Definition 7.4.7. *If A is an $m \times n$ matrix, then $A^T A$ is the **Gram matrix** or **Gramian** of A.*

Theorem 7.4.8. *If A is an $m \times n$ matrix then:*

$\left.\begin{array}{l} \text{(a) } \operatorname{null}(A^T A) = \operatorname{null}(A) \\ \text{(b) } \operatorname{rank}(A^T A) = \operatorname{rank}(A) \end{array}\right\}$ Prob. Set

(c) *$A^T A$ is invertible iff A has linearly independent columns*

Proof. Let $A = [\underset{\sim}{a}_1 \cdots \underset{\sim}{a}_n]$

First note $A^T A$ is $n \times n$. Thm 1.5.7 implies

$A^T A$ is invertible $\iff$ $A^T A \underset{\sim}{x} = \underset{\sim}{0}$ only for $\underset{\sim}{x} = \underset{\sim}{0}$

$\iff \operatorname{null}(A^T A) = \{\underset{\sim}{0}\}$

$\iff \operatorname{null}(A) = \{\underset{\sim}{0}\}$ (Part a)

$\iff \sum_{i=1}^{n} x_i \underset{\sim}{a}_i = \underset{\sim}{0}$ only $\underset{\sim}{x} = \underset{\sim}{0}$

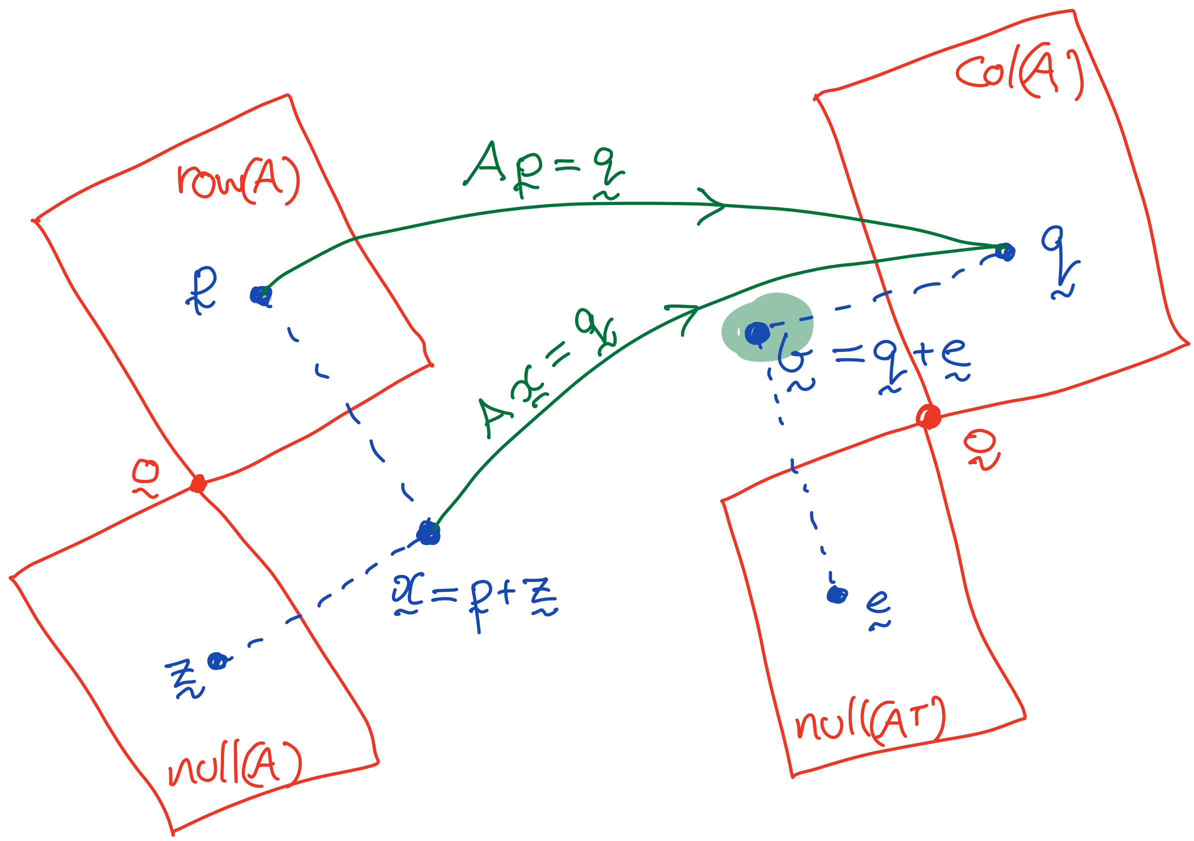
$\iff$ the columns of A are L.I.

$\square$

$\square$

$A \in M_{mn}$

$\mathbb{R}^n = \text{row}(A) \oplus \text{null}(A)$ $\qquad$ $\mathbb{R}^m = \text{col}(A) \oplus \text{null}(A^T)$



$A\underset{\sim}{x} = \underset{\sim}{b}$ has no sol$^{\underline{n}}$s when $\underset{\sim}{e} \neq \underset{\sim}{0}$

But $\underset{\sim}{x} = \underset{\sim}{p} + \underset{\sim}{z}$ is a LS sol$^{\underline{n}}$ to $A\underset{\sim}{x} = \underset{\sim}{b}$

for all $\underset{\sim}{z} \in \text{null}(A)$, for <u>any</u> $\underset{\sim}{e} \in \text{null}(A^T)$

7.5 Application: Data fitting (A&R §6.5)

Suppose we are interested in how a quantity y depends upon another quantity x, and suppose we have a theoretical hunch that y should depend linearly on x, so $y = a + b\,x$ for some choice of a and b. Suppose further that we perform an experiment and obtain n pairs of values $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. We would like to use our data to determine a and b.

If, as would be ideal, our data did all lie on a single line we would have

$$\begin{aligned} y_1 &= a + b\,x_1 \\ y_2 &= a + b\,x_2 \\ \vdots \quad & \quad \vdots \\ y_n &= a + b\,x_n \end{aligned}$$

This can be expressed more succinctly in matrix form as $\mathbf{y} = M\mathbf{v}$, where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \qquad M = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \qquad \mathbf{v} = \begin{bmatrix} a \\ b \end{bmatrix}.$$

Because of experimental error however, we cannot expect the data to fall perfectly on a straight line. Therefore, $M\mathbf{v} = \mathbf{y}$ will generally have no exact solution, and we will instead be interested in obtaining a least squares solution, $\mathbf{v} = \mathbf{v}_*$, which is a choice of (a, b) that minimizes $\|\mathbf{y} - M\mathbf{v}\|$.

Theorem 7.4.5 implies that our task is to solve $M^T M\mathbf{v} = M^T\mathbf{y}$.

In fact, it can be shown that the column vectors of M are linearly independent iff the n data points do not all lie on a vertical line in the xy-plane. Theorem 7.4.8 then implies that under this (very reasonable) assumption on the data, the least squares problem has a unique solution, $\mathbf{v} = (M^T M)^{-1} M^T \mathbf{y}$.

Since we are fitting to a linear function, $M^T M$ will always be a 2×2 matrix in this case, so that $(M^T M)^{-1}$ is very straightforward to calculate.

If we express the squared error $\|\mathbf{y} - M\mathbf{v}\|^2$ in terms of its components we obtain

$$\|\mathbf{y} - M\mathbf{v}\|^2 = d_1^2 + d_2^2 + \ldots + d_n^2,$$

where $d_i = y_i - a - bx_i$, and a and b are the components of the least squares solution for $\mathbf{v}$.

The quantity d_i is often called a **residual**, and gives the difference between the value of y predicted by the model and the value found experimentally in the ith datum.

E.g. 7.5.1. *Find the line of best fit for the data:*

$$(0,2),(1,3),(3,7),(6,12).$$

$$\underset{\sim}{y} = \begin{pmatrix} 2 \\ 3 \\ 7 \\ 12 \end{pmatrix}, \quad M = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 3 \\ 1 & 6 \end{pmatrix}$$

$$\Rightarrow M^T M = \begin{pmatrix} 4 & 10 \\ 10 & 46 \end{pmatrix} \quad \& \quad M^T \underset{\sim}{y} = \begin{pmatrix} 24 \\ 96 \end{pmatrix}$$

So solving $M^T M \underline{v} = M^T \underset{\sim}{\underline{y}} \Rightarrow$

$$v_2 = 12/7, \quad v_1 = 12/7$$

so line of best fit is

$$y = \frac{12}{7} + \frac{12}{7}x$$

The technique we have discussed for least squares fitting to linear functions can be easily generalized to any polynomial. Suppose we again have n data points $(x_1, y_1), \ldots, (x_n, y_n)$ and suppose we believe these data to be described by a polynomial of degree m. The corresponding system is

$$\begin{aligned} y_1 &= a_0 + a_1\, x_1 + \ldots + a_m\, x_1^m \\ y_2 &= a_0 + a_1\, x_2 + \ldots + a_m\, x_2^m \\ \vdots & \qquad\qquad\qquad\qquad \vdots \\ y_n &= a_0 + a_1\, x_n + \ldots + a_m\, x_n^m \end{aligned}$$

which can be expressed in matrix form as $\mathbf{y} = M\mathbf{v}$ with

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \qquad M = \begin{bmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^m \\ 1 & x_2 & x_2^2 & \ldots & x_2^m \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^m \end{bmatrix}, \qquad \mathbf{v} = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix}.$$

E.g. 7.5.2. *Find the quadratic polynomial that best fits the four points*

$$(2,0),(3,-10),(5,-48),(6,-76).$$

$$\underset{\sim}{y} = \begin{pmatrix} 0 \\ -10 \\ -48 \\ -76 \end{pmatrix}, \quad M = \begin{pmatrix} 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 5 & 25 \\ 1 & 6 & 36 \end{pmatrix}, \quad \underset{\sim}{v} = \begin{pmatrix} v_0 \\ v_1 \\ v_2 \end{pmatrix}$$

Solving $M^T M \underline{v} = M^T \underline{y} \implies \underset{\sim}{v} = \begin{pmatrix} 2 \\ 5 \\ -3 \end{pmatrix}$

So best fit quadratic is

$$y = 2 + 5x - 3x^2.$$

7.6 Application: Linear Regression

A common problem in statistics is the following. We suspect that the mean y of a random quantity can be described by a polynomial in another quantity x,

$$y = a_0 + a_1 x + \ldots + a_k x^k.$$

This curve is called a **linear regression** curve in statistics.

We are given experimental data (x_1, y_1, s_1), ..., (x_n, y_n, s_n), and it is believed that

$$y_i = a_0 + a_1 x_i + \ldots + a_k x_i^k + \epsilon_i \tag{7.6}$$

where ϵ_i has mean 0 and standard deviation s_i. We wish to estimate the parameters a_i.

If the s_i are known to be all equal, then statistical considerations imply that the best estimate for the a_i is simply the least-squares fit of the curve (7.6) to the pairs (x_1, y_1), (x_2, y_2), ..., (x_n, y_n), as described in Section 7.5.

Often, however, the s_i will not all be the same. E.g. if the data is generated from Monte Carlo simulations the s_i will typically be distinct. In this more general case, statistical considerations suggest that rather than minimizing the distance between $\mathbf{y}$ and $M\mathbf{v}$ with respect to the dot product, it is more appropriate to consider a weighted inner product. In particular, we define the diagonal matrix S whose diagonal entries are $1/s_i^2$, and introduce the inner product

$$\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^T S \mathbf{v}.$$

Problem 7.6.1 (Weighted least squares)**.** *Let* $w_1, w_2, \ldots, w_m > 0$ *and consider the corresponding weighted inner product on* $\mathbb{R}^m$*, as defined in E.g. 7.1.5. Given a linear system* $A\mathbf{x} = \mathbf{b}$ *of* m *equations in* n *unknowns, find a vector* $\mathbf{x}$ *that minimizes* $\|\mathbf{b} - A\mathbf{x}\|$ *with respect to this weighted inner product on* $\mathbb{R}^m$*. We call such an* $\mathbf{x}$ *a* ***weighted least squares solution*** *of the system, we call* $\mathbf{b} - A\mathbf{x}$ *the* ***weighted least squares error vector*** *and* $\|\mathbf{b} - A\mathbf{x}\|$ *the* ***weighted least squares error****.*

Theorem 7.6.2. *Let* $w_1, \ldots, w_m > 0$ *and consider the corresponding weighted inner product. A vector* $\mathbf{x}$ *is a weighted least squares solution to* $A\mathbf{x} = \mathbf{b}$ *iff it is a solution of the associated* ***weighted normal system***

$$A^T W A\mathbf{x} = A^T W \mathbf{b}, \tag{7.7}$$

where W *is the diagonal matrix with* ii *entry* w_i*. Moreover, the weighted normal system is always consistent.*

Remark 7.6.3. *In fact, Theorem 7.6.2 can be generalized further by replacing the diagonal matrix* W *with any matrix* C *for which* $\mathbf{x}^T C\mathbf{x} > 0$ *for all nonzero* $\mathbf{x}$*. In the statistical context, the off diagonal entries in* C *are related to correlations between the data points.*

To fit a polynomial (regression curve)

$$y = a_0 + a_1 x + \ldots + a_k x^k$$

to given experimental data $(x_1, y_1, s_1), \ldots, (x_n, y_n, s_n)$ we define

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, M = \begin{bmatrix} 1 & x_1 & \ldots & x_1^m \\ 1 & x_2 & \ldots & x_2^m \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & \ldots & x_n^m \end{bmatrix}, S = \begin{bmatrix} s_1^{-2} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & s_m^{-2} \end{bmatrix}, \mathbf{v} = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix}$$

and solve $M^T S M\mathbf{v} = M^T S\mathbf{y}$ for $\mathbf{v}$.

Exercise 7.6.4. *We again consider the experimental data of E.g. 7.5.1, but this time the experimenter has also provided us with the standard deviation for each measurement.*

x	y	s
0	2	0.25
1	3	0.5
3	7	1
6	12	1

Use weighted least squares to fit the data to a line.

$$M = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 3 \\ 1 & 6 \end{pmatrix}, \quad S = \begin{pmatrix} 16 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \underset{\sim}{y} = \begin{pmatrix} 2 \\ 3 \\ 7 \\ 12 \end{pmatrix}$$

Solve $M^T S M \underset{\sim}{v} = M^T S \underset{\sim}{y} \implies \underset{\sim}{y} = \frac{574}{303} + \frac{497}{303} x$

8 Orthogonalization

8.1 Orthogonal matrices (A&R §7.1)

Definition 8.1.1. *A square matrix A is said to be* ***orthogonal*** *if*

$$A^{-1} = A^T.$$

E.g. 8.1.2. *Let*

$$A = \frac{1}{3}\begin{bmatrix} 1 & 2 & 2 \\ 2 & -2 & 1 \\ -2 & -1 & 2 \end{bmatrix}$$

Show that A is orthogonal.

$AA^T = I \implies A$ is orthogonal.

E.g. 8.1.3. *Let*

$$R_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}.$$

Show that R_θ is orthogonal.

$$R_\theta R_\theta^T = \begin{bmatrix} \cos^2\theta + \sin^2\theta & 0 \\ 0 & \cos^2\theta + \sin^2\theta \end{bmatrix} = I$$

$\implies R_\theta$ is orthogonal.

Theorem 8.1.4. *Let A be an $n \times n$ matrix. The following are equivalent statements.*

(a) A is orthogonal

(b) The row vectors of A form an orthonormal basis for $\mathbb{R}^n$ with respect to the dot product

(c) The column vectors of A form an orthonormal basis for $\mathbb{R}^n$ with respect to the dot product

Proof. Let $A \in M_{nn}$ & let A have rows $\underset{\sim}{r}_1, \dots, \underset{\sim}{r}_n$. If $1 \le i, j \le n$ then

$$(AA^T)_{ij} = \sum_{k=1}^{n} (A)_{ik} (A^T)_{kj} = \sum_{k=1}^{n} (A)_{ik} (A)_{jk}$$

$$= \sum_{k=1}^{n} (\underset{\sim}{r}_i)_k (\underset{\sim}{r}_j)_k = \underset{\sim}{r}_i \cdot \underset{\sim}{r}_j$$

So: A is orthogonal iff $AA^T = I$

$\Leftrightarrow (AA^T)_{ij} = \begin{cases} 1 & i=j \\ 0 & i \neq j \end{cases} \Leftrightarrow \underset{\sim}{r}_i \cdot \underset{\sim}{r}_j = \begin{cases} 1 & i=j \\ 0 & i \neq j \end{cases}$

$\Leftrightarrow$ the rows of A form an orthonormal set (wrt dot product). So (a) $\Leftrightarrow$ (b)

- Now observe that A is orthogonal $\Leftrightarrow$

$A^{-1} = A^T \Leftrightarrow (A^T)^{-1} = A \Leftrightarrow (A^T)^{-1} = (A^T)^T$

$\Leftrightarrow A^T$ is orthogonal $\Leftrightarrow$ rows of A^T form an orthonormal basis

$\Leftrightarrow$ columns of A form an orthonormal basis

So (a) $\Leftrightarrow$ (c). $\square$