Università del Salento

Appunti di Metodi Statistici e Computazionali

per gli studenti della Laurea Triennale in Fisica

F.Strafella
Dipartimento di Matematica e Fisica

Indice

Capitolo 1

Operazioni matematiche di base

In estrema sintesi possiamo indicare tre procedure numeriche che sono centrali per la soluzione di moltissimi problemi matematici connessi al calcolo di modelli elaborati per descrivere i più disparati sistemi fisici.
Queste sono la **differenziazione**, la **quadratura** e la **ricerca di radici** e corrispondono ad eseguire in modo numerico le operazioni che nei corsi di analisi matematica abbiamo conosciuto con i nomi rispettivamente di: derivate, integrali, ricerca degli zeri per una data funzione.
Supponiamo di avere quindi una $f(x)$: con la differenziazione vogliamo stimare il valore della derivata ad un dato valore della x, con la quadratura vogliamo eseguire l'operazione inversa che corrisponde a calcolare l'integrale in un intervallo definito della x, con la ricerca di zeri vogliamo cercare i valori della x per cui la nostra funzione va a zero.
Se la forma analitica della f è nota allora si possono quasi sempre ricavare espressioni esplicite per le derivate di f e per gli integrali definiti che permettono di calcolare facilmente le derivate in un generico punto e l'integrale definito tra due punti. Se però non conosciamo la forma analitica della nostra funzione ma abbiamo solo i suoi valori $f(x_i)$ in un certo numero di punti x_i siamo costretti ad usare formule approssimate per esprimere derivate ed integrali usando ciò che abbiamo e cioè i valori $f(x_i)$ della nostra funzione incognita in un numero discreto di punti x_i. Per quanto riguarda gli zeri di una funzione, nonsempre possono essere trovati analiticamente e pertanto in questo campo la ricerca di soluzioni numeriche è praticamente la soluzione migliore.
Il modo in cui cercheremo di raggiungere i notri obiettivi (derivare, integrare, trovare zeri) al calcolatore, si fonda sulla possibilità di approssimare la nostra f con una funzione semplice (del tipo di un polinomio di primo o secondo grado) su cui le nostre operazioni matematiche si possano eseguire facilmente.

1.1 Differenziazione

Supponiamo di voler calcolare la derivata nel punto x=0 della nostra funzione f della quale conosciamo il valore in una serie di punti x_i equamente distanziati intorno ad x=0 con spaziatura h come mostrato in Fig. 1.1. In notazione indicheremo i valori della funzione nei vari punti con:

$$f_i = f(x_n) \qquad \text{con} \qquad x_n = nh \ (n = 0, \pm 1, \pm 2, ...)$$

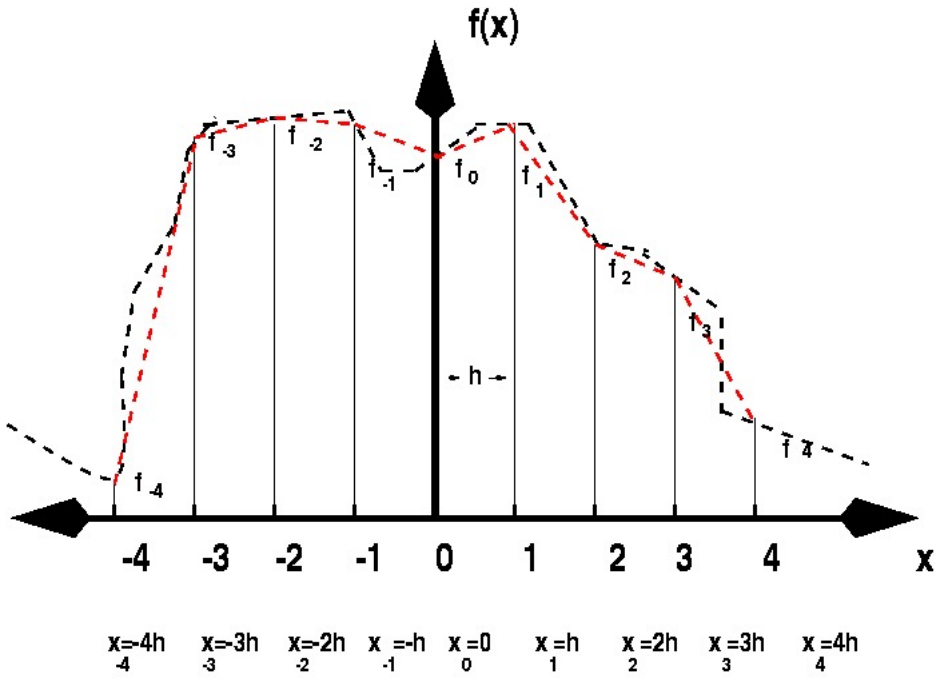

Figura 1.1: Il tratteggio indica: in nero la funzione incognita di cui vorremmo calcolare la derivata; in rosso la spezzata che connette i valori a noi noti della funzione nei punti x_n.

La Figura 1.1 illustra la situazione che ora andiamo a discutere iniziando con un'espansione in serie di Taylor della nostra f in un intorno di x_0:

$$f(x) = f(x_0) + (x - x_0)f'(x_0) + \frac{(x - x_0)^2}{2!}f''(x_0) + \frac{(x - x_0)^3}{3!}f'''(x_0) + ... \quad (1.1)$$

dove abbiamo usato gli apici per indicare il grado della derivata della funzione. Per semplificare la notazione scegliamo $x_0 = 0$, senza per questo togliere gener-

alità alla nostra discussione[1]:

$$f(x) = f_0 + xf' + \frac{x^2}{2!}f'' + \frac{x^3}{3!}f''' + ... \tag{1.2}$$

nella quale le derivate sono valutate nel punto $x = x_0 \equiv 0$ ed abbiamo adottato la notazione f_0 per indicare la $f(x_0)$. È facile ora verificare che se $x = \pm h$ valgono le seguenti due relazioni:

$$f_{\pm 1} \equiv f(x = \pm h) = f_0 \pm hf' + \frac{h^2}{2}f'' \pm \frac{h^3}{6}f''' + O(h^4), \tag{1.3}$$

$$f_{\pm 2} \equiv f(x = \pm 2h) = f_0 \pm 2hf' + 2h^2 f'' \pm \frac{4h^3}{3}f''' + O(h^4), \tag{1.4}$$

dove con $O(h^4)$, detto **errore di troncamento**, abbiamo indicato tutti i restanti termini della serie di Taylor di ordine superiore o uguale al quarto. Per convenzione diremo quindi che se l'errore di troncamento va come $O(h^{n+1})$, la formula di integrazione corrispondente sarà accurata all'ordine n o, equivalentemente, diremo che è accurata a meno di termini di ordine $n + 1$.

Se ora consideriamo le approssimazioni della funzione nei punti $+h$ e $-h$ e le sottraiamo, isolando la derivata prima otteniamo:

$$f' = \frac{f_1 - f_{-1}}{2h} - \frac{h^2}{6}f''' + O(h^4) \tag{1.5}$$

L'uso di questa espressione per la derivata prima richiede comunque la conoscenza della derivata terza e sarebbe impraticabile se non considerassimo che il termine in f''' va a zero come h^2 e quindi per h sufficientemente piccolo possiamo riscrivere la precedente come:

$$f' \simeq \frac{f_1 - f_{-1}}{2h} \tag{1.6}$$

Questa è la cosiddetta derivata *a tre punti* che corrisponde al valore esatto nel caso in cui la funzione f sia un polinomio di secondo grado per il quale la derivata terza diventa nulla ed il troncamento prima effettuato non ha quindi conseguenze numeriche. Nella relazione 1.6 è anche facile rivedere la definizione di derivata come limite di un rapporto incrementale per $h \to 0$. La Fig. 1.2 esemplifica la geometria usata per ricavare la darivata a tre punti.

[1]Si noti che una traslazione opportuna della f lungo l'asse delle x potrà portare qualsiasi punto sullo zero, senza per questo cambiare la derivata nel punto. Il ragionamento che segue vale quindi per ogni punto e non solo per $x = 0$.

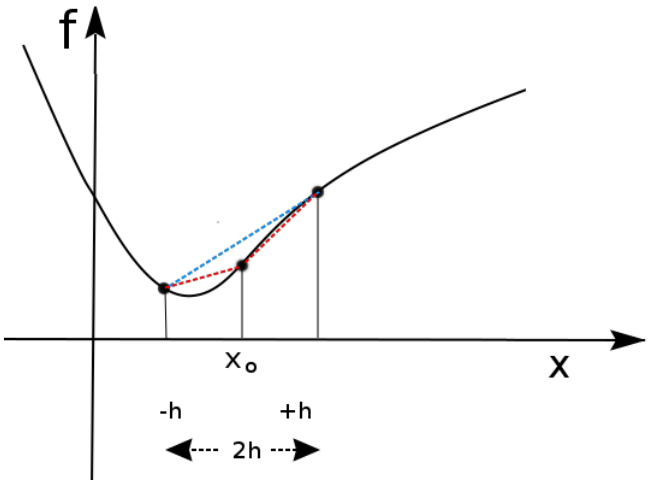

Figura 1.2: La derivata in x_0 di una generica $f(x)$ viene valutata numericamente conoscendo il valore della funzione in due punti adiacenti equispaziati ad una distanza h. L'integrale della stessa funzione nell'intervallo $[-h, +h]$ viene approssimato come somma delle aree dei due trapezi rettangoli delimitati dai segmenti tratteggiati in rosso. È evidente che che al diminuire di h i valori calcolati per la derivata e l'integrale saranno sempre più accurati.

Per calcolare derivate superiori si possono ancora sfruttare le relazioni (1.3) e (1.4). Per esempio, combinandole in modo opportuno, possiamo verificare che vale la

$$f_1 - 2f_0 + f_{-1} = h^2 f'' + O(h^4)$$

che ci permette poi di ricavare la derivata seconda f'', ottenendo:

$$f'' \simeq \frac{f_1 - 2f_0 + f_{-1}}{h^2} \tag{1.7}$$

1.2 Quadratura (integrali definiti)

Storicamente il termine **quadratura** è stato usato per indicare la ricerca del quadrato che avesse la stessa area di un dato cerchio. Nel calcolo numerico con il termine quadratura intendiamo il calcolo dell'integrale (e quindi dell'area sottostante) di una funzione $f(x)$ tra due limiti $x_a < x_b$ che definiscono l'*intervallo di integrazione*. Si tratta quindi del calcolo di un integrale definito. È evidente che, dato un intervallo di integrazione, potremo sempre suddividerlo in un numero N (pari) di sottointervalli di ampiezza h. Per semplificare la notazione useremo $a \equiv x_a$ e $b \equiv x_b$ ottenendo per il numero di intervalli

$$N = \frac{b-a}{h}$$

su ognuno dei quali ci porremo il problema della quadratura. Sarà sufficiente allora sviluppare una procedura di integrazione per un generico sottointervallo, diciamo da $-h$ a $+h$, per poi estenderla a tutti i sottointervalli, sommandone infine i singoli risultati. Il problema dall'integrazione si riduce così alla somma di tante integrazioni quanti sono i sottointervalli individuati. Esprimiamo allora la somma delle nostre integrazioni così :

$$\int_a^b f(x)dx = \int_a^{a+2h} f(x)dx + \int_{a+2h}^{a+4h} f(x)dx + ... + \int_{b-2h}^{b} f(x)dx \tag{1.8}$$

e cerchiamo poi di approssimare la $f(x)$ all'interno di ogni intervallo di integrazione con una funzione che si possa integrare esattamente. È chiaro che, fissato p.es. l'intervallo $[-h,\ +h]$ da considerare, il risultato dell'integrazione numerica sarà tanto più accurato quanto più i valori della funzione approssimante sono simili a quelli della nostra $f(x)$.

L'approssimazione più semplice è ovviamente quella lineare, che corrisponde a considerare la $f(x)$ in ogni sottointervallo approssimata da una linea retta congiungente i punti noti della $f(x)$ tra loro adiacenti. Nelle Fig. 1.1 ed 1.2 questa approssimazione è mostrata con il tratteggio rosso. Usando questa rappresentazione possiamo ora approssimare l'integrale della $f(x)$ nell'intervallo $[-h, +h]$ con la somma delle aree dei due trapezi adiacenti, ottenendo:

$$\int_{-h}^{+h} f(x)dx = \frac{h}{2}(f_{-1} + 2f_0 + f_1) + O(h^3) \tag{1.9}$$

che è poi la cosiddetta *regola trapezoidale* per il cacolo di integrali. Per valutarne l'accuratezza si consideri che usando un solo passo, e quindi un solo trapezoide, si avrebbe un'incertezza dell'ordine di $(b-a)^3$, per cui se dividiamo l'intervallo in sottointervalli di più piccola ampiezza h avremo una diminuzione dell'incertezza ed un risultato finale accurato a meno di termini dell'ordine di $O(h^3)$.

Sull'errore commesso integrando con un trapezio

L'errore commesso nel calcolo di un integrale è valutabile in generale come

differenza tra valore vero e valore calcolato. Nel caso della integrazione con un unico trapezio possiamo scrivere l'errore così :

$$E = \int_a^b f(x)dx - \left[\frac{f(a)+f(b)}{2}(b-a)\right] \tag{1.10}$$

dove il secondo termine rappresenta l'approssimazione usata per calcolare l'integrale (come l'area del rettangolo con base $(b-a)$ e altezza pari alla media dei valori della f negli estremi dell'intervallo). Se ora indichiamo con $\tilde{f}$ l'approssimazione lineare alla f tra i punti a e b possiamo anche esprimere l'errore commesso come:

$$\begin{aligned} E = & \int_a^b (f(x) - \tilde{f}(x))\ dx \\ = & \int_a^b \frac{f(x)-\tilde{f}(x)}{(x-a)(x-b)}(x-a)(x-b)\ dx \end{aligned} \tag{1.11}$$

Siccome notiamo che nell'intervallo di integrazione:

- il termine $(x-a)(x-b)$ non cambia di segno;
- la funzione $f(x) - \tilde{f}(x)$ è continua

possiamo utilizzare il teorema del valor medio che ci assicura che esiste un punto intermedio $x = \eta$ tale che l'integrale precedente possa essere riscritto come

$$E = \frac{f(\eta)-\tilde{f}(\eta)}{(\eta-a)(\eta-b)} \int_a^b (x-a)(x-b)\ dx \tag{1.12}$$

dove il termine che precede l'integrale è ormai una costante opportunamente determinata. Sviluppando l'integrale a destra con pochi passaggi si ottiene che vale $(a-b)^3/6$, il che corrisponde a dire che l'errore nella valutazione dell'integrale col metodo trapezoidale su un intervallo (a,b) va con il cubo dell'intervallo.

Se quindi suddividiamo l'intervallo di integrazione in $N = (b-a)/h$ intervalli di ampiezza h, avremo che ogni intervallo contribuirà ad un errore sull'integrazione che va come h^3.

Ora è intuitivo che, se usiamo una serie di Taylor per meglio approssimare la funzione integranda, potremo essere più accurati nel valutare il valore vero dell'integrale magari usando anche un minor numero di passi. Se quindi riconsideriamo la

serie di Taylor in Eq. 1.2 e ci ricordiamo che abbiamo già ricavato delle approssimazioni numeriche per la derivata prima (eq.1.6) e seconda (eq.1.7), possiamo ri-esprimere la nostra funzione nell'intervallo di integrazione $[-h, +h]$ come serie di Taylor utilizzando le nostre approssimazioni:

$$f(x) = f_0 + x\left(\frac{f_1 - f_{-1}}{2h}\right) + x^2\left(\frac{f_1 - 2f_0 + f_{-1}}{2h^2}\right) + O(h^3) \tag{1.13}$$

Questa può essere integrata facilmente per ottenere:

$$\begin{aligned}\int_{-h}^{+h} f(x) &= f_0 2h \ + \ 0 \ + \ \frac{h}{3}(f_1 - 2f_0 + f_{-1}) + O(h^5) \\ &= \frac{h}{3}(f_1 + 4f_0 + f_{-1}) + O(h^5)\end{aligned} \tag{1.14}$$

che rappresenta la cosiddetta regola di integrazione **alla Simpson**.

Si noti che nell'integrazione della (1.13) i termini con potenze dispari scompaiono perchè antisimmetrici e per questo, nel valutare il termine di troncamento $O(...)$, l'integrale in x^3 si annullerà e rimarrà l'integrale in x^4 che darà come risultato un termine del tipo $O(h^5)$, come riportato nella (1.14). È anche interessante notare che il termine $O(h^5)$ contiene al suo interno anche la derivata quarta dello sviluppo in serie e quindi se la f'''' fosse nulla (come nel caso di polinomi di grado inferiore al terzo) questo termine andrebbe a zero e l'integrale sarebbe esatto visto che anche tutte le altre derivate dello sviluppo in serie sarebbero nulle.

La regola precedente ci dà quindi il mattone con cui costruire numericamente un integrale definito suddividendo l'intervallo di integrazione in sottointervalli di ampiezza $2h$ ed adottando la scomposizione mostrata in eq.(1.8). Quindi, dato l'intervallo $[a, b]$, potremo riscrivere:

$$\begin{aligned}\int_a^b f(x)dx = \frac{h}{3}[&(f_a + 4f_{a+h} + f_{a+2h}) + (f_{a+2h} + 4f_{a+3h} + f_{a+4h}) \\ &+ ... + (f_{b-2h} + 4f_{b-h} + fb)] \\ = \frac{h}{3}[&f_a + 4f_{a+h} + 2f_{a+2h} + 4f_{a+3h} + ... + 4f_{b-h} + f_b]\end{aligned} \tag{1.15}$$

che viene detta **formula di Simpson** per l'integrazione.

Sulla quadratura "alla Simpson"

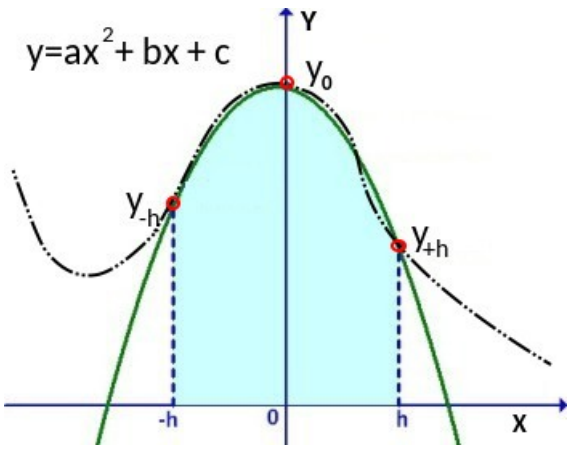

Figura 1.3: La possibilità di approssimare localmente una funzione generica (linea tratteggiata) con una parabola (linea verde continua) è alla base dello sviluppo della regola di Simpson per l'integrazione. Il grado di approssimazione sarà tanto migliore quanto più piccola sarà l'ampiezza h dell'intervallo.

Proviamo a ricavare la formula per l'integrazione alla Simpson in modo intuitivo, utilizzando il fatto che vogliamo approssimare la funzione integranda usando polinomi quadratici. Adottiamo per questo una parabola di equazione $y = ax^2 + bx + c$ che passi attraverso tre punti $(-h, y_{-h})$, $(0, y_0)$, (h, y_{+h}) della nostra funzione come mostrato in Figura 1.3.

L'integrale della parabola tra i limiti $-h, +h$ sarà data da:

$$\begin{aligned} A &= \int_{-h}^{+h} ax^2 + bx + c \; dx = \left(\frac{ax^3}{3} + \frac{bx^2}{2} + cx \right) \Big|_{-h}^{+h} \\ &= \frac{2ah^3}{3} + 2ch = \frac{h}{3}(2ah^2 + 6c) \end{aligned} \tag{1.16}$$

Se ora sfruttiamo l'appartenenza dei nostri tre punti ad una parabola possiamo scrivere le seguenti condizioni:

$$\begin{aligned} y_{-h} &= ah^2 - bh + c \\ y_0 &= c \\ y_{+h} &= ah^2 + bh + c \end{aligned} \tag{1.17}$$

Organizziamo ora opportunamente le precedenti condizioni per costruire un termine molto simile a quello ricavato per l'area della parabola:

$$y_{-h} + 4y_0 + y_{+h} = (ah^2 - bh + c) + 4c + (ah^2 + bh + c) = 2ah^2 + 6c \tag{1.18}$$

tanto che possiamo riscrivere l'area sotto la parabola in (1.16) come:

$$A = \frac{h}{3}(y_{-h} + 4y_0 + y_{+h}) \tag{1.19}$$

Se ora allarghiamo il campo ad un intervallo di integrazione $[a, b]$, che suddividiamo in $N = (b-a)/h$ passi di integrazione come in (1.8), ed andiamo a sommare su tutti gli intervalli otteniamo:

$$\begin{aligned}\int_a^b f(x)\ dx &= \frac{h}{3}\ [(y_a + 4y_{a+h} + y_{a+2h}) + (y_{a+2h} + 4y_{a+3h} + y_{a+4h}) + ...] \\ &= \frac{h}{3}[y_a + 4y_{a+h} + 2y_{a+2h} + 4y_{a+3h} + 2y_{a+4h} + ...]\end{aligned} \tag{1.20}$$

risultato del tutto simile a quello in eq. 1.15.

1.2.1 Quadratura "alla Gauss"

Un altro modo di affrontare il problema della quadratura è di considerare il valore dell'integrale come dato dal prodotto tra i valori della funzione calcolata in alcuni punti prescelti ed opportuni pesi da definirsi:

$$\int_a^b f(x)dx \approx \sum_{i=1}^{n} w_i f(x_i) \tag{1.21}$$

In fondo il metodo di Simpson prima discusso rientra proprio in un caso di questo tipo in cui i pesi sono già noti: infatti una volta fissate le x_i avevamo le w_i per integrare la nostra $f(x)$. Tuttavia, nelle formule di quadratura discusse finora è sempre necessario valutare la funzione calcolata negli estremi di integrazione ed in una serie di punti equispaziati tra i due estremi. Questo approccio è certamente valido quando la funzione integranda varia "ragionevolmente" nell'intervallo di integrazione, ma diventa un evidente svantaggio nel caso in cui la funzione integranda sia rapidamente variabile in parti dell'intervallo o anche possa essere singolare (ma ancora integrabile) negli estremi di integrazione.

Sugli integrali impropri

Quando la funzione integranda non è continua nell'intervallo di integrazione,

oppure almeno uno degli estremi di integrazione non è finito, l'integrale si dice "improprio". Questi casi vengono di solito calcolati analiticamente modificando il limite di integrazione "problematico" e facendo poi tendere il risultato ottenuto al vero limite originale. Un'altro approccio è quello di sfruttare un cambiamento di variabile favorevole che riporti i limiti di integrazione al finito.

Si possono verificare due casi di integrale improprio:

- di primo tipo: uno dei limiti di integrazione è all'infinito, p.es:

$$\int_1^\infty \frac{1}{x^2}dx = \lim_{t\to\infty}\int_1^t \frac{1}{x^2}dx \ \ldots \ = 1$$

- di secondo tipo: la funzione presenta punti di discontinuità nell'intervallo di integrazione, p.es.:

$$\int_0^4 \frac{1}{\sqrt{x}}dx = \lim_{\epsilon\to 0}\int_\epsilon^4 \frac{1}{\sqrt{x}}dx \ \ldots \ = 4$$

oppure:

$$\int_0^2 \frac{1}{\sqrt{4-x^2}}dx = \lim_{\epsilon\to 0}\int_0^{2-\epsilon} \frac{1}{\sqrt{4-x^2}}dx \ \ldots \ = \frac{\pi}{2}$$

Nei casi illustrati prima i risultati riportati sono ottenibili in termini analitici. In termini numerici invece è chiaro che si impone l'uso di una tecnica di integrazione che non coinvolga il calcolo della funzione ai limiti di integrazione. In casi simili si preferisce quindi usare il metodo di Gauss.

Per affrontare questo tipo di situazioni da un punto di vista numerico si sono sviluppati i **metodi di Gauss** che si basano sull'utilizzo dei valori della funzione in una serie di punti x_i opportunamente scelti, ma non necessariamente equispaziati o contenenti i limiti.

Un esempio a due punti

Come esempio introduttivo alla filosofia dell'integrazione "alla Gauss" immaginiamo di voler ottenere l'integrale di una f nell'intervallo [-1, +1] considerando due soli punti scelti in modo tale che possiamo scrivere:

$$\int_{-1}^{+1} f(x)dx = w_1(x_1)f(x_1) + w_2 f(x_2) \tag{1.22}$$

dove come w_1 e w_2 abbiamo indicato degli opportuni pesi da modulare in modo che l'uguaglianza sia soddisfatta. Se inoltre facciamo la richiesta di avere un risultato esatto quando la f è un polinomio di grado minore o uguale a 3, possiamo scrivere le 4 possibilità che in questo caso si possono presentare:

$$\begin{array}{lll} f(x)=1 & \longrightarrow & 1\cdot w_1 + 1\cdot w_2 = \int_{-1}^{+1} 1dx = 2 \\ f(x)=x & \longrightarrow & x_1\cdot w_1 + x_2\cdot w_2 = \int_{-1}^{+1} xdx = 0 \\ f(x)=x^2 & \longrightarrow & x_1^2\cdot w_1 + x_2^2\cdot w_2 = \int_{-1}^{+1} x^2dx = 2/3 \\ f(x)=x^3 & \longrightarrow & x_1^3\cdot w_1 + x_2^3\cdot w_2 = \int_{-1}^{+1} x^3dx = 0 \end{array} \tag{1.23}$$

Queste sono in pratica 4 condizioni da soddisfare contemporaneamente ed una possibile soluzione che implica la scelta di x_i nell'intervallo di integrazione è:

$$x_1 = -\frac{1}{\sqrt{3}},\ x_2 = +\frac{1}{\sqrt{3}},\ w_1 = 1,\ w_2 = 1 \tag{1.24}$$

e quindi, sostituendo nella formula per la quadratura (1.22), otteniamo:

$$\int_{-1}^{+1} f(x)dx = 1\cdot f(-\frac{1}{\sqrt{3}}) + 1\cdot f(\frac{1}{\sqrt{3}}) \tag{1.25}$$

È interessante notare che le ascisse ottenute in questo esempio di quadratura coincidono con le radici del polinomio di Legendre di ordine n=2, motivo per cui questo metodo viene spesso riferito come regola di **quadratura di Gauss-Legendre** a due punti.
La differenza tra i metodi a punti equispaziati (trapezi o Simpson) e quello di Gauss-Legendre a due punti è mostrata in modo intuitivo in Figura 1.4.

Sui polinomi di Legendre

Questi polinomi sono soluzioni di una particolare equazione differenziale ordinaria di secondo grado che gioca un ruolo importante in vari campi della Fisica. Questa è appunto detta "equazione di Legendre" ed ha forma:

$$(1-x^2)y'' - 2xy' + ky = 0 \quad \text{per} -1 < x < 1$$

Le soluzioni sono esprimibili come polinomi dati da:

$$y = P_n(x) = \frac{1}{2^n\ n!}\ \frac{d^n}{dx^2}(x^2-1)^n \tag{1.26}$$

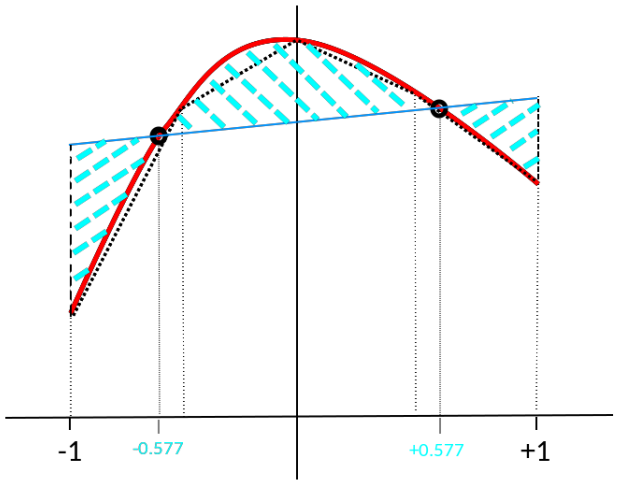

Figura 1.4: Integrazione di Gauss-Legendre a due punti. L'integrale della funzione viene calcolato esattamente dall'area del trapezio che passa per i due punti indicati da un cerchietto ad $x = \pm 1/\sqrt{3} \simeq \pm 0.577$. Se nelle regioni più esterne il trapezio contiene più area della curva, nella parte centrale invece esclude un'area esattamente uguale a quella inglobata sui bordi. In questo modo si ottiene che l'integrale calcolato per la curva è esatto. I punti di riferimento qui sono due, per funzioni più complicate si potranno usare più punti (vedi Tab. 1.1).

dove con $n = 1, 2, 3, ...$ si indica il grado del polinomio. Esplicitando i primi 6 polinomi si ottiene:

$$\begin{aligned} P_0(x) &= 1 \\ P_1(x) &= x \\ P_2(x) &= \frac{1}{2}(3x^2 - 1) \\ P_3(x) &= \frac{1}{2}(5x^3 - 3x) \\ P_4(x) &= \frac{1}{8}(35x^4 - 30x^2 + 3) \\ P_5(x) &= \frac{1}{8}(63x^5 - 70x^3 + 15x) \end{aligned} \tag{1.27}$$

che graficati come in Figura 1.5 mostrano subito che P_0 non ha zeri, P_1 ha uno zero in $x = 0$, P_2 ha due zeri in $x = \pm 1/\sqrt{3}$ e gli altri hanno il comportamento mostrato in figura fino al grado 4. Per il calcolo numerico sono utilissime due formule di ricorrenza che per calcolare polinomi e pesi per un dato polinomio usano i risultati ottenuti per polinomi di grado inferiore. Per il polinomio di grado n abbiamo:

$$nP_n = (2n - 1)\ x\ P_{n-1} - (n - 1)\ P_{n-2} \tag{1.28}$$

e per il calcolo dei pesi associati ai vari zeri del polinomio di grado n possiamo usare:

$$w_i = \frac{2}{(1 - x_i^2)\ [P_n'(x_i)]^2} \tag{1.29}$$

dove la derivata prima è anch'essa esprimibile usando polinomi:

$$P_n^{'} = \frac{n}{x^2 - 1}(xP_n - P_{n-1}). \tag{1.30}$$

Nelle precedenti relazioni P_{n-1} e P_{n-2} sono rispettivamente i polinomi di grado $n - 1$ ed $n - 2$.

Ricordiamo che siamo partiti richiedendo che questo metodo, usando due soli punti, possa calcolare **esattamente** integrali di un polinomio fino al grado 3. Se

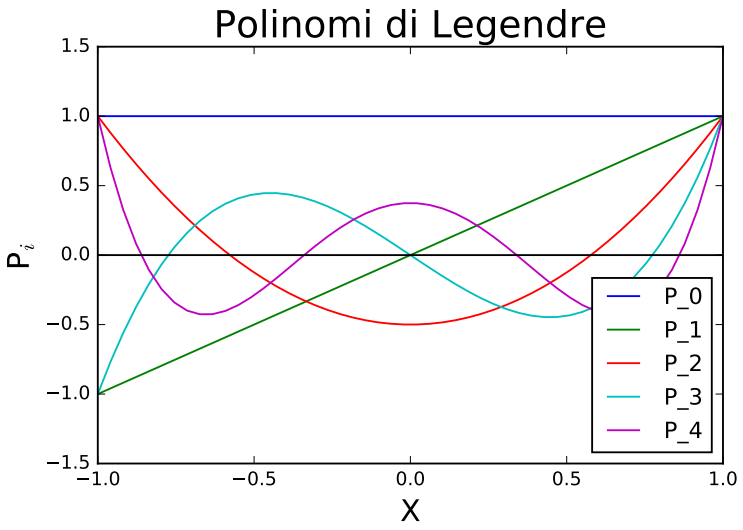

Figura 1.5: Polinomi di Legendre fino al quarto grado. GLi zeri, corrispondenti alle intersezioni con l'asse x, sono riportati in Tab. 1.1.

estendiamo l'approccio a più punti potremmo verificare che, usando n punti, si possono valutare **esattamente** integrali di funzioni polinomiali di grado $(2n-1)$. Per far questo bisognerà scegliere le radici dei polinomi di Legendre di grado n come punti su cui valutare la funzione con opportuni pesi. Queste informazioni sono ormai diffusamente tabulate e, come esempio, in tabella 1.1 sono riportati alcuni di questi valori per i primi sei gradi del polinomio usato, notando che i valori delle x corrispondono alle intersezioni con l'asse delle ascisse che si vedono in Figura 1.5.
Una prima valutazione approssimata di questi valori può anche essere fatta utilizzando una semplice espressione:

$$x_i = \cos \frac{\pi(i + 0.75)}{n + 0.5} \tag{1.31}$$

dove n è il numero di nodi (o grado del polinomio), ed i indica l'i-esimo zero con $i = 0, ..., n-1$. Questa formula appare particolarmente conveniente nel preparare una procedura di integrazione numerica che debba individuare di volta in volta l'i-esimo zero al variare del grado n del polinomio[2].
Su questa falsariga è possibile aumentare il numero di punti da utilizzare e quindi anche il grado del polinomio che si riesce ad integrare esattamente, sempre

[2]Come si vedrà in seguito per calcolare in modo numerico gli zeri di una funzione è necessario avere un'idea preliminare della regione in cui lo zero in questione debba essere cercato. Nel nostro caso la zona dove cercare può essere individuata a partire dall'approssimazione indicata.

x	w
n=2	
0.577350269189626	1.000000000000000
n=3	
0.000000000000000	0.888888888888889
0.774596669241483	0.555555555555556
n=4	
0.339981043584856	0.652145154862546
0.861136311594053	0.347854845137454
n=5	
0.000000000000000	0.5688888888888
0.538469310105683	0.478628670499366
0.906179845938664	0.236926885056189
n=6	
0.238619186083197	0.467913934572691
0.661209386466265	0.360761573048139
0.932469514203152	0.171324492379170

Tabella 1.1: Punti e coefficienti per la quadratura di Gauss-Legendre. Per semplicità sono riportati i soli punti positivi: p.es. per $n = 2$ i punti saranno: $x_1 = 0.577...$ ed $x_2 = -x_1 = -0.577...$ con pesi tutti e due unitari; per $n = 3$ considereremo invece i punti $x_0 = 0.00..., x_1 = 0.774..., x_2 = -x_1 = -0.774...$ con pesi rispettivi $w_0 = 0.888..., w_1 = 0.555..., w_2 = 0.555...$

ricordando che per usare la regola di Gauss-Legendre bisogna prima effettuare un opportuno cambio di variabile che riduca i limiti di integrazione all'intervallo [-1,+1].

Sui limiti di integrazione

Quando i limiti a e b non sono all'infinito, la sostituzione che modifica i limiti di integrazione ad un intervallo scelto da noi α e β è data da:

$$t = \frac{(\beta - \alpha)x + \alpha b - \beta a}{b - a}$$

Se ora imponiamo per i nostri limiti $\alpha = -1$ e $\beta = +1$ otteniamo:

$$t = \frac{2x - b - a}{b - a} \quad \text{e quindi:} \quad x = \frac{t \cdot (b - a) + (b + a)}{2} \tag{1.32}$$

In definitiva l'integrale equivalente tra [-1,+1] diventa:

$$\int_a^b f(x)dx = \int_{-1}^{+1} f(\frac{t \cdot (b - a) + (b + a)}{2}) \cdot \frac{b - a}{2} dt \tag{1.33}$$

Se invece uno dei due limiti è all'infinito si potrà usare uno di questi metodi:

- sostituire il limite infinito con uno finito, avendo cura di valutare l'errore che si compie nel trascurare la coda dell'integrale. Per esempio, se si può calcolare analiticamente la parte restante dell'integrale si potrà valutare il "pezzo mancante".
- cambiare la variabile x nella nuova variabile t per rendere finiti i limiti di integrazione, p.es.:

$$\begin{array}{lll} t = e^{-x} & \longrightarrow & x = -ln(t) \\ & & \text{per } x = \infty \rightarrow t = 0 \\ & & \text{per } x = 0 \rightarrow t = 1 \\ & \text{oppure:} & \\ t = x/(x+1) & \longrightarrow & x = t/(1-t) \\ & & \text{per } x = \infty \rightarrow t = 1 \\ & & \text{per } x = 0 \rightarrow t = 0 \end{array}$$

Notiamo infine che oltre al metodo di Gauss-Legendre, di cui abbiamo discusso, sono stati sviluppati altri metodi sempre basati sulla stessa idea: un integrale può essere espresso come sommatoria dei prodotti tra i valori che la funzione assume in punti scelti ed i corrispondenti pesi anch'essi scelti opportunamente. Per rendersi meglio conto delle potenzialità di questi metodi consideriamo la (1.21) e moltiplichiamo la f per una nuova funzione p ottenendo:

$$\int_a^b p(x)f(x)dx \approx \sum_{i=1}^{n} w_i f(x_i) \tag{1.34}$$

in cui è ovvio che, quando $p(x) = 1$, si ricade nel caso già discusso. Riferendoci a questa scrittura possiamo dire che la caratteristica delle formule di quadratura di questo tipo è di essere esatte se $f(x)$ è un polinomio di grado non superiore a $2n - 1$. Sia le x_i che le w_i che compaiono nella (1.34) dipendono dalla scelta della $p(x)$ e danno luogo a diversi tipi di integrazione. Le x_i in particolare sono gli zeri di un polinomio di grado n che varia al variare di $p(x)$. La tabella 1.2 mostra e riassume la situazione in diversi casi che si possono incontrare usando l'integrazione "alla Gauss".
Si capisce dalle prescrizioni di questa elencazione che per usare una di queste formule bisognerà prima ridurre il nostro integrale all'intervallo di integrazione indicato usando una opportuna sostituzione di variabile.

funzione; limiti di integrazione		Formule di	x_i zeri di polinomi di
$p(x) = 1; a = -1, b = +1$	$\rightarrow$	Gauss-Legendre	Legendre di grado n
$p(x) = e^{-x}; a = 0, b = +\infty$	$\rightarrow$	Gauss-Laguerre	Laguerre di grado n
$p(x) = e^{-x^2}; a = -\infty, b = +\infty$	$\rightarrow$	Gauss-Hermite	Hermite di grado n
$p(x) = \frac{1}{\sqrt{1-x^2}}; a = -1, b = +1$	$\rightarrow$	Gauss-Chebyshev	Chebyshev di grado n

Tabella 1.2

1.3 Ricerca di zeri

Ci poniamo ora il problema di trovare una soluzione per l'equazione

$$f(x) = 0 \tag{1.35}$$

con $f(x)$ funzione di una sola variabile x. In termini più formali diremmo che data una funzione $f(x)$, vogliamo trovare un valore $x = x_0$ tale che $f(x_0) = 0$.

Se il numero x_0 esiste allora viene detto **radice dell'equazione** (1.35) oppure **zero della funzione** $f(x)$. In quanto detto finora abbiamo fatto riferimento ad una funzione del tutto generica e quindi le considerazioni che svilupperemo saranno utilizzabili sia per risolvere equazioni lineari che non-lineari. Va infatti sottolineato che, in generale, non è facile trovare un'espressione analitica per le radici di un equazione. Per fare l'esempio dei semplici polinomi, abbiamo a disposizione delle espressioni analitiche per trovare le radici fino al quarto grado, mentre per trovare soluzioni di polinomi di grado maggiore l'unica possibilità che abbiamo è di usare metodi numerici.

Le idee per risolvere il problema della ricerca di radici con un calcolatore si basano tutte su metodi iterativi che possiamo riassumere così :

> A partire da una qualche stima o approssimazione iniziale di una radice x_0 costruiamo una sequenza $\{x_k\}$ di nuovi valori della x che soddisfano ad un criterio predefinito, verificando che la sequenza costruita converga effettivamente ad una radice della (1.35).

Tutto ciò in presenza di due importanti aspetti della procedura: un criterio di **convergenza** ed uno di **stop** dell'iterazione. Mentre i criteri di convergenza possono dipendere dal particolare approccio usato nel cercare le radici, i criteri di stop fanno riferimento sostanzialmente ad una terna di possibili condizioni che si possono incontrare nell'iterazione. Indicando con ϵ la tolleranza che pretendiamo

nel nostro calcolo e con $x = x_k$ lo specifico valore della x corrispondente alla k-esima iterazione della nostra procedura, si impone uno **stop** all'iterazione quando si verifica una delle seguenti circostanze:

- il valore della funzione risulta minore o uguale alla tolleranza: $|f(x_k)| \leq \epsilon$
- la variazione relativa della x tra due iterazioni consecutive è minore o uguale alla tolleranza: $|(x_{k-1} - x_k)/x_k| \leq \epsilon$
- il numero di iterazioni k è maggiore o uguale ad N, con N numero massimo di iterazioni ritenuto accettabile per la nostra procedura.

1.3.1 Metodo delle bisezione

Come suggerito dal titolo, il metodo che ora andiamo a descrivere per la ricerca degli zeri di una funzione si basa su ripetute bisezioni dell'intervallo contenente lo zero cercato. Supponiamo che $f(x)$ sia nota e abbia radici [3] reali nell'intervallo $[a, b]$, il metodo consiste nel procedere secondo i seguenti passi:

- dividiamo l'intervallo $[a, b]$ in due parti uguali con $c = (a + b)/2$ punto medio. Se in c la funzione è nulla (o è entro il livello di tolleranza richiesto) allora in c avremo trovato il nostro zero. Altrimenti concluderemo che lo zero si trova in uno dei due intervalli $[a, c]$ o $[c, b]$.
- tra questi due nuovi intervalli individuiamo quello che contiene la radice e suddividiamolo a sua volta in due parti ripetendo quanto fatto al punto precedente.
- continuiamo il processo della bisezione finchè la radice non si trova "intrappolata" in un intervallo tanto piccolo da rientrare entro il limite di tolleranza richiesto.

Per completare il quadro dobbiamo stabilire quale sia il semi-intervallo che, dopo la bisezione, conterrà lo zero cercato. A questo scopo utilizziamo il fatto che la nostra funzione f è continua nell'intervallo $[a, b]$ e quindi, dato un valore m compreso tra i valori della funzione ai bordi $f(a)$ ed $f(b)$, cioè $f(a) < m < f(b)$, esiste un punto x_m in cui $f(x_m) = m$. Questo ci garantisce che se $f(a)$ ed $f(b)$ sono di segno opposto, allora da qualche parte entro l'intervallo la funzione dovrà necessariamente attraversare lo zero per poter cambiare di segno. Quindi, prima

[3] Cosa di cui ci possiamo accertare verificando che la funzione cambi di segno nello stesso intervallo.

di iniziare la procedura di bisezione dobbiamo già aver individuato un intervallo entro il quale la nostra funzione cambia segno.
La stessa procedura la possiamo comunque implementare in modo da partire da un valore della $f(x)$ in un punto iniziale di prova x_{ini} sicuramente minore dello zero che andiamo a cercare. Usando un passo h predefinito andiamo poi a calcolare la funzione al passo successivo $f(x_{\text{ini}+h})$ e verifichiamo se la funzione cambia segno. In questo modo stiamo praticamente esplorando il comportamento della funzione $f(x)$ all'aumentare della x e quindi potremo incontrare due possibili situazioni:

- $f(x_{\text{ini}+h})$ mantiene lo stesso segno di $f(x_{\text{ini}})$: non siamo ancora nelle vicinanze di uno zero e quindi proviamo a calcolare la f esplorando punti successivi, sempre con passo di ampiezza h, proseguendo fino ad osservare un cambiamento del segno;

- $f(x_{\text{ini}+h})$ cambia segno rispetto ad $f(x_{\text{ini}})$: in questo caso nell'ultimo passo esplorato abbiamo superato lo zero cercato ed andremo quindi a bisecare l'ultimo segmento in modo da restringere meglio la regione che contiene lo zero.

La procedura descritta è ovviamente iterativa e quindi, se la funzione mantiene lo stesso segno nell'intervallo esplorato, dovremo procedere ad esplorare l'intervallo successivo che avrà sempre ampiezza h ma con il punto iniziale aggiornato dalla relazione $x_{ini} = x_{ini} + h$. Fatto questo potremo fermare la procedura quando una delle condizioni descritte prima per lo stop sarà soddisfatta. La Fig. 1.6 illustra un caso in cui la funzione ha due zeri e quindi diventa importante capire se vogliamo individuare il primo, il secondo o tutti e due i punti di zero. È evidente che un minimo di conoscenza preliminare dell'andamento generale della funzione ci aiuterà a definire il punto di partenza x_A o x_B più opportuno per raggiungere i nostri scopi.

La rapidità della procedura di convergenza dipenderà dal fatto che ogni volta che bisechiamo l'intervallo la sua ampiezza diminuisce di un fattore 2 e quindi, dopo n iterazioni sappiamo che la radice è compresa entro un intervallo di dimensione $\epsilon_{n+1} = \epsilon_n/2$. Usando questa proprietà possiamo prevedere che le iterazioni necessarie per ottenere lo zero con una certa tolleranza ϵ saranno:

$$n = \log_2(\frac{\epsilon_0}{\epsilon}) \tag{1.36}$$

dove con ϵ_0 abbiamo indicato la dimensione dell'intervallo $[a, b]$ adottato inizialmente.

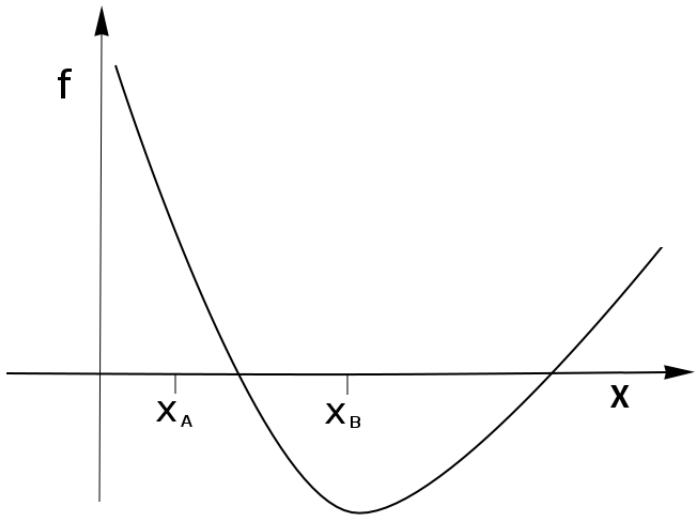

Figura 1.6: La funzione possiede due zeri. Se si esplora l'asse X verso destra alla ricerca del primo zero sarà decisivo il partire da X_A o da X_B. Se invece la nostra tecnica usa la bisezione di un intervallo, allora sarà necessario aver individuato prima quale zero si vuole individuare per scegliere opportunamente X_A ed X_B a cavallo dello zero cercato.

1.3.2 Metodo della secante

Un altro metodo utile per la ricerca di zeri è quello cosiddetto della **secante** che ha il pregio di essere più rapidamente convergente di quello della bisezione, sempre nel caso di funzioni continue non rapidamente variabili nelle vicinanze dello zero. Rispetto al metodo della secante ci sono similitudini e differenze:

- come nel caso della bisezione la funzione deve essere localmente approssimativamente lineare, cioè si deve comportare "bene" nella regione in cui cerchiamo gli zeri;

- similmente al metodo della bisezione, ad ogni iterazione si determina un nuovo punto x_{k+1} che va ridurre la dimensione del campo di ricerca dello zero;

- diversamente dalla bisezione in cui di volta in volta si valuta quale degli estremi dell'intervallo di ricerca sostituire, qui il nuovo intervallo viene sempre definito dalle ultime due valutazioni ottenute.

La fig.1.7 illustra il meccanismo di funzionamento del metodo della secante applicato ad una generica funzione.

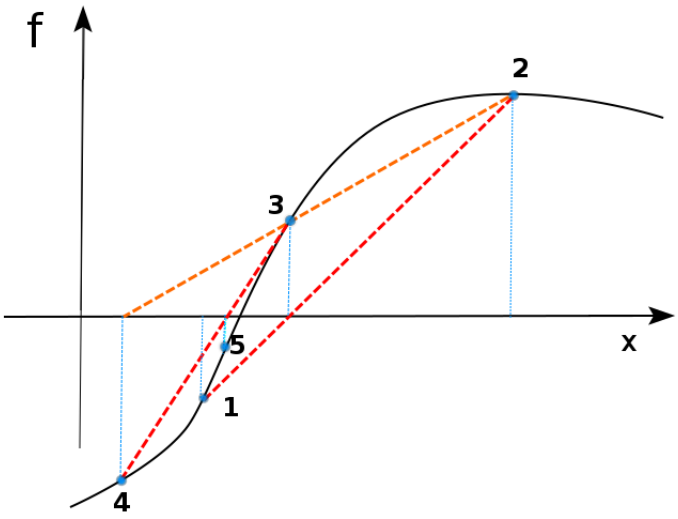

Figura 1.7: Scelti due punti iniziali 1 e 2 tra i quali si trova lo zero della funzione, la loro congiungente incrocia le ordinate in un punto che corrisponde al punto 3 sulla funzione. Congiungendo i punti 2 e 3 si incrocia l'asse delle ordinate in un punto che corrisponde al punto 4 sulla funzione. Congiungento 3 e 4 .... Iterando questa procedura ci si avvicina sempre di più allo zero della funzione.

1.3.3 Metodo di Newton-Raphson

Questo metodo per il calcolo degli zeri di una funzione si può usare ogni volta che oltre alla funzione conosciamo anche la sua derivata prima. Il procedimento adottato in questo caso parte da un punto iniziale x_i e considera l'intersezione tra la tangente alla funzione in quel punto con l'asse delle ordinate, individuando in questo modo una nuova posizione x_{i+1} in cui valutare la funzione e la sua derivata. Iterando questo processo si potrà quindi arrivare ad individuare lo zero della nostra funzione entro l'accuratezza richiesta, sempre che la funzione sia localmente lineare in prossimità della zero. Il senso di questo metodo si può capire a partire dalla serie di Taylor già incontrata (1.1) che, se troncata al solo termine lineare diventa:

$$f(x) = f_0 + f'(x - x_0) + ... \tag{1.37}$$

Se nella precedente volessimo annullare la $f(x)$ e quindi trovare lo zero, dovremmo spostarci rispetto ad x_0 di un passo pari a

$$(x - x_0) = -\frac{f_0}{f'}$$

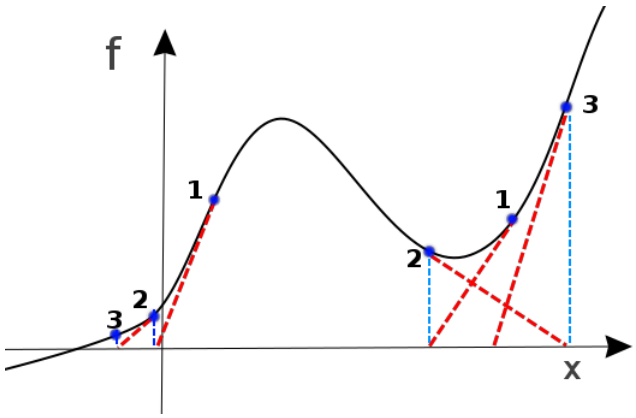

Figura 1.8: Usando il metodo di Newton-Raphson e partendo dal punto 1 si ottiene, per iterazione, la sequenza di punti indicati con i numeri 1,2,3. La sequenza a destra non converge allo zero, mentre quella a sinistra tende correttamente allo zero della funzione, mostrando che il metodo può avere successo solo se parte da un punto all'interno del "bacino di attrazione" dello zero.

che geometricamente corrisponde proprio alla distanza tra x_0 ed il punto in cui la tangente allo stesso punto interseca l'asse x. Quindi, usando questa espressione, otteniamo una valutazione di un nuovo punto sull'asse delle x che possiamo generalizzare così :

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} \tag{1.38}$$

Iterando questa procedura andremo ad individuare una sequenza di punti che convergerà verso lo zero della funzione come illustrato in Figura 1.8.
Come per gli altri metodi, anche per questo è necessario rendersi conto dei problemi che si possono generare da una incauta scelta dei punti di partenza e/o con funzioni complicate. La Fig.1.8 illustra un caso di scelta favorevole ed uno sfavorevole, quest'ultimo dovuto ad una scelta inappropriata del punto di partenza dell'iterazione rispetto all'andamento della funzione.

1.4 Equazioni differenziali ordinarie

Le **equazioni differenziali ordinarie** sono quelle che coinvolgono funzioni di una sola variabile e le loro derivate. Siccome i modelli della Fisica sono spesso più facilmente formulati usando equazioni differenziali, la soluzione di queste ultime è chiaramente di grande interesse per noi. Qui ci occuperemo solo di quelle di primo grado (in cui compaiono le sole derivate prime) perchè una equazione di

grado superiore può sempre scomporsi in una serie di equazioni di primo grado. Infatti, se consideriamo una generica equazione del secondo ordine

$$\frac{d^2y}{dx^2} + q(x)\frac{dy}{dx} = r(x) \tag{1.39}$$

possiamo facilmente verificare che è possibile ridurla a due equazioni di primo grado equivalenti:

$$\begin{aligned} \frac{dy}{dx} &= z(x) \\ \frac{dz}{dx} &= r(x) - q(x)z(x) \end{aligned} \tag{1.40}$$

Per fare un esempio pratico, se consideriamo l'equazione del moto della meccanica abbiamo un'espressione di secondo grado:

$$m\frac{d^2x}{dt^2} = F(x) \tag{1.41}$$

e, ricordando che il momento è $p = m(dx/dt)$, la possiamo riscrivere come due equazioni di primo grado:

$$\frac{dx}{dt} = \frac{p}{m} \quad ; \quad \frac{dp}{dt} = F(x) \tag{1.42}$$

Quindi, estrapolando da questo esempio, possiamo dire che il problema della soluzione di un'equazione differenziale di ordine n si riduce alla soluzione di un sistema di n equazioni accoppiate del primo ordine.

L'altro ingrediente importante per ottenere soluzioni fisiche è la considerazione delle condizioni al contorno che, in generale, danno luogo a due categorie di problemi:

- problema ai valori iniziali: nel qual caso il valore di y è dato al punto iniziale x_{ini} e si vuole calcolare il valore ad un dato punto finale x_{fin};

- problema con condizioni al contorno: il valore della y è specificato in più di un punto (nel caso più semplice si tratta dei punti iniziale e finale) e si richiede di calcolare l'andamento della funzione nei punti intermedi.

Qui ci interesseremo al più semplice dei due, cioè al caso in cui si impone il solo valore iniziale. Un esempio tipico lo abbiamo nel caso del moto di un punto materiale in cui, conoscendo posizione e momento iniziale della particella, si voglia conoscere il moto successivo sulla base delle (1.42).

La logica alla base della soluzione di questi problemi si basa sulla riscrittura delle derivate in termini finiti per indicare i quali useremo Δx e Δy al posto di dx e dy, ottenendo infine delle formule algebriche per valutare il cambiamento della funzione Δy al variare di Δx della variabile indipendente. È intuitivo che al diminuire del passo Δx otterremo una sempre migliore approssimazione alla equazione differenziale originale.

1.4.1 metodo di Eulero

Mettendo in pratica letteralmente quanto abbiamo detto finora corrisponde a seguire il cosiddetto **metodo di Eulero** che, anche se sconsigliabile ai fini pratici perchè inefficiente, è utile per impadronirsi dei concetti importanti in questo tipo di problemi.
Si tratta di approssimare con differenze un'equazione del tipo

$$\frac{dy}{dx} = g(x, y(x)) \tag{1.43}$$

tenendo conto che, per semplificare la scrittura, nel seguito potremo poi usare y al posto di $y(x)$. Allora, in un generico punto x_i possiamo scrivere:

$$\frac{y_{i+1} - y_i}{h} + O(h) = g(x_i, y_i) \tag{1.44}$$

e da questa ricavare la regola di ricorrenza per y_{i+1}:

$$y_{i+1} = y_i + hg(x_i, y_i) + O(h^2) \tag{1.45}$$

che corrisponde a trovare la successiva y_{i+1} come somma della attuale y_i più il contributo di un termine che rappresenta la stima dell'integrale sulla funzione g, fatta localmente[4] su un piccolo intervallo h. Come si vede l'accuratezza con cui si valuta la funzione al passo successivo $i+1$ è dell'ordine di $O(h^2)$ per cui, siccome suddividiamo il cammino su cui integrare l'equazione in N passi, avremo che il risultato finale sarà accurato a meno di $O(h)$, proprio perchè $N \propto h^{-1}$. L'errore globale quindi dimunuisce solo linearmente all'aumentare del numero di passi e per questo motivo il metodo descritto viene considerato piuttosto inefficiente.

[4] utilizzando l'approssimazione trapezoidale

1.4.2 Aumentare l'efficienza

Il metodo di Eulero permette quindi una soluzione in cui l'errore di approssimazione va come h e quindi, diminuendo h si potrà ottenere una soluzione sempre più accurata. Tuttavia, il conseguente aumento proporzionale del numero di operazioni richieste per completare la soluzione rende sconveniente questa strada e quindi, per ottenere una più rapida diminuzione dell'errore di approssimazione O, conviene riconsiderare l'espansione in serie di Taylor della nostra y nell'intorno y_i per calcolare la y_{i+1}:

$$y_{i+1} = y_i + hy_i' + \frac{1}{2}h^2 y_i'' + O(h^3) \tag{1.46}$$

Dalla (1.43) otteniamo l'espressione delle derivate:

$$\begin{aligned} y' &= g(x,y) \\ y'' &= \frac{dg}{dx} + \frac{dg}{dy}\frac{dy}{dx} = \frac{dg}{dx} + \frac{dg}{dy}g \end{aligned} \tag{1.47}$$

Se ora sostituiamo queste espressioni per le derivate nella precedente relazione otteniamo:

$$y_{i+1} = y_i + hg + \frac{1}{2}h^2\left[\frac{dg}{dx} + \frac{dg}{dy}g\right] + O(h^3) \tag{1.48}$$

dove sia la g che le sue derivate sono calcolate in x_i, y_i. Vediamo quindi che questa relazione di ricorrenza è associata ad un errore che va come $O(h^3)$ e quindi, considerando che l'integrazione richiede un numero di passi $N \propto h^{-1}$, otterremo un errore globale del tipo $O(h^2)$ che è comunque migliore dell'approccio di Eulero.

Quindi, una volta note le derivate prima e seconda della funzione g possiamo usare questo metodo ed in linea di principio possiamo anche migliorare l'efficienza considerando derivate di ordine superiore nell'espansione in serie. Adottando quest'ultima possibilità si potrà migliorare, ma pagando il prezzo di un'algebra più complicata.

1.4.3 Metodo Runge-Kutta

Si tratta di un metodo efficiente e largamente utilizzato che, diversamente da quanto richiesto dal metodo discusso prima, non ha bisogno della conoscenza delle derivate della funzione e per questo viene largamente preferito nella pratica. L'idea è di risolvere il problema suddividendolo in più passi nei quali poter

applicare un metodo di tipo "Eulero pesato" che sostanzialmente ci permette di usare i soli valori della funzione senza dover anche calcolare le sue derivate.
Per rendersi conto del meccanismo di funzionamento consideriamo l'equazione differenziale (1.43) e immaginiamo di suddividere l'intervallo in cui cerchiamo soluzioni in una successione di più sottointervalli. La soluzione analitica su un generico sottointervallo si può scrivere:

$$y_{i+1} = y_i + \int_{x_i}^{x_{i+1}} g(x, y)dx \tag{1.49}$$

che corrisponde a trasformare la soluzione dell'equazione differenziale nell'equivalente problema integrale. Se ora stabiliamo un intervallo di integrazione di ampiezza h e approssimiamo la g con una serie di Taylor come in eq.(1.46), sviluppata intorno al punto mediano $x_m = x_i + h/2$, possiamo riscrivere l'integrale come:

$$\begin{aligned} y_{i+1} &= y_i + \int_{-h/2}^{+h/2} g(x, y)dx \\ &= y_i + \int_{-h/2}^{+h/2} \left[g(x_m, y_m) + xg'(x_m, y_m) + \frac{x^2}{2} g''(x_m, y_m) + ... \right] dx \end{aligned} \tag{1.50}$$

Ricordando poi che g, g', g'' sono costanti rispetto all'integrazione ed integrando le varie parti della somma otteniamo:

$$y_{i+1} = y_i + h\ g(x_m, y(x_m)) + O(h^3) \tag{1.51}$$

Notiamo come nell'ultima relazione il secondo termine a destra corrisponda proprio ad approssimare l'integrale con l'area di un parallelepipedo rettangolo con base h ed altezza pari al valore della funzione nel punto medio. Si nota ancora che l'errore associato a questo approccio va come h^3 ed è il risultato dalla integrazione del termine quadratico dell'espansione in serie poichè l'integrazione del termine lineare si annulla[5]. A questo punto per procedere a calcolare y_{i+1} avremmo bisogno di conoscere la $g(x_m, y(x_m))$ nel punto medio dell'intervallo, quantità che al momento non conosciamo perchè non abbiamo ancora calcolato la $y(x_m)$. Per superare questa difficoltà possiamo però stimare il valore della y_m utilizzando la formula di Eulero (1.45) e adottando un passo $h/2$ che ci permette di scrivere:

$$y(x_m) = y(x_i) + h/2\ g(x_i, y(x_i)) + O(h^2) \tag{1.52}$$

[5]analogamente a quanto già discusso per l'accuratezza della (1.14), l'integrazione del termine lineare produce un termine in x^2 che si annulla sull'intervallo di integrazione $[-h/2, +h/2]$.

che, come si vede, contiene solo termini noti al punto i-esimo dell'iterazione. Questa relazione, insieme alla (1.50) ci permette di scrivere la (1.49) come:

$$\begin{aligned} y_{i+1} &= y_i + hg(x_m, y(x_m)) + O(h^3) \\ &= y_i + hg(x_m, y(x_i) + h/2\ g(x_i, y(x_i))) + O(h^3) \end{aligned} \tag{1.53}$$

nella quale, come si vede, tutti i termini sono calcolabili al passo i-esimo, essendo noto l'intervallo h. Questo approccio è detto **Runge-Kutta del secondo ordine** per via dell'accuratezza finale e per facilitarne l'uso è utile organizzare il calcolo secondo lo schema:

$$\begin{aligned} (\Delta y)_0 &= h\ g(x_i, y(x_i)) \\ (\Delta y)_1 &= h\ g(x_i + h/2, y_i + (\Delta y)_0/2) \\ y_{i+1} &= y_i + (\Delta y)_1 + O(h^3) \end{aligned} \tag{1.54}$$

in cui si vede che per avanzare al passo $(i+1)$-esimo dell'integrazione si fa uso dei soli valori già calcolati per il passo i-esimo. Il senso geometrico di questo metodo può essere apprezzato se si nota che i termini $(\Delta y)_0$ e $(\Delta y)_1$ non sono altro che le variazioni della y ricavate rispettivamente a partire dalla derivata della g nei punti iniziale e mediano (vedi Figura 1.9).
Si noti che il metodo Runge-Kutta è sviluppabile ulteriormente per ottenere una migliore accuratezza utilizzando un maggior numero di punti intermedi nell'intervallo di integrazione. Il metodo più diffuso è il cosiddetto Runge-Kutta del quarto ordine che utilizza derivate calcolate in 4 punti. Lo schema di calcolo per questo metodo diventa:

$$\begin{aligned} k_1 &= g(x_i, y(x_i)) \\ k_2 &= g(x_i + h/2, y(x_i) + k_1 h/2) \\ k_3 &= g(x_i + h/2, y(x_i) + k_2 h/2) \\ k_4 &= g(x_i + h, y(x_i) + k_3 h) \\ y_{i+1} &= y_i + h(k_1/6 + k_2/3 + k_3/3 + k_4/6) + O(h^5) \end{aligned} \tag{1.55}$$

che, come si vede, è un metodo accurato al 4 ordine che viene di solito preferito a quello del secondo ordine che abbiamo prima ricavato. Il Runge-Kutta del quarto ordine è infatti ritenuto come un buon compromesso tra velocità e accuratezza del calcolo ed è pertanto largamente utilizzato. Infine, è interessante notare che il risultato appare ora legato ad una media pesata dei prodotti $hk_{1,2,3,4}$ che non sono altro che degli incrementi associati alle pendenze ricavate nei 4 punti utilizzati nel calcolo.

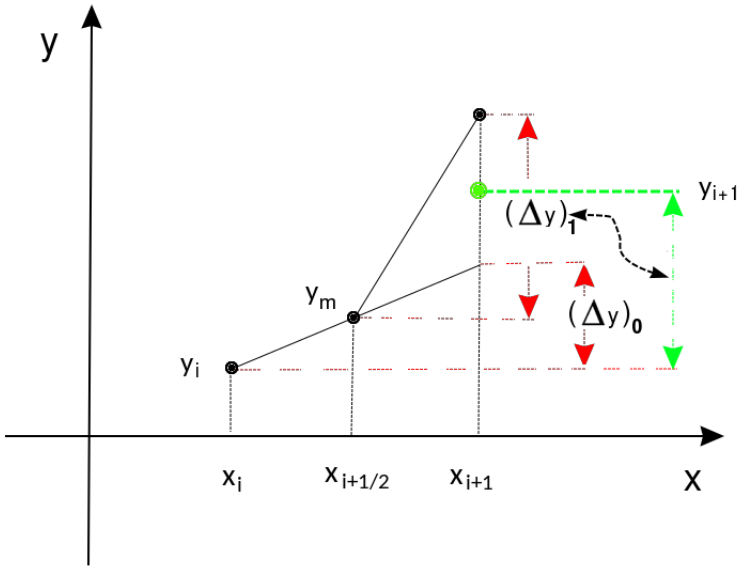

Figura 1.9: Schema del metodo Runge-Kutta di ordine 2. I simboli si riferiscono alle relazioni (1.54) che determinano il punto finale y_{i+1} (in verde) come somma del valore della y al passo iniziale iniziale più l'incremento $(\Delta y)_1$ che si calcola una volta noto l'incremento $(\Delta y)_0$.

Capitolo 2

Numeri casuali e metodo *Monte Carlo*

Familiarizzare con i numeri casuali (in letteratura **random numbers**) è di grande utilità per prepararsi allo sviluppo di modelli al calcolatore. I numeri casuali infatti giocano un ruolo importante nello sviluppo di modelli in fisica e non solo. Sono coinvolti in vari aspetti che vanno dal calcolo di integrali complicati fino al campionamento, alla simulazione, al test di modelli.

Ogni volta che un processo avviene senza avere alcuna connessione con altri processi siamo portati a dire che si tratta di un processo casuale. Questo però non vuol dire che il processo stesso non abbia una sua struttura interna. Per fare un esempio, supponiamo di estrarre a caso un numero da una tombola e, dopo averlo rimesso in gioco, ripetere questa operazione un gran numero di volte. Con il procedere delle estrazioni potremo osservare che il numero di volte in cui è stato estratto un dato numero tenderà ad avvicinarsi al numero di volte in cui è stato estratto ogni altro numero producendo così una distribuzione degli eventi abbastanza uniforme. Questo fatto si può rappresentare in forma di istogramma dei numeri estratti che si dovrebbe mostrare abbastanza "piatto" proprio ad indicare che tutti i numeri hanno la stessa probabilità di essere estratti.

Un altro esempio potrebbe essere il lancio di 6 monete uguali. Anche in questo caso, ripetendo l'operazione un gran numero di volte, potremo osservare quante volte otteniamo rispettivamente 6, 5, 4, 3, 2, 1, 0 monete con "testa" rivolta verso l'alto. Scopriremo che, diversamente da prima, l'istogramma corrispondente sarà a forma di campana e ricorderà una distribuzione di tipo gaussiano centrata sul numero 3. In tutti e due gli esempi abbiamo quindi usato un processo

casuale (estrarre i numeri o lanciare le monete) per verificare che le estrazioni o i lanci seguono una loro struttura interna che si riflette nelle diverse distribuzioni.

Sia la distribuzione uniforme che quella Gaussiana sono perfettamente definite analiticamente e sono caratterizzate da un valor medio e da una **varianza** (la cui radice chiamiamo **deviazione standard**). Nel nostro caso però abbiamo a che fare con dati sperimentali e quindi siamo costretti a fare delle valutazioni sia della media che della varianza a partire dagli stessi dati. Per questo ricaviamo una stima di queste quantità a partire dalle relazioni:

$$\begin{aligned} \bar{x} &= \frac{\sum_1^N x_i}{N} \\ \sigma &= \frac{\sum_1^N (x_i - \bar{x})^2}{N-1} \end{aligned} \tag{2.1}$$

dove con N abbiamo indicato il numero totale di misure a disposizione. Se usiamo un calcolatore possiamo generare una sequenza di numeri casuali usando un algoritmo basato sulla seguente formula di ricorrenza:

$$x_{i+1} \; = \; (A \; x_i \; + \; B) \mod M \qquad \text{per } i = 0, 1, 2, ... \tag{2.2}$$

In questa formula, anche detta "generatore lineare congruente" (in letteratura "linear congruential generator"), ci sono quattro parametri e cioè: A, B, M, ed x_0, quest'ultimo detto **seme o seed** che è necessario per far partire la formula di ricorrenza da un valore iniziale della x. Questi parametri sono numeri interi positivi scelti in modo che

$$A, B, x_0 < M \quad \text{con} \;\; A \; \neq \; 0$$

cosicchè la serie costruita con questa formula avrà valori estratti nell'intervallo $0 < \; x_i \; < M - 1$.
Tuttavia, i numeri "casuali" che così generiamo vengono calcolati da un algoritmo deterministico e quindi non dovremmo chiamarli effettivamente casuali. Dovremmo pertanto piuttosto chiamarli **pseudo-casuali** anche se l'algoritmo ricorsivo che li estrae li fa sembrare casuali. Proprio per la natura deterministica degli algoritmi utilizzati, la serie prima o poi si ripeterà segnalando appunto che in realtà si tratta di numeri non assolutamente casuali ma solo **pseudo-casuali**. Infatti una serie di numeri casuali dovrebbe avere periodo infinito e per avvicinarsi il più possibile a questo sono stati sviluppati vari algoritmi per rendere il periodo di ricorrenza dei numeri estratti dal calcolatore il più lungo possibile. Oltre ad avere un periodo di ricorrenza molto lungo, dei buoni numeri pseudo-casuali

dovrebbero anche essere tra loro il meno correlati possibile ed a questo scopo sono stati sviluppati dei test per verificarne il grado di correlazione. Considerazioni di questo tipo, all'interno di ogni linguaggio adatto al calcolo numerico, hanno portato allo sviluppo di algoritmi per generare numeri pseudo-casuali che siano il più possibile assimilabili ai veri numeri casuali

2.1 Probabilità non uniforme

Finora abbiamo discusso un modo in cui un calcolatore può fornirci numeri pseudo-casuali che siano distribuiti con uguale probabilità in un dato intervallo di valori. Spesso però, nel simulare un fenomeno naturale, abbiamo bisogno di estrarre a caso valori di grandezze la cui distribuzione di probabilità non sia uniforme ma abbia una sua precisa connotazione. Per esempio, nel caso della radiazione che arriva da una stella lontana l'esperienza ci dice che il numero di fotoni al secondo che raccogliamo fluttua intorno ad un valore medio, in un modo ben descritto da una funzione di probabilità di Poisson. Se volessimo simulare un simile fenomeno dovremmo poter estrarre "a caso" valori non più distribuiti in modo uniforme ma piuttosto distribuiti secondo una Poissoniana.

Siccome questo discorso si può generalizzare a tutte le possibili distribuzioni di probabilità, è utile elaborare qualche algoritmo che permetta di estrarre numeri casuali distribuiti secondo una predefinita funzione di probabilità (PDF, da Probability Density Function). Questo scopo è raggiungibile in più modi, qui ne discutiamo uno basato sulla conoscenza della stessa funzione di probabilità che vogliamo simulare durante la generazione dei numeri casuali.

2.1.1 metodo dell'inversione

Questo metodo è basato sulla conoscenza della cumulativa (CDF, da Cumulative Density Function) della funzione di probabilità che vogliamo usare per simulare le variabili casuali di nostro interesse. In estrema sintesi il metodo si basa sul fatto che, per ogni variabile casuale x, c'è una CDF che, una volta normalizzata all'unità, è una funzione $F(x)$ non-decrescente con $0 \leq F(x) \leq 1$ per tutti i valori di x.
Questo fatto suggerisce di usare la CDF per stabilire una corrispondenza inversa tra i numeri casuali estratti con probabilità uniforme (che abbiamo incontrato prima) e la CDF stessa. I seguenti passi illustrano intuitivamente il metodo, nel caso in cui si voglia simulare una PDF di tipo triangolare come mostrato in Figura 2.1:

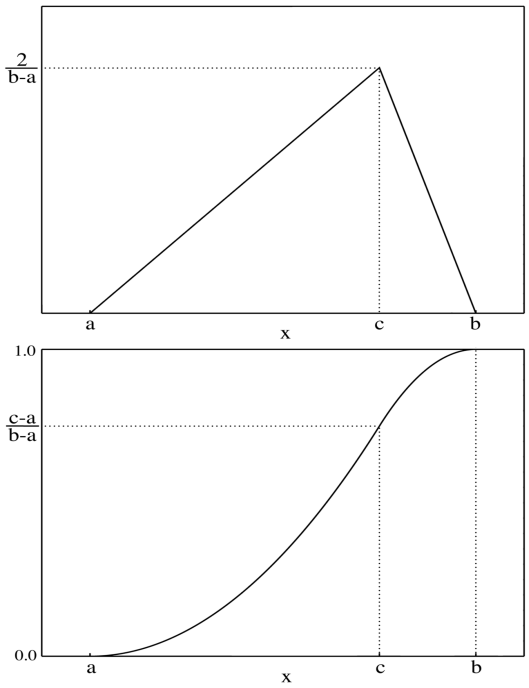

Figura 2.1: Distribuzione di probabilità (PDF, sopra) e corrispondente cumulativa (CDF, sotto). Si noti che la PDF in alto ha area unitaria, mentre i valori della CDF in basso sono compresi tra 0 ed 1.

La funzione cumulativa normalizzata rappresenta la probabilità di ottenere un valore X minore della variabile x in modo che

$$F(x) \;=\; Prob(X < x) = \frac{\int_{-\infty}^{x} P(x)dx}{\int_{-\infty}^{+\infty} P(x)dx} \tag{2.3}$$

- graficare la funzione cumulativa, normalizzata all'unità, per la PDF che vogliamo simulare (vedi Fig. 2.1);

- estrarre un numero nell'intervallo [0, 1] da un generatore casuale con probabilità uniforme

- riportare il numero casuale estratto tra [0,1] sull'asse y della cumulativa normalizzata tra [0,1];

- ricavare il valore corrispondente sull'asse x della cumulativa.

In questo modo da una distribuzione uniforme di numeri casuali possiamo simulare una nuova distribuzione per effetto della proiezione sull'asse x della cumulativa dei valori via via estratti. È quindi intuitivo che se la PDF in Figura 2.1 fosse una costante avremmo una cumulativa (in basso) che sarebbe una linea retta con pendenza 1 e quindi la mappatura della coordinata y sulla x produrrebbe gli stessi valori senza cambiare la forma della distribuzione dei numeri casuali di partenza.

2.1.2 Metodo del rigetto

Si basa su una interpretazione geometrica della PDF e può essere usato per generare campioni di valori di una qualsiasi variabile casuale x che soddisfi a queste condizioni:

- può assumere valori solo su un intervallo finito [a,b];

- ha una PDF che è limitata, cioè non diverge per nessun valore della variabile casuale

Se ora indichiamo con c il massimo valore raggiunto dalla PDF della nostra variabile, per generare valori casuali della nostra x usando il metodo del rigetto operiamo secondo questa procedura:

1) includiamo la PDF che vogliamo usare entro un rettangolo che abbia come altezza il valore massimo c assunto dalla PDF e come base l'intervallo di variabilità consentito alla x, come in Figura **??**;

2) generiamo due numeri casuali dalla nostra distribuzione uniforme iniziale e normalizziamoli ai lati del rettangolo prima definito (moltiplicandoli uno per $(b-a)$ e l'altro per c;

3) usiamo i due numeri ottenuti come coordinate (x, y) di un punto interno al rettangolo prima costruito intorno alla PDF;

4) se il punto ricadrà al disotto della PDF ne accetteremo la x, altrimenti verrà ignorato e si ritornerà al punto precedente ripetendo la procedure quanto necessario fino ad estrarre l'insieme di valori richiesto.

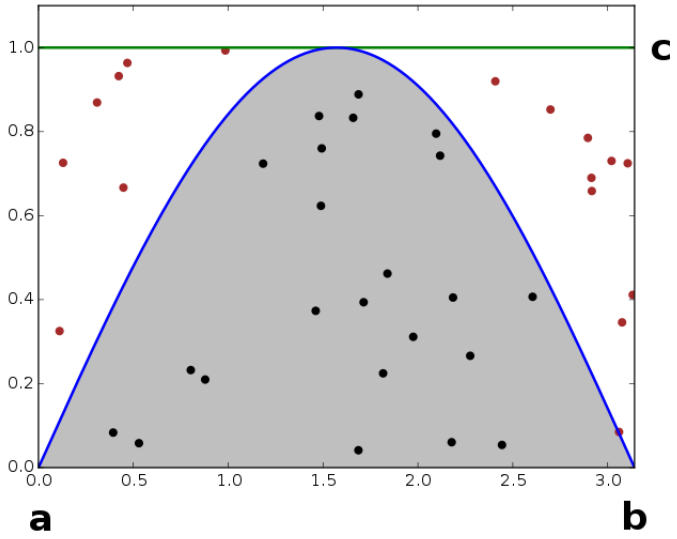

Figura 2.2: Generazione di numeri casuali secondo una predefinita distribuzione: metodo del rigetto. Due numeri casuali estratti da una distribuzione uniforme sono usati per individuare un punto nel piano. Le ascisse dei punti che ricadono sotto la curva costituiscono un insieme di numeri casuali distribuiti secondo la stessa curva.

2.2 Il calcolo di π come esempio di *Monte Carlo*

Il problema che andiamo ora a discutere è anche detto problema dall'**ago di Buffon**. Si tratta di un classico problema posto per la prima volta nel 18-esimo secolo da G.L.Leclerque, conte di Buffon (da cui il nome) ed è un esempio di applicazione di concetti probabilistici alla geometria. Ce ne occupiamo perchè

si presta bene a far vedere come si possono sfruttare fenomeni casuali (o loro simulazioni) per risolvere problemi anche intricati.
Il problema si può esprimere così :

> dato un piano suddiviso da linee parallele uniformemente distanziate a distanza d l'una dall'altra a formare una griglia, trovare la probabilità che, lanciando a caso un "ago" di lunghezza l questo, cadendo sul piano, vada ad intersecare una delle linee tracciate sullo stesso piano.

La Figura 2.3 mostra la geometria che corrisponde a questo problema e ci aiuta ad intuire che si possono verificare due casi a seconda che la lunghezza l dell'ago sia minore o maggiore della spaziatura d tra le linee tracciate sul piano.

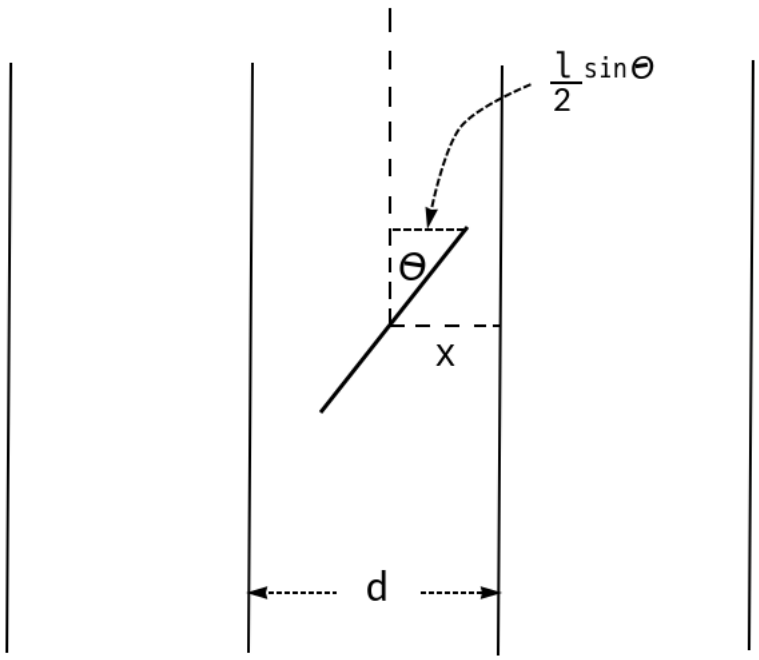

Figura 2.3: Geometria per il problema dell'ago di Buffon nel caso in cui l'ago sia di lunghezza minore della separazione tra linee adiacenti.

Dopo il lancio dell'ago sul piano la sua posizione può essere descritta da due variabili: la distanza x del centro dell'ago dalla linea più vicina e l'orientazione Θ dell'ago rispetto alla stessa linea (Fig. 2.3), intendendo con Θ l'angolo acuto.

Si noti che considerare la linea più vicina corrisponde a constatare che il centro dell'ago potrà trovarsi ad una distanza dalla linea compresa tra 0 e $d/2$.

Relazioni analitiche

Considerando ora le probabilità associate a queste variabili, facciamo le seguenti ipotesi:

- la posizione in cui andrà a cadere l'ago sarà casuale: la probabilità che il centro dell'ago si trovi ad una distanza x dalla linea più vicina è uniforme;

- la probabilità di trovare l'ago orientato secondo un certo angolo Θ, con $0 \leq \Theta \leq \pi/2$, è anch'essa uniforme.

Dato che le due distribuzioni di probabilità sono uniformi ricaviamo facilmente l'espressione per le due probabilità:

$$\begin{aligned} \text{per la } x: \quad & \int_0^{d/2} P(x)dx = 1 \longrightarrow \quad P(x) = 2/d \\ \text{per } \Theta: \quad & \int_0^{\pi/2} P(\Theta)d\Theta = 1 \longrightarrow \quad P(\Theta) = 2/\pi \end{aligned} \tag{2.4}$$

che, nell'ipotesi di indipendenza tra le due variabili x e Θ, possiamo usare per ottenere la probabilità combinata come prodotto:

$$P(x, \Theta) = P(x)P(\Theta) = \frac{4}{d\ \pi} \tag{2.5}$$

Usando questa espressione possiamo ora valutare la probabilità che ha un ago di intersecare una linea in due possibili casi:

- **lunghezza dell'ago minore della distanza tra due linee:** $l < d$
  dato un angolo Θ si vede dalla Fig. 2.3 che c'è una distanza massima dalla linea più vicina $x_{\max} = (l/2)sin\Theta$ al di là della quale l'ago non sarà più sovrapposto alla linea. Per valutare P_{sov}, ovvero la probabilità totale per la sovrapposizione ago-linea dobbiamo allora integrare la (2.5) entro limiti di integrazione appropriati:

$$P_{\text{sov}} = \int_{\Theta=0}^{\pi/2} \int_{x=0}^{(l/2)sin\Theta} \frac{4}{d\ \pi}\ dx\ d\Theta \tag{2.6}$$

cioè integreremo su tutti gli angoli, ma solo sulle distanze che, angolo per angolo, consentono la sovrapposizione ago-linea. Il risultato sarà:

$$P_{\text{sov}} = \frac{4}{d\ \pi}\ \int_{\Theta=0}^{\pi/2} \frac{l}{2} sin\Theta\ d\Theta\ =\ \frac{2\ l}{d\ \pi} \tag{2.7}$$

- **lunghezza dell'ago maggiore della distanza tra due linee:** $l > d$
in questo caso si deve considerare che il limite di integrazione $(l/2)sin\Theta$ usato prima per la variabile x non è più corretto perchè può assumere valori maggiori di $d/2$. Siccome il massimo possibile per la distanza tra centro dell'ago e linea vicina è proprio $d/2$ dovremo tenerne conto dividendo l'intervallo di integrazione sulle x in due parti:

  - una che corrisponde a tutti gli angoli tali che $(l/2)sin\Theta < d/2$, condizione che determina l'angolo critico $\Theta_c = \arcsin(d/l)$
  - un'altra che comprende tutti gli altri casi con $\Theta > \Theta_c$ in cui l'ago incrocia sempre la linea vicina.

Ricalcolando ora la probabilità con questi vincoli abbiamo:

$$P_s = P(x)P(\Theta) = \frac{2}{\pi}\left[\int_{\Theta=0}^{\Theta_c} \frac{l\ sin\Theta}{d} d\Theta + \int_{\Theta_c}^{\pi/2} 1\ d\Theta\right] \tag{2.8}$$

dove il fattore $2/\pi$ che precede corrisponde al valore della $P(\Theta)$ (vedi eq.2.4) mentre l'espressione tra parentesi quantifica la $P(x)$ che è divisa in due pezzi: il primo termine tiene conto delle situazioni simili a quelle del caso $l < d$ (come in eq.2.7), mentre il secondo termine rappresenta il contributo dei casi con angoli maggiori di Θ_c per i quali l'ago incrocia certamente la linea.
Sviluppando gli integrali si ottiene:

$$\begin{aligned} P_{\text{sov}} =& \frac{2}{\pi}\left[\frac{l}{d}\left[-cos\Theta\right]_0^{\arcsin(d/l)} + \left[\Theta\right]_{\arcsin(d/l)}^{\pi/2}\right] \\ =& \frac{2}{\pi}\left[\frac{l}{d}\left(-cos(arcsin(\frac{d}{l})) + 1\right) + \frac{\pi}{2} - \arcsin(\frac{d}{l})\right] \end{aligned} \tag{2.9}$$

Ricordando che $\cos(\arcsin(x)) = \sqrt{1-x^2}$ e che $\arcsin(x) + \arccos(x) = \pi/2$, la precedente diventa:

$$P_{\text{sov}} = \frac{2}{\pi}\left[\frac{l}{d}\left(1 - \sqrt{1 - \frac{d^2}{l^2}}\right) + \arccos(\frac{d}{l})\right] \tag{2.10}$$

dalla quale possiamo verificare che nel limite $l \to d$ la probabilità tende a $2/\pi$ mentre per $l \to \infty$ diventa 1, cioè aghi di lunghezza infinita sicuramente attraversano una linea.

Calcolo numerico

Ora che abbiamo le espressioni della probabilità che gli aghi incrocino una linea della griglia nei due casi ($l < d$ ed $l > d$), notiamo che il valore di π è sempre coinvolto. Questa circostanza ci offre l'occasione di calcolare π sfruttando una valutazione che possiamo fare "sperimentalmente" della probabilità $P_{\rm sov}$ che appare nelle formule (2.7, 2.10). P.es. dalla (2.7) ricaviamo:

$$\pi = \frac{2l}{dP_{\rm sov}} \tag{2.11}$$

per cui lanciando a caso un ago sulla griglia un numero $N_{\rm tot}$ di volte potremmo ottenere una stima della probabilità $\mathrm{P}_{\rm sov}$ semplicemente contando quante volte l'ago lanciato a caso andrà a cadere in modo tale da incrociare una linea. Questo è un modo abbastanza intuitivo (anche detto *frequentista*) di valutare la probabilitaà che stimiamo dal rapporto:

$$P_{\rm sov} = \frac{N_{\rm sov}}{N_{\rm tot}} \tag{2.12}$$

dove con $N_{\rm sov}$ abbiamo indicato il numero di volte che l'ago si sovrappone ad una linea della griglia.

Siccome la lunghezza dell'ago (l) e la distanza (d) tra due righe sono quantità note, non ci resta che valutare la $P_{\rm sov}$, cosa che possiamo fare a tavolino immaginando di eseguire l'esperimento.

In modo analogo è evidentemente possibile calcolare π usando la relazione (2.10) in cui le quantità d ed l sono note e si va a valutare $P_{\rm sov}$ con la stessa tecnica di calcolo usata prima.

Capitolo 3

Analisi delle componenti principali (PCA)

Questa parte di lezioni intende illustrare un metodo di analisi dei dati che va sotto il nome di *Principal Component Analysis* (PCA) ad eventuali lettori con un minimo background nell'ambito del calcolo matriciale. Per facilitare il raggiungimento di una buona consapevolezza di quando, come e perchè l'analisi PCA funziona, seguiremo per quanto possibile un approccio intuitivo che baseremo su alcuni esempi.

Una osservazione o un esperimento vengono efficacemente descritti e hanno "senso fisico" solo quando siamo in grado di acquisire/conoscere i valori di un certo numero di grandezze che abbiamo in qualche modo deciso di utilizzare per caratterizzare il fenomeno indagato. Per esempio, se studiamo i temporali potremmo essere interessati a registrare l'ora, la data, la durata, il numero di fulmini, la quantità d'acqua precipitata, la temperatura al suolo, la pressione atmosferica iniziale e finale, la dimensione dell'area interessata dalla pioggia, la presenza o assenza di grandine, la fase della Luna, ed altre possibili grandezze che riteniamo possano essere legate o possano condizionare in qualche modo il fenomeno che vogliamo studiare.

Questa elencazione, volutamente lunga, fa capire come possiamo facilmente trovarci a dover gestire una grande quantità di dati, dai quali vorremmo provare ad estrarre le informazioni essenziali per indagare sulla natura del fenomeno studiato. L'analisi delle componenti principali (che d'ora in avanti indicheremo con PCA (da *Principal Component Analysis*) viene utilizzata in questi contesti come uno strumento che sfrutta l'**algebra lineare** per individuare le correlazioni di maggiore importanza che si presentano tra tutti i dati accumulati durante un esperimento o, meglio, una serie di esperimenti.

Lo scopo di queste note è duplice: da una parte quello di fornire un esempio intuitivo che faccia comprendere il meccanismo di funzionamento del metodo PCA; d'altra parte quello di discutere la faccenda in modo abbastanza completo, includendo per questo degli elementi di algebra lineare. Questa infatti fornisce gli strumenti di calcolo che trovano una importante applicazione nel risolvere il problema pratico di dare ai dati sperimentali una organizzazione che sia efficiente senza sacrificarne il contenuto informativo.
Il nostro obiettivo è infatti quello di diminuire, per quanto possibile, la dimensionalità[1] dei nostri dati, senza per questo rinunciare al grosso dell'informazione presente negli stessi dati. In pratica vogliamo ridurre il numero di variabili in gioco, senza però perdere la capacità di decrittare il "racconto" che i dati ci danno del fenomeno che ci interessa.
L'approccio che useremo sarà di tipo didattico e quindi solo occasionalmente ci potrà essere qualche indicazione sulle possibili prove che, in modo un po' più rigoroso, riguardano il campo dell'algebra lineare. Queste prove, comunque, possono essere ignorate dal lettore interessato alla sola applicazione pratica del metodo.

3.1 Un caso immaginario

Come abbiamo visto, dopo un dato esperimento ci troviamo in generale a dover maneggiare una gran mole di dati la cui lettura/interpretazione è resa complicata proprio dalla loro grande quantità e dalla loro eventuale ridondanza[2] In senso più tecnico potremmo sospettare che i dati raccolti durante il nostro esperimento non siano del tutto *linearmente indipendenti* gli uni dagli altri.
Per fare chiarezza sarebbe opportuno individuare quali tra le grandezze misurate siano più significative per descrivere il fenomeno osservato o, ancora più in generale, quale combinazione tra grandezze misurate sia più adatta ad una descrizione che mira ad essere la più semplice possibile.
Per fare un esempio pratico immaginiamo una situazione sperimentale ipotetica in cui tre telecamere riprendono, da postazioni diverse, una scena in cui c'è una molla di massa trascurabile fissata da una sua estremità ad una parete. All'altro capo della molla è fissata una massa che, poggiata su un piano senza attrito, viene fatta oscillare. Sappiamo dalla meccanica che la dinamica di un sistema del

[1] Con il termine **dimensionalià** vogliamo indicare il numero di variabili casuali considerate per descrivere un dato esperimento/fenomeno.

[2] La ridondanza qui si riferisce alla possibilità che i nostri dati siano "troppi" nel senso che alcuni di essi potrebbero semplicemente rappresentare una ripetizione di informazioni già acquisite su una data variabile del sistema. Per esempio, acquisire la densità di un gas perfetto è ridondante se sono già note temperatura e pressione.

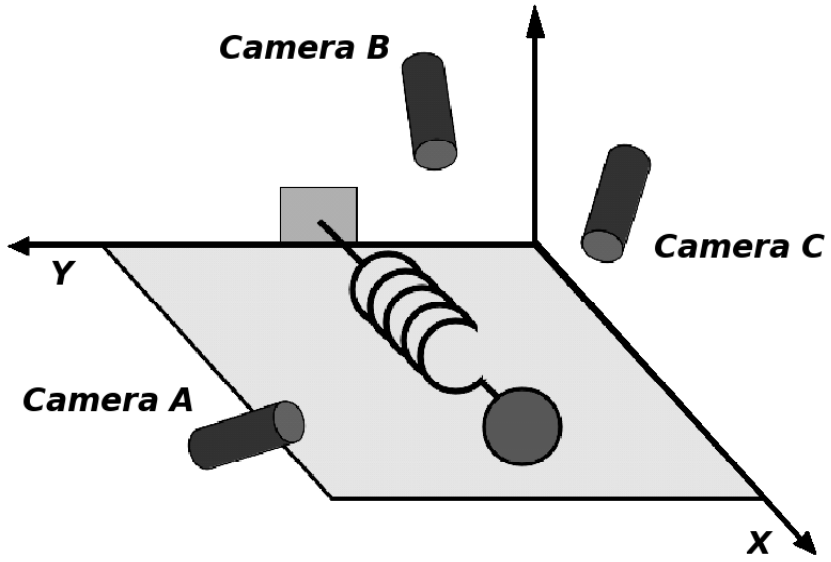

Figura 3.1: Schema dell'esperimento allestito.

genere è monodimensionale e quindi può essere perfettamente descritta usando un'unica variabile associata allo spostamento x.
Se ora noi ignoriamo la meccanica e ci comportiamo ingenuamente, eseguendo l'esperimento potremo acquisire una quantità di immagini bidimensionali da cui possiamo ricavare i valori di coppie di coordinate x, y che individuano la posizione della massa oscillante ad un dato istante di tempo t per ognuna delle tre telecamere usate. Si potrebbe obiettare che posizionando opportunamente una telecamera potremmo eliminare ogni complicazione misurando esattamente quello che serve, cioè la dinamica lungo l'asse x del sistema di riferimento in Fig. 3.1. Tuttavia questo non è quello che succede in generale, visto che nei casi reali noi non sappiamo *a priori* quale tipo di misura sia meglio in grado di catturare la dinamica di un nostro generico sistema fisico. In un caso reale poi interviene un'altra complicazione che è legata alla presenza del "rumore" (che qui intendiamo come incertezza) in ogni segnale misurato, cosa di cui si dovrà anche tener conto per una più verosimile interpretazione dei segnali.

3.1.1 Cambio di base

Da quanto abbiamo detto fin qui si può già intuire che il cambiamento del sistema di riferimento su cui proiettiamo i nostri dati è la chiave per semplificare il nostro

problema. Si può pensare infatti che, procedendo per prove ed errori, prima o poi riusciremo ad individuare un posizionamento di una telecamera tale da vedere il sistema muoversi lungo una sola direzione, riducendo quindi il problema alla considerazione di una sola variabile. L'obiettivo del metodo PCA è proprio l'individuazione di un nuovo set di variabili v', ottenute a partire da combinazioni delle variabili misurate v^m, che possa descivere piuù efficacemente il fenomeno studiato. Con la locuzione "più efficacemente" intendiamo dire che il numero di variabili nuove $N(v')$ necessarie per descrivere compiutamente il sistema dovrà essere minore del numero di variabili misurate nell'esperimento $N(v^m)$. Per rappresentare questo concetto, d'ora in avanti diremo che la nuova descrizione (o anche il set di dati) dovrà avere una minore dimensionalità.
Ritornando ora al nostro ipotetico esperimento da studiare è evidente che le nostre tre telecamere ci daranno un insieme di dati di dimensionalità 6 (2 coordinate per ognuna delle immagini ottenute dalle 3 telecamere) che saranno certamente ridondanti, visto cha già sappiamo che il nostro problema è intrinsecamente unidimensionale. Quello che allora ci aspettiamo è che l'uso del metodo della PCA sia in grado di suggerirci che l'essenza del moto nel nostro esperimento è unidimensionale, anche se partiamo dai nostri dati osservativi che sono invece 6-dimensionali.

Organizzazione dei dati

Prima di procedere conviene a questo punto stabilire un criterio organizzativo per i nostri dati in modo da potervi fare riferimento nel seguito, ogni volta che sarà necessario. Questo criterio è illustrato in Tabella 3.1 dove si vede che una riga j (con $1 \leq j \leq m$) individua un dato parametro misurato n volte, mentre una colonna i (con $1 \leq i \leq n$) individua un particolare set di valori per gli m parametri da noi considerati. Questi ultimi saranno i valori delle grandezze fisiche che abbiamo ottenuto al tempo t_i nello studio del fenomeno/osservazione che ci interessa.

Una base "naive"

In generale allora possiamo immaginare che una osservazione ad un dato istante di tempo t_i produca un set di m misure corrispondenti alle m variabili che abbiamo deciso di osservare durante l'esperimento. Queste costituiscono un vettore di m componenti che possiamo imaginare di rappresentare in uno spazio m-dimensionale. Eseguendo n osservazioni, ognuna ad un dato tempo t_i con

j ↓	**i**	→	**1**	**2**	**3**	...	**n**
	t_i	→	t_1	t_2	t_3	...	t_n
1	par A	→	$A_{1,1}$	$A_{1,2}$	$A_{1,3}$	...	$A_{1,n}$
2	par B	→	$B_{2,1}$	$B_{2,2}$	$B_{2,3}$	...	$B_{2,n}$
3	par C	→	$C_{3,1}$	$C_{3,2}$	$C_{3,3}$	...	$C_{3,n}$
...	⋮	→	⋮	⋮	⋮	⋮	⋮
m	par Z	→	$Z_{m,1}$	$Z_{m,2}$	$Z_{m,3}$	...	$Z_{m,n}$

Tabella 3.1: Schema di organizzazione dei nostri dati in forma di matrice. Il numero **m** di righe rappresenta il numero di parametri con cui caratterizziamo gli esperimenti/osservazioni, mentre il numero **n** di colonne è pari al numero di esperimenti/osservazioni che abbiamo condotto. Con un po' di immaginazione possiamo quindi costruirci uno spazio con **m** dimensioni in cui rappresentiamo gli **n** punti corrispondenti al valore che i parametri assumono nelle **n** colonne della matrice.

$1 \leq i \leq n$, saremo in grado di individuare n vettori nello spazio m-dimensionale che "fotografano" il comportamento del nostro sistema nelle varie osservazioni fatte. In modo equivalente potremmo dire che, data una base ortonormale nello spazio m-dimensionale, ogni osservazione è rappresentata da una particolare combinazione lineare di questi vettori di base. Una scelta "naive" di una base **B** è la matrice identità **I**:

$$\begin{bmatrix} \mathbf{b_1} \\ \mathbf{b_2} \\ \vdots \\ \mathbf{b_m} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 1 \end{bmatrix} = \mathbf{I} \tag{3.1}$$

che è una matrice quadrata in cui ogni riga j rappresenta le m componenti del vettore di base $\mathbf{b_j}$. Come si vede i vettori $\mathbf{b_j}$ hanno una sola componente diversa da zero così che ogni prodotto tra due di essi è nullo. Quindi, siccome il loro modulo è pari ad 1, rappresentano i versori di una base ortonormale.

Riassumendo, durante l'esperimento si acquisiscono tre immagini ad ogni istante di tempo t_i e da ognuna di queste immagini vengono estratti due valori (x, y) che messi insieme per tutte le telecamere costituiscono dei vettori a 6 componenti: $(x_A, y_A, x_B, y_B, x_C, y_C)$. In questo modo, se costruiamo un vettore colonna del

tipo:

$$\mathbf{x} = \vec{x} = \begin{bmatrix} x_A \\ y_A \\ x_B \\ y_B \\ x_C \\ y_C \end{bmatrix} \tag{3.2}$$

vediamo che questo rappresenta l'insieme dei coefficienti nella base "naive" stabilita prima. Come prova basterà moltiplicare la base (3.1) per il vettore (3.2) per ottenere le componenti del vettore rappresentativo di questa osservazione nello spazio a sei dimensioni che abbiamo definito per collocare le nostre misure. Si noterà che le componenti in questo caso sono ovviamente pari alle misure ottenute e quindi adottare la base naive (3.1) non cambia la rappresentazione delle nostre misure.

Una nuova base

A questo punto ci dobbiamo domandare se è possibile individuare un'altra base che, usando una opportuna combinazione lineare dei vettori della base iniziale, possa rappresentare in modo più efficiente l'insieme dei dati sperimentali acquisiti. Notare che qui va sottolineato l'aggettivo **lineare** perchè il metodo PCA è basato proprio sulla notevole semplificazione, indotta da questa proprietà, del metodo che ci porta a definire una base migliore per una rappresentazione più semplice dei nostri dati. Con queste premesse possiamo quindi già dire che il metodo PCA non farà altro che ri-esprimere i dati come combinazione lineare di una nuova base che ci proponiamo di individuare come "migliore"di quella banale mostrata in (3.1).
Prima di procedere ricordiamo subito che nel trattare con matrici 2D adotteremo alcune convenzioni di uso comune:

- il primo indice si riferisce alle righe, il secondo alle colonne della matrice
- le operazioni tra matrici rispettano le regole usuali, in particolare la moltiplicazione segue la regola delle righe×colonne

Siano ora $\mathbf{X}$ ed $\mathbf{Y}$ due matrici n colonne per m righe, legate tra loro da una trasformazione lineare $\mathbf{P}$ che possiamo rappresentare così

$$\mathbf{P}\ \mathbf{X} = \mathbf{Y} \tag{3.3}$$

L'equazione precedente corrisponde proprio ad un cambiamento di base e può essere interpretata in vari modi:

1) $\mathbf{P}$ è una matrice che, moltiplicata ad $\mathbf{X}$ la trasforma in $\mathbf{Y}$

2) da un punto di vista geometrico $\mathbf{P}$ produce una rotazione più uno *stretching*[3] che trasforma $\mathbf{X}$ in $\mathbf{Y}$

3) le righe della matrice $\mathbf{P}$, $\{\mathbf{p_1}, ..., \mathbf{p_m}\}$, sono un set di nuovi vettori di base su cui rappresentare le colonne di $\mathbf{X}$ (che corrispondono ai set di dati raccolti durante l'esperimento, come mostrato nella (3.2).

Per esplicitare il senso dell'ultima interpretazione definiamo le seguenti quantità:

- $\mathbf{p}_j$ siano le righe di $\mathbf{P}$ che sono formate da m elementi. Quindi $\mathbf{P}$, con m righe ed m colonne, è quadrata (come in 3.1);
- $\mathbf{x}_i$ ed $\mathbf{y}_i$ siano rispettivamente le n colonne di $\mathbf{X}$ ed $\mathbf{Y}$, ognuna formata da m elementi (ricordiamo: m sono i nostri 6 parametri ottenuti per ogni osservazione effettuata; n invece individua il numero di osservazioni effettuate, ognuna al suo tempo t_i)

e scriviamo esplicitamente il prodotto (3.3):

$$\begin{aligned} \mathbf{P}\ \mathbf{X} &= \begin{bmatrix} \mathbf{p}_1 \\ \vdots \\ \mathbf{p}_m \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 & \dots & \mathbf{x}_n \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{p}_1\mathbf{x}_1 & \vdots & \mathbf{p}_1\mathbf{x}_n \\ \vdots & \ddots & \vdots \\ \mathbf{p}_m\mathbf{x}_1 & \vdots & \mathbf{p}_m\mathbf{x}_n \end{bmatrix} = \mathbf{Y} \end{aligned} \tag{3.4}$$

Nella precedente notiamo che una colonna della nuova matrice $\mathbf{Y}$ è composta dal prodotto scalare tra le diverse righe della $\mathbf{P}$ ed i dati corrispondenti ad una stessa colonna delle $\mathbf{X}$. Siccome i dati contenuti in una generica colonna $\mathbf{x}_i$ individuano una precisa osservazione, è come dire che stiamo "proiettando" i dati osservati su una base $\mathbf{P}$ attraverso il prodotto scalare. In altre parole non facciamo altro che scrivere in $\mathbf{Y}$ le componenti dei nostri dati $\mathbf{X}$ sulla nuova base $\mathbf{P}$.

Ora che abbiamo familiarizzato con il meccanismo di trasformazione dei dati rispetto ad una nuova base, ci resta da individuare quale sia la base più appropriata per i nostri scopi. Evidentemente la risposta è legata alle caratteristiche che vorremmo per la $\mathbf{Y}$ che sarà la nuova rappresentazione dei nostri dati.

[3] In 2D uno stretching corrisponde ad una trasformazione che dilata/comprime le distanze lungo una data direzione, lasciando invariate quelle nella direzione perpendicolare.

3.1.2 Varianza e covarianza

Ora entriamo nel cuore del problema: a partire dai nostri dati originariamente "ingarbugliati" vogliamo individuare il modo più significativo in cui possiamo esprimere il loro comportamento medio. Per trovare una soluzione procederemo in modo intuitivo, cercando di chiarire di volta in volta le assunzioni che facciamo, per poi illustrare la procedura matematica che ci permette di decifrare i dati. Partiamo quindi considerando che i dati originali sono "ingarbugliati" da due potenziali responsabili che ora andiamo a discutere: il *rumore* (noise) e la *ridondanza*.

Rumore (Noise)

In generale i dati sperimentali sono tanto più significativi (utilizzabili) quanto minore è il contributo del rumore alle misure acquisite. Un parametro cruciale utile per rappresentare la qualità delle nostre misure è il rapporto tra il segnale effettivamente attribuibile alla grandezza fisica misurata ed il rumore che si sovrappone alla misura. Se definiamo allora il rapporto tra la varianza del segnale e quella del rumore come il rapporto segnale/rumore SNR (Signal to Noise Ratio) possiamo scrivere:

$$\mathrm{SNR} = \frac{\sigma^2_{\mathrm{signal}}}{\sigma^2_{\mathrm{noise}}} \tag{3.5}$$

Con questa definizione diremo che un SNR$>>$ 1 indicherà misure di buona qualità, mentre un SNR$\approx$ 1 indicherà misure significativamente contaminate dal rumore. Per rappresentare graficamente questo concetto in Figura 3.2 vediamo come si distribuiscono le misure ottenute da una delle telecamere del nostro esperimento in Figura 3.1.
Idealmente ci aspettiamo che ogni telecamera riveli una relazione perfettamente lineare tra X ed Y, per cui ogni allontanamento da questa linearità lo attribuiamo alla presenza del rumore che sempre coesiste con i dati sperimentali. Le due direzioni principali individuate dai punti possono in questo caso essere interpretate come le segnature del segnale (nella direzione di elongazione massima) e del rumore (in direzione trasversale al segnale). Quindi, per chiarire il senso che diamo alle σ nella (3.5), riportiamo in Figura 3.2 due vettori di lunghezza pari alle varianze dei dati lungo le due direzioni principali. Il rapporto tra i moduli dei due vettori è proprio il rapporto SNR prima definito che qui possiamo interpretare graficamente come una misura di quanto sia "gonfiata" la distribuzione dei punti nel grafico. È evidente che ad un SNR$\gg 1$ corrisponderà un caso in cui i punti si distribuiscono strettamente intorno alla linea del segnale in modo da rendere $\sigma^2_{\mathrm{noise}} \ll \sigma^2_{\mathrm{signal}}$. Nel seguito assumeremo che il nostro apparato strumentale sia

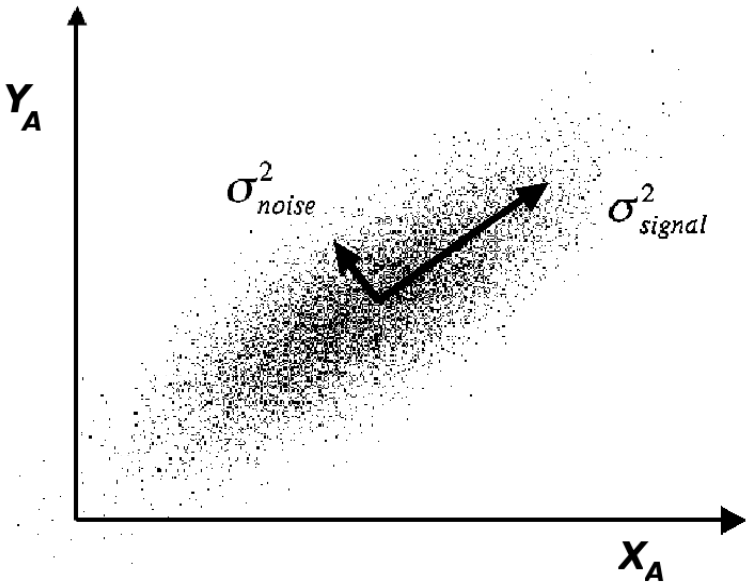

Figura 3.2: Distribuzione delle misure di posizione X ed Y ottenute dalle immagini della telecamera A.

abbastanza efficiente da consentire misure con un alto SNR. Stiamo quindi andando a considerare una situazione in cui i parametri (segnali) acquisiti sono tanti ma sono anche "buoni", nel senso che hanno un buon SNR [4].

Ridondanza

Questo problema emerge chiaramente nel caso dell'esperimento che stiamo considerando. Infatti abbiamo posizionato più sensori (telecamere) che in pratica poi misurano la stessa grandezza, ovvero la posizione di una massa nel tempo. L'informazione acquisita sarà perciò ridondante o, detto in modo diverso, inutilmente abbondante. Una rappresentazione grafica di questo concetto la possiamo realizzare immaginando un plot in cui riportiamo punti corrispondenti a misure ottenute per due diverse grandezze registrate, come mostrato in Figura 3.3.
L'idea che guida il metodo PCA è strettamente legata a questa caratteristica dei dati: vogliamo infatti scoprire quale e quanta ridondanza c'è nei nostri dati per

[4]Notare che qui stiamo considerando come rumore non l'incertezza di una singola misura ma l'allontanamento dei punti osservati dalla tendenza media che essi mostrano lungo l'asse di elongazione principale.

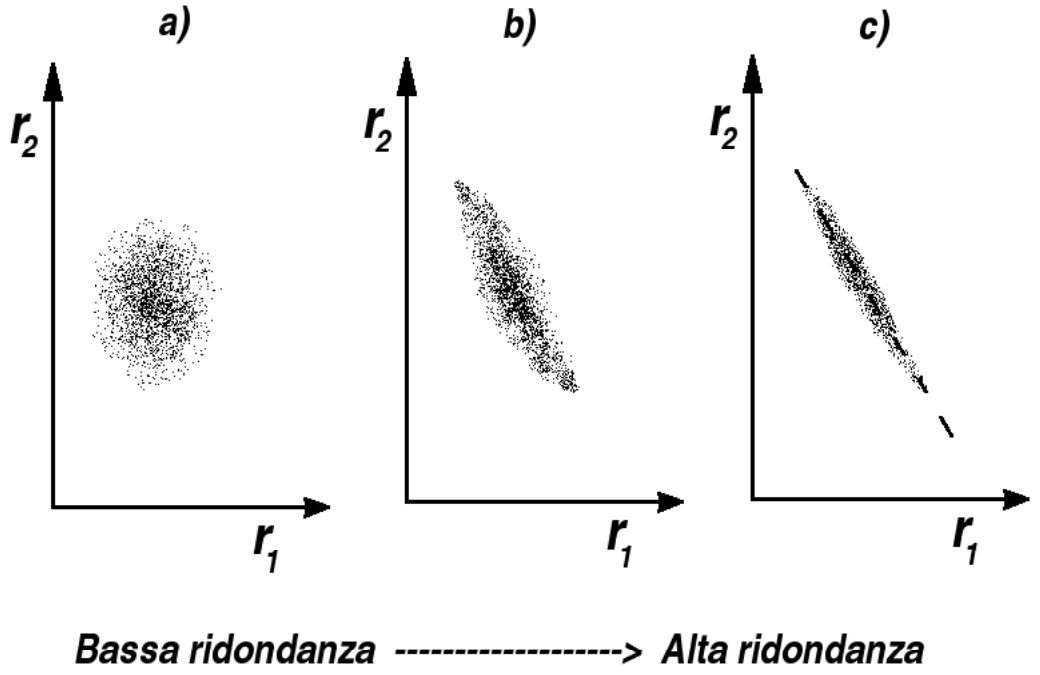

Figura 3.3: Esempi di distribuzione dei dati che si possono osservate registrando due grandezze (r_1, r_2). I casi a), b) e c), rappresentano situazioni a ridondanza crescente. È evidente che nel caso c) basterebbe misurare una delle due grandezze rappresentate per avere informazioni anche sull'altra, rendendo quindi ridondante il considerarle tutte e due. In questi casi conviene descrivere i dati con una sola variabile, che però possa descrivere tutta la dinamica del sistema osservato (legata alla varianza nella direzione del best fit). Sarà quindi una combinazione lineare di r_1, r_2 che massimizza la varianza.

poter poi riesprimere gli stessi dati in modo più conciso, riducendone in questo modo la dimensionalità.

Matrice di covarianza

Abbiamo già visto nella relazione (3.5) che abbiamo usato il rapporto tra varianze per definire il SNR. Dalla Figura 3.2 notiamo però che per valutare le varianze σ^2_{signal} e σ^2_{noise} relative a ciò che noi stiamo considerando "segnale" e "rumore", dobbiamo considerare che i dati sperimentali sono distribuiti sia in X che in Y cosicchè abbiamo a che fare con due variabili contemporaneamente. Se ora immaginiamo un esperimento in cui abbiamo ottenuto due set di misure simultanee A e B (p.es. voltaggio e corrente; pressione e temperatura; oppure x_A e y_A con la telecamera A nel nostro esempio) a media nulla[5] abbiamo:

$$A = \{a_1, a_2, \ldots, a_n\} \quad , \quad B = \{b_1, b_2, \ldots, b_n\} \tag{3.6}$$

le varianze di A e B sono rispettivamente definite da:

$$\sigma^2_A = \langle a_i^2 \rangle_i \quad , \quad \sigma^2_B = \langle b_i^2 \rangle_i \tag{3.7}$$

dove $\langle \ldots \rangle_i$ indica la media dei valori "..." valutata su tutti gli indici $1 < i < n$ e dove si è tenuto conto che A e B sono a media nulla, cioè è stata già fatta la riduzione $a_i \leftarrow (a_i - \bar{a})$. La generalizzazione della varianza quando si usano due variabili combinate è detta *covarianza* ed è definita da:

$$\sigma^2_{AB} = \langle a_i b_i \rangle_i \tag{3.8}$$

Si noti che:

- $\sigma^2_{AB} \to 0$ solo se A e B sono completamente scorrelati dando luogo a termini positivi e negativi che nella media tendono ad annullarsi.
- $\sigma^2_{AB} = \sigma^2_A$, se $A \equiv B$

Proviamo ora ad esprimere la covarianza in forma matriciale e per far questo adottiamo la convenzione che considera le componenti delle misure per i parametri A e B come elementi di vettori di tipo riga:

$$\mathbf{a} = [a_1, a_2, \ldots, a_n] \quad , \quad \mathbf{b} = [b_1, b_2, \ldots, b_n] \tag{3.9}$$

[5]Per ottenere una rappresentazione del segnale a media nulla, basterà sottrarre il valor medio a tutti i valori del set di misure. Questo elimina dai dati il segnale medio costante per meglio evidenziare la variabilità e semplificare la trattazione.

Con questa notazione possiamo esprimere la covarianza come un prodotto di matrici:

$$\sigma^2_{AB} = \frac{\mathbf{ab^T}}{n-1} \tag{3.10}$$

avendo indicato con l'apice $^{\mathbf{T}}$ la matrice trasposta. Ricordando infatti la regola "righe $\times$ colonne" per la moltiplicazione tra matrici, abbiamo che:

$$\mathbf{ab^T} = a_1 b_1 + a_2 b_2 + \ldots, a_n b_n$$

cosicchè, dividendo per $n-1$ come in eq. (3.10), otteniamo proprio la covarianza definita dalla eq. (3.8) [6].

Covarianza tra più misure

Per generalizzare il discorso ad un numero arbitrario di misure (nel nostro esempio avremmo n misure ognuna con 6 grandezze misurate: x_A, y_A, x_B, y_B, x_C, y_C). Le indicheremo in forma compatta riassumendole in altrettanti vettori di tipo riga come in (3.9). Per questo useremo il simbolo $\mathbf{x}_j$ per riferirci al parametro j-esimo che sarà composto a sua volta da n singole misure. Con queste posizioni allora avremo più vettori di tipo riga $\mathbf{x}_j$, di n elementi, che concorrono a definire una matrice che contiene tutta l'informazione su un esperimento:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_m \end{bmatrix} \tag{3.11}$$

Per come è stata costruita, le componenti di questa matrice hanno una interpretazione immediata che riflette l'organizzazione della Tab. 3.1:

- una riga j corrisponde a tutte le misure di un particolare tipo (p.es. la temperatura) ottenute in vari istanti di tempo $t_1, t_2, t_3, \ldots, t_n$;

- una colonna i corrisponde all'insieme delle diverse misure (p.es. temperatura, pressione, voltaggio, ...) ottenute ad un dato istante di tempo t_i nel corso dell'esperimento i-esimo.

[6] Usando dati sperimentali il fattore di normalizzazione per la sommatoria nella 3.8 non è 1/n, ma piuttosto 1/(n-1). Per stimare la varianza è infatti necessario aver già stimato la media, togliendo un "grado di libertà" agli stessi dati

Siamo ora in grado di generalizzare la covarianza, introdotta con la eq. 3.10 tra due set di misure, per definire una "matrice di covarianza" $\mathbf{S_X}$ in questo modo:

$$\mathbf{S_X} \equiv \frac{1}{n-1}\mathbf{X}\mathbf{X^T} \tag{3.12}$$

Sempre usando la regola del prodotto "righe×colonne" possiamo verificare che moltiplicando la prima riga di $\mathbf{X}$ per la prima colonna di $\mathbf{X^T}$ otteniamo la varianza della grandezza espressa dalla prima riga. Infatti, siccome nella trasposta le righe e le colonne sono scambiate, la prima colonna di $\mathbf{X^T}$ sarà uguale alla prima riga di $\mathbf{X}$ cosicchè l'operazione descritta corrisponde alla eq. 3.10 con $\mathbf{a}$=$\mathbf{b}$, rappresentando quindi la varianza $\sigma^2_{1,1}$ dei valori del parametro rappresentato nella prima riga di $\mathbf{X}$.
Il prossimo passo è di moltiplicare la prima riga di $\mathbf{X}$ per la seconda colonna di $\mathbf{X^T}$ dando luogo al termine di covarianza $\sigma^2_{1,2}$ tra la prima e la seconda grandezza fisica utilizzate nell'esperimento. Proseguendo nel calcolo avremo riempito con valori di covarianza una nuova matrice che risulterà quadrata:

$$\mathbf{S_X} = \begin{bmatrix} \sigma^2_{1,1} & \sigma^2_{1,2} & \sigma^2_{1,3} & \dots & \sigma^2_{1,m} \\ \sigma^2_{2,1} & \sigma^2_{2,2} & \sigma^2_{2,3} & \dots & \sigma^2_{2,m} \\ \sigma^2_{3,1} & \sigma^2_{3,2} & \sigma^2_{3,3} & \dots & \sigma^2_{3,m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma^2_{m,1} & \sigma^2_{m,2} & \sigma^2_{m,3} & \dots & \sigma^2_{m,m} \end{bmatrix} \tag{3.13}$$

È da notare che:

- è una matrice $m \times m$, e quindi quadrata, avendo messo in relazione tra di loro le m grandezze fisiche acquisite nell'esperimento;
- la sua diagonale è occupata dalle varianze delle singole grandezze rappresentate nell'esperimento;
- le posizioni fuori dalla diagonale contengono le covarianze tra tutte le coppie di grandezze;
- la matrice è simmetrica in quanto per costruzione $\sigma^2_{i,j} = \sigma^2_{j,i}$;
- un valore grande della covarianza indica una situazione ben correlata del tipo di quella rappresentata nel grafico c) di Figura 3.3.

Possiamo quindi dire che $\mathbf{S_X}$ contiene informazione su tutte le correlazioni esistenti tra coppie di grandezze misurate. Questo è ora il punto: siccome sospettiamo che i nostri dati possano essere affetti da ridondanza, per cercare di ridurla ci apprestiamo a manipolare la $\mathbf{S_X}$ per ottenere una nuova matrice, che indicheremo con $\mathbf{S_Y}$, che abbia delle particolari proprietà.

Diagonalizzazione della matrice di covarianza

Siccome il nostro scopo è di ridurre la ridondanza dei dati, vorremmo trovare delle nuove variabili che mostrino una grande varianza e quindi una grande sensibilità (cioè varino molto) al variare delle condizioni dell'esperimento. Queste nuove variabili vorremmo che siano sempre rappresentative dei nostri dati ma che abbiano una covarianza più piccola possibile per minimizzare quella che abbiamo chiamato ridondanza. Per questo ci aspettiamo che le variabili più indicate per descrivere le nostre osservazioni siano quelle che producono covarianze piccole o nulle tra le diverse misure, come nel caso esemplificato nel grafico a) di Figura 3.3. In un caso del genere ci aspettiamo che la matrice $\mathbf{S_Y}$ di covarianza debba mostrare i corrispondenti valori non-diagonali nulli o molto piccoli.

In questo senso possiamo allora collegare la riduzione di ridondanza con la diminuzione o annullamento delle covarianze, cosa che possiamo ottenere con la diagonalizzazione della matrice di covarianza dei dati originali $\mathbf{S_X}$. Per raggiungere quindi il nostro scopo dobbiamo trovare una nuova proiezione dei nostri dati (su una diversa base) che produca una nuova matrice di dati $\mathbf{Y}$ tale che la sua covarianza $\mathbf{S_Y}$ abbia gli elementi non diagonali nulli (idealmente) o almeno molto piccoli (in pratica).

Questo semplificherà l'analisi di nostri dati perchè una matrice diagonale possiede tutti gli elementi fuori dalla diagonale pari a zero e quindi basterà considerare i soli elementi diagonali per valutare quali siano le nuove variabili più importanti per descrivere il comportamento del sustema studiato. In pratica si individueranno delle nuove variabili[7] che producono la maggiore varianza σ^2_{signal} nei dati e quindi massimizzano il rapporto S/N in eq. (3.5). Tutto ciò ci permette di dire che le nuove variabili saranno più sensibili di quelle originali e quindi più adatte a rivelare e/o descrivere il comportamento del sistema in esame.

[7]Queste saranno combinazioni lineari delle variabili originariamente acquisite negli esperimenti

Possiamo quindi concludere che il metodo PCA si basa essenzialmente sulla ricerca di una base di vettori $\{\mathbf{p}_1,\ldots,\mathbf{p}_m\}$ ortonormali ($\mathbf{p}_i\mathbf{p}_j{=}\delta_{i,j}$) che individuano le direzioni più appropriate per rappresentare il nostro insieme di dati nel modo più essenziale possibile e quindi evitando la ridondanza.

Procedura seguita

Immaginiamo ora di avere uno spazio m dimensionale in cui abbiamo rappresentato i punti corrispondenti alle nostre m grandezze, misurate n volte (come p.es. nello schema di Tabella 3.1). Avremo così una distribuzione di n punti che somiglierà ad una nuvola rappresentata in uno spazio m-dimensionale, più o meno simile a quella mostrata in Figura 3.2 nel caso più facilmente visualizzabile di due sole dimensioni.
Per ridurre la ridondanza il metodo PCA si muove in questo spazio m dimensionale con i seguenti passi:

- individua una direzione nella quale le proiezioni dei punti rappresentati hanno la maggiore varianza (p.es. nella Figura 3.2 questa direzione è quella indicata da σ_{signal}). A questa direzione viene associato il primo vettore della base $\mathbf{p}_1$;

- analogamente al passo precedente si va ad individuare la seconda direzione nella quale la varianza è maggiore (tra tutte quelle rimaste) in modo da individuare così la direzione verso cui puntare il successivo vettore di base $\mathbf{p}_2$;

- iterando il punto precedente individueremo via via i successivi vettori di base che saranno ordinati, per costruzione, rispetto al valore della varianza cioè rispetto all'importanza di quella direzione nell'evidenziare la variabilità dei dati (da cui il termine utilizzato di "componenti principali")

Il fatto che la PCA usi una base ortonormale rende il problema trattabile con i metodi dell'algebra lineare con i quali cercheremo soluzioni algebriche. Queste dovranno tradurre in pratica la nostra richiesta per l'estrazione delle componenti principali dai nostri dati.

Assunzioni e Limiti

Prima di procedere ribadiamo una serie di cose importanti per avere una più piena consapevolezza dei vantaggi e dei limiti del metodo PCA.

1- **Linearità**.
Il cambiamento di base è realizzato attraverso proiezioni lineari dei dati sulla nuova base ortonormale. Per completezza diciamo che c'è anche la possibilità di usare trasformazioni non-lineari ed il metodo che ne deriva viene detto "kernel PCA".

2- **Sufficienza di Media e Varianza**:
Il metodo PCA assume che la media e la varianza descrivano completamente la distribuzione di probabilità dei dati trattati. Si noti che la sola distribuzione a media zero che soddisfa questa richiesta è la Gaussiana, quindi il metodo è strettamente applicabile solo a dati "Gaussiani". D'altra parte la gaussianità dei dati garantisce anche che il nostro modo di definire sia il rapporto segnale/rumore (SNR) che la matrice di covarianza, corrisponde ad una descrizione completa rispettivamente del rumore e della ridondanza.

3- **Varianza come sinonimo di importanza**:
Le grandezze che mostrano la varianza più grande sono quelle che hanno SNR maggiore. In questo senso, le componenti principali sono quelle con le varianze maggiori perchè sono associate a dinamiche dei dati potenzialmente più interessanti. Al contrario le componenti con minore varianza sono associate piuttosto al rumore. Si noti che questa caratteristica può essere sfruttata per ottenere una riduzione del rumore nei dati attraverso una ricostruzione del segnale che usi le sole componenti interessanti, evitando quelle più rappresentative del rumore.

4- **Componenti principali ortogonali**:
Questa assunzione corrisponde ad una semplificazione del nostro problema perchè ci permette di accedere a tecniche di decomposizione basate su operazioni ben note di algebra lineare che andremo ad utilizzare nel seguito.

3.2 PCA e Autovettori di Covarianza

Per ottenere una riduzione della ridondanza dei nostri dati percorreremo due possibili strade. Nella prima, cercheremo una soluzione utilizzando l'algebra lineare. Nella seconda introdurremo una soluzione detta SVD (da Singular Value Decomposition) che coinvolge un più generale e importante metodo di decomposizione.

3.2.1 Soluzione #1: PCA classica

Questa soluzione deriva dal metodo che usa gli autovettori per decomporre una matrice. Riprendiamo la nostra matrice $\mathbf{X}$ dell'eq. 3.11 che ha dimensioni $m \times n$, con m che indica la grandezze misurata ed n il numero di campioni di una stessa grandezza acquisiti durante gli esperimenti/osservazioni. Ci proponiamo di:

> trovare una matrice ortonormale $\mathbf{P}$ sulla quale proiettare i nostri dati $\mathbf{X}$ attraverso il prodotto tra matrici[8] ottenendo la matrice delle proiezioni
>
> $$\mathbf{Y} = \mathbf{PX}$$
>
> Questa deve essere tale che la sua matrice di covarianza (eq. 3.12)
>
> $$\mathbf{S_Y} \equiv \frac{1}{n-1}\mathbf{YY^T}$$
>
> sia diagonalizzata. In questo caso interpreteremo le righe di $\mathbf{P}$ (cioè i vettori della nuova base) come le componenti principali di $\mathbf{X}$, cioè saranno le componenti principali dei nostri dati.

Cominciamo allora riscrivendo la matrice di covarianza $\mathbf{S_Y}$ in termini della nuova base ortonormale che abbiamo scelto di chiamare $\mathbf{P}$:

$$\begin{aligned} \mathbf{S_Y} &= \tfrac{1}{n-1}\mathbf{YY^T} &&= \tfrac{1}{n-1}(\mathbf{PX})(\mathbf{PX})^\mathbf{T} \\ &= \tfrac{1}{n-1}\mathbf{PXX^TP^T} &&= \tfrac{1}{n-1}\mathbf{PAP^T} \end{aligned} \tag{3.14}$$

dove nell'ultimo passaggio abbiamo introdotto la matrice $\mathbf{A} \equiv \mathbf{XX^T}$ che è simmetrica per costruzione come abbiamo già notato nel calcolo della matrice di covarianza (vedi eq. 3.12 e 3.13). A questo punto ricordiamo dall'algebra lineare che una matrice simmetrica come la nostra $\mathbf{A}$ è diagonalizzata da una matrice ortogonale costruita con i suoi autovettori , secondo la relazione

$$\mathbf{A} = \mathbf{EDE^T} \tag{3.15}$$

dove $\mathbf{D}$ rappresenta una matrice diagonale ed $\mathbf{E}$ è una matrice costruita con gli autovettori di $\mathbf{A}$ organizzati in colonne. La matrice $\mathbf{A}$ ha $r \leq m$ autovettori

[8] Analogamente al prodotto scalare tra vettori, il prodotto interno tra matrici $\mathbf{PX}$ corrisponde alla proiezione della matrice $\mathbf{X}$ sulla $\mathbf{P}$

[8] Da un teorema di algebra lineare.

ortonormali, con r detto *rango* della matrice. Se il rango di $\mathbf{A}$ fosse $r < m$ la matrice è detta "degenerata" intendendo con questo che le sue righe (o colonne) non sono tutte linearmente indipendenti. Il rango è anche dato dal massimo numero di linee o colonne linearmente indipendenti per cui, se $r < m$ abbiamo una indicazione che alcune righe (o colonne) della matrice sono combinazione lineare di altre righe (o colonne) e quindi non aggiungono nuova informazione rispetto alle altre righe (o colonne) da cui sono derivate.

Possiamo allora immaginare che i dati in realtà occupino (si possono rappresentare in) un sottospazio di dimensione $r < m$ e quindi il problema andrebbe riformulato in modo da usare solo i parametri linearmente indipendenti ottenendo quindi una matrice $\mathbf{A}'$ di dimensione $r < m$. Tuttavia si può rimediare aggiungendo agli r vettori ortonormali, che si trovano usando gli autovettori della matrice $\mathbf{A}$ diversi da zero, m-r vettori addizionali usati per "riempire" la base ortonormale che stiamo costruendo. Questi vettori addizionali non avranno comunque effetto sulla soluzione finale perchè le varianze associate alle loro direzioni saranno nulle, non essendo i nostri dati distribuiti lungo questi nuovi vettori di base aggiuntivi. A questo punto siamo in grado di procedere in questo modo:

- scegliamo di costruire una matrice $\mathbf{P}$ in modo che ciascuna riga $\mathbf{p}_i$ sia un autovettore della $\mathbf{A}$=$\mathbf{X}\mathbf{X}^{\mathrm{T}}$;
- per quanto stabilito nell'eq. 3.15 il punto precedente implica che possiamo scrivere: $\mathbf{P}\equiv\ \mathbf{E}^{\mathrm{T}}$, trasposta perchè in $\mathbf{P}$ gli autovettori sono organizzati in righe mentre in $\mathbf{E}$ lo sono in colonne;
- sostituiamo nella eq. 3.15 ottenendo $\mathbf{A} = \mathbf{P^T D P}$;
- usiamo la proprietà delle matrici ortogonali per cui la matrice inversa $\mathbf{P}^{-1}$ è uguale alla trasposta $\mathbf{P^T}$ (vedi Appendice).

Possiamo usare ora queste considerazioni per sviluppare la eq. 3.14:

$$\begin{aligned} \mathbf{S_Y} &= \tfrac{1}{n-1}\mathbf{P A P^T} &&= \tfrac{1}{n-1}\mathbf{P}(\mathbf{P^T D P})(\mathbf{P})^{\mathbf{T}} \\ &= \tfrac{1}{n-1}(\mathbf{P P^T})\mathbf{D}(\mathbf{P P^T}) &&= \tfrac{1}{n-1}(\mathbf{P P^{-1}})\mathbf{D}(\mathbf{P P^{-1}}) \qquad (3.16) \\ &= \tfrac{1}{n-1}\mathbf{D} \end{aligned}$$

da cui appare evidente che la scelta fatta per $\mathbf{P}$ è quella che diagonalizza $\mathbf{S_Y}$.
A chiusura di questo primo metodo ricordiamo il senso di quanto abbiamo esposto finora:

- le componenti principali di $\mathbf{X}$(la matrice contenente i dati) sono gli autovettori di $\mathbf{XX^T}$ che vanno a costituire poi le righe di $\mathbf{P}$. Quindi le righe di $\mathbf{P}$ sono i nuovi vettori che costituiscono la base più conveniente per i nostri dati;

- l'i-esimo valore diagonale di $\mathbf{S_Y}$ rappresenta la varianza di dei dati $\mathbf{X}$ proiettati sul nuovo vettore di base $\mathbf{p}_i$.

In definitiva il calcolo della PCA di un set di dati $\mathbf{X}$ richiede di:

1- normalizzare tutte le misure sottraendo le rispettive medie, in modo da trattare con segnali a media nulla;

2- calcolare gli autovettori della matrice di covarianza $\mathbf{XX^T}$. Questi definiscono la nuova base $\mathbf{P}$ più conveniente su cui proiettare i dati originali ottenendo la nuova matrice $\mathbf{Y}$;

3- calcolare la matrice $\mathbf{S_Y}$ i cui elementi diagonali rappresentano le varianze dei dati rispetto ai vettori $\mathbf{p}_i$ della nuova base $\mathbf{P}$.

Quest'ultimo punto permette poi di usare il valore della varianza per quantificare l'importanza delle varie componenti dei dati ottenute sulla nuova base.

> **In conclusione**: La tecnica PCA fu introdotta già nel 1901 come strumento per l'analisi di dati multivariati. Le componenti principali sono rappresentate dagli autovettori della matrice di covarianza dei dati. La proiezione dei dati sugli assi definiti dalle componenti principali viene anche detta trasformata di Hotelling o anche di Karhunen-Loeve. Come abbiamo visto il metodo consiste nel trovare la trasformazione ortogonale che diagonalizza la matrice di covarianza dei dati.