Università di Pisa

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE
Corso di Laurea in Ingegneria robotica e dell'automazione

Progetto per il corso di identificazione dei sistemi incerti

Identificazione di sistemi mediante modello NARMAX e approccio bayesiano

Candidato:
Giorgio Mirenda

Anno Accademico 2013-2014

Contents

Lista dei simboli

$\propto$ uguaglianza a meno di una costante di normalizzazione

Chapter 1

Introduzione

*T*he 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 (sopratutto se non lineare), una fase importante è stabilire la struttura del modello. Spesso infatti non sono incognite solo le costanti della legge che regola il processo ma non si conoscono neppure quali e quante operazioni effettuare sui segnali in gioco.
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. Quindi la sola identificazione parametrica è sufficiente 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 dell'accuratezza della descrizione ottenuta applicando i principi primi delle scienze (rischio di tralasciare fenomeni importanti)
- le leggi che ottengo applicando i principi primi delle scienze non sono eccessivamente complesse per lo scopo per cui servono

In tutti gli altri casi si ipotizza una classe (abbastanza vasta) di strutture di modello e si lascia al processo di identificazione l'onere di scegliere quella che più 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 .
La classe di modelli scelta è quella dei nonlineari, auto-regressivi, 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 non lineare 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 $\mathcal{M}$ 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 $m \in \mathcal{M}$ la probabilità a posteriori

$$P(M = m|Y)$$

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

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

la definizione corrisponde all'idea ragionevole

$$P(M = m|Y) \propto P(M = m, Y)$$

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

$$\int P(M = m|Y)dm = \int k \cdot P(M = m, Y)dm = k \cdot P(Y) := 1$$

da cui si ricava il valore di k

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

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

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

dunque mettendo insieme le equazioni (2.1) e la (2.2) si ottiene la formula di Bayes

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

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

$$P(M = m|Y) = \frac{P(Y|M = m)P(M = m)}{\int P(Y|M = m)P(M = m)dm} \tag{2.4}$$

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 Algoritmo Metropolis Hasting

Supponiamo di voler estrarre campioni da una distribuzione $\psi^*(x)$ nota, difficile da campionare direttamente.
Talvolta è possibile farlo mediante trasformazione deterministica di campioni estratti da una distribuzione più semplice come quella uniforme (si veda l'appendice 5.2).
Quando questo non è possibile si può 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 alla stessa distribuzione di regime indipendentemente dalla distribuzione di partenza.
Le condizioni necessarie per l'ergodicità della catena sono:

- **irriducibilità** : la probabilità di visitare ciascuno stato a partire da ciascuno stato è strettamente positiva

- **aperiodicità** : una catena è periodica se può ritornare in un certo stato solo a istanti multipli di un qualche intero maggiore di 1. Una catena è aperiodica se non è 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 $p(x'|x)$ (probabilità di transizione dal vecchio stato x al nuovo stato x') della catena, in modo che la catena sia ergodica e che la distribuzione di regime sia proprio $\psi^*(x)$. Una condizione sufficiente non necessaria è che la distribuzione obiettivo soddisfi la condizione di reversibilità

$$\psi^*(x)p(x'|x) = \psi^*(x')p(x|x') \tag{2.5}$$

Ricordando poi che una densità di probabilità è detta stazionaria se è autofunzione dell'operatore di transizione della catena

$$\psi^* = \psi^* P \tag{2.6}$$

$$\int \psi^*(x)p(x'|x)dx = \psi^*(x') \tag{2.7}$$

è semplice dimostrare che se vale la condizione di reversibilità allora la distribuzione è stazionaria infatti

$$\int \psi^*(x)p(x'|x)dx = \int \psi^*(x')p(x|x')dx = \psi^*(x') \int p(x|x')dx = \psi^*(x') \tag{2.8}$$

Se si sceglie un kernel arbitrario uguale ad una probabilità di transizione facile da campionare $p(x'|x) = s(x'|x)$ solitamente la condizione di reversibilità non è soddisfatta. Solitamente si ha uno sbilanciamento che significa che alcune transizioni sono più probabili in un certo verso

$$\psi^*(x)s(x'|x) > \psi^*(x')s(x|x') \tag{2.9}$$

In tal caso, viste le distribuzioni di partenza e la probabilità di transizione scelta è più 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 probabilità di transizione moltiplicata per un fattore correttivo $p(x'|x) = \gamma(x'|x)s(x'|x)$.
In particolare si può interpretare $\gamma(x'|x)$ come la probabilità di accettare la mossa $x \to x'$. Imponendo quindi che valga

$$\psi^*(x)\gamma(x'|x)s(x'|x) = \psi^*(x')\gamma(x|x')s(x|x') \tag{2.10}$$

si ricava

$$\frac{\gamma(x'|x)}{\gamma(x|x')} = \frac{\psi^*(x')s(x|x')}{\psi^*(x)s(x'|x)} \tag{2.11}$$

Perchè si abbia coerenza con la eq(2.9) bisogna che il rapporto al membro sinistro sia minore di 1. In particolare posso scegliere

- $\gamma(x|x') = 1$ che equivale ad accettare tutte le transizioni $x \to x'$ che sono meno frequenti
- $\gamma(x'|x) = min(1, \frac{\psi^*(x')s(x|x')}{\psi^*(x)s(x'|x)})$ fattore di riduzione delle transizioni più frequenti

L'algoritmo MH è quindi così definito

Algorithm 1 Algoritmo di Metropolis Hasting

```
inizializza X_0 = x_0;
for t = 0 ... T do
    Estraggo la proposta per il nuovo stato x' ~ s(x'|X_t)
    Calcolo la probabilità di accettare la mossa γ(x'|x) = min(1, ψ*(x')s(X_t|x') / ψ*(X_t)s(x'|X_t))
    Estraggo una variabile da una distribuzione uniforme u ~ U(0,1)
    if u ≤ γ(x'|X_t) then
        La catena transita nel nuovo stato propostoX_{t+1} = x'
    else
        La catena rimane nel vecchio stato X_{t+1} = X_t
    end if
end for
```

2.3 Reversible Jump Monte Carlo Markov Chain

Nelle sezioni precedenti la transizione della catena associava (in maniera aleatoria) uno stato di $X \subseteq \mathbb{R}^n$ ad uno stato di $X' \subseteq \mathbb{R}^n$.
Cosa succede se la transizione dovesse associare uno stato di $X \subseteq \mathbb{R}^n$ ad uno stato di $X' \subseteq \mathbb{R}^m$ con $m \neq n$?
La questione scaturisce dal fatto che per gli scopi dell'identificazione dei sistemi, restringersi al caso $m = n$ è 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 coefficienti dello sviluppo.
Quando l'algoritmo suggerisca di aggiungere o eliminare uno (o più) termini dello sviluppo è 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 dimensionalità ad un'altro con dimensionalità diversa.
Si pensi quindi di enumerare le tipologie di modello (anche infinite), si avrà che l'insieme delle possibili strutture di modello è rappresentabile come un insieme di indici

$$\mathcal{K} = \{1 \quad , 2, \quad , \ldots , \quad k, \quad , \ldots \} \subseteq \mathbb{N} \tag{2.12}$$

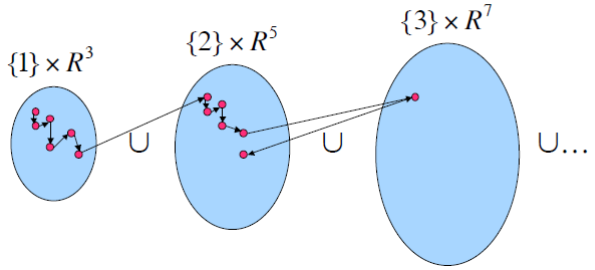
a ciascun modello è associato uno spazio dei coefficienti dello sviluppo (o in generale dei parametri del modello).
Si ha quindi che a ciascuna tipologia di modello è associato uno spazio $X_k \subseteq \mathbb{R}^{n(k)}$

dove $n(k)$ è il numero di termini previsti dal modello indicizzato da k.

Si può quindi pensare che la ricerca del modello avvenga in una unione di sottospazi definita da

$$X = \bigcup_{k \in K} (\{k\} \times X_k) \tag{2.13}$$

L'inclusione esplicita dell'indice k si ha perchè spesso è necessario che l'algoritmo vi acceda, specialmente quando esistono più modelli diversi a cui corrisponde lo stesso spazio dei parametri.
Lo stato della catena è 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 è 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 densità di regime ψ^*.



Cii limitiamo a considerare i casi in cui, per qualche distribuzione di probabilità di partenza la probabilità 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 è detta di *equilibrio bilanciato* , se verificata da una certa probabilità è una condizione sufficiente affinchè questa sia stazionaria ma non è necessaria, possono in generale esistere delle probabilità stazionarie che non verificano la condizione di equilibrio bilanciato ma per semplificare lo studio non analizzeremo questo caso generale.

Imporre l'equilibrio bilanciato è la chiave per imporre una specifica probabilità di regime.
Devono ancora valere le proprietà di irriducibilità e aperiodicità affinchè la probabilità di regime sia unica e sia raggiunta a partire da qualunque distribuzione iniziale. Tipicamente gli articoli che introducono la teoria dell'RJMCMC riportano il seguente protocollo (ideato da Peter Green nel 1995) che serve a fare in modo che la catena converga alla distribuzione desiderata nonostante i cambi di dimensionalità tra i sottospazi.
Spesso per attraversare lo spazio X è necessario avere a disposizione diverse tipologie di mosse.
Una tipologia di mossa associa la struttura del modello attuale ad una nuova struttura.

Affinchè la condizione di equilibrio bilanciato sia soddisfatta è necessario che se esiste la mossa $(x \to x')$ esista anche la mossa inversa $(x' \to x)$.
Si indichi con $j_m(x)$ la probabilità che dallo stato x sia scelta la mossa di tipo m e si indichi con $j_m^{-1}(x')$ la probabilità che dallo stato x' sia scelta la mossa di tipo m inversa.

Il problema è che non ha senso paragonare probabilità definite su spazi di dimensione diversa.
Per rendere *invisibile* alla catena il cambio di dimensionalità basta immergere i due spazi di dimensione diversa in uno stesso spazio di dimensione maggiore e definire in tale spazio la probabilità.

Si adotti quindi il seguente protocollo:
Supponiamo che la catena risieda attualmente nello stato x
Si estragga la tipologia si mossa.
Si generino r numeri casuali $u \in \mathbb{R}^r$ da una specificata densità g.
Si costruisca il nuovo stato attraverso una funzione deterministica h di x e di u.

$$(x', u') = h(x, u)$$

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

$$(x, u) = h'(x', u')$$

l'equazione di equilibrio bilanciato si può scrivere dunque come

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

La sommatoria sulle possibili mosse deriva dal fatto che ad ogni iterazione la tipologia di mossa è esclusiva. Una condizione sufficiente non necessaria perchè valga il bilancio è che esso valga mossa per mossa ovvero

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

se inoltre la funzione h è un diffeomorfismo vale la formula classica del cambio di variabili e quindi si può passare dall'uguaglianza integrale all'uguaglianza delle integrande.
Dal momento che la mossa è 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

$$j_m(x)\psi^*(x)g(u)\gamma(x,x') = j_m^{-1}(x')\psi^*(x')g(u')\gamma(x',x')\left|\frac{\partial(\theta_{k'},u')}{\partial(\theta_k,u)}\right|$$

Condizione necessaria affinchè h sia un diffeomorfismo è che

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

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

$$\gamma(x, x') = min\left\{1, \frac{j_m^{-1}(x')\psi^*(x')g(u')}{j_m(x)\psi^*(x)g(u)} \left|\frac{\partial(\theta_{k'}, u')}{\partial(\theta_k, u)}\right|\right\}$$

Chapter 3

Il modello

3.1 Modello NARMAX

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

$$y_t = f(y_{t-1}, \ldots, y_{t-n_y}, u_{t-1}, \ldots, u_{t-n_u}, y_{t-n_y}, e_{t-1}, \ldots, e_{t-n_e}) + e_t \tag{3.1}$$

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 da $\mathcal{N}(0, \sigma_e^2)$.
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 3.1 come combiaione di termini di processo e di rumore,

$$y_t = \sum_{j=1}^{M_p} (a_j y_{t-\delta_{y,j}}^{k_{y,j}} u_{t-\delta_{u,j}}^{k_{u,j}}) + \sum_{j=1}^{M_e} b_j (y_{t-\delta_{y,j}}^{k_{y,j}} u_{t-\delta_{u,j}}^{k_{u,j}} e_{t-\delta_{e,j}}^{k_{e,j}}) + e_t \tag{3.2}$$

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

3.2 Equazione di regressione

Data una sequenza di N ingressi

$$\{u(1) \quad u(2) \quad u(3) \quad u(4) \ldots u(N)\} \tag{3.3}$$

e N uscite

$$\{y(1) \quad y(2) \quad y(3) \quad y(4) \quad \ldots y(N)\} \tag{3.4}$$

si hanno a disposizione $N - \max_j\{n_u, n_y, n_e\}$ equazioni .

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

$$\mathcal{P} = \{y^2(t-1), y(t-2)u(t)\} \tag{3.5}$$

e l'insieme dei termini di rumore

$$\mathcal{E} = \{e^3(t-1)u(t-1)\} \tag{3.6}$$

e si avessero a disposizione gli ingressi

$$u = \{u(1) \quad u(2) \quad u(3) \quad u(4) \quad u(5) \quad u(6)\} \tag{3.7}$$

e le uscite

$$u = \{y(1) \quad y(2) \quad y(3) \quad y(4) \quad y(5) \quad y(6)\} \tag{3.8}$$

Si avrebbero a disposizione le equazioni di regressione

$$\begin{bmatrix} y(6) \\ y(5) \\ y(4) \\ y(3) \end{bmatrix} = \begin{bmatrix} y^2(5) & y(4)u(6) \\ y^2(4) & y(3)u(5) \\ y^2(3) & y(2)u(4) \\ y^2(2) & y(1)u(3) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} + \begin{bmatrix} e^3(5)u(5) \\ e^3(4)u(4) \\ e^3(3)u(3) \\ e^3(2)u(2) \end{bmatrix} b_1 + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{bmatrix} \tag{3.9}$$

che possono essere scritte compattamente in forma matriciale come

$$\mathbf{y} = \mathbf{P}_k\mathbf{a}_k + \mathbf{E}_q\mathbf{b}_q + \epsilon \tag{3.10}$$

3.3 Densità di probabilità a posteriori del modello

Nello specifico del modello NARMAX la formula di Bayes diventa

$$p(k, q, \mathbf{P}_k, \mathbf{E}_q, \mathbf{a}_k, \mathbf{b}_q|\mathbf{y}) \propto p(y|k, q, \mathbf{P}_k, \mathbf{E}_q, \mathbf{a}_k, \mathbf{b}_q, \sigma_e^2)p(k, q, \mathbf{P}_k, \mathbf{E}_q\mathbf{a}_k, \mathbf{b}_q) \tag{3.11}$$

Di seguito si specificano le espressioni per i fattori del membro a destra

3.3.1 Likehood

Se il rumore è gaussiano a media nulla anche la verosimiglianza è una gaussiana, funzione del vettore dei residui, di stessa varianza. Infatti come si deduce dall'equazione di regressione (3.10) si ha

$$p(y|k, q, \mathbf{P}_k, \mathbf{E}_q, \mathbf{a}_k, \mathbf{b}_q, \sigma_e^2) = p(\mathbf{P}_k\mathbf{a}_k + \mathbf{E}_q\mathbf{b}_q + \epsilon|k, q, \mathbf{P}_k, \mathbf{E}_q, \mathbf{a}_k, \mathbf{b}_q, \sigma_e^2) = p(\epsilon|k, q, \mathbf{P}_k, \mathbf{E}_q, \mathbf{a}_k, \mathbf{b}_q, \sigma_e^2) \tag{3.12}$$

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

$$p(\mathbf{e}|k, q, \mathbf{P}_k, \mathbf{E}_q, \mathbf{a}_k, \mathbf{b}_q, \sigma_e^2) = p(\epsilon|k, q, \mathbf{P}_k, \mathbf{E}_q, \mathbf{a}_k, \mathbf{b}_q, \sigma_e^2) = \frac{1}{\sqrt{2\pi\sigma_e^2}^N} exp\left(-\frac{1}{2\sigma_e^2}\epsilon^T\epsilon\right) \tag{3.13}$$

dunque in definitiva

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

con

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

3.3.2 Scelta delle distribuzioni a priori e iperparametri

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

3.3.3 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 più alto $p(X|\zeta)$ rappresenta la probabilità a priori (rispetto alle misure) della variabile X .
L'ipotesi sul'andamento della funzione f si suppone quindi già 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 è quindi indipendente dalla scelta del posterior $p(\zeta|X)$, infatti

$$p(\zeta|X) \propto p(X|\zeta)p(\zeta) \tag{3.14}$$

La distribuzione $p(X|\zeta)$ è quindi un operatore che mappa una funzione in un'altra funzione. E' possibile per tale operatore trovare una sorta di *autofunzione*, nel senso più lato di distribuzione appartenente a una certa famiglia (normale, uniforme, poisson 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 *distribuzioni 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 è finito e ridotto. (Ad esempio l'evoluzione di una gaussiana può essere rappresentata concisamente dall'evoluzione del suo valor medio e della sua varianza)

Sei 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.

3.3.4 Probabilità del modello a priori

Modificando leggermente la (3.11) per tenere conto anche degli iperparametri (destritti nel dettaglio in questa sezione) si ha

$$p(k, q, \mathbf{P}_k, \mathbf{E}_q, \mathbf{a}_k, \mathbf{b}_q, \sigma_a^2, \sigma_b^2, \lambda_a, \lambda_b|\mathbf{y}) \propto p(y|k, q, \mathbf{P}_k, \mathbf{E}_q, \mathbf{a}_k, \mathbf{b}_q, \sigma_e^2) p(k, q, \mathbf{P}_k, \mathbf{E}_q \mathbf{a}_k, \mathbf{b}_q, \sigma_a^2, \sigma_b^2, \lambda_a, \lambda_b) \tag{3.15}$$

fattorizzabile come

$$\begin{aligned} p(k, q, \mathbf{P}_k, \mathbf{E}_q, \mathbf{a}_k, \lambda_a, \sigma_a^2, \mathbf{b}_q, \lambda_b, \sigma_b^2) = p(\mathbf{a}_k|\mathbf{P}_k, k, \sigma_a^2) p(k|\lambda_a) p(\mathbf{P}_k) p(\lambda_a) p(\sigma_a^2) \times \\ p(\mathbf{b}_q|\mathbf{E}_q, q, \sigma_b^2) p(q|\lambda_b) p(\mathbf{E}_q) p(\lambda_b) p(\sigma_b^2) \end{aligned} \tag{3.16}$$

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 a media nulla e covarianza isotropica ovvero

$$p(\mathbf{a}_k|k, \mathbf{P}_k, \sigma_a^2) = p(\mathbf{a}_k|k, \sigma_a^2) \sim \mathcal{N}(\mathbf{0}, \sigma_a^2 \mathbf{I}) \tag{3.17}$$

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

$$\mathcal{IG}(x|, \alpha, \beta) \sim x^{-\alpha-1} exp\left(-\frac{\beta}{x}\right) \tag{3.18}$$

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

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

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

□

Vettore dei parametri dei termini di rumore

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

$$p(\mathbf{b}_q|k, \mathbf{E}_q, \sigma_b^2) = p(\mathbf{b}_q|q, \sigma_b^2) \sim \mathcal{N}(\mathbf{0}, \sigma_b^2\mathbf{I}) \tag{3.19}$$

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.

$$p(\sigma_b^2) = \mathcal{IG}(\sigma_b^2|\alpha_b, \beta_b) \Rightarrow \tag{3.20}$$

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

Numero di termini di processo

La probabilità a priori del numero di termini di processo è rappresentata da una distribuzione di Poisson troncata, il parametro λ_a.

$$p(k|\lambda_a) = \frac{\frac{\lambda_a^k}{k!}}{\sum_{i=0}^{N} \frac{\lambda_a^i}{i!}} \tag{3.21}$$

Con N numero massimo di termini di processo.
Questo tipo di distribuzione è più indicato rispetto ad una distribuzione uniforme, perchè l'iperparametro λ_a può essere interpretato come il numero medio di termini ipotizzando un numero di termini possibili sufficientemente alto infatti

$$\mathbb{E}[k|\lambda_a] = \frac{\sum_{k=0}^{N} k\frac{\lambda_a^k}{k!}}{\sum_{i=0}^{N} \frac{\lambda_a^i}{i!}} = \frac{\sum_{k=1}^{N} k\frac{\lambda_a^k}{k!}}{\sum_{i=0}^{N} \frac{\lambda_a^i}{i!}} = \lambda_a \frac{\sum_{j=0}^{N-1} \frac{\lambda_a^j}{j!}}{\sum_{i=0}^{N} \frac{\lambda_a^i}{i!}} = \lambda_a \frac{\sum_{j=0}^{N-1} \frac{\lambda_a^j}{j!}}{\frac{\lambda_a^N}{N!} + \sum_{i=0}^{N-1} \frac{\lambda_a^i}{i!}} = \lambda_a \frac{1}{1 + \frac{\frac{\lambda_a^N}{N!}}{\sum_{i=0}^{N-1} \frac{\lambda_a^i}{i!}}} \tag{3.22}$$

si noti che

$$\lim_{N\to\infty} \frac{\frac{\lambda_a^N}{N!}}{\sum_{i=0}^{N-1} \frac{\lambda_a^i}{i!}} = 0 \tag{3.23}$$

L'iperparametro λ_a è estratto da una distribuzione gamma

$$p(\lambda_a) = \mathcal{GA}(\lambda_a|\alpha_{\lambda_a}, \beta_{\lambda_b}) = \lambda_a^{(\alpha_{\lambda_a}-1)} exp\left(-\frac{\lambda_a}{\beta_{\lambda_a}}\right) \tag{3.24}$$

Di seguito dimostro che la distribuzione gamma è coniugata per il parametro della poisson

Proof.

$$p(\lambda_a|k) \propto p(k|\lambda_a)p(\lambda_a) \tag{3.25}$$

$$\propto \frac{\lambda_a^k}{k!}\lambda_a^{(\alpha_{\lambda_a}-1)}exp\left(-\frac{\lambda_a}{\beta_{\lambda_a}}\right)$$

$$\propto \frac{1}{k!}\mathcal{GA}(\lambda_a|\alpha_{\lambda_a}+k,\beta_{\lambda_b}) \tag{3.26}$$

□

Numero di termini di rumore

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

$$p(k|\lambda_a) = \frac{\frac{\lambda_b^q}{q!}}{\sum_{i=0}^{N}\frac{\lambda_b^i}{i!}} \tag{3.27}$$

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

$$p(\lambda_b) = \mathcal{GA}(\lambda_b|\alpha_{\lambda_b},\beta_{\lambda_b}) = \lambda_b^{(\alpha_{\lambda_b}-1)}exp\left(-\frac{\lambda_b}{\beta_{\lambda_b}}\right) \tag{3.28}$$

3.3.5 Matrice di regressione dei termini di processo e di rumore

La matrice di regressione dei termini è considerata una va aleatoria uniformemente distribuita in modo che nessun termine di modello nè di rumore sia, a priori delle misure, più probabile di altri

$$p(\mathbf{P}_k) \propto 1 \tag{3.29}$$

$$p(\mathbf{E}_q) \propto 1 \tag{3.30}$$

Chapter 4

Algoritmo RJMCMC per l'identificazione di modelli NARMAX

Algorithm 2: RJMCMC for NARMAX identification

---

fissa i parametri di tuning c e σ_e^2
inizializza $(k^{(0)}, q^{(0)}, \mathbf{P}_k^{(0)}, \mathbf{E}_q^{(0)}, \mathbf{a}_k^{(0)}, \mathbf{b}_q^{(0)}, \lambda_a^{(0)}, \lambda_b^{(0)}, \sigma_a^{2(0)}, \sigma_b^{2(0)})$

for $i = 1 : N_{iter}$ **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 $\lambda_a^{(i)} \sim p(\lambda_a | k^{(i)})$

if $i > 2$ **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 $(z_q \leq B_q^{(i)} + D_q^{(i)})$ **then**
Effettua mossa di morte (Algoritmo 2)
else
Aggiorna i parametri (Algoritmo 3)

Aggiorna la varianza dei parametri
end if
Estrai $\lambda_b^{(i)} \sim p(\lambda_b|q^{(i)})$
end if
Calcola l'errore residuo $\epsilon^{(i)} = \mathbf{y} - \mathbf{P}_k^{(i)}\mathbf{a}_k^{(i)} - \mathbf{E}_q^{(i)}\mathbf{b}_q^{(i)}$
Aggiornamento di $\mathbf{E}_q^{(i)}$
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 è lo scalare nullo. Gli iperparametri delle poisson troncate vengono inizializzati a quello che ci si aspetta essere il numero 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 è critica perchè essi verranno aggiornati con l'evoluzione della catena e si adatteranno al valore più appropriato. La varianza σ_e^2 del rumore bianco invece viene inizializzata e non viene più aggiornata, tale parametro rappresenta infatti un parametro di tuning dell'algoritmo essendo in qualche modo una metrica di affidabilità delle misure, è sensato dunque che esso influisca direttamente sulla probabilità di accettazione delle mosse: abbassare tale parametro equivale a ritenere l'incertezza bassa quindi il rapporto di accettazione sarà più selettivo e le mosse proposte verranno scartate più frequentemente.
L'algoritmo prevede un numero fissato a priori di iterazioni. In corrispondenza di ciascuna iterazione la catena di markov andrà a risiedere in uno stato che rappresenta un modello del sistema. L'assunzione di **ergodicità** della catena permette di ricostruire la probabilità 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 però adottare due accorgimenti:

- il numero di iterazioni deve essere abbastanza elevato da fare in modo che valga con buona approssimazione l'ipotesi di ergodicità. Se le statistiche si raccolgono su poche iterazioni le probabilità 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 perchè 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 rumore 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.

4.1 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 dimensionalità 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

$$B_k = \begin{cases} 1 & k = 0 \\ 0 & k = M_p \\ c \cdot min\left\{1, \frac{p(k+1|\lambda_a)}{p(k|\lambda_a)}\right\} & altrimenti \end{cases} \tag{4.1}$$

La mossa di morte viene estratta con probabilità

$$D_k = \begin{cases} 0 & k = 0 \\ c \cdot min\left\{1, \frac{p(k-1|\lambda_a)}{p(k|\lambda_a)}\right\} & altrimenti \end{cases} \tag{4.2}$$

La mossa di aggiornamento viene estratta con probabilità

$$U_k = 1 - B_k - D_k \tag{4.3}$$

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

$$B_k p(k|\lambda) = D_{k+1} p(k+1|\lambda) \tag{4.4}$$

La (4.4) è 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 probabilità del numero di termini convergerebbe alla distribuzione a priori. In realtà nelle prossime sezioni si aggiungerà un meccanismo di accettazione o rifiuto delle mosse che va di fatto ad alterare la probabilità di regime del numero di termini in modo da tenere conto anche dell'errore di regressione (informazione delle misure) piùttosto che delle sole informazioni a priori.

4.2 Estrazione di una tipologia di mossa

L'estrazione della tipologia di mosse deve avvenire con le probabilità B_k, U_k, D_k calcolate come descritto nella sezione precedente. Un modo semplice per imporre tale probabilità è 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 probabilità cercate basta integrare tra gli estremi opportuni la ddp della variabile uniforme che è un l'impulso rettangolare

$$P\left(0 < z_k \leq B_k^{(i)}\right) = \int_0^{B_k^{(i)}} rect(z-0.5)dz = \int_0^{B_k^{(i)}} dx = B_k^{(i)} \tag{4.5}$$

$$P\left(B_k^{(i)} < z_k \leq B_k^{(i)} + D_k^{(i)}\right) = \int_{B_k^{(i)}}^{B_k^{(i)}+D_k^{(i)}} rect(z-0.5)dz = \int_0^{D_k^{(i)}} dx = D_k^{(i)} \tag{4.6}$$

$$P\left(B_k^{(i)} + D_k^{(i)} < z_k \leq 1\right) = \int_{B_k^{(i)}+D_k^{(i)}}^{1} rect(z-0.5)dz = \int_{B_k^{(i)}+D_k^{(i)}}^{1} dx = 1 - B_k^{(i)} - D_k^{(i)} = U_k^{(i)} \tag{4.7}$$

con

$$rect(z) = \begin{cases} 0 & abs(z) \leq 0.5 \\ 1 & abs(z) > 0.5 \end{cases} \tag{4.8}$$

4.3 Mossa di nascita

Algorithm 3: Mossa di nascita

1: All'iterazione i estrai casualmente un termine $p^(i)$ da quelli non ancora selezionati
2: Calcola la quantità r_a
3: Calcola il rapporto di accettazione della mossa di nascita $\gamma_{birth}^{(k)}$
4: Estrai $z_b \sim \mathcal{U}[0,1]$
5: **if** $(z_b \leq \gamma_{birth}^{(k)})$ **then**
6: $k := k + 1$
7: $\mathcal{P}_k^{(i)} = \{\mathcal{P}_k^{(i-1)} \cup p^{(i)}\}$
8: aggiornare i parametri con il valor medio della proposal $\mathbf{a}_k := \mu_{a,k'}$
9: **else**
10: Aggiorna i parametri usando (algoritmo 3)
11: Aggiorna la varianza σ_a^2
12: **end if**

Il termine da aggiungere nello sviluppo viene scelto casualmente (con probabilità uniforme) tra i termini disponibili ovvero non già presenti nell'attuale modello
Una volta scelto il termine candidato è 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 probabilità

$$\gamma_{birth}^{(k)} = \min\{1, r_a\} \tag{4.9}$$

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

$$r_a = \frac{\psi^*(x')g'(u')}{\psi^*(x)g(u)} \left| \frac{\partial(\theta_{k'}, u')}{\partial(\theta_k, u)} \right| \tag{4.10}$$

si ha che $\psi^*(x')$ è 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

$$\psi^*(x') = p(k', \mathbf{P}_{k'}, \mathbf{a}_{k'} | \mathbf{y}, \lambda_a, \sigma_a^2) \tag{4.11}$$

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

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

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

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

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

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

dunque

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

Per semplificare r_a e evitare di estrarre un vettore $\mathbf{a}_k$ dalla distribuzione si può 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

$$r_a = \frac{p(k', \mathbf{P}_{k'}|y, \lambda_a, \sigma_a^2)}{p(k, \mathbf{P}_k|y, \lambda_a, \sigma_a^2)} = \frac{\sigma_a^{-k'} \sqrt{det(C_{a,k'})} exp(\frac{1}{2}\mu_{a,k'}^T C_{a,k'}^{-1} \mu_{a,k'}) p(k'|\lambda_A)}{\sigma_a^{-k} \sqrt{det(C_{a,k})} exp(\frac{1}{2}\mu_{a,k}^T C_{a,k}^{-1} \mu_{a,k}) p(k|\lambda_A)} \tag{4.12}$$

dove

$$C_{a,k} = \sigma_e^{-2} \mathbf{P}_k^T \mathbf{P}_k + \sigma_a^{-2} I_k \tag{4.13}$$

$$\mu_{a,k} = \sigma_e^{-2} C_{a,k} \mathbf{P}_k^T (\mathbf{y} - \mathbf{E}_q \mathbf{b}_q) \tag{4.14}$$

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 (4.14).
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.

4.3.1 Mossa di morte

Algorithm 4: Mossa di morte

1: All'iterazione i estrai casualmente un termine $p^{(i)}$ da quelli attualmente selezionati
2: Calcola la quantità r_a

3: Calcola il rapporto di accettazione della mossa di morte $\gamma_{death}^{(k)}$
4: Estrai $z_d \sim \mathcal{U}[0,1]$
5: **if** $(z_d \leq \gamma_{death}^{(k)})$ **then**
6: $k := k - 1$
7: $\mathcal{P}_k^{(i)} = \mathcal{P}_k^{(i-1)} / \{p^{(i)}\}$
8: aggiornare i parametri con il valor medio della proposal $\mathbf{a}_k := \mu_{a,k'}$
9: **else**
10: Aggiorna i parametri usando (algoritmo 3)
11: Aggiorna la varianza σ_a^2
12: **end if**

Il termine da eliminare dallo sviluppo viene scelto casualmente (con probabilità uniforme) tra i termini già presenti nell'attuale modello
Una volta scelto il termine candidato è 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 probabilità

$$\gamma_{death}^{(k)} = \min\left\{1, \frac{1}{r_a}\right\} \tag{4.15}$$

Con r_a ottenuto mediante l'equazione (4.12)
Per imporre la probabilità di accettazione viene estratto un numero da una distribuzione $z_d \sim \mathcal{U}[0,1]$, si accetta la mossa se $z_d \leq \gamma_{death}^{(k)}$ e si rifiuta se $z_d > \gamma_{death}^{(k)}$.
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 mediante la (4.14).
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.

4.4 Aggiornamento dei parametri

Algorithm 5: Aggiornamento dei parametri

1: **for** m=1:k **do**
2: Estrarre un candidato $\hat{a}_m^{(i)} \sim \mathcal{N}(a_m^{(i-1)}, C_{m,m})$ per $a_m^{(i)}$
3: Porre $a_{-m}^{(i)} = a_{-m}^{(i-1)}$

4: Calcolare la probabilità di accettazione

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

5: Estrai $z \sim \mathcal{U}[0,1]$
6: **if** $z \leq \alpha(\hat{a}_m^{(i)}|a_m^{(i-1)})$ **then**
7: $a_m^{(i)} = \hat{a}_m^{(i)}$
8: **else**
9: $a_m^{(i)} = a_m^{(i-1)}$
10: **end if**
11: **end for**

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

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

che usando Bayes può essere espressa come

$$p(\mathbf{a}_k|\mathbf{y}, k, \mathbf{P}_k, \mathbf{E}_q, \sigma_a^2) \propto p(\mathbf{y}|\mathbf{P}_k, \mathbf{E}_q, \mathbf{a}_k, \mathbf{b}_q, \sigma_e^2)p(\mathbf{a}_k|k, \sigma_a^2) \propto \prod_t exp\left(\frac{\epsilon_t^2}{2\sigma_e^2}\right) p(\mathbf{a}_k|k, \sigma_a^2)$$

con

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

e

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

I parametri sono aggiornati sequenzialmente usando un algoritmo di tipo MH che, invece di campionare la posterior multivriata la approssima campionando la probabilità dei singoli coefficienti condizionate rispetto agli altri. Dato però che anche le singole probabilità condizionate non sono semplici da campionare, si ricorre ad una densità 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 probabilità che è calcolata alla maniera di MH.

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

Nella espressione sopra si è omesso il simbolo k che indica la struttura di modello a cui si riferisce il vettore di coefficienti. Se non c'è 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 è 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 probabilità $1 - \alpha(\hat{a}_m^{(i)}|a_m^{(i-1)})$ il coefficiente proposto viene rifiutato e rimane pari al vecchio valore.

4.4.1 Equazione del candidato di Besag

La formula utilizzata per semplificare l'espressione del rapporto r_a è detta *formula del candidato*. Il nome le è 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 svolgimento di un esame da parte di uno studente del quale però non ricordava più il nome. Non sapremo quindi mai chi fu il primo ad averla pensata. Nonostante anche Besag tralasci la dimostrazione della formula, questa non è complessa e deriva in sostanza dalla formula di Bayes applicata opportunamente alla densità congiunta delle tre variabili in gioco.
Si supponga di aver raccolto un vettore di misurazioni $\mathbf{x}$,sia θ un vettore di parametri del modello, ci si chiede quale sia la densità di probabilità di una nuova misura z condizionata alle vecchie misure ovvero $p(z|\mathbf{x})$
tesi:

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

dimostrazione: Dalla definizione di ddp condizionata

$$p(z|\mathbf{x}) = \frac{p(z,x)}{p(x)} \tag{4.16}$$

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

$$p(z,x,\theta) = p(z,x,\theta)$$

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

$$p(\theta|z,x)p(z,x) = p(z,x|\theta))p(\theta)$$

utilizzando l'indipendenza della nuova misura rispetto alle precedenti

$$p(\theta|z,x)p(z,x) = p(z|\theta)p(x|\theta)p(\theta)$$

da cui

$$p(z,x) = \frac{p(z|\theta)p(x|\theta)p(\theta)}{p(\theta|z,x)} \tag{4.17}$$

Dunque la 4.16 diventa

$$p(z|x) = \frac{p(z|\theta)p(x|\theta)p(\theta)}{p(\theta|z,x)p(x)} \tag{4.18}$$

utilizzando Bayes

$$p(z|x) = \frac{p(z|\theta)\frac{p(\theta|x)p(x)}{p(\theta)}p(\theta)}{p(\theta|z,x)p(x)} \tag{4.19}$$

Da cui la tesi

$$p(z|x) = \frac{p(z|\theta)p(\theta|x)}{p(\theta|z,x)} \tag{4.20}$$

ponendo

$$z := (k, \mathbf{P}_k)$$
$$x := (y, \lambda_a, \sigma_a^2)$$
$$\theta := (\mathbf{a}_k)$$

si ha

$$p(k, \mathbf{P}_k|y, \lambda_a, \sigma_a^2) = \frac{p(k, \mathbf{P}_k|\mathbf{a}_k)p(\mathbf{a}_k|y, \lambda_a, \sigma_a^2)}{p(\mathbf{a}_k|k, \mathbf{P}_k, y, \lambda_a, \sigma_a^2)} \tag{4.21}$$

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

$$p(k, \mathbf{P}_k|y, \lambda_a, \sigma_a^2) = \frac{p(k, \mathbf{P}_k, \mathbf{a}_k|y, \lambda_a, \sigma_a^2)}{p(\mathbf{a}_k|y, k', \mathbf{P}_{k'}, \sigma_a^2)} \tag{4.22}$$

Da cui la semplificazione per l'acceptance ratio

$$ra = \frac{p(k', \mathbf{P}_{k'}|y, \lambda_a, \sigma_a^2)}{p(k, \mathbf{P}_k|y, \lambda_a, \sigma_a^2)} \tag{4.23}$$

dunque sono interessato a calcolare

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

che può essere ottenuta come probabilità marginale

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

Usando Bayes l'integranda diventa

$$\begin{aligned}
&\frac{1}{p(y)} p(y|k, \mathbf{P}_k, \mathbf{a}_k, \lambda_A, \sigma_a^2, q, \mathbf{E}_q, \lambda_B, \mathbf{b}_q, \sigma_e^2) p(k, \mathbf{P}_k, \mathbf{a}_k|\lambda_a, \sigma_a^2) \\
=&\frac{1}{p(y)} p(k|\lambda_A) p(\mathbf{P}_k) p(y|k, \mathbf{P}_k, \mathbf{a}_k, \lambda_A, \sigma_a^2, q, \mathbf{E}_q, \lambda_B, \mathbf{b}_q, \sigma_e^2) p(\mathbf{a}_k|k, \sigma_a^2) \\
=&\frac{1}{p(y)} p(k|\lambda_A) p(\mathbf{P}_k) \mathcal{N}(\mathbf{P}_k \mathbf{a}_k + \mathbf{E}_q \mathbf{b}_q, \sigma_e^2 I_k) \mathcal{N}(0, k, \sigma_a^2 I_k) \\
=&\mathcal{K} exp[\frac{1}{2} \sigma_e^{-2} (2\mathbf{y}^T \mathbf{P}_k \mathbf{a}_k + 2\mathbf{y}^T \mathbf{E}_q \mathbf{b}_q - \mathbf{b}_q^T \mathbf{E}_q^T \mathbf{E}_q \mathbf{b}_q - 2\mathbf{a}_k^T \mathbf{P}_k^T \mathbf{E}_q \mathbf{b}_q - \mathbf{y}^T \mathbf{y}) \\
&- \frac{1}{2} \mathbf{a}_k^T (\sigma_e^{-2} \mathbf{P}_k^T \mathbf{P}_k + \sigma_a^{-2} I_k) \mathbf{a}_k]
\end{aligned}$$

$$\mathcal{K} = \frac{1}{p(y)} \frac{\sigma_e^{-N}}{\sqrt{2\pi}^N} \frac{\sigma_a^{-k}}{\sqrt{2\pi}^k} p(k|\lambda_A) p(\mathbf{P}_k)$$

Si moltiplichi e si divida per $\mathcal{N}(\mathbf{a}_k|\mu, C)$

$$\sqrt{(2\pi)^k det(C)} exp[\frac{1}{2}\sigma_e^{-2}(2\mathbf{y}^T\mathbf{P}_k\mathbf{a}_k + 2\mathbf{y}^T\mathbf{E}_q\mathbf{b}_q - \mathbf{b}_q^T\mathbf{E}_q^T\mathbf{E}_q\mathbf{b}_q - 2\mathbf{a}_k^T\mathbf{P}_k^T\mathbf{E}_q\mathbf{b}_q - \mathbf{y}^T\mathbf{y}) -$$
$$\frac{1}{2}\mathbf{a}_k^T(\sigma_e^{-2}\mathbf{P}_k^T\mathbf{P}_k + \sigma_a^{-2}I_k)\mathbf{a}_k + \frac{1}{2}(\mathbf{a}_k - \mu)^T C^{-1}(\mathbf{a}_k - \mu)]\mathcal{N}(\mathbf{a}_k|\mu, C)d\mathbf{a}_k$$
$$= \sqrt{(2\pi)^k det(C)} exp[\frac{1}{2}\sigma_e^{-2}(2\mathbf{y}^T\mathbf{P}_k\mathbf{a}_k + 2\mathbf{y}^T\mathbf{E}_q\mathbf{b}_q - \mathbf{b}_q^T\mathbf{E}_q^T\mathbf{E}_q\mathbf{b}_q - 2\mathbf{a}_k^T\mathbf{P}_k^T\mathbf{E}_q\mathbf{b}_q - \mathbf{y}^T\mathbf{y}) -$$
$$\frac{1}{2}\mathbf{a}_k^T(\sigma_e^{-2}\mathbf{P}_k^T\mathbf{P}_k + \sigma_a^{-2}I_k)\mathbf{a}_k + \frac{1}{2}(\mathbf{a}_k^T C^{-1}\mathbf{a}_k - 2\mathbf{a}_k^T C^{-1}\mu - \mu^T C^{-1}\mu)]\mathcal{N}(\mathbf{a}_k|\mu, C)d\mathbf{a}_k$$

raccogliendo i termini simili

$$= \sqrt{(2\pi)^k det(C)} exp[\frac{1}{2}\sigma_e^{-2}(2\mathbf{y}^T\mathbf{E}_q\mathbf{b}_q - \mathbf{b}_q^T\mathbf{E}_q^T\mathbf{E}_q\mathbf{b}_q - \mathbf{y}^T\mathbf{y})\frac{1}{2}\mathbf{a}_k^T(\sigma_e^{-2}\mathbf{P}_k^T\mathbf{P}_k$$
$$+ \sigma_a^{-2}I_k - C^{-1})\mathbf{a}_k + \mathbf{a}_k^T(\sigma_e^{-2}\mathbf{P}_k^T\mathbf{y} - \sigma_e^{-2}\mathbf{P}_k^T\mathbf{E}_q\mathbf{b}_q - C^{-1}\mu) + \frac{1}{2}(\mu^T C^{-1}\mu)]\mathcal{N}(\mathbf{a}_k|\mu, C)d\mathbf{a}_k$$

annullando l'espressione tra parentesi che costituisce la matrice della forma quadratica in $\mathbf{a}_k$ e l'espressione tra parentesi che costituisce la matrice che moltiplica $\mathbf{a}_k^T$,si ricava l'espressione per la media e la covarianza

$$C = \sigma_e^{-2}\mathbf{P}_k^T\mathbf{P}_k + \sigma_a^{-2}I_k$$

$$\mu = \sigma_e^{-2}C\mathbf{P}_k^T(\mathbf{y} - \mathbf{E}_q\mathbf{b}_q)$$

grazie al quale l'integranda si semplifica in modo tale che tutta la dipendenza da ak sia attribuita alla gaussiana di parametri μ e C
Calcolando l'integrale si ottiene

$$\begin{aligned} p(k, \mathbf{P}_k|y, \lambda_a, \sigma_a^2) =& p(k|\lambda_A)p(\mathbf{P}_k)\mathcal{K} \cdot exp[\frac{1}{2}\sigma_e^{-2}(+2\mathbf{y}^T\mathbf{E}_q\mathbf{b}_q - \mathbf{b}_q^T\mathbf{E}_q^T\mathbf{E}_q\mathbf{b}_q - \mathbf{y}^T\mathbf{y}) \\ &+ \frac{1}{2}(\mu^T C^{-1}\mu)] \int \mathcal{N}(\mathbf{a}_k|\mu, C)d\mathbf{a}_k \\ =& p(k|\lambda_A)p(\mathbf{P}_k)\mathcal{K} \cdot exp[\frac{1}{2}\sigma_e^{-2}(+2\mathbf{y}^T\mathbf{E}_q\mathbf{b}_q - \mathbf{b}_q^T\mathbf{E}_q^T\mathbf{E}_q\mathbf{b}_q - \mathbf{y}^T\mathbf{y}) \\ &+ \frac{1}{2}(\mu^T C^{-1}\mu)] \end{aligned}$$

da cui

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

Nell'approccio scelto i modelli di processo e di rumore vengono aggiornati separatamente e in sequenza durante le iterazioni.

Chapter 5

Appendici

5.1 APPENDICE: Catene di Markov a stato discreto o continuo

Nelle successive sezioni verranno brevemente presentati i concetti e il formalismo delle catene di Markov a stato discreto per poi passare a quelle a stato continuo. Questo servirà alla comprensione dell'algoritmo utilizzato per campionare la densità di probabilità; esso infatti costruisce una catena di Markov avente la distribuzione desiderata come distribuzione di equilibrio. Lo stato della catena dopo un certo numero di passi è quindi usato come campione estratto e, con i campioni estratti nelle varie iterazioni, si costruisce poi un istogramma che approssima la densità desiderata.
In particolare verrà descritto l'algoritmo di Metropolis Hasting e successivamente l'algoritmo RJMCMC che ne rappresenta una estensione al caso in cui lo stato cambi dimensionalità durante alcune transizioni della catena.

5.1.1 Processo markoviano

Un processo stocastico di Markov è un processo nel quale la probabilità di transizione che determina il passaggio ad uno stato di sistema dipende unicamente dallo stato di sistema immediatamente precedente e non dal "come" si è giunti a tale stato. Tale processo prende il nome dal matematico russo Andrej Andreevič Markov che per primo ne sviluppò la teoria.
Formalmente può essere scritto come

$$P(X(t_{n+1}) \leq x_{n+1} | X(t_n) = x_n, X(t_{n-1}) = x_{n-1}, \ldots, X(t_0) = x_0) = P(X(t_{n+1}) \leq x_{n+1} | X(t_n) = x_n) \tag{5.1}$$

5.1.2 Matrici stocastiche

Una matrice stocastica $P = P[i, j]$ è una matrice tale che

- ogni elemento $P[i, j]$ è non negativo.

- la somma di ogni riga $P[i, \cdot]$ è unitaria

Sia lo spazio di stato $S := \{1, \ldots, n\}$ ogni riga $P[i, \cdot]$ può essere vista come una distribuzione di probabilità sullo spazio S.
Se una matrice P è stocastica anche P^k lo è.

5.1.3 Catene di markov a stato discreto

Una matrice stocastica descrive la dinamica di una catena di markov che prende valori nello spazio di stato S.
Formalemente un processo di Markov a stato discreto $\{X_t\}$ che prende valori in S è una catena di Markov con matrice di transizione P se

$$p\{X_{t+1} = j | X_t = i\} = P[i, j] \tag{5.2}$$

per ogni $t \geq 0$ e $i, j \in S$.
Questa definizione richiede che $\{X_t\}$ abbia la proprietà di Markov ovvero che per ogni t valga

$$\mathbb{P}\{X_{t+1} \,|\, X_t\} = \mathbb{P}\{X_{t+1} \,|\, X_t, X_{t-1}, \ldots\} \tag{5.3}$$

Quindi lo stato X_t è una descrizione completa della configurazione del sistema al tempo t.
Per costruzione $P[i, j]$è la probabilità di andare dallo stato i allo stato j in un passo (unità di tempo).
$P[i, \cdot]$ è quindi la distribuzione di probabilità di X_{t+1} condizionatamente a $X_t = i$;

5.1.4 Distribuzione marginale

Sia

1. $\{X_t\}$ una catena di Markov con matrice di transizione P
2. si denoti la distribuzione di X_t come ψ_t (nota)

Ci si chiede quale sia la distribuzione di X_{t+1} o più in generale di X_{t+m} .
Cerchiamo quindi di risalire a ψ_{t+m} partendo dal caso $m = 1$
Come si risale a ψ_{t+1} dati ψ_t e P?
Per ogni$j \in S$ (stato di arrivo) possiamo decomporre $p(X_{t+1} = j)$ come segue sfruttando la legge della probabilità totale:

$$p\{X_{t+1} = j\} = \sum_{i \in S} p\{X_{t+1} = j \,|\, X_t = i\} \cdot p\{X_t = i\} \tag{5.4}$$

In particolare sono state enumerate tutte le possibilità di arrivo in j e le probabilità di ogni caso sono state sommate.
Usando la definizione della matrice di transizione

$$\psi_{t+1}[j] = \sum_{i \in S} P[i, j] \psi_t[i] \tag{5.5}$$

Si hanno tante equazioni quanti sono gli stati $j \in S$

Se pensiamo a ψ_{t+1} e ψ_t come vettori riga le equazioni sopra possono essere incluse in un'unica equazione matriciale.

$$\psi_{t+1} = \psi_t P \tag{5.6}$$

Ripetendo m volte si ha

$$\psi_{t+m} = \psi_t P^m \tag{5.7}$$

5.1.5 Distribuzioni stazionarie

Dato un operatore di transizione P, possono esistere alcune distribuzioni che rimangono invariate rispetto alla sua applicazione.
Tali distribuzioni sono dette stazionarie.

Formalmente ψ^* su S è stazionaria per P se

$$\psi^* = \psi^* P \tag{5.8}$$

Si evince anche che vale

$$\psi^* = \psi^* P^t \forall t \tag{5.9}$$

Questo sottolinea un fatto importante, ovvero che se ψ_0 è una distribuzione stazionaria, allora X_t avrà la stessa distribuzione per ogni t.

Quindi una distribuzione stazionaria ha l'importante interpretazione come valore di regime del processo stocastico.

Matematicamente una distribuzione stazionaria è un autovettore dell'operatore P con autovalore 1, detto anche *punto fisso*. Per ogni matrice stocastica P esiste almeno una distribuzione stazionaria (teorema di Brouwer) .

Una condizione sufficiente per l'unicità della distribuzione stazionaria è l'ergodicità. Una matrice stocastica P è detta ergodica se esiste un intero positivo m tale che gli elementi di P^m siano strettamente positivi. Inoltre sotto le condizioni di ergodicità vale un importante risultato.

$\forall\, j \in S$

$$\frac{1}{n} \sum_{t=1}^{n} \mathbf{1}\{X_t = j\} \to \psi^*[j] \quad \text{as } n \to \infty \tag{5.10}$$

Dove

1. $\mathbf{1}\{X_t = j\} = 1$ se $X_t = j$ e zero altrimenti

2. la convergenza si ha con probabilità 1

3. il risultato non dipende da ψ_0

La frazione di tempo in cui la catena sosta nello stato j converge a $\psi^*[j]$ quando il tempo tende all'infinito.

5.1.6 Estensione al caso stato continuo

Solitamente i processi di markov sono definiti su uno spazio di stato discreto, tuttavia, se vogliamo includere nello stato della catena anche i coefficienti dello sviluppo dobbiamo estendere i processi al caso continuo.
`http://quant-econ.net/jl/stationary_densities.html`

La famiglia delle distribuzioni $P[i, \cdot]$ sarà rimpiazzata da famiglie di densità $p(x, \cdot)$ una per ogni $x \in S$
In modo analogo al caso discreto, $p(x, \cdot)$ è da intendere come la densità X_{t+1} dato $X_t = x$
Più formalmente un *kernel* stocastico S è una funzione $p\colon S \times S \to \mathbb{R}$ con le proprietà

1. $p(x, y) \geq 0$ for all $x, y \in S$

2. $\int_S p(x, y)dy = 1$ for all $x \in S$

Per il caso discreto valeva che

$$\psi_{t+1}[j] = \sum_{i \in S} P[i, j]\psi_t[i] \tag{5.11}$$

Nel caso continuo bisogna sostituire la sommatoria con un integrale e le masse di probabilità con densità di probabilità.

$$\psi_{t+1}(y) = \int p(x, y)\psi_t(x)\, dx, \qquad \forall y \in S \tag{5.12}$$

E' conveniente vedere questo processo di aggiornamento come un operatore (funzione che manda funzioni in funzioni). Sia $\mathcal{D}$ l'insieme di tutte le densità su S e sia P l'operatore da $\mathcal{D}$ in se' stesso che prende la densità ψ e la mappa nella nuova densità ψP definita da

$$(\psi P)(y) = \int p(x, y)\psi(x)dx \tag{5.13}$$

dunque

$$\psi_{t+1}(y) = (\psi_t P)(y) \tag{5.14}$$

Siccome si intende che la relazione vaga per ogni y si può omettere la variabile e scrivere una espressione simile a quella per il caso discreto.

$$\psi_{t+1} = \psi_t P \tag{5.15}$$

5.1.7 Stazionarietà

Analogamente al caso discreto, dato un kernel stocastico p , si definisce distribuzione stazionaria ψ^* su S, se è un punto fisso rispetto all'operatore P.
In formule

$$\psi^*(y) = \int p(x,y)\psi^*(x)\,dx, \qquad \forall y \in S \tag{5.16}$$

Come nel caso discreto se ψ^* è stazionario per l'operatore P, e la distribuzione diX_0 è ψ^* allora X_t avrà la stessa distribuzione per ogni t. Quindi ψ^* è l'equivalente stocastico della condizione di regime.

5.2 Appendice: Metodo analitico di campionamento

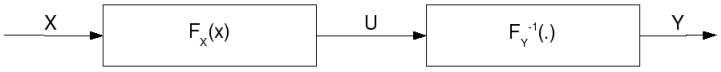
Sia data una variabile casuale X con distribuzione $F_X(x)$. Determinare la funzione

$$g : \mathbb{R} \to \mathbb{R}$$

tale che la $Y = g(X)$ abbia la distribuzione assegnata $F_Y(y)$.
La soluzione di questo problema si articola in due passi:

- passaggio dalla $F_X(x)$ monotona alla variabile casuale U uniforme in $(0,1)$
- passaggio da Ualla variabile casuale Y con $F_Y(y)$monotona.

Con uno schema a blocchi:



Si verifica che: $U = F_X(x)$ è uniforme.
Infatti, si ha

$$F_U(u) = P(U \le u) = P(F_X(x) \le u) = P(X \le x) = U$$

$$Y = F_Y^{-1}(u)$$

Infatti, si ha

$$F_Y(y) = P(Y \le y) = P(F_Y^{-1}(u) \le y) = P(U \le F_Y(y)) = F_U(F_Y(y)) = F_Y(y)$$