CAPITOLO 3

PROBLEMA DEI MINIMI QUADRATI

Il problema dei minimi quadrati lineare consiste sostanzialmente nella risoluzione del seguente sistema lineare sopradimensionato, cioè con più equazioni che incognite:

$$Ax=b$$

dove $A\in\mathcal{L}(R^n,R^m)$, $x\in R^n$, $b\in R^m$ e $m\geq n$.

Conviene includere il caso m=n perche alcuni risultati di questo capitolo, applicati alle matrici quadrate, saranno nuovi rispetto a quanto esposto nel capitolo 6 relativo ai sistemi lineari.

Appare evidente che un tale sistema non può avere soluzione per ogni vettore $b\in R^m$. D'altra parte è altrettanto evidente che per qualche vettore b ci possano essere una o più soluzioni. Basta infatti ricordare che, per ogni x, il prodotto Ax ($\in R^m$) è una combinazione delle colonne della matrice A e quindi appartiene al **range di A**, che sarà indicato con $\mathcal{R}(A)$ (sottospazio di R^m generato dalle n colonne di A). La dimensione di $\mathcal{R}(A)$ è in generale $\leq n$, mentre risulta uguale ad n se e solo se le colonne di A sono linearmente indipendenti, cioè se la matrice A è di rango massimo. Di conseguenza la soluzione esiste se e soltanto se $b\in\mathcal{R}(A)$ e, nel caso che esista, sarà unica se e soltanto se A ha rango massimo.

L'argomento di questo capitolo sarà lo studio del caso $b\notin\mathcal{R}(A)$. Un esempio tipico di problema che conduce ad un sistema sovradimensionato è l'approssimazione di dati sperimentali.

Approssimazione di dati sperimentali ed interpolazione sopradimensionata

Supponiamo di sapere che un fenomeno (fisico, naturale, sociale,...) si esprime attraverso una funzione y=f(t) rappresentabile come combinazione lineare di certe funzioni di base $f(t)=a_1u_1(t)+...+a_nu_n(t)$. Per esempio la posizione s di un corpo in funzione del tempo t in un moto rettilineo uniforme si esprime con un polinomio algebrico di primo grado

$$s=p_1(t)=s_0+vt,$$

dove s_0 è la posizione al tempo 0 e v è la velocità. Nel caso invece di un moto uniformemente accelerato la posizione s si esprime con un polinomio di secondo grado

$$s=p_2(t)=s_0+v_0t+\frac{at^2}{2}$$

dove s_0 e v_0 sono, rispettivamente, la posizione e la velocità al tempo 0 ed *a* e' l'accelerazione. In quest'ultimo caso conoscendo la posizione e la velocità iniziale s_0 e v_0 nonchè l'accelerazione a, si può determinare la posizione del corpo ad ogni istante t. Viceversa se non sono noti i parametri s_0, v_0 ed a, essi possono essere determinati attraverso il rilevamento sperimentale delle posizioni s_1, s_2 ed s_3 assunte dal corpo in tre istanti diversi t_1 ,t_2 e t_3 imponendo le seguenti *condizioni interpolatorie*:

$$s_1=s_0+v_0t_1+\frac{at_1^2}{2}$$

$$s_2=s_0+v_0t_2+\frac{at_2^2}{2}$$

$$s_3=s_0+v_0t_3+\frac{at_3^2}{2}$$

Come si vedrà in un successivo capitolo, questo sistema ammette una unica soluzione (s_0, v_0, a) e, per tali valori, ogni altra posizione s_i rilevata sperimentalente al tempo t_i dovrebbe, in linea di principio, soddisfare l'equazione $s_i=s_0+v_0t_i+\frac{at_i^2}{2}$.

In altre parole, se eseguiamo m osservazioni sperimentali $s_1,...s_m$, agli istanti $t_1,...t_m$, esse dovrebbero soddisfare il sistema

$$s_1=s_0+v_0t_1+\frac{at_1^2}{2}$$

$$s_2=s_0+v_0t_2+\frac{at_2^2}{2}$$

.....

$$s_m=s_0+v_0t_m+\frac{at_m^2}{2}$$

In realtà, a causa dei possibili errori sperimentali, ciò non si verifica, il chè è come dire che gli m punti (t_i,s_i), i=1,...,m, del piano non stanno tutti sulla stessa parabola. Ogni sottosistema fatto da tre di quelle equazioni ammette una soluzione, ma non esiste una soluzione del sistema completo. Avendo pari fiducia nelle osservazioni eseguite, ci si pone

il problema di come individuare i parametri (s_0, v_0, a) tenendo conto di tutti i dati sperimentali a disposizione cioè: come trovare una terna (s_0, v_0, a) che possa essere vista come "soluzione" dell'intero sistema sopradimensionato costituito da m (>3) equazioni e 3 incognite.

Chiamando r il **vettore residuo** di componenti:

$$r_1=s_1-(s_0+v_0t_1+\frac{at_1^2}{2})$$

$$r_2=s_2-(s_0+v_0t_2+\frac{at_2^2}{2})$$

.....

$$r_m=s_m-(s_0+v_0t_m+\frac{at_m^2}{2});$$

si chiama **soluzione ai minimi quadrati** la terna (s_0, v_0, a) che minimizza la norma euclidea del residuo o, equivalentemente, il suo quadrato $||r||_2^2$. Dimostreremo che tale soluzione esiste sempre e stabiliremo sotto quale condizione essa è unica.

Nell'esempio precedente il modello quadratico del fenomeno era assunto come dato certo. Era l'incertezza sui dati forniti dalle osservazioni che suggeriva l'opportunità di eseguire un numero sovrabbondante di esperimenti e calcolare la soluzione ai minimi quadrati. Il seguente esempio, formalmente analogo ma sostanzialmente diverso, si riferisce al caso in cui l'incertezza risiede nel modello da adottare per rappresentare il fenomeno y=f(x) del quale supponiamo invece di conoscere delle valutazioni certe $y_i=f(x_i)$ i=1,...,m.

Il censimento della popolazione negli Stati Uniti dal 1900 al 1970 ha fornito i seguenti dati

1900	abitanti:	75.994.575
1910		91.972.266
1920		105.710.620
1930		122.775.046
1940		131.669.275
1950		150.697.361
1960		179.323.175
1970		203.235.298

Da questi dati vogliamo avere delle valutazioni sulla popolazione negli anni intermedi ai censimenti e, almeno nel medio termine, una previsione sull'andamento demografico successivamente al 1970. Non c'è alcuna legge basata su argomentazioni di tipo sociologico od economico che indichi rigorosamente quale tipo di funzione è la più adatta ad approssimare i nostri dati. Possiamo solo prevedere una generica crescita della popolazione e decidiamo quindi di approssimare i nostri dati con una parabola o una cubica. Anche in questo caso imponendo tutti i vincoli interpolatori si perviene ad un sistema sopradimensionato di 7 equazioni in 3 o 4 incognite (i coefficienti della parabola o della cubica) per il quale si può adottare come soluzione quella ai minimi quadrati.

Non è questo il capitolo nel quale indagare sulla "bontà" del modello adottato (parabola, cubica o altro ancora) e delle rispettive soluzioni fornite attraverso il metodo dei minimi quadrati nei confronti dei dati da approssimare. In questo capitolo si vedrà che un generico sistema lineare sovradimensionato ammette sempre almeno una soluzione ai minimi quadrati. Si vedrà inoltre sotto quali condizioni essa è unica e come può essere calcolata nel modo più stabile.

Soluzione ai minimi quadrati .

Ribadiamo la definizione di soluzione ai minimi quadrati. Dato il sistema lineare

$$Ax=b$$

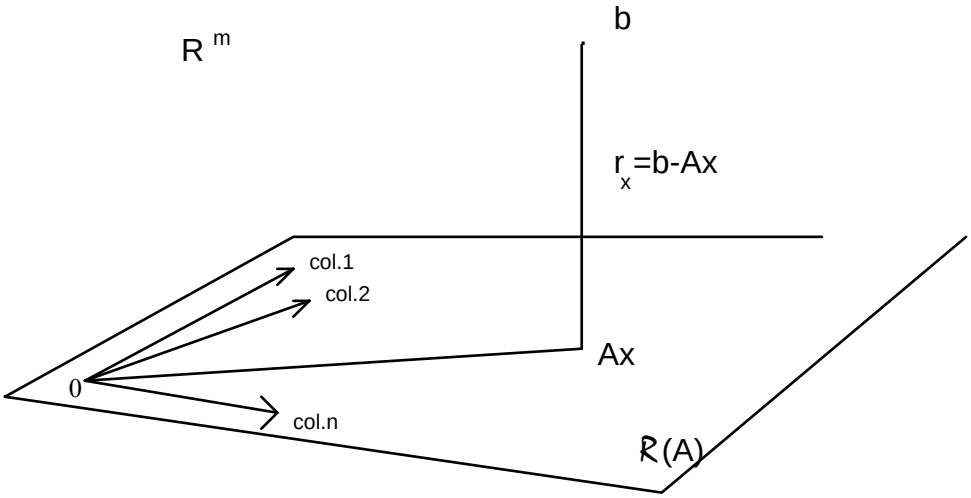
dove $A\in\mathcal{L}(R^n,R^m)$, $x\in R^n$, $b\in R^m$ e $m\geq n$, si definisce *soluzione ai minimi quadrati* il vettore x tale che

$$||Ax-b||_2^2 \leq ||Ay-b||_2^2 \quad \forall y\in R^n$$

o, se si preferisce, tale che il corrispondente vettore $Ax\in\mathcal{R}(A)$ sia di minima distanza da b tra tutti i vettori Ay di $\mathcal{R}(A)$. Il vettore Ax è detto **elemento di minimo scarto**.

Come abbiamo già detto all'inizio è interessante il caso $b\notin\mathcal{R}(A)$ per il quale il sistema dato non ammette soluzione esatta.

Poichè la norma euclidea è dedotta da un prodotto scalare che a sua volta definisce una nozione di ortogonalità in R^m , ($u\perp v \Leftrightarrow u^Tv=0$) è intuitivo che l'elemento di minimo scarto da b in $\mathcal{R}(A)$ è la proiezione ortogonale di b su $\mathcal{R}(A)$, cioè l'elemento Ax tale che il vettore Ax-b è ortogonale ad $\mathcal{R}(A)$.



La condizione di ortogonalità si esprime attraverso il sistema:

$$A^T(Ax-b)=0$$

oppure

$$A^TAx=A^Tb \qquad \text{con} \qquad A^TA\in\mathcal{L}(R^n,R^n)\ , \quad x\in R^n,\ A^Tb\in R^n$$

detto **sistema di equazioni normali**.

La precedente asserzione, formulata intuitivamente, è dimostrata nel seguente teorema.

Teorema 4.1. *Supponiamo che esista un vettore* $x\in R^n$ *tale che*:

$$A^T(Ax-b)=0.$$

Allora :

$$||Ax-b||_2^2 \le ||Ay-b||_2^2 \quad \forall y\in R^n.$$

Dim. Sia r_y: $=Ay-b = Ay-Ax+Ax-b = A(y-x)+r_x$. Passando alle norme si ha:

$$||r_y||_2^2 = ||A(y-x) + r_x ||_2^2 = (A(y-x) + r_x)^T(A(y-x) + r_x) =$$

$$||A(y-x)||_2^2 + r_x^TA(y-x) + (A(y-x))^Tr_x + ||r_x||_2^2=$$

$$||A(y-x)||_2^2 + ||r_x||_2^2 \ge ||r_x||_2^2.$$

Quindi se esiste una soluzione del sistema di equazioni normali, allora x è una soluzione dell'equazione Ax-b=0 *nel senso dei minimi quadrati*. Per quanto riguarda la risolubilità del sistema di equazioni normali, vale il seguente teorema:

Teorema 4.2. *La matrice* A^TA *è non singolare se e solo se le colonne di* A *sono linearmente indipendenti (cioè se* A *ha rango massimo)*.

Dim. Supponiamo che le colonne di A siano linearmente dipendenti; ciò significa che $\exists\ x \neq 0$ tale che Ax=0 ed anche $A^TAx=0$. In tal caso A^TA sarebbe singolare. Viceversa, supponiamo che A^TA sia singolare. Allora $\exists\ x \neq 0$ tale che $A^TAx=0$ ed anche $x^TA^TAx=0$. Per le proprietà del prodotto scalare sarebbe Ax=0 con $x \neq 0$, da cui si deduce che le colonne di A sono linearmente dipendenti.

Quindi se A è di rango massimo, esiste ed è unica la soluzione ai minimi quadrati per ogni vettore b. Possiamo ricapitolare dicendo che:

Per ogni matrice A di rango massimo la soluzione ai minimi quadrati esiste sempre ed è unica. Essa si ottiene risolvendo il sistema di equazioni normali $A^TAx=A^Tb$ *.*

Per quanto riguarda il condizionamento del problema dei minimi quadrati, accontentiamoci di analizzare il condizionamento della matrice A^TA del sistema di equazioni normali che dobbiamo risolvere. Detti $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n \geq 0$ i suoi autovalori, si definiscono **valori singolari di A** i numeri reali

$$\sigma_i = \sqrt{\lambda_i} \qquad i=1,...,n$$

per i quali si ha ancora

$$\sigma_1 \geq \sigma_2 \geq ... \sigma_n \geq 0$$

Il condizionamento di A^TA è quindi:

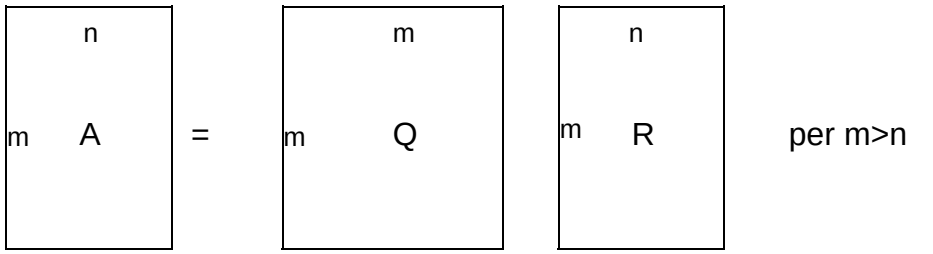
$$K_2(A^TA) = \begin{cases} \dfrac{\sigma_1^2}{\sigma_n^2} & \text{se} \quad \sigma_n \neq 0 \\ \infty & \text{se} \quad \sigma_n = 0 \end{cases}$$

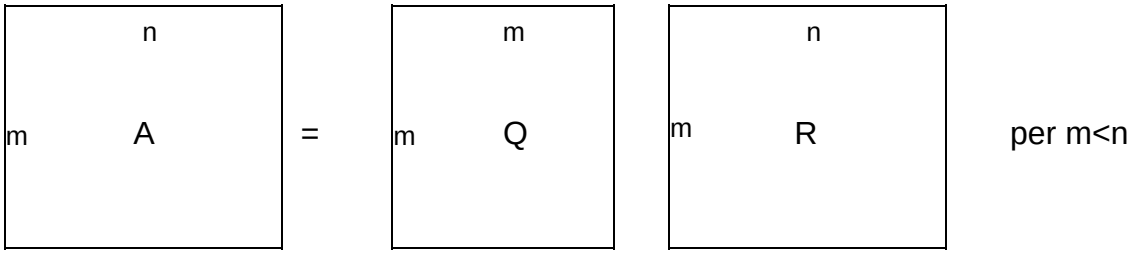
<u>La fattorizzazione QR.</u>

La fattorizzazione che ora verrà descritta vale per una matrice di dimensione *m x n* qualunque. Sia data dunque una matrice $A \in \mathcal{L}(C^n, C^m)$ le cui colonne verranno indicate con $a_1, a_2, ..., a_n$. Generalizziamo la definizione di matrice triangolare superiore al caso $m \neq n$ chiamando ancora elementi diagonali di A gli elementi $a_{11}, a_{22}, ..., a_{rr}$, r=min (m,n), e dicendo che A è triangolare superiore se tutti gli elementi sotto la diagonale sono nulli. Esse avranno la seguente forma:

$$\begin{Vmatrix} a_{11} & . & . & . & a_{1n} \\ & a_{22} & & & . \\ & & . & & . \\ & & & . & . \\ & & & & a_{nn} \end{Vmatrix} \text{ per } m>n, \qquad \begin{Vmatrix} a_{11} & . & . & . & . & . & a_{1n} \\ & . & & & & & . \\ & & . & & & & . \\ & & & . & & & . \\ & & & & a_{mm} & . & a_{mn} \end{Vmatrix} \text{ per } m<n.$$

Teorema 4.3. *Per ogni matrice* $A \in \mathcal{L}(C^n, C^m)$ *esiste una matrice unitaria* $Q \in \mathcal{L}(C^m, C^m)$ *ed una matrice triangolare superiore* $R \in \mathcal{L}(C^n, C^m)$ *tali che* $A=QR$.





Risoluzione stabile del problema dei minimi quadrati.

Al punto precedente abbiamo visto che la matrice $A \in \mathcal{L}(R^n, R^m)$ si può fattorizzare in $A=QR$ con $Q(\in \mathcal{L}(R^m, R^m))$ ortogonale, e R $(\in \mathcal{L}(R^n, R^m))$ triangolare superiore. Il sistema $Ax=b$ da risolvere nel senso dei minimi quadrati si può dunque scrivere come:

$$QRx=b$$

Siccome le trasformazioni ortogonali non cambiano la norma euclidea, è equivalente minimizzare $||QRx-b||_2^2$ oppure $||Q^T(QRx-b)||_2^2=||Rx-Q^Tb||_2^2$. Le componenti del vettore residuo $Rx-Q^Tb$ sono:

$$\begin{aligned} r_{11}x_1+\ & r_{12}x_2+ && + r_{1n}x_n - b'_1 \\ & r_{22}x_2+ && + r_{2n}x_n - b'_2 \\ & && + r_{nn}x_n - b'_n \\ & && - b'_{n+1} \\ & && - b'_m \end{aligned}$$

dove le ultime m-n componenti non dipendono da x, ed è evidente che la sua norma sarà minima se è minima la norma delle sue prime n componenti. Nel caso $r_{ii}\neq 0$ per ogni i, ciò si ottiene risolvendo il sistema triangolare:

$$\begin{aligned} r_{11}x_1+\ & r_{12}x_2+ && r_{1n}x_n = b'_1 \\ & r_{22}x_2+ && r_{2n}x_n = b'_2 \\ & && r_{nn}x_n = b'_n \end{aligned}$$

la cui matrice, appartenente a $\mathcal{L}(R^n,R^n)$, chiameremo ora con R'.

Osserviamo ora che la matrice R' ha un condizionamento migliore della matrice A^TA per cui sarà sempre preferibile passare attraverso la fattorizzazione QR e risolvere il precedente sistema triangolare piuttosto che risolvere direttamente il sistema di equazioni normali. Infatti, poichè

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

si ha

$$K_2(A^TA)=K_2(R^TR).$$

D'altra parte

$$R^TR=R'^TR'$$

e, come abbiamo già visto precedentemente per le matrici quadrate,

$$K_2(R'^TR')=K_2^2(R'),$$

per cui

$$K_2(A^TA)=K_2^2(R').$$

Essendo l'indice di condizionamento di qualsiasi matrice un numero ≥ 1, risulta

$$K_2(A^TA)\geq K_2(R') .$$

Esempio

Consideriamo i dati esposti all'inizio del capitolo sulla popolazione degli Stati Uniti negli anni dal 1900 al 1970 ed interpoliamoli con un polinomio di secondo grado in forma canonica:

$$p(t)=c_1+c_2t+c_3t^2.$$

La matrice A del sistema sopradimensionato (8×3) ha i seguenti valori singolari:

$$\sigma_1 =0.106\ \ 10^8 \qquad \sigma_2 =0.648\ \ 10^2 \qquad \sigma_3 =0.346\ \ 10^{-3}$$

e quindi la matrice A^TA del sistema di equazioni normali ha un indice di condizionamento molto elevato :

$$K(A^TA)=0.1\ 10^{22}.$$

Escludiamo dunque l'uso del sistema normale e applichiamo la tecnica della fattorizzazione QR per la quale l'indice di condizionamento è

$$\sigma_1 /\sigma_n=0.306\ \ 10^{11}.$$

Essa fornisce i seguenti valori dei coefficienti della parabola:

in semplice precisione (8 cifre significative):

$$c_1 = - 0.372\ \ 10^5 \qquad c_2 =0.368\ \ 10^2 \qquad c_3 = - 0.905\ \ 10^{-2}$$

con i quali si ha il valore estrapolato: p(1980)=145.21 milioni,
ed in doppia precisione (16 cifre significative):

$$c_1 = 0.375\ 10^5 \qquad c_2 = -\ 0.402\ 10^2 \qquad c_3 = 0.108\ 10^{-1}$$

per i quali si ottiene p(1980)= 227.78 milioni.

I due valori differiscono tra loro in modo consistente e quindi i risultati ottenuti in semplice precisione sono inaccettabili (vedremo che quello ottenuto in doppia precisione è corretto). Anche utilizzando la fattorizzazione QR, l'indice di condizionamento σ_1 / σ_n è ancora troppo grande per la semplice precisione.

Un modo per abbassare ulteriormente l'indice di condizionamento consiste nell'adottare rappresentazioni alternative per il polinomio approssimante.

Si considerino le espressioni:

$$p(t)=c_1+c_2(t-1900)+c_3(t-1900)^2.$$

e

$$p(t)=c_1+c_2\left(\frac{t-1935}{10}\right)+c_3\left(\frac{t-1935}{10}\right)^2.$$

e si calcoli, per entrambe, l'indice di condizionamento della matrice A^TA la soluzione ai minimi quadrati. Per entrambi i casi si troverà la stima p(1980)= 227.78 milioni che risulta quindi accettabile.

La decomposizione in valori singolari: SVD.

Si dimostra che per ogni matrice $A\in\mathcal{L}(R^n,R^m)$ esiste una matrice ortogonale $U\in\mathcal{L}(R^m,R^m)$, una matrice ortogonale $V \in\mathcal{L}(R^n,R^n)$ ed una matrice diagonale $\Sigma\in\mathcal{L}(R^n,R^m)$ con i valori singolari $\sigma_1\geq\sigma_2\geq...\sigma_n$ sulla diagonale, tale che $A=U\Sigma V^T$:

$$A = U \begin{Vmatrix} \sigma_1 & & & & \\ & \sigma_2 & & 0 & \\ & & .. & & \\ & & & .. & \\ & & & & \sigma_n \\ & 0 & & & \end{Vmatrix} V^T$$

Tale decomposizione di A, detta **decomposizione in valori singolari** o più brevemente **SVD** (Singular Value Decomposition), trova molte applicazioni tra le quali il calcolo della soluzione di norma minima per il problema dei minimi quadrati.

Poichè U è ortogonale, si ha:

$$||Ax-b||_2^2 = ||U\Sigma V^T x-b||_2^2 = ||U^T(U\Sigma V^T x-b)||_2^2 = ||\Sigma V^T x-U^T b||_2^2.$$

Con il cambio di variabile $V^T x = z$ e $U^T b = d$, si ha infine:

$$||Ax-b||_2^2 = ||\Sigma z-d||_2^2$$

dove

$$||\Sigma z-d||_2^2 = (\sigma_1 z_1-d_1)^2 +... + (\sigma_n z_n-d_n)^2 + d_{n+1}^2+...+d_m^2.$$

E' evidente che la soluzione z che minimizza $||\Sigma z-d||_2^2$ è data, nel caso che i valori singolari siano tutti non nulli, da $z_1 = \frac{d_1}{\sigma_1}$,..., $z_n = \frac{d_n}{\sigma_n}$; in tal caso il residuo è $d_{n+1}^2+...+d_m^2$.. Se viceversa qualche valore singolare è nullo , diciamo $\sigma_i = 0$, allora z_i è arbitrario ed il residuo è $d_i^2+d_{n+1}^2+...+d_m^2$ per ogni valore di z_i. Le altre componenti di z sono indipendenti dalla scelta di z_i e quindi il vettore z di norma minima si ottiene per $z_i = 0$. Per ogni z, il corrispondente vettore x=Vz conserva la norma di z e quindi x è la soluzione di norma minima per il problema originale Ax=b.

Si osservi che l'utilizzo della SVD per il calcolo della soluzione ai minimi quadrati richiede la risoluzione di un sistema lineare la cui matrice è Σ il cui indice di condizionamento, σ_1 / σ_n , è uguale a quello della matrice R'. Se σ_n è molto piccolo l'indice di condizionamento può risultare ancora troppo alto e fornire una soluzione inaccettabile. Può risultare più stabile porre $\sigma_n = 0$ e calcolare la soluzione di norma minima. Infatti ciò comporta la risoluzione di un sistema di dimensione n-1 ottenuto dal precedente togliendo l'ultima riga e l'ultima colonna. Il suo indice di condizionamento

risulta allora σ_1/σ_{n-1}. Il miglioramento dell'indice di condizionamento può compensare la perturbazione introdotta dalla soppressione dell'ultimo valore singolare.
Esempio:

Riprendiamo l'esempio sulla popolazione degli Stati Uniti negli anni dal 1900 al 1970 e consideriamo la SVD dove i valori singolari sono ancora

$\sigma_1 = 0.106\ 10^8$ $\sigma_2 = 0.648\ 10^2$ $\sigma_3 = 0.346\ 10^{-3}$

e l'indice di condizionamento è $\sigma_1/\sigma_n = 0.306\ 10^{11}$.

I valori dei coefficienti c_1, c_2, c_3 che si ottengono in semplice precisione sono gli stessi del metodo QR ed abbiamo osservato che, a causa dell'alto valore dell'indice di condizionamento, non sono accettabili. In particolare si otteneva p(1980)=145.21 milioni.
Proviamo allora a porre $\sigma_3=0$ e calcoliamo la soluzione di norma minima. Si ottiene:

in semplice precisione:

$c_1 = -0.166\ 10^{-2}$ $c_2 = -0.162\ 10^1$ $c_3 = -0.869\ 10^{-3}$

con i quali si ha il valore estrapolato: p(1980)=214.96 milioni,

ed in doppia precisione:

$c_1 = -0.167\ 10^{-2}$ $c_2 = -0.162\ 10^1$ $c_3 = -0.871\ 10^{-3}$

con i quali si ha il valore estrapolato: p(1980)=212.91 milioni.
Anche la semplice precisione fornisce questa volta un valore "accettabile", di gran lunga migliore di quello ottenuto tenendo conto di tutti i valori singolari.

Un'altra applicazione interessante della SVD è la compressione dei dati. Per n=m=500, la matrice A è costituita da 250.000 coefficienti. Supponiamo che essi abbiano valori compresi tra 0 ed 1 e rappresentino, in modo crescente, le varie tonalità di grigio comprese tra il bianco (=0) ed il nero (=1) in un fotogramma quadrato costituito, appunto, da 250.000 punti. Supponiamo che un satellite esegua una sequenza, a distanza molto ravvicinata, di tali fotogrammi e che li debba inviare a terra. A volte è comodo, a costo di una perdita di precisione dell'immagine, poter ridurre la massa di dati da trasmettere. Ciò può essere fatto attraverso la SVD nel seguente modo.

Si osservi che indicando con u_i e con v_i le colonne di U e V rispettivamente, si ha:

$$A = U\Sigma V^T = \sum_{i=1}^{n} \sigma_i u_i v_i^T \ .$$

cioè la matrice A è data dalla somma di n matrici di rango 1. Se $\sigma_i \approx 0$ per i>r, allora si può troncare la precedente somma ed approssimare A con

$$A \approx \sum_{i=1}^{r} \sigma_i u_i v_i^T .$$

Nel nostro esempio, se potessimo accontentarci dei primi 20 termini, sarebbe sufficiente inviare a terra 20 colonne di U e di V e 20 valori singolari: in tutto 20020 dati. (vedi esempio numerico della cartella FINGPRINT))