1 best solution for $A \cdot x = b$: Least square method

One important point is that $A \cdot x = b$ are usually large systems with a large number of equations and relatively smaller number of unknowns. Thus, the equation is most probably not solvable.

1.1 what is "best" and why ?

The term "best" refers to the following error vector $e = b - A \cdot \alpha$. where $\alpha = A \cdot (A^T \cdot A)^{-1} \cdot A^T$. The best solution x is the one minimizing the norm of the error: $|e|^2 = e^T \cdot e$. The solution uses calculus:

1.2 $\frac{d}{dx}(x^T \cdot A \cdot x)$

$$\frac{d}{dx}(x^T \cdot A \cdot x) = \begin{bmatrix} \frac{d}{dx_1}(x^T \cdot A \cdot x) \\ \frac{d}{dx_2}(x^T \cdot A \cdot x) \\ .. \\ \frac{d}{dx_n}(x^T \cdot A \cdot x) \end{bmatrix}$$

$$x = \begin{bmatrix} x_1 \\ x_2 \\ .. \\ x_n \end{bmatrix} \qquad \text{define } x$$

$$x^T \cdot A \cdot x = \begin{bmatrix} x_1 & x_2 & .. & x_n \end{bmatrix} \cdot \begin{bmatrix} a_{11} & a_{12} & .. & a_{1n} \\ a_{21} & a_{22} & .. & a_{2n} \\ .. & .. & .. & .. \\ a_{n1} & a_{n2} & .. & a_{nn} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ .. \\ x_n \end{bmatrix} \qquad \text{substitute } x \text{ for its matrix form}$$

$$= \begin{bmatrix} \sum_{i=1}^n x_i \cdot a_{i1} & \sum_{i=1}^n x_i \cdot a_{i2} & .. & \sum_{i=1}^n x_i \cdot a_{in} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ .. \\ x_n \end{bmatrix}$$

$$= x_1 \cdot \sum_{i=1}^n x_i \cdot a_{i1} + x_2 \cdot \sum_{i=1}^n x_i \cdot a_{i2} + .. + x_n \cdot \sum_{i=1}^n x_i \cdot a_{in}$$

Let's now consider for simplicity $d = \frac{d}{dx_1}(x^T \cdot A \cdot x)$ as the parameters are symmetric. then

$$
\begin{aligned}
d &= \frac{d}{dx_1}(x_1 \cdot \sum_{i=1}^{n} x_i \cdot a_{i1} + x_2 \cdot \sum_{i=1}^{n} x_i \cdot a_{i2} + .. + x_n \cdot \sum_{i=1}^{n} x_i \cdot a_{in}) \\
&= 2 \cdot x_{11} \cdot a_{11} + \sum_{i=2}^{n} x_i \cdot a_{i1} + x_2 \cdot a_1 2 + .. + x_n \cdot a_{in}
\end{aligned}
$$

$$\frac{d}{dx_1}(x^T \cdot A \cdot x) = \sum_{i=1}^{n} x_i \cdot a_{1i} + \sum_{i=1}^{n} x_i \cdot a_{i1}$$

$$\frac{d}{dx_j}(x^T \cdot A \cdot x) = \sum_{i=1}^{n} x_i \cdot a_{ji} + \sum_{i=1}^{n} x_i \cdot a_{ij} \qquad \text{more generally}$$

$$\frac{d}{dx}(x^T \cdot A \cdot x) = (A^T + A) \cdot x \qquad \text{vectorized form}$$

additionally let's consider the term

$$\frac{d}{dx}(x \cdot A^T) = A \tag{1}$$

where A is a column vector.

1.3 proof

Having all the components of the proof set, let's put the pieces together. we have:

$$e^T \cdot e = (b - A \cdot \alpha)^T \cdot (b - A \cdot \alpha) = x^T A^T A x - x^T A^T b - b^T A x + b^T b \tag{2}$$

minimizing the error is equivalent to minimizing:

$$x^T A^T A x - x^T A^T b - b^T A x$$

using the proofs considered above:

$$\frac{d}{dx} e = 2 \cdot A^T A x - 2 \cdot A^T b \tag{3}$$

The vector x achieving the minimum error is

$$x = (A^T A)^{-1} \cdot A^T b \tag{4}$$

2 Orthogonal matrices

2.1 Properties

if we have vectors $x_1, x_2, .., x_n$ non-zero vectors, such $x_i^T \cdot x_j = 0 \quad \forall i \neq j$ then they are independent. The proof is as follows

$$
\begin{aligned}
Q &= \begin{bmatrix} x_1 & x_2 & .. & x_n \end{bmatrix} \\
Q^T \cdot Q &= \begin{bmatrix} |x_1|^2 & 0 & 0 & .. & 0 \\ 0 & |x_2|^2 & 0 & .. & 0 \\ .. & & & & \\ 0 & 0 & 0 & .. & |x_n|^2 \end{bmatrix} && \text{using } x_i^T \cdot x_j = 0 \\
Q^T \cdot Q \textit{ is invertible } &\implies N(Q^T \cdot Q) = Z. \\
&\implies N(Q) = Z \qquad \text{using } N(Q^T \cdot Q) = N(Q)
\end{aligned}
$$

$N(Q) = Z \implies$ the columns are independent. Thus, the claim is proved. If Q is a square orthonormal matrix , then we have $Q^T \cdot Q = Q \cdot Q^T = I$.

2.2 Grim-schmidt Orthogonalization

The main goal of this process is to turn a matrix A with ***independent columns*** into a orthonormal matrix Q such that $C(Q) = C(A)$. Assuming:

$$
\begin{aligned}
A &= \begin{bmatrix} a_1 & a_2 & .. & a_n \end{bmatrix} && \text{define } A \\
Q &= \begin{bmatrix} q_1 & q_2 & .. & q_n \end{bmatrix} && \text{define } Q \\
q_1 &= a_1 \\
q_i &= a_i - \sum_{j=1}^{i-1} \frac{q_j \cdot q_j^T}{q_j^T \cdot q_j} \cdot a_i && \forall\ i > 1 \qquad \text{orthogonality} \\
q_i &\leftarrow \frac{q_i}{|q_i|^2} && \text{ensure a length of 1}
\end{aligned}
$$

2.2.1 orthogonality

Let's prove that $q_i^T \cdot q_j = 0 \; \forall \; i \neq j$. Let's first prove it for $n = 3$. let a, b, c independent vectors, and A, B, C the orthonormal vectors, then:

$$\begin{aligned}
A &= a \\
B &= b - \frac{A \cdot A^T}{A^T \cdot A} \cdot b \\
C &= c - \frac{A \cdot A^T}{A^T \cdot A} \cdot c - \frac{B \cdot B^T}{B^T \cdot B} \cdot c \\
A^T \cdot B &= A^T \cdot b - A^T \cdot \frac{A \cdot A^T}{A^T \cdot A} \cdot b \\
&= A^T \cdot b - A^T \cdot b = 0 && A \text{ and } B \text{ are orthogonal} \\
A^T \cdot C &= A^T (c - \frac{A \cdot A^T}{A^T \cdot A} \cdot c - \frac{B \cdot B^T}{B^T \cdot B} \cdot c) \\
&= A^T \cdot c - A \cdot \frac{A \cdot A^T}{A^T \cdot A} \cdot c - 0 && \text{use } A \text{ and } B \text{ orthogonality} \\
&= A^T \cdot b - A^T \cdot b = 0 && A \text{ and } C \text{ orthogonal}
\end{aligned}$$

using the same argument, it is easy to see that B and C are orthogonal.
The second step is to prove the following $P(n) \implies P(n+1)$ with $P(n)$: result true for n.
Assuming $a_1, a_2, ...a_n$ and $q_1, q_2.., q_n$ satisfying our assumption. Let a_{n+1} a vector independent of the rest of the vectors and:

$$\begin{aligned}
q_{n+1} &= a_{n+1} - \sum_{j=1}^{n} \frac{q_j \cdot q_j^T}{q_j^T \cdot q_j} \cdot a_{n+1} && \text{define } q_{n+1} \\
q_k^T \cdot a_{n+1} = q_k^T \cdot a_{n+1} - q_k^T \cdot \frac{q_k \cdot q_k^T}{q_k^T \cdot q_k} \cdot a_{n+1} - q_k^T \cdot & \sum_{j=1, j \neq k}^{n} \frac{q_j \cdot q_j^T}{q_j^T \cdot q_j} \cdot a_{n+1} && \text{substitution} \\
q_k^T \cdot \sum_{j=1, j \neq k}^{n} \frac{q_j \cdot q_j^T}{q_j^T \cdot q_j} \cdot a_{n+1} = \sum_{j=1, j \neq k}^{n} q_k^T \cdot & \frac{q_j \cdot q_j^T}{q_j^T \cdot q_j} \cdot a_{n+1} = 0 && q_i^T \cdot q_j = 0 \; \forall \; i \neq j \\
q_k^T \cdot a_{n+1} &= 0
\end{aligned}$$

Thus the claim is proved.

2.2.2 The same column space

Let's consider the following:

$$
\begin{aligned}
q_i &= a_i - \sum_{j=1}^{i-1} \frac{q_j \cdot q_j^T}{q_j^T \cdot q_j} \cdot a_i && \forall\ i > 1 \\
&= \sum_{j=1}^{i-1} q_j \cdot \frac{q_j^T \cdot a_i}{q_j^T \cdot q_j} && \text{a slight reformulation} \\
q_j \cdot \frac{q_j^T \cdot a_i}{q_j^T \cdot q_j} &= q_j \cdot \alpha && \text{with } \alpha \in \mathbf{R} \text{ as } q_j \text{ and } a_j \text{ are column vectors} \\
q_i &= a_i - \sum_{j=1}^{i-1} q_j \cdot \alpha_j && \text{reformulate } q_i
\end{aligned}
$$

as $q_1 = a_1$, then it is easy to see that q_i is a linear combination of other vectors a_j. Then any linear combination of q_i belongs to $C(A)$. and since q_i are independent (as they are one to one orthogonal), they span the whole $C(A)$. thus $C(A) = Q(A)$