

Dispensa Modelli Statistici II Inferenza Bayesiana

Parte I

a.a. 2017-2018

Fulvia Pennoni

Università degli Studi di Milano-Bicocca (IT)

fulvia.pennoni@unimib.it

Introduzione all'inferenza Bayesiana

Il teorema di Bayes (Bayes, 1773) si utilizza per aggiornare l'informazione disponibile in base a nuove conoscenze sull'evento di interesse. Nella definizione di **probabilità condizionata** che coinvolge 2 eventi (A e B) si considera

$$P(A|B) = \frac{P(A, B)}{P(B)} \quad (1)$$

dove $P(B)$ = probabilità di B (detta *prior*); $P(A,B) = P(A \cap B)$ probabilità dell'evento congiunto di A e di B.

Il fatto che si pone una **condizione** indica che nel calcolo si intende una probabilità in base al verificarsi di un altro evento B associato ad A: *quando B si è GIA' verificato*, o nel caso in cui *si verifichi B*. Si tratta di quantificare se e quanta informazione aggiunge ad un evento (A) il fatto di poter osservare un altro evento (B) che è legato ad (A).

Moltiplicando entrambi i lati della equazione 1 per $P(B)$ si ottiene

$$P(B)P(A|B) = P(A \cap B)$$

e scrivendo analogamente si ottiene $P(A \cap B) = P(B|A)P(A)$

Si definisce *Bayes'rule* la seguente

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

Supponendo 3 eventi indicati con A_1 , A_2 e A_3 si può scrivere seguendo la regola precedente che

$$P(A_1, A_2, A_3) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2)$$

Si definisce *law of total probability* (regola della probabilità complessiva) la seguente: sia A_1, \dots, A_n una partizione dello spazio S (ovvero A_i sono considerati eventi disgiunti e tali che la loro unione produce S) dove $P(A_i) > 0$ per ogni i allora

$$P(B) = \sum_{i=1}^n P(B|A_i)P(A_i).$$

Il teorema della probabilità complessiva permette una scomposizione in somme di prodotti composte dalla prior e da una probabilità condizionata. Indica che una probabilità non condizionata può essere ricostruita attraverso una somma pesata delle probabilità condizionate dell'evento in oggetto rispetto alle probabilità a priori di un altro evento (A) quando lo spazio della variabile può essere suddiviso in n parti disgiunte.

Se si considera un ulteriore evento (E) supposto associato con A e B tale che $P(A \cap E) > 0$ e $P(B \cap E) > 0$ la Bayes'rule si scrive nel modo seguente:

$$P(A|B, E) = \frac{P(B|A, E)P(A|E)}{P(B|E)}$$

e la law of total probability si scrive

$$P(B|E) = \sum_{i=1}^n P(B|A_i, E)P(A_i|E).$$

Richiami:

- Indipendenza tra due eventi A,B: $P(A \cap B) = P(A)P(B)$, per $P(A) > 0, P(B) > 0$ si ha che l'indipendenza $P(A|B) = P(A)$ e vale lo stesso per $P(B|A) = P(B)$;
- Indipendenza condizionata tra A e B rispetto ad E: $P(A \cap B|E) = P(A|E)P(B|E)$

Esempio Bayes' billards

In un articolo Stigler (1982) riprende l'esempio proposto da Bayes per illustrare la Bayes'rule. Nell'esempio si dimostra che per ogni intero k e n con $0 \leq k \leq n$ vale la seguente:

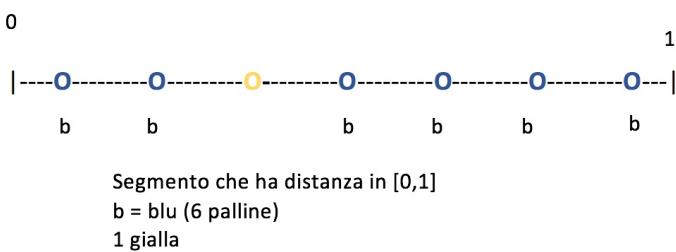
$$P(X = k) = \int_0^1 \binom{n}{k} x^k (1-x)^{n-k} dx = \frac{1}{n+1}$$

ovvero la distribuzione marginale di X è una distribuzione uniforme discreta.

Nelle seguenti due dimostrazioni si illustra senza calcoli che sia la parte destra (dimostrazione 1) e la parte sinistra (dimostrazione 2) dell'equazione precedente sono uguali a $\frac{1}{n+1}$.

Dimostrazione 1 (Parte destra equazione):

Si dispone di n palline colorate blu e di 1 pallina gialla che si tirano presso un segmento di lunghezza predefinita



La posizione delle palline genera delle realizzazioni di numeri pseudo casuali ($p \sim \text{Unif}(0,1)$). La variabile casuale X che conta il numero di palline blu che sono state posizionate prima della pallina gialla (a sinistra), X è una variabile casuale discreta

con valori in $0, 1, \dots, n$. Per calcolare la probabilità di $X = k$ condizionata rispetto a p occorre riferirsi alla probabilità del numero di palline Blu a sinistra della gialla $P(X = k|B = p)$ che è la probabilità definita da una variabile casuale Binomiale $Bin(n, p)$ se si assume ogni lancio come una prova indipendente con distribuzione di Bernoulli con probabilità p .

Si considera la regola delle probabilità totali: si condiziona per $B = p$, pertanto è possibile scrivere:

$$P(X = k) = \int_0^1 P(X = k|B = p)f(p)dp$$

dove $f()$ è la funzione di densità di B . Dalle considerazioni precedenti risulta che

$$P(X = k) = \int_0^1 \binom{n}{k} p^k (1-p)^{n-k} dp = \frac{1}{n+1}.$$

Per $n = 6$

$$P(X = 2) = \frac{1}{6+1} = \frac{1}{7} = 0,14.$$

Dimostrazione 2 (Parte destra dell'equazione): Si dispone di $n + 1$ palline tutte blu. Si tirano a caso nel segmento precedente. Si sceglie una pallina a caso tra quelle tirate e si colora di giallo. Definendo $(X = k)$ come il numero delle palline blu a sinistra della pallina gialla se pallina colorata di giallo è scelta in modo casuale risulta

$$P(X = k) = \frac{1}{n+1}$$

per $k = 0, 1, \dots, n$. Ovvero X ha la **stessa distribuzione** nei due lati dell'equazione. Si noti che il valore dell'integrale prescinde da k (numero di palline blu prima della gialla, numero dei successi).

Considerando α, β interi positivi e sostituendo per $k = (\alpha - 1)$ e $n - k = (b - 1)$

si ottiene la densità di una variable casuale Beta dove:

$$B(\alpha, \beta) = \frac{1}{(\alpha + \beta - 1) \binom{\alpha + \beta - 2}{a - 1}} = \frac{(\alpha - 1)!(\beta - 1)!}{(\alpha + \beta - 1)!}.$$

Richiami sulla variabile casuale Beta

$$X \sim Be(\alpha, \beta)$$

$$f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} \quad 0 \leq x \leq 1 \quad \alpha, \beta > 0$$

$$\text{con funzione Beta pari a } B(\alpha, \beta) = \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx$$

$$E(X) = \left(\frac{\alpha}{\alpha+\beta}\right) \text{ e} \\ var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \text{ La varianza si scrive anche come } var(X) = \frac{E(X)[1-E(X)]}{\alpha+\beta+1}.$$

Generalmente:

- Se $\alpha < 1$ e $\beta < 1$ si ha una forma a U;
- Se $\alpha = \beta$ la densità è simmetrica.
- Se $\alpha = 1$ e $\beta = 1$ la $B(\alpha, \beta) = Unif(0, 1)$.

Richiami esempio

Rispetto all'esempio precedente $p \sim Beta(1, 1) = Unif(0, 1)$; la distribuzione condizionata dell'evento ($X = k$) numero di palline blu prima della pallina gialla è Binomiale $P(X|p) = Bin(n, p)$.

Per ottenere la distribuzione a posteriori di p si utilizza la regola di Bayes:

$$f(p|X = k) = \frac{P(X = k|p)f(p)}{P(X = k)}.$$

Si dimostra che

$$f(p|X = k) \propto p^{\alpha+k-1}(1-p)^{\beta+n-k-1}.$$

ovvero è una $Beta(\alpha + k, \beta + n - k)$.

Si nota in generale che partendo da una distribuzione a priori Beta per il parametro p fissando dei valori per α e β si ottiene una distribuzione a posteriori di tipo Beta e si dice che la famiglia Beta è *coniugata* alla Binomiale. La distribuzione a posteriori di p viene utilizzata quale stima per il parametro oppure si considera il valore atteso o la moda. Da notare che la stima di massima verosimiglianza di p è pari a $\frac{k}{n}$.

Cf. la simulazione per l'esempio nella parte di applicazione.

Definizioni 1

In un contesto parametrico la teoria della regola di Bayes precedente viene impiegata per stimare $\boldsymbol{\theta} \in \Theta$. La **probabilità iniziale (prior)** $p(\boldsymbol{\theta}|H)$ è la funzione di densità/probabilità del vettore dei parametri che dipende (è condizionata) dall'informazione iniziale H . Nel contesto sperimentale o osservazionale sono congetture sul valore del parametro o sono informazioni ricavate da studi precedenti.

Disponendo di informazioni campionare sulla variabile casuale di interesse \mathbf{X} e considerando il parametro $\boldsymbol{\theta}$ la prior viene sostituita dalla distribuzione detta **probabilità a posteriori (posterior)** che risulta la seguente

$$p(\boldsymbol{\theta}|\mathbf{X}, H) = \frac{p(\boldsymbol{\theta}, \mathbf{X}|H)}{p(\mathbf{X}|H)} = \frac{p(\mathbf{X}|\boldsymbol{\theta}, H)p(\boldsymbol{\theta}|H)}{p(\mathbf{X}|H)} \quad (2)$$

con $p(\mathbf{X}|H) = \int_{\Theta} p(\boldsymbol{\theta}, \mathbf{X}|H)d\boldsymbol{\theta}$ se $\boldsymbol{\theta}$ ha supporto continuo. (I valori al denominatore si ottengono in base alla definizione di probabilità condizionata).

Esempio

L'esempio seguente riguarda una situazione concreta di applicazione del teorema e del calcolo della costante k detta di normalizzazione.

Un dottore visitando un paziente implicitamente assegna una probabilità al fatto che abbia una certa malattia, supponiamo si tratti di tonsillite. Dalla visita il dottore stabilisce che forse (con probabilità pari a 0.7) il paziente ha la malattia. Formalizzando si ha

$$p(\theta = 1|H) = 0.7$$

Il dottore dispone di un test (es. tampone) sapendo che il test precedentemente è risultato positivo 95 volte su 100 quando c'è effettivamente la tonsillite ma è risultato positivo anche in assenza di malattia per 40 casi su 100. Occorre considerare le seguenti quantità della variabile di interesse (presenza/assenza della malattia) riferite a due valori distinti del parametro

$$p(X = 1|\theta = 1) = 0.95.$$

$$p(X = 1|\theta = 0) = 0.4.$$

E' possibile sulla base delle due informazioni precedenti definire le seguenti probabilità sul parametro condizionate ai valori osservati e alle informazioni iniziali (numeratore dell'equazione (2))

$$p(\theta = 1|X = 1) \propto p(X = 1|\theta = 1)p(\theta = 1) = 0.95(0.7) = 0.665$$

mentre

$$p(\theta = 0|X = 1) \propto p(X = 1|\theta = 0)p(\theta = 0) = 0.40(0.30) = 0.120.$$

Si noti che le precedenti probabilità sono riferite al parametro condizionatamente a ciò che si è verificato e si noti che si calcolano come prodotto tra il supporto che offrono i dati osservati (i test svolti in precedenza) al valore del parametro e la prior che è dovuta all'esperienza del dottore.

Affinché si tratti di una probabilità *a-posteriori* complessiva è necessario il calcolo della costante k che pone il vincolo della somma pari a 1

$$k(0.665) + k(0.120) = 1$$

da cui $k = 1/0.785$.

Pertanto

$$p(\theta = 1|X = 1) = 0.665k = 0.665/0.785 = 0.847$$

e

$$p(\theta = 0|X = 1) = 0.120k = 0.120/0.785 = 0.153$$

la probabilità iniziale di avere la tonsillite varia (si incrementa nel caso corrente) in base alle informazioni disponibili.

Definizioni 2

Si dispone del vettore $\mathbf{x} = (x_1, \dots, x_n)$ di n osservazioni la cui distribuzione (ovvero il modello statistico) sottostante è funzione di alcuni parametri $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$. Assegnando anche al parametro una distribuzione di densità o di probabilità si enunciano delle supposizioni sul parametro in base a delle conoscenze sul problema oggetto di studio. Questa distribuzione è detta *prior* (a priori). Potendo scrivere

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

è possibile ricavare la distribuzione *posterior* (a posteriori) ovvero la distribuzione dei parametri condizionata ai dati osservati (rispetto a \mathbf{x}) come

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{x})}.$$

I valori osservati \mathbf{x} sono fissi pertanto da densità/probabilità $p(\mathbf{x}|\boldsymbol{\theta})$ è funzione solo di $\boldsymbol{\theta}$ ed è indicata anche con $\ell(\boldsymbol{\theta}; \mathbf{x})$ o funzione di verosimiglianza. Si scrive infatti anche

$$p(\boldsymbol{\theta}|\mathbf{x}) \propto \ell(\boldsymbol{\theta}, \mathbf{x})p(\boldsymbol{\theta})$$

dove il simbolo \propto che indica "proporzionale", per indicare che la distribuzione a posteriori si ottiene moltiplicando la funzione definita a priori per la verosimiglianza.

Il risultato va aggiustato con la [costante di normalizzazione](#) per ottenere una densità/probabilità.

Ad esempio si considera la costante di normalizzazione k

$$p(\boldsymbol{\theta}|\mathbf{x}) = \textcolor{blue}{k} p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

e si pone il vincolo affinché sia una funzione di densità

$$1 = \int_{\Theta} \textcolor{blue}{k} p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$$

e

$$\frac{1}{k} = \text{E}_{\boldsymbol{\theta}}[p(\mathbf{x}|\boldsymbol{\theta})].$$

Il valore atteso precedente identifica la distribuzione predittiva di \mathbf{X} : ovvero la distribuzione attesa della variabile casuale rispetto alla scelta che è stata effettuata per la densità *a priori*. L'adeguatezza della prior si verifica anche considerando la distribuzione predittiva [prima](#) di osservare \mathbf{X} , confrontando i valori realizzati rispetto quelli previsti e valutando in questo modo lo schema inferenziale adottato.

Nel seguente si mostra la natura *sequenziale* del ragionamento Bayesiano: prendendo in considerazione [la prior](#), [la verosimiglianza](#) e la [densità a posteriori](#), per un vettore di osservazioni iniziali $\mathbf{x}_1 = (x_{11}, \dots, x_{1n})$

$$p(\boldsymbol{\theta}|\mathbf{x}_1) \propto \ell_1(\boldsymbol{\theta}; \mathbf{x}_1)p(\boldsymbol{\theta}).$$

e poi un altro vettore di realizzazioni $\mathbf{x}_2 = (x_{21}, \dots, x_{2n})$ indipendenti da quelle ottenute in precedenza si ha

$$p(\boldsymbol{\theta}|\mathbf{x}_1, \mathbf{x}_2) \propto \ell_2(\boldsymbol{\theta}; \mathbf{x}_2)p(\boldsymbol{\theta}|\mathbf{x}_1)$$

$$\propto \ell_1(\boldsymbol{\theta}; \mathbf{x}_1)\ell_2(\boldsymbol{\theta}; \mathbf{x}_2)p(\boldsymbol{\theta})$$

e generalizzando ad una sequenza di dati osservati in successione $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ si

ha

$$p(\boldsymbol{\theta} | \mathbf{x}_n, \mathbf{x}_{n-1}, \dots, \mathbf{x}_1) \propto \left(\prod_{i=1}^n \ell_i(\boldsymbol{\theta}; \mathbf{x}_i) p(\boldsymbol{\theta}) \right).$$

dove la densità a posteriori è definita rispetto ai prodotti delle rispettive verosimiglianze.

Esempio Bayes'Billard generalizzato

L'esempio delle palline illustrato in precedenza è un caso particolare del seguente contesto. Si dispone di un campione casuale di ampiezza n di realizzazioni \mathbf{x} da una variabile casuale di Bernoulli ($0 < \theta < 1$) ogni singola prova ha probabilità costante pari a θ di successo (θ è una variabile aleatoria). Nel caso di dati dicotomici (successo/insuccesso) la distribuzione di densità assegnata alla proporzione di successi θ è una $Beta(\alpha, \beta)$. Il modello statistico si denota

$$X_1, X_2, \dots, X_n \sim Be(\theta) \quad \theta \sim Beta(\alpha, \beta)$$

ovvero X ha una distribuzione di Bernoulli con parametro θ pertanto

$$P(X = x | \Theta = \theta) = \theta^x (1 - \theta)^{1-x} \quad x = 0, 1.$$

Quando si osserva $\mathbf{x} = (x_1, x_2, \dots, x_n)$ la funzione di verosimiglianza per \mathbf{x} fissato è

$$\ell(\theta, \mathbf{x}) = \prod_{i=1}^n [\theta^{x_i} (1 - \theta)^{1-x_i}] = \theta^k (1 - \theta)^{n-k}$$

dove $k = \sum_{i=1}^n x_i$ indica il numero degli eventi ricercati (successi). Si noti che la quantità è proporzionale ad una distribuzione $Beta(k + 1, n - k + 1)$.

Se secondo le ipotesi del ricercatore la distribuzione Beta è quella che identifica i possibili valori assumibili dal parametro distribuzione a priori è la Beta definita da due parametri $\theta \sim Beta(\alpha, \beta)$. Fissando dei valori per i parametri α e β

$$p(\theta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1}$$

ed applicando la Bayes'rule si ottiene

$$p(\theta|\mathbf{x}) \propto \frac{1}{B(\alpha, \beta)} \theta^{k+\alpha-1} (1-\theta)^{n-k+\beta-1}$$

Si dimostra che la costante di normalizzazione è pari a $B(\alpha, \beta)^{-1} \times B(k + \alpha, n - k + \beta)^{-1}$ e pertanto anche la distribuzione a posteriori è una distribuzione di densità Beta con i seguenti valori dei parametri $Beta(k + \alpha, n - k + \beta)$. Si noti che il valore atteso risulta pari a $\frac{k+\alpha}{n+\alpha+\beta}$ mentre la moda pari a $\frac{k+\alpha-1}{n+\alpha+\beta-2}$.

Il modello è definito Beta-Binomiale e si stabilisce che la famiglia Beta è *coniugata* dalla Binomiale. Ovvero la Binomiale fa sì che partendo da una distribuzione a priori Beta assegnata in base all'oggetto si studio si ottenga la stessa forma funzionale per la distribuzione a posteriori.

Esempio

Esito rispetto ad un referendum si/no. Disponendo di vari esiti di sondaggi e considerando il parametro p riguardante la probabilità di esito positivo possiamo ipotizzare $p \sim Beta(\alpha, \beta)$. In tal caso la distribuzione condizionata del numero di "si" (X) al valore del parametro rispetto alla popolazione degli aventi diritto al voto è una Binomiale $X|p \sim Bin(n, p)$. La distribuzione a posteriori si ottiene applicando la Bayes'rule come visto nelle precedenti sezioni.

Scambiabilità

Nel seguito si richiama la definizione di scambiabilità delle componenti in quanto per gli esempi si intende sfruttare questa proprietà. Se si dispone di una permutazione di $\{1, \dots, n\}$ degli indici delle variabili X_1, \dots, X_n queste sono dette **scambiabili** quando tutte le possibili permutazioni ($n!$) presentano la stessa distribuzione n -dimensionale. Dalla definizione si deduce che tutte le distribuzioni marginali di ogni permutazione possibile devono essere le stesse.

Inoltre una sequenza infinita di variabili casuali X_1, X_2, \dots , è detta scambiabile se ogni $n - pla$ di variabili aleatorie scelta dalla successione risulta scambiabile.

Una definizione equivalente è la seguente: Le variabili aleatorie X_1, \dots, X_n si dicono scambiabili se la funzione di ripartizione congiunta di una qualunque permutazione X_{i_1}, \dots, X_{i_n} delle componenti coincide con quella delle variabili aleatorie originarie.

Il seguente teorema denominato [teorema di rappresentazione](#) o di [scambiabilità](#) per successioni di variabili casuali è stato proposto da Bruno De Finetti (1937). Viene enunciato nel seguito in riferimento ad eventi che possono assumere solo valori 0 oppure 1.

Definizione: Siano X_1, X_2, \dots una successione di variabili scambiabili che assumono solo i valori 0 oppure 1 con distribuzione di probabilità P e sia S_n per ogni intero n pari a $S_n = X_1 + X_2 + \dots + X_n$. Esiste una funzione di ripartizione $F(\theta)$ tale che per ogni n-pla (x_1, x_2, \dots, x_n) si ha

$$P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \int_0^1 \prod_{i=1}^n [\theta^{x_i} (1-\theta)^{1-x_i}] dF(\theta)$$

ovvero

$$P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \int_0^1 [\theta^{s_n} (1-\theta)^{n-s_n}] dF(\theta)$$

con

$$F(\theta) = \lim_{n \rightarrow \infty} P\left(\frac{S_n}{n} \leq \theta\right)$$

e

$$\theta = \lim_{n \rightarrow \infty} \frac{S_n}{n}$$

E' possibile notare che sotto l'ipotesi di scambiabilità le variabili osservabili possono essere considerate indipendenti e somiglianti condizionatamente al valore di θ . In pratica nel caso illustrato in precedenza di variabili binarie non è necessario ai fini inferenziali esplicitare P ma basta specificare $F(\theta)$, ovvero la distribuzione iniziale del parametro che induce la distribuzione probabilistica della successione X_1, X_2, \dots . Non è necessario conoscere tutte i risultati delle prove ma il numero totale dei successi. Si noti che l'elicitazione della prior deve avvenire rispetto al comportamento

asintotico di $F(\theta)$. In pratica se vale il teorema le variabili casuali X_1, X_2, \dots, X_n possono essere considerate indipendenti e somiglianti.

Esempio

Nel seguente esempio riguardante l'emofilia (malattia congenita ereditaria che inibisce il cromosoma X detta X-linked recessiva, cf. Wiki) tratto da (Gelman, Carlin, Stern e Rubin, 1995) si mostrano le quantità coinvolte nel procedimento Bayesiano rispetto alla probabilità per una donna di risultare portatrice del gene della malattia. La donna è portatrice di questo gene perché il cromosoma X impedisce l'espressione della malattia. La donna ne è colpita solo se figlia di padre emofilitico e madre portatrice sana.

Se la donna X ha un fratello affetto da emofilia (malattia congenita ereditaria), significa che la mamma è portatrice di due geni: quello nocivo e quello neutro. Si è interessati alla probabilità che la donna X sia o meno portatrice del gene considerando che il padre della donna non è affetto da questa malattia. Se si considera $\theta = 1$: la donna X è portatrice del gene e $\theta = 0$: la donna X non è portatrice del gene, la distribuzione a priori per θ in base alle considerazioni esposte risulta tale che $p(\theta = 1) = p(\theta = 0) = 1/2$.

L'informazione a priori viene aggiornata considerando ad esempio la donna X abbia 2 figli. Se Y_1 identifica la malattia del primo figlio (0 assente, 1 presente) e Y_2 identifica la malattia del secondo figlio: se donna è portatrice i figli hanno una probabilità pari a 0.5 di sviluppare la malattia (considerando nullo il tasso di mutazione). Sotto le ipotesi che la distribuzione congiunta $f(y_2, y_1)$ delle due variabili che identificano la malattia dei figli soddisfi il teorema di rappresentazione di de Finetti, ovvero risulti invariante rispetto a permutazioni degli indici e che i due fratelli possano sviluppare la malattia in modo indipendente (i figli non sono gemelli identici) si calcola

$$p(y_1 = 0, y_2 = 0 | \theta = 1) = 0.5(0.5) = 0.25$$

e

$$p(y_1 = 0, y_2 = 0 | \theta = 0) = 1(1) = 1.$$

Se la donna non è portatrice c'è una probabilità prossima a 1 che i figli non abbiano il gene.

La distribuzione a posteriori rispetto alla probabilità che la donna (mamma) sia portatrice è data da

$$p(\theta = 1 | \mathbf{y}) = \frac{p(y|\theta = 1)p(\theta = 1)}{p(y|\theta = 1)p(\theta = 1) + p(y|\theta = 0)p(\theta = 0)}$$

che risulta pari a

$$p(\theta = 1 | \mathbf{y}) = \frac{0.25(0.5)}{0.25(0.5) + 1(0.5)} = \frac{0.125}{0.525} = 0.20.$$

Si definisce *prior odds* la quantità data dal rapporto $0.5/0.5 = 1$ rispetto al fatto che la mamma sia portatrice. Mentre si chiama *posterior odds* la quantità data dal rapporto $0.25/1 = 0.25$ che la mamma sia portatrice secondo la distribuzione a posteriori.

Bayes'rule applicata ai modelli miscuglio

Nei modelli di mistura finita illustrati nelle dispense (Pennoni, 2017, Parte Modelli Statistici II) si utilizza l'algoritmo Expectation-Maximization per la stima dei parametri. Si considerano nel seguito solo miscugli di densità di Gauss e si illustrano le quantità coinvolte nei due passi dell'algoritmo:

- **passo E:** alla generica iterazione l si applica la regola di Bayes considerando che la distribuzione *a posteriori* dei dati mancanti si può esprimere come

$$f(\mathbf{u} | \mathbf{y}) = \frac{f(\mathbf{y} | \mathbf{u}) f(\mathbf{u})}{f(\mathbf{y})} \quad (3)$$

dove $f(\mathbf{u})$ indica la distribuzione *a priori* dei dati mancanti. Si calcolano ad esempio le probabilità *a-posteriori* che l'osservazione i appartenga alla componente j condizionatamente al valore y_i e al valore dei parametri $\boldsymbol{\theta}^l$ all'iterazione corrente come

$$\tau_{ij}^{(l)} = \frac{f_j(y_i | \phi_j^l) \pi_j^l}{f(y_i; \boldsymbol{\theta}^l)}$$

per $i = 1, \dots, N$ e $j = 1, \dots, k$. In questo passo si sostituiscono le stime di massima verosimiglianza che esistono in forma chiusa;

- al **passo M** alla $(l + 1)$ -esima iterazione si massimizza la log-verosimiglianza rispetto ai valori del passo E e si ottiene ad esempio il peso della j -esima componente nel modo seguente

$$\pi_j^{l+1} = \frac{1}{N} \sum_{i=1}^N \tau_{ij}^{(l)} \quad j = 1, \dots, k.$$

Le medie e le varianze si stimano considerando che $\boldsymbol{\theta}^{(l+1)}$ è la soluzione dell'equazione

$$\sum_j \sum_n \tau_{ij}^{(l)} \frac{\partial \log f(y_i; \boldsymbol{\theta}^l)}{\partial \theta} = 0.$$

La funzione di log-verosimiglianza è lineare e la soluzione esiste in forma chiusa ed è la seguente

$$\begin{aligned} \mu_j^{(l+1)} &= \frac{\sum_i \tau_{ij}^{(l)} y_i}{\sum_i \tau_{ij}^{(l)}}; \\ \sigma_j^{2(l+1)} &= \frac{\sum_i \tau_{ij}^{(l)} (y_i - \mu_j^{(l+1)})^2}{\sum_i \tau_{ij}^{(l)}}. \end{aligned}$$

Per costruzione le componenti sono delle ellissi centrate sulla media data la scomposizione spettrale di $\boldsymbol{\Sigma}$ (matrice di varianza e covarianza). Ad esempio per la componente j -esima risulta

$$\boldsymbol{\Sigma}_j = \lambda_j \mathbf{D}_j \mathbf{A}_j \mathbf{D}_j^\top$$

dove \mathbf{D} fornisce l'orientamento essendo la matrice ortogonale degli autovettori; \mathbf{A}

determina la forma della densità essendo una matrice diagonale con valori che sono proporzionali agli autovalori posti in ordine decrescente; e lo scalare λ_j determina il volume.

Scelta della prior

Nel seguito si elencano i principali metodi di scelta della distribuzione iniziale proposti in letteratura. La scelta generalmente avviene attraverso:

- i) la specificazione soggettiva (in base alle conoscenze del settore);
- ii) la specificazione secondo una distribuzione di densità o di probabilità;
- iii) la relazione in base ad una famiglia coniugata;
- iv) la scelta di distribuzione a priori non informativa;
- v) distribuzioni gerarchiche.

Nel seguente si introducono brevemente alcune considerazioni e qualche esempio per ogni punto precedente. Si noti che si parla di **inferenza robusta** dal punto vista dell'approccio Bayesiano quando l'inferenza a posteriori (le considerazioni sulla distribuzione del parametro) risulta simile anche se si utilizzano diverse distribuzioni a priori che soddisfano le conoscenze iniziali.

- i) Nel caso di specificazione **soggettiva** si considera il vettore dei parametri θ non noto e si utilizza ad esempio l'istogramma per assegnare la probabilità ad ogni intervallo discreto. Si considera $P(X \leq q_\alpha) = \alpha$ with $\alpha \in [0, 1]$ in modo da assegnare una probabilità ad ogni quartile della distribuzione. Nel caso siano plausibili diverse distribuzioni a priori si la probabilità relativa di alcuni valori del parametro rispetto ad altri. Berger, (1995) ha proposto il calcolo della quantità

$$P(\theta = \theta_0 | \theta = \theta_0 \text{ or } \theta = \theta_1) = \frac{p(\theta_0)}{p(\theta_0) + p(\theta_1)}$$

dove $P(\theta)$ indica la distribuzione a priori, per poter assegnare dei pesi all'insieme dei valori plausibili per θ a priori.

In questo contesto si parla di l'*elicitazione*: si considera la conoscenza del fenomeno o i pareri da parte di esperti sul problema d'interesse. Pertanto l'inferenza comprende i seguenti steps:

- la definizione del problema;
- la definizione delle misure di sintesi;
- la stima della densità/probabilità congiunta;
- la validazione delle procedura inferenziale.

ii) Nel caso di specificazione della prior attraverso una specifica distribuzione probabilistica ci si riferisce generalmente ad una famiglia parametrica. Ad esempio se l'ipotesi di distribuzione simmetrica per i valori del parametro è plausible (la densità diminuisce più velocemente quando si tratta di valori del parametro più lontani dalla moda e i valori lontani dalla moda hanno probabilità irrilevanti) allora la distribuzione di Gauss viene assunta come prior. I parametri della distribuzione a priori μ, σ sono definiti *iperparametri*.

La conoscenza a priori circa il valore del parametro può essere utilizzata per specificare una densità a priori della particolare forma funzionale scelta, come nel seguente esempio.

Esempio

Si consideri il seguente esempio che riguarda l'impatto dell'osteoporosi curata con un nuovo farmaco. Occorre stabilire la dimensione ottimale di un campione per impostare un trial clinico per un nuovo farmaco. Il parametro d'interesse è riferito alla possibilità di fratture nell'arco di 5 anni per il paziente curato con il farmaco. Se il dottore stabilisce le seguenti probabilità *a priori* in fase di elicitation $p_{0.25} = 0.47, p_{0.5} = 0.61, p_{0.75} = 0.74$ a cui si aggiungono quelle riferite ai quantili seguenti $p_{0.1} = 0.35$ e $p_{0.9} = 0.83$ per poter utilizzare la distribuzione Beta occorre trovare gli iperparametri $\alpha, \beta > 0$ condizionati a

questi valori

$$\min_{\alpha, \beta} \left[\sum_{i=1}^5 [Beta(p, \alpha, \beta) - p_i]^2 \right].$$

dove $p = c(0.10, 0.25, 0.5, 0.75, 0.9)$ sono i quantili che devono assumere i valori specificati sopra. Minimizzando la distanza di cui sopra (cf. dispense parte applicata) si ricavano i seguenti valori dei parametri della distribuzione Beta $\hat{\alpha} = 2.1$ e $\hat{\beta} = 1.4$, infatti per la Beta(2.1, 1.4) si ha $\hat{\pi}_{0.75} = 0.79$. \diamond

- iii) Si dice che una distribuzione di probabilità iniziale è **coniugata** al modello utilizzato o, equivalentemente, alla funzione di verosimiglianza, se la forma funzionale della distribuzione iniziale e della distribuzione finale sono uguali.

Nel caso specifico della famiglia coniugata Gaussiana:

- distribuzione delle osservazioni $l(X; \theta) \sim N(\theta, \sigma^2)$;
- prior $p(\theta) \sim N(\mu, \tau^2)$;
- posterior $p(\theta|x) \sim N(\mu_1, \tau_1^2)$.

Il vantaggio di questa situazione è che aggiungendo osservazioni occorre modificare soltanto i parametri della distribuzione a posteriori.

Definizione: Sia $F = \{p(x|\theta), \theta \in \Theta\}$ la famiglia di campionamento o di osservazione. Un insieme di distribuzioni P si definisce **famiglia coniugata** rispetto a F se per ogni $p \in F$ e per ogni probabilità a priori $p(\theta) \in P$ la distribuzione a posteriori risulta $p(\theta|x) \in P$. \diamond

Si richiede implicitamente che tutte le distribuzioni marginali abbiano la stessa forma distributiva (cf. Teorema di rappresentazione De Finetti). Occorre che ogni possibile valore di θ abbia una probabilità positiva anche se molto bassa (Lindley Cromwell's rule).

Tra le famiglie coniugate si distinguono le famiglie *naturali* rispetto al modello di campionamento riferito alla log-verosimiglianza. Ad esempio, la classe di

tutte le distribuzioni $B(\alpha, \beta)$ con parametri che assumono valori interi rappresenta una *famiglia coniugata naturale* per lo schema di campionamento di Bernoulli. Esiste la classe delle famiglie coniugate nell'ambito della famiglia esponenziale.

Esempio 1 Siano Y_1, \dots, Y_n variabili casuali indipendenti ed distribuite come una $Poisson(\lambda)$ con $\lambda > 0$. La probabilità congiunta risulta

$$f(\mathbf{y}|\lambda) = \prod_i^n \frac{e^{-\lambda} \lambda^{y_i}}{y_i!}$$

dove $i = 1, \dots, n$, e la verosimiglianza ha il kernel della famiglia Gamma

$$l(\lambda, \mathbf{y}) \propto e^{-n\lambda} \lambda^{\sum y_i}.$$

La distribuzione a priori per λ nella famiglia coniugata è $\lambda \sim Gamma(\alpha, \beta)$ tale che la densità a posteriori risulta

$$p(\lambda|\mathbf{y}) \propto \lambda^{\alpha + \sum y_i - 1} \exp(-(\beta + n)\lambda)$$

tale che $p(\lambda|\mathbf{y}) \sim Gamma(\alpha + \sum y_i, \beta + n)$ (cf. esempio delle dispense parte delle applicazioni).

Esempio 2 Per la distribuzione coniugata Gaussiana introdotta in precedenza quando il parametro di interesse è media la prior è data da $\theta \sim N(\mu, \tau^2)$ e supponendo che $p(X|\theta) \sim N(\theta, \sigma^2)$ con σ^2 noto la distribuzione a posteriori $p(\theta|X = x)$ è una distribuzione di Gauss ($\theta|X = x) \sim N(\mu_1, \tau_1^2)$ dove

$$\mu_1 = \frac{\frac{1}{\tau^2}}{\frac{1}{\sigma^2} + \frac{1}{\tau^2}} \mu + \frac{\frac{1}{\sigma^2}}{\frac{1}{\sigma^2} + \frac{1}{\tau^2}} x \quad e \quad \tau_1^2 = \frac{1}{\frac{1}{\sigma^2} + \frac{1}{\tau^2}}$$

e μ_1 si scrive anche in funzione di τ^1 nel seguente modo

$$\mu_1 = (\sigma^{-2}x + \mu\tau^{-2})\tau_1^2.$$

Ovvero la media a posteriori è una media pesata rispetto alla distribuzione a priori e alla verosimiglianza. Il peso della media a priori sarà tanto più elevato quanto più è bassa la variabilità stabilita da τ^2 (i valori a priori sono poco variabili). Se la precisione $1/\sigma^2$ è bassa (i dati mostrano una variabilità ridotta) allora il peso delle informazioni campionarie sarà rilevante. Nelle applicazioni quando la varianza τ^2 viene imposta molto elevata la specificazione della prior è detta non informativa.

- La **precisione** è definita come il reciproco della varianza e viene assunta per quantificare l'informazione disponibile;
- La **precisione a posteriori** è data dalla somma della precisione della funzione di verosimiglianza e dalla precisione della distribuzione a priori.

Dimostrazione: La precedente si dimostra notando che

$$p(\theta|x) \propto \ell(\theta; x)p(\theta)$$

$$\begin{aligned} p(\theta|x) &\propto \exp \left\{ -\frac{1}{2\sigma^2}(x-\theta)^2 - \frac{1}{2\tau^2}(x-\mu)^2 \right\} \\ &\propto \exp \left\{ -\frac{\theta^2}{2\sigma^2} - \frac{\theta^2}{1\tau^2} + \frac{x\theta}{\sigma^2} + \frac{\mu\theta}{\tau^2} \right\} \\ &\propto \exp \left\{ -\frac{\theta^2}{2} \left(\frac{1}{\sigma^2} + \frac{1}{\tau^2} \right) + \theta \left(\frac{x}{\sigma^2} + \frac{\mu}{\tau^2} \right) \right\} \end{aligned}$$

Ponendo $\tau_1^2 = (\tau^{-2} + \sigma^{-2})^{-1}$ e $\mu_1 = (\sigma^{-2}x + \mu\tau^{-2})\tau_1^2$ si ha che

$$\begin{aligned}
p(\theta|x) &\propto \exp\left(-\frac{\theta^2}{2\tau_1^2} + \frac{\theta\mu_1}{\tau_1^2}\right) \\
&\propto \exp\left(-\frac{1}{2\tau_1^2}(\theta - \mu_1)^2\right) \\
&\propto \frac{1}{\sqrt{2\pi\tau_1^2}} \exp\left(-\frac{1}{2\tau_1^2}(\theta - \mu_1)^2\right).
\end{aligned}$$

◊

La media \bar{X} è una statistica sufficiente per θ ed ha la seguente distribuzione $\bar{X} \sim N(\theta, \frac{\sigma^2}{n})$ pertanto il risultato precedente si generalizza per n valori nel modo che segue:

$$\mu_1 = \frac{n\frac{1}{\sigma^2}\bar{x} + \frac{1}{\tau^2}\mu}{\frac{1}{n\sigma^2} + \frac{1}{\tau^2}} \quad \tau_1^2 = \frac{1}{\frac{1}{n\sigma^2} + \frac{1}{\tau^2}}.$$

- iv) Una prior non informativa (reference prior) si utilizza nell'ambito dell'inferenza Bayesiana quando occorre assegnare maggior peso alle osservazioni e non alle informazioni soggettive sulla prior. Inizialmente sono state proposte prior basate sulla distribuzione uniforme in modo da assegnare stessa probabilità iniziale ad ogni punto del supporto del parametro: $p(\theta) \propto k$ con k costante. Tuttavia questa scelta comporta che la distribuzione iniziale non risulta appropriata quando il campo di variazione di θ non è limitato ed inoltre non soddisfa la proprietà di invarianza rispetto a trasformazioni del parametro.

La prior proposta da Jeffreys è tale che

$$p(\theta) \propto I[(\theta)]^{1/2}$$

con $\theta \in \Theta$ e con I che denota il valore atteso dell'informazione di Fisher (derivata seconda della log-verosimiglianza cambiata di segno). Benché questa classe di distribuzioni a priori ha la proprietà di invarianza alcune volte con-

duce a distribuzioni a priori improprie che tuttavia comportano distribuzioni a posteriori che sono proprie. (Questo metodo di condurre l'inferenza è simile al il metodo della massima verosimiglianza in cui quest'ultima è viene penalizzata, Firth, 1993). Tuttavia l'uso della prior proposta da Jeffrey non soddisfa il principio di verosimiglianza: anche utilizzando due stesse verosimiglianze se l'informazione di Fisher risultante è diversa, l'inferenza Bayesiana conduce a due diverse distribuzioni a posteriori.

Bernardo nel 1979 ha proposto la definizione di reference prior che è data dalla distribuzione che massimizza l'informazione che non sarebbe possibile ricavare sul parametro d'interesse in base ad un numero infinito di replicazioni dell'esperimento ovvero

$$p(\theta) = \arg \max_{p_n(\theta)} I(\mathbf{X}_\infty, \theta)$$

dove $I(\mathbf{X}_\infty, \theta) = \lim_{n \rightarrow \infty} I(\mathbf{X}_n, \theta)$.

Regioni di credibilità

Si tratta di un insieme di valori della distribuzione a posteriori in grado di fornire una stima intervallare per il parametro di interesse simile a quella ottenibile nell'inferenza classica con l'intervallo di confidenza.

Definizione: Sia $\boldsymbol{\theta}$ definito Θ una regione $\mathbf{C} \subset \Theta$ è una [una regione di credibilità](#) o una regione di confidenza Bayesiana

$$P(\boldsymbol{\theta} \in \mathbf{C} | \mathbf{x}) \geq 1 - \alpha.$$

dove $(1 - \alpha)$ indica il livello di credibilità. ◊

La regione è definita in $[c_1, c_2]$ e non è sempre un intervallo univoco. Si richiede che sia α che \mathbf{C} siano relativamente piccoli. La regione di confidenza ha la proprietà di invarianza rispetto a trasformazioni del parametro. La lunghezza dell'intervallo è inversamente proporzionale all'altezza della densità: le distribuzioni a posteriori

con densità maggiormente concentrate permettono di ricavare intervalli più corti rispetto a quelle con più alta variabilità.

Definizione: Una **una regione di credibilità o una regione Bayesiana di confidenza** avente la massima densità a posteriori (highest posterior density, HPD) per θ al livello $100(1 - \alpha)\%$ è tale che

$$C = \{\theta \in \Theta : p(\theta|x) \geq k(\alpha)\}$$

con $k(\alpha)$ costante più elevata tale che $P(\theta \in C|x) \geq 1 - \alpha$. \diamond

Si noti che la regione di credibilità nell'esempio indicato nella Figura 1 seguente è pari all'unione delle due regioni C_1 e C_2 .

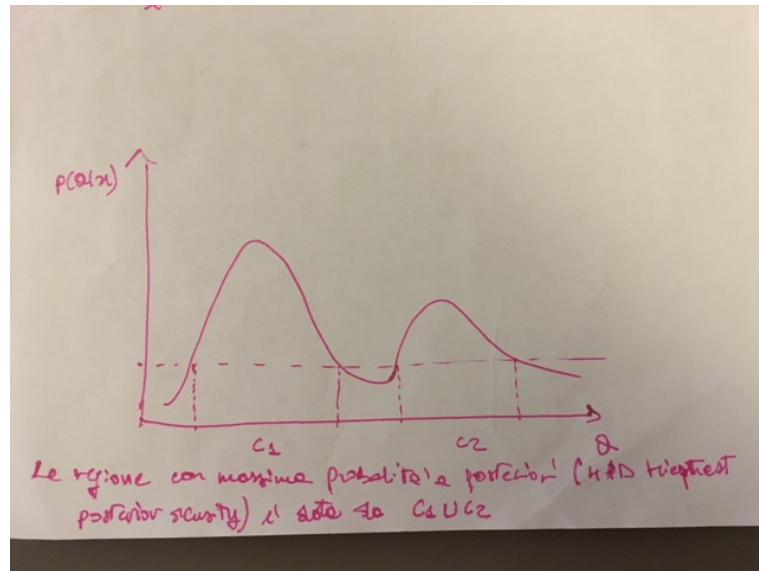


Figure 1: Esempio di regione con massima probabilità a posteriori.

Quando si sceglie una prior non informativa la regione di credibilità potrebbe portare ad un intervallo di credibilità o cui estremi corrispondono agli estremi dell'intervallo di confidenza ottenuto considerando la quantità pivotale nell'inferenza classica.

Esempio Nell'esempio 2 illustrato in precedenza l'intervallo di credibilità per θ al livello $100(1-\alpha)\%$ corrispondente alla massima probabilità a posteriori si determina

$$(\mu_1 - z_{\alpha/2}\tau_1, \mu_1 + z_{\alpha/2}\tau_1)$$

dato che la distribuzione a posteriori è simmetrica pertanto si ha

$$1 - \alpha = P\left(z_{\alpha/2} < \frac{\theta - \mu_1}{\tau_1}\right) < z_{\alpha/2}$$

condizionamente ad un campione di realizzazioni da $\mathbf{x} = (x_1, \dots, x_n)$ supposte generate da $X \sim N(\theta, \sigma^2)$.

Se si è interessati alla **distribuzione predittiva** ovvero quella rispetto ad osservazioni successive questa si determina come densità marginale di \mathbf{x} . Il calcolo avviene considerando il valore atteso della distribuzione a posteriori. Per maggiori dettagli sugli aspetti applicativi si rimanda a Albert (2009).

Riferimenti

- Albert, J. (2009). *Bayesian computation with R*. Springer-Verlag, New-York.
- Bayes T., Price M. (1763). *An Essay towards solving a Problem in the Doctrine of Chances*, Philosophical Transactions of the Royal Society of London, **53**, 370-418.
- Berger, J. (1985). *Statistical Decision theory and Bayesian Analysis*. Springer-Verlag, New-York.
- De Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Ann. Inst. H. Poincarè*, **7**, 1–68.
- Migon, H. S., Gamerman, D., and Louzada, F. (2014). *Statistical inference: an integrated approach*. Chapman & Hall.

- Gelman, A., Carlin J. B., Stern, H. S. and Rubin, D. B. (2005). *Bayesian Data Analysis*. Chapman & Hall, London.
- Gelman, S., Jones A. and Meng, X. L. (Eds.). (2011). *Handbook of Markov Chain Monte Carlo*. CRC press.
- Gelfand, A. E., and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, **85**, 398-409.
- Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97-109.
- Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, **6**, 721-741.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, **21**, 1087-1092.
- Stigler, S. M. (1982). Thomas Bayes's bayesian inference. *Journal of the Royal Statistical Society. Series A* , 250-258.
- Tanner, M. A., and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, **82**, 528-540.

Dispensa Modelli Statistici II e Inferenza Bayesiana

Parte II

a.a. 2017-2018

Fulvia Pennoni

Università degli Studi di Milano-Bicocca (IT)

fulvia.pennoni@unimib.it

1 Metodi Markov Chain Monte Carlo

I metodi di Monte Carlo (MC) sfruttano le proprietà delle catene di Markov (metodi MC e Markov Chain Monte Carlo) per generare delle realizzazioni di ampie dimensioni che si considerano approssimazioni indipendenti e somiglianti dalla distribuzione *a posteriori* dei parametri d'interesse del modello. Questi permettono pertanto di sviluppare l'inferenza dal punto di vista Bayesiano quando il calcolo analitico non è proponibile.

Nel seguito inizialmente si enunciano alcune delle definizioni utili per comprendere le caratteristiche principali di un processo stocastico di Markov. Si introduce il modello di Markov per dati longitudinali ed il modello di transizione definito sulle variabili latenti (latent Markov model) sempre per dati longitudinali. Nell'ultima sezione si introducono gli algoritmi noti come Metropolis-Hastings e Gibbs sampler che permettono di estrarre osservazioni dalla distribuzione che approssima la distribuzione *a posteriori*.

2 Catene di Markov

Si considera una sequenza temporale di variabili casuali $Y^{(0)}, Y^{(1)}, \dots, Y^{(t)}$ dove $t, t = 1, \dots, T$ indica l'**istante** considerato. Nel seguito $\mathbf{Y}^{(t)} = (Y^{(0)}, Y^{(1)}, \dots, Y^{(t)})$ definisce un **processo stocastico**.

Quando la dipendenza generata dalla struttura temporale è ignorata e le variabili casuali sono assunte con stessa forma distributiva si può scrivere

$$P(Y^{(t)}|Y^{(t-1)}, \dots, Y^{(0)}) = P(Y^{(t)}).$$

In base alla proposta di Markov (1906) la dipendenza temporale viene caratterizzata in modo semplice come segue

$$P(Y^{(t)}|Y^{(t-1)}, \dots, Y^{(0)}) = P(Y^{(t)}|Y^{(t-1)}) \quad \forall t \geq 1.$$

I processi stocastici che soddisfano questa relazione di dipendenza condizionata hanno la proprietà di Markov. La famiglia di variabili casuali che soddisfa questo insieme di relazioni definisce un **processo stocastico Markoviano** a tempo continuo o discreto. Un processo stocastico. È detto Markoviano quando la sua struttura di dipendenza è basata su regole probabilistiche che definiscono il modo in cui la variabile precedente influenza la successiva. Si parla pertanto di **catena** di Markov. La distribuzione congiunta delle variabili casuali definisce le caratteristiche del processo stocastico.

Si dimostra che partendo da una sequenza di variabili casuali indipendenti $X^{(1)}, X^{(2)}, \dots, X^{(t)}, \dots$ con identica distribuzione ad esempio $N(0, 1)$ è possibile

generare una nuova sequenza di tipo Markoviano nel modo seguente

$$\begin{aligned}
 Y^{(1)} &= X^{(1)} \\
 Y^{(2)} &= \alpha(X^{(1)}) + (1 - \alpha^2)X^{(2)} \\
 &\vdots \\
 Y^{(t)} &= \alpha^{(t-1)}X^{(1)} + \sqrt{(1 - \alpha^2)}\left(\alpha^{(t-2)}X^{(2)} + \alpha^{(t-3)}X^{(3)} + \dots + \alpha^{(t)}X^{(t)}\right)
 \end{aligned}$$

con α numero reale tale che $|\alpha| < 1$. Si dimostra che la distribuzione condizionata di $Y^{(t)}$ ai valori precedenti $Y^{(1)}, \dots, Y^{(t-1)}$ è un processo stocastico Markoviano con distribuzione di Gauss con media pari a $\alpha(Y^{t-1})$ e varianza pari a $(\sqrt{(1 - \alpha^2)})$. L'intero processo è un processo di Markov multidimensionale.

La catena di Markov si contraddistingue per delle proprietà enunciate di seguito. La **stazionarietà** della catena si ha quando

$$P(Y^{(1)}, \dots, Y^{(t)}) = P(Y^{(1+m)}, \dots, Y^{(t+m)})$$

per $t \geq 1$ e $m \geq 0$.

La catena di Markov è detta **omogenea** quando vale la proprietà di Markov e la distribuzione condizionata $P(Y^{t+1}|Y^t)$ non dipende da t ovvero

$$P(Y^{(t+1)}|Y^{(t)}) = P(Y^{(2)}|Y^{(1)})$$

per $t \geq 1$.

In generale ci si riferisce a $\mathbf{Y}^{(t)}$ come un processo definito da una sequenza di stati che iniziano con $Y^{(0)}$. Lo **spazio degli stati** del processo è l'insieme dei valori che il processo può assumere. Ogni componente è definita dallo **stato** del processo. Per indicare gli stati la probabilità relativa ad una coppia di stati si indica come

$$P(Y^{(2)} = j | P(Y^{(1)} = k))$$

probabilità di passaggio dallo stato k di $Y^{(1)}$ allo stato j di $Y^{(2)}$. Le probabilità di transizione p_{jk} definisco la probabilità di spostamento tra gli stati. Pertanto il processo è

detto **stazionario o omogeneo** rispetto al tempo quando le probabilità di transizione

$$p_{jk} = P(Y^{(t)} = k | Y^{(t-1)} = j)$$

non cambiano al passare del tempo. Ovvero p_{jk} non dipende da t . Le probabilità di transizione sono definite positive $p_{jk} > 0$ e tali che per $j \geq 1$ e $k \leq t$. Il comportamento della catena si determina solo in relazione alle probabilità iniziali per ogni stato e di transizione tra gli stati.

Considerando lo scenario delle prove ripetute in cui si osserva una realizzazione di $Y^{(t-1)} = j$ si considera la probabilità di $Y^{(t)} = k$ alla prova successiva calcolata in base alla probabilità che lo stato k ha assunto rispetto alle prove effettuate in precedenza che si denota come a^k nel seguito. Con tali probabilità si costruisce una **matrice di transizione** che è una matrice di dimensione $m \times m$. Ad esempio, la matrice 2×2 diventa

$$\boldsymbol{\Pi} = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix}, \quad j, k = 1, 2$$

in cui ogni riga definisce una la probabilità condizionata per uno stato. Si tratta di una matrice detta stocastica tale che è non negativa ed le cui righe sommano a 1. Per questo la matrice è detta matrice stocastica per riga.

Ad esempio, se si scrive

$$\boldsymbol{\Pi} = \begin{pmatrix} 1/3 & 2/3 \\ 1/2 & 1/2 \end{pmatrix}, \quad j, k = 1, 2$$

si nota che partendo dallo stato $j = 1$ (riga 1) si transita da p_1 a p_{11} con una probabilità pari a $1/3$ si transita invece a p_{12} partendo sempre da p_1 con una probabilità pari a $2/3$. Nella seconda riga invece le probabilità di passaggio sono identiche.

Uno stato i è detto **transiente** se c'è una probabilità positiva che partendo da i non si ritorni mai a i . E' anche detto **ricorrente** se c'è una probabilità positiva e pari 1 di transitare a i .

La catena è detta **irriducibile** se considerando una coppia di stati i, j è possibile passare da i a j oppure da j a i in un numero finito di passi, altrimenti la catena è detta riducibile. Se la catena è irriducibile ogni stato è ricorrente ed è possibile visitarlo più volte.

Per un processo stocastico definito come **irriducibile e aperiodico** (omogeneo) è possibile

definire una distribuzione **stazionaria o di equilibrio**. La distribuzione di equilibrio definisce il comportamento asintotico del processo: le probabilità finali del processo dopo un certo periodo di tempo. Quest'ultima è la distribuzione marginale del processo e si calcola risolvendo il sistema di equazioni di Chapman-Kolmogorov che permettono di ricavare la matrice di transizione dopo n passi partendo da certe probabilità iniziali per ciascun stato. Nelle equazioni la probabilità di andare da i a j in $n+m$ passi si calcola prodotto della probabilità di andare da i a k in n passi e poi di andare da k a j in m passi.

Tale distribuzione di equilibrio è definita come limite del prodotto tra le probabilità iniziali $\boldsymbol{\pi}$ e di transizione $\boldsymbol{\Pi}$ al crescere del numero di passi:

$$\lim_{n \rightarrow \infty} \hat{\boldsymbol{\pi}} \boldsymbol{\Pi}^n.$$

Il valore di $\boldsymbol{\pi}$ deve soddisfare la seguente uguaglianza

$$\hat{\boldsymbol{\pi}} \boldsymbol{\Pi} = \hat{\boldsymbol{\pi}}$$

che definisce la distribuzione equilibrio, invariante o steady-state distribution per la matrice di transizione $\boldsymbol{\Pi}$. La quota di tempo speso in uno stato si calcola partendo dalla distribuzione di equilibrio.

Si dimostra che se **A è una matrice stocastica** allora avrà certamente un autovalore pari a 1 e tale autovalore è il più grande. La distribuzione di equilibrio può anche essere ricostruita rispetto agli autovalori corrispondenti all'autovettore più grande e normalizzati. Si denota come autovalore di una matrice A , $n \times n$ quel numero reale λ tale che

$$A\mathbf{v} = \lambda\mathbf{v}$$

dove il vettore colonna \mathbf{v} $n \times 1$ con valori diversi da zero è detto autovettore.

Richiami

Nota Regola di moltiplicazione tra 2 matrici: Date A e B due $m \times m$ matrici stocastiche l'elemento generico della matrice C risultante dal loro prodotto è definito come $c_{ij} =$

$\sum_{k=1}^m a_{ik}b_{kj}$. Ad esempio, il prodotto di A $m \times n$ per B $n \times r$ restituisce C $m \times r$

$$\begin{pmatrix} 0 & 1/2 & 1/2 \\ 0 & 0 & 1/2 \end{pmatrix} \begin{pmatrix} 7 \\ 8 \\ 9 \end{pmatrix} = \begin{pmatrix} 0 \cdot 7 + 1/2 \cdot 8 + 1/2 \cdot 9 \\ 0 \cdot 7 + 0 \cdot 8 + 1/2 \cdot 9 \end{pmatrix} = \begin{pmatrix} 8.5 \\ 4.5 \end{pmatrix}$$

La somma di due matrici A e B è una matrice C i cui elementi sono pari alla somma degli elementi di A e B situati nella corrispondenti posizioni.

$$\begin{pmatrix} 0 & 1/2 & 1/2 \\ 0 & 0 & 1/2 \end{pmatrix} + \begin{pmatrix} 0 & 1/2 & 1/2 \\ 0 & 0 & 1/2 \end{pmatrix} = \begin{pmatrix} 0 & 1/4 & 1/4 \\ 0 & 0 & 1/4 \end{pmatrix}$$

Esempio di passeggiata casuale

Un processo stocastico è definito passeggiata casuale (unidimensional random walk) se il processo È tale che per lo stato n vale:

$$X_n = X_{n-1} + W_n$$

dove W_n È una variabile casuale di Bernoulli a valori ± 1 . Ovvero il valore assunto dal processo dipende dal precedente e si ha solo una probabilità p di spostarsi verso una direzione e di conseguenza $1 - p$ di spostarsi nell'altra direzione. Non È possibile pertanto restare nello stesso stato.

Scrivendo nel modo seguente

$$X_n = X_0 + \sum_{i=1}^n W_i$$

si nota che $P(X_n = j | X_{n-1} = j + 1) = p$ e $P(X_n = j | X_{n-1} = j - 1) = 1 - p$.

Il processo è definito dalla distribuzione iniziale (si suppone come nell'esempio seguente di partire dallo stato 4) e dalla matrice di transizione. Ad esempio, per un processo con 6

possibili stati la passeggiata casuale è definita da

$$\boldsymbol{\pi} = \left(\begin{array}{cccccc} 0 & 0 & 0 & 1 & 0 & 0 \end{array} \right)$$

e dalla matrice di transizione che ha elementi pari a 0 nella diagonale principale esclusi quelli riferiti ai due stati assorbenti (non è possibile rimanere nello stesso stato eccetto per i due stati iniziale e finale). Gli stati ammettono passaggi solo verso stati adiacenti

$$\boldsymbol{\Pi} = \left(\begin{array}{cccccc} 1 & 0 & 0 & 0 & 0 & 0 \\ 1-p & 0 & p & 0 & 0 & 0 \\ 0 & 1-p & 0 & p & 0 & 0 \\ 0 & 0 & 1-p & 0 & p & 0 \\ 0 & 0 & 0 & 1-p & 0 & p \\ 0 & 0 & 0 & 0 & 0 & 1 \end{array} \right).$$

3 Modello di transizione (Transition model)

Nel seguito si illustra il modello detto anche Markov chain model che utilizza le proprietà del processo esposto in precedenza nel contesto dei dati longitudinali. Per una trattazione dettagliata si veda Bartolucci, Farcomeni e Pennnoni (2013). I dati longitudinali si presentano quando si dispone di osservazioni ripetute su svariati periodi temporali per un insieme di unità. Differiscono dalle serie storiche in quanto nei dati in serie vi sono tantissimi periodi temporali di osservazione ma in generale si tratta di un'unica unità statistica o poche unità. Il livello informativo dei dati longitudinali È superiore a quello contenuto nei dati cross section in quanto incrementano la possibilità di rilevare la variabilità dell' evento di interesse e di studiarne l'evoluzione temporale. Nel contempo presentano una struttura più complessa rispetto ad altre tipologie di dati, ad esempio, una survey pianificata su più tempi presenta delle caratteristiche peculiari che riguardano: le unità (individui), il periodo temporale (tempi, le rispettive frequenze) e le variabili di interesse (risposta e covariate). Non è inoltre scontato che tutte le unità possano/vogliano fornire i dati nelle indagini successive alla prima, per cui i dati potrebbero risultare incompleti.

Nel modello di transizione per dati categoriali le variabili $Y^{(1)}, \dots, Y^{(T)}$ che definiscono il processo stocastico \mathbf{Y} sono variabili risposta che presentano c categorie dove

$c = 0, \dots, c - 1$. In base alle proprietà esposte in precedenza il modello grafico (grafo direzionato aciclico) risulta composto da archi che uniscono solo i nodi (le variabili) immediatamente seguenti in senso temporale come nel grafo successivo dove gli archi rappresentano le probabilità condizionate che definiscono la catena. Rispetto alle ipotesi di

$$Y^{(1)} \longrightarrow Y^{(2)} \longrightarrow \dots \longrightarrow Y^{(T)}$$

Markovianità del primo ordine ed omogeneità della catena la distribuzione congiunta si determina con le probabilità iniziali di ogni stato (categoria) e di transizione tra i vari stati

$$f(\mathbf{y}) = \boldsymbol{\pi}_{(y^{(0)})} \boldsymbol{\Pi}_{y^{(t)}|y^{(t-1)}}$$

dove nel seguito le probabilità iniziali $\pi_{(y^{(0)})}$ si denotano come π_y e le probabilità di transizione $\pi_{y^{(t)}|y^{(t-1)}}$ come $\pi_{y|\bar{y}}$. I parametri nel modello di transizione sono pertanto relativi alle probabilità di iniziare in ogni stato e di transitare di stato in stato. I vincoli imposti su tali parametri sono che la somma delle probabilità iniziali sia pari a 1 e che la matrice di transizione sia una matrice stocastica. Il modello presenta pertanto un numero di parametri liberi (degrees of freedom) pari alla seguente somma che coinvolge il numero di categorie della variabile risposta c

$$c - 1 + c(c - 1)$$

in cui $c - 1$ È il numero dei parametri riferiti alle probabilità iniziali mentre i restanti sono quelli della matrice di transizione omogenea nel tempo.

Nell'ambito della stima di massima verosimiglianza esistono le stime in forma chiusa per i parametri. Disponendo di un campione con n unit? e considerando le configurazioni di risposta osservate (pattern di risposta) \mathbf{y} e le rispettive frequenze $n_{\mathbf{y}}$ la funzione di log-verosimiglianza risulta pari a

$$\ell(\boldsymbol{\theta}) = \sum_{\mathbf{y}} n_{\mathbf{y}} \log f(\mathbf{y})$$

dove $\boldsymbol{\theta}$ raccoglie tutti i parametri del modello. La verosimiglianza precedente si può

scrivere come somma che coinvolge i rispettivi parametri nel modo seguente

$$\ell(\boldsymbol{\theta}) = \sum_{y=0}^{c-1} a_y^{(1)} \log \pi_y + \sum_{\bar{y}=0}^{c-1} \sum_{y=0}^{c-1} a_{\bar{y}y} \log \pi_{y|\bar{y}}$$

dove con $a_y^{(1)}$ si indica il numero di unit? che hanno risposto y nel periodo iniziale mentre con $a_{\bar{y}y}$ si indica il numero di unit? che hanno risposto \bar{y} al tempo $t - 1$ e y al tempo t . Le soluzioni per l'equazione precedente sono le seguenti

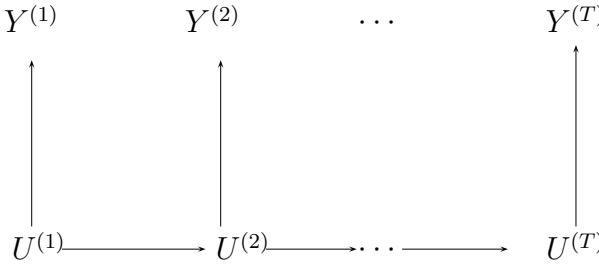
$$\hat{\pi}_y = \frac{a_y^{(1)}}{n} \quad \hat{\pi}_{y|\bar{y}} = \frac{a_{\bar{y}y}}{a_{\bar{y}}}$$

che per $\bar{y}, y = 0, \dots, c - 1$ sono le stime di massima verosimiglianza dei parametri del modello.

4 Modello latente di Markov

Il modello che suppone la struttura di dipendenza Markoviana su delle variabili non osservate o latenti si chiama modello *latente* di Markov (hidden Markov model). In questo modello è possibile ipotizzare che la dipendenza osservata ad istanti temporali successivi sia dovuta a componenti che la generano denotate con $U^{(t)}$ (Unobserved) nel seguito, e tali che rendono le variabili risposta indipendenti. In particolare il modello è utile nel caso in cui le misurazioni disponibili (variabili osservate) rappresentano in modo indiretto l'oggetto dello studio. Ad esempio, la soddisfazione è un concetto latente che potrebbe essere misurato attraverso le risposte fornite dagli individui tramite un questionario. In questo caso gli stati del processo latente rappresentano i diversi livelli di soddisfazione/insoddisfazione nella popolazione. La struttura di dipendenza/indipendenza è raffigurata con il seguente grafo aciclico (Directed Acyclic Graph, DAG). Si associa pertanto una componente non osservata ad ogni componente del processo $Y^{(t)}$ in ogni istante temporale $t, t = 1, \dots, T$ e si assume il processo latente \boldsymbol{U} con struttura di Markov del primo ordine.

Rispetto all'esempio sulla soddisfazione notiamo che quando è nota la configurazione di risposta rispetto al processo latente la risposta fornita da un individuo ad una domanda del questionario non è più utile per fare inferenza sulla risposta successiva. La variabile latente rappresenta la soddisfazione (variabile risposta) mentre le risposte alle domande



sono proxies della variabile latente. Si dice anche la variabile risposta è misurata con errori (measurement errors), per maggiori dettagli si rimanda a Wiggins (1955).

Il modello precedente come altri modelli con variabili latenti si compone di due parti definite: *modello di misura* che coinvolge la parametrizzazione delle variabili risposta \mathbf{Y} condizionate alla corrispondente variabile non osservata $U^{(t)}$ e il *modello latente* che coinvolge i parametri delle variabili non osservate \mathbf{U} ovvero le probabilità iniziali e di transizione della catena. Per una trattazione più dettagliata si veda la Sezione 3.2 del libro Bartolucci, Farcomeni, Pennoni (2012).

I parametri del modello di misura sono denotati come segue in quanto si tratta di probabilità condizionate

$$\phi_{y|u} = f_{Y^{(t)}|U^{(t)}}(y|u), \quad t = 1, \dots, T, \quad u = 1, \dots, k, \quad y = 0, \dots, c - 1,$$

si noti che, rispetto al modello di Markov illustrato nella sezione precedente, il processo di Markov che è definito sulle variabili non osservate e come nel modello di Markov precedente è un processo a stati discreti, con numero di stati pari a k . La variabile latente È infatti assunta categoriale con $u, u = 1, \dots, k$ stati latenti. Il numero degli stati latenti in questo caso si interpreta anche come possibili gruppi di popolazioni sottostanti che caratterizzano il fenomeno osservato. Come per i modelli miscuglio finiti illustrati nelle dispense del primo modulo in cui occorreva stimare il numero delle componenti anche in questo caso nelle situazioni di interesse il numero k può essere determinato in base ai criteri d'informazione.

I parametri del modello latente sono pertanto identici a quelli della sezione precedente tuttavia si denotano nel modo seguente in quanto riferiti adesso alla componente non

osservata: le probabilità iniziali come

$$\pi_u = f_{U^{(1)}}(u), \quad u = 1, \dots, k,$$

e le probabilità di transizione (sotto l'ipotesi di catena omogenea) come

$$\pi_{u|\bar{u}} = f_{U^{(t)}|U^{(t-1)}}(u|\bar{u}), \quad t = 2, \dots, T, \quad \bar{u}, u = 1, \dots, k;$$

in cui u è riferita a $U^{(t)}$, mentre \bar{u} a $U^{(t-1)}$. Pertanto il numero dei parametri liberi del modello coinvolge le seguenti quantità che sono riferite al modello misura (c categorie della variabile risposta) ed al processo latente con k stati

$$k(c-1) + (k-1) + k(k-1)$$

La distribuzione di probabilità del processo latente $f_{\mathbf{U}}(\mathbf{u})$ definisce la probabilità *a priori* per il vettore \mathbf{U} . La distribuzione $f_{\mathbf{U}}(\mathbf{u})$ si determina

$$f_{\mathbf{U}}(\mathbf{u}) = \pi_{u^{(1)}} \pi_{u^{(t)}|u^{(t-1)}}.$$

La distribuzione delle variabili osservate $f_{\mathbf{Y}}(\mathbf{y})$ in base alle assunzioni precedenti (variabili non osservate discrete con struttura di Markov del primo ordine e di indipendenza locale) si ricostruisce come

$$f_{\mathbf{Y}}(\mathbf{y}) = \sum_{\mathbf{u}} f_{\mathbf{Y}|\mathbf{U}}(\mathbf{y}|\mathbf{u}) f_{\mathbf{U}}(\mathbf{u})$$

in cui nel contesto in esame $f_{\mathbf{Y}|\mathbf{U}}(\mathbf{y}|\mathbf{u})$ indica la distribuzione di probabilità condizionate ovvero distribuzione condizionata di \mathbf{Y} rispetto a \mathbf{U} è riferita al modello di misura e risulta

$$f_{\mathbf{Y}|\mathbf{U}}(\mathbf{y}|\mathbf{u}) = \prod_{t=1}^T \phi_{y^{(t)}|u^{(t)}},$$

per ogni realizzazione \mathbf{y} di \mathbf{Y} . Mentre la distribuzione delle variabili osservate

$$f_{\mathbf{Y}}(\mathbf{y}) = \sum_{\mathbf{u}} \phi_{y^{(1)}|u^{(1)}} \cdots \phi_{y^{(T)}|u^{(T)}} \pi_{u^{(1)}} \pi_{u^{(2)}|u^{(1)}} \cdots \pi_{u^{(T)}|u^{(T-1)}}$$

Applicando la regola di Bayes si ottiene la distribuzione di probabilità *a posteriori* per

le variabili non osservate nel modo seguente

$$f_{U|Y}(\mathbf{u}|\mathbf{y}) = \frac{f_{Y|U}(\mathbf{y}|\mathbf{u})f_U(\mathbf{u})}{f_Y(\mathbf{y})}$$

in base alla quale È possibile assegnare ad ogni unità un rispettivo stato latente per ogni periodo considerato.

Analogamente al modello illustrato in precedenza disponendo di un campione di n unità indipendenti $i = 1, \dots, n$ la funzione di log-verosimiglianza si determina come

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^n \log f_Y(\mathbf{y}_i),$$

con $\boldsymbol{\theta}$ che rappresenta il vettore dei parametri. Tuttavia analogamente ai modelli miscuglio illustrati nella parte 1 delle dispense occorre massimizzare la log-verosimiglianza dei *dati completi*, ovvero quella che coinvolge anche le variabili latenti. Questa si esprime come

$$\ell^*(\boldsymbol{\theta}) = \sum_{i=1}^n \log f_{U,Y}(u_i, y_i)$$

e si scomponete nella somma seguente

$$\ell^*(\boldsymbol{\theta}) = \sum_{i=1}^n \log f_{Y|U}(\mathbf{y}_i|\mathbf{u}_i) + \sum_{i=1}^n \log f_U(\mathbf{u}_i).$$

Nel caso specifico della parametrizzazione adottata in precedenza quando si hanno k stati discreti la log-verosimiglianza completa rispetto al pattern di frequenze si scrive nel modo seguente

$$\ell^*(\boldsymbol{\theta}) = \sum_{u=1}^k \sum_{y=0}^{c-1} a_{uy} \log \phi_{y|u} + \sum_{u=1}^k b_u^{(1)} \log \pi_u + \sum_{\bar{u}=1}^k \sum_{u=1}^k b_{\bar{u}u} \log \pi_{u|\bar{u}},$$

in cui a_{uy} È la frequenza di risposta [numero di unit?](#) che si trovano nello stato latente u hanno configurazione di riposta y per la variabile $Y^{(t)}$, mentre b_u rappresenta la frequenza dello stato latente u , e $b_{\bar{u}u}$ la frequenza delle transizioni dallo stato latente \bar{u} allo stato latente u .

Utilizzando l'algoritmo Expectation Maximization (Baum *et al.*, 1970, Dempster *et al.*,

1977, Welch, 2003) illustrato nelle dispense precedenti si massimizza la verosimiglianza dei dati completi in modo iterativo con il passo E ed M:

- **E-step:** calcola il valore atteso condizionato di $\ell^*(\boldsymbol{\theta})$ ai dati osservati e ad un valore iniziale per i parametri. I valori attesi si denotano con \hat{a}_{uy} , \hat{b}_u , and $\hat{b}_{\bar{u}u}$ che si calcolano in base alla probabilità a posteriori delle variabili latenti;
- **M-step:** si massimizza la log-verosimiglianza dei dati completi in cui ogni valore viene sostituito con quello atteso calcolato al passo E. Si utilizzano le formule esplicite se disponibili oppure si usa l'algoritmo Newton-Raphson.

La convergenza si valuta come nei modelli mistura in base alla verosimiglianza relativa che può essere espressa come segue:

$$\frac{\ell(\boldsymbol{\theta})^{(s)} - \ell(\boldsymbol{\theta})^{(s-1)}}{|\ell(\boldsymbol{\theta})^{(s)}|} < \epsilon,$$

con $\boldsymbol{\theta}^{(s)}$ si indica il vettore dei parametri stimati che si ottiene alla fine dell'iterazione s -esima al passo M dell'algoritmo e ϵ rappresenta il livello di tolleranza scelto per la convergenza (per esempio 10^{-8}). I principali due criteri di informazione AIC (Akaike, 1973) e BIC (Schwarz, 1978) sono utili sia quando è di interesse valutare l'ipotesi di omogeneità del processo latente sia quando si intende scegliere il numero degli stati latenti k . In tal caso come per le componenti del modello mistura la selezione avviene scegliendo il modello che presenta il valore più basso per l'indice di adattamento (cf. dispense parte I, per come si determinano).

4.1 Modello latente di Markov con covariate

In questo sezione si accenna brevemente al modello latente di Markov in cui le covariate (variabili esplicative) hanno influenza su modello di misura, per maggiori dettagli si rimanda al Capitolo 5 del libro Bartolucci, Farcomeni, Pennoni (2012). Quando le covariate influenzano direttamente le variabili risposta $Y^{(t)}$, le variabili latenti $U^{(t)}$ hanno il ruolo di spiegare l'eterogeneità non osservata ovvero l'eterogeneità che rimane tra le unità rispetto a quella spiegata dalle covariate disponibili per l'analisi. Il vettore delle covariate si denota con $\mathbf{X}^{(t)}$ per indicare che queste possono essere diverse a seconda del periodo temporale

$t = 1, \dots, T$. Il modello di misura si denota includendo le covariate

$$\phi_{y|u,\mathbf{x}}^{(t)} = f_{Y^{(t)}|U^{(t)},\mathbf{X}^{(t)}}(y|u,\mathbf{x}), \quad t = 1, \dots, T, u = 1, \dots, k, y = 0, \dots, c - 1.$$

Le assunzioni di indipendenza locale e di dipendenza del primo ordine sono formulate condizionatamente al vettore delle covariate osservate.

Si utilizza la parametrizzazione che si adotta nel modello lineare generalizzato. Nel seguito si considera solo la seguente parametrizzazione formulata per variabili risposta ordinali con $c - 1$ categorie. Si tratta di una formulazione del modello di misura con logits che coinvolgono tutte le categorie della variabile risposta (logit cumulati). Si confrontano le categorie di ordine inferiore (denominatore) con quelle di ordine maggiore (numeratore) e si parametrizza il modello di misura nel modo seguente

$$\log \frac{p(Y^{(t)} \geq y|U^{(t)}, \mathbf{X}^{(t)})}{p(Y^{(t)} < y|U^{(t)}, \mathbf{X}^{(t)})} = \log \frac{\phi_{y|u\mathbf{x}}^{(t)} + \dots + \phi_{c-1|u\mathbf{x}}^{(t)}}{\phi_{0|u\mathbf{x}}^{(t)} + \dots + \phi_{y-1|u\mathbf{x}}^{(t)}} = \mu_y + \alpha_u + \mathbf{x}'\boldsymbol{\beta}, \quad (1)$$

per $t = 1, \dots, T$, $u = 1, \dots, k$, and $y = 1, \dots, c - 1$. Si noti che il parametro α_u è specifico per ogni stato latente u mentre μ_u è un intercetta specifica di ogni categoria della variabile risposta, il vettore $\boldsymbol{\beta}'$ rappresenta il vettore dei parametri riferiti alle covariate. I parametri del modello latente sono gli stessi della formulazione precedente sotto l'ipotesi di omogeneità nel tempo del processo stocastico sottostante. Benchè simile al modello ad effetti casuali il modello in (1) assume che le covariate possano avere una specifica dinamica variabile nel tempo.

5 Algoritmo Metropolis-Hastings

Questo algoritmo è stato proposto da Metropolis et al. (1953, *Journal of chemical physics*) inizialmente per un problema bidimensionale con $n = 10$. È stato in seguito esteso da Hastings (1970, *Biometrika*). Si basa sul fatto che è possibile utilizzare un processo stocastico di Markov in cui la matrice di transizione è omogenea nel tempo ed in cui ogni stato può essere raggiunto da tutti gli altri in un numero finito di passi senza che vi siano stati assorbenti. Le traiettorie generate da questa catena dopo molte iterazioni possono essere assunte come determinazioni della distribuzione di equilibrio.

Per descrivere la distribuzione a posteriori del parametro si generano dei valori possibili

θ' tali da avere una solida distribuzione di densità (probabilità) per ogni punto del supporto del parametro da stimare θ . Questa quantità viene anche chiamata kernel di transizione della catena ovvero definisce la distribuzione condizionata che determina le transizioni tra gli stati. Si sceglie una distribuzione di campionamento (jumping distribution) tale che possa generare dei passaggi di stato con diversa verosimiglianza sotto l'ipotesi di simmetria (reversibilità) che si può esprimere nel modo seguente

$$q(\theta|\theta') = q(\theta'|\theta)$$

in modo che la catena risulti irriducibile. Il kernel di transizione $q(\theta'|\theta)$ deve essere definito nel supporto della distribuzione a posteriori. Hasting (1970) ha proposto una variante in cui l'assunzione di simmetria non è più richiesta in cui occorre un termine che permette una compensazione. Peskun (1970) ha generalizzato il metodo considerando come distribuzione di campionamento la Gaussiana e la Poisson.

Si considera una catena di Markov $\mathbf{X}^{(t)} = (X_1^{(1)}, \dots, X_p^{(t)})$ con spazio degli stati discreto o continuo. Il valore iniziale in $t = 0$ $\mathbf{X}^{(0)} = \mathbf{x}_0$ viene stabilito per ottenere la prima realizzazione della catena campionando da una generica distribuzione g tale che $f(\mathbf{x}^{(0)}) > 0$ con f che denota la distribuzione di interesse (distribuzione a posteriori). Ad ogni iterazione $t > 0$:

- si campiona un possibile valore (candidato) \mathbf{X}^* da una distribuzione $g(.|\mathbf{x}^{(t)})$;
- si calcola il rapporto tra

$$R(\mathbf{x}^t, \mathbf{X}^*) = \frac{f(\mathbf{x}^*)q(\mathbf{x}^*|\mathbf{x}^{t-1})}{f(\mathbf{x}^t)q(\mathbf{x}^t|\mathbf{x}^{t-1})}$$

dove $R(\mathbf{x}^{(t)}, \mathbf{X}^*)$ è detto rapporto Metropolis-Hastings che è sempre definito per costruzione

- si stabilisce se il valore campionato può essere accettato come valore per $\mathbf{X}^{(t+1)}$

$$\mathbf{X}^{(t+1)} = \mathbf{X}^*$$

in base alla probabilità definita da $\min\{R(\mathbf{x}^{(t)}, \mathbf{X}^*), 1\}$. La regola decisionale porta ad accettare $\mathbf{X}^{(t+1)} = \mathbf{x}^*$ oppure a rifiutare il valore candidato \mathbf{x}^* e pertanto $\mathbf{X}^{(t+1)} = \mathbf{x}^{(t)}$ e si incrementa t per ripetere la procedura.

Esempio

Sia $\mathbf{s} = (s_1, s_2, \dots, s_M)$ la distribuzione stazionaria della catena e P la matrice di transizione di una catena di Markov definita sullo stesso spazio degli stati ma con distribuzione stazionaria diversa da quella di \mathbf{s} . Occorre modificare P in modo che questa assuma la distribuzione stazionaria di riferimento. Il procedimento è il seguente:

$$a_{ij} = \min\left(\frac{s_j p_{ji}}{s_i p_{ij}}, 1\right)$$

si sceglie un evento binario avente probabilità a_{ij} per l'evento successo che se si realizza determina il passaggio (jump) verso lo stato j producendo $X_{n+1} = j$ altrimenti si resta nella stessa posizione tale che $X_{n+1} = i$. Si dimostra che se $s_i q_{ij} = s_j q_{ij}$ vale la condizione di reversibilità. La scelta della distribuzione è importante perché determina anche la velocità di convergenza alla distribuzione stazionaria. La catena deve esplorare l'intero spazio degli stati.

Nel caso di famiglia Gaussiana (cf. *iii*) nelle dispense) $Y|\theta \sim N(\theta, \sigma^2)$ con σ^2 noto, con la prior per $\theta \sim N(\mu, \tau^2)$ dove entrambe sono costanti note allora la distribuzione a posteriori è data da $f(\theta|y) \propto f(y|\theta)f(\theta)$ e si determina in modo analitico.

Tuttavia si può impiegare l'algoritmo (a scopo illustrativo) per costruire una catena di Markov la cui distribuzione stazionaria è proprio $f(\theta|y)$. Essendo lo spazio degli stati continuo, si genera una realizzazione da $\epsilon \sim N(0, d^2)$ dove d è una costante scelta a piacere e si fissa

$$\theta' = \theta^0 + \epsilon_n$$

il rapporto di Metropolis-Hastings

$$R(\boldsymbol{\theta}^t, \boldsymbol{\theta}^*) = \frac{f(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{t-1})}{f(\boldsymbol{\theta}^t)q(\boldsymbol{\theta}^t|\boldsymbol{\theta}^{t-1})}$$

diventa

$$= \frac{\exp(-\frac{1}{2\sigma^2}(x - \theta')^2) \exp(-\frac{1}{2\tau^2}(\theta' - \mu)^2)}{\exp(-\frac{1}{2\sigma^2}(x - \theta^0)^2) \exp(-\frac{1}{2\tau^2}(\theta^0 - \mu)^2)}.$$

Il valore di questo rapporto viene utilizzato per assegnare la probabilità all'evento successo: la probabilità di successo si determinata in base al $\min\{R(\mathbf{x}^{(t)}, \mathbf{X}^*), 1\}$ e si campiona dalla distribuzione di Bernoulli. Se il risultato è l'evento successo si decide di

accettare il candidato θ' come valore realizzato della distribuzione a posteriori altrimenti si rifiuta il candidato e si procede campionando un nuovo candidato. \diamond

La distribuzione da cui vengono originati i valori candidati comporta che la catena abbia le proprietà di irriducibilità e di aperiodicità essenziali per la convergenza dell'algoritmo. La distribuzione marginale di $\mathbf{X}^{(t+1)}$ è la distribuzione a posteriori f di interesse ed è la distribuzione stazionaria della catena. Nell'esempio precedente occorre scegliere la varianza d della distribuzione da cui si genera il valore candidato. Se d è grande rispetto a quello della distribuzione a posteriori questo implica un elevato tasso di rifiuto per i valori candidati. Di conseguenza il numero delle iterazioni necessarie per poter esplorare tutto lo spazio degli stati della distribuzione a posteriori si incrementa.

Per determinare se l'algoritmo ha eseguito un numero di passi sufficiente in modo tale che il risultato delle realizzazioni sia prossimo alla distribuzione di equilibrio si considerano le seguenti procedure:

- si valuta la distanza necessaria tra le osservazioni affinché queste possano essere considerate indipendenti;
- si valuta il tasso di accettazione: un tasso di accettazione ottimale che rende plausibile la distribuzione da cui si generano i valori x' è compreso tra il 25% ed il 45% (meglio se 45%);
- si ripete la procedura partendo da diversi valori iniziali creando catene multiple;
- si utilizzano delle analisi grafiche per i valori generati: *sample path* che consiste nel disegnare i valori del parametro ad ogni iterazione (trace plot) e si verifica se i valori si concentrano quasi subito verso un valore medio;
- si produce il grafico della funzione di autocorrelazione dei valori a diversi lag. Ci si attende una decadimento della correlazione tra le realizzazioni abbastanza rapido.

Dato che per le proprietà della catena di Markov la distribuzione target si ottiene come limite si parla di *burn in* quando si eliminano le prime realizzazioni della catena. Il periodo di burn in (ovvero quante realizzazioni devono essere eliminate) dipende da alcuni fattori tra cui la lunghezza delle realizzazioni. Una regola stabilisce di eliminare la prima metà delle realizzazioni. In altri studi si rileva che non è necessario eliminare osservazioni

(ovvero che non serve definire il periodo di burn in). WinBUGS (Bayesian Inference using Gibbs sampling) è uno dei software creati appositamente per questo tipo di analisi.

6 Richiami sulle catene di Markov

Per un processo stocastico con spazio degli stati finito la proprietà di ricorrenza della catena di Markov permette di soddisfare le proprietà asintotiche. Nel caso in cui lo spazio degli stati è illimitatamente continuo occorre che la catena di Markov soddisfi la ricorrenza secondo la definizione di Harris (1956) in modo che la probabilità di visita di un certo sottospazio A infinite volte sia pari a uno. La definizione è la seguente: l'insieme A è detto ricorrente secondo Harris se

$$P_x(\eta_B \rightarrow \infty) = 1$$

$\forall x \in A$ con $B \in A$ e η_B identifica il numero di passaggi in A .

Se la distribuzione invariante della catena di un processo definito su di uno spazio di stati continuo non soddisfa la richiesta precedente esiste la probabilità che la catena si arresti in un intervallo dello spazio degli stati che non è prossimo al punto definito di convergenza rispetto alla posizione di equilibrio partendo da qualsiasi punto iniziale.

- Quando il processo raggiunge la distribuzione stazionaria, si dice che il processo è per sempre in equilibrio o oscilla nel sottospazio definito dalla distribuzione marginale del processo.
- La proprietà di ricorrenza permette di definire delle restrizioni in modo tale da essere certi che gli stati al limite possano essere visitati infinite volte.
- La stazionarietà permette di assegnare una componente costante alla struttura probabilistica della catena che definisce le posizioni del processo, ed assicura che (eventualmente) la struttura probabilistica rimanga costante.
- Il periodo della catena è la lunghezza di tempo che il processo richiede per ripetere lo stesso ciclo di realizzazioni. Tuttavia non si hanno garanzie su quale tipo di distribuzione si ottiene al limite.
- La proprietà di ergodicità assicura che se la catena ha raggiunto lo stato ergodico i valori campionati possono essere considerati delle determinazioni della distribuzione

a posteriori. Il teorema che definisce l'ergodicità è interpretato nel contesto di un processo stocastico come il teorema centrale del limite dato che asserisce che ogni funzione della distribuzione a posteriori può essere stimata attraverso delle realizzazioni campionarie ottenute da una catena di Markov assunta come ergodica. Infatti con i valori medi campionari si ricavano estimatori consistenti per i parametri. La proprietà di ergodicità permette di assumere l'indipendenza del processo dalle condizioni iniziali in cui si trova la catena (cf. tavole di sopravvivenza demografiche).

La velocità di convergenza può essere geometrica o uniforme. Se consideriamo il teorema centrale del limite per uno stimatore di plug-in avente una varianza finita la distribuzione limite è Gaussiana:

$$\sqrt{n} \frac{\hat{h}(\theta_i) - h(\theta)}{\sqrt{Var(\hat{h}(\theta_i))}} \longrightarrow N(0, 1)$$

per $n \rightarrow \infty$ dove $\hat{h}(\theta_i) = \frac{1}{n} \sum_{j=i+1}^{i+n} h(\theta_j)$ è la media.

Riferimenti

- Akaike, H. (1973). Information Theory and an Extension of the maximum likelihood principle. *Second International Symposium on Information Theory*, (Petrov, B.N. and Csaki, F. Eds), 267-281.
- Bartolucci, F., Pandolfi S., Pennoni F. **LMeST** (2017): An R package for latent Markov models for longitudinal categorical data. *Journal of Statistical Software*, **81**, 1- 38. <https://www.jstatsoft.org/article/view/v081i04>.
- Bartolucci, F., Farcomeni A., Pennoni, F. (2014). Latent Markov models: a review of a general framework for the analysis of longitudinal data with covariates, *with discussion*, *Test*, **23**, 433-465.
- Bartolucci, F., Farcomeni A., Pennoni F. (2014). Rejoinder on: Latent Markov Models: a review of a general framework for the analysis of longitudinal data with covariates, *Test*, **23**, 484-496.

- Bartolucci, F., Farcomeni, A., Pennoni, F. (2013). *Latent Markov Models for Longitudinal Data*, Chapman and Hall/CRC press, Boca Raton.
<https://www.taylorfrancis.com/books/9781466583719>
- Blitzstein J. K., Hwang J. (2015). *Introduction to probability*. CRC press, Boca Raton.
- Baum L., Petrie T., Soules G., Weiss N. (1970). A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *The Annals of Mathematical Statistics*, **41**, 164–171.
- Dempster, A. P. Laird N. M., and Rubin D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.
- Harris, T., (1956). The existence of a stationary measures for certain Markov processes. *In proceedings 3rd Berkeley Symp. Math. Stats. Prob.* Vol. 2, 113-124, University of California, Berkely.
- Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97-109.
- Gelman, S., Jones A. and Meng, X. L. (Eds.). (2011). *Handbook of Markov Chain Monte Carlo*. CRC press.
- Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, **6**, 721-741.
- Gelfand, A. E., and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, **85**, 398-409.
- Durrett R. (1999). *Essential of Stochastic Processes*, Springer-Verlag, New-York.
- Kirkwood J. R. (2015). *Markov Processes*, Chapman and Hall/CRC press, Boca Raton.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, **21**, 1087-1092.

- Schwarz G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.
- Tanner, M. A., and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, **82**, 528-540.
- Welch L. R. (2003) . Hidden Markov models and the Baum-Welch algorithm. IEEE Information Theory Society Newsletter, **53**, 1–13.
- Wiggins L. M. (1955). *Mathematical models for the Analysis of Multiwave Panels*, In: Ph.D. Dissertation, Columbia University, Ann Arbor, MI.

Dispensa Modelli Statistici II Inferenza Bayesiana

Parte III

a.a. 2017-2018

Fulvia Pennoni

Università degli Studi di Milano-Bicocca (IT)

fulvia.pennoni@unimib.it

Utilizzo della procedura proc MCMC di SAS[®]

In questa parte delle dispense si illustrano le principali caratteristiche della procedura Markov Chain Monte Carlo (proc MCMC) di SAS. Le slides sono tratte dal corso BAYESIAN ANALYSIS WITH SAS[®] 2014 SAS INC.

- Nella prima slide di Figure 1 sono elencati i principali statements della procedura proc MCMC: PARAMS per i parametri del modello, PRIOR per i parametri della distribuzione a priori, MODEL per la funzione di log-verosimiglianza e PREDDIST per la distribuzione a posteriori. Si noti dallo schema di impostazione della sintassi che sono possibili molte opzioni anche definite dall'utente.
- La procedura restituisce in output le quantità elencate nella slide di Figure 2: i campioni della distribuzione a posteriori, le principali statistiche della stessa, le misure diagnostiche e le analisi grafiche. Si noti che la procedura permette di richiedere varie statistiche di sintesi per la distribuzione a posteriori.
- Nelle slides di Figure 3 e Figure 4 si illustrano le procedure di campionamento implementate nella procedura. Si noti in Figure 3 che viene utilizzato il cam-

pionamento Gibbs e sono state implementate diversi algoritmi: Metropolis-Hastings (centrato sul valore), Metropolis, Random Walk Metropolis e Independent Metropolis.

- Nella slide di Figura 5 si mostra che la procedura automaticamente esegue il *burn in*: toglie le prime realizzazioni generate dalla catena di Markov per rendere irrilevante l'influenza dei valori iniziali sulla distribuzione a posteriori.
- Le seguenti slide in Figura 6 (good mixing) e 7 (poor mixing) riportano alcune analisi grafiche fornite dall'output della procedura circa le autocorrelazioni tra valori generati e la densità a posteriori stimata. Mostrano le principali differenze di come appare il trace plot (numero di iterazioni vval ori del parametro, il grafico della funzione di autocorrelazione e la stima della densità a posteriori) nel caso in cui la catena produce subito delle osservazioni che risultano indipendenti (good mixing) (in Figura 6) e nel caso in cui non si evince l'indipendenza neanche dopo 12000 iterazioni (in Figura 6). Nella Figura 6 la funzione delle autocorrelazioni non presenta valori elevati diversamente da quella della Figura 7. La densità a posteriori presenta una forma più irregolare nel caso di poor mixing.
- Oltre alle statistiche di adattamento come illustrato in Figura 8 riferite ai criteri AIC e BIC (cf. Dispense Parte Teorica Modelli) la procedura restituisce il valore del criterio di informazione denominato Deviance Information Criterion (Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van der Linde, A. 2002, Bayesian Measures of Model Complexity and Fit, with discussion, *Journal of the Royal Statistical Society, Series B*, 64, 583–616).
- Nella slide di Figura 9 si fornisce gli statement per alcune opzioni della procedura. Si noti che tra le varie impostazioni con l'input **NBI** (number of burn in) è possibile stabilire il numero delle iterazioni che si intendono scartare (altrimenti il numero di default è pari a 2000). E' inoltre possibile richiedere svariati grafici per la diagnosi grafica della distribuzione a posteriori con l'opzione **PLOTS**. La procedura permette di avere varie statistiche di sintesi per la dis-

distribuzione a posteriori.

- Nella slide di Figura 10 si evince che tra le statistiche di sintesi c'è anche l'intervallo di credibilità con la massima densità a posteriori.
- Nella slide di Figura 11 si richiama la sintassi per definire i valori iniziali per i parametri del modello. Questi sono introdotti con **PARAMS** seguito dal nome del parametro e dai valori iniziali per gli stessi. Si noti che è possibile anche non assegnare un valore iniziale.
- Nella slide di Figura 12 la distribuzione iniziale è assegnata con **PRIOR** seguita dal nome del parametro e dalla tilde prima del nome della distribuzione con i rispettivi parametri. Si illustrano 2 statements che permettono di specificare le distribuzioni iniziali dei parametri (Normale e Gamma). Si noti che la specificazione è tale che la prior congiunta è il prodotto di ogni prior sui tre parametri del modello.
- Nella slide di Figura 13 si illustrano le caratteristiche del **model** statement. La forma della funzione di verosimiglianza viene descritta nella parte **MODEL**. Prima del **MODEL** si assegna la relazione per il parametro e poi si scrive il modello ad esempio **y ~ normal**. Nella Figura 14 si illustra la sintassi che specifica il modello.
- Nella Figura 15 si illustra la famiglia coniugata, i parametri e la corrispondente distribuzione a priori.

Esempio applicativo

Il seguente esempio è tratto dal materiale di cui sopra, (i dati sono simili a quelli presenti in Hosmer, D.W. and Lemeshow, S. (2000). *Applied Logistic Regression* 2nd edition, New York: John Wiley & Sons). Si tratta di un modello logistico generalizzato per una variabile risposta dicotomica $y \in \{0, 1\}$

$$Y \sim Bernoulli(p)$$

la funzione di link

$$g(p) = \log \left(\frac{p}{1-p} \right)$$

modellizza i log-odds dell'evento successo come funzione lineare delle covariate.

In generale si utilizza una distribuzione di Gauss per ogni parametro in β (coefficienti della regressione) se non si hanno particolari informazioni a priori si stabilisce che la media a priori è pari a zero e la varianza piuttosto alta (prior non informativa). Le distribuzioni a priori sono assegnate in modo indipendente per ogni parametro.

Nei dati è presente una variabile dicotomica per indicare l'evento basso peso alla nascita (se il peso alla nascita è inferiore di 2500grammi) che rappresenta la variabile risposta binaria (1 basso, 0 alto) del modello lineare generalizzato. Le covariate sono le seguenti:

- consumo di alcolici durante la gravidanza (1 sì);
- familiarità per l'ipertensione (1 sì);
- peso della mamma (unità di misura inglese) rilevato nell'ultimo periodo mestruale;
- precedenti travagli prematuri (0 nessuno; 1 uno; 2 due ...).

Nel seguente si illustra la sintassi di SAS necessaria per stimare il modello quando la prior è assegnata in modo indipendente ad ogni parametro ed è non informativa ($\mu = 0, \sigma^2 = 1$).

- Nella slide di Figura 16 si inseriscono in input valori iniziali dei parametri e si utilizza un modello parametrico per le distribuzioni a priori dei parametri. Per stimare il modello ed ottenere il relativo output si impostano le seguenti opzioni:

riga 40 `diag=all` per richiedere in output tutte le statistiche diagnostiche della procedura;

riga 40 `dic` per richiedere in output il criterio di informazione denominato DIC;

riga 41 `plot(smooth)` per richiedere che venga inserita nei grafici di trace plot anche la curva di tendenza;

riga 41 `seed` per fissare un valore che permette di replicare le analisi;

riga 43 `params` per definire i coefficienti del modello di regressione lineare generalizzato ed impostare tutti i valori iniziali pari a 0;

riga 45 `prior` per assegnare ad ogni parametro una distribuzione iniziale non informativa tale che $\beta_j \sim N(0, 100)$ con $j = 0, \dots, 4$;

riga 47 `p = logistic` per specificare il tipo di logit per il modello lineare generalizzato e le covariate presenti nel modello;

riga 52 `model` per specificare che la variabile risposta basso peso (low) è assunta con distribuzione di Bernoulli di parametro `p`.

- Nel seguito in Figura 17 ed in Figura 18 si mostrano i risultati dell'output per il modello stimato. Si nota dalla Figura 17 che i dati sono relativi a 189 neonati, per il campionamento è stato utilizzato l'algoritmo Metropolis. I grafici diagnosticci in Figura 18 evidenziano tutti che la catena non ha raggiunto la distribuzione di equilibrio. Occorre specificare un numero più elevato di iterazioni (ad esempio 400000) e modificare anche il periodo di burn in.
- In Figura 19 si impostano le opzioni aggiuntive per stimare di nuovo il modello in modo da utilizzare anche l'algoritmo quasi-Newton per la stima della matrice di covarianza iniziale. Nello statement precedente:

riga 71 `propcov=quanew` per richiedere l'algoritmo quasi-Newton;
 riga 72 `nbi=5000` permette di togliere le prime 5000 realizzazioni (burn in period);
 riga 73 `ntu=5000` permette di stabilire il numero di campioni da utilizzare per la ricerca di una distribuzione adeguata per l'algoritmo Metropolis. All'inizio della procedura di stima c'è la fase definita *tuning phase*;
 riga 74 `nmc=400000` definisce il numero di campioni che occorre simulare dalla catena di Markov;
 riga 75 `thin=10` specifica che una ogni 10 realizzazioni devono essere considerate per l'output finale in modo da ridurre l'autocorrelazione;
 riga 76 `mchistory=brief` permette di ottenere delle statistiche aggiuntive per monitorare le realizzazioni delle catene di Markov.

- In Figura 20 viene illustrato l'output della tuning history che permette di valutare il tasso di accettazione (RWM) dell'algoritmo Metropolis. Come riportato nella parte teorica delle dispense questo deve essere compreso tra 20% e 45% ed è pari a circa 0.30 pertanto si considera accettabile. Circa il 70% dei valori campionari proposti vengono rifiutati.
- In Figura 21 vengono riportate le stime dei parametri e la rispettiva regione riferita alla massima densità. $\hat{\beta}_1$ e $\hat{\beta}_2$ sono entrambi positivi (riferiti al consumo di alcool e alla predisposizione per l'ipertensione) e sono le covariate che maggiormente hanno influenza sul basso peso del neonato. Per coloro che consumano alcool l'odds di avere un neonato sottopeso è circa 2 volte l'odds di coloro che non consumano alcool. Per coloro che hanno familiarità per l'ipertensione è l'odds è circa 7 volte quello di chi non presenta questa caratteristica.
- Dai grafici di diagnosi per due parametri riportati in Figura 22 e 23 non si evidenziano criticità della distribuzione di equilibrio della catena. Non si notano comportamenti anomali del trace plot o autocorrelazioni elevate. I valori

del trace plot sono centrati rispetto al valore medio e la variabilità sembra costante.

Slides

Typical PROC Call is a Mixture of Statements and DATA Step Language

PROC MCMC *options*;

PARMS; define parameters.

PRIOR; declare prior distributions

Programming statements; } define log-likelihood function
 MODEL;

PREDIST; posterior prediction

RANDOM; random effects

UDS; User-Defined Sampler

run;

Figure 1:

What Does the Procedure Produce

- samples from the posterior
- posterior statistics (mean, s.d., HPD, etc)
- convergence diagnostics (ESS, Geweke, MCSE, etc)
- graphical display (trace plot, ACF plot, KDE plot)



Figure 2:

Sampling Methods in SAS

Gibbs: Proposal changed to match the posterior conditional distributions.	Metropolis-Hastings: Asymmetric proposal distribution centered on current value.
ARMS: Uses an envelope function and a squeezing function to arrive at posterior samples.	Metropolis: Proposal changed to symmetric centered on current value.
Gamerman: Proposal distribution includes iteration of the iterative weighted least squares (IWLS) algorithm.	Random Walk Metropolis: Proposal no longer centered on current value but added to it.
	Independent Metropolis: Proposal has no dependence on the current value in any way.

Figure 3:

Sampling Algorithm Hierarchy

	Continuous Parameters	Discrete Parameters
When Applicable	Conjugate Direct	Conjugate Direct Inverse CDF
All Others	RWM RWM-t HMC NUTS slice	Discrete RWM Geometric RWM

Figure 4:

Burn-In and Thinning

- *Burn-in* refers to the practice of discarding an initial portion of a Markov chain sample so that the effect of the initial values on the posterior inference is minimized.
- *Thinning* refers to the practice of keeping every k^{th} simulated draw from each sequence in order to reduce sample autocorrelations.
- Autocorrelations do not lead to biased Monte Carlo estimates, but rather it is an indicator of poor sampling efficiency.

Figure 5:

Diagnostic Plots – Good Mixing

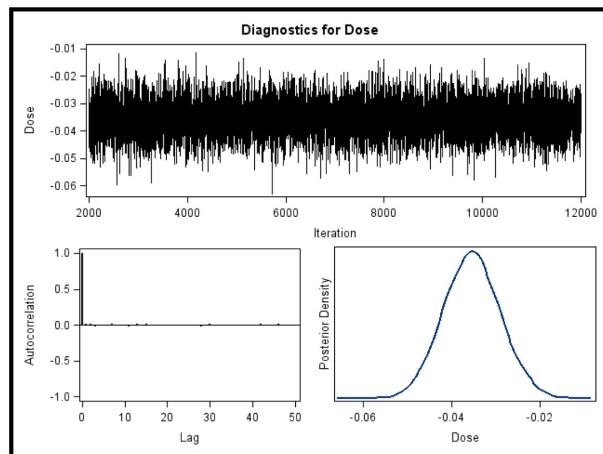


Figure 6:

Diagnostic Plots – Poor Mixing

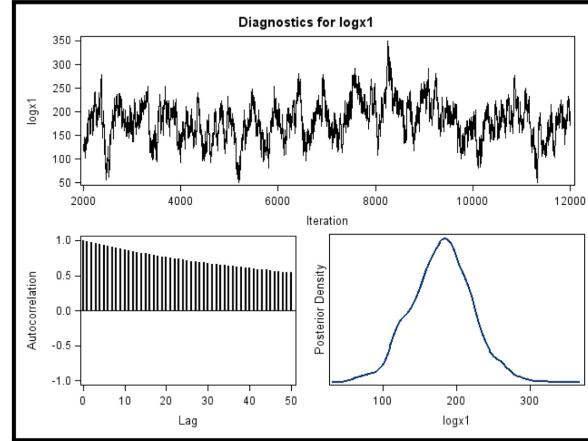


Figure 7:

Deviance Information Criterion (DIC)

- *Deviance Information Criterion (DIC)* is a Bayesian alternative to AIC and BIC.
- It is a statistic where the smaller value indicates a better fit to the data set.
- DIC can be applied to non-nested models and models that have random effects.

Figure 8:

PROC MCMC Statement Options

Option	Description
DATA=	name of the input data set
OUTPOST=	name of the output data set for posterior samples
NBI=	number of burn-in iterations
NMC=	number of MCMC iterations
THIN=	thinning of the Markov chain
SEED=	random number generator seed
STATISTICS=	posterior statistics
DIAGNOSTICS=	convergence diagnostics
PLOTS=	diagnostics plots
DIC	computes deviance information criterion (DIC)

Figure 9:

Posterior Summaries

The posterior summaries include the following:

- Posterior mean, standard deviation, and percentiles
- Equal-tail and highest posterior density intervals
- Covariance and correlation matrices
- Deviance information criterion (DIC)

Figure 10:

PARMS Statement Examples

```
parms alpha 0 beta 1
```

Declares α and β to be model parameters and assigns initial value of 0 to α and 1 to β .

```
parms alpha 0 beta;
```

Assigns initial value of 0 to α and leaves β uninitialized.

```
parms (alpha beta) 1;
```

Assigns 1 as initial values to both α and β .

Figure 11:

PRIOR Statement Example

```
prior beta0 beta1 ~ normal(mean=0,  
                           var=1e6);  
prior sigma2 ~ igamma(shape=2.001,  
                       scale=1.001);
```

This code specifies the following joint prior distribution:

$$\pi(\beta_0, \beta_1, \sigma^2) = \pi(\beta_0) * \pi(\beta_1) * \pi(\sigma^2)$$

Figure 12:

MODEL Statement

- The MODEL statement is used to specify the conditional distribution of the data given the parameters (the likelihood function).
- You must specify a single dependent variable or a list of dependent variables, a tilde, and a distribution with its arguments.
- The dependent variables can be either variables from the data set or functions of variables in the program.
- Multiple MODEL statements are allowed for defining models with multiple independent components.

Figure 13:

MODEL Statement Examples

```
mu=beta0 + beta1*x1;
model y ~ normal(mu,var=sigma2);
```

This code specifies $f(y_i | \mu_i, \sigma^2) = \phi(y_i | \mu_i, \sigma^2)$

$$\mu_i = \beta_0 + \beta_1 X_i$$

```
w=log(y);
model w ~ normal(alpha,var=1);
```

This code specifies $f(\log(y_i) | \alpha, 1) = \phi(\log(y_i) | \alpha, 1)$

Figure 14:

Conjugate Pairs

Family	Parameter	Prior
Normal with known μ	σ^2	Inverse gamma family
Normal with known μ	τ	gamma family
Normal with known σ^2 , σ , or τ	μ	normal
Multivariate normal with known Σ	μ	multivariate normal
Multivariate normal with known μ	Σ	Inverse Wishart
Multinomial	P	Dirichlet
Binomial/binary	ρ	beta
Poisson	λ	gamma family

Figure 15:

```

38 /* proc MCMC */
39
40 proc mcmc data=work.lowbw diag=all dic
41           plots(smooth)=all seed=27513;
42
43 parms (beta0 beta1 beta2 beta3 beta4) 0;
44
45 prior beta: ~ normal(0, var=100);
46
47 p = logistic(beta0
48             +beta1*alcohol
49             +beta2*hist_hyp
50             +beta3*mother_wt
51             +beta4*prev_preterm);
52 model low ~ binary(p);
53
54 title "Analisi Bayesiana per il peso dei neonati ";
55
56 run;
57
58
59 /*Per chiudere il file pdf aperto per l'output*/
60 /*ods pdf close;*/
61
62 ods pdf close;

```

Figure 16:

Analisi Bayesiana per il peso dei neonati

La procedura MCMC

Numero osservazioni lette	189
Numero osservazioni usate	189

Parametri				
Blocco	Parametro	Metodo di campionamento	Valore iniziale	Distribuzione a priori
1	beta0	N-Metropolis	0	normal(0, var=100)
	beta1		0	normal(0, var=100)
	beta2		0	normal(0, var=100)
	beta3		0	normal(0, var=100)
	beta4		0	normal(0, var=100)

Figure 17:

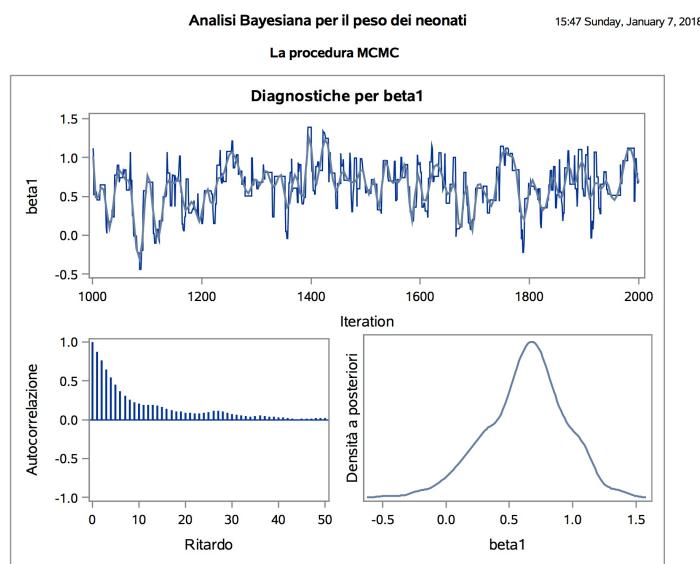


Figure 18:

```

65 /* proc MCMC 2 */
66
67 proc mcmc data=work.lowbw
68         outpost=birthout
69         diag=all
70         dic
71         propcov=quanew
72         nbi=5000
73         ntu=5000
74         nmc=400000
75         thin=10
76         mchistory=brief
77         plots(smooth)=all
78         seed=27513
79         stats=all;
80 parms (beta0 beta1 beta2 beta3 beta4) 0;
81
82 prior beta: ~ normal(0, var=100);
83
84 p = logistic(beta0
85                 +beta1*alcohol
86                 +beta2*hist_hyp
87                 +beta3*mother_wt
88                 +beta4*prev_pretrm);
89
90 model low ~ binary(p);
91
92 title "Analisi Bayesiana 2 per il peso dei neonati";
93
94 run;

```

Figure 19:

Analisi Bayesiana 2 per il peso dei neonati

15:47 Sunday, January 7, 2018

La procedura MCMC

Numero osservazioni lette	189
Numero osservazioni usate	189

Parametri				
Blocco	Parametro	Metodo di campionamento	Valore iniziale	Distribuzione a priori
1	beta0	N-Metropolis	0	normal(0, var=100)
	beta1		0	normal(0, var=100)
	beta2		0	normal(0, var=100)
	beta3		0	normal(0, var=100)
	beta4		0	normal(0, var=100)

Cronologia dell'ottimizzazione		
Fase	Scala	Tasso di accettazione RWM
1	2.3800	0.2854
2	2.3800	0.2868

Cronologia degli scarti		
Scala	Tasso di accettazione RWM	
2.3800	0.3026	

Figure 20:

Analisi Bayesiana 2 per il peso dei neonati

15:47 Sunday, January 7, 2018

La procedura MCMC

Parametro	N	Media	Deviazione standard	Percentili		
				25	50	75
				1.1927	1.7845	
beta0	40000	1.2141	0.8598	0.6239		
beta1	40000	0.6902	0.3371	0.4599	0.6908	0.9196
beta2	40000	1.8948	0.7229	1.4042	1.8701	2.3685
beta3	40000	-0.0190	0.00678	-0.0235	-0.0187	-0.0143
beta4	40000	-0.0533	0.1718	-0.1676	-0.0506	0.0626

Intervalli a posteriori					
Parametro	Alfa	Intervallo con code uguali		Intervallo HPD	
beta0	0.050	-0.4162	2.9449	-0.4594	2.8891
beta1	0.050	0.0295	1.3440	0.0294	1.3430
beta2	0.050	0.5292	3.3739	0.4968	3.3314
beta3	0.050	-0.0328	-0.00625	-0.0323	-0.00587
beta4	0.050	-0.3982	0.2771	-0.3917	0.2820

Figure 21:

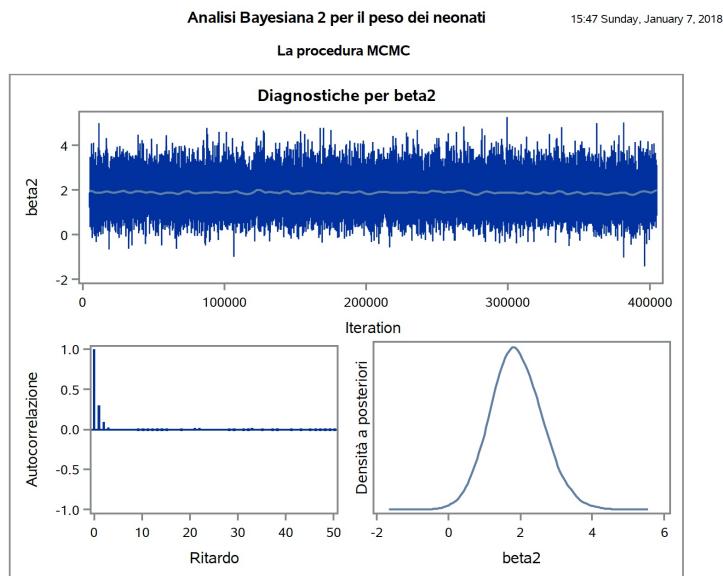


Figure 22:

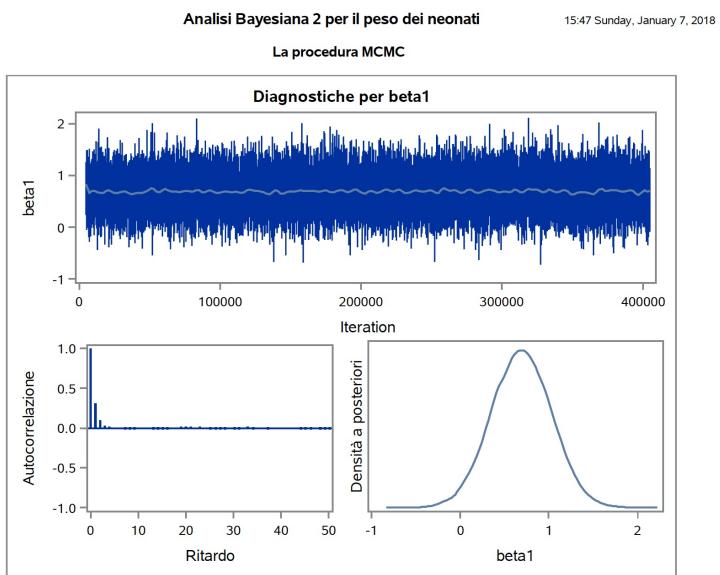


Figure 23: