $T^2 \propto x^3$

Notice that we can take square roots on both sides and write the third law as

$$T = Cx^{3/2}$$

The figure below shows the orbital period vs the mean distance from the sun of the first four planets. The curve is $T = 0.0011x^{3/2}$.



Let's see how we can use the tools we have developed to fit a curve like this to the data. Suppose we suspect that the data follows some law like this, but we want to show that this is the best fit. So we can try to fit the data to a curve

$$T = Cx^m \tag{0.5}$$

where T is the period and x is the distance from the sun. If we can show that the best fit is when $m = 3/2$, that would be strong evidence for the third law.

So we try to fit a curve

$$T = Cx^m$$

to the data. In other words, we try to find the best values of C and m so that the error between the curve and the data is as small as possible. This is not a line, so we cannot use least squares directly. But, if we take logs of both sides, we get

$$m \ln x + \ln C = \ln T$$

which is linear in m and $\ln C$. Now suppose we have data for 4 of the planets. Then we get the system of 4 equations

$$\begin{pmatrix} \ln x_1 & 1 \\ \ln x_2 & 1 \\ \ln x_3 & 1 \\ \ln x_4 & 1 \end{pmatrix} \begin{pmatrix} m \\ \ln C \end{pmatrix} = \begin{pmatrix} \ln T_1 \\ \ln T_2 \\ \ln T_3 \\ \ln T_4 \end{pmatrix}$$

This is a system of the form $A\mathbf{v} = \mathbf{b}$, where $\mathbf{v} = \begin{pmatrix} m \\ \ln C \end{pmatrix}$. Solving the normal equations

$$A^T A \mathbf{v} = A^T \mathbf{b}$$

gives us an m that is very close to $3/2$! (Exercise.)

Planet	Orbital period (yr)	Ave dist from sun (million miles)
Mercury	0.240846	36
Venus	0.615	67.2
Earth	1	93
Mars	1.881	141.6
Jupiter	11.86	483.6
Saturn	29.456	886.7

Polynomial fits to data

Other kinds of curve fits can be put into a least-squares form as well. Suppose that we want to fit a parabola through a number of data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Let's call the parabola

$$p(x) = ax^2 + bx + c$$

Our goal is to find the "best" parabola, which means finding the best values of a, b and c. If the parabola fits exactly through all the points then

$$p(x_i) = y_i, \quad i = 1, 2, \ldots, n,$$

but generally the data will not lie exactly on a parabola. To find the best parabola, we set up the inconsistent system $p(x_i) = y_i$, which can be written as

$$\begin{aligned} ax_1^2 + bx_1 + c &= y_1 \\ ax_2^2 + bx_2 + c &= y_2 \\ &\vdots \\ ax_n^2 + bx_n + c &= y_n \end{aligned}$$

Or, in matrix-vector form as $A\mathbf{v} = \mathbf{y}$, where

$$A = \begin{pmatrix} x_1^2 & x_1 & 1 \\ x_2^2 & x_2 & 1 \\ \vdots & \vdots & \vdots \\ x_n^2 & x_n & 1 \end{pmatrix}, \quad \mathbf{v} = \begin{pmatrix} a \\ b \\ c \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

For fitting $p(x) = ax^2 + bx + c$, we get $A\mathbf{v} = \mathbf{y}$, where

$$A = \begin{pmatrix} x_1^2 & x_1 & 1 \\ x_2^2 & x_2 & 1 \\ \vdots & \vdots & \vdots \\ x_n^2 & x_n & 1 \end{pmatrix}, \quad \mathbf{v} = \begin{pmatrix} a \\ b \\ c \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

The matrix A above is called a *Vandermonde matrix*. As long as the x_i's are distinct, the columns will be LI. We then need to find the best approximation to $A\mathbf{c} = \mathbf{y}$. This is done by solving the normal equations

$$A^T A\mathbf{v} = A^T\mathbf{y}$$

Example

Find the best parabola through the points $(0,0), (1,2), (2,1), (3,0)$.

In this case the Vandermonde matrix and $\mathbf{y}$ are

$$A = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 1 & 1 \\ 4 & 2 & 1 \\ 9 & 3 & 1 \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} 0 \\ 2 \\ 1 \\ 0 \end{pmatrix}$$

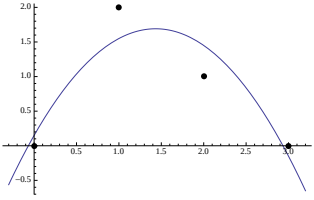
Therefore, the normal equations $A^T A\mathbf{c} = A^T\mathbf{y}$ are

$$\begin{pmatrix} 98 & 36 & 14 \\ 36 & 14 & 6 \\ 14 & 6 & 4 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 6 \\ 4 \\ 3 \end{pmatrix}$$

Solving this system by elimination gives $a = -3/4,\ b = 43/20,\ c = 3/20$. So the least-squares parabola through the 4 points is

$$p(x) = -\frac{3}{4}x^2 + \frac{43}{20}x + \frac{3}{20}$$

The least-squares parabola and the four points are shown below.



PROBLEM: A^TA is numerically unstable. We don't want to find the least squares solution of

$$X\beta = y$$

by finding the solution of

$$X^TX\beta = X^Ty$$

A much more stable and efficient method uses a clever implementation of the QR factorization due to Householder. Gram-Schmidt orthogonalization of the columns of A leads to the QR factorization of any matrix A as

$$\boxed{A = QR}$$

where Q is an $m \times m$ orthogonal matrix and R is upper triangular. We can also write the economy version of the QR factorization as

$$A = QR$$

where Q is $m \times n$ and R is $n \times n$.

To find the least squares solution of

$$X\beta = y$$

we write X as $X = QR$, then multiply both sides by $X^T = R^T Q^T$:

$$\begin{aligned} X^T X\beta &= X^T y \\ R^T Q^T QR\beta &= R^T Q^T y \\ R^T R\beta &= R^T Q^T y \\ R\beta &= Q^T y \end{aligned}$$

So we solve

$$R\beta = Q^T y$$

This is easy to solve since R is upper triangular. Yet, we don't actually find the QR factorization either! By performing a series of *Householder reflections*, we can reduce the problem to this upper triangular problem.

Householder Reflections

Let u be a unit vector and define

$$H = I - 2uu^T$$

Then

$$\begin{aligned} H^T H &= (I - 2uu^T)(I - 2uu^T) \\ &= I - 4uu^T + 4uu^T uu^T \\ &= I \end{aligned}$$

so H is orthogonal.

Moreover,

$$Hu = (I - 2uu^T)u = u - 2u = -u$$

and, if $v \perp u$, then

$$Hv = (I - 2uu^T)v = v - 2uu^T v = v$$

Thus H reflects vectors across the plane perpendicular to u.

Given x, if u bisects the angle between x and an axis, then Hx will lie on that axis.

So set

$$e_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

and let

$$v = x - \|x\|e_1, \qquad u = \frac{v}{\|v\|}$$

Then

$$\|v\|^2 = \|x\|^2 - 2\|x\|x_1 + \|x\|^2 = 2(\|x\|^2 - \|x\|x_1)$$
$$v^T x = \|x\|^2 - \|x\|x_1$$

So

$$uu^T x = \frac{1}{\|v\|^2}(v^T x)v = \frac{1}{2}v$$

and therefore

$$Hx = x - 2uu^T x = x - v = \|x\|e_1$$

$$Hx = \begin{pmatrix} \|x\| \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Likewise, setting

$$e_k = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix}$$

to be the kth standard basis vector and

$$v = x - \|x\| e_k, \qquad u = \frac{v}{\|v\|}$$

we have

$$Hx = \|x\| e_k$$

In other words, we put the length of x in the kth position of the vector.

QR FACTORIZATION

For a given A, find the Householder reflection $H_1, H_2, \ldots, H_n$ so that

$$H_1 A = \begin{pmatrix} \star & \star & \cdots & \star \\ 0 & \star & \cdots & \star \\ & & \vdots & \\ 0 & \star & \cdots & \star \end{pmatrix}, \qquad H_2 A = \begin{pmatrix} \star & \star & \star & \cdots & \star \\ 0 & \star & \star & \cdots & \star \\ 0 & 0 & \star & \cdots & \star \\ & & & \vdots & \\ 0 & 0 & \star & \cdots & \star \end{pmatrix}, \qquad \ldots$$

Then

$$H_n \cdots H_1 A = \begin{pmatrix} \star & \cdots & \star \\ & \ddots & \\ \mathbf{0} & & \star \\ & \mathbf{0} & \end{pmatrix} = R$$

is upper triangular.

Since orthogonal matrices are closed under multiplication,

$$Q^T = H_n \cdots H_1$$

is orthogonal. Thus, we have

$$A = QR$$

where

$$Q = H_1 H_2 \cdots H_n$$