1. Minimizzazione non lineare

Spesso la funzione che si introduce nel χ^2 non è lineare nei parametri. Ne abbiamo avuto un esempio negli esercizi 9.6 e 9.7, dove il modello introdotto nella funzione da minimizzare era rappresentato dalle funzioni:

$$\mu(x_i;\mu,\sigma) = 1000 \cdot \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right] \cdot 5 \ ,$$

$$\mu(x_i;p) = N \, \frac{10!}{x_i!(10-x_i)!} \, p^{x_i}(1-p)^{10-x_i} \ ,$$

che, come si vede, non sono affatto lineari nei parametri μ, σ e p.

Abbiamo a suo tempo già risolto questi esercizi, utilizzando un codice di minimizzazione non lineare, commentandone dal punto di vista statistico i risultati. In questo paragrafo intendiamo descrivere sommariamente gli algoritmi utilizzati da questi programmi, senza però entrare in eccessivi dettagli di calcolo numerico, che sono trattati in altri testi come [3] o [12].

Il problema di risolvere è dunque quello di trovare il minimo di una funzione di p variabili, ad esempio $\chi^2(\boldsymbol{\theta})$, che sono i parametri liberi del fit.

Gli algoritmi più efficienti sono basati sullo sviluppo della funzione da minimizzare in serie di Taylor. Poiché nell'intorno del minimo le derivate prime tendono ad annullarsi, in genere lo sviluppo è fatto fino alla derivata seconda. In una dimensione avremo ovviamente:

$$\chi^2(\theta) = \chi^2_0 + b\,\theta + \frac{1}{2} c\,\theta^2 \ , \tag{1.1}$$

dove b e c sono rispettivamente la derivata prima e seconda del χ^2 calcolate nel punto iniziale θ_0, che per comodità abbiamo preso pari a zero, mentre $\chi^2_0 = \chi^2(\theta_0)$. Se $c > 0$, la funzione ha un minimo in $\theta = -b/c$; infatti

$$\frac{\mathrm{d}\chi^2}{\mathrm{d}\theta} = b + c\,\theta = 0 \quad \Longrightarrow \quad \theta = -\frac{b}{c} \ . \tag{1.2}$$

Condizione necessaria per essere in un intorno del minimo è che $c > 0$. Quando $c < 0$ bisogna procedere per passi nel verso dato da $-b$ finché $c > 0$; infatti, il segno della derivata prima indica il verso di crescita della funzione,

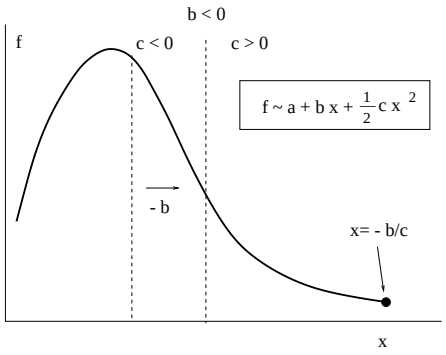

Fig. 1.1. Ricerca del minimo in una dimensione

il segno opposto quello di diminuzione. Questa proprietà è valida anche in più dimensioni se si considera il vettore gradiente della funzione: *l'opposto del gradiente è un vettore che punta sempre verso il minimo.* Operando quindi per passi successivi col *metodo del gradiente*, si va in un punto dove la derivata seconda è positiva e poi ci si sposta nel minimo (1.2); si continua così finché la funzione non riprende a crescere, trovando poi il minimo con delle approssimazioni paraboliche. Il procedimento è schematizzato in fig. 1.1.

La generalizzazione in più dimensioni di questo metodo si basa sull'espressione:

$$\chi^2(\boldsymbol{\theta}) = \chi_0^2 + \left[\boldsymbol{\nabla}\chi^2(\boldsymbol{\theta_0})\right]^\dagger \cdot \boldsymbol{\theta} + \boldsymbol{\theta}^\dagger G \boldsymbol{\theta} \ , \tag{1.3}$$

dove il simbolo $\boldsymbol{\nabla}$ indica il gradiente e G la matrice quadrata delle derivate seconde del χ^2 rispetto ai parametri. La condizione di minimo equivalente alla (1.2) è ora data da:

$$\boldsymbol{\theta} = -G^{-1}\boldsymbol{\nabla}\chi^2(\boldsymbol{\theta_0}) \ , \tag{1.4}$$

mentre la condizione sulle derivate seconde diventa:

$$\boldsymbol{\theta}^\dagger G \boldsymbol{\theta} > 0 \ , \tag{1.5}$$

il che significa che *la matrice G deve essere definita positiva.*

Analogamente al caso monodimensionale appena discusso, i codici di minimizzazione in genere verificano se sono o meno in una regione dove G è definita positiva. Tra le condizioni necessarie e sufficienti perché una matrice sia definita positiva, quelle più usate richiedono che tutti gli autovalori siano positivi, oppure che tutti i determinanti delle sottomatrici quadrate comprendenti l'elemento g_{11} (dette matrici *upper left*, superiori sinistre) siano positivi. In alternativa, viene usato il metodo del gradiente finché la condizione non è verificata, poi si applica la (1.4), continuando ad incrementare i parametri finché la funzione comincia a crescere. In pratica questa procedura viene quasi sempre realizzata ridefinendo la matrice G come:

$$G^{-1} \quad \rightarrow \quad (G + \lambda I)^{-1} \ , \tag{1.6}$$

dove, nel caso G non sia definita positiva, λ è più grande del valore assoluto dell'autovalore più negativo di G. Quando λ è molto grande, G diventa trascurabile e la (1.6) corrisponde in pratica al metodo del gradiente. Man mano che gli autovalori di G diventano meno negativi λ diminuisce e la ricerca del minimo utilizza progressivamente il metodo dello sviluppo al second'ordine [12].

Vediamo ora, un po' più in dettaglio, questi principi applicati alla minimizzazione di funzioni tipo χ^2. Lo sviluppo (1.3) può essere scritto esplicitamente come:

$$\chi^2 \simeq \chi_0^2 + \sum_j \frac{\partial \chi^2}{\partial \theta_j}\,\theta_j + \frac{1}{2}\sum_{jk} \frac{\partial^2 \chi^2}{\partial \theta_j \partial \theta_k}\,\theta_j \theta_k \ . \tag{1.7}$$

Nel caso che la funzione χ^2 sia del tipo:

$$\chi^2(\boldsymbol{\theta}) = \sum_i \frac{[y_i - f(x_i, \boldsymbol{\theta})]^2}{\sigma_i^2} \equiv \sum_i H_i^2 \ ,$$

dove l'indice i si riferisce ai punti misurati, le derivate seconde assumono la forma:

$$\begin{aligned} \frac{\partial^2 \chi^2}{\partial \theta_j \partial \theta_k} &= \frac{\partial}{\partial \theta_j}\frac{\partial}{\partial \theta_k}\sum_i H_i^2 = \frac{\partial}{\partial \theta_j}\sum_i 2\,H_i \frac{\partial H_i}{\partial \theta_k} \\ &= 2\sum_i \frac{\partial H_i}{\partial \theta_j}\frac{\partial H_i}{\partial \theta_k} + 2\sum_i H_i\,\frac{\partial^2 H_i}{\partial \theta_j \partial \theta_k} \ . \end{aligned} \tag{1.8}$$

Se si trascura il secondo termine, si ottiene l'approssimazione:

$$\frac{\partial^2 \chi^2}{\partial \theta_j \partial \theta_k} \simeq 2\sum_i \frac{\partial H_i}{\partial \theta_j}\frac{\partial H_i}{\partial \theta_k} = 2\sum_i \frac{1}{\sigma_i^2}\frac{\partial f_i}{\partial \theta_j}\frac{\partial f_i}{\partial \theta_k} \ . \tag{1.9}$$

L'esperienza ha mostrato che con questa approssimazione vi sono molti vantaggi: l'algoritmo risulta più rapido, accurato quanto gli algoritmi che ricorrono a sviluppi di ordine superiore e la *matrice delle derivate seconde è sempre definita positiva* [5, 7].

A questo punto vi consigliamo di rivedere la (10.94), che mostra come i codici di minimizzazione si servono delle stesse approssimazioni utilizzate nel teorema 10.3.

È importante notare che annullare le derivate seconde rispetto ai parametri θ della funzione

$$H_i(\boldsymbol{\theta}) = \frac{y_i - f(x_i, \boldsymbol{\theta})}{\sigma_i}$$

corrisponde a supporre che le derivate seconde della funzione $f(x, \boldsymbol{\theta})$ siano nulle, ovvero che f sia esprimibile, nell'intorno di un valore di partenza $\boldsymbol{\theta}^*$, con uno sviluppo di Taylor al primo ordine *lineare nei parametri*:

$$f(x,\boldsymbol{\theta}) = f(x,\boldsymbol{\theta}^*) + \sum_k \left.\frac{\partial f}{\partial \theta_k}\right|_{\theta_k^*} \Delta\theta_k \ , \tag{1.10}$$

dove $\Delta\theta = (\theta - \theta^*)$. In questo caso la (1.7) è una relazione esatta, in quanto le derivate del χ^2 di ordine superiore al secondo sono nulle. Se ora teniamo presente le formule viste nel paragrafo precedente sui minimi quadrati lineari e operiamo nelle (10.63, 10.65) le sostituzioni:

$$\begin{aligned} \theta_k &\rightarrow \Delta\theta_k = \theta_k - \theta_k^* \ , \\ y_i &\rightarrow y_i - f(x_i,\boldsymbol{\theta}^*) \ , \\ f_k(x_i) &\rightarrow \left.\frac{\partial f}{\partial \theta_k}\right|_{\theta_k^*} \ , \end{aligned}$$

vediamo che le equazioni risolutive, passo per passo, sono identiche alle (10.75-10.77). La matrice F della (10.70), in accordo con la (**??**), contiene ora le p derivate prime di $f(x,\boldsymbol{\theta})$ calcolate nel punto $\boldsymbol{\theta}^*$ e negli n punti sperimentali x_i, e la matrice $\Sigma = (F^\dagger W F)$ della (10.78), che in questo caso viene detta matrice di curvatura, contiene le derivate seconde del χ^2 espresse come prodotti delle derivate prime di f e dei pesi $1/\sigma_i^2$. Dalla (1.9) discende infatti la (10.94). Una volta trovato il minimo rispetto ai valori $\Delta\theta_k$, il fit si sposta nel nuovo punto θ_k. Nel punto di minimo, gli errori sui parametri vengono ottenuti attraverso la matrice degli errori (10.83), come nel caso lineare.

Quando la funzione da minimizzare dipende dalla funzione di verosimiglianza, si può scegliere di minimizzarla nella forma

$$-2\ln L(\boldsymbol{\theta}) + C \equiv -2\sum_i \ln f(x_i,\boldsymbol{\theta}) + C \ . \tag{1.11}$$

La ripetizione del calcolo appena fatto mostra che, linearizzando la funzione f, la matrice delle derivate seconde della (1.11) ha elementi che vengono approssimati da

$$\frac{\partial^2\chi^2}{\partial\theta_j\partial\theta_k} \simeq 2\sum_i \frac{1}{f_i^2}\frac{\partial f_i}{\partial\theta_j}\frac{\partial f_i}{\partial\theta_k} \ . \tag{1.12}$$

Si può mostrare che anche in questo caso l'approssimazione restituisce una matrice definita positiva [5, 7], che, lo ricordiamo, approssima la matrice di informazione di Fisher campionaria.

La discussione precedente ci ha mostrato che in un problema di minimizzazione del χ^2 un minimo dovrebbe sempre esistere, perché la matrice G della (1.3) è sempre definita positiva . Tuttavia, quando si deve risolvere un problema complicato con molti parametri, spesso i programmi segnalano la mancanza di questa condizione. In questo caso una soluzione viene comunque ottenuta forzando il parametro λ della (1.6) ad assumere valori grandi. Va sottolineato che in questo caso la soluzione *non è affidabile*, soprattutto per ciò che riguarda il calcolo degli errori. Se vi capitasse di imbattervi in queste difficoltà, dovete cercare di rimuovere in tutti i modi le cause che rendono la matrice di curvatura non definita positiva. Quelle più comuni sono:

a) *una parametrizzazione o una forma funzionale errata della funzione* $f(x, \boldsymbol{\theta})$ *che rappresenta il valore atteso*: occorre allora verificare che non vi siano parametri ridondanti, ovvero parametri le cui stime hanno coefficiente di correlazione vicino ad uno. Questo si può vedere dalla matrice degli errori (o delle correlazioni) che il programma fornisce in uscita. L'eliminazione dei parametri in eccesso in genere risolve il problema;
b) *un valore irrealistico dei parametri iniziali*: a volte il problema è ben posto, ma ai parametri viene assegnato un valore iniziale molto lontano dal minimo. Il programma può allora incontrare, nello spazio multidimensionale dei parametri, una regione molto irregolare, dove non vi sono minimi definiti. Se il programma trova da solo la regione giusta, smetterà di segnalare che la matrice G non è definita positiva e nell'intorno del minimo la soluzione sarà affidabile. Se questo non succede, bisogna ripartire con un valore diverso e più realistico dei parametri;
c) *errori numerici*: nel caso di molti parametri, gli errori numerici dovuti all'accuratezza del calcolatore usato possono diventare dominanti, e fare fallire l'algoritmo. È pertanto sempre consigliabile eseguire calcoli di questo tipo utilizzando almeno la doppia precisione.

Da tutta la discussione precedente risulta chiaro che gli algoritmi di minimizzazione trovano *il primo minimo che incontrano*, non quello assoluto. Ci sono dei metodi sicuri per sapere se quello trovato e un minimo relativo oppure assoluto? Purtroppo no. Questo problema è spesso il cruccio più grande di chi deve minimizzare funzioni complicate con molti parametri liberi. Occorre pertanto una estrema cura nella scelta del "punto di partenza", cioè nei valori da assegnare inizialmente ai parametri, che vanno scelti utilizzando al massimo la conoscenza che si ha a priori del problema. È spesso utile anche ripetere la minimizzazione più volte cambiando i valori iniziali e verificare che la soluzione trovata per il minimo sia stabile.

Per concludere, ricapitoliamo il ruolo della matrice delle derivate seconde prodotta dai codici di minimizzazione: la sua inversa, calcolata nel punto di minimo, contiene una stima delle varianze e covarianze delle componenti di $\hat{\boldsymbol{\theta}}$. Nel caso dei minimi quadrati non lineari e nel caso di stimatori di massima verosimiglianza, gli errori sono valutati solo in modo approssimato, sfruttando l'approssimazione asintotica gaussiana della distribuzione degli stimatori; per i minimi quadrati lineari l'espressione è invece esatta e restituisce proprio la (10.82).

Tutti questi errori consentono di costruire intervalli di confidenza simmetrici attorno alle componenti di $\hat{\boldsymbol{\theta}}$, il cui livello di confidenza risulta esatto nel caso lineare gaussiano, come ad esempio nelle (10.39, 10.40). Tuttavia, se il modello non è lineare, i livelli di confidenza corrispondenti agli errori ottenuti dalla matrice possono scostarsi in modo sensibile da quelli gaussiani, anche se essi sono quasi sempre riportati in uscita dai codici di minimizzazione. Per questo è spesso preferibile, anche per l'invarianza rispetto a riparametrizzazioni (par. 9.5), adottare un approccio come quello della (9.45) e della (10.95),

quando valgono le condizioni per le quali $\chi^2(\hat{\boldsymbol{\theta}})$ oppure $2[\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\hat{\boldsymbol{\theta}})]$ hanno distribuzione asintotica χ^2. Alcuni codici di minimizzazione forniscono, in genere su richiesta, le soluzioni della (9.45) e della (10.95).

Se le correlazioni tra i parametri sono importanti e il modello è fortemente non lineare, assumere che nell'intorno del punto di minimo la funzione da minimizzare sia parabolica può portare ad errori non trascurabili nel calcolo degli intervalli di confidenza. In questi casi è bene controllare la correttezza della procedura utilizzata con dati simulati, come nei problemi 10.9 e 10.10.

Bibliografia

1. A. Azzalini. *Inferenza Statistica*. Springer-Verlag Italia, Milano, 2001.
2. P. Baldi. *Calcolo delle Probabilità e Statistica*. McGraw Hill Italia, Milano, 1998.
3. P.R. Bevington and D.K. Robinson. *Data reduction and error analysis for the physical sciences*. McGraw-Hill, New York, 1992.
4. H. Cramer. *Mathematical Methods of Statistics*. Princeton University Press, Princeton, 1951.
5. W.T. Eadie, D. Drijard, F. James, M. Roos, and B. Sadoulet. *Statistical Methods in experimental Physics*. North Holland, Amsterdam, 1971.
6. P. Galeotti. *Probabilità e Statistica*. Levrotto e Bella, Torino, 1984.
7. F. James. Minuit reference manual. Technical Report CERN D506, CERN, Ginevra, 1992.
8. M. Lybanon. Comment on "least squares when both variables have uncertainties". *American Journal of Physics*, 52:276–278, 1984.
9. A.M. Mood, F.A. Graybill, and D.C. Boes. *Introduzione alla Statistica*. McGraw-Hill Italia, Milano, 1991.
10. J. Orear. Least squares when both variables have uncertainties. *American Journal of Physics*, 50:912–916, 1982.
11. L. Piccinato. *Metodi per le decisioni statistiche*. Springer-Verlag Italia, Milano, 1996.
12. W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Wetterling. *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, Cambridge, second edition, 1992.
13. G.A.F. Seber and C.J. Wild. *Nonlinear Regression*. Wiley Interscience, New York, 1989.