

# Seminario di calcolo scientifico: Applicazioni nella geodesia del problema ai minimi quadrati

Mele Giampaolo

May 8, 2011

## Abstract

Il problema ai minimi quadrati è alla base di molte applicazioni, in questo seminario tratterò soprattutto l'applicazione al problema del fitting di una curva o superficie e in modo approfondito nel caso di un'ellissoide. In questo contesto è immediata l'applicazione nel campo della geodesia dato che la terra può esser approssimata con un ellissoide e il problema diventa quello di trovare l'ellissoide che meglio la approssima. Mostrerò brevemente i primi tentativi storici di risoluzione di tali problemi quali quello di Laplace e quello di Gauss. Dopo il breve excursus storico enuncerò e dimostrerò alcuni teoremi che saranno utili per risolvere il problema ai minimi quadrati. Successivamente mostrerò il problema ai minimi quadrati vincolato con un'altra applicazione alla geodesia (correzione minima di misure). Infine mostrerò il problema totale ai minimi quadrati nella sua interpretazione geometrica e algebrica mostrando un metodo per risolverlo e presenterò il problema aperto del fitting di varietà affini (problema ai minimi quadrati non lineare).

## 1 Introduzione

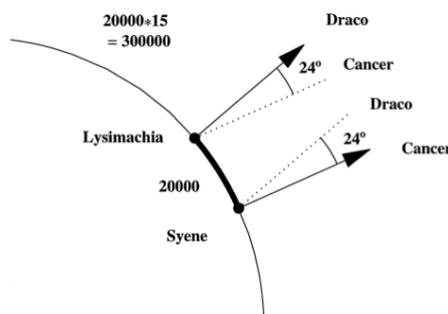
Il problema dei minimi quadrati è alla base di molti metodi di fitting di curve e superfici. Fare il fitting di una superficie vuol dire, data una classe di superfici, trovare quella che meglio approssima i dati che abbiamo a disposizione (ad esempio da esperimenti e misurazioni). L'esempio più semplice di fitting per una curva è quello della regressione lineare, in questo caso i dati sono  $n$  punti e il problema consiste nel trovare la retta che meglio approssima la distribuzione di tali punti. Per  $n = 1$  tutte le rette passanti per il punto dato vanno bene, per  $n = 2$  abbiamo una sola retta che passa per i due punti, per  $n > 2$  il problema cessa di essere banale. Chiaramente possiamo immaginare, dati  $n$  punti di voler trovare una superficie diversa da una retta che meglio li approssima, ad esempio nel piano possiamo cercare una circonferenza che meglio approssima la distribuzione di questi punti, per  $n = 1, 2, 3$  il problema è banale come per la regressione lineare, per  $n > 3$  il problema diventa più difficile. Quindi, stabilito cosa vuol dire fare il fitting di una curva o una superficie, in questo lavoro si mostrerà un esempio chiave che è quello del fitting della superficie della terra, quindi il problema di determinare la forma della terra e poi di fare il fitting delle misurazioni. In conclusione verrà mostrata una generalizzazione della regressione lineare, ovvero

fare il fitting di un iperpiano in  $\mathbb{C}^n$ , noto come problema generalizzato ai minimi quadrati, e si mostrerà qualche esempio di fitting non lineare mostrando come il problema sia malcondizionato e sarà presentato un problema aperto.

## 2 Forma della terra: primi tentativi

Il problema del determinare la forma della terra è stato affrontato sin dalle origini, alcune teorie sono state formulate da filosofi, altre erano canoni religiosi, ad esempio le prime teorie erano che la terra fosse un disco circolare piatto oppure un piano infinito; chiaramente queste teorie sono state smentite con lo sviluppo tecnologico e scientifico. La prima ipotesi ragionevole che è stata fatta è quella che la terra fosse una sfera, stabilito ciò, è nato il problema di misurarne la circonferenza. Il primo tentativo è quello di Archimede e Posidio nel III secolo a.C., questi sono partiti dal presupposto che

- Il Sole è tanto distante da considerare paralleli i raggi su Alessandria e su Syene
- Syene è situata sul tropico del Cancro
- Syene e Alessandria (in Grecia) sono sullo stesso meridiano



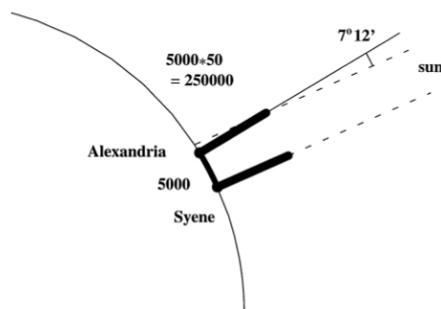
La distanza tra Syene e Lysimachia è circa 20000 stadi e la declinazione delle due costellazioni è circa 24 gradi (1/15 di un cerchio), quindi la circonferenza della terra è circa

$$15 \times 20000 = 300000 \text{ stadi.}$$

Il secondo tentativo è stato fatto da Eratostene nel II secolo a.C., la misurazione è stata fatta con il presupposto che

- Il Sole è tanto distante da considerare paralleli i raggi su Alessandria e su Syene
- Syene è situata sul tropico del Cancro
- Syene e Alessandria (in Grecia) sono sullo stesso meridiano

Durante il solstizio d'estate Eratostene calcolò l'angolo di elevazione del Sole ad Alessandria, misurando l'ombra proiettata da un bastone piantato in terra, ricavando approssimativamente un valore di 1/50 di circonferenza (cioè  $7^\circ 12'$ ), mentre a Syene il bastone non proiettava alcuna ombra (i raggi erano paralleli al bastone).



Dato che la distanza tra Syene e Alessandria è 5000 stadi si ricava che la circonferenza della terra è

$$50 \times 5000 = 250000 \text{ stadi.}$$

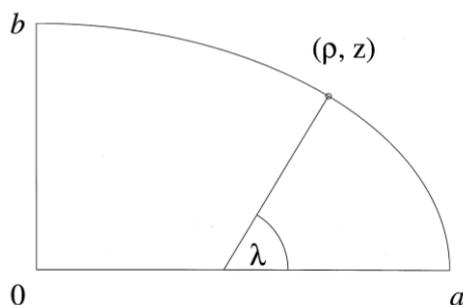
Seppur le ipotesi che sono state fatte per fare le due misure non sono del tutto corrette, i risultati sono troppo diversi.

Archimede e Posidonio	300 000 stadi
Eratostene	250 000 stadi

Considerando che 1 stadio è circa 157.5 metri, le misure fatte sono incoerenti, ciò ha indotto a pensare che la forma della terra non sia perfettamente sferica.

### 3 Fitting di un'ellissoide

Abbandonata l'idea che la terra potesse esser di forma perfettamente sferica si è pensato di approssimare la sua forma con un ellissoide, quindi nasce il problema del determinare tale ellissoide, ovvero trovare il semiasse maggiore e il semiasse minore, quindi fare il fitting di questa superficie. Diamo ora una formulazione matematica del problema, consideriamo l'ellissoide in coordinate cilindriche



Quindi un punto  $(x, y, z)$  sull'ellissoide sarà determinato dalle coordinate  $(\rho, \lambda, z)$  dove  $\lambda$  è l'angolo tra la normale alla superficie per il punto che stiamo considerando e il piano equatoriale mentre  $\rho = \sqrt{x^2 + y^2}$ . Osserviamo dapprima che l'origine è situata in uno dei fuochi, quindi imponendo l'appartenenza all'ellissoide troviamo

$$\rho = \frac{a \cos(\lambda)}{\sqrt{1 - e^2 [\sin(\lambda)]^2}} \quad z = \frac{a \sigma \sin(\lambda)}{\sqrt{1 - e^2 [\sin(\lambda)]^2}}$$

dove si è posto  $a \geq b > 0$ ,  $e^2 = 1 - (b/a)^2$  è il quadrato dell'eccentricità mentre  $\sigma^2 = 1 - e^2 = (b/a)^2$ . Inoltre svolgendo i calcoli si trova come varia la lunghezza d'arco rispetto all'angolo

$$ds = \frac{a\sigma^2}{\left\{1 - e^2 [\sin(\lambda)]^2\right\}^{3/2}} d\lambda = a\sigma^2 \left\{1 + \frac{3}{2}e^2 [\sin(\lambda)]^2 + \frac{3 \cdot 5}{2^3}e^4 [\sin(\lambda)]^4 + \dots\right\} d\lambda$$

nella formula precedente è stato fatto lo sviluppo in serie di Taylor. Daltronde sappiamo che  $e^2 < 0.01$ , quindi se prendiamo solo i primi due termini otteniamo un errore relativo più piccolo di 0.00025, infatti se prendiamo i termini successivi ai primi due nella serie

$$\sum_{k=2}^{\infty} \prod_{l=1}^k (2l+1) \frac{|e \sin(\lambda)|^{2k}}{2^k \cdot k!} < \frac{15e^4}{8} \left(1 + \frac{7e^2}{6}\right) \sum_{k=0}^{\infty} e^k < 0.00025$$

pertanto posto  $c_0 = a\sigma^2$  e  $c_1 = \frac{3}{2}a\sigma^2e^2$  troviamo la relazione

$$\frac{\Delta s}{\Delta \lambda} = c_0 + c_1 [\sin(\lambda)]^2$$

quindi la misura delle lunghezze di  $n$  archi di un meridiano producono  $n$  equazioni, la cui soluzione è un'approssimazione di  $c_0$  e  $c_1$  che determinano l'ellissoide, ovvero da queste riusciamo a ricavare la lunghezza dei semiassi e quindi a fare il fitting dell'ellissoide.

Quindi abbiamo un sistema sovradeterminato dato che ci sono solo due variabili e un numero arbitrario di equazioni in base a quante misurazioni si fanno, questo tipo di problema è noto come problema ai minimi quadrati. Faremo spesso riferimento a questa tabella dovuta a Laplace che per primo ha sviluppato un metodo per affrontare questo tipo di problemi.

$\Delta s/\Delta \lambda = c_0 + c_1[\sin(\lambda)]^2$ ;	location;	latitude $\lambda$ ;	arc $\Delta \lambda$ ;
$25538.85 = c_0 + c_1 * 0.00000$ ;	Peru;	$00.0000^\circ$ ;	$3.4633^\circ$ ;
$25666.65 = c_0 + c_1 * 0.30156$ ;	Good Hope;	$37.0093^\circ$ ;	$1.3572^\circ$ ;
$25599.60 = c_0 + c_1 * 0.39946$ ;	Pennsylvania;	$43.5556^\circ$ ;	$1.6435^\circ$ ;
$25640.55 = c_0 + c_1 * 0.46541$ ;	Italy;	$47.7963^\circ$ ;	$2.4034^\circ$ ;
$25658.28 = c_0 + c_1 * 0.52093$ ;	France;	$51.3327^\circ$ ;	$10.7487^\circ$ ;
$25683.30 = c_0 + c_1 * 0.54850$ ;	Austria;	$53.0926^\circ$ ;	$3.2734^\circ$ ;
$25832.25 = c_0 + c_1 * 0.83887$ ;	Lapland;	$73.7037^\circ$ ;	$1.0644^\circ$ .

## 4 Precursori del metodo dei minimi quadrati

Il problema del fitting di un'ellissoide si riduce a quello del risolvere un sistema lineare sovradeterminato, come vedremo più avanti ci sono vari modi per risolverlo ma è interessante vedere i primi approcci che ci sono stati a questo problema, ne vedremo in dettaglio due: quello di Laplace e quello di Gauss.

## 4.1 Metodo di Laplace

Come annunciato Laplace fu il primo ad affrontare questo tipo di problemi, il suo metodo consiste nel minimizzare la media dei valori assoluti degli errori con il vincolo che la somma degli errori sia zero, ipotesi ragionevole se le misure sono molte e abbastanza accurate. Mostriamo il metodo di Laplace con un esempio, si consideri il seguente sistema

$$\begin{cases} x + y = 4 \\ x - y = 0 \\ x - y = 2 \\ x + y = 6 \end{cases}$$

questo è un sistema sovradeterminato come quelli che possono uscire dal fitting di un'ellissoide. Abbiamo detto che la somma degli errori deve essere nulla, quindi in questo caso se  $A$  è la matrice dei coefficienti,  $b$  i termini noti e  $x = (x, y)$  le variabili, posto  $r = Ax - b$  dobbiamo imporre  $r = 0$ , quindi

$$x + y + x - y + x - y + x + y = 4 + 0 + 2 + 6$$

quindi troviamo  $x = 3$ , dobbiamo ora minimizzare la media dei valori assoluti degli errori

$$\frac{|(3+y) - 4| + |3 - y| + |(3-y) - 2| + |(3+y) - 6|}{4} = \begin{cases} 2 - y > 1 & \text{se } y < 1 \\ 1 = 1 & \text{se } 1 \leq y \leq 3 \\ y - 2 > 1 & \text{se } 3 < y \end{cases}$$

quindi il minimo è raggiunto per  $1 \leq y \leq 3$ , pertanto la soluzione secondo il metodo di Laplace è  $(x, y) \in \{3\} \times [1, 3]$ .

Osserviamo che le colonne dei coefficienti sono ortogonali, in tal caso è possibile risolvere il problema  $Ax = b$  moltiplicando prima per  $A(:, 1)$  e risolvere rispetto alla  $x$  e poi per  $A(:, 2)$  e risolvere rispetto alla  $y$ , pertanto il risultato in questo caso è  $x = 3$  e  $y = 2$ . Osserviamo che il metodo appena descritto funziona ogni volta che abbiamo  $n$  equazioni con la proprietà che la matrice associata al sistema lineare abbia le colonne ortogonali a due a due.

## 4.2 Metodo di Gauss

Il problema ai minimi quadrati consiste nel risolvere il sistema  $Ax = b$  la matrice  $A$  ha più righe che colonne, pertanto il sistema è sovradeterminato quindi la parola risolvere non ha ancora un significato preciso dato che la soluzione potrebbe non esistere, daltronde abbiamo già visto nella sezione precedente come Laplace ha affrontato il problema e come sia possibile risolverlo se le colonne di  $A$  sono a due a due ortogonali. Diamo ora una formulazione un po più accurata in termini di fitting di una funzione lineare, consideriamo i punti  $P_i = (a_{i,1}, \dots, a_{i,n})$  e i valori  $b_i$  per  $i = 1, \dots, m$ , cerchiamo una funzione lineare

$$u : \mathbb{R}^n \rightarrow \mathbb{R} \quad \text{tale che} \quad u(P_i) = b_i$$

ovvero il problema consiste nello stimare i coefficienti  $x_1, \dots, x_n$  tali che

$$\begin{cases} a_{1,1}x_1 + \dots + a_{1,n}x_n = b_1 \\ \vdots \\ a_{m,1}x_1 + \dots + a_{m,n}x_n = b_m \end{cases}$$

chiaramente non è detto che tale funzione lineare esista, se abbiamo  $m = n$  allora la funzione  $u$  esiste ed è unica, se  $m < n$  allora esistono più di una funzione lineare che soddisfa le condizioni, se  $m > n$  una tale funzione  $u$  potrebbe non esistere, quindi cercheremo una funzione lineare che più si avvicina a quei valori. Mostriamo ora come Gauss ha risolto il problema.

Consideriamo dunque il problema  $Ax = b$ , per la situazione che stiamo considerando è sensato considerare il vettore  $b = (b_i)$  con componenti che sono in realtà la speranza di variabili aleatorie  $b_i = \mathbb{E}[B_i]$ . Consideriamo la matrice  $V = (V_{i,j})$  con  $V_{i,j} = \mathbb{E}[(B_i - b)(B_j - b_j)]$  detta matrice delle covarianze, osserveremo subito che la soluzione del sistema lineare  $Ax = b$  (se esiste) è ancora una variabile aleatoria. Consideriamo le matrici  $L$  tali che  $LA = I$  (se esistono), allora abbiamo che se  $A\bar{x} = b$  allora

$$\bar{x} = I\bar{x} = LA\bar{x} = Lb$$

pertanto se conosciamo la matrice  $L$  allora conosciamo anche la soluzione  $\bar{x}$ , il problema risolto da Gauss è stato quello di determinare questa matrice con l'ulteriore condizione che la covarianza  $U = \mathbb{E}[(x - \bar{x})(x - \bar{x})^H]$  sia minima. Osserviamo che la matrice seguente rispetta tutte le proprietà richieste

$$L = (A^H V^{-1} A)^{-1} A^H V^{-1}$$

il fatto che  $LA = I$  è ovvio, bisogna mostrare che la covarianza è minima. Infatti facendo il conto esplicito, se  $K$  è una matrice tale che  $KA = I$ , allora

$$\begin{aligned} U &= \mathbb{E}[(x - \bar{x})(x - \bar{x})^H] = \mathbb{E}[(x - Kb)(x - Kb)^H] \\ &= \mathbb{E}\left\{[X - KAx - K(b - Ax)][X - KAx - K(b - Ax)]^H\right\} \\ &= \mathbb{E}\left\{[K(b - Ax)][K(b - Ax)]^H\right\} \\ &= K\mathbb{E}\{(b - Ax)(b - Ax)^H\}K^H = KVK^H \\ &= LVL^H + (K - L)V(L - K)^H + (K - L)V(K - L)^H \end{aligned}$$

dalla condizione  $KA = I$  e dalla definizione stessa di  $L$  segue che i termini di mezzo sono nulli, quindi

$$U = LVL^H + (K - L)V(K - L)^H$$

quindi  $U$  è hermitiana semidefinita positiva. Per ogni vettore  $z$  vale

$$z^H U z = z^H LVL^H z + z^H (K - L)V(K - L)^H z \geq z^H LVL^H z$$

e il minimo viene raggiunto per  $K = L$  che è quello che volevamo mostrare. Inoltre troviamo subito che la

$$U = (A^H V^{-1} A)^{-1}$$

### 4.3 Qualche risultato numerico

Abbiamo fin ora mostrato due metodi per risolvere il problema dei sistemi sovradeterminati, quello di Laplace e quello di Gauss, ma teniamo a mente che questi metodi sono nati per risolvere il problema del fitting di una superficie o di una funzione lineare. Usiamo la tabella che è nella sezione 3 e applichiamo il metodo di Gauss, dobbiamo costruire dunque la matrice  $V$ , osserviamo che non abbiamo nessuna correlazione tra i termini assegnati e i diversi archi, ovvero la covarianza tra le diverse misurazioni è nulla, quindi la matrice  $V$  sarà diagonale e possiamo vederla come  $V = W \cdot W$  con

$$W = \text{diag}(3.4633, 1.3572, 1.6435, 2.4034, 10.7487, 3.2734, 1.0644)$$

come risultato otteniamo

$$c_0 = 25534.47 \quad c_1 = 242.81$$

ovvero

$$\begin{aligned} e^2 &= 0.006339 \\ \sigma^2 &= 0.006339 \\ f &= 1 - b/a = 1 - \sigma = 0.003175 \\ a &= 100 \ 170.25 \ m \end{aligned}$$

e ricaviamo anche la matrice della covarianza

$$U = \begin{pmatrix} 0.007701 & -0.147686 \\ -0.147686 & 0.309706 \end{pmatrix}$$

osserviamo che per come abbiamo impostato il problema abbiamo trattato le lunghezze degli archi come variabili aleatorie mentre la latitudine l'abbiamo considerata esatta, vedremo più avanti che è possibile cambiare approccio con il problema ai minimi quadrati generalizzato.

Se invece affrontiamo il problema con il metodo di Laplace otteniamo

$$\begin{aligned} f &= 0.00335281066474 \\ a &= 6 \ 378 \ 137 \ m \end{aligned}$$

osserviamo che soprattutto il valore di  $a$  è troppo diverso ma il tutto è giustificato dal fatto che i metodi usati sono diversi, il valore più sensato è quello ottenuto da Laplace.

## 5 Problema ai minimi quadrati

Fin ora non abbiamo ancora enunciato formalmente il problema dei minimi quadrati, abbiamo solo detto che consiste nel trovare la soluzione in un certo senso minima di un sistema sovradeterminato e abbiamo mostrato come questo problema nasca dal fare il fitting di una superficie e come questo problema è stato affrontato da Laplace e da Gauss. Daremo ora la formulazione completa di tale problema, mostreremo come risolverlo e ne faremo l'analisi dell'errore.

Sia  $A \in \mathbb{C}^{m \times n}$  con  $m \geq n$ , consideriamo il sistema  $Ax = b$ , il problema ai minimi quadrati consiste nel cercare i vettori  $x \in \mathbb{C}^n$  tali che il residuo  $r = \|Ax - b\|$  sia minimo. Non abbiamo specificato la norma ma senza perdere di generalità per il principio di equivalenza delle norme i vettori  $x$  non dipendono dalla norma scelta, al massimo cambia il valore di  $r$ , in questo contesto considereremo la norma euclidea.

## 5.1 Risoluzione del problema ai minimi quadrati con la fattorizzazione QR

Consideriamo il problema ai minimi quadrati  $Ax = b$  con  $A \in \mathbb{C}^{m \times n}$  con  $n \geq m$ , se  $A = QR$  con  $Q$  matrice unitaria ed  $R$  matrice triangolare superiore, allora

$$Ax = b$$

$$Q^H Ax = Q^H b$$

$$Rx = Q^H b$$

Daltronde la norma euclidea è preservata perchè  $Q$  è una matrice unitaria, quindi

$$\|Ax - b\|_2 = \|Rx - Q^H b\|_2$$

La matrice  $R$  avrà  $n$  colonne linearmente indipendenti e zeri sotto l' $r$ -esima riga dove  $r$  è il rango di  $A$ . Allora  $\|Rx - Q^H b\|$  raggiunge il minimo se e soltanto se  $\tilde{x} = x(1 : r)$  è l'unica soluzione delle prime  $n$  equazioni date dal sistema  $Rx = Q^H b$ .

Per ottenere una fattorizzazione  $QR$  di una matrice  $A$  con rango  $r$  possiamo agire nel seguente modo

- Troviamo una matrice di permutazione  $P \in \mathbb{C}^{n \times n}$  tale che la matrice  $AP$  abbia le prime  $r$  colonne linearmente indipendenti
- Procediamo con le trasformazioni di Householder, troviamo una successione di matrici  $S_1, \dots, S_r$  con la seguente proprietà
  - $S_1(PA)e_1 = \lambda_1 e_1$  ovvero la matrice  $S_1(PA)$  ha come prima colonna un multiplo del primo vettore della base canonica
  - in generale  $S_i \dots S_1(PA)$  è una matrice tale per  $j \leq i$ , la colonna  $j$ -esima ha zeri dalla posizione  $i$  in poi.

Le matrici  $S_1, \dots, S_r$  sono matrici unitarie di Householder quindi in conclusione otteniamo  $Q = S_1 \dots S_r$  e di conseguenza avremo  $R = S_r \dots S_1 A$  che conterrà zeri dall' $r$ -esima riga in poi.

(Per una descrizione completa delle matrici di Householder si veda l'appendice).

Consideriamo ora il cambio di coordinate  $z = P^{-1}x$ , allora

$$Ax = b$$

$$(Q^H AP)(P^{-1}x) = Q^H b$$

$$Rz = Q^H b$$

A questo punto se imponiamo  $z_{r+1} = \dots = z_n = 0$  e che  $z_1, \dots, z_r$  risolvano le prime  $r$  equazioni allora la soluzione è unica ed è la soluzione del problema ai minimi quadrati a meno di ritornare alle coordinate iniziali, ovvero  $x = Pz$ . Che  $x$  minimizzi  $\|Ax - b\|_2$  è chiaro considerando che  $Q$  è unitaria, inoltre è possibile calcolare  $\|Ax - b\|_2$  dove  $x$  è la soluzione appena definita

$$\|Ax - b\|_2 = \|Rx - Q^H b\|_2 = \|(Rx - Q^H b)_{r+1}, \dots, (Rx - b)_n\|_2$$

## 5.2 Problema ai minimi quadrati vincolato

Nella geodesia spesso è necessario trattare il problema ai minimi quadrati vincolato, ovvero abbiamo:  $C \in \mathbb{C}^{k \times n}$ ,  $E \in \mathbb{C}^{l \times n}$ ,  $d \in \mathbb{C}^k$ ,  $f \in \mathbb{C}^l$ ; il problema consiste nel determinare un vettore  $x \in \mathbb{C}^n$  che minimizza  $\|Ex - f\|_2$  ma che soddisfa  $Cx = d$ . Cercheremo di ricondurre il problema ai minimi quadrati vincolato ad un problema ai minimi quadrati non vincolato, per farlo consideriamo una base ortonormale  $(q_1, \dots, q_k, q_{k+1}, \dots, q_n)$  dove  $(q_{k+1}, \dots, q_n)$  è una base ortonormale del nucleo di  $C$  che quindi ci da una parametrizzazione dello spazio delle soluzioni del sistema  $Cx = d$  il che ci riconduce ad un problema ai minimi quadrati non vincolato nel sottospazio  $(q_1, \dots, q_k)$ , mostriamo come fare questo passaggio nel concreto. Supponiamo che tutte le matrici siano di rango massimo, sia  $Q = (q_1 | \dots | q_n)$  allora  $C^H = QR$ , mentre  $C = LQ^H$  dove chiaramente  $R$  è una matrice triangolare superiore mentre  $L$  è una matrice triangolare inferiore con  $L = R^H$ . Dato che  $R = Q^H C^H$  ha zeri sotto la  $k$ -esima riga segue che in  $Q^H$  tutte le righe  $q_{k+1}^H, \dots, q_n^H$  sono perpendicolari a tutte le colonne di  $C^H$  quindi segue che le righe  $q_1^H, \dots, q_k^H$  generano lo stesso spazio che generano le colonne di  $C^H$ , quindi  $Q$  ci da il cambiamento di base che ci serve, infatti, posto  $w = Q^H x$  abbiamo che il sistema diventa

$$\begin{pmatrix} L \\ EQ \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} C \\ E \end{pmatrix} Q(Q^H x) = \begin{pmatrix} d \\ f \end{pmatrix}$$

Quindi esiste esattamente una soluzione  $w_1$  del sistema

$$L \begin{pmatrix} w_1 \\ 0 \end{pmatrix} = d$$

Quindi il problema iniziale si riduce a determinare un vettore  $w_2 \in \mathbb{C}^l$  che minimizza

$$\left\| \begin{pmatrix} L \\ EQ \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} - \begin{pmatrix} d \\ f \end{pmatrix} \right\|_2 = \left\| EQ \begin{pmatrix} 0 \\ w_2 \end{pmatrix} - \left[ f - EQ \begin{pmatrix} w_1 \\ 0 \end{pmatrix} \right] \right\|_2$$

Dunque abbiamo ricondotto il problema ai minimi quadrati vincolato ad un problema ai minimi quadrati non vincolato.

## 5.3 Esempio di problema ai minimi quadrati vincolato

Si può dimostrare come conseguenza del teorema di Gauss-Bonnet e facendo alcuni conti che se approssimiamo la terra con un ellissoide (di cui facciamo

il fitting) si giunge alla conclusione che la somma degli angoli interni di un triangolo è  $180^\circ 0' 1.749''$ . De Krayenhof misurò i seguenti angoli

$$\alpha = 50^\circ 58' 15.238'' \text{ in Harlingen}$$

$$\beta = 82^\circ 47' 15.351'' \text{ in Leeuwarden}$$

$$\gamma = 46^\circ 14' 27.202'' \text{ in Ballum}$$

in questo caso abbiamo che  $\alpha + \beta + \gamma = 179^\circ 59' 57.791''$  quindi ci sono degli errori nelle misure dato che la somma non è corretta. E' possibile correggere questi errori impostando un problema ai minimi quadrati vincolato, ovvero si tratterà di risolvere

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} = \begin{pmatrix} 50^\circ 58' 15.238'' \\ 82^\circ 47' 15.351'' \\ 46^\circ 14' 27.202'' \end{pmatrix}$$

con il vincolo

$$\alpha + \beta + \gamma = 180^\circ 0' 1.749''$$

ovvero cerchiamo la più piccola correzione dei dati tale che la somma degli angoli interni sia corretta, applicando il metodo mostrato prima troviamo che la soluzione è

$$\alpha = 50^\circ 58' 16.557333''$$

$$\beta = 82^\circ 47' 16.670333''$$

$$\gamma = 46^\circ 14' 28.521333''$$

## 6 Fattorizzazioni SVD

Consideriamo ora un altro approccio per risolvere il problema ai minimi quadrati, quello delle fattorizzazioni SVD

**Teorema 6.1.** *Per ogni matrice  $A \in \mathbb{C}^{m \times n}$  di rango  $r$  esistono due matrici unitarie  $U$  e  $V$  e una matrice diagonale  $\Sigma$  tale che*

$$A = U \Sigma V^H = u_1 \sigma_1 v_1^H + \dots + u_r \sigma_r v_r^H = \tilde{U} \tilde{\Sigma} \tilde{V}^H$$

$U \in \mathbb{C}^{m \times m}$  è unitaria, le prime  $r$  colonne formano una base ortonormale dello spazio generato dalle colonne di  $A$ , mentre le restanti  $m - r$  colonne formano una base ortonormale del nucleo di  $A^H$ .

$V \in \mathbb{C}^{n \times n}$  è unitaria, le prime  $r$  colonne formano una base ortonormale dello spazio generato dalle righe di  $A$ , mentre le restanti  $n - r$  colonne formano una base ortonormale del nucleo di  $A$

$\Sigma \in \mathbb{C}^{m \times n}$  è diagonale nel senso che  $\Sigma_{k,l} = 0$  per  $k \neq l$  mentre  $\Sigma_{i,i} = \sigma_i$  con

$$\sigma_1 \geq \dots \geq \sigma_r > 0 = \sigma_{r+1} = \dots = \sigma_{\min\{m,n\}}$$

Inoltre vale che per  $1 \leq j \leq r$

$$Av_j = \sigma_j u_j$$

$$A^H u_j = \sigma_j v_j$$

mentre per  $r+1 \leq j \leq n$   $Av_j = 0$  quindi definiamo

$$\tilde{U} = (u_1, \dots, u_r) \in \mathbb{C}^{m \times r}$$

$$\tilde{V} = (v_1, \dots, v_r) \in \mathbb{C}^{n \times r}$$

$$\tilde{\Sigma} = \text{diag}(\sigma_1, \dots, \sigma_r) \in \mathbb{C}^{r \times r}$$

*dimostrazione:* consideriamo la matrice semidefinita positiva  $AA^H \in \mathbb{C}^{n \times n}$  con autovalori

$$\lambda_1 \geq \dots \lambda_r > 0 = \lambda_{r+1} = \dots = \lambda_n$$

e sia  $V = (v_1, \dots, v_r, v_{r+1}, \dots, v_n)$  una base di autovettori ortonormali. Definiamo  $\sigma_j := \sqrt{\lambda_j}$  e  $u_j = (1/\sigma_j)Av_j$  a questo punto svolgendo i calcoli si verifica la tesi.  $\square$

In generale chiameremo i  $\sigma_1, \dots, \sigma_r$  valori singolari,  $v_1, \dots, v_n$  vettori singolari destri mentre  $u_1, \dots, u_m$  vettori singolari sinistri.

**Definizione 1.**  $A = U\Sigma V^H$  è detta *decomposizione ai valori singolari di A*, i  $\sigma_1, \dots, \sigma_r$  sono i *valori singolari*,  $v_1, \dots, v_n$  sono i *vettori singolari destri* mentre  $u_1, \dots, u_m$  sono i *vettori singolari sinistri*.

## 6.1 Risoluzione del problema ai minimi quadrati con le fattorizzazioni SVD

Una fattorizzazione SVD di una matrice  $A$  ci permette di risolvere il problema ai minimi quadrati.  $\tilde{U}^H b$  proietta  $b$  ortogonalmente nello spazio generato dalle colonne di  $A$ , pertanto

$$\|Ax - \tilde{U}^H b\|_2 \leq \|Ax - b\|_2 \quad \forall x \in \mathbb{C}^n$$

dato che  $\tilde{U}^H b$  si trova nello spazio generato dalle colonne di  $A$  allora esiste una soluzione  $x \in \mathbb{C}^n$  di  $Ax = \tilde{U}^H b$ .

Inoltre tutte le soluzioni di questo sistema differiscono al più per un vettore del nucleo di  $A$ , quindi se cerchiamo la più corta (inteso come norma euclidea) allora dobbiamo fare la proiezione ortogonale nel complementare del nucleo di  $A$ , quindi dobbiamo prendere  $x^\dagger = \tilde{V}^H x$ . Mostriamo ora come trovare esplicitamente il vettore  $x^\dagger$  che sarà la soluzione del problema ai minimi quadrati

$$Ax = b$$

$$(\tilde{U}\tilde{\Sigma}\tilde{V}^H)x = b$$

$$\tilde{\Sigma}(\tilde{V}^H x) = \tilde{U}^H b$$

$$\tilde{V}^H x = \tilde{\Sigma}^{-1}\tilde{U}^H b$$

$$x^\dagger = (\tilde{V}\tilde{\Sigma}^{-1}\tilde{U}^H)b$$

**Definizione 2.** La matrice pseudoinversa di  $A$  (oppure inversa di Penrose) si definisce come

$$A^\dagger := \tilde{V}\tilde{\Sigma}^{-1}\tilde{U}^H$$

Quindi la soluzione di un problema ai minimi quadrati  $Ax = b$  sarà  $x^\dagger = A^\dagger b$

## 7 Norme e numero di condizionamento

Le norme e il numero di condizionamento di una matrice servono per stimare la propagazione degli errori durante la computazione.

**Definizione 3.** Date due norme  $\|\cdot\|_p$  su  $\mathbb{C}^n$  e  $\|\cdot\|_q$  su  $\mathbb{C}^m$  queste due inducono una norma  $\|\cdot\|_{p,q}$  sullo spazio delle matrici  $\mathbb{C}^{m \times n}$  definita come

$$\begin{aligned} \|A\|_{p,q} &:= \max \{ \|Au\|_q \text{ tale che } u \in \mathbb{C}^n, \|u\|_p = 1 \} \\ &= \max \{ \|Au\|_q / \|u\|_p \text{ tale che } u \in \mathbb{C}^n, u \neq 0 \} \end{aligned}$$

**Lemma 7.1.** Per tutti i numeri reali  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_{n-1} \geq \sigma_n \geq 0$  vale

$$\begin{aligned} \min_{\|x\|_2=1} \sum_{i=1}^n (\sigma_i x_i)^2 &= \min_{1 \leq i \leq n} \sigma_i^2 = \sigma_n^2 \\ \max_{\|x\|_2=1} \sum_{i=1}^n (\sigma_i x_i)^2 &= \max_{1 \leq i \leq n} \sigma_i^2 = \sigma_1^2 \end{aligned}$$

*dimostrazione:* Dimostriamo la prima. Sappiamo che  $\sum_{i=1}^n x_i^2 = 1$  quindi abbiamo  $x_n^2 = 1 - \sum_{i=1}^{n-1} x_i^2$ , quindi andando a sostituire

$$\begin{aligned} \sum_{i=1}^n (\sigma_i x_i)^2 &= \sigma_n^2 x_n^2 + \sum_{i=1}^{n-1} (\sigma_i x_i)^2 = \sigma_n^2 \left( 1 - \sum_{i=1}^{n-1} x_i^2 \right) + \sum_{i=1}^{n-1} \sigma_i^2 x_i^2 \\ &= \sigma_n^2 + \sum_{i=1}^{n-1} (\sigma_i^2 - \sigma_n^2) x_i^2 \geq \sigma_n^2 \end{aligned}$$

dove l'uguaglianza vale se e soltanto se  $x_i = 0$  per  $\sigma_i \neq \sigma_n$ .

L'altra formula si dimostra in modo speculare.  $\square$

**Definizione 4.** Data una matrice  $A$  quadrata e invertibile definiamo il numero di condizionamento

$$K_2(A) = \|A\|_2 \|A^{-1}\|_2 = \sigma_1 / \sigma_n$$

questo numero fornisce una stima superiore e una stima inferiore della differenza  $\|\tilde{x} - x\|_p$  dove  $x$  è la soluzione del sistema  $Ax = b$  e  $\tilde{x}$  invece è la soluzione del sistema perturbato  $A\tilde{x} = \tilde{b}$

**Proposizione 7.2.** Per ogni matrice  $A \in \mathbb{C}^{m \times n}$  vale che  $\|A\|_{2,2} = \sigma_1$  ovvero il più grande valore singolare, mentre il numero di condizionamento è  $K_2(A) = \sigma_1 / \sigma_n$

*dimostrazione:* Sia data una decomposizione ai valori singolari  $A = U\Sigma V^H$ . Per ogni vettore  $x \in \mathbb{C}^n$  con  $\|x\|_2 = 1$ , sia  $w = V^H x$ , allora essendo  $V$  unitaria abbiamo  $\|w\|_2 = 1$ , quindi

$$\begin{aligned} \|A\|_2^2 &= x^H A^H A x = x^H (V \Sigma^H U^H) (U \Sigma V^H) x = x^H V \Sigma^H \Sigma V^H x \\ &= \|\Sigma w\|_2^2 = \sum_{i=1}^n (\sigma_i w_i)^2 \leq \sigma_1^2 \end{aligned}$$

dove il massimo valore viene raggiunto per  $w = e_1$  o equivalentemente per  $x = v_1$ , quindi vale anche  $K_2(A) = \sigma_1/\sigma_n$   $\square$

**Definizione 5.** Data una qualsiasi matrice  $A \in \mathbb{C}^{m \times n}$  e date due norme  $\|\cdot\|_p$  su  $\mathbb{C}^n$  e  $\|\cdot\|_q$  su  $\mathbb{C}^m$  definiamo il numero di condizionamento  $K_{p,q}(A) = \|A\|_{p,q} \|A^\dagger\|_{q,p}$ .

**Proposizione 7.3.** Data una matrice  $A$  quadrata e invertibile e dati  $b, \tilde{b}, x$  e  $\tilde{x}$  con  $Ax = b$  e  $A\tilde{x} = \tilde{b}$  vale

$$\frac{1}{K_{p,q}(A)} \frac{\|\tilde{b} - b\|_q}{\|b\|_q} \leq \frac{\|\tilde{x} - x\|_p}{\|x\|_p} \leq K_{p,q}(A) \frac{\|\tilde{b} - b\|_q}{\|b\|_q}$$

*dimostrazione:* Basta usare  $\|b\|_q = \|Ax\|_q \leq \|A\|_{p,q} \cdot \|x\|_p$  e  $\|\tilde{x} - x\|_p = \|A^{-1}(\tilde{b} - b)\|_p \leq \|A^{-1}\|_{q,p} \cdot \|\tilde{b} - b\|_q$   $\square$

Abbiamo dunque visto cosa succede se risolviamo un sistema dove consideriamo il termine noto perturbato, vediamo ora cosa succede se invece supponiamo che la matrice sia perturbata

**Proposizione 7.4.** Data una matrice quadrata invertibile  $A$  e data un'altra matrice  $C$  con la proprietà che  $\|A - C\| \leq 1/\|A^{-1}\|$ , se  $Ax = b$  e  $Cw = b$  allora

$$\frac{\|w - x\|}{\|x\|} \leq \frac{K(A)}{1 - K(A) \cdot \|A - C\|/\|A\|} \cdot \frac{\|A - C\|}{\|A\|}$$

Queste due proposizioni mostrano come il numero di condizionamento di una matrice quadrata può dare una stima dell'errore considerando che in aritmetica floating point non risolviamo mai esattamente il sistema ma sempre un sistema perturbato, quindi sappiamo quanto la soluzione che abbiamo trovato dista dalla soluzione vera. Daltronde il numero di condizionamento influisce anche nella stima dell'errore in un problema ai minimi quadrati come dice la seguente proposizione

**Proposizione 7.5.** Per tutte le matrici  $A \in \mathbb{C}^{m \times n}$  e  $C \in \mathbb{C}^{m \times n}$  con rango massimo  $r = n \leq m$  e per tutti i vettori  $b \in \mathbb{C}^m$  e  $d \in \mathbb{C}^m$ , se esiste un numero positivo  $\epsilon$  tale che

$$K_2(A)\epsilon \leq 1$$

$$\|A - C\|_2 \leq \epsilon \|A\|_2$$

$$\|b - d\|_2 \leq \epsilon \|b\|_2$$

allora le soluzioni  $\tilde{x} \in \mathbb{C}^n$  e  $\tilde{z} \in \mathbb{C}^n$  dei problemi ai minimi quadrati  $Ax = b$  e  $Cz = d$  soddisfano

$$\frac{\|x - z\|_2}{\|x\|_2} \leq \frac{K_2(A)\epsilon}{1 - K_2(A)\epsilon} \left( 2 + [1 + K_2(A)] \frac{\|b - Ax\|_2}{\|A\|_2 \cdot \|x\|_2} \right)$$

e anche

$$\frac{\|(b - Ax) - (d - Az)\|_2}{\|b\|_2} \leq [1 + 2K_2(A)]\epsilon$$

Il seguente teorema dice che per ogni matrice quadrata invertibile  $A$  la matrice singolare più vicina ha distanza  $1/\|A^{-1}\|$

**Teorema 7.6.** *Per ogni matrice quadrata invertibile  $A$  ed ogni norma indotta vale che*

$$\min_{\det(S)=0} \|A - S\| = \frac{1}{\|A^{-1}\|}$$

*dimostrazione:* Per ogni matrice  $S$  non invertibile esiste un vettore  $z \neq 0$  tale che  $Sz = 0$ , quindi

$$\|A - S\| \geq \frac{\|(A - S)z\|}{\|z\|} = \frac{\|Az\|}{\|z\|} = \frac{\|A^{-1}\| \|Az\|}{\|A^{-1}\| \|z\|} \geq \frac{\|A^{-1}Az\|}{\|A^{-1}\| \|z\|} = \frac{1}{\|A^{-1}\|}$$

Dobbiamo ora mostrare che questo minimo viene effettivamente raggiunto. Esiste un vettore  $y \neq 0$  con  $\|A^{-1}y\| = \|A^{-1}\| \|y\|$ , possiamo inoltre scegliere un funzionale lineare  $w$  tale che

$$w(A^{-1}y) = \|w\| \|A^{-1}y\| = 1$$

sia  $w^*$  la matrice associata a  $w$  nella base canonica tale che  $w(z) = w^*z$  per ogni vettore  $z$  e definiamo

$$S := A - yw^*$$

dobbiamo ora mostrare che  $S$  è singolare e che ha la distanza minima da  $A$ . Che sia singolare segue da

$$S(A^{-1}y) = (A - yw^*) \cdot (A^{-1}y) = y - y \cdot 1 = 0$$

inoltre

$$\begin{aligned} \|A - S\| &= \max \{ \|(yw^*)x\| \text{ tale che } \|x\| = 1 \} \\ &= \max \{ \|y(w^*x)\| \text{ tale che } \|x\| = 1 \} \\ &= \|y\| \cdot \max \{ w^*x \text{ tale che } \|x\| = 1 \} \\ &= \|y\| \|w^*\| \\ &= \|y\| \frac{1}{\|A^{-1}y\|} \\ &= \|y\| \frac{1}{\|A^{-1}\| \|y\|} \\ &= \frac{1}{\|A^{-1}\|} \end{aligned}$$

□

## 8 Approssimazione con matrici di rango inferiore

Il seguente è noto come teorema di Schmidt, Mirsky, e Weyl e ci dice come approssimare una matrice  $C \in \mathbb{C}^{m \times n}$  di rango  $r$  con una matrice  $S \in \mathbb{C}^{m \times n}$  di rango  $k < r$  in modo da minimizzare la norma di Frobenius  $\|C - S\|_F$

**Teorema 8.1.** Per ogni matrice  $C \in \mathbb{C}^{m \times n}$  con  $\sigma_1 \geq \dots \geq \sigma_r > 0$  e con

$$C = \sum_{i=1}^r \sigma_i u_i v_i^H$$

e per ogni matrice  $S \in \mathbb{C}^{m \times n}$  di rango  $k \leq r$  vale

$$\|C - S\|_F^2 \geq \sum_{i=k+1}^r \sigma_i^2$$

dove il minimo viene raggiunto per

$$S = \sum_{i=1}^k \sigma_i u_i v_i^H$$

Nota per il prof: a lezione abbiamo dimostrato questo teorema per la norma indotta dalla norma Euclidea e la dimostrazione è un po più corta, se vuole porto quella

*dimostrazione.* Se  $k = r$  non c'è nulla da dimostrare dato che  $S = C$ .

Facciamo dunque la dimostrazione nel caso  $k < r$ .

Data una matrice  $A$  definiamo  $\sigma_i(A)$  come i valori singolari di  $A$ , mentre  $u_i(A)$  e  $v_i(A)$  sono i vettori singolari sinistri e destri di  $A$ , definiamo la matrice

$$A_k = \sum_{i=1}^k \sigma_i(A) u_i(A) v_i(A)^H$$

**Passo 1** Mostriamo ora che  $\sigma_1(C - S) \geq \sigma_{k+1}(C)$ .

Se  $S$  ha rango  $k$  allora avrà una decomposizione ai valori singolari  $S = \sum_{i=1}^k \tau_i w_i z_i^H = W \tau Z^H$ . Consideriamo lo spazio  $Z^\perp$  ortogonale allo spazio generato dai vettori  $z_1, \dots, z_k$ , quindi  $\dim(Z^\perp) = n - k > n - (k + 1) \geq n - r$ . Dato che lo spazio  $V$  generato dai primi  $k + 1$  vettori colonna  $v_1, \dots, v_{k+1}$  ha appunto dimensione  $k + 1$  segue che  $Z^\perp \cap V \neq \{0\}$ , quindi esiste un vettore di coefficienti  $\gamma \in \mathbb{C}^{k+1}$  dove a meno di normalizzare abbiamo  $\|\gamma\|_2 = 1$  tale che  $x := V\gamma = \sum_{i=1}^{k+1} \gamma_i v_i \in Z^\perp \cap V$  quindi si avrà  $Zx = Sx = 0$ . Sia  $\tilde{\gamma} := (\gamma^H, 0^H)^H \in \mathbb{C}^n$  quindi

$$\begin{aligned} \sigma_1^2(C - S) &\geq x^H (C - S)^H (C - S) x \\ &= x^H C^H C x \\ &= \tilde{\gamma}^H V^H V \Sigma U^H U \Sigma V^H V \tilde{\gamma} \\ &= \tilde{\gamma}^H \Sigma^2 \tilde{\gamma} = \sum_{i=1}^{k+1} (\gamma_i \sigma_i)^2 \geq \sigma_{k+1}^2 \end{aligned}$$

**Passo 2** Mostriamo ora che  $\sigma_1(A - B) \geq |\sigma_1(A) - \sigma_1(B)|$ .

Questa è una conseguenza della disuguaglianza triangolare inversa, infatti

$$\sigma_1(A - B) = \|A - B\|_2 \geq \left| \|A\|_2 - \|B\|_2 \right| = |\sigma_1(A) - \sigma_1(B)|$$

**Passo 3** Generalizziamo il passo 2 per tutti gli altri valori singolari. Sappiamo che per ogni matrice  $G$  e per ogni indice  $l$  abbiamo che  $\sigma_1(G - G_l) = \sigma_{l+1}(G)$ . Di conseguenza per tutte le matrici  $G, H \in \mathbb{C}^{m \times n}$  e per tutti gli indici  $k$  ed  $l$  abbiamo

$$\begin{aligned}\sigma_{l+1}(G) + \sigma_{k+1}(H) &= \sigma_1(G - G_l) + \sigma_1(H - H_k) \geq \sigma_1([G - G_l] + [H - H_k]) \\ &= \sigma_1([G + H] - [G_l + H_k]) \geq \sigma_{l+k+1}(G + H)\end{aligned}$$

dove l'ultimo passaggio è giustificato dal fatto che il rango di  $G_l + H_k$  è minore o uguale a  $l + k$ . Nel caso particolare in cui  $A := G + H$  e  $B := H$  allora

$$\sigma_{l+1}(A - B) \geq \sigma_{l+k+1}(A) - \sigma_{k+1}(B)$$

ponendo

$$G := C - S \quad \text{e} \quad H := S$$

otteniamo

$$\sigma_{l+1}(C - S) + 0 = \sigma_{l+1}(C - S) + \sigma_{k+1}(S) \geq \sigma_{l+1+k}(C)$$

**Conclusion** Se  $S$  ha rango  $k$  allora preso  $G := C - S$  e  $H := S$ , usando il punto precedente

$$\sigma_{l+1}(C - S) + 0 = \sigma_{l+1}(C - S) + \sigma_{k+1}(S) \geq \sigma_{l+1+k}(C)$$

Quindi concludo ottenendo

$$\|C - S\|_F^2 = \sum_{i=1}^r \sigma_i^2(C - S) \geq \sum_{i=1}^r \sigma_{i+k}^2(C) = (\sigma_{k+1} + \dots + \sigma_r^2)(C)$$

L'uguaglianza si ottiene per

$$S = \sum_{i=1}^k \sigma_i u_i v_i^H$$

□

## 9 Problema totale ai minimi quadrati

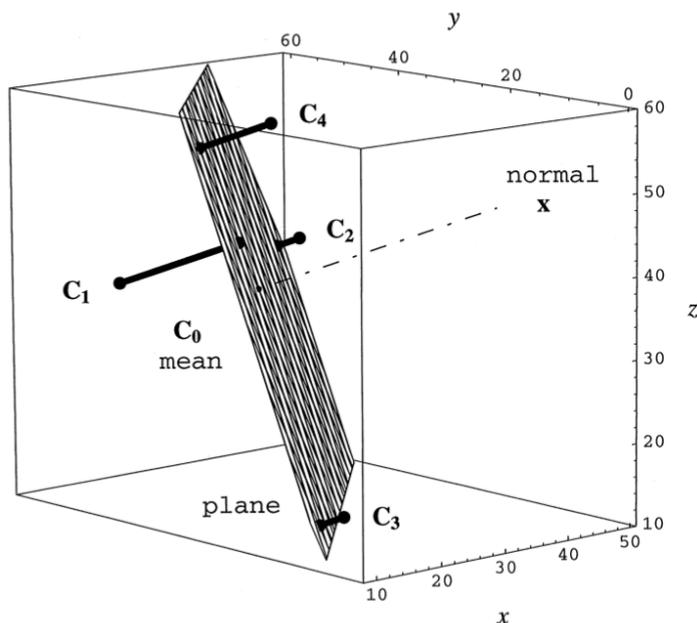
Quello che fin ora abbiamo affrontato è il problema ordinario ai minimi quadrati, ovvero abbiamo considerato il sistema sovradeterminato  $Ax = b$  e abbiamo cercato quell'  $\tilde{x}$  che rende minima la discrepanza  $\|b - \tilde{b}\|$  dove  $\tilde{b} := A\tilde{x}$ . Il problema ai minimi quadrati totale ammette più formulazioni ma l'idea è di perturbare anche la matrice, ma lo vedremo dopo.

### 9.1 Formulazione geometrica

E' sicuramente interessante dare la formulazione geometrica del problema totale ai minimi quadrati dato che quando generalizzeremo al problema non lineare ai minimi quadrati faremo riferimento a questa formulazione.

Il problema totale ai minimi quadrati è la generalizzazione della regressione lineare, ovvero dati i punti  $c_1, \dots, c_m \in \mathbb{C}^{n+1}$  cerchiamo l'iperpiano  $H$  (quindi un sottospazio di dimensione  $n$ ) che minimizza la media delle distanze dei punti dall'iperpiano.

Specifichiamo che la distanza di un punto  $P$  dall'iperpiano  $H$  è da intendersi nel seguente modo: si prende la retta  $n$  normale all'iperpiano  $H$  che passa per il punto  $P$  (esiste ed è unica) e si calcola la distanza tra  $n \cap H$  e  $P$ , questa sarà la nostra distanza punto-iperpiano (vedremo nella generalizzazione che non sempre riusciremo ad usare questa definizione se al posto di iperpiano ci mettiamo una varietà, ma lo affronteremo bene dopo).



Per identificare un iperpiano ci è sufficiente identificare il suo vettore normale e un suo punto. Quindi riconduciamo il problema a trovare un  $c_0 \in H$  e un vettore  $x \perp H$  con  $x \neq 0$  che minimizzano la seguente funzione

$$D(x, c_0; c_1, \dots, c_m) := \sum_{i=1}^m \frac{|\langle c_i - c_0, x \rangle|^2}{\langle x, x \rangle}$$

Per semplificare la notazione possiamo considerare la seguente matrice

$$C_{c_0} = \begin{pmatrix} c_1^H - c_0^H \\ \vdots \\ c_m^H - c_0^H \end{pmatrix}$$

Di conseguenza

$$D(x, c_0; c_1, \dots, c_m) = \frac{\|C_{c_0}x\|_2^2}{\|x\|_2^2}$$

**Definizione 6.** Dati  $c_1, \dots, c_m \in \mathbb{C}^{n+1}$  definiamo il centroide come

$$\bar{c} = \frac{1}{m} \sum_{i=1}^m c_i$$

**Lemma 9.1.** Per ogni vettore normale (all'iperpiano)  $x \in \mathbb{C}^{n+1}$ , per ogni punto  $c_0 \in \mathbb{C}^{n+1}$  e per ogni insieme di punti  $c_1, \dots, c_m \in \mathbb{C}^{n+1}$  vale

$$D(x, c_0; c_1, \dots, c_m) \geq D(x, \bar{c}; c_1, \dots, c_m)$$

di conseguenza l'iperpiano di cui vogliamo fare il fitting deve passare attraverso il centroide.

*dimostrazione:* Consideriamo il vettore  $w := C_{c_0}x$ , quindi abbiamo che  $w_i = \langle c_i - c_0, x \rangle$  e quindi

$$D(x, c_0; c_1, \dots, c_m) = \frac{\|w\|_2^2}{\|x\|_2^2}$$

Consideriamo il vettore  $z := C_{\bar{c}}x$ , quindi abbiamo che  $z_i = \langle c_i - \bar{c}, x \rangle$  e quindi

$$D(x, \bar{c}; c_1, \dots, c_m) = \frac{\|z\|_2^2}{\|x\|_2^2}$$

Definiamo per comodità  $1 := (1, \dots, 1) \in \mathbb{C}^m$  e  $h = \langle x, \bar{c} - c_0 \rangle$  quindi con questa notazione possiamo esprimere

$$w = z + h1$$

Osserviamo che  $z \perp 1$  infatti

$$\langle z, 1 \rangle = 1^H (C_{\bar{c}}x) = (1^H C_{\bar{c}})x = \left( m\bar{c}^H - \sum_{j=1}^m c_j^H \right) x = 0 * x = 0$$

Usando il fatto che  $z \perp 1$  possiamo concludere usando il teorema di pitagora, quindi da  $w = z + h1$  abbiamo

$$\begin{aligned} D(x, c_0; c_1, \dots, c_m) &= \|w\|_2^2 / \|x\|_2^2 = (\|z\|_2^2 + h^2 \|1\|_2^2) / \|x\|_2^2 = \\ &= D(x, \bar{c}; c_1, \dots, c_m) + h^2 m / \|x\|_2^2 \\ &\geq D(x, \bar{c}; c_1, \dots, c_m) \end{aligned}$$

L'uguaglianza si ha solo se  $h = 0$  ovvero  $\langle x, c - c_0 \rangle = 0$  ovvero se  $c$  appartiene all'iperpiano passante per  $\bar{c}$  e avente come vettore normale  $x$ .  $\square$

All'inizio di questa sezione abbiamo detto che per individuare un iperpiano ci è sufficiente conoscere un suo punto e il suo vettore normale, il punto lo abbiamo trovato con il lemma precedente, con il lemma che segue calcoliamo il vettore normale, risulterà che tale vettore ottimale (che minimizza la solita funzione) sarà il vettore singolare destro rispetto al valore singolare più piccolo di  $C_{c_0}$

**Lemma 9.2.** Dato un insieme di punti  $c_0, c_1, \dots, c_m \in \mathbb{C}^{n+1}$ , sia  $v$  un vettore singolare destro della matrice  $C_{c_0}$  relativo al più piccolo valore singolare  $\sigma$ . Allora vale che per ogni vettore non nullo  $x \in \mathbb{C}^{n+1}$

$$D(x, c_0; c_1, \dots, c_m) \geq D(v, c_0; c_1, \dots, c_m)$$

Con l'uguaglianza se e soltanto se  $x$  è pure un vettore singolare destro relativo al più piccolo valore singolare della matrice. In particolare vale che

$$D(v, c_0; c_1, \dots, c_m) = \sigma^2$$

*dimostrazione.* Consideriamo la solita funzione da minimizzare  $D(x, c_0, c_1, \dots, c_m) = \|C_{c_0}x\|_2^2/\|x\|_2^2$ , segue che  $D$  è minima se prendiamo un vettore unitario  $v = x/\|x\|_2$  che minimizza  $\|C_{c_0}v\|_2$ , dalla teoria vista sulle fattorizzazioni SVD segue che questo deve coincidere con un vettore singolare destro relativo al più piccolo valore singolare della matrice e si avrà  $D(v, c_0; c_1, \dots, c_m) = \|C_{c_0}v\|_2^2 = \sigma^2$   $\square$

Da questi due lemmi segue immediatamente il seguente

**Teorema 9.3.** *Dati  $c_1, \dots, c_m \in \mathbb{C}^{n+1}$  allora per risolvere il problema totale ai minimi quadrati è sufficiente determinare il centroide  $\bar{c}$  dei punti e calcolare un vettore singolare destro della matrice  $C_{\bar{c}}$ , quindi l'iperpiano centrato avrà  $v$  come vettore normale e passerà per  $\bar{c}$  (è dunque univocamente determinato) e vale che la somma dei quadrati delle distanze dei punti dall'iperpiano è  $\sigma^2$*

*dimostrazione.* Ovvvia conseguenza dei due lemmi precedenti  $\square$

Se la matrice  $C_{\bar{c}}$  ha il valore singolare più piccolo ripetuto, ovvero  $\sigma = \sigma_{k+1} = \dots = \sigma_{k+l}$  allora si avrà un sottospazio  $V_\sigma$  generato dai corrispondenti valori singolari  $v_{k+1}, \dots, v_{k+l}$ , in questa situazione esiste una famiglia  $\mathcal{H}$  di iperpiani che soddisfano il problema totale ai minimi quadrati, in particolare un iperpiano che soddisfa il problema  $H \in \mathcal{H}$  avrà come vettore normale un certo  $v \in V_\sigma$  e conterrà tutti gli assi  $\bar{c} + V_\sigma^\perp$ . In particolare nel caso degenerare in cui  $\sigma = 0$  i punti  $c_1, \dots, c_m$  apparterranno all'insieme  $\cap \{H \text{ tale che } H \in \mathcal{H}\}$ .

## 9.2 Formulazione algebrica

La formulazione algebrica del problema totale ai minimi quadrati è la seguente: dato il sistema  $Ax = b$  si cerca la matrice  $\tilde{A}$  e il vettore  $\tilde{b}$  che minimizzano la norma di Frobenius

$$\|[\tilde{A}; \tilde{b}] - [A; b]\|_F$$

soggetto alla condizione che il sistema  $\tilde{A}\tilde{x} = \tilde{b}$  abbia soluzione (intesa in senso classico). Per la teoria vista siamo già in grado di risolvere questo problema, infatti posto

$$C = [A; b]$$

$$\tilde{C} = [\tilde{A}; \tilde{b}]$$

il problema diventa quello di determinare la matrice  $\tilde{C}$  che minimizza  $\|\tilde{C} - C\|_F$  soggetto alla condizione che il nucleo di  $\tilde{C}$  contenga un vettore della forma  $(x^H, -1)^H$ . Per il teorema visto prima (approssimazione con matrici di rango inferiore) possiamo scegliere

$$\tilde{C} = \sum_{i=1}^n \sigma_i u_i u_i^* = [A; b] - \sigma_{n+1} u_{n+1} v_{n+1}^H$$

quindi  $v_{n+1}$  genera il nucleo di  $\tilde{C}$ . Ci sono ora due casi: se  $((v_{n+1})_{n+1}) \neq 0$  allora il problema ammette soluzione

$$\begin{pmatrix} \tilde{x} \\ -1 \end{pmatrix} = \frac{-1}{(v_{n+1})_{n+1}} v_{n+1}$$

Se invece  $(v_{n+1})_{n+1} = 0$  allora bisogna procedere con un'approssimazione di rango ancora più piccolo.

A volte nelle applicazioni capita di dover considerare esatte le prime  $k$  colonne della matrice  $A$  e soggette a correzione le altre  $n - k$ . Sia dunque

$$A_1 := (A(; 1), \dots, A(; k))$$

$$A_2 := (A(; k + 1), \dots, A(; n))$$

se  $A_1$  ha rango  $l \leq r$  e se  $Q$  è la proiezione ortogonale nello spazio generato dalle colonne di  $A_1$  e  $Q^\perp$  è la proiezione ortogonale sul complemento ortogonale (dello spazio generato dalle colonne di  $A_1$ ) allora la matrice di rango  $r$  che minimizza la norma di Frobenius

$$\|[A_1; \tilde{A}_2] - [A_1; A_2]\|_F$$

è definita in termini della decomposizione ai valori singolari di  $Q^\perp A_2$

$$Q^\perp A_2 = \sum_{i=1}^{n-k} \tau_i w_i z_i^H$$

dalla formula

$$\tilde{A}_2 = Q A_2 + \sum_{i=1}^{r-l} \tau_i w_i z_i^H$$

Se applichiamo questo metodo per risolvere il problema del fitting dell'ellissoide

$$c_0 + c_1 [\sin(\lambda)]^2 = \Delta s$$

dove i dati li prendiamo dalla tabella mostrata prima e supponiamo i coefficienti  $c_0$  corretti allora facendo un confronto

	TLS con prima colonna fissata	OLS	stima attuale
$e^2$	0.010210	0.006339	0.00669437999013

## 10 Problema ai minimi quadrati non lineare e il problema aperto del fitting di una varietà

Concludo il seminario presentando un problema aperto, quello del fare il fitting di una varietà, quindi in questo caso si tratterà di risolvere il problema totale ai minimi quadrati non lineare. La formulazione di tale problema è identica alla formulazione geometrica del problema totale ai minimi quadrati sostituendo alla parola iperpiano la parola varietà. Per varietà limitiamoci agli esempi più semplici ad esempio i luoghi di zeri di funzioni  $C^1$  o se si vuole semplificare possiamo limitarci a considerare luoghi di zeri di polinomi (infatti i polinomi approssimano le funzioni continue). Tanto per fare un esempio un ellissoide è una varietà intesa come luogo di zeri di un polinomio (in questo caso di più variabili). Quindi in generale abbiamo  $c_1, \dots, c_m \in \mathbb{R}^n$  e una classe di varietà  $\mathcal{V}$  (ad esempio tutti gli ellissoidi), fare il fitting di una varietà vuol dire trovare la varietà  $V \in \mathcal{V}$  che minimizza la somma dei quadrati delle distanze dei punti

dalla varietà. Osserviamo che per determinare la distanza punto-varietà intesa come

$$d(x, V) = \min \{d(x, y) \text{ tale che } y \in V\}$$

non è scontato come nel caso lineare, daltronde possiamo ulteriormente semplificare il tutto e supporre che i vettori normali della varietà non si intersechino tra di loro (il che equivale a chiedere l'invertibilità della mappa di Gauss), se siamo in questa ipotesi possiamo ripetere esattamente il ragionamento fatto nel caso lineare e la distanza di un punto dalla varietà si calcola trovando il vettore normale  $n$  alla varietà che passa per  $x$  e in tal caso

$$d(x, V) = \min \{d(x, y) \text{ tale che } y \in n \cap V\}$$

dove in genere  $n \cap V$  è un insieme finito. Un esempio pratico è un'ellisse in  $\mathbb{R}^2$ , tutte le condizioni sono soddisfatte e  $|n \cap V| = 2$ , non è difficile trovare che per determinare la distanza di un punto da un'ellisse è necessario trovare gli zeri di un polinomio di quarto grado. Non è neanche difficile trovare casi ancora più degeneri dove per trovare la distanza punto varietà è necessario risolvere problemi mal condizionati come trovare gli zeri di un polinomio di grado elevato. Pertanto è già un problema determinare la distanza punto varietà. In generale non ci sono ancora teoremi che garantiscono convergenza o buone tecniche di fitting per varietà pertanto questo resta ancora un problema aperto.