Materiale Integrativo per il corso di

Identificazione dei Modelli e Analisi dei Dati

1. Stima dei parametri

Nel Capitolo 2 si è visto che il problema della stima di una misura di probabilità in una famiglia di misure, sulla base di un vettore di osservazioni, può essere effettuata secondo il criterio della massima verosimiglianza. Dopo una rigorosa formulazione, il problema di stima viene ricondotto al problema della ricerca del massimo per il funzionale di verosimiglianza. Nel caso di una famiglia di misure parametrizzata con un vettore $\vartheta \in I\!R^m$, la stima di massima verosimiglianza della misura si ottiene calcolando quel valore di ϑ che rende massimo il funzionale di verosimiglianza. Al valore ottimale $\hat{\vartheta}$ così ottenuto viene dato il nome di stima di massima verosimiglianza del vettore di parametri ϑ.

Si consideri il caso di osservazione di una variabile aleatoria Y che assume valori in $I\!R^r$, definita su una terna statistica $(\Omega, \mathcal{F}, \mathcal{P}_\vartheta)$ in cui la famiglia di misure di probabilità è descrthitta dal vettore di parametri $\vartheta \in \Theta$, dove $\Theta \subseteq I\!R^m$ è un insieme di valori ammissibili per il parametro ϑ. La variabile Y induce su $(I\!R^r, \mathcal{B}(I\!R^r))$ una famiglia di misure di probabilità, che sarà indicata con $\mathcal{P}_{Y,\vartheta}$. Nel caso in cui tutte le misure della famiglia $\mathcal{P}_{Y,\vartheta}$ siano assolutamente continue rispetto alla misura di Lebesgue su $I\!R^r$, il funzionale di verosimiglianza $L(\vartheta)$ può essere definito come la derivata di Radon-Nicodym della generica misura di $\mathcal{P}_{Y,\vartheta}$ rispetto alla misura di Lebesgue della variabile aleatoria Y valutata nell'osservazione y. Si ricordi che tale derivata non è altro che la densità di probabilità, e pertanto

$$L(\vartheta) = \frac{d\mathcal{P}_{Y,\vartheta}}{d\lambda}(y) = p_Y(y; u, \vartheta). \tag{1.1}$$

La stima di massima verosimiglianza $\hat{\vartheta}$ del vettore dei parametri è quel valore di $\vartheta \in \Theta$ che rende massima la densità $p_Y(y; u, \vartheta)$, valutata nell'osservazione y

$$\hat{\vartheta} : L(\hat{\vartheta}) = \max_{\vartheta \in \Theta} L(\vartheta). \tag{1.2}$$

Da quanto esposto si nota che il problema della stima di Massima Verosimiglianza di parametri di distribuzioni si riduce al problema della ricerca di massimi di funzioni. Nella maggioranza delle applicazioni non è possibile la risoluzione algebrica in forma chiusa di un problema del tipo (1.2) e pertanto occorre ricorrere ai metodi della Teoria dell'Ottimizzazione, scienza che si occupa dello studio di algoritmi numerici per la soluzione dei problemi di ricerca del massimo (o del minimo) di funzioni.

In generale un metodo numerico per la soluzione di un problema di massimo è un algoritmo, che può essere implementato su di un calcolatore, che a partire da un valore iniziale ϑ^0 per il vettore di parametri, valore ammissibile ma non ottimo, produce una sequenza $\{\vartheta^k\}$ che al crescere di k tende al valore ottimale $\hat{\vartheta}$ (in un tempo infinito o, se si è fortunati, in un tempo finito). La Teoria dell'Ottimizzazione si occupa dello studio degli algoritmi iterativi di soluzione, caratterizzandoli in termini di velocità di convergenza e di complessità computazionale.

Una classe particolare di algoritmi è quella degli algoritmi che utilizzano la conoscenza del gradiente della funzione da massimizzare per il calcolo del $(k+1)$-esimo elemento della successione a partire dal k-esimo.

Per capire in che modo la conoscenza del gradiente possa aiutare a risolvere il problema numerico della ricerca dell'ottimo si consideri un algoritmo ideale *a tempo continuo*, ovvero un algoritmo che invece di produrre una sequenza discreta $\{\vartheta^k\}$ asintoticamente convergente a $\hat{\vartheta}$, produca una traiettoria $\{\vartheta(t)\}$ continua nel tempo. Per semplicità di trattazione si supporrà $\Theta = I\!R^m$.

Ipotesi: $L(\vartheta)$ sia una funzione $C^1(I\!R^m, I\!R)$, ed esista unico $\hat{\vartheta}$ tale che:

$$L(\hat{\vartheta}) > L(\vartheta), \quad \forall \vartheta \neq \hat{\vartheta}, \tag{1.3}$$

$$\left.\frac{dL(\vartheta)}{d\vartheta}\right|_{\hat{\vartheta}} = 0 \qquad \text{e} \qquad \frac{dL(\vartheta)}{d\vartheta} \neq 0, \quad \forall \vartheta \neq \hat{\vartheta}. \tag{1.4}$$

Tesi: l'equazione differenziale

$$\dot{\vartheta}(t) = \alpha \left(\frac{dL}{d\vartheta}(\vartheta(t))\right)^{\mathrm{T}}, \quad \text{con } \alpha > 0, \tag{1.5}$$

per qualunque valore iniziale $\vartheta(0) = \vartheta^0$ produce una traiettoria tale che

$$\lim_{t\to\infty} \vartheta(t) = \hat{\vartheta}. \tag{1.6}$$

La convergenza dell'algoritmo *ideale* sopra riportato è facilmente dimostrabile considerando l'equazione differenziale (1.5) come un sistema dinamico e mostrando che il punto $\hat{\vartheta}$ è per esso un punto di equilibrio globalmente asintoticamente stabile. A tale scopo è possibile impiegare il Teorema di Lyapunov, utilizzando la seguente funzione definita positiva

$$V(\vartheta) = L(\hat{\vartheta}) - L(\vartheta). \tag{1.7}$$

Dalla (1.3) è facile verificare che $V(\vartheta)$ è definita positiva e che si annulla solo in $\hat{\vartheta}$. Il calcolo della derivata di $V(\vartheta)$ sulle traiettorie del sistema (1.5) porta a

$$\dot{V}(\vartheta(t)) = \frac{dV}{d\vartheta}\dot{\vartheta}(t) = -\alpha \left\| \frac{dL}{d\vartheta}(\vartheta(t)) \right\|^2. \tag{1.8}$$

Dalla condizione (1.4) è facile verificare che $\dot{V}(\vartheta)$ è definita negativa e si annulla solo in $\hat{\vartheta}$. Resta pertanto dimostrata la stabilità asintotica globale del punto di equilibrio $\hat{\vartheta}$.

Da quanto detto segue che qualunque sia $\vartheta(0)$, si ha la convergenza asintotica di $\vartheta(t)$ a $\hat{\vartheta}$.

È già stato sottolineato che l'algoritmo (1.5) è un algoritmo *ideale*, perché a tempo continuo. Se però si riuscisse a costruire una buona approssimazione a tempo discreto dell'equazione differenziale (1.5) si potrebbe comunque ottenere una sequenza $\{\vartheta^k\}$ tale da approssimare sempre meglio la stima ottima $\hat{\vartheta}$ al crescere di k.

L'*Algoritmo del Gradiente*, uno tra i più utilizzati per la risoluzione dei problemi di ottimizzazione, può essere considerato un'approssimazione a passo variabile dell'equazione differenziale (1.5).

Algoritmo del Gradiente

$$\vartheta^{k+1} = \vartheta^k + \alpha_k \left(\frac{dL}{d\vartheta}(\vartheta^k) \right)^{\mathrm{T}}. \tag{1.9}$$

Il coefficiente α_k può essere scelto volta per volta al passo k secondo diversi criteri, e proprio il criterio di scelta di questo coefficiente gioca un ruolo fondamentale per la convergenza dell'algoritmo.

Se oltre alle condizioni (1.3) (sul funzionale L) e (1.4) (sul gradiente di L) anche la seguente condizione sulle derivate seconde di L è soddisfatta

$$\frac{d^2 L(\vartheta)}{d\vartheta^2} < 0, \quad \forall \vartheta \in I\!R^m \tag{1.10}$$

(si noti che la (1.10) è una condizione globale di concavità rivolta verso il basso), allora il seguente algoritmo ideale

$$\dot{\vartheta}(t) = - \left(\frac{d^2 L}{d\vartheta^2} \right)^{-1} \left(\frac{dL}{d\vartheta} \right)^{\mathrm{T}} \Bigg|_{\vartheta(t)}, \tag{1.11}$$

fornisce, per qualunque valore iniziale, una traiettoria asintoticamente tendente al valore ottimo $\hat{\vartheta}$. La dimostrazione viene fatta prendendo la stessa funzione di Lyapunov scelta per dimostrare la convergenza dell'algoritmo (1.5). La derivata della funzione di Lyapunov lungo le traiettorie del sistema (1.11) è pari a

$$\dot{V}(\vartheta) = \frac{dL}{d\vartheta} \left(\frac{d^2 L}{d\vartheta^2} \right)^{-1} \frac{dL}{d\vartheta}^{\mathrm{T}}, \tag{1.12}$$

ed in base all'ipotesi (1.10) è definita negativa, e si annulla solamente in $\hat{\vartheta}$ (si ricordi che se una matrice è simmetrica e definita negativa, la sua inversa esiste ed è simmetrica e definita negativa).

Il ben noto *Metodo di Newton* può essere interpretato come una discretizzazione dell'algorimo (1.11):

Metodo di Newton

$$\vartheta^{k+1} = \vartheta^k - \left(\frac{d^2 L(\vartheta)}{d\vartheta^2} \right)^{-1}_{\vartheta^k} \left(\frac{dL}{d\vartheta}(\vartheta^k) \right)^{\mathrm{T}}. \tag{1.13}$$

Questo algoritmo, quando applicabile, assicura una velocità di convergenza superiore a quella fornita dall'algoritmo del gradiente. D'altro canto può essere computazionalmente oneroso il calcolo della matrice delle derivate seconde. È facile verificare (e il lettore lo verifichi!) che se il funzionale è una funzione quadratica del vettore di parametri, del tipo cioè $L(\vartheta) = \vartheta^{\mathrm{T}} A \vartheta + b^{\mathrm{T}} \vartheta + c$, con $A < 0$, allora l'algoritmo del gradiente fornisce la soluzione ottima in un solo passo. È per quasto motivo che se ϑ^k si trova in un intorno di $\hat{\vartheta}$ in cui il funzionale può essere bene approssimato da una funzione quadratica la velocità di convergenza dell'algoritmo da quel punto in poi risulta molto elevata.

Evidentemente se non sono soddisfatte le condizioni (1.3) e (1.4) insorgono difficoltà (sia teoriche che numeriche o algoritmiche) nel calcolo della stima ottima $\hat{\vartheta}$.

Un problema di tipo algoritmico si ha quando non è soddisfatta la seconda parte della condizione (1.4), ovvero nel caso in cui il gradiente si annulli anche in punti diversi dal massimo per il funzionale $L(\vartheta)$. In questo caso l'algoritmo ideale (1.5) può convergere a punti di stazionarietà diversi dall'ottimo (la derivata della funzione di Lyapunov (1.7) risulta semidefinita negativa, e pertanto il punto $\hat{\vartheta}$ risulta semplicemente stabile).

Nel caso in cui il massimo del funzionale $L(\vartheta)$ esiste ma non è unico viene meno la condizione (1.3). Esiste allora un insieme di valori di ϑ che rende massimo $L(\vartheta)$, ed i parametri in questo insieme vengono detti indistinguibili. In questo caso è possibile che i parametri del problema non siano stati scelti in modo oculato: una diversa scelta per il vettore dei parametri potrebbe portare a verificare le condizioni (1.3) e (1.4).

È bene sottolineare che in alcune situazioni anziché massimizzare il funzionale di verosimiglianza $L(\vartheta)$ si preferisce minimizzarne il logaritmo cambiato di segno. Si ha pertanto

$$\hat{\vartheta} : J(\hat{\vartheta}) = \min_{\vartheta \in \Theta} J(\vartheta). \tag{1.14}$$

in cui

$$J(\vartheta) = -\ln L(\vartheta). \tag{1.15}$$

In questo caso l'algoritmo del gradiente si scrive

$$\vartheta^{k+1} = \vartheta^k - \alpha_k \left(\frac{dJ}{d\vartheta}(\vartheta^k) \right)^{\mathrm{T}}. \tag{1.16}$$

Di particolare interesse è il problema di stimare ϑ nel caso in cui se ne osservi una funzione non lineare $f(\vartheta)$ con rumore di misura gaussiano. Il problema viene

formalizzato nel modo seguente

$$Y(\omega) = f(\vartheta) + N(\omega), \qquad Y(\omega), N(\omega) \in I\!R^r,\ \vartheta \in I\!R^m, \tag{1.17}$$

in cui N è un vettore aleatorio gaussiano a media nulla e covarianza Ψ_N definita positiva. Il problema di stimare ϑ secondo il criterio della massima verosimiglianza consiste nel massimizzare la densità (Gaussiana) del vettore aleatorio Y valutata nell'osservazione y. Si ha cioè

$$L(\vartheta) = p_Y(y; u, \vartheta) = p_N(y - f(\vartheta)), \tag{1.18}$$

in cui $p_N(\cdot)$ è la densità (Gaussiana) del rumore di misura N. Si ha quindi

$$L(\vartheta) = \frac{1}{(2\pi)^{r/2}|\Psi_N|^{1/2}} \exp\Big(-\frac{1}{2}(y - f(\vartheta))^{\mathrm{T}} \Psi_N^{-1} (y - f(\vartheta))\Big). \tag{1.19}$$

Per poter sperare di avere un valore unico per il massimo del funzionale di verosimiglianza occorre ipotizzare che il numero di misure sia superiore al numero dei parametri, e cioè che $r > m$.

Nel caso qui considerato la matrice di covarianza Ψ_N è supposta nota, e pertanto risulta conveniente ricercare il minimo del logaritmo del funzionale $L(\vartheta)$ cambiato di segno, eliminando preliminarmente gli addendi costanti $\ln(2\pi)^{r/2}$ e $\ln|\Psi_N|^{1/2}$. La stima di massima verosimiglianza di ϑ è pertanto data dalla soluzione del problema (1.14), in cui

$$J(\vartheta) = \frac{1}{2}(y - f(\vartheta))^{\mathrm{T}} \Psi_N^{-1} (y - f(\vartheta)). \tag{1.20}$$

Il gradiente del funzionale $J(\vartheta)$ è pari a

$$\frac{dJ(\vartheta)}{d\vartheta} = -(y - f(\vartheta))^{\mathrm{T}} \Psi_N^{-1} \frac{df(\vartheta)}{d\vartheta}. \tag{1.21}$$

La condizione di annullamento del gradiente

$$-(y - f(\hat{\vartheta}))^{\mathrm{T}} \Psi_N^{-1} \frac{df(\vartheta)}{d\vartheta}\Big|_{\hat{\vartheta}} = 0, \tag{1.22}$$

che è una condizione necessaria di ottimo, può essere interpretata come una condizione di ortogonalità, nel prodotto scalare $\langle x, y\rangle = x^{\mathrm{T}} \Psi_N^{-1} y$, tra l'errore residuo $(y - f(\hat{\vartheta}))$ e le colonne della matrice $df(\vartheta)/d\vartheta$, che sono le derivate di $f(\vartheta)$ rispetto alle componenti del vettore di parametri

$$\frac{dJ(\vartheta)}{d\vartheta}\Big|_{\hat{\vartheta}} = 0, \quad \Leftrightarrow \quad -(y - f(\hat{\vartheta}))^{\mathrm{T}} \Psi_N^{-1} \frac{df(\vartheta)}{d\vartheta_j}\Big|_{\hat{\vartheta}} = 0, \qquad j = 1, \ldots, m. \tag{1.23}$$

In questa applicazione è possibile dare una nozione di indistinguibilità dei parametri indipendente dall'osservazione y.

Definizione 1.1. *Due parametri ϑ_a e ϑ_b sono detti* **indistinguibili** *se $f(\vartheta_a) = f(\vartheta_b)$.*

È evidente che due parametri indistinguibili dànno lo stesso valore per il funzionale $J(\vartheta)$. Per una discussione locale dell'indistinguibilità è importante il seguente teorema:

Teorema 1.2. *Sia Θ un sottoinsieme aperto di $\mathbb{R}^m$ e sia $f(\vartheta) \in C^2(\Theta, \mathbb{R}^r)$, con $r > m$. Sia $S(\bar{\vartheta}, \epsilon) \subset \Theta$ un intorno sferico di raggio $\epsilon > 0$ di un punto $\bar{\vartheta} \in \Theta$, tale che*

$$\text{rango}\left(\frac{df(\vartheta)}{d\vartheta}\right) < m, \quad \forall \vartheta \in S(\bar{\vartheta}, \epsilon). \tag{1.24}$$

Allora esiste una curva $\theta(\rho) \in C^1((-1,1), \mathbb{R}^m)$ tale che

$$\theta(0) = \bar{\vartheta}, \quad \text{e} \quad \begin{cases} \theta(\rho) \in S(\bar{\vartheta}, \epsilon), \\ f(\theta(\rho)) = f(\bar{\vartheta}), \\ \dfrac{d\theta(\rho)}{d\rho} \neq 0, \end{cases} \quad \forall \rho \in (-1,1). \tag{1.25}$$

Definizione 1.3. *Sia Θ un sottoinsieme aperto di $\mathbb{R}^m$. Un valore $\bar{\vartheta} \in \Theta$ per il vettore di parametri ϑ è detto* **localmente identificabile** *se esiste un intorno sferico di raggio ϵ di $\bar{\vartheta}$ in cui non vi sono parametri indistinguibili da $\bar{\vartheta}$, ovvero*

$$\exists \epsilon > 0 : \qquad f(\vartheta) \neq f(\bar{\vartheta}) \qquad \forall \vartheta \in S(\bar{\vartheta}, \epsilon) \setminus \{\bar{\vartheta}\} \tag{1.26}$$

Dalla definzione data si può affermare che un valore $\bar{\vartheta} \in \mathbb{R}^m$ non è localmente identificabile se

$$\forall \epsilon > 0, \ \exists \vartheta^\epsilon \in S(\bar{\vartheta}, \epsilon) : \quad f(\vartheta^\epsilon) = f(\bar{\vartheta}). \tag{1.27}$$

In questo caso è possibile mostrare che se $f(\vartheta) \in C^2(\Theta, \mathbb{R}^r)$ allora per ogni $\epsilon > 0$ esiste una curva $\theta(\rho) \in C^1((-1,1), \mathbb{R}^m)$ che gode delle proprieta (1.25). In altre parole, se un parametro $\bar{\vartheta} \in \Theta$ non è localmente identificabile esiste almeno una curva continua di parametri indistinguibili passante per $\bar{\vartheta}$, interamente contenuta in $S(\bar{\vartheta}, \epsilon)$.

È banale verificare la proprietà inversa, e cioè che se esiste una curva continua di parametri indistinguibili passanti per $\bar{\vartheta}$, allora il parametro $\bar{\vartheta}$ non è localmente identificabile. Si può pertanto affermare che:

la non identificabilità locale di un parametro $\bar{\vartheta}$ implica ed è implicata dall'esistenza di (almeno) una curva $\theta(\rho)$ di parametri indistinguibili passante per $\bar{\vartheta}$.

Derivando la funzione $f(\theta(\rho))$ rispetto a ρ, e tenendo conto del fatto che essa è costante e pari a $f(\hat{\vartheta})$, si ha

$$\frac{df(\vartheta)}{d\vartheta}\Big|_{\theta(\rho)} \frac{d\theta(\rho)}{d\rho} = 0, \qquad \forall \rho \in (-1,1). \tag{1.28}$$

Poichè $d\theta(\rho)/d\rho \neq 0$, segue che deve necessariamente essere

$$\frac{d\theta(\rho)}{d\rho} \in \mathcal{N}\left(\frac{df(\vartheta)}{d\vartheta}\Big|_{\theta(\rho)}\right). \tag{1.29}$$

Segue quindi che se il parametro $\bar{\vartheta}$ non è localmente identificabile, allora

$$\text{rango}\Big(\frac{df}{d\vartheta}\Big)_{\theta(\rho)} < m, \quad \forall \rho \in (-1,1). \tag{1.30}$$

Tenendo conto di quanto discusso finora, possiamo dare il seguente risultato:

Teorema 1.4. *Sia Θ un sottoinsieme aperto di $\mathbb{R}^m$, e sia $f(\vartheta) \in C^2(\Theta, \mathbb{R}^r)$, con $r > m$. Condizione necessaria e sufficiente affinché tutti i valori dei parametri in $\bar{\Theta}$, sottoinsieme aperto di Θ, siano localmente identificabili è che la matrice delle derivate $df/d\vartheta$ abbia le colonne indipendenti $\forall \vartheta \in \bar{\Theta}$, a meno di un insieme Q di punti isolati di $\bar{\Theta}$.*

In formule la condizione del teorema si scrive

$$\text{rango}\Big(\frac{df}{d\vartheta}\Big) = m \quad \forall \vartheta \in \bar{\Theta} \setminus Q, \tag{1.31}$$

1.1: Stima dei parametri nei sistemi lineari

Nel capitolo 3 del libro viene trattato il problema della stima dei parametri di un sistema lineare e stazionario a tempo continuo, senza rumore di stato, con osservazioni rumorose a tempo discreto. Nel caso in cui lo stato iniziale $x(0)$ del sistema sia noto, questo problema è sostanzialmente un problema del tipo (1.17), come si può vedere confrontando l'equazione di misura (3.1.17) riportata sul libro con l'equazione (1.17) di questi appunti. Qui tratteremo più in dettaglio il caso di stato iniziale $x(0) = 0$. In questo caso il termine $H_N(x(0), u, \vartheta)$ coincide con il termine $F_N(u, \vartheta)$ definito nella formula (3.1.26) del libro, formula peraltro riportata in maniera errata nel testo. L'esatta espressione del vettore $F_N(u, \vartheta)$ è la seguente

$$F_N(u,\vartheta) = \begin{bmatrix} D(\vartheta)u(0) \\ \vdots \\ \int_0^{N\Delta} C(\vartheta)e^{A(\vartheta)(N\Delta-\tau)}B(\vartheta)u(\tau)\,d\tau + D(\vartheta)u(N\Delta) \end{bmatrix}, \tag{**3.1.26**}$$

e consiste nel vettore dei valori della risposta forzata dell'uscita negli istanti di tempo $k\Delta$, per $k = 0, 1, \ldots, N$ (il lettore che ha studiato la Teoria dei Sistemi è senz'altro in grado di ricavare tale vettore anche nel caso di sistemi a tempo discreto, e lo faccia!). In altre parole $F_N(u, \vartheta)$ può essere scritto come

$$F_N(u,\vartheta) = \begin{bmatrix} y(0;u,\vartheta) \\ y(\Delta;u,\vartheta) \\ \vdots \\ y(N\Delta;u,\vartheta) \end{bmatrix}, \tag{1.32}$$

in cui

$$y(t;u,\vartheta) = \int_0^t C(\vartheta)e^{A(\vartheta)(t-\tau)}B(\vartheta)u(\tau)\ d\tau + D(\vartheta)u(t). \tag{1.33}$$

Indicando con $J_N(\vartheta)$ il funzionale nell'applicazione in esame (coerentemente con le notazioni del libro) si ha

$$J_N(\vartheta) = \frac{1}{2}(z_N - F_N(u,\vartheta))^{\mathrm{T}}\Psi_N^{-1}(z_N - F_N(u,\vartheta)). \tag{1.34}$$

Considerando anche il fatto che la matrice di covarianza Ψ_N è diagonale a blocchi (vedi l'espressione (3.1.20) del libro) per il funzionale vale l'espressione (3.1.23) del libro in cui però occorre porre $x(0) = 0$. Si ottiene quindi

$$J_N(\vartheta) = \frac{1}{2}\sum_{k=0}^{N}(z(k\Delta) - y(k\Delta;u,\vartheta))^{\mathrm{T}}(GG^{\mathrm{T}})^{-1}(z(k\Delta) - y(k\Delta;u,\vartheta)) \tag{1.35}$$

Da un punto di vista algoritmico il valore numerico della $J_N(\vartheta)$ si ottiene svolgendo una simulazione del sistema nell'intervallo $[0, N\Delta]$, avendo fissato il valore desiderato ϑ del parametro, e utilizzando i valori numerici dell'uscita calcolata negli istanti $k\Delta$ nell'espressione (1.35).

Calcolando il gradiente di $J_N(\vartheta)$ si ottiene

$$\frac{dJ_N(\vartheta)}{d\vartheta} = -(z_N - F_N(u,\vartheta))^{\mathrm{T}}\Psi_N^{-1}\frac{dF_N(u,\vartheta)}{d\vartheta}. \tag{1.36}$$

Si rende quindi necessario il calcolo della derivata di $F_N(u,\vartheta)$ rispetto al vettore di parametri ϑ. Occorre pertanto calcolare le matrici

$$\frac{dy(k\Delta;u,\vartheta)}{d\vartheta} = \left[\begin{array}{ccc} \dfrac{dy(k\Delta;u,\vartheta)}{d\vartheta_1} & \cdots & \dfrac{dy(k\Delta;u,\vartheta)}{d\vartheta_m} \end{array}\right], \quad k = 0,1,\ldots,N. \tag{1.37}$$

A questo punto si può osservare che la componente j-esima del vettore delle derivate di $y(k\Delta;u,\vartheta)$ coincide con l'uscita del sistema di sensibilità rispetto al parametro ϑ_j, così come è definito nel paragrafo 3.3 del libro (formule (3.3.3.) e (3.3.4) del libro), nell'istante $k\Delta$. Si ha cioè

$$\frac{dy(k\Delta;u,\vartheta)}{d\vartheta} = [\, y_{\vartheta_1}(k\Delta) \quad \cdots \quad y_{\vartheta_m}(k\Delta)\,], \tag{1.38}$$

(per semplicità nei vettori y_{ϑ_j} si è omessa la dipendenza da u e da ϑ) e quindi

$$\frac{dF_N(u,\vartheta)}{d\vartheta} = \left[\begin{array}{ccc} y_{\vartheta_1}(0) & \cdots & y_{\vartheta_m}(0) \\ y_{\vartheta_1}(\Delta) & \cdots & y_{\vartheta_m}(\Delta) \\ \vdots & \cdots & \vdots \\ y_{\vartheta_1}(N\Delta) & \cdots & y_{\vartheta_m}(N\Delta) \end{array}\right], \tag{1.39}$$

Si capisce pertanto la necessità di implementare il sistema di sensibilità qualora si vogliano utilizzare algoritmi numerici che impieghino il gradiente del funzionale $J(\vartheta)$ da minimizzare. Nel libro è riportato il solo sistema di sensibilità dell'uscita (forzata) rispetto ad un parametro scalare. Nel caso di un parametro vettoriale $\vartheta \in I\!R^m$ il sistema di sensibilità è dato da

$$\begin{aligned}
\begin{bmatrix} \dot{x} \\ \dot{x}_{\vartheta_1} \\ \dot{x}_{\vartheta_2} \\ \vdots \\ \dot{x}_{\vartheta_m} \end{bmatrix} &= \begin{bmatrix} A & 0 & 0 & \cdots & 0 \\ A_{\vartheta_1} & A & 0 & \cdots & 0 \\ A_{\vartheta_2} & 0 & A & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ A_{\vartheta_m} & 0 & 0 & \cdots & A \end{bmatrix} \begin{bmatrix} x \\ x_{\vartheta_1} \\ x_{\vartheta_2} \\ \vdots \\ x_{\vartheta_m} \end{bmatrix} + \begin{bmatrix} B \\ B_{\vartheta_1} \\ B_{\vartheta_2} \\ \vdots \\ B_{\vartheta_m} \end{bmatrix} u \\
y_\vartheta = \begin{bmatrix} y_{\vartheta_1} \\ y_{\vartheta_2} \\ \vdots \\ y_{\vartheta_m} \end{bmatrix} &= \begin{bmatrix} C_{\vartheta_1} & C & 0 & \cdots & 0 \\ C_{\vartheta_2} & 0 & C & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ C_{\vartheta_m} & 0 & 0 & \cdots & C \end{bmatrix} \begin{bmatrix} x \\ x_{\vartheta_1} \\ x_{\vartheta_2} \\ \vdots \\ x_{\vartheta_m} \end{bmatrix} + \begin{bmatrix} D_{\vartheta_1} \\ D_{\vartheta_2} \\ \vdots \\ D_{\vartheta_m} \end{bmatrix} u.
\end{aligned} \tag{1.40}$$

Dai valori y_{ϑ_j} calcolati negli istanti $k\Delta$ è quindi possibile calcolare il gradiente di $F_N(u, \vartheta)$ (formula (1.39)), e successivamente il gradiente di $J_N(\vartheta)$.

Un'espressione più maneggevole per il gradiente di $J_N(\vartheta)$ è la seguente

$$\frac{dJ_N(\vartheta)}{d\vartheta} = -\sum_{k=0}^{N} (z(k\Delta) - y(k\Delta; u, \vartheta))^{\mathrm{T}} (GG^{\mathrm{T}})^{-1} \frac{dy(k\Delta; u, \vartheta)}{d\vartheta}. \tag{1.41}$$

Per riassumere quanto finora esposto vediamo come sia possibile a questo punto utilizzare un algoritmo di ottimizzazione che richieda il calcolo del gradiente del funzionale da minimizzare. Assumiamo di disporre di un algoritmo di ottimizzazione che richieda il calcolo, per ogni valore fissato del parametro ϑ, sia del valore del funzionale $J_N(\vartheta)$ che del gradiente $dJ_N(\vartheta)/d\vartheta$.

I valori del funzionale $J_N(\vartheta)$ e del gradiente $dJ_N(\vartheta)/d\vartheta$ vengono calcolati utilizzando le espressioni (1.35) e (1.41). Si è già detto che per il calcolo di J_N in corrispondenza di un particolare valore di ϑ è sufficiente simulare il sistema nell'intervallo $[0, N\Delta]$.

Per calcolare il gradiente di J_N occorre calcolare le derivate derivate $dy(k\Delta; u, \vartheta)/d\vartheta$. Queste possono essere ottenute simulando il sistema di sensibilità nell'intervallo $[0, N\Delta]$, in corrispondenza dello stesso valore del parametro ϑ, ed utilizzarne le uscite negli istanti $k\Delta$ nell'espressione (1.41).

Nel paragrafo precedente è stata data una definizione di indistinguibilità indipendente dall'osservazione effettuata. Si ricordi che nel contesto della stima dei parametri di un sistema lineare l'osservazione è stata indicata con z_N. Secondo la definizione data, due parametri ϑ_a e ϑ_b sono indistinguibili se

$$F_N(u, \vartheta_a) = F_N(u, \vartheta_b). \tag{1.42}$$

L'assenza di curve in $\mathbb{R}^m$ costituite da parametri indistinguibili, e quindi l'identificabilità locale dei parametri in un insieme aperto $\Theta \subseteq \mathbb{R}^m$ è garantita se e solo se le colonne della matrice delle derivate $dF_N/d\vartheta$ sono indipendenti in tutto Θ, eventualmente a meno di un insieme Q di punti isolati.

Si ha quindi

$$\text{rango}\left(\frac{dF_N(u,\vartheta)}{d\vartheta}\right) = m \quad \forall \vartheta \in \Theta \setminus Q \quad \Leftrightarrow \quad \begin{array}{c}\text{Tutti i parametri in } \Theta \text{ sono} \\ \text{localmente identificabili}\end{array} \tag{1.43}$$

Si osservi che la nozione di indistinguibilità finora riportata non dipende dall'osservazione z_N, ma può dipendere dall'ingresso u applicato, in quanto due parametri indistinguibili per un dato ingresso potrebbero risultare distinguibili qualora venisse applicato un ingresso diverso. Addirittura potrebbe accadere che due parametri possono risultare indistinguibili se l'uscita viene misurata con un certo intervallo di campionamento Δ, mentre potrebbero risultare distinguibili utilizzando un diverso passo di campionamento.

È per questo motivo che nel libro di testo il problema dell'identificabilità viene impostato a prescindere dal passo di campionamento delle misure. In altre parole, le definizioni date, e quindi i teoremi dimostrati, hanno valore nel caso di osservazioni a tempo continuo, e cioè supponendo di conoscere le misure dell'uscita in tutti gli istanti di tempo in sottointervallo di $[0, \infty)$.

Ad esempio, la definizione 3.2.1 data nel libro è una definizione di indistinguibilità rispetto ad un insieme di esperimenti e prescinde dall'intervallo di tempo con cui vengono campionate le uscite.

Tale definizione, particolarizzata al caso in esame in queste note ($x(0) = 0$) e con le notazioni qui introdotte, diventa:

Definizione 3.2.1.b. *La coppia di valori $\vartheta, \alpha \in \Theta$ del vettore dei parametri è detta* **indistinguibile** *rispetto ad un insieme di ingressi U_T se:*

$$y(t; u, \vartheta) = y(t; u, \alpha), \quad \forall t \in [0, T], \tag{1.44}$$

per ogni $u \in U_T$. In caso contrario la coppia è detta **distinguibile**.

Chiariamo il significato della proposizione "in caso contrario", dando la definizione di distinguibilità:

Definizione 1.5. *La coppia di valori $\vartheta, \alpha \in \Theta$ del vettore dei parametri è detta* **distinguibile** *rispetto ad un insieme di ingressi U_T se $\exists u \in U_T$ e se esiste un sottointervallo I di $[0, T]$ tale che*

$$y(t; u, \vartheta) \neq y(t; u, \alpha), \quad \forall t \in I. \tag{1.45}$$

D'altronde, se due parametri non risultano distinguibili a partire dalla conoscenza delle misure in tutto l'intervallo di tempo $[0, T)$, non è sperabile che risultino distinguibili a partire dalla conoscenza delle misure nei soli istanti $k\Delta$, per $k = 0, 1, \ldots, N$.

Da quanto discusso si deduce che le condizioni di identificabilità nel caso di osservazioni a tempo continuo sono condizioni solo *necessarie* per la *identificabilità* dei parametri nel caso di osservazioni a tempo discreto. Viceversa le condizioni di *indistinguibilità* nel caso di osservazioni a tempo continuo sono sufficienti per l'*indistinguibilità* con osservazioni a tempo discreto.

L'analogo del teorema (1.4) relativo al caso di osservazioni a tempo discreto, è il seguente:

Teorema 1.6. *Sia Θ un sottoinsieme aperto di $\mathbb{R}^m$, e sia $y(\vartheta; u, \vartheta)$, data dalla* (1.33) *per un ingresso fissato, una funzione di classe C^2 nella variabile ϑ per $t \in [0, T]$. Allora il vettore di parametri ϑ è localmente identificabile in un sottoinsieme aperto $\bar{\Theta}$ di Θ se e solo se non esiste una funzione $k \in C^1(\bar{\Theta}, \mathbb{R}^m)$ diversa da zero in $\bar{\Theta} \setminus Q$, con Q insieme di punti isolati, tale che*

$$\frac{dy(t; u, \vartheta)}{d\vartheta} k(\vartheta) = 0 \quad \forall \vartheta \in \bar{\Theta} \setminus Q, \ \forall t \in [0, T]. \tag{1.46}$$

A differenza dalla condizione data dal teorema (1.4), che è facilmente verificabile con una simulazione, la condizione di questo teorema non è facilmente verificabile. È comunque evidente che se esiste $k(\vartheta)$ tale da annullare il gradiente di $y(t; u, \vartheta)$ per ogni $t \in [0, T]$ allora la stessa $k(\vartheta)$ annulla il vettore $dF_N/d\vartheta$. Questo è un riscontro del fatto che l'indistinguibilità dei parametri nel caso di osservazioni a tempo continuo implica l'indistinguibilità nel caso di osservazioni a tempo discreto.

Una condizione necessaria per la non esistenza di una funzione vettoriale $k(\vartheta)$ che soddisfi la (1.46) può essere ottenuta considerando le proprietà della matrice dei coefficienti di Markov

$$R(\vartheta) = \begin{bmatrix} D(\vartheta) \\ C(\vartheta)B(\vartheta) \\ \vdots \\ C(\vartheta)A^{2n-1}(\vartheta)B(\vartheta) \end{bmatrix}. \tag{1.47}$$

definita nel caso di sistemi ad un ingresso ed una uscita nell formula (3.2.22) del libro. Si ricordi che l'eguaglianza delle matrici dei coefficienti di Markov in corrispondenza a due valori per il parametro implica ed è implicata dall'uguaglianza delle risposte impulsive del sistema (Teoremi 3.2.3-3.2.6)

$$R(\vartheta_a) = R(\vartheta_b) \quad \Leftrightarrow \quad w(t, \vartheta_a) = w(t, \vartheta_b), \ \forall t \in [0, \infty) \tag{1.48}$$

Si consideri ora il vettore dei coefficienti di Markov $\widetilde{R}(\vartheta)$, definito nel libro di testo con la formula (3.2.24), che può essere pensato costruito sovrapponendo in un unica colonna le colonne della matrice di dei coefficienti di Markov $R(\vartheta)$.

Si ha il seguente teorema:

Teorema 1.7. *Sia Θ un sottoinsieme aperto di $\mathbb{R}^m$. Il vettore di parametri ϑ è localmente identificabile in un sottoinsieme aperto $\bar{\Theta}$ di Θ se e solo se lo Jacobiano del vettore dei coefficienti di Markov $\widetilde{R}(\vartheta)$ ha rango m in $\bar{\Theta} \setminus Q$, con Q insieme di punti isolati.*

È possibile mostrare che la condizione data da questo teorema è una condizione necessaria per la non esistenza di una funzione vettoriale $k(\vartheta)$ che soddisfi la (1.46), e quindi è una condizione necessaria per l'identificabilità rispetto a un dato ingresso.

Infatti, se il vettore dei coefficienti di Markov ha rango inferiore a m in un intorno $S(\bar{\vartheta}, \epsilon)$ di un punto $\bar{\vartheta}$, allora esiste una curva $\theta(\rho)$ che soddisfa le condizioni (1.25), con $f(\vartheta) = R(\vartheta)$, e cioè tale che

$$R(\theta(\rho)) = R(\bar{\vartheta}), \quad \forall \rho \in (-1, 1). \tag{1.49}$$

Valgono allora le seguenti implicazioni

$$\begin{aligned}
&\text{rango}\left(\frac{d\widetilde{R}(\vartheta)}{d\vartheta}\right) < m, \quad \forall \vartheta \in S(\bar{\vartheta}, \epsilon) \\
&\quad \Leftrightarrow \quad \exists \theta(\rho) : R(\theta(\rho)) = R(\bar{\vartheta}), \qquad \forall \rho \in (-1, 1) \\
&\quad \Leftrightarrow \quad w(t, \theta(\rho)) = w(t, \bar{\vartheta}) \qquad \forall t \in [0, \infty) \quad \forall \rho \in (-1, 1) \\
&\quad \Rightarrow \quad y(t; u, \theta(\rho)) = y(t; u, \bar{\vartheta}) \qquad \forall t \in [0, \infty) \quad \forall \rho \in (-1, 1) \\
&\quad \Rightarrow \quad \frac{dy(t; u, \theta(\rho))}{d\rho} = 0 \qquad \forall t \in [0, \infty) \quad \forall \rho \in (-1, 1) \\
&\quad \Rightarrow \quad \left(\frac{dy(t; u, \vartheta)}{d\vartheta}\right)_{\theta(\rho)} \frac{d\theta}{d\rho} = 0 \qquad \forall t \in [0, \infty) \quad \forall \rho \in (-1, 1)
\end{aligned} \tag{1.50}$$

Da quanto discusso segue che se in tutto un aperto $\bar{\Theta}$ le colonne dello Jacobiano di $\widetilde{R}(\vartheta)$ non sono indipendenti, allora esiste un vettore $k(\vartheta)$ che soddisfa alla (1.46) ($k(\vartheta)$ è sempre la derivata di una curva di parametri indistinguibili), e quindi in $\bar{\Theta}$ esistono parametri che non sono localmente indistinguibili.

Esempio.

Si consideri il seguente sistema

$$\begin{aligned}
&\dot{x}(t) = A(\vartheta)x(t) + B(\vartheta)u(t), \quad x(0) = 0, \\
&y(t) = C(\vartheta)x(t),
\end{aligned} \tag{1.51}$$

in cui

$$\begin{aligned}
&A = \begin{bmatrix} 0 & 1 \\ \vartheta_1 & \vartheta_2 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ \vartheta_3 \end{bmatrix}, \\
&C = \begin{bmatrix} \vartheta_4 & 0 \end{bmatrix}, \quad D = 0.
\end{aligned} \tag{1.52}$$

Le osservazioni sono a tempo discreto

$$z(k\Delta) = y(k\Delta) + GN_k, \quad k = 0, 1, \ldots, N. \tag{1.53}$$

con $\{N_k\}$ sequenza gaussiana bianca standard.

Si desidera stimare il vettore di parametri $\vartheta = [\vartheta_1 \ \vartheta_2 \ \vartheta_3 \ \vartheta_4]^{\mathrm{T}}$ secondo il criterio della masima verosimiglianza. È bene verificare preliminarmente l'identificabilità locale dei parametri. La matrice dei coefficienti di Markov (vedi la formula 3.2.22 del libro) è la seguente (in questo esempio $n = 2$, e quindi $2n - 1 = 3$)

$$R(\vartheta) = \begin{bmatrix} D \\ CB \\ CAB \\ CA^2B \\ CA^3B \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vartheta_3\vartheta_4 \\ \vartheta_2\vartheta_3\vartheta_4 \\ (\vartheta_1 + \vartheta_2^2)\vartheta_3\vartheta_4 \end{bmatrix}. \tag{1.54}$$

L'identificabilità locale dei parametri può essere verificata controllando che il rango della derivata della matrice dei coefficienti di Markov sia pari ad m, e cioè pari a 4. In questo esempio si ha

$$\frac{dR(\vartheta)}{d\vartheta} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \vartheta_4 & \vartheta_3 \\ 0 & \vartheta_3\vartheta_4 & \vartheta_2\vartheta_4 & \vartheta_2\vartheta_3 \\ \vartheta_3\vartheta_4 & 2\vartheta_2\vartheta_3\vartheta_4 & (\vartheta_1 + \vartheta_2^2)\vartheta_4 & (\vartheta_1 + \vartheta_2^2)\vartheta_3 \end{bmatrix}. \tag{1.55}$$

La matrice $dR/d\vartheta$ è singolare $\forall\vartheta \in \mathbb{R}^4$ (il vettore $k(\vartheta) = [\ 0 \ \ 0 \ \ \vartheta_3 \ \ -\vartheta_4]^{\mathrm{T}}$ è nel nucleo di $dR/d\vartheta$).

Questo vuol dire che non esistono parametri identificabili in $\mathbb{R}^4$. Per poter procedere all'identificazione del sistema con esperimenti, occorre riparametrizzare le incognite del sistema. Si osservi che il problema di singolarità è dovuto al fatto che le ultime due colonne della derivata della matrice di Markov sono singolari, e cioè che le sensibilità della risposta impulsiva rispetto al parametro ϑ_3 e al parametro ϑ_4 non sono linearmente indipendenti (i parametri ϑ_3 e ϑ_4 si moltiplicano sempre negli elementi della matrice di Markov). D'altronde la funzione di trasferimento del sistema (1.51) è pari a

$$W(s, \vartheta) = \frac{\vartheta_3\vartheta_4}{s^2 - \vartheta_2 s - \vartheta_1}, \tag{1.56}$$

e da questa si capisce che con esperimenti a partire da stato inizale nullo non è possibile stimare sia ϑ_3 che ϑ_4: ci si deve limitare a stimare solamente il prodotto $\vartheta_3\vartheta_4$. Conviene allora porre nel sistema (1.51) $\vartheta_4 = 1$, e procedere con la stima di ϑ_3. Si ha in questo caso

$$R(\vartheta) = \begin{bmatrix} 0 \\ 0 \\ \vartheta_3 \\ \vartheta_2\vartheta_3 \\ (\vartheta_1 + \vartheta_2^2)\vartheta_3 \end{bmatrix}, \qquad \vartheta = \begin{bmatrix} \vartheta_1 \\ \vartheta_2 \\ \vartheta_3 \end{bmatrix} \in \mathbb{R}^3 \tag{1.57}$$

e

$$\frac{dR(\vartheta)}{d\vartheta} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & \vartheta_3 & \vartheta_2 \\ \vartheta_3 & 2\vartheta_2\vartheta_3 & (\vartheta_1 + \vartheta_2^2) \end{bmatrix}. \tag{1.58}$$

Indicando con $I(\vartheta_3 \neq 0)$ l'insieme di punti di $\mathbb{R}^3$ tali che $\vartheta_3 = 0$, allora si vede facilmente che la matrice $dR/d\vartheta$ ha rango 3 in tutto $\mathbb{R}^3 \setminus I(\vartheta_3 \neq 0)$, che è un insieme aperto. Pertanto, tutti i parametri in questo insieme risultano essere localmente identificabili. A questo punto è possibile procedere alla stima mediante algoritmi numerici (del tipo gradiente, per esempio).