Chapter 1

Introduzione

The number of things you don't know is
one of the things you don't know

Peter Green

Quando si vuole identificare un sistema dinamico bisogna

- stabilire la struttura del legge dinamica
- stabilire il valore numerico delle costanti del modello

La prima richiesta consiste nel decidere il tipo e il numero di operazioni sui segnali di ingresso e di stato che sono in grado di riprodurre accettabilmente (e possibilmente con una legge compatta) la dinamica del sistema. La seconda richiesta consiste invece nel determinare il valore numerico dei parametri che intervengono.

I casi in cui è sufficiente una identificazione parametrica sono i casi in cui si ha già la struttura della legge. Per ottenere la legge (a meno del valore delle costanti) è necessario uno studio del sistema a partire dai principi primi che regolano tutti i fenomeni che sono coinvolti. In definitiva l'idengtificazione parametrica è sufficiente solo se sono verificate le seguenti condizioni

- è **possibile** ottenere una descrizione del sistema applicando i principi primi delle scienze
- è **vantaggioso** ottenere una descrizione del sistema applicando i principi primi delle scienze (tempo speso, rischio di errori etc)
- ci si accontenta della descrizione ottenuta applicando i principi primi delle scienze (rischio di tralasciare fenomeni importanti)
- la complessità delle leggi che si ottengono applicando i principi primi delle scienze non è tale da renderle inutilizzabili

Quando una di queste condizioni non è soddisfatta è possibile inserire nel processo di identificazione anche la ricerca della struttura del modello. A tale scopo

si ipotizza una classe (abbastanza vasta) di strutture di modello e si lascia al processo di identificazione l'onere di scegliere quella che maggiormente si adatta alle misure. I metodi Bayesiani in particolare sono appropriati per la selezione dei modelli perchè forniscono naturalmente (senza analisi aggiuntve) un contesto in cui si quantifica l'incertezza associata all'identificazione. L'approccio bayesiano infatti, non fornisce una singola descrizione del sistema ma fornisce una distribuzione di probabilità associata al set di modelli possibili. Dunque risolvere il problema di inferenza bayesiana conduce a scelte più ponderate rispetto alla semplice estrazione del modello che si adatta meglio ai dati reali. Analizzando la distribuzione prodotta si può per esempio individuare un modello poco meno accurato del migliore ma preferirlo perchè meno complesso .
Nella seguente trattazione la classe di modelli scelta è quella dei nonlineari, autoregressivi, a media mobile con ingressi esogeni (NARMAX); essa è una popolare classe modelli ingresso-uscita spesso usata nell'identificazione di sistemi non lineari in vari ambiti dell'ingegneria (e non) poichè ha il pregio di produrre leggi compatte ma accurate. Come si vedrà nei successivi capitoli, ciascun modello NARMAX può essere visto come lo sviluppo su una opportuna base polinomiale della legge che regola il sistema. Ciò che distingue un modello da un altro è il numero e l'insieme dei termini presenti nello sviluppo.
La ricerca esaustiva tra tutte le possibili combinazioni di termini NARMAX è tuttavia quasi sempre computazionalmente improponibile , così si è puntato negli ultimi anni a sviluppare metodi di campionamento efficaci che, avvalendosi di una catena di Markov opportunamente costruita, esplorano in maniera non esaustiva l'insieme dei modelli ottenendo comunque statistiche significative.

Chapter 2

Identificazione bayesiana e metodi di campionamento Monte Carlo

2.1 Approccio Bayesiano per l'identificazione

L'inferenza bayesiana è un approccio all'inferenza statistica in cui le probabilità non sono interpretate come frequenze ma piuttosto come livelli di fiducia nel verificarsi di un dato evento. Il nome deriva dal teorema di Bayes, che costituisce il fondamento di questo approccio. Gli statistici bayesiani sostengono che i metodi dell'inferenza bayesiana rappresentano una formalizzazione del metodo scientifico, che normalmente implica la raccolta di dati che avvalorano o confutano una data ipotesi. Queste caratteristiche rendono l'approccio bayesiano un utile ausilio per discriminare tra alternative in conflitto e dunque un ottimo strumento per l'identificazione dei sistemi. Il metodo usa una stima del grado di fiducia in una data ipotesi prima dell'osservazione dei dati al fine di associare un valore numerico al grado di fiducia in quella stessa ipotesi successivamente all'osservazione dei dati. Si supponga di voler identificare un sistema scegliendo il più adatto in un insieme M di modelli. Sia M una variabile aleatoria che rappresenta la legge che regola il vero sistema. L'obiettivo dell'identificazione bayesiana è quello di calcolare per ogni $m \in M$ la probabilità a posteriori

$$P(M = m|Y) \tag{2.1}$$

dunque complessivamente determinare la distribuzione della probabilità sull'insieme dei modelli condizionatamente alle misure. La probabilità condizionata è stata definita in termini della probabilità congiunta e marginale dei due eventi

$$P(M = m|Y) = P(M = m, Y)P(Y) \tag{2.2}$$

la definizione corrisponde all'idea ragionevole

$$P(M = m|Y) \propto P(M = m, Y) \tag{2.3}$$

Avendo però ristretto l'insieme di supporto su cui è definita la metrica di probabilità, è necessario imporre che valgano ancora gli assiomi della probabilità, in

particolare chiamando k la costante di proporzionalità e integrando su tutti i casi possibili si ha

$$P(M = m|Y)dm = k \cdot P(M = m, Y)dm = k \cdot P(Y) := 1 \tag{2.4}$$

da cui si ricava il valore di k

$$k = \frac{1}{P(Y)} \tag{2.5}$$

ottenendo la definizione di probabilità condizionata. Scambiando i ruoli delle variabili aleatorie si ha anche che

$$P(Y|M = m) = P(M = m, Y)P(M = m) \tag{2.6}$$

dunque mettendo insieme le equazioni [EQUAZIONI] si ottiene la formula di Bayes

$$P(M = m|Y) = \frac{P(Y|M = m)P(M = m)}{P(Y)} \tag{2.7}$$

usando il teorema della probabilità totale si pu'o inoltre esprimere la costante al denominatore in funzione delle quantità al numeratore ottenendo

$$P(M = m|Y) = \frac{P(Y|M = m)P(M = m)}{\int P(Y|M = m)P(M = m)dm} \tag{2.8}$$

Nei problemi di identificazione l'espressione della probabilità a posteriori è quasi sempre impossibile da valutare analiticamente sopratutto per colpa dell'integrale al denominatore che deve essere calcolato su tutti i possibili modelli; si ricorre quindi a metodi numerici di tipo Monte Carlo basati sul campionamento della distribuzione. Spesso risulta impossibile anche campionare direttamente la distribuzione e se ne deve ottenere una stima accettando o scartando opportunamente i campioni estratti da una distribuzione più semplice.

2.2 Idea dei metodi Monte Carlo

I metodi Monte Carlo sfruttano il seguente ragionamento: se si vuole campionare una distribuzione con densità $p(x)$ (nel nostro caso la posterior sui modelli) si ha

$$p(x) = p(x) \otimes \delta(x) = \int_{\xi} p(\xi)\delta(x - \xi)d\xi \tag{2.9}$$

se poi $p(x)$ è fattorizzabile come

$$p(x) = q(x)g(x) \tag{2.10}$$

tale che $q(x)$ sia una densità di probabilità ovvero

- $q(x) > 0$ per ogni m
- $\int_m q(x) = 1$

allora

$$p(x) = \int_{\xi} q(\xi)g(\xi)\delta(x - \xi) = E[g(Q)\delta(x - Q)] \tag{2.11}$$

con Q variabile aleatoria avente come densità $q(\cdot)$. Usando uno stimatore si può approssimare il valore atteso come

$$E[g(Q)\delta(x-Q)] \simeq \frac{1}{N}\sum_{k=1}^{N} g(q_k)\delta(x-q_k) \tag{2.12}$$

con $q_k \sim q(\cdot)$ campione estratto dalla distribuzione di densità di proposal $q(\cdot)$ In pratica si estraggono numerosi modelli da una distribuzione di proposal $q(\cdot)$, e si costruisce un istogramma pesato con i pesi

$$g(q_k) = \frac{p(q_k)}{q(q_k)} \tag{2.13}$$

Si noti che nella formula dei pesi la desnsità $p(\cdot)$ è solamente una funzione da valutare in uno specifico modello, cosa che quasi sempre è fattibile perchè richiede una informazione che è locale nello spazio dei modelli, al contrario del problema di campionare direttamente la densità che richiederebbe una conoscenza della funzione su tutto lo spazio.

2.3 Importance sampling: scelta della proposal

In teoria le precedenti considerazioni sono già sufficienti per campionare la distribuzione, in pratica la scelta di una proposal opportuna è critica per l'accuratezza dell'algoritmo. Nella realtà, avendo a disposizione un tempo limitato, e quindi un numero limitato di campionisi otterrà una approssimazione della distribuzione cercata. La proposal è determinante nell'imporre alcuni criteri di approssimazione
Partendo dal presupposto ovvio che non è possibile esplorare esaustivamente lo spazio dei modelli, bisogna fare delle assuzioni su quali regioni dello spazio dei modelli esplorare maggiormente (e quindi descrivere) e quali invece trascurare. E' sensato disinteressarci delle regioni in cui i modelli hanno bassa probabilità. Ai fini dell'identificazione, non ci interesserà mai confrontare due modelli con probabilità bassa. Si può benissimo sbagliare o ignorare il rapporto reciproco tra le loro probabilità per concentrarsi sulla descrizione delle regioni dove i modelli hanno probabilità più alta. Questo induce a scegliere la proposal quanto più simile alla posterior, in modo da estrarre campioni prevalentemente dove la posterior è alta. Tale criterio è detto di *importance sampling.*

Al limite se riuscissi ad essere sicuro che sto campionando da $q(\cdot) = p(\cdot)$ avrei che i pesi si ridurrebbero a

$$g(q_k) = \frac{p(q_k)}{q(q_k)} = \frac{p(q_k)}{p(q_k)} = 1 \tag{2.14}$$

e dunque

$$p(x) \simeq \frac{1}{N}\sum_{k=1}^{N} \delta(x-q_k) \tag{2.15}$$

Ovvero potrei avere una approssimazione di $p(\cdot)$ semplicemente valutando la frequenza relativa con cui è stato estratto $q_k = x$. Attenzione, l'ipotesi di

sapere che $q(\cdot) = p(\cdot)$ non vuol dire che conosco la forma della funzione su tutto lo spazio, ma solo che

- so valutare la funzione in un particolare modello
- ho un metodo (anche indiretto) per campionare la funzione

Gli algoritmi Monte Carlo Markov Chain infatti ottengono i modelli q_k come stati di una catena di Markov opportunamente costruita sullo spazio dei modelli. A tale catena si chiede di avere come unica probabilità di regime proprio $p(\cdot)$.

Dopo un transitorio iniziale la catena finirà di fatti per campionare i suoi stati dalla distribuzione di regime $p(\cdot)$.

Chapter 3

Algoritmo Metropolis Hasting

Supponiamo di voler estrarre campioni da una distribuzione p(m) nota, difficile da campionare direttamente. Si pu'o chiedere che la distribuzione obiettivo sia distribuzione di regime di una catena di Markov. Una catena si dice ergodica se dopo un certo tempo (transitorio iniziale) , essa converge all'unica distribuzione di regime indipendentemente dalla distribuzione di partenza. Le condizioni necessarie per l'ergodicit'a della catena sono:

- **irriducibilit'** : la probabilit'a di visitare ciascuno stato a partire da ciascuno stato 'e strettamente positiva
- **aperiodicità** : una catena 'e periodica se pu'o ritornare in un certo stato solo a istanti multipli di un qualche intero maggiore di 1. Una catena 'e aperiodica se non 'e periodica.

In poche parole, fissato un qualsiasi istante abbastanza grande e un qualsiasi stato, deve essere possibile una storia temporale della catena che la porta a risiedere in quello stato a quell'istante. La questione diventa quindi come scegliere il kernel

$$s(x'|x)$$

(probabilit'a di transizione dal vecchio stato m al nuovo stato m') della catena, in modo che la catena sia ergodica e che la distribuzione di regime sia proprio p(m). Una condizione sufficiente non necessaria 'e che la distribuzione obiettivo soddisfi la condizione di reversibilit'a

$$p(x)s(x'|x) = p(x')s(x|x') \tag{3.1}$$

L'operatore di transizione è quello che nelle catene a stato discreto e finito era rappresentato dalla matrice di transizione mentre nelle catene a stato continuo è il funzionale

$$Tr[\bullet] = \int \bullet(x)s(x'|x)dx \tag{3.2}$$

Una densit'a di probabilit'a $\hat{p}$ 'e detta stazionaria se 'e autofunzione dell'operatore di transizione della catena, ovvero

$$Tr[\hat{p}] = \hat{p} \tag{3.3}$$

e semplice dimostrare che se vale la condizione di reversibilit‘a allora la distribuzione ‘e stazionaria infatti

$$Tr[\hat{p}] = \int \hat{p}(x)s(x'|x)dx = \int \hat{p}(x')s(x|x')dx =$$
$$\hat{p}(x') \int s(x|x')dx = \hat{p}(x') \cdot 1 = \hat{p}(x')$$

Se si sceglie un kernel arbitrario uguale ad una probabilit‘a di transizione facile da campionare $s(x'|x) = q(x'|x)$ solitamente la condizione di reversibilit‘a non ‘e soddisfatta. Solitamente si ha uno sbilanciamento che significa che alcune transizioni sono pi‘u probabili in un certo verso.

$$\hat{p}(x)s(x'|x) > \hat{p}(x')s(x|x') \tag{3.4}$$

In tal caso, viste le distribuzioni di partenza e la probabilit‘a di transizione scelta ‘e pi‘u probabile osservare la transizione $x \to x'$ piuttosto che la transizione $x' \to x$ Si cerca quindi di ristabilire l'equilibrio scegliendo come kernel la probabilit‘a di transizione moltiplicata per un fattore correttivo

$$q(x'|x) = \gamma(x'|x)q(x'|x) \tag{3.5}$$

In particolare si pu‘o interpretare $\gamma(x'|x)$ come la probabilit‘a di accettare la mossa $x \to x'$. Imponendo quindi che valga

$$\hat{p}(x)\gamma(x'|x)q(x'|x) = \hat{p}(x')\gamma(x|x')q(x|x') \tag{3.6}$$

si ricava

$$\frac{\gamma(x'|x)}{\gamma(x|x')} = \frac{\hat{p}(x')q(x|x')}{\hat{p}(x)q(x'|x)} \tag{3.7}$$

Perch‘e si abbia coerenza con la eq[CITARE EQUAZIONE] bisogna che il rapporto al membro sinistro sia minore di 1. In particolare si può scegliere

- $\gamma(x|x') = 1$ che equivale ad accettare tutte le transizioni $m' \to m$ che sono meno frequenti
- $\gamma(x'|x) = min\left\{1, \frac{\hat{p}(x')q(x|x')}{\hat{p}(x)q(x'|x)}\right\}$ fattore di riduzione delle transizioni pi‘u frequenti

L'algoritmo MH è così definito

Algoritmo Metropolis Hastings

Inizializza $X_0 = x_0$
for $t = 0, 1, 2 \dots T$ **do**
$i \leftarrow 0$
Estraggo la proposta per il nuovo stato $x^t \sim q(x^t|X_t)$
Calcolo la probabilità di accettare la mossa $\gamma(x'|x) = min\left\{1, \frac{\hat{p}(x^t)q(X_t|x^t)}{\hat{p}(X_t)q(x^t|X_t)}\right\}$
Estraggo una variabile da una distribuzione uniforme $u \sim U(0,1)$
if $u \leq \gamma(x^t|X_t)$ **then**
La catena transita nel nuovo stato proposto $X_{t+1} = x^t$
else
La catena rimane nel vecchio stato $X_{t+1} = X_t$
end if
end for

Chapter 4

Reversible Jump Monte Carlo Markov Chain

Nelle sezioni precedenti la transizione della catena associava (in maniera aleatoria) uno stato di $X \subset \mathbb{R}^n$ ad uno stato di $X' \subset \mathbb{R}^n$. Cosa succede se la transizione dovesse associare uno stato $M \subset \mathbb{R}^n$ ad uno stato di $X \subset \mathbb{R}^m$ con $m = n$? La questione scaturisce dal fatto che per gli scopi dell'identificazione dei sistemi, restringersi al caso $m = n$ 'e molto limitativo e corrisponde a conoscere in partenza il numero dei parametri da identificare. Per i sistemi non lineari spesso si considerano modelli del sistema che sono sviluppi polinomiali di equazioni alle differenze e si lascia all'algoritmo di identificazione l'onere di determinare quali e quanti termini dello sviluppo includere. In base al numero di termini scelti si avranno da scegliere anche altrettanti coeffici- enti dello sviluppo. Quando l'algoritmo suggerisca di aggiungere o eliminare uno (o pi'u) termini dello sviluppo 'e necessario aggiornare anche la dimensione del vettore dei coefficienti. Ecco quindi che acquistano senso mosse che portano lo stato della catena da uno spazio con una certa dimensionalit'a ad un'altro con dimensionalit'a diversa. Si pensi quindi di enumerare le tipologie di modello (anche infinite), si avr'a che l'insieme delle possibili strutture di modello 'e rappresentabile come un insieme di indici

$$\mathcal{K} = \{1, 2, \dots k \dots\} \subset \mathbb{N}$$

Ciascun modello 'e associato uno spazio dei coefficienti dello sviluppo (o in generale dei parametri del modello). Si ha quindi che a ciascuna tipologia di modello 'e associato uno spazio

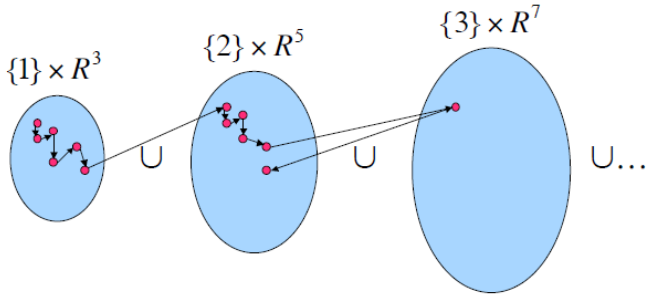
$$X_k \subset \mathbb{R}^{n(k)}$$

dove $n(k)$ 'e il numero di termini previsti dal modello indicizzato da k. Si pu'o quindi pensare che la ricerca del modello avvenga in una unione di sottospazi definita da

$$X = \bigcup_{k \in \mathcal{K}} (\{k\} \times X_k) \tag{4.1}$$

L'inclusione esplicita dell'indice k si ha perch'e spesso 'e necessario che l'algoritmo vi acceda, specialmente quando esistono pi' u modelli diversi a cui corrisponde

lo stesso spazio dei parametri. Lo stato della catena ‘e quindi rappresentabile come coppia $x = (k, \theta_k)$ dove k indica l'indice della struttura di modello e θ_k indica i parametri associati a tale struttura, il numero di parametri dipende da k. Se pur lo spazio appena costruito sia un po' inusuale, l'algoritmo MH ‘e ancora valido, si tratta di vedere come costruire una catena di Markov che abbia lo stato in questo insieme X e che abbia una desiderata densit‘a di regime $\hat{p}(\cdot)$.



Ci limitiamo a considerare i casi in cui, per qualche distribuzione di probabilit‘a di partenza la probabilit‘a che la catena risieda in un sottoinsieme $A \subset X$ e muova verso un certo $B \subset X$ sia la stessa qualora si scambino i ruoli di A e di B. Questa richiesta ‘e detta di equilibrio bilanciato , se verificata da una certa probabilit‘a ‘e una condizione sufficiente affinch‘e questa sia stazionaria.
Spesso per attraversare lo spazio X ‘e necessario avere a disposizione diverse tipolo- gie di mosse. Una tipologia di mossa associa la struttura del modello attuale ad una nuova strut- tura. Affinch‘e la condizione di equilibrio bilanciato sia soddisfatta ‘e necessario che se es- iste la mossa $(x \to x')$ esista anche la mossa inversa $(x' \to x)$. Si indichi con $j_m(x)$ la probabilit‘a che dallo stato x sia scelta la mossa di tipo m e si indichi con $j_m^{-1}(x)$la probabilit‘a che dallo stato x sia scelta la mossa di tipo m inversa. Il problema ‘e che non ha senso paragonare probabilit‘a definite su spazi di dimen- sione diversa. Per rendere invisibile alla catena il cambio di dimensionalit‘a basta immergere i due spazi di dimensione diversa in uno stesso spazio di dimensione maggiore e definire in tale spazio la probabilit‘a. Si adotti quindi il seguente protocollo (ideato da Peter Green nel 1995):

- Si estragga la tipologia di mossa.
- Si generino r numeri casuali $u \in \mathbb{R}^r$ da una specificata densit‘a g.
- Si costruisca il nuovo stato attraverso una funzione deterministica h di x e di u.
$$(x', u') = h(x, u) \tag{4.2}$$
In questo caso sono stati indicati con u gli r numeri che sono estratti casualmente dalla distribuzione g quando si effettua la mossa inversa da x' a x usando la funzione deterministica inversa
$$(x, u) = h^{-1}(x', u') \tag{4.3}$$
l'equazione di equilibrio bilanciato si pu‘o scrivere dunque come

$$\sum_m \int_{(x,x')\in A\times B} j_m(x)\hat{p}(x)g_m(u)\gamma_m(x'|x)dxdu = \sum_m \int_{(x,x')\in A\times B} j_m^{-1}(x')\hat{p}(x')g_m(u')\gamma_m(x'|x)dx'du'$$

La sommatoria sulle possibili mosse deriva dal fatto che ad ogni iterazione la tipolo- gia di mossa ‘e esclusiva. Una condizione sufficiente non necessaria perch‘e valga il bilancio ‘e che esso valga mossa per mossa ovvero

$$\int_{(x,x')\in A\times B} j_m(x)\hat{p}(x)g_m(u)\gamma_m(x'|x)dxdu = \int_{(x,x')\in A\times B} j_m^{-1}(x')\hat{p}(x')g_m(u')\gamma_m(x'|x)dx'du'$$

se inoltre la funzione h ‘e un diffeomorfismo vale la formula classica del cambio di variabili e quindi si pu‘o passare dall'uguaglianza integrale all'uguaglianza delle integrande. Dal momento che la mossa ‘e fissata si ha che k e k' sono costanti dunque il cambio di variabili richiesto riguarda solo il vettore dei parametri e delle variabili ausiliarie.

$$(\theta'_k, u') \to (\theta_k, u)$$

Applicandolo si ottiene

$$j_m(x)\hat{p}(x)g_m(u)\gamma_m(x'|x) = j_m^{-1}(x')\hat{p}(x')g_m(u')\gamma_m(x'|x)\left|\frac{\partial(\theta_{k'}, u')}{\partial(\theta_k, u)}\right|$$

Condizione necessaria affinch‘e h sia un diffeomorfismo ‘e che

$$dim(\theta_{k'}) + dim(u') = dim(\theta_k) + dim(u) \tag{4.4}$$

detta condizione di *dimension matching*. Si ha dunque in tal caso che

$$\gamma_m(x'|x) = \min\left\{1, \frac{j_m^{-1}(x')\hat{p}(x')g_m(u')\gamma_m(x'|x)}{j_m(x)\hat{p}(x)g_m(u)}\left|\frac{\partial(\theta_{k'}, u')}{\partial(\theta_k, u)}\right|\right\} \tag{4.5}$$

Chapter 5

Il modello

5.1 Modello Narmax

Grazie all'uso di un modello esplicito di rumore, il modello NARMAX 'e in grado di modellare l'effetto di rumori non bianchi, correlati a causa delle nonlinearit'a del sistema. Un sistema SISO tempo discreto pu'o essere rappresentato da un modello NARMAX descritto da

$$y_t = f(y_{t-1}, \ldots, y_{t-n_y}, y_{t-1}, \ldots, y_{t-n_y}, e_{t-1}, \ldots, e_{t-n_e}) + e_t \tag{5.1}$$

dove si assume che $f(\cdot)$ sia una funzione non lineare sconosciuta, $u_t \in \mathbb{R}$ e $y_t \in \mathbb{R}$ siano gli ingressi e uscite del sistema e $e_t \in \mathbb{R}$ denota un termine di rumore estratto dalla distribuzione normale $\mathcal{N}\left(0, \sigma_e^2\right)$.
Gli ordini della dinamica sono rispettivamente n_u, n_y, n_e. Decomponiamo la funzione f $(\cdot)$ in una somma di funzioni di base polinomiali e riesprimiamo la 5.1 come combiaione di termini di processo e di rumore,

$$y_t = \sum_{j=1}^{M_p} \left(a_j y_{t-\delta_{y,j}}^{k_{y,j}} u_{t-\delta_{u,j}}^{k_{u,j}} \right) + \sum_{j=1}^{M_e} \left(b_j y_{t-\delta_{y,j}}^{k_{y,j}} u_{t-\delta_{u,j}}^{k_{u,j}} e_{t-\delta_{e,j}}^{k_{e,j}} \right) + e_t \tag{5.2}$$

dove il modello di processo 'e composto da M_p monomi combinazione di soli termini di uscita e di rumore e il modello di rumore 'e composto da M_e monomi combinazione di ingresso, uscita e rumore. Con $k_{y,j}$, $k_{u,j}$, $k_{e,j} \in N$ si indicano le potenze che compaiono nei monomi, con $\delta_{y,j}$, $\delta_{e,j} \in \mathbb{N} - \{0\}$ e $\delta_{u,j} \in \mathbb{N}$ il ritardo delle varie grandezze.

5.2 Equazione di regressione

Data una sequenza di N ingressi

$$\{u(1)u(2)u(3)u(4)\ldots u(N)\} \tag{5.3}$$

e N uscite

$$\{y(1)y(2)y(3)y(4)\ldots y(N)\} \tag{5.4}$$

si hanno a disposizione

$$N - max\{n_u, n_y, n_e\} \tag{5.5}$$

equazioni.
Per esempio se l'insieme dei termini di processo scelti fosse

$$\mathcal{P} = \{y^2(t-1), y(t-2)u(t)\} \tag{5.6}$$

e l'insieme dei termini di rumore

$$\mathcal{E} = e^3(t-1)u(t-1) \tag{5.7}$$

e si avessero a disposizione gli ingressi

$$u = \{u(1), u(2), u(3), u(4), u(5), u(6)\} \tag{5.8}$$

e le uscite

$$y = \{y(1), y(2), y(3), y(4), y(5), y(6)\} \tag{5.9}$$

Si avrebbero a disposizione le equazioni di regressione

$$\begin{bmatrix} y(6) \\ y(5) \\ y(4) \\ y(3) \end{bmatrix} = \begin{bmatrix} y^2(5) & y(4)u(6) \\ y^2(4) & y(3)u(5) \\ y^2(3) & y(2)u(4) \\ y^2(2) & y(1)u(3) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} + \begin{bmatrix} e^3(5)u(5) \\ e^3(4)u(4) \\ e^3(3)u(3) \\ e^3(2)u(2) \end{bmatrix} b_1 + \begin{bmatrix} \epsilon(6) \\ \epsilon(5) \\ \epsilon(4) \\ \epsilon(3) \end{bmatrix} \tag{5.10}$$

che possono essere scritte compattamente in forma matriciale come

$$\mathbf{y} = \mathbf{P}_k \mathbf{a}_k + \mathbf{E}_q \mathbf{b}_q + \epsilon \tag{5.11}$$

5.3 Densit'a di probabilit'a a posteriori del modello

$$p(k, q, \mathbf{P}_k, \mathbf{E}_q, \mathbf{a}_k, \mathbf{b}_q | \mathbf{y}) \propto p(\mathbf{y} | k, q, \mathbf{P}_k, \mathbf{E}_q, \mathbf{a}_k, \mathbf{b}_q, \sigma_e^2) p(k, q, \mathbf{P}_k, \mathbf{E}_q, \mathbf{a}_k, \mathbf{b}_q) \tag{5.12}$$

5.3.1 Likehood

Se il rumore 'e gaussiano a media nulla anche la verosimiglianza 'e una gaussiana, fun- zione del vettore dei residui, di stessa varianza. Infatti come si deduce dall'equazione di regressione [CITARE EQUAZIONE] si ha

$$p(\mathbf{y} | k, \mathbf{P}_k, eq, \mathbf{a}_k, \mathbf{b}_q, \sigma_e^2) = p(\mathbf{P}_k \mathbf{a}_k + \mathbf{E}_q \mathbf{b}_q + \epsilon | k, q, \mathbf{P}_k, \mathbf{E}_q, \mathbf{a}_k, \mathbf{b}_q, \sigma_e^2) = p(\epsilon | k, q, \mathbf{P}_k, \mathbf{E}_q, \mathbf{a}_k, \mathbf{b}_q, \sigma_e^2) \tag{5.13}$$

e supponendo di aver identificato correttamente i termini e i parametri in modo tale da avere come errore residuo (ineliminabile) il rumore di misura si ha

$$p(\mathbf{e} | k, q, \mathbf{P}_k, \mathbf{E}_q, \mathbf{a}_k, \mathbf{b}_q, \sigma_e^2) = p(\epsilon | k, q, \mathbf{P}_k, \mathbf{E}_q, \mathbf{a}_k, \mathbf{b}_q, \sigma_e^2) = \frac{1}{\sqrt{2\pi\sigma_e^2}^N} exp\left(-\frac{1}{2\sigma_e^2}\epsilon^T \epsilon\right) \tag{5.14}$$

dunque in definitiva

$$p(\mathbf{y} | k, \mathbf{P}_k, eq, \mathbf{a}_k, \mathbf{b}_q, \sigma_e^2) = \frac{1}{\sqrt{2\pi\sigma_e^2}^N} exp\left(-\frac{1}{2\sigma_e^2}\epsilon^T \epsilon\right) \tag{5.15}$$

con

$$\epsilon = \mathbf{y} - \mathbf{P}_k \mathbf{a}_k - \mathbf{E}_q \mathbf{b}_q \tag{5.16}$$

5.4 Scelta delle distribuzioni a priori e iperparametri

In assenza di preferenze, la scelta delle distribuzioni a priori solitamente ricade su distribuzioni poco informative ovvero densit'a con un supporto ampio e varianza grande.
Per aumentare la flessibilit'a si adotta una struttura gerarchica in cui gli stessi parametri delle distribuzioni a priori sono a loro volta realizzazioni di variabili aleato- rie. Per convenienza si sceglie di rappresentare i parametri con una distribuzione a prori coniugata della likehood.

5.5 Il concetto di distribuzioni coniugate

Sia X una variabile che modella il sistema estratta da una distribuzione $p(X|\zeta) = f(\zeta, \cdot)$ dipendente dal parametro ζ. Al livello gerarchico pi' u alto $p(X|\zeta)$ rappresenta la probabilit'a a priori (rispetto alle misure) della variabile X . L'ipotesi sull'andamento della funzione f si suppone quindi gi'a fatta a tale livello gerarchico. Al livello gerarchico inferiore la stessa $p(X|\zeta)$ rappresenta invece la likehood (*dalla forma nota!!*) del parametro ζ . La scelta della forma del prior $p(\zeta)$ non 'e quindi indipendente dalla scelta del posterior $p(\zeta|X)$, infatti

$$p(\zeta|X) \propto p(X|\zeta)p(\zeta) \tag{5.17}$$

La distribuzione $p(X|\zeta)$ 'e quindi un operatore che mappa una funzione in un'altra funzione. E' possibile per tale operatore trovare una sorta di *autofunzione*, nel senso pi' u lato di distribuzione appartenente a una certa famiglia (normale, uniforme, pois- son etc) che viene mappata dall'operatore in un'altra funzione ancora appartenete alla medesima famiglia. Data una certa likehood le sue *autofunzioni* in questo senso, vengono dette dis- tribuzioni coniugate. Se i prior sono scelti tra le distribuzioni coniugate, quando (mediante l'applicazione dell'operatore likehood) evolvono non cambiano la forma d'onda ma solo i parametri che la descrivono. Quindi la descrizione di una dinamica su uno spazio funzionale a infinite dimensioni viene rappresentata dalla dinamica in uno spazio il cui numero di dimensioni 'e finito e ridotto. (Ad esempio l'evoluzione di una gaussiana pu'o essere rappresentata concisamente dall'evoluzione del suo valor medio e della sua varianza).
Se i prior non sono scelti tra le distribuzioni coniugate, invece, ad ogni applicazione dell'operatore likehood cambia l'intera forma d'onda della distribuzione diventando impossibile da descrivere analiticamente.

5.5.1 Probabilit'a del modello a priori

Modificando leggermente la [CITARE EQUAZIONE] per tenere conto anche degli iperparametri (de- stritti nel dettaglio in questa sezione) si ha

$$p(k, q, \mathbf{P}_k, eq, \mathbf{a}_k, \mathbf{b}_q, \sigma_a^2, \sigma_b^2, \lambda_a, \lambda_b|\mathbf{y}) \propto p(\mathbf{y}|k, q, \mathbf{P}_k, ak, \mathbf{b}_q, ve)p(k, q, \mathbf{P}_k, \mathbf{E}_q, \mathbf{a}_k, \mathbf{b}_q, \sigma_a^2, \sigma_b^2, \lambda_a, \lambda_b) \tag{5.18}$$

fattorizzabile come

$$\begin{aligned} p(k, q, \mathbf{P}_k, eq, \mathbf{a}_k, \mathbf{b}_q, \sigma_a^2, \sigma_b^2, \lambda_a, \lambda_b|\mathbf{y}) =& p(\mathbf{a}_k|\mathbf{P}_k, k, \sigma_a^2)p(k|\lambda_a)p(\mathbf{P}_k)p(\lambda_a)p(\sigma_a^2)\times \\ & p(\mathbf{b}_q|\mathbf{E}_q, q, \sigma_b^2)p(q|\lambda_b)p(\mathbf{E}_q)p(\lambda_b)p(\sigma_b^2) \end{aligned} \tag{5.19}$$

Di seguito vediamo nel dettaglio come sono state modellate le distribuzioni a priori

Vettore dei parametri dei termini di modello

Il vettore di parametri $\mathbf{a}_k$ si assume distribuito come una gaussiana multidimensionale isotropica a media nulla ovvero

$$p(\mathbf{a}_k|k, \mathbf{P}_k, \sigma_a^2) = p(\mathbf{a}_k|k, \sigma_a^2) \sim \mathcal{N}\left(\mathbf{0}, \sigma_a^2\mathbf{I}\right) \tag{5.20}$$

Questa scelta implica che la varianza della gaussiana sia distribuita come una gamma inversa in modo da essere coniugata. La distribuzione gamma inversa 'e definita come segue

$$\mathcal{IG}(x|\alpha, \beta) = x^{-\alpha-1} \exp\left(-\frac{\beta}{x}\right) \tag{5.21}$$

Di seguito si prova che la distribuzione gamma inversa 'e effettivamente coniugata della gaussiana per l'iperparametro varianza

Proof. 'e sufficiente applicare la regola di Bayes e ottenere ancora una gamma inversa

$$\begin{aligned}
p(\sigma_a^2|k, \mathbf{a}_k) &\propto p(\mathbf{a}_k, k, \sigma_a^2) \cdot p(\sigma_a^2) \\
&\propto \frac{1}{\sqrt{\sigma_a^2}^k} exp\left(\frac{-\mathbf{a}_k^T\mathbf{a}_k}{2\sigma_a^2}\right) \sigma_a^{2(-\alpha_a-1)} exp\left(-\frac{\beta_a}{\sigma_a^2}\right) \\
&\propto \sigma_a^{2(-\alpha_a-1-\frac{1}{2}k)} exp\left(-\frac{\beta_a + \frac{1}{2}\mathbf{a}_k^T\mathbf{a}_k}{\sigma_a^2}\right) \\
&\propto \sigma_a^{2(-\alpha_a-1-\frac{1}{2}k)} exp\left(-\frac{\beta_a + \frac{1}{2}\mathbf{a}_k^T\mathbf{a}_k}{\sigma_a^2}\right) \\
&\propto \mathcal{IG}(\sigma_a^2|\alpha + \frac{1}{2}k, \beta_a + \frac{1}{2}\mathbf{a}_k^T\mathbf{a}_k)
\end{aligned}$$

□

Vettore dei parametri dei termini di rumore

La modellazione del vettore dei parametri dei termini di rumore 'e analoga a quella trattata nella sottosezione precedente. Il vettore di parametri $\mathbf{b}_q$ si assume dis- tribuito come una gaussiana multidimensionale isotropica a media nulla

$$p(\mathbf{b}_q|k, \mathbf{E}_q, \sigma_b^2) = p(\mathbf{b}_q|q, \sigma_b^2) \sim \mathcal{N}\left(\mathbf{0}, \sigma_b^2\mathbf{I}\right) \tag{5.22}$$

Questa scelta implica che la varianza della gaussiana sia distribuita come una gamma inversa in modo da essere coniugata. Con dimostrazione del tutto analoga alla sezione precedente si dimostra che la varianza della gaussiana deve essere una gamma inversa.

$$\begin{aligned}
p(\sigma_b^2) = \mathcal{IG}(\sigma_b^2|\alpha_b, \beta_b) \Rightarrow p(\sigma_b^2|q, \mathbf{b}_q) &\propto p(\mathbf{b}_q, q, \sigma_b^2) \cdot p(\sigma_b^2) \\
&\propto \mathcal{IG}(\sigma_b^2|\alpha_b + \frac{1}{2}q, \beta_b + \frac{1}{2}\mathbf{b}_q^T\mathbf{b}_q)
\end{aligned}$$

Numero di termini di processo

La probabilit‘a a priori del numero di termini di processo ‘e rappresentata da una distribuzione di Poisson troncata, il parametro λ_a.

$$p(k|\lambda_a) = \frac{\frac{\lambda_a^k}{k!}}{\sum_{i=0}^{N} \frac{\lambda_a^i}{i!}} \tag{5.23}$$

Con N numero massimo di termini di processo. Questo tipo di distribuzione ‘e pi‘ u indicato rispetto ad una distribuzione uniforme, perch‘e l'iperparametro λ_a pu‘o essere interpretato come il numero medio di termini ipotizzando un numero di termini possibili sufficientemente alto infatti

$$\mathbb{E}[k|\lambda_a] = \frac{\sum_{k=0}^{N} k\frac{\lambda_a^k}{k!}}{\sum_{i=0}^{N} \frac{\lambda_a^i}{i!}} = \frac{\sum_{k=1}^{N} k\frac{\lambda_a^k}{k!}}{\sum_{i=0}^{N} \frac{\lambda_a^i}{i!}} = \lambda_a \frac{\sum_{j=0}^{N-1} k\frac{\lambda_a^j}{j!}}{\sum_{i=0}^{N} \frac{\lambda_a^i}{i!}} = \lambda_a \frac{\sum_{j=0}^{N-1} k\frac{\lambda_a^j}{j!}}{\sum_{i=0}^{N-1} \frac{\lambda_a^i}{i!} + \frac{\lambda_a^N}{N!}} = \lambda_a \frac{1}{1 + \frac{\frac{\lambda_a^N}{N!}}{\sum_{i=0}^{N-1} \frac{\lambda_a^i}{i!}}} \tag{5.24}$$

si noti che

$$\lim_{N\to\infty} \frac{\frac{\lambda_a^N}{N!}}{\sum_{i=0}^{N-1} \frac{\lambda_a^i}{i!}} = 0 \tag{5.25}$$

L'iperparametro λ_a ‘e estratto da una distribuzione gamma

$$p(\lambda_a) = \mathcal{GA}(\lambda_a|\alpha_{\lambda_a}, \beta_{la}) = \lambda_a^{(\alpha_{\lambda_a}-1)} exp\left(-\frac{\lambda_a}{\beta_{\lambda_a}}\right) \tag{5.26}$$

Di seguito dimostro che la distribuzione gamma ‘e coniugata per il parametro della poisson *Proof.*

$$\begin{aligned} p(\lambda_a|k) &\propto p(k|\lambda_a)p(\lambda_a) \\ &\frac{\lambda_a^k}{k!}\lambda_a^{(\alpha_{\lambda_a}-1)} exp\left(-\frac{\lambda_a}{\beta_{\lambda_a}}\right) \qquad \propto \frac{1}{k!}\mathcal{GA}(\lambda_a|\alpha_{\lambda_a}+k, \beta_{\lambda_a}) \end{aligned}$$

□

Numero di termini di rumore

Analogamente a quanto detto nella sezione precedente, il numero dei termini di rumore ‘e modellato come una variabile aleatoria Poisson troncata

$$p(q|\lambda_b) = \frac{\frac{\lambda_b^q}{q!}}{\sum_{i=0}^{N} \frac{\lambda_b^i}{i!}} \tag{5.27}$$

e il suo iperparametro λ_b ‘e modellato come una variabile aleatoria gamma-distribuita

$$p(\lambda_b) = \mathcal{GA}(\lambda_b|\alpha_{\lambda_b}.\beta_{\lambda_b}) = \lambda_b^{\alpha_{\lambda_b}-1)} exp\left(-\frac{\{}{\beta_{\lambda_a}}\right) \tag{5.28}$$

Matrice di regressione dei termini di processo e di rumore

La matrice di regressione dei termini ‘e considerata una va aleatoria uniformemente distribuita in modo che nessun termine di modello n‘e di rumore sia, a priori delle misure, pi‘ u probabile di altri

$$p(\mathbf{P}_k) \propto 1 \tag{5.29}$$

$$p(\mathbf{E}_q) \propto 1 \tag{5.30}$$

Chapter 6

Algoritmo RJMCMC per l'identificazione di modelli NARMAX

L'algoritmo MH è così definito

Algorithm RJMCMC for NARMAX identification

fissa i parametri di tuning c e σ_e^2

inizializza$\left(k^{(0)}, q^{(0)}, \mathbf{P}_k^{(0)}, \mathbf{E}_q^{(0)}, \mathbf{a}_k^{(0)}, \mathbf{b}_q^{(0)}, \lambda_a^{(0)}, \lambda_b^{(0)}, \sigma_a^{(0)}, \sigma_b^{(0)}\right)$

for $i = 1 : N_{\text{iter}}$ **do**
  Estrai $z_k \sim \mathcal{U}[0,1]$
  if $z_k \leq B_k^{(i)}$ **then**
    Effettua la mossa di nascita (Algoritmo 1)
  else if $z_k \leq B_k^{(i)} + D_k(i)$ **then**
    Effettua mossa di morte (Algoritmo 2)
  else
    Aggiorna i parametri (Algoritmo 3)
    Aggiorna la varianza dei parametri
  end if
  Estrai $\lambda_b^{(i)} \sim p(\lambda_b | q^{(i)})$

  if $i >$ **then**
    Estrai $z_q \sim \mathcal{U}[0,1]$
    if $z_q \leq B_q^{(i)}$ **then**
      Effettua la mossa di nascita (Algoritmo 1)
    else if $z_k \leq B_q^{(i)} + D_q(i)$ **then**
      Effettua mossa di morte (Algoritmo 2)
    else
      Aggiorna i parametri (Algoritmo 3)
      Aggiorna la varianza dei parametri

end if
end if

Calcola l'errore residuo $\epsilon^{(i)} = \mathbf{y} - \mathbf{P}_k^{(i)}\mathbf{a}_k^{(i)} - \mathbf{E}_q^{(i)}\mathbf{b}_q^{(i)}$:assumilo come stima del segnale di rumore

Aggiornamento di $\mathbf{E}_q^{(i)}$ in base alla nuova stima

end for

L'algoritmo parte con l'inizializzazione di parametri e iperparametri: solitamente si parte con un modello vuoto ovvero con nessun termine (di rumore e di processo), le matrici di regressione sono in tal caso (per convenzione) una singola colonna di elementi nulli e il vettore dei parametri 'e lo scalare nullo. Gli iperparametri delle poisson troncate vengono inizializzati a quello che ci si aspetta essere il nu- mero medio di termini nello sviluppo, mentre le varianze dei coefficienti vengono inizializzate ad un valore non piccolo in modo da avere inizialmente dei prior molto dispersi e quindi poco informativi. L'inizializzazione di questi parametri non 'e crit- ica perch'e essi verranno aggiornati con l'evoluzione della catena e si adatteranno al valore pi' u appropriato. La varianza σ_e^2 del rumore bianco invece viene inizializzata e non viene pi' u aggiornata, tale parametro rappresenta infatti un parametro di tun- ing dell'algoritmo essendo in qualche modo una metrica di affidabilit'a delle misure, 'e sensato dunque che esso influisca direttamente sulla probabilit'a di accettazione delle mosse: abbassare tale parametro equivale a ritenere l'incertezza bassa quindi il rapporto di accettazione sar'a pi' u selettivo e le mosse proposte verranno scartate pi' u frequentemente. L'algoritmo prevede un numero fissato a priori di iterazioni. In corrispondenza di ciascuna iterazione la catena di markov andr'a a risiedere in uno stato che rappre- senta un modello del sistema. L'assunzione di **ergodicità** permette di ricostruire la probabilit'a che la catena risieda in un particolare stato a partire dal numero di iterazioni medio in cui la catena risiede in quello stato. E' necessario per'o adottare due accorgimenti:

- il numero di iterazioni deve essere abbastanza elevato da fare in modo che valga con buona approssimazione l'ipotesi di ergodicit'a. Se le statistiche si raccol- gono su poche iterazioni le probabilit'a che si ottengono sono molto influenzate dalla particolare realizzazione della catena.

- il numero medio di iterazioni in corrispondenza del quale la catena risiede in un particolare stato deve essere calcolato escludendo le prime iterazioni della catena (il cosiddetto periodo di burn-in ossia di riscaldamento), questo perch'e le prime iterazioni sono fortemente influenzate dalla particolare inizializzazione dei parametri e dello stato.

Nel contesto della singola iterazione eseguono due algoritmi analoghi: l'evoluzione del modello di processo e l'evoluzione del modello di rumore. Entrambi prevedono una fase in cui si seleziona la tipologia di mossa, una fase in cui si effettua la mossa scelta e infine una fase in cui si estrae un nuovo iperparametro per il numero medio di termini. Dopo che hanno eseguito questi algritmi si calcola il nuovo errore di regressione come differenza tra le uscite misurate del sistema e le uscite predette dal modello. Assumendo poi che il modello attuale sia capace di descrivere esattamente l'uscita misurata, si considera l'errore di regressione come se fosse l'attuale campione di ru- more gaussiano bianco e si aggiorna

la matrice di regressione del rumore. Per ottenere l'output dell'identificazione basta tenere in memoria (a partire da quando si ritiene esaurito il transitorio iniziale) i modelli visitati dalla catena e costruire progressivamente gli istogrammi sul numero di termini, sull'identificativo dei termini e sul coefficiente associato a ciascun identificativo dei termini. Di seguito si illustrano le tipologie di mossa previste, il meccanismo di scelta della tipologia di mossa e nelle sezioni seguenti si entra nel dettaglio di come vengono effettuate le mosse.

6.0.2 Tipologia di mosse

Le mosse che modificano lo stato della catena, nell'algoritmo dell'articolo sono:

- Nascita:
  Viene selezionato un nuovo termine tra quelli rimasti e viene inserito nello sviluppo polinomiale, togliendolo dall'insieme dei termini disponibili per una futura mossa di nascita. Il numero di termini del modello passa da k a $k' = k + 1$.
  I coefficienti dello sviluppo vengono estratti nuovamente

- Morte:
  Viene selezionato un nuovo termine tra quelli presenti nello sviluppo polinomiale e viene eliminato reinserendolo nell'insieme dei termini disponibili per una successiva mossa di nascita. Il numero di termini del modello passa da k a $k' = k - 1$
  I coefficienti dello sviluppo vengono estratti nuovamente

- Aggiornamento:
  Non si ha cambio di dimensionalit'a $k' = k$ Vengono solamente cambiati i coefficienti dello sviluppo e la varianza della distribuzione da cui sono estratti.

Mosse analoghe sono previste per i termini di rumore. Ad ogni iterazione della catena,viene estratta una delle tre mosse.
La mossa di nascita viene estratta con probabilità B_k dove

$$B_k = \begin{cases} 1 & k = 0 \\ 0 & k = M_p \\ c \cdot \min\left\{1, \frac{p(k+1|\lambda_a}{p(k|\lambda_a)}\right\} & textaltrimenti \end{cases} \tag{6.1}$$

La mossa di morte viene estratta con probabilità

$$D_k = \begin{cases} 0 & k = 0 \\ c \cdot \min\left\{1, \frac{p(k-1|\lambda_a}{p(k|\lambda_a)}\right\} & textaltrimenti \end{cases} \tag{6.2}$$

La mossa di aggiornamento viene estratta con probabilit'a

$$U_k = 1 - B_k - D_k \tag{6.3}$$

La costante c serve per regolare la frequenza relativa tra mosse che cambiano la dimensionalit'a (morte e nascita) e la mossa di aggiornamento.
La scelta delle probabilit'a di nascita e di morte garantisce che

$$B_k p(k|\lambda) = D_{k+1} p(k+1|\lambda) \tag{6.4}$$

La (6.4) ‘e una equazione di equilibrio bilanciato per il solo numero di termini. Questo vuol dire che se si avesse solo l'informazione del numero di termini (nessuna informazione sui coefficienti o sul tipo di termini,quindi nessuna idea sull'errore di regressione) la probabilit‘a del numero di termini convergerebbe alla distribuzione a priori. In realt‘a nelle prossime sezioni si aggiunger‘a un meccanismo di accettazione o rifiuto delle mosse che va di fatto ad alterare la probabilit‘a di regime del numero di termini in modo da tenere conto anche dell'errore di regressione (informazione delle misure) pi‘ uttosto che delle sole informazioni a priori.

6.0.3 Estrazione di una tipologia di mossa

L'estrazione della tipologia di mosse deve avvenire con le probabilit‘a B k , U k , D k calcolate come descritto nella sezione precedente. Un modo semplice per imporre tale probabilit‘a ‘e la seguente:

Estrai $z_k \in \mathcal{U}[0,1]$
if $z_k \leq B_k$ **then**
  Effettuare la mossa di nascita
else if $B_k < z_k \leq B_k + D_k$ **then**
  Effettuare la mossa di morte
else
  Effettuare la mossa di aggiornamento
end if

per mostrare che si ottengono effettivamente le probabilit‘a cercate basta integrare tra gli estremi opportuni la ddp della variabile uniforme che ‘e un l'impulso rettangolare

$$P\left(0 < z_k \leq B_k^{(i)}\right) = \int_0^{B_k^{(i)}} rect(z-0.5)dz = \int_0^{B_k^{(i)}} dz = B_k^{(i)} \tag{6.5}$$

$$P\left(B_k^{(i)} < z_k \leq B_k^{(i)} + D_k^{(i)}\right) = \int_{B_k^{(i)}}^{B_k^{(i)}+D_k^{(i)}} rect(z-0.5)dz = \int_0^{D_k^{(i)}} dz = D_k^{(i)} \tag{6.6}$$

$$P\left(B_k^{(i)} + D_k^{(i)} < z_k \leq 1\right) = \int_{B_k^{(i)}+D_k^{(i)}}^{1} rect(z-0.5)dz = \int_{B_k^{(i)}+D_k^{(i)}}^{1} dz = 1-B_k^{(i)}-D_k^{(i)} = U_k^{(i)} \tag{6.7}$$

con

$$rect(z) = \begin{cases} 0 & abs(z) \leq 0.5 \\ 1 & abs(z) > 0.5 \end{cases} \tag{6.8}$$

6.1 Mossa di nascita

Algorithm 3 Mossa di nacita

All'iterazione i estrai casualmente un termine di processo $p^{(i)}$ quelli non ancora selezionati

Calcola la quantità r_a
Calcola il rapporto di accettazione della mossa di nascita $\gamma_{birth}^{(k)}$
Estrai $z_b \in \mathcal{U}[0,1]$
if $z_b \geq \gamma_{birth}^{(k)}$ **then**
  k:=k+1
  $\mathcal{P}_k^{(i)} = \mathcal{P}_k^{(i-1)} \cup \{p^{(i)}\}$
  aggiornare i parametri con il valor medio della proposal $\mathbf{a}_k := \mu_{a,k'}$
else
  Aggiorna i parametri usando (l'algoritmo 3)
  Aggiorna la varianza σ_a^2
end if

Il termine da aggiungere nello sviluppo viene scelto casualmente (con probabilit‘a uniforme) tra i termini disponibili ovvero non gi‘a presenti nell'attuale modello Una volta scelto il termine candidato ‘e necessario decidere se accettare o meno la mossa di nascita alla maniera di RJMCMC, in modo da far convergere la catena all'equilibrio rappresentato dalla posterior dei modelli.
In particolare si accetta la mossa di nascita con probabilit‘a

$$\gamma_{birth}^{(k)} = min\{1, r_a\} \tag{6.9}$$

a tal fine viene estratto un numero da una distribuzione $z_b \in \mathcal{U}[0,1]$,si accetta la mossa se $z_b \geq \gamma_{birth}^{(k)}$ birth e si rifiuta se $z_b > \gamma_{birth}^{(k)}$. Il rapporto di accettazione r_a viene calcolato come

$$r_a = \frac{\hat{p}(x')g'(u')}{\hat{p}(x)g(u)} \left| \frac{\partial(\theta_{k'}, u')}{\partial(\theta_k, u)} \right| \tag{6.10}$$

si ha che $\hat{p}$(x) ‘e la distribuzione target della catena (la posterior) calcolata nel nuovo stato . Visto che il modello di rumore e gli iperparametri si tengono fissi durante la transizione si ha che

$$\hat{p}(x') = p(k', \mathbf{P}_{k'}, \mathbf{a}_{k'} | \mathbf{y}, \lambda_a, \sigma_a^2) \tag{6.11}$$

si noti che le variabili $k, \mathbf{P}_k$, identificano complessivamente una particolare struttura del modello che ‘e fissata dalla tipologia di mossa, non si ha quindi da effettuare parti- colari richieste per essi mentre bisogna imporre la condizione di *dimension matching* sui vettori dei parametri e delle variabili ausiliarie chiedendo che la trasformazione

$$(\mathbf{a}_k, u) \rightarrow (\mathbf{a}_{k'}, u') \tag{6.12}$$

sia un diffeomorfismo, si scelgono quindi le variabili ausiliarie in modo che valga.

$$dim(\mathbf{a}_k) + dim(u) = dim(\mathbf{a}_{k'}) + dim(u') \tag{6.13}$$

Una possibile scelta ‘e $u = \mathbf{a}_k$ e $u' = \mathbf{a}_{k'}$. Il jacobiano del cambio di variabili ‘e

$$\left| \frac{\partial(\mathbf{a}_k, \mathbf{a}_{k'})}{\partial(\mathbf{a}_{k'}, \mathbf{a}_k)} \right| = \left| \frac{\partial \begin{bmatrix} 0 & I_k \\ I_{k'} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{a}_{k'} \\ \mathbf{a}_k \end{bmatrix}}{\partial(\mathbf{a}_{k'}, \mathbf{a}_k)} \right| = \begin{vmatrix} 0 & I_k \\ I_{k'} & 0 \end{vmatrix} = |-1| = 1 \tag{6.14}$$

dunque

$$r_a = \frac{p(k', \mathbf{P}_{k'}, \mathbf{a}_{k'}|\mathbf{y}, \lambda_a, \sigma_a^2) p(\mathbf{a}_k|\mathbf{y}, k', \mathbf{P}_{k'}, \sigma_a^2)}{p(k, \mathbf{P}_k, \mathbf{a}_k|\mathbf{y}, \lambda_a, \sigma_a^2) p(\mathbf{a}_{k'}|\mathbf{y}, k', \mathbf{P}_{k'}, \sigma_a^2)} \tag{6.15}$$

Per semplificare r_a e evitare di estrarre un vettore $\mathbf{a}_k$ dalla distribuzione si pu'o far uso dell'equazione del candidato di Besag (Si veda in seguito).
Con l'equazione di Besag ed alcuni passaggi algebrici, l'espressione di r_a si semplifica

$$r_a = \frac{p(k', \mathbf{P}_{k'}|\mathbf{y}, \lambda_a, \sigma_a^2)}{p(k, \mathbf{P}_k|\mathbf{y}, \lambda_a, \sigma_a^2)} = \frac{\sigma_a^{-k'} \sqrt{det(C_{a,k'})} exp(\frac{1}{2} \mu_{a,k'}^T C_{a,k'}^{-1} \mu_{a,k'}) p(k'|\lambda_a)}{\sigma_a^{-k} \sqrt{det(C_{a,k'})} exp(\frac{1}{2} \mu_{a,k}^T C_{a,k}^{-1} \mu_{a,k}) p(k|\lambda_a)} \tag{6.16}$$

dove

$$C_{a,k} = \sigma_e^{-2} \mathbf{P}_k^t \mathbf{P}_k + \sigma_a^{-2} I_k \tag{6.17}$$

$$\mu_{a,k} = \sigma_e^{-2} C_{a,k} (\mathbf{P}_k^T (\mathbf{y} - \mathbf{E}_q \mathbf{b}_q) \tag{6.18}$$

Dunque se la mossa di nascita viene accettata, il termine proposto viene aggiunto allo sviluppo incrementando di conseguenza il conteggio del numero di termini. I coefficienti dello sviluppo vengono aggiornati con gli elementi del vettore valor medio calcolato mediante la [CITARE EQUAZIONE] Se invece la mossa viene rifiutata i termini dello sviluppo rimangono gli stessi mentre si aggiornano i coefficienti dello sviluppo e viene estratta una nuova varianza per i coefficienti.

6.1.1 Mossa di morte

Algorithm 3 Mossa di morte

All'iterazione i estrai casualmente un termine di processo $p^{(i)}$ da quelli attualmente selezionati
Calcola la quantità r_a
Calcola il rapporto di accettazione della mossa di morte $\gamma_{death}^{(k)}$
Estrai $z_d \in \mathcal{U}[0, 1]$
if $z_d \geq \gamma_{death}^{(k)}$ **then**
    k:=k-1
    $\mathcal{P}_k^{(i)} = \mathcal{P}_k^{(i-1)} - \{p^{(i)}\}$
    aggiornare i parametri con il valor medio della proposal $\mathbf{a}_k := \mu_{a,k'}$
else
    Aggiorna i parametri usando (l'algoritmo 3)
    Aggiorna la varianza σ_a^2
end if

Il termine da eliminare dallo sviluppo viene scelto casualmente (con probabilit'a uniforme) tra i termini gi'a presenti nell'attuale modello Una volta scelto il termine candidato 'e necessario decidere se accettare o meno la mossa di morte alla maniera di RJMCMC, in modo da far convergere la catena all'equilibrio rappresentato dalla posterior dei modelli In particolare si accetta la mossa di morte con probabilit'a

$$\gamma_{birth}^{(k)} = min \left\{ 1, \frac{r_a}{r_a} \right\} \tag{6.19}$$

Con r a ottenuto mediante l'equazione [CITARE EQUAZIONE] Per imporre la probabilit'a di accettazione viene estratto un numero da una distribuzione $z_d \in \mathcal{U}[0,1]$,si accetta la mossa se $z_d \geq \gamma_{death}^{(k)}$ birth e si rifiuta se $z_b > \gamma_{death}^{(k)}$. Dunque se la mossa di morte viene accettata, il termine proposto viene eliminato dallo sviluppo e reinserito tra i termini disponibili per le future mosse di nascita,viene decrementato di conseguenza il conteggio del numero di termini. I coefficienti dello sviluppo vengono aggiornati con gli elementi del vettore valor medio calcolato me- diante la [CITA]. Se invece la mossa viene rifiutata i termini dello sviluppo rimangono gli stessi mentre si aggiornano i coefficienti dello sviluppo e viene estratta una nuova varianza per i coefficienti.

6.1.2 Aggiornamento dei parametri

Algorithm 5 Aggiornamento dei parametri

for m=1:k **do**
Estrarre un candidato $\hat{a_m}^{(i)} \sim \mathcal{N}(a_m^{(i-1)}, C_{m,m})$ per $a_m^{(i)}$
Porre $a_{-m}^{(i)} = a_{-m}^{(i-1)}$
Calcolare la probabilità di accettazione

$$\alpha(\hat{a_m}^{(i)}|a_m^{(i-1)}) = \min\left\{1, \frac{p(\hat{a_m}^{(i)}|a_{-m}^{(i-1)}, \mathbf{y})q(a_{-m}^{(i-1)}|a^{(i)})}{p(a_m^{(i-1)}|a_{-m}^{(i-1)}, \mathbf{y})q(\hat{a_m}^{(i)}|a^{(i-1)}}\right\}$$

$z \sim \mathcal{U}[0,1]$
if $z \geq \alpha(\hat{a_m}^{(i)}|a_m^{(i-1)})$ **then**
$a_m^{(i)} = \hat{a}_m^{(i-1)}$
else$a_m^{(i)} = a_m^{(i-1)}$
end if
end for

L'aggiornamento dei parametri viene fatto campionando la ddp a posteriori dei parametri stessi

$$p(\mathbf{a}_k|\mathbf{y}, k, \mathbf{P}_k, \mathbf{E}_q, \sigma_a^2) \tag{6.20}$$

che usando Bayes pu'o essere espressa come

$$p(\mathbf{a}_k|\mathbf{y}, k, \mathbf{P}_k, eq, \sigma_a^2) \propto p(\mathbf{y}|\mathbf{P}_k, \mathbf{E}_q, \mathbf{a}_k, \mathbf{b}_q, \sigma_e^2)p(\mathbf{a}_k|k, \sigma_a^2) \propto exp\left\{-\frac{\epsilon^T\epsilon}{2\sigma_e^2}p(\mathbf{a}_k|k, \sigma_a^2)\right\} \tag{6.21}$$

con

$$\epsilon = \mathbf{y} - \mathbf{P}_k\mathbf{a}_k - \mathbf{E}_q\mathbf{b}_q \tag{6.22}$$

e

$$p(\mathbf{a}_k|k, \sigma_a^2) = \mathcal{N}\left(\mathbf{0}, \sigma_a^2\mathbb{I}\right) \tag{6.23}$$

I parametri sono aggiornati sequenzialmente usando un algoritmo di tipo MH che, invece di campionare la posterior multivriata la approssima campionando la proba- bilit'a dei singoli coefficienti condizionate rispetto agli altri. Dato per'o che anche le singole probabilit'a condizionate non sono semplici da campionare, si ricorre ad una densit'a di proposal scelta come gaussiana di varianza $C_{m,m}$ (l' m-esimo elemento della diagonale della matrice $C_{a,k}$) e valor medio il vecchio coefficiente $a_m^{(i-1)}$. Il coefficiente proposto viene accettato (sostituito al posto

del corrispondente vecchio coefficiente) con una probabilit‘a che ‘e calcolata alla maniera di MH.

$$\alpha(\hat{a}_m^{(i)}|a_m^{(i-1)}) = min\left\{1, \frac{p(\hat{a}_m^{(i)}|a_{-m}^{(i-1)}, \mathbf{y})q(a_{-m}^{(i-1)}|a^{(i)})}{p(a_{-m}^{(i-1)}|a_{-m}^{(i-1)}, \mathbf{y})q(\hat{a}_m^{(i)}|a^{(i)})}\right\} \tag{6.24}$$

Nella espressione sopra si ‘e omesso il simbolo k che indica la struttura di modello a cui si riferisce il vettore di coefficienti. Se non c'‘e il pedice si indica tutto il vettore dei coefficienti. Il simbolo di accento circonflesso indica l'elemento estratto dalla proposal. Il pedice m indica che si sta parlando dell'm-esimo elemento del vettore, il pedice -m indica il vettore di tutti i coefficienti escluso l'm-esimo. L'apice indica l'iterazione di RJMCMC in cui ‘e stato calcolato il termine: quelli che hanno apice (i-1) sono i vecchi coefficienti mentre quelli con apice (i) sono quelli calcolati nell'attuale iterazione. Con probabilit‘a $1 - \alpha(\hat{a}_m^{(i)}|a_m^{(i-1)}$ il coefficiente proposto viene rifiutato e rimane pari al vecchio valore.

6.2 Equazione del candidato di Besag

La formula utilizzata per semplificare l'espressione del rapporto r_a 'e detta formula del candidato. Il nome le 'e stato attribuito dal professore e statista Julian Ernst Besag a fine anni '80 quando era docente alla University of Durham. Egli riporta la formula in un articolo spiegando di averla letta (senza dimostrazione) nello svol- gimento di un esame da parte di uno studente del quale per'o non ricordava pi' u il nome. Non sapremo quindi mai chi fu il primo ad averla pensata. Nonostante anche Besag tralasci la dimostrazione della formula, questa non 'e complessa e deriva in sostanza dalla formula di Bayes applicata opportunamente alla densit'a congiunta delle tre variabili in gioco. Si supponga di aver raccolto un vettore di misurazioni x ,sia θ un vettore di parametri del modello, ci si chiede quale sia la densit'a di probabilit'a di una nuova misura z condizionata alle vecchie misure ovvero $p(z|x)$
tesi:

$$p(z|x) = \frac{p(z|\theta)p(\theta|x)}{p(\theta|z,x)} \tag{6.25}$$

dimostrazione: Dalla definizione di ddp condizionata

$$p(z|x) = \frac{p(z,x)}{p(x)} \tag{6.26}$$

Per sviluppare il numeratore 'e necessario prima di tutto dimostrare un passaggio intermedio Si consideri la tautologia

$$p(x,\theta,z) = p(z,\theta,z) \tag{6.27}$$

dove ciascun membro 'e la ddp congiunta delle tre variabili in esame. utilizzando la definizione di ddp condizionata

$$p(\theta|z,x)p(z,x) = p(z,x|\theta)p(\theta) \tag{6.28}$$

utilizzando l'indipendenza della nuova misura rispetto alle precedenti

$$p(\theta|z,x)p(z,x) = p(z|\theta)p(x|\theta)p(\theta) \tag{6.29}$$

da cui

$$p(z,x) = \frac{p(z|\theta)p(x|\theta)p(\theta)}{p(\theta|z,x)} \tag{6.30}$$

utilizzando Bayes

$$p(z|x) = \frac{p(z|\theta)\frac{p(\theta|x)p(x)}{p(\theta)}}{p(\theta|z,x)p(x)} \tag{6.31}$$

Da cui la tesi

$$p(z|x) = \frac{p(z|\theta)p(\theta|x)}{p(\theta|z,x)} \tag{6.32}$$

ponendo

$$z := (z, \mathbf{P}_k) \tag{6.33}$$

$$x := (y, \lambda_a, \sigma_a^2) \tag{6.34}$$

$$\theta := (\mathbf{a}_k) \tag{6.35}$$

si ha

$$p(k, \mathbf{P}_k|y, \lambda_a, \sigma_a^2) = \frac{p(k, \mathbf{P}_k|\mathbf{a}_k)p(\mathbf{a}_k|y, \lambda_a, \sigma_a^2)}{p(\mathbf{a}_k|k, \mathbf{P}_k, y\lambda_a, \sigma_a^2)} \tag{6.36}$$

usando la definizione di probabilit‘a condizionata si nota come il prodotto al numer- atore non sia altro che la probabilit‘a congiunta delle variabili k, $\mathbf{P}_k$ e $\mathbf{a}_k$, ottenendo quindi

$$p(k, \mathbf{P}_k|y, \lambda_a, \sigma_a^2) = \frac{p(k, \mathbf{P}_k, \mathbf{a}_k|y, \lambda_a, \sigma_a^2)}{p(\mathbf{a}_k|y, k', \mathbf{P}_{k'}, \sigma_a^2)} \tag{6.37}$$

Da cui la semplificazione per l'acceptance ratio

$$r_a = \frac{p(k', \mathbf{P}_{k'}|\mathbf{y}, \lambda_a, \sigma_a^2)}{p(k, \mathbf{P}_k|\mathbf{y}, \lambda_a, \sigma_a^2)} \tag{6.38}$$

. Dovrò quindi calcolare

$$p(k, \mathbf{P}_k|\mathbf{y}, \lambda_a, \sigma_a^2) \tag{6.39}$$

che può essere ottenuta come probabilità marginale

$$p(k, \mathbf{P}_k|\mathbf{y}, \lambda_a, \sigma_a^2) = \int p(k, \mathbf{P}_k, \mathbf{a}_k|y\lambda_a, \sigma_a^2)d\mathbf{a}_k \tag{6.40}$$

Usando Bayes l'integranda diventa