Capitolo 3 – Covarianza, correlazione, best-fit lineari e non lineari

1) Covarianza e correlazione

Ad un problema si associa spesso più di una variabile quantitativa (es.: di una persona possiamo determinare peso e altezza, oppure QI e reddito; di un segnale spettroscopico si può operare una deconvoluzione in termini di posizione dei picchi e loro ampiezza, etc.).

Consideriamo il caso di due variabili x, y e indichiamo con (x_i, y_i) i valori relativi al caso i-esimo di un insieme di N dati. Per ciascuno dei set (x_i) e (y_i) possiamo definire media e varianza come visto nel capitolo 1

$$<x> \quad \sigma_x^2 \qquad\qquad <y> \quad \sigma_y^2$$

Possiamo ora definire una quantità, chiamata **covarianza**, come

$$\sigma_{xy} \equiv \frac{1}{N}\sum_{i=1}^{N}(x_i - < x >)(y_i - < y >) = \qquad (1)$$

$$= \frac{1}{N}\sum_{i=1}^{N} x_i y_i - < x >< y > - < x >< y > + < x >< y > = < xy > - < x >< y >$$

Se non c'è nessuna relazione tra le due variabili si dice che x e y sono **indipendenti** e la loro **covarianza è nulla**.

Esempio: una persona più alta della media, non è detto che abbia un QI elevato. Viceversa, una persona più alta della media peserà probabilmente più della media. La covarianza, in questo caso, è un *numero positivo* perché nella maggior parte dei casi $(x_i - < x >)$ e $(y_i - < y >)$ saranno ambedue positivi (per persone alte e pesanti) o ambedue negativi (per persone piccole e leggere).

Se x e y sono variabili *non indipendenti* che ha senso sommare o sottrarre, la varianza di $x \pm y$ non è più data semplicemente da $\sigma_x^2+\sigma_y^2$. Infatti

$$N\sigma_{x\pm y}^2 \equiv \sum_{i=1}^{N}\left(x_i \pm y_i - \left(< x > \pm < y >\right)\right)^2 = \sum_{i=1}^{N}\left(\left(x_i - < x >\right) \pm \left(y_i - < y >\right)\right)^2 =$$

$$= \sum_{i=1}^{N}\left(\left(x_i - < x >\right)^2 + \left(y_i - < y >\right)^2 \pm 2\left(x_i - < x >\right)\left(y_i - < y >\right)\right) =$$

$$= N\left(\sigma_x^2 + \sigma_x^2 \pm 2\sigma_{xy}\right) \qquad (2)$$

Le interrelazioni tra due variabile possono essere descritte da una quantità *legata alla covarianza* e detta **coefficiente di correlazione** ρ

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \qquad (3)$$

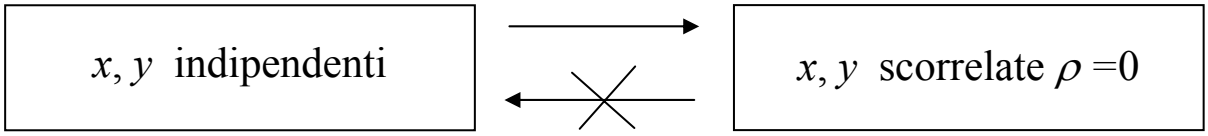
che ha la proprietà di avere **valore assoluto** sempre ≤ 1.

- Se ρ = 1 le variabili si dicono **correlate positivamente** in modo perfetto.

- Se ρ = −1 le variabili si dicono **correlate negativamente** in modo perfetto.

- Se x e y sono **indipendenti**, allora i valori attesi di covarianza e correlazione sono nulli.

N.B. Se la correlazione ha *valore atteso nullo*, le variabili si dicono **scorrelate**, ma non è detto che siano indipendenti. Infatti un sottogruppo potrebbe avere *covarianza negativa* ed essere compensato esattamente da un altro sottogruppo a *covarianza positiva*.



Covarianza e *correlazione* possono essere definite sia su un insieme di coppie di dati (x,y) sia per una distribuzione continua $f(x,y)$. Così come la varianza misura l'ampiezza delle fluttuazioni attorno al valore medio, la correlazione misura il *collegamento* tra fluttuazioni di x e fluttuazioni di y.

2) Best-fit di dati sperimentali

L'andamento dei *dati sperimentali* o un *modello teorico* possono suggerire che le variabili x e y siano legate da una relazione *nota* del tipo

$$y = f(x, \mathrm{A}, \mathrm{B}, \ldots \mathrm{N})$$

dove gli A, B, …, N sono parametri in generale non noti. Il problema, in questo caso, consiste nel determinare l'insieme dei valori dei parametri per cui la curva $y = f(x)$ "passi il più vicino possibile" ai punti (x_i, y_i). La curva così ottenuta prende il nome di **best-fit** o **curva di regressione dei minimi quadrati**.

Effettuare una **regressione di *y* su *x*** significa prendere in considerazione le **differenze tra ordinate** di *punto* e *curva*

$$\delta y = y_i - f(x_i) \tag{4}$$

e **minimizzare la somma delle deviazioni al quadrato** (applicazione del principio dei minimi quadrati)

$$S_y(\mathrm{A}, \mathrm{B}, \ldots, \mathrm{N}) = \sum_{i=1}^{n} (y_i - f(x_i))^2 \tag{5}$$

Viceversa, se è definibile la funzione inversa $x = f^{-1}(y)$, si può effettuare la **regressione di *x* su *y*** minimizzando la somma dei quadrati delle deviazioni δx_i delle ascisse

$$S_x(\mathrm{A}, \mathrm{B}, \ldots, \mathrm{N}) = \sum_{i=1}^{n} \left(x_i - f^{-1}(y_i)\right)^2 \tag{6}$$

2.1 La regressione lineare

Il caso più importante nella pratica è quello in cui la *relazione attesa* tra le variabili x e y è **di tipo lineare**

$$y = \mathrm{A} + \mathrm{B}x \tag{7}$$

e si minimizzano le deviazioni delle ordinate. In questo caso si parla di **best-fit** (o **regressione**) **lineare**.

N.B. Alcuni calcolatori tascabili usano la notazione A, B; altri indicano A come *intercept* (intercetta con l'asse y) e B come *slope* (pendenza della retta).

La soluzione del problema di *regressione lineare di y su x* è esprimibile in termini di **medie, varianze e covarianza** calcolate sul set di dati (x_i, y_i). Per ottenere A e B si impone che sia nulla la media delle deviazioni δy

$$< \delta y >=< \left(y - \left(\mathrm{A} + \mathrm{B}x\right)\right) >=< y > -\mathrm{A} - \mathrm{B} < x >= 0 \tag{8}$$

$$\Rightarrow \qquad < y >= \mathrm{A} + \mathrm{B} < x > \tag{9}$$

e che sia nulla la derivata parziale $\dfrac{\partial S_y(\mathrm{A},\mathrm{B})}{\partial \mathrm{B}}$, come deve essere se S_y ha un minimo in (A, B)

$$\frac{\partial}{\partial \mathrm{B}} \sum_{i=1}^{n} \left(y_i - \mathrm{A} - \mathrm{B}x_i\right)^2 = \sum_{i=1}^{n} 2\left(y_i - \mathrm{A} - \mathrm{B}x_i\right)x_i = 2\sum_{i=1}^{n} \left(x_i y_i - \mathrm{A}x_i - \mathrm{B}x_i^2\right) = 0$$

$$\Rightarrow \qquad < xy >= \mathrm{A} < x > + \mathrm{B} < x^2 > \tag{10}$$

Mettendo a sistema la (9) e la (10) si ottiene

$$< xy >= \left(< y > -\mathrm{B} < x >\right) < x > +\mathrm{B} < x^2 > \quad \Rightarrow$$

$$\mathrm{B} = \frac{< xy > - < x >< y >}{< x^2 > - < x >^2} = \frac{\sigma_{xy}}{\sigma_x^2} \tag{11}$$

$$\mathrm{A} =< y > -\mathrm{B} < x > \tag{12}$$

N.B. I calcolatori tascabili utilizzano relazioni di questo tipo per il calcolo di A e B.

2.1.1 *Interpretazione degli scostamenti tra le ordinate sperimentali e quelle della retta di best-fit* ***(nota per i più esperti)***

Indichiamo con un asterisco i valori assunti dalle ordinate della retta di best-fit in corrispondenza ai valori della variabile indipendente x_i

$$y_i^* = \mathrm{A} + \mathrm{B}x_i \tag{13}$$

e costruiamo la varianza di y^*-y, *quantità a media nulla* che misura come le ordinate dei punti si scostano dalle corrispondenti ordinate della retta di best-fit

$$N\sigma^2_{y^*-y} = \sum_{i=1}^{N}\left(y_i - \mathrm{A} - \mathrm{B}x_i\right)^2 = \sum_{i=1}^{N}\left(y_i - \left(< y > -\mathrm{B} < x >\right) - \mathrm{B}x_i\right)^2 =$$

$$= \sum_{i=1}^{N}\left(\left(y_i - < y >\right) - \mathrm{B}\left(x_i - < x >\right)\right)^2 = \sum_{i=1}^{N}\begin{pmatrix}\left(y_i - < y >\right)^2 + \mathrm{B}^2\left(x_i - < x >\right)^2 - \\ - 2\mathrm{B}\left(y_i - < y >\right)\left(x_i - < x >\right)\end{pmatrix}$$

(14)

da cui, utilizzando le definizioni di *varianza*, *covarianza* e la equazione (11) si ottiene

$$\sigma^2_{y^*-y} = \sigma_y^2 + \mathrm{B}^2\sigma_x^2 - 2\mathrm{B}\sigma_{xy} = \sigma_y^2 + \frac{\sigma_{xy}}{\sigma_x^4}^2 \sigma_x^2 - 2\frac{\sigma_{xy}}{\sigma_x^2}\sigma_{xy} =$$

$$= \sigma_y^2 - \frac{\sigma_{xy}^2}{\sigma_x^2} = \sigma_y^2\left(1 - \frac{\sigma_{xy}^2}{\sigma_x^2\sigma_y^2}\right) = \sigma_y^2\left(1-\rho^2\right)$$

(15)

La quantità $\frac{\sigma^2_{y^*-y}}{\sigma_y^2} = \left(1-\rho^2\right)$ è interpretabile come **frazione** della varianza di y che è **associata alle deviazioni dalla retta** di best-fit. Invece ρ^2 ha il significato di parte della varianza di y_i dovuta alle variazioni di x_i.

Il quadrato del coefficiente di correlazione spiega la parte della varianza di y che è dovuta alle variazioni della variabile indipendente.

N.B. L'analisi di correlazione non ha senso con meno di tre punti sperimentali. Per N=1 varianza e covarianza sono nulle. Per N=2 il coefficiente di correlazione vale sempre ±1.

2.2 La regressione non-lineare

Supponiamo di avere una qualsiasi relazione teorica tra due quantità x e y e n parametri fisici $k_1, \ldots, k_n$

$$y = f(x_1, k_1, \ldots, k_n)$$

In aggiunta, supponiamo di avere m coppie di punti (x_i, y_i) determinati sperimentalmente. Partendo dalla *relazione teorica* e dai *punti sperimentali*, il set di parametri fisici può essere determinato sotto la condizione che sia **minima la somma dei quadrati degli scarti tra le ordinate teoriche e quelle sperimentali** (principio dei minimi quadrati)

$$S = \sum_{i=1}^{m} \left(y_{i,theor} - y_{i,exp} \right)^2 \tag{16}$$

L'equazione (1) può essere scritta come

$$S = \sum_{i=1}^{m} \left(f\left(x_i, k_1, \ldots, k_n\right) - y_i \right)^2 \tag{17}$$

Dove la grandezza S *rappresenta* una "ipersuperficie" nello spazio n-dimensionale dei parametri k. Il minimo della funzione S viene ottenuto ponendo uguali a zero le sue derivate parziali rispetto ai parametri k_n

$$\frac{\partial S}{\partial k_1} = \frac{\partial S}{\partial k_2} = \ldots = \frac{\partial S}{\partial k_n} = 0 \tag{18}$$

N.B. La funzione f prende il nome di "funzione oggetto"

Si ottiene così un sistema di n equazioni differenziali non-lineari accoppiate, del tipo

$$\frac{\partial S}{\partial k_i} = 2\sum_{i=1}^{m}\left(f\left(x_i, k_1, ..., k_n\right) - y_i\right) \times \frac{\partial f\left(x_i, k_1, ..., k_n\right)}{\partial k_i} = 0 \qquad (19)$$

la cui soluzione fornisce il set di parametri $k_1, \ldots, k_n$ che **offrono il miglior accordo possibile** tra *dati sperimentali* e *modello teorico*

best-fit non lineare multi-parametrico

Problemi (seri):

1. Come si fa a risolvere il set di equazioni differenziali alle derivate parziali?
2. Esiste un solo minimo della funzione S?

2.3 Metodi di ottimizzazione (cenni)

Definizioni:

1) *Vettore dei parametri*

$\boldsymbol{k} = (k_1, \ldots, k_n)$
n = numero di gradi di libertà del sistema (degrees of freedom)

2) *Configurazione di partenza*

Set dei valori inizialmente assegnati alle componenti del vettore $\boldsymbol{k}$
$\boldsymbol{k}^{(0)} = (k_1^{(0)}, \ldots, k_n^{(0)})$

3) *Vincoli*

$a_i \leq k_i \leq b_i$ (ottimizzazione vincolata)

2.3.1 Ottimizzatori deterministici (Higher-Order Deterministic Optimization methods, HODOM)

I metodi HODOM costruiscono una **sequenza** di "soluzioni" $\boldsymbol{k}^{(p)}$ mediante *l'equazione ricorsiva*

$$\boldsymbol{k}^{(p+1)} = \boldsymbol{k}^{(p)} + a^{(p)}\boldsymbol{s}^{(p)} \tag{20}$$

a partire da una scelta iniziale $\boldsymbol{k}^{(0)}$. Qui p indica il numero d'ordine della iterazione corrente, $a^{(p)}$ indica lo "*step*" (a), e $\boldsymbol{s}^{(p)}$ la "*slope*" (b).

(a) Il valore dello *step* determina la **precisione del processo di ottimizzazione**, nel senso che step più piccoli corrispondono a una più fine esplorazione della ipersuperficie dello "spazio delle configurazioni" (funzione S). Ovviamente, ad uno *step più piccolo corrispondono tempi di calcolo più lunghi*.

(b) La scelta del tipo di "*slope*" caratterizza i vari metodi proposti in letteratura. Ad esempio, nel caso in cui la slope coincida con la *direzione* lungo cui la **funzione decresce più rapidamente**

$$\boldsymbol{s}^{(p)} = -\text{grad}\, f(\boldsymbol{k}^{(p)}) \tag{21}$$

si ha il metodo cosiddetto di "*steepest-descent*".
Altre scelte comuni della *dipendenza funzionale della slope* si basano su un **calcolo alle differenze finite** delle derivate della *funzione errore S* e dànno origine ai metodi di Newton, quasi-Newton, gradiente coniugato, etc.

N.B. risolvere una equazione differenziale tramite una *stima alle differenze finite* significa **sostituire** al differenziale ∂x l'incremento Δx e, di conseguenza, sostituire l'integrale con una sommatoria su un numero **finito** di termini.

2.3.2 Ottimizzatori statistici (Zeroth-Order Stochastic Optimization Methods, ZOSOM)

Questi metodi sono basati su una tecnica conosciuta come *simulated annealing* ("tempera" simulata) e, pur avendo un **comune fondamento matematico**, possono assumere nomi diversi: *algoritmi genetici*, *strategie evolutive*, *Montecarlo*, etc.

In generale, questi metodi sono caratterizzati dall'avere definito alla p-esima iterazione un set di μ vettori "padri", $\boldsymbol{k}_1^{(p)},...,\boldsymbol{k}_\mu^{(p)}$, da cui vengono generati λ vettori "figli", $\boldsymbol{y}_1^{(p)},...,\boldsymbol{y}_\lambda^{(p)}$ sulla base di una certa *funzione densità di probabilità* ρ. Ad esempio, la probabilità di avere un certo "figlio" vicino al punto $\boldsymbol{k}_1$ può essere scelta **proporzionale** alla *funzione multidimensionale Gaussiana*

$$\rho(\boldsymbol{k},m,d)=\exp\left(-\sum_i\left(\frac{\boldsymbol{k}_i-m_i}{d_i}\right)^2\right) \tag{22}$$

dove m_i è una media operata sui vettori "padri" $\boldsymbol{k}_i$ e d_i sono le componenti di un vettore "varianza" aggiustato ad ogni iterazione.

1) Si ha una strategia di tipo (μ+λ) se alla iterazione p+1 i nuovi "padri" saranno costituiti dai μ vettori che forniscono i più piccoli valori della funzione errore S, nell'ambito di un set comprendente sia i "padri" sia i "figli" dell'iterazione p-esima.

2) Si ha una strategia di tipo (μ,λ) se i migliori μ vettori sono scelti nell'insieme dei "figli" (ovviamente, in questo caso dovremo avere $\lambda > \mu$).

N.B. Il metodo di Montecarlo è un esempio di strategia (1+1).

Regole di modifica delle varianze, d_i

- Se i vettori “padri” sono vicini al minimo della funzione *S* e l'intervallo di ricerca è molto maggiore della distanza dal minimo stesso, soltanto una piccola frazione dei “figli” sarà migliore dei “padri” e, di conseguenza, si dovrà ridurre il valore di $|d|$;
- Se, al contrario, i “figli” sono spesso migliori dei “padri”, occorrerà aumentare la varianza per esplorare regioni sufficientemente lontane dalla media *m*.

I parametri della *strategia evolutiva* su cui è necessario operare per **massimizzare l'efficienza** dell'ottimizzatore sono, tipicamente,

1. la lunghezza della storia evolutiva,
2. la modifica percentuale della varianza,
3. il tasso critico di successi che determina un aumento dell'intervallo di ricerca.

Vantaggi degli ottimizzatori di tipo ZOSOM rispetto agli HODOM

1. sono più efficienti nel raggiungere un *minimo globale* o, perlomeno, un *minimo locale molto stabile*;
2. la precisione della soluzione è superiore, non essendo influenzata dal calcolo numerico delle derivate (differenze finite);
3. il metodo può essere applicato anche se la derivata della funzione *f* non esiste;
4. non si hanno soluzioni oscillatorie o percorsi a zig-zag (a causa della *natura statistica* della generazione dei vettori);
5. il trattamento delle condizioni al contorno (vincoli) è solitamente molto più semplice rispetto agli HODOM: basta eliminare dall'evoluzione i vettori con componenti al di fuori del dominio previsto;
6. poichè la tecnica di ricerca casuale, per sua natura, **esplora simultaneamente** tutte le *dimensioni dello spazio dei parametri*, il tempo di calcolo è sostanzialmente indipendente dal numero dei parametri stessi.