Least squares method for solving overdetermined $Ax=b$

Note Title

We will consider 2 cases — A is full-rank
A is rank-deficient.

Problem: To find a "solution" for an overdetermined system $Ax=b$ i.e. a system with m eqns. & n unknowns, where $m > n$.

Such a problem does not have a solution in general.

The aim is to minimize the residual $r = b - Ax$.
i.e. to find x such that r is as small as possible (note that r cannot be zero in general).

The problem can be stated as-

given $A \in \mathbb{C}^{m \times n}$, $m \geq n$, $b \in \mathbb{C}^m$, find $x \in \mathbb{C}^n$ such that $\|b - Ax\|_2$ is minimized.

Such problems occur commonly in data fitting.
In this context, LSP fall into 2 categories - linear LSP & nonlinear LSP.
We will consider polynomial LSP (little more general than linear LSP but less general than non-linear LSP)

Polynomial least squares fitting

Problem: Given m distinct points $x_1, \ldots, x_m$ & data $y_1, \ldots, y_m$ at these points, $\{(x_1, y_1), \ldots, (x_m, y_m)\}$ find a polynomial $p(x)$ which fits this information in such a way that -

$p(x_1) = y_1$
$p(x_2) = y_2$
$\vdots$
$p(x_m) = y_m$

$$\sum |p(x_i) - y_i|^2 \text{ is minimized.}$$

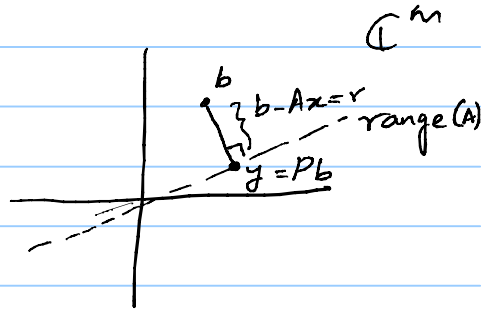
Suppose $p(x) = c_0 + c_1 x + \cdots + c_{n-1} x^{n-1}$ $(n \le m)$

The data fitting problem is to solve the system

$$\underbrace{\begin{bmatrix} 1 & x_1 & \cdots & x_1^{n-1} \\ 1 & x_2^2 & \cdots & x_2^{n-1} \\ \vdots & & & \\ 1 & x_m & \cdots & x_m^{n-1} \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_{n-1} \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}}_{b} \qquad (*)$$

To solve $(*)$ we want to find a point $Ax \in \text{range}(A)$
Such that $r = b - Ax$ is minimized.

Geometrically, this will happen when
x is such that $Ax = Pb$,
where P is the projection of $\mathbb{C}^m$
onto range(A).



$\therefore$ $r = b - Ax$ must be orthogonal to range(A).

Thm: Let $A \in \mathbb{C}^{m \times n}$, $b \in \mathbb{C}^m$ $(m \ge n)$

A vector $x \in \mathbb{C}^n$ minimizes $\|r\|_2 = \|b - Ax\|_2$

$\iff$ $r \perp$ range(A) i.e. $A^* r = 0$.

Proof: Note that for any $z \in \mathbb{C}^m$, $\exists$ unique $y \in$ range(A) & $w \in$ null(A^T) such that $z = y + w$.
$(\mathbb{C}^m = \text{range}(A) \oplus \text{null}(A^T))$.

Write $b = b_1 + b_2$, where $b_1 \in$ range A, $b_2 =$ null A^T.

$Ax - b = Ax - b_1 - b_2$, here $Ax - b_1 \in$ range A
$b_2 \perp (Ax - b_1)$

$$\therefore \underbrace{\|Ax - b\|_2^2}_{} = \underbrace{\|Ax - b_1\|_2^2}_{} + \underbrace{\|b_2\|_2^2}_{}$$

Since b_2 is fixed, $\|Ax-b\|$ is minimized

$\Leftrightarrow$ $\|Ax-b_1\|$ is minimized.

$\Leftrightarrow$ $Ax-b_1=0$ i.e. $Ax=b_1$

$\Leftrightarrow$ $Ax-b = b_2 \in \text{null}(A^T)$.

i.e. $r \in \text{null}(A^T)$.

$\therefore r \perp$ range A.

Lemma: $r \perp \text{range}(A) \Leftrightarrow \exists\, x$ minimizing $r=Ax-b$

$\Leftrightarrow A^*Ax = A^*b$.

normal system of eqns. for $Ax=b$.

Pf: We know that $r \perp \text{range}(A) \Leftrightarrow A^*r=0$

Let P be the orthogonal projection onto range A,

then $P = A(A^*A)^{-1}A^*$

Consider $Pb = \left(A(A^*A)^{-1}A^*\right)b = A(A^*A)^{-1}A^*Ax$

$= AA^{-1}A^{*-1}A^*Ax$.

$= Ax$.

Thus, P projects b onto Ax

i.e. x is the factor for which r is minimized.

The system of equations $A^*Ax = A^*b$ is called system of "normal equations" for the given system $Ax=b$.

Theorem: The system of equations $A^*Ax = A^*b$ has a unique solution $\Leftrightarrow$ A has full rank.

Proof: If $A^*Ax = A^*b$ does not have a unique solution

i.e. A^*A is singular, then $A^*Ay=0$ for some $y \neq 0$.

$\Rightarrow y^*A^*Ay=0$

$\Rightarrow \langle Ay, Ay\rangle = 0 \Rightarrow Ay=0$,

which is a contradiction since A is full rank.

Conversely, if A is not full rank, then

$\exists\ y \neq 0$ s.t. $Ay = 0 \quad \Rightarrow A^*Ay = 0$

$\Rightarrow A^*A$ is singular,
which is a contradiction

To show that $y = Pb$ is the unique point that minimizes $\|b - y\|_2$, let $z \neq y$ be another such point, then $z \perp b \rightarrow z - y \perp b - y$

$$\Rightarrow \|b - z\|_2^2 = \|b - y\|_2^2 + \underbrace{\|y - z\|_2^2}_{>0}$$

$$\Rightarrow \|b - z\|_2^2 > \|b - y\|_2^2,$$

which is a contradiction.

To summarize:

If A has full rank, then the solution x to the LSP (*) is unique and is given by $Ax = Pb$

i.e. $x = \underbrace{(A^*A)^{-1}A^*}\, b.$

this matrix is called the pseudo-inverse of A, denoted by A^+.

$\therefore$ The LSP ($Ax = b$) for a full-rank matrix A reduces to computing $x = A^+ b$.

There are 3 popular methods of (*):

(I) Solve the normal equations -

A full rank $\Rightarrow A^*A$ is HPD, $\therefore A^*Ax = A^*b$ can be solved using Cholesky factorization.

Algorithm: ① form A^*A & A^*b.

② Compute Cholesky factorization of A^*A as $A^*A = R^*R$ (R is upper Δ^r.)

③ Solve the lower Δ^r system $R^*w = A^*b$ for w

④ Solve the upper Δ^r system $Rx = w$ for x.

Operation count is dominated by the first 2 steps, which require $mn^2 + \frac{n^3}{3}$ flops
(mn^2 for A^*A; $\frac{n^3}{3}$ for Cholesky decomp.)

(II) Using QR factorization :

Let $A = QR$ be the QR factorization, then $P = QQ^*$, so $y = Pb = QQ^*b$

Since $y \in \text{range}(A)$, the system $Ax = y$ has an exact solution

range: $Ax = Pb$

$Ax = y = Pb$

$\therefore QRx = QQ^*b$

$\therefore Rx = Q^*b$

$P = A(A^*A)^{-1}A^*$
$\vdots$
$= QQ^*$.

Algorithm: ① Compute $A = QR$
② Compute Q^*b
③ Solve the upper Δ^r system $Rx = Q^*b$ for x.

Operation count is dominated by the first step
$\sim 2mn^2 - \frac{2n^3}{3}$ flops (using Householder's method.)

(III) Using SVD -
Let $A = U\Sigma V^*$
Then $P = UU^*$ (check: $P = A(A^*A)^{-1}A^*$).

$\therefore Ax = Pb \implies U\Sigma V^* x = UU^* b$

$\implies \Sigma V^* x = U^* b.$

Algorithm: 1) Compute SVD $A = U\Sigma V^*$.

2) Compute vector $U^* b$

3) Solve the diagonal system

$\Sigma w = U^* b$ for w $(\because V^* x = w)$

4) Compute $x = Vw$

Operation count is dominated by the first step.

Some remarks about full rank LSP.

① The choice of the norm for measuring the residual $r = Ax - b$ is important. In theory one could choose any induced matrix norm; different norms would produce different optimum solutions.
Minimization in the 1-norm & ∞-norm is complicated by the fact that the function $f(x) = \|Ax-b\|_p$ is not differentiable for these values of p.

On the other hand, the 2-norm is good for the foll. reasons:
(i) $\phi(x) = \frac{1}{2} \cdot \|Ax-b\|_2^2$ is a differentiable function of x & so the minimizers of ϕ satisfy the gradient eqn. $\nabla \phi(x) = 0$

{small example: $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix}$

$$Ax=b: \quad \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}$$

i.e.
$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 &= b_1 \\ a_{21}x_1 + a_{22}x_2 &= b_2 \\ a_{31}x_1 + a_{32}x_2 &= b_3. \end{aligned}\right\}$$

$$\phi(x) = \frac{1}{2}\|Ax-b\|_2^2 = \frac{1}{2}\left((a_{11}x_1 + a_{12}x_2 - b_1)^2 + (a_{21}x_1 + a_{22}x_2 - b_2)^2 + (a_{31}x_1 + a_{32}x_2 - b_3)^2\right)$$

$$\nabla\phi(x) = \begin{pmatrix} (a_{11}x_1 + a_{12}x_2 - b_1)a_{11} + (\quad)a_{21} + (\quad)a_{31} \\ (a_{11}x_1 + a_{12}x_2 - b_1)a_{12} + (\quad)a_{22} + (\quad)a_{32} \end{pmatrix}$$

$$= \begin{pmatrix} a_{11} & a_{21} & a_{31} \\ a_{12} & a_{22} & a_{32} \end{pmatrix} \begin{pmatrix} a_{11}x_1 + a_{12}x_2 - b_1 \\ a_{21}x_1 + a_{22}x_2 - b_2 \\ a_{31}x_1 + a_{32}x_2 - b_3 \end{pmatrix}$$

$$= A^T(Ax - b)$$

So $\nabla \phi(x) = 0$ means $A^T A x = A^T b$.

(ii) The 2-norm is invariant under unitary / orthogonal transformations. So, in theory, to solve a system $Ax = b$, one could solve $Q^T A x = Q^T b$ for some orthogonal matrix Q. (pre-conditioning).

(2) Even when A is full rank, trouble can be expected if A is "nearly rank-deficient" i.e. columns of A are nearly dependent -

eg: $A = \begin{pmatrix} 1 & 0 \\ 0 & 10^{-6} \\ 0 & 0 \end{pmatrix}$ (Refer: section 5.3.1. of G-L)

Rank-deficient LSPs.

If A is rank-deficient, then there are infinitely many solutions to the LSP:

Consider the system $Ax = b$, where $A \in \mathbb{R}^{m \times n}$, $m \geq n$ & $r = \text{rank } A < n$.

$Ax - b = r$.

$A(x+z) - b$
$= Ax + Az - b$
$= Ax - b = r$.

Suppose x is a minimiser of $Ax - b$, then $x + z$ is also a minimiser for any $z \in \text{null}(A)$.

(note that rank $A < n \Rightarrow$ null (A) is non-empty)

In this case, first the (numerical) rank of A must be determined & then the solution can be identified.

To solve a rank-deficient LSP, we use "complete orthogonal factorization" – if Q & Z are orthogonal matrices such that $Q^T A Z = \begin{bmatrix} T_{11} & 0 \\ 0 & 0 \end{bmatrix}$ (r rows, $m-r$ rows; r cols, $n-r$ cols), $r = \operatorname{rank} A$.

(So $A = Q \begin{bmatrix} T_{11} & 0 \\ 0 & 0 \end{bmatrix} Z^T$) ↙

Then $\|Ax-b\|_2^2 = \|AZZ^Tx-b\|_2^2 = \|(Q^TAZ)Z^Tx - Q^Tb\|_2^2$

$= \|T_{11}w - c\|_2^2 + \|d\|_2^2$

where $Z^Tx = \begin{pmatrix} w \\ y \end{pmatrix}\begin{matrix} \}r \\ \}n-r \end{matrix}$, $Q^Tb = \begin{pmatrix} c \\ d \end{pmatrix}$

If x is to be a minimizer of $\|Ax-b\|_2^2$ then we must have $\|T_{11}w - c\|_2^2 = 0$ i.e. $w = T_{11}^{-1}c$.

We can choose y to be zero, so that $Z^Tx = \begin{pmatrix} T_{11}^{-1}c \\ 0 \end{pmatrix}\begin{matrix} \}r \\ \}n-r \end{matrix}$

so $x_{LS} = Z\begin{pmatrix} T_{11}^{-1}c \\ 0 \end{pmatrix}$. ✱

The SVD is a particularly "revealing" complete orthogonal factorization.

Theorem: Suppose $U^TAV = \Sigma$ is the SVD of $A \in \mathbb{R}^{m\times n}$ ($r < \min\{m,n\}$), with $r = \operatorname{rank} A$. If $U = [u_1|\cdots|u_m]$, $V = [v_1|\cdots|v_n]$ and $b \in \mathbb{R}^m$, then

$$(x_{LS})_j = \sum_{i=1}^{r} \frac{u_i^Tb}{\sigma_i} v_{ji}$$

($A: \mathbb{R}^n \to \mathbb{R}^m$)

minimizes $\|Ax-b\|_2$, and it has the smallest 2-norm among all minimizers.

Proof: (This is just a re-wording of the argument on the earlier page in the language of SVD -)

$$\Sigma = \begin{bmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_r & \\ & & & 0 \\ & & & & \ddots \\ & & & & & 0 \end{bmatrix}_{m \times n}, \quad w = \begin{pmatrix} w_1 \\ \vdots \\ w_m \end{pmatrix}$$

$$\|Ax - b\|_2^2 = \|\underbrace{U^T A V}_{\Sigma} \underbrace{V^T x}_{w} - U^T b\|_2^2$$

$$= \|\Sigma w - U^T b\|_2^2, \quad \text{where } w = V^T x$$

$$= \sum_{i=1}^{m} (\sigma_i w_i - u_i^T b)^2$$

$$= \sum_{i=1}^{r} (\sigma_i w_i - u_i^T b)^2 + \underbrace{\sum_{i=r+1}^{m} (u_i^T b)^2}_{\neq 0, \text{ no control over this term}}, \quad \text{since } \sigma_{r+1} = \cdots = \sigma_m = 0.$$

To minimize the above sum, we choose w such that

$$\sigma_i w_i - u_i^T b = 0 \text{ for } 1 \le i \le r \quad \& \quad \underbrace{w_i = 0 \text{ for } i > r.}$$

this will ensure that x has minimal norm.

i.e. $$w_i = \begin{cases} \dfrac{u_i^T b}{\sigma_i} & \text{for } 1 \le i \le r \\ 0 & \text{for } i > r. \end{cases}$$

$$\therefore x_{LS} = Vw = V \begin{pmatrix} \frac{u_1^T b}{\sigma_1} \\ \vdots \\ \frac{u_r^T b}{\sigma_r} \\ 0 \\ \vdots \\ 0 \end{pmatrix}_{n \times 1} = \begin{pmatrix} \sum_{i=1}^{r} \frac{u_i^T b}{\sigma_i} v_{1i} \\ \vdots \\ \sum_{i=1}^{r} \frac{u_i^T b}{\sigma_i} v_{ni} \end{pmatrix}. \quad \square$$

Note ① As in the case of full rank LSPs, this solution can also be expressed as

$$x_{LS} = A^+ b, \quad \text{(Check!)}$$

where $A^+ = V \Sigma^+ U^T$ is the pseudo-inverse of A

$$\left(\Sigma^+ = \text{diag}\left(\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_r}, 0, \ldots 0 \right) \right).$$

② QR with column pivoting can also be used to solve rank-deficient LSPs (coming soon)

Summary: $Ax = b$.

Full rank LSP : unique soln. given by

$$x_{LS} = A^{+} b.$$

Rank-deficient LSP : there are infinitely many solns.

$x_{LS} = A^{+} b$, this is minimum in 2-norm.