Chap 3. Orthogonality

3.1 Orthogonal Vectors and Subspaces

For $x=(x_1,\cdots,x_n)$, $\|x\|=\sqrt{x^T\cdot x}$: the length of x.

$$\|x\|^2=x_1^2+x_2^2+\cdots+x_n^2$$

$x, y\in\mathbb{R}^n$: orthogonal $\Leftrightarrow$ $x^T\cdot y=x_1y_1+\cdots+x_ny_n=0$

the inner product of x and y.

3A

The inner product $x^T\cdot y=0$ iff x and y are orthogonal.

If $x^Ty>0$, their angle is less than $90°$.

If $x^T\cdot y<0$, their angle is greater than $90°$.

Useful Fact: If nonzero vectors $v_1,\cdots,v_k$ are mutually orthogonal, then those vectors are linearly independent.

Pf: Suppose $c_1v_1+\cdots+c_kv_k=0$.

$$\Rightarrow \quad v_i(c_1v_1+\cdots+c_kv_k)=c_iv_i^T\cdot v_i=0$$

Since v_i are nonzero, $v_i^T\cdot v_i\neq 0$ and $c_i=0$ ∎

The coordinate vectors $e_1,\cdots,e_n\in\mathbb{R}^n$ are mutually orthogonal unit vectors. When they are rotated, the result is a new orthonormal basis: a new system of mutually orthogonal unit vectors in $\mathbb{R}^n$. In $\mathbb{R}^2$, $v_1=(\cos\theta,\sin\theta)$, $v_2=(-\sin\theta,\cos\theta)$ form an orthonormal basis.

Orthogonal Subspaces

3B Two subspaces V and W of $\mathbb{R}^n$ are orthogonal if every vector $v\in V$ is orthogonal to every $w\in W$:

$v^T\cdot w=0$ for all v and w.

3C Fundamental Theorem of orthogonality

The row space is orthogonal to the nullspace (in $\mathbb{R}^n$).
The column space is orthogonal to the left nullspace (in $\mathbb{R}^m$).

Proof $x \in N(A) \Leftrightarrow Ax = 0$

$v \in C(A^T) \Leftrightarrow v = A^T z$ for some z

$\Rightarrow v^T x = z^T A x = 0$

$\therefore N(A) \perp C(A^T)$ ⌟

Definition (orthogonal Complement)

V: a subspace of $\mathbb{R}^n$.

$V^\perp$: the orthogonal complement of V is the space of all vectors orthogonal to V.

↳ "V perp".

3D Fundamental Theorem of Linear Algebra, Part II

The nullspace is the orthogonal complement of the row space (in $\mathbb{R}^n$)
The left nullspace is the orthogonal complement of the column space in $\mathbb{R}^m$.

3E $Ax = b$ is solvable iff $y^T b = 0$ whenever $y^T A = 0$

Proof $b \in C(A) \Leftrightarrow b \in (N(A^T))^\perp$ ⌟

Ex:

$$A = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ -1 & 0 & 1 \end{bmatrix} \Rightarrow y = (1,1,1)^T \in N(A^T)$$

since $y^T A = 0$.

Thus, $Ax = b$ is solvable $\Leftrightarrow b_1 + b_2 + b_3 = 0$.

↳ $y^T b$.

The Matrix and the Subspaces

If $W = V^\perp$, then $V = W^\perp$ and $\dim V + \dim W = n$.
In other words $V^{\perp\perp} = V$. The whole space $\mathbb{R}^n$ can be decomposed into two perpendicular parts.

3F From the row space to the column space, A is actually invertible. Every vector b in the column space comes from exactly one vector x_r in the row space.

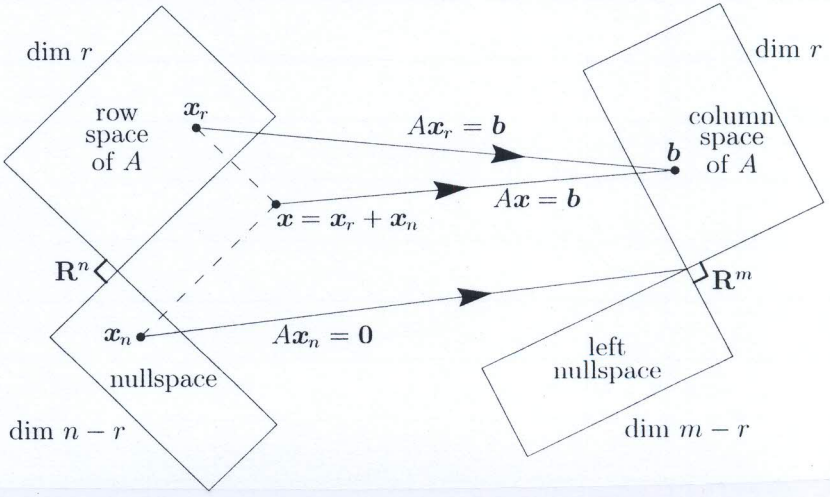
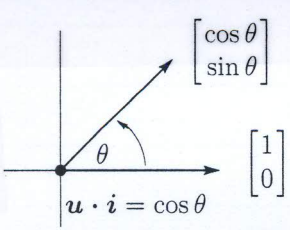


Figure 3.4 The true action $Ax = A(x_{\text{row}} + x_{\text{null}})$ of any m by n matrix.

Too. $b \in C(A)$, $b = Ax = Ax_r$, where $x = x_r + x_n$ and $Ax_n = 0$.
Let x_r' in the row space and $Ax_r' = b$,
then $A(x_r - x_r') = b - b = 0$, and $x_r - x_r' \in N(A)$
and $x_r - x_r'$ is also in the row space $\Rightarrow x_r = x_r'$.
Exactly one vector in the row space is carried to b. ∎

0. A^T goes from $\mathbb{R}^m$ to $\mathbb{R}^n$ and from $C(A)$ to $C(A^T)$.
But $A^T \neq A^{-1}$ in general. A^T moves the spaces correctly, but not the individual vectors. When A^{-1} fails to exist, the best substitute is the pseudoinverse A^+:
$A^+Ax = x$ for $x \in C(A^T)$ and $A^+y = 0$ for $y \in N(A^T)$.

3.2 Cosines and Projections onto Lines



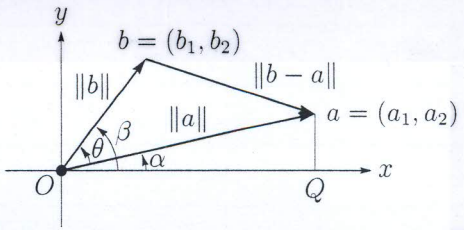


Figure 3.6 The cosine of the angle $\theta = \beta - \alpha$ using inner products.

Cosine formula: $\cos\theta = \cos\beta\cos\alpha + \sin\beta\sin\alpha = \dfrac{a_1b_1 + a_2b_2}{\|a\|\,\|b\|}$

$\dfrac{b_1}{\|b\|}$ $\dfrac{a_1}{\|a\|}$ $\dfrac{b_2}{\|b\|}$ $\dfrac{a_2}{\|a\|}$

3G $\cos\theta = \dfrac{a^Tb}{\|a\|\,\|b\|}$

Law of Cosines: $\|b-a\|^2 = \|b\|^2 + \|a\|^2 - 2\|b\|\,\|a\|\cos\theta$.

$\Rightarrow b^Tb - 2a^Tb + a^Ta = b^Tb + a^Ta - 2\|b\|\,\|a\|\cos\theta$

$\therefore a^Tb = \|a\|\,\|b\|\cos\theta$

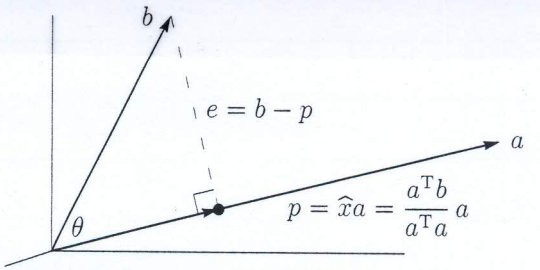


Figure 3.7 The projection p of b onto a, with $\cos\theta = \dfrac{Op}{Ob} = \dfrac{a^Tb}{\|a\|\,\|b\|}$.

3H The projection of b onto the line in the direction of a is

$p = \hat{x}a = \dfrac{a^Tb}{a^Ta}a$ $\quad \left[\because (b-\hat{x}a)\perp a \Leftrightarrow a^T(b-\hat{x}a) = 0\right]$

3Q All vectors a and b satisfy the Schwartz inequality:

$$|a^T b| \le \|a\| \|b\|,$$

which is $|\cos\theta| \le 1$ in $\mathbb{R}^n$.

Too:

$$\left\| b - \frac{a^T b}{a^T a} a \right\|^2 = b^T b - 2\frac{(a^T b)^2}{a^T a} + \left(\frac{a^T b}{a^T a}\right)^2 a^T a$$

$$= \frac{(b^T b)(a^T a) - (a^T b)^2}{a^T a} \ge 0 \quad \square$$

The equality holds iff b is a multiple of a.

Ex1: Project $b = (1,2,3)$ onto the line through $a = (1,1,1)$.

$$\hat{x} = \frac{a^T b}{a^T a} = \frac{6}{3} = 2 \quad \text{and} \quad p = \hat{x} a = (2,2,2).$$

$$\cos\theta = \frac{\|p\|}{\|b\|} = \frac{\sqrt{12}}{\sqrt{14}} \quad \text{and} \quad \cos\theta = \frac{a^T b}{\|a\| \|b\|} = \frac{6}{\sqrt{3}\sqrt{14}}.$$

$\Rightarrow 6 \le \sqrt{3}\sqrt{14}$ and $\sqrt{36} \le \sqrt{42}$.

The cosine is less than 1 since b is not parallel to a.

Projection Matrix of Rank 1

$$p = a \cdot \frac{a^T b}{a^T a} \Rightarrow P = \frac{a \cdot a^T}{a^T a} : \text{projection matrix}$$

① P is a symmetric matrix

② Its square is itself: $P^2 = P$.

Ex2: $a = (1,1,1)$

$$P = \frac{a \cdot a^T}{a^T a} = \frac{1}{3}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}\begin{bmatrix} 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$$

Remark on scaling:

The projection matrix is the same if a is scaled.

$$a=\begin{bmatrix}\alpha\\ \alpha\\ \alpha\end{bmatrix} \text{ gives } P=\frac{1}{3\alpha^2}\begin{bmatrix}\alpha\\ \alpha\\ \alpha\end{bmatrix}[\alpha\ \alpha\ \alpha]=\begin{bmatrix}\frac{1}{3}&\frac{1}{3}&\frac{1}{3}\\ \frac{1}{3}&\frac{1}{3}&\frac{1}{3}\\ \frac{1}{3}&\frac{1}{3}&\frac{1}{3}\end{bmatrix}$$

If a has a unit length, $P=aa^T$ $(a^Ta=1)$.

Ex3: Project onto the θ-direction in the xy-plane.

The line goes through $a=(\cos\theta,\sin\theta)$.

$$P=\frac{aa^T}{a^Ta}=\underbrace{\frac{1}{c^2+s^2}}_{1}\begin{bmatrix}c\\ s\end{bmatrix}[c\ s]=\begin{bmatrix}c^2&cs\\ cs&s^2\end{bmatrix}$$

- This matrix P was discovered in Section 2.6.
  Now we know P in any dimensions.
- To project b onto a, multiply by the projection matrix
  $p=Pb$.

3.3 Projections and Least Squares

Least squares solution for $ax = b$ by minimizing
$E^2 = \|ax - b\|^2 = (a_1x - b_1)^2 + \cdots + (a_mx - b_m)^2$.

$$\frac{1}{2}\frac{dE^2}{dx} = (a_1x - b_1)a_1 + \cdots + (a_mx - b_m)a_m = 0$$

3K The least squares solution to a problem $ax = b$ in one unknown is $\hat{x} = a^Tb/a^Ta$.

Orthogonality of a and e: $a^T(b - \hat{x}a) = a^Tb - \frac{a^Tb}{a^Ta}a^Ta = 0$

e: the error vector

Least-Squares Problem with Several Variables

$Ax = b$, where A is an $m \times n$ matrix.

$\Rightarrow E = \|Ax - b\|$: error, which is the distance from b to the point Ax in the column space.

- Locate the point $p = A\hat{x}$ that is closer to b than any other point in the column space.
- The error $e = b - A\hat{x}$ must be perpendicular to the column space.

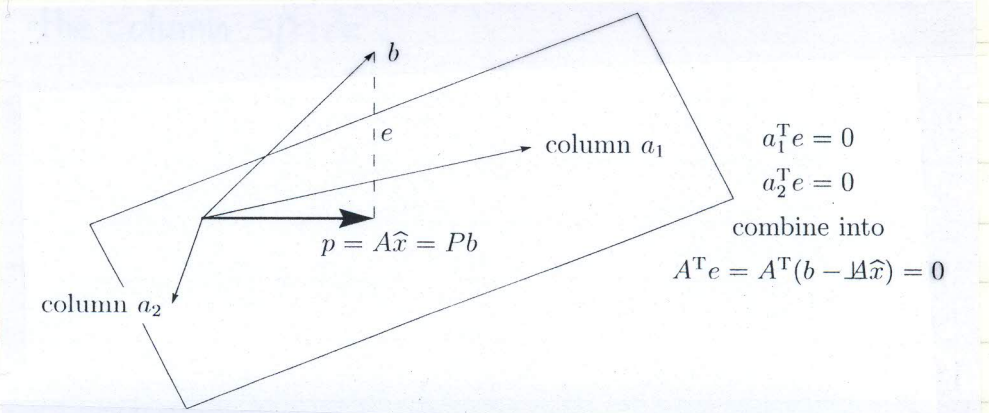


Figure 3.8 Projection onto the column space of a 3 by 2 matrix.

① All vectors perpendicular to the column space ITE in the left nullspace. Thus the error $e = b - A\hat{x}$ must be in the nullspace of A^T:

$$A^T(b - A\hat{x}) = 0 \quad \text{or} \quad A^TA\hat{x} = A^Tb.$$

② The ~~error~~ vector must be perpendicular to each column $a_1, \cdots, a_n$ of A:

$$\begin{cases} a_1^T(b - A\hat{x}) = 0 \\ \vdots \\ a_n^T(b - A\hat{x}) = 0 \end{cases} \quad \text{or} \quad \begin{bmatrix} a_1^T \\ \vdots \\ a_n^T \end{bmatrix} \begin{bmatrix} b - A\hat{x} \end{bmatrix} = 0.$$

This is again $A^T(b - A\hat{x}) = 0$ and $A^TA\hat{x} = A^Tb$.

0. Taking partial derivatives of $E^2 = (Ax - b)^T(Ax - b)$ gives the same $2A^TAx - 2A^Tb = 0$.
↳ symmetric square matrix

3L When $Ax = b$ is inconsistent, its least-squares solution minimizes $\|Ax - b\|^2$:

Normal equations: $A^TA\hat{x} = A^Tb$.

A^TA is invertible exactly when the columns of A are linearly independent! Then,

Best estimate $\hat{x}$: $\hat{x} = (A^TA)^{-1}A^Tb$.

The projection of b onto the column space is the nearest point $A\hat{x}$:

Projection: $\hat{p} = A\hat{x} = A(A^TA)^{-1}A^Tb$.

Ex: $A = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 0 & 0 \end{bmatrix}$, $b = \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix}$. $Ax = b$ has no solution

$A^TA\hat{x} = A^Tb$ gives the best $\hat{x}$.

The projection of $b = (4,5,6)$ is $p = (4,5,0)$.

Solving the normal equations:

$$A^TA = \begin{bmatrix} 1 & 1 & 0 \\ 2 & 3 & 0 \end{bmatrix}\begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 5 \\ 5 & 13 \end{bmatrix}$$

$$\hat{x} = (A^TA)^{-1}A^Tb = \begin{bmatrix} 13 & -5 \\ -5 & 2 \end{bmatrix}\begin{bmatrix} 1 & 1 & 0 \\ 2 & 3 & 0 \end{bmatrix}\begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}.$$

Projection $p = A\hat{x} = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 4 \\ 5 \\ 0 \end{bmatrix}$.

Remarks: $\rightarrow b = Ax$ (a combination of the columns)

① $b \in C(A) \Rightarrow p = A(A^TA)^{-1}A^TAx = Ax = b$.

$\therefore$ The projection of b is still b.

② $b \in N(A^T) \Rightarrow$ ($N(A^T)$: left nullspace) $p = A(A^TA)^{-1}A^Tb = A(A^TA)^{-1}0 = 0$ ($A^Tb = 0$)

$\therefore$ b projects to the zero vector

③ When A is square and invertible, $C(A) = \mathbb{R}^n$ (the whole space).

A: invertible $\Rightarrow p = A(A^TA)^{-1}A^Tb$

$= A \cdot A^{-1}(A^T)^{-1}A^Tb = b$.

$\therefore$ Every vector projects to itself, $p = b$, $\hat{x} = x$.

④ A has only one column $\Rightarrow A^TA = a^Ta$ and $\hat{x} = a^Tb/a^Ta$.

The Cross-Product Matrix A^TA

- A^TA has the same nullspace as A.

  ① $Ax = 0 \Rightarrow A^TAx = 0$

  ② $A^TAx = 0 \Rightarrow x^TA^TAx = 0$, $\|Ax\|^2 = 0$, $Ax = 0$.

  → the same is true for AA^T

3M If A has independent columns, then A^TA is square, symmetric, and invertible.

Projection Matrices

(We assume the independence of the columns of A in what follows)

$P = A(A^TA)^{-1}A^T$: projection matrix

o. $p = Pb$ is the component of b in the column space, and the error $e = b - Pb = (I-P)b$ is in $N(A^T)$, the orthogonal complement of $C(A)$.
$(I-P)$ is also a projection matrix!
$\therefore$ Pb and $(I-P)b$ are two perpendicular components of b.

3N The projection matrix $P = A(A^TA)^{-1}A^T$ has two basic properties: (i) $P^2 = P$ and (ii) $P^T = P$. Conversely, any symmetric matrix with $P^2 = P$ represents a projection.

<pf>

o. $P^2 = A(A^TA)^{-1}A^TA(A^TA)^{-1}A^T = A(A^TA)^{-1}A^T = P$

$P^T = (A^T)^T((A^TA)^{-1})^TA^T = A((A^TA)^T)^{-1}A^T$
$= A(A^TA)^{-1}A^T = P$.

o. For the converse, from $P^2 = P$ and $P^T = P$, we can show the error vector $b - Pb$ is orthogonal to the column space. For any vector $Pc \in C(P)$,

$(b-Pb)^TPc = b^T(I-P)^TPc = b^T(P-P^2)c = 0$.

Thus, $b - Pb$ is orthogonal to the space, and Pb is the projection onto the column space. ∎

Ex1: A: an invertible 4×4 matrix. $\Rightarrow C(A) = \mathbb{R}^4$.

$P = A(A^TA)^{-1}A^T = AA^{-1}(A^T)^{-1}A^T = I$.

I: symmetric, $I^2 = I$, $b - Ib = 0$.

The projection onto the whole space is the identity matrix.

Least-Squares Fitting of Data

In a series of experiments, we expect the output b to be a linear function of the input t. We look for a straight line: $b = C + Dt$, with two unknowns C, D.

$$\begin{cases} C + Dt_1 = b_1 \\ C + Dt_2 = b_2 \\ \vdots \\ C + Dt_m = b_m \end{cases} \quad \text{or} \quad \underbrace{\begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_m \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} C \\ D \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}}_{b}$$

The best solution $(\hat{C}, \hat{D})$ is the $\hat{x}$ that minimizes $E^2 = \|b - Ax\|^2 = (b_1 - C - Dt_1)^2 + \cdots + (b_m - C - Dt_m)^2$.

The vector $p = A\hat{x}$ is as close as possible to b. Of all straight lines $b = C + Dt$, we choose the one that best fits the data. On the graph, the errors are the vertical distances $b - C - Dt$ to the straight line (not perpendicular distances!)

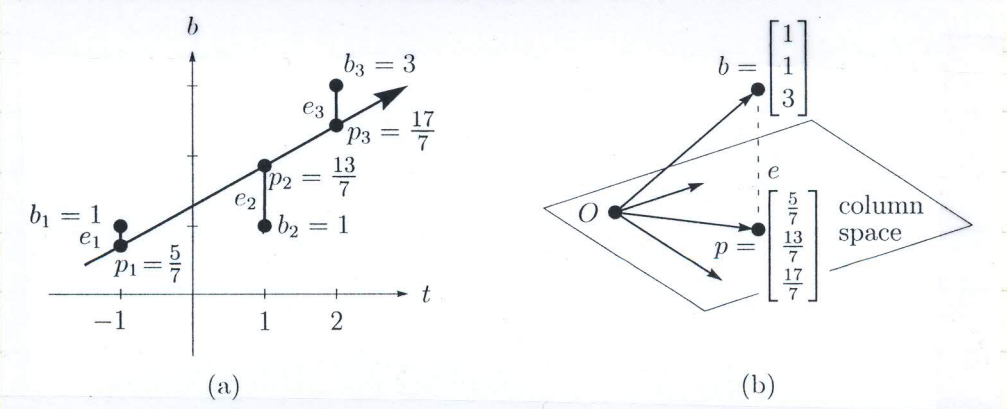


Figure 3.9 Straight-line approximation matches the projection p of b.

Ex 2: Three measurements:

$b_1=1$ at $t_1=-1$; $b_2=1$ at $t_2=1$; $b_3=3$ at $t_3=2$

$$A\mathbf{x}=\mathbf{b} \text{ is } \begin{cases} C-D=1 \\ C+D=1 \\ C+2D=3 \end{cases} \text{ or } \begin{bmatrix} 1 & -1 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}\begin{bmatrix} C \\ D \end{bmatrix}=\begin{bmatrix} 1 \\ 1 \\ 3 \end{bmatrix}$$

$$A^TA\mathbf{x}=A^T\mathbf{b} \text{ is } \begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix}\begin{bmatrix} C \\ D \end{bmatrix}=\begin{bmatrix} 5 \\ 6 \end{bmatrix}$$

$\therefore \hat{C}=\frac{9}{7}$ and $\hat{D}=\frac{4}{7}$ and the best line is $\frac{9}{7}+\frac{4}{7}t$

30 The measurements $b_1,\cdots,b_m$ are given at distinct points $t_1,\cdots,t_m$. Then the straight line $\hat{C}+\hat{D}t$ which minimizes E^2 comes from least squares:

$$A^TA\begin{bmatrix} \hat{C} \\ \hat{D} \end{bmatrix}=A^T\mathbf{b} \text{ or } \begin{bmatrix} m & \Sigma t_i \\ \Sigma t_i & \Sigma t_i^2 \end{bmatrix}\begin{bmatrix} \hat{C} \\ \hat{D} \end{bmatrix}=\begin{bmatrix} \Sigma b_i \\ \Sigma t_i b_i \end{bmatrix}$$

Remark: Given a mixture of two radioactive chemicals with known half-lives λ and μ, we want to know their unknown amounts C and D: $b=Ce^{-\lambda t}+De^{-\mu t}$.

$$A\mathbf{x}=\mathbf{b} \text{ is } \begin{cases} Ce^{-\lambda t_1}+De^{-\mu t_1}\approx b_1 \\ \vdots \\ Ce^{-\lambda t_m}+De^{-\mu t_m}\approx b_m . \end{cases}$$

The least-squares principle will give optimal $\hat{C}$ and $\hat{D}$.

o. But, if we knew the amounts C and D, and were trying to discover the decay rates λ and μ, this is a problem in nonlinear least squares. Setting the derivatives of E^2 to zero will give nonlinear equations for the optimal λ and μ.

Weighted Least Squares

The estimate $\hat{x}$ of weight from two observations $x = b_1$ and $x = b_2$:

$$\begin{bmatrix} 1 \\ 1 \end{bmatrix} [x] = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \Rightarrow A^T A \hat{x} = A^T b, \quad 2\hat{x} = b_1 + b_2$$

Weighted error: $E^2 = w_1^2 (x - b_1)^2 + w_2^2 (x - b_2)^2$

$$\frac{dE^2}{dx} = 2\left[w_1^2 (x - b_1) + w_2^2 (x - b_2)\right] = 0 \text{ at } \hat{x}_w = \frac{w_1^2 b_1 + w_2^2 b_2}{w_1^2 + w_2^2}$$

The least squares solution to $WA\hat{x} = Wb$ is $\hat{x}_w$

Weighted Normal Equation: $(A^T W^T W A)\hat{x}_w = A^T W^T W b$.

- The projection $A\hat{x}_w$ is still the point in the column space that is closest to b under a new meaning of closeness. The perpendicularity test involves $(Wy)^T(Wx) = 0$ instead of $y^T x = 0$. The matrix $W^T W$ appears in the middle. In this new sense, the projection $A\hat{x}_w$ and the error $b - A\hat{x}_w$ are again perpendicular.
- The inner product of x and y is generalized to $y^T C x$, where $C = W^T W$ is a symmetric matrix. For an orthogonal matrix $W = Q$, $C = Q^T Q = I$ and the inner product is not new.

For any invertible matrix W, these rules define a new inner product and length:

Weighted by W: $(x, y)_W = (Wy)^T (Wx)$ and $\|x\|_W = \|Wx\|$.

o. If the errors in the b_i are independent of each other, and their variances are σ_i^2, then the right weights are $w_i = 1/\sigma_i$. A more accurate measurement, with a smaller variance, gets a heavier weight.

o. The observations may not be independent. Then W has off-diagonal terms. The best unbiased matrix $C = W^T W$ is the inverse of the covariance matrix — whose i,j entry is the expected value of (error in b_i) times (error in b_j). The main diagonal of C^{-1} contains the variances σ_i^2; → the average of (error in b_i)^2.

Ex 3:

Two bridge partners both guess (after the bidding) the total number of spades they hold. For each guess, the errors $-1, 0, 1$ might have equal probability $\frac{1}{3}$. Then the expected error is zero and the variance is $\frac{2}{3}$.

$$E(e) = \tfrac{1}{3}(-1) + \tfrac{1}{3}(0) + \tfrac{1}{3}(1) = 0$$
$$E(e^2) = \tfrac{1}{3}(-1)^2 + \tfrac{1}{3}(0)^2 + \tfrac{1}{3}(1)^2 = \tfrac{2}{3}.$$

The two guesses are dependent but not identical. Say the chance that they are both too high or both too low is zero, but the chance of opposite errors is $\frac{1}{3}$. Then $E(e_1 e_2) = \frac{1}{3}(-1)$, and the inverse of the covariance matrix is $W^T W$:

$$\begin{bmatrix} E(e_1^2) & E(e_1 e_2) \\ E(e_1 e_2) & E(e_2^2) \end{bmatrix}^{-1} = \begin{bmatrix} \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} \end{bmatrix}^{-1} = C = W^T W.$$

This matrix goes into the middle of the weighted normal equations.

3.4 Orthogonal Bases and Gram-Schmidt

3P The vectors $q_1, \cdots, q_n$ are orthonormal if

$$q_i^T q_j = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$$

A matrix with orthonormal columns will be called Q.

The standard basis $e_1, \cdots, e_n$ consists of the columns of I. We can rotate these vectors to form other orthonormal bases. A subspace of $\mathbb{R}^n$ may not contain the standard vectors e_i, but an orthonormal basis can be constructed for the space. This construction is known as Gram-Schmidt orthogonalization

Orthogonal Matrices

3Q If Q has orthonormal columns, then $Q^TQ = I$.
An orthogonal matrix is a square matrix with orthonormal columns. Then $Q^T = Q^{-1}$. (Note that $Q^TQ = I$ even if Q is rectangular. But then Q^T is only a left-inverse.)

Ex1: $$Q = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \Rightarrow Q^T = Q^{-1} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$$

Ex2: $$P = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \Rightarrow P^{-1} = P^T = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

- A reflection also forms an orthogonal matrix. Geometrically, an orthogonal Q is either a rotation or the product of a rotation and a reflection.

3R Multiplication by any Q preserves lengths:
$\|Qx\| = \|x\|$ for every vector x.
It also preserves inner products and angles, since
$(Qx)^T(Qy) = x^T Q^T Q y = x^T y$.

o. Write b as a combination $b = x_1 q_1 + x_2 q_2 + \cdots + x_n q_n$.
$q_i^T b = x_i q_i^T q_i \Rightarrow x_i = q_i^T b$
$\therefore\ b = (q_1^T b) q_1 + (q_2^T b) q_2 + \cdots + (q_n^T b) q_n$
o. $x_1 q_1 + \cdots + x_n q_n = b \Rightarrow Qx = b \Rightarrow x = Q^T b = Q^T b$.
o. $\|b\|^2 = (q_1^T b)^2 + (q_2^T b)^2 + \cdots + (q_n^T b)^2 = \|Q^T b\|^2$.

Remarks: Since $Q^T = Q^{-1}$, we have $Q \cdot Q^T = I$. The rows of a square matrix are orthonormal whenever the columns are.

Rectangular Matrices with Orthonormal Columns

3S If Q has orthonormal columns, the least-square problem is (easy)
$Qx = b$: rectangular system with no solution for most b.
$Q^T Q \hat{x} = Q^T b$: normal equation for the best $\hat{x}$ – in which $Q^T Q = I$.
$\hat{x} = Q^T b$: $\hat{x}_i = q_i^T b$
$p = Q\hat{x}$: the projection of b is $(q_1^T b) q_1 + \cdots + (q_n^T b) q_n$
$p = QQ^T b$: the projection matrix is $P = QQ^T$

Ex3: Projection of $b = (x, y, z)$ onto the xy-plane is $p = (x, y, 0)$

$q_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$ and $(q_1^T b) q_1 = \begin{bmatrix} x \\ 0 \\ 0 \end{bmatrix}$; $q_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$ and $(q_2^T b) q_2 = \begin{bmatrix} 0 \\ y \\ 0 \end{bmatrix}$

$\Rightarrow\ P = q_1 q_1^T + q_2 q_2^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ and $P \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} x \\ y \\ 0 \end{bmatrix}$

Ex4: When the measurement times average to zero, fitting a straight line leads to orthogonal columns.

Take $t_1=-3$, $t_2=0$, $t_3=3$. The attempt to fit $y=C+Dt$ leads to

$$\begin{cases} C+Dt_1=y_1 \\ C+Dt_2=y_2 \\ C+Dt_3=y_3 \end{cases} \quad \text{or} \quad \begin{bmatrix} 1 & -3 \\ 1 & 0 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} C \\ D \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

We can project y separately onto each column:

$$\hat{C} = \frac{[1\ 1\ 1][y_1\ y_2\ y_3]^T}{1^2+1^2+1^2}, \quad \hat{D} = \frac{[-3\ 0\ 3][y_1\ y_2\ y_3]^T}{(-3)^2+0^2+3^2}$$

$\hat{C}=(y_1+y_2+y_3)/3$ is the best fit by a horizontal line, whereas $\hat{D}t$ is the best fit by a straight line through the origin. The sum $\hat{C}+\hat{D}t$ is the best fit by any straight line.

If the average $\bar{t}=(t_1+\cdots+t_m)/m$ is not zero, then the time can be shifted by $\bar{t}$. Let $y=C+Dt=c+d(t-\bar{t})$.

$$\hat{c} = \frac{[1\ \cdots\ 1][y_1\ \cdots\ y_m]^T}{1^2+1^2+\cdots+1^2} = \frac{y_1+\cdots+y_m}{m}$$

$$\hat{d} = \frac{[(t_1-\bar{t})\ \cdots\ (t_m-\bar{t})][y_1\ \cdots\ y_m]^T}{(t_1-\bar{t})^2+\cdots+(t_m-\bar{t})^2} = \frac{\Sigma(t_i-\bar{t})y_i}{\Sigma(t_i-\bar{t})^2}$$

The best $\hat{c}$ is the mean and we also get a convenient formula for $\hat{d}$. The earlier A^TA had the off-diagonal entries Σt_i, and shifting the time by $\bar{t}$ made these entries zero. The shift is an example of the Gram-Schmidt process.

The Gram-Schmidt Process

Given three independent vectors a, b, c, let $q_1 = a/\|a\|$.
Second vector $B = b - (q_1^T b) q_1$ and $q_2 = B/\|B\|$.
Third vector $C = c - (q_1^T c) q_1 - (q_2^T c) q_2$ and $q_3 = C/\|C\|$.

Ex 5: Gram-Schmidt

$$a = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad c = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} \Rightarrow q_1 = a/\sqrt{2} = \begin{bmatrix} 1/\sqrt{2} \\ 0 \\ 1/\sqrt{2} \end{bmatrix}$$

$$B = b - (q_1^T b) q_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} - \frac{1}{\sqrt{2}} \begin{bmatrix} 1/\sqrt{2} \\ 0 \\ 1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} 1/2 \\ 0 \\ -1/2 \end{bmatrix} \Rightarrow q_2 = \begin{bmatrix} 1/\sqrt{2} \\ 0 \\ -1/\sqrt{2} \end{bmatrix}$$

$$C = c - (q_1^T c) q_1 - (q_2^T c) q_2 = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} - \sqrt{2} \begin{bmatrix} 1/\sqrt{2} \\ 0 \\ 1/\sqrt{2} \end{bmatrix} - \sqrt{2} \begin{bmatrix} 1/\sqrt{2} \\ 0 \\ -1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = q_3$$

Orthonormal basis $Q = [q_1 \; q_2 \; q_3] = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 0 & 0 & 1 \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 \end{bmatrix}$

3T The Gram-Schmidt process starts with independent vectors $a_1, \cdots, a_n$, and ends with orthonormal vectors $q_1, \cdots, q_n$. At step j, it subtracts from a_j its components in the directions $q_1, \cdots, q_{j-1}$:

$$A_j = a_j - (q_1^T a_j) q_1 - \cdots - (q_{j-1}^T a_j) q_{j-1}.$$

Then q_j is the unit vector $A_j/\|A_j\|$.

Remark on the calculation

It is easier to compute the orthogonal a, B, C without forcing their lengths to equal one. Then square roots enter only at the end, when dividing by those lengths.

The Factorization $A = QR$

$$a = (q_1^T a)\cdot q_1$$
$$b = (q_1^T b)\cdot q_1 + (q_2^T b) q_2$$
$$c = (q_1^T c)\cdot q_1 + (q_2^T c) q_2 + (q_3^T \cdot c)\cdot q_3$$

$$QR \text{ factors}: \quad A = \begin{bmatrix} a & b & c \end{bmatrix} = \begin{bmatrix} q_1 & q_2 & q_3 \end{bmatrix} \begin{bmatrix} q_1^T a & q_1^T b & q_1^T c \\ & q_2^T b & q_2^T c \\ & & q_3^T c \end{bmatrix} = QR$$

Ex:
$$A = \begin{bmatrix} 1 & 1 & 2 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 0 & 0 & 1 \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 \end{bmatrix} \begin{bmatrix} \sqrt{2} & 1/\sqrt{2} & \sqrt{2} \\ & 1/\sqrt{2} & \sqrt{2} \\ & & 1 \end{bmatrix} = QR$$

The length of a, B, C

34 Every $m \times n$ matrix with independent columns can be factored into $A = QR$. The columns of Q are orthonormal, and R is upper triangular and invertible. When $m = n$ and all matrices are square, Q becomes an orthogonal matrix.

The orthogonalization simplifies the least-squares problem $Ax = b$. The normal equations become easier since

$$A^TA = R^TQ^TQR = R^TR.$$

The fundamental equation $A^TA\hat{x} = A^Tb$ simplifies to a triangular system

$$R^TR\hat{x} = R^TQ^Tb \quad \text{or} \quad R\hat{x} = Q^Tb.$$

Instead of solving $QRx = b$, we solve $R\hat{x} = Q^Tb$ by back-substitution. The real cost is the mn^2 operations of Gram-Schmidt, which are needed to find Q and R.

Function Spaces

① Lengths and Inner Products.

$$\|f\|^2 = \int_0^{2\pi} (f(x))^2 dx, \quad (f,g) = \int_0^{2\pi} f(x)g(x)dx$$

The Schwartz inequality is still satisfied:

$$|(f,g)| \leq \|f\| \cdot \|g\|$$

② Gram-Schmidt for Functions.

There is no interval $[a,b]$ on which $(1,x^2) = \int_a^b x^2 dx = 0$.
Therefore the closest parabola to $f(x)$ is not the sum of its projections onto $1, x, x^2$. On the interval $[0,1]$,

$$A^TA = \begin{bmatrix} (1,1) & (1,x) & (1,x^2) \\ (x,1) & (x,x) & (x,x^2) \\ (x^2,1) & (x^2,x) & (x^2,x^2) \end{bmatrix} = \begin{bmatrix} \int 1 & \int x & \int x^2 \\ \int x & \int x^2 & \int x^3 \\ \int x^2 & \int x^3 & \int x^4 \end{bmatrix} = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{bmatrix}$$

This is the ill-conditioned Hilbert matrix with a large inverse.
The situation becomes impossible if we add a few more axes.
It is virtually hopeless to solve $A^TA\hat{x} = A^Tb$ for the closest polynomial of degree 10. More precisely, it is hopeless to solve this by Gauss elimination. Every roundoff error would be amplified by more than 10^{13}.

The right idea is to switch to orthogonal axes (by Gram-Schmidt). On the interval $[-1,1]$, we have

$$(1,x) = \int_{-1}^{1} x\,dx = 0, \quad (x,x^2) = \int_{-1}^{1} x^3 dx = 0.$$

Starting with $v_1 = 1$, $v_2 = x$,

Orthogonalize $v_3 = x^2 - \frac{(1,x^2)}{(1,1)}1 - \frac{(x,x^2)}{(x,x)}x = x^2 - \frac{1}{3}$ ((x,x^2) → 0)

$1, x, x^2 - \frac{1}{3}$: Legendre polynomials

The closest polynomial is now computable by projecting onto each of the Legendre polynomials.