UNIVERSITÀ
DEGLI STUDI
FIRENZE

Scienze Matematiche Fisiche e Naturali:
Laurea magistrale in Informatica

Corso di inferenza statistica bayesiana

Quaderno degli esercizi

Autore:
Andrea Moscatelli

Docenti:
Fabio Corradi
Francesco Stingo

Anno accademico 2017-2018

Indice

Capitolo 1

Processi stocastici e scambiabilità

1.1 Esercizio 1 di pagina 55 - Daboni Wedlin

(Testo dell'esercizio ripreso dal libro [2]).
Gli eventi E_1, E_2, E_3, E_4, E_5 sono giudicati scambiabili. Sono assegnate le seguenti probabilità:

$P(E_2) = \frac{1}{2}$

$P(E_3 \wedge E_5) = \frac{1}{4}$,

$P(E_1 \wedge \overline{E}_2 \wedge E_3 \wedge \overline{E}_4 \wedge E_5) = P(E_1 \wedge \overline{E}_2 \wedge \overline{E}_3 \wedge \overline{E}_4 \wedge \overline{E}_5) = P(E_1 \wedge E_2 \wedge E_3 \wedge E_4 \wedge E_5) = \frac{1}{30}$.

Si calcolino:

$P(E_2 \wedge E_3 \wedge E_4)$

$P(E_1 \wedge E_2 \wedge E_3 \wedge E_4)$

$P(E_1 \wedge E_2 \wedge \overline{E}_3 \wedge \overline{E}_4 \wedge \overline{E}_5)$

Svolgimento:

Ci troviamo nel caso di un *processo di alternativa semplice limitato* con 5 eventi $(E_1, E_2, E_3, E_4, E_5)$, ai quali corrispondono 5 variabili aleatorie $(X_1, X_2, X_3, X_4, X_5)$ sotto l'ipotesi di scambiabilità. Si parla di processo di alternativa semplice limitato poichè gli eventi in questione sono di numero limitato, di tipo elementare (vero-falso) e le variabili indicatrici associate sono di tipo 0-1.
L'ipotesi di scambiabilità garantisce che dati n eventi, non necessariamente indipendenti, la probabilità che se ne realizzino esattamente h su n non è dipendente

dall'ordine degli eventi stessi ovvero, permutando l'ordine delle variabili indicatrici, la probabilità della loro realizzazione resta immutata.

Usando la terminologia dei processi di alternativa semplice indicheremo con:

- ω_h^n la probabilità che dati n eventi se ne realizzino esattamente h, indipendentemente dal loro ordine
- $\frac{\omega_h^n}{\binom{n}{h}}$ la probabilità di una singola *traiettoria* formata da una determinata sequenza di h successi su n prove. Infatti di tali possibili traiettorie ne avremo $\binom{n}{h}$.
- ω_h la probabilità di h successi su h prove.

Possiamo a questo punto riscrivere i dati del nostro problema come segue:

$$\begin{aligned} P(E_2) = \omega_1 &= \frac{1}{2} \\ P(E_3 \wedge E_5) = \omega_2 &= \frac{1}{4} \\ P(E_1 \wedge \overline{E}_2 \wedge E_3 \wedge \overline{E}_4 \wedge E_5) &= \frac{\omega_3^5}{\binom{5}{3}} = \frac{1}{30} \\ P(E_1 \wedge \overline{E}_2 \wedge \overline{E}_3 \wedge \overline{E}_4 \wedge \overline{E}_5) &= \frac{\omega_1^5}{\binom{5}{1}} = \frac{1}{30} \\ P(E_1 \wedge E_2 \wedge E_3 \wedge E_4 \wedge E_5) = \omega_5 &= \frac{1}{30}. \end{aligned}$$

Ciò che dobbiamo calcolare sarà quindi:

$$\begin{aligned} P(E_2 \wedge E_3 \wedge E_4) &= \omega_3 \\ P(E_1 \wedge E_2 \wedge E_3 \wedge E_4) &= \omega_4 \\ P(E_1 \wedge E_2 \wedge \overline{E}_3 \wedge \overline{E}_4 \wedge \overline{E}_5) &= \frac{\omega_2^5}{\binom{5}{2}} \end{aligned}$$

Dalla teoria dei processi scambiabili sappiamo che le ω_h e le ω_h^n sono legate da specifiche relazioni e che è possibile ricavare le une dalle altre. Un modo per ottenere le serie richieste dai dati forniti è quello di sfruttare la seguente equazione:

$$\omega_h^n = \binom{n}{h}(-1)^{n-h} \cdot \Delta^{n-h} \cdot \omega_h \qquad h \leq n$$

Possiamo quindi riscrivere le quantità richieste e date dal problema in funzione delle ω_h:

$$\frac{\omega_2^5}{\binom{5}{2}} = (-1)^3\Delta^3\omega_2 = -(\omega_5 - 3\omega_4 + 3\omega 3 - \omega_2)$$
$$\frac{\omega_1^5}{\binom{5}{1}} = (-1)^4\Delta^4\omega_1 = \omega_5 - 4\omega_4 + 6\omega 3 - 4\omega_2 + \omega_1 = -4\omega_4 + 6\omega_3 - \frac{7}{15} = \frac{1}{30}$$
$$\frac{\omega_3^5}{\binom{5}{3}} = (-1)^2\Delta^2\omega_3 = \omega_5 - 2\omega_4 + \omega_3 = -2\omega_4 + \omega_3 + \frac{1}{30} = \frac{1}{30}$$

Mettendo a sistema le ultime due equazioni potremo ricavare quanto segue:

$$\begin{cases} -4\omega_4 + 6\omega_3 - \frac{7}{15} = \frac{1}{30} \\ -2\omega_4 + \omega_3 + \frac{1}{30} = \frac{1}{30} \end{cases} \implies \begin{cases} \omega_3 = \frac{1}{8} \\ \omega_4 = \frac{1}{16} \end{cases}$$

$$\frac{\omega_2^5}{\binom{5}{2}} = -\frac{1}{30} + \frac{3}{16} - \frac{3}{8} + \frac{1}{16} = \frac{7}{240}$$

In conclusione, le quantità richieste saranno:

$$P(E_2 \wedge E_3 \wedge E_4) = \omega_3 = \frac{1}{8}$$
$$P(E_1 \wedge E_2 \wedge E_3 \wedge E_4) = \omega_4 = \frac{1}{16}$$
$$P(E_1 \wedge E_2 \wedge \overline{E}_3 \wedge \overline{E}_4 \wedge \overline{E}_5) = \frac{\omega_2^5}{\binom{5}{2}} = \frac{7}{240}$$

1.2 Esercizio 2.6 - Hoff

(Testo dell'esercizio ripreso dal libro [1]).
Conditional independence: Suppose events A and B are conditionally independent given C, which is written $A \perp B|C$. Show that this implies that $A^c \perp B|C$, $A \perp B^c|C$ and $A^c \perp B^c|C$, where A^c means "not A". Find an example where $A \perp B|C$ holds but $A \perp B|C^c$ does not hold.

Svolgimento

Prima di procedere allo svolgimento di questo esercizio è necessario ricordare alcuni concetti fondamentali di probabilità su eventi.
Dati due eventi A e B, possiamo riscrivere la proprietà dell'evento A come segue

$$\begin{aligned} P(A) &= P(A \cap B) + P(A \cap B^c) \\ &= P(A, B) + P(A, B^c) \end{aligned}$$

e per il teorema di Bayes sappiamo che

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

di conseguenza anche la seguente equazione è immediatamente verificata

$$\begin{aligned} P(A|C) &= P(A,B|C) + P(A,B^c|C) \\ &= \frac{P(A,B,C)}{P(C)} + \frac{P(A,B^c,C)}{P(C)} \end{aligned} \tag{1.1}$$

Ricordiamo anche che la nozione di *indipendenza condizionata* fra due eventi A e B dato un evento C, ovvero $A \perp B|C$, implica che:

$$P(A,B|C) = P(A|C) \cdot P(B|C) \tag{1.2}$$

Sapendo che la seguente equazione è sempre vera,

$$P(A,B|C) = P(B|C) \cdot P(A|B,C)$$

possiamo quindi derivare anche la seguente proprietà:

$$\begin{aligned} P(B|C) \cdot P(A|B,C) = P(A,B|C) = P(A|C) \cdot P(B|C) \\ P(A|B,C) = P(A|C) \end{aligned} \tag{1.3}$$

Partendo da queste assunzioni, l'esercizio ci chiede quindi, dati $A \perp B|C$, di dimostrare le seguenti implicazioni:

1) $A^c \perp B|C \implies P(A^c,B|C) = P(A^c|C) \cdot P(B|C)$
2) $A \perp B^c|C \implies P(A,B^c|C) = P(A|C) \cdot P(B^c|C)$
3) $A^c \perp B^c|C \implies P(A^c,B^c|C) = P(A^c|C) \cdot P(B^c|C)$

Usando le proprietà appena citate procediamo a dimostrare le 3 equazioni:

1)

$$\begin{aligned} P(A^c|C) \cdot P(B|C) &= (1 - P(A|C)) \cdot P(B|C) \\ &= P(B|C) - \underbrace{P(A|C) \cdot P(B|C)}_{= P(A,B|C) \text{ per la (1.2)}} \\ &= P(B|C) - P(A,B|C) \\ &= \frac{P(B,C)}{P(C)} - \frac{P(A,B,C)}{P(C)} \\ \text{...per la proprietà (1.1)} &= \frac{P(A^c,B,C)}{P(C)} = P(A^c,B|C) \end{aligned}$$

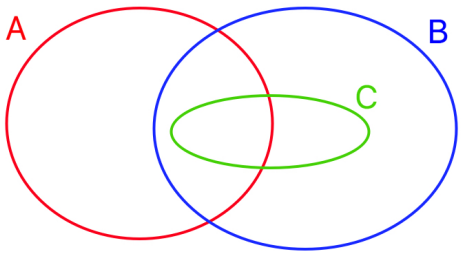
2)

$$\begin{aligned} P(A|C) \cdot P(B^c|C) &= P(A|C) \cdot (1 - P(B|C)) \\ &= P(A|C) - P(A|C) \cdot P(B|C) \\ &= P(A|C) - P(A, B|C) \\ &= \frac{P(A, C)}{P(C)} - \frac{P(A, B, C)}{P(C)} \\ &= \frac{P(A, B^c, C)}{P(C)} = P(A, B^c|C) \end{aligned}$$

3)

$$\begin{aligned} P(A^c|C) \cdot P(B^c|C) &= (1 - P(A|C)) \cdot P(1 - P(B|C)) \\ &= 1 - P(A|C) - P(B|C) + P(A, B|C) \\ &= 1 - \frac{P(A, C)}{P(C)} - \frac{P(B, C)}{P(C)} + \frac{P(A, B, C)}{P(C)} \\ &= \frac{P(C) - P(A, C) - P(B, C) + P(A, B, C)}{P(C)} \\ &= \frac{P(A^c, B^c, C)}{P(C)} = P(A^c, B^c|C) \end{aligned}$$

L'esercizio ci chiede adesso di trovare un esempio per il quale sia verificata $A \perp B|C$ ma non $A \perp B|C^c$. Per rispondere a questo problema possiamo rappresentare il nostro esempio in forma grafica:



Graficamente possiamo vedere come la proprità (1.3) sia rispettata per $A \perp B|C$ ma non per $A \perp B|C^c$.

Figura 1.1: L'aria di $P(A|B,C)$ e di $P(A|C)$ (entrambe in celeste) coincidono
.

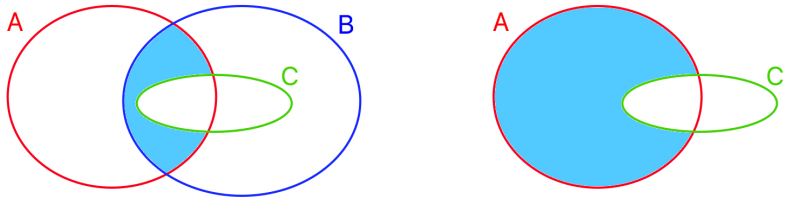


Figura 1.2: L'aria di $P(A|B,C^c)$ e di $P(A|C^c)$ (entrambe in celeste) **non** coincidono
.

Capitolo 2

Predittiva a posteriori

2.1 Esercizio 3.7 - Hoff

(Testo dell'esercizio ripreso dal libro [1]).
Posterior prediction: Consider a pilot study in which $n_1 = 15$ children enrolled in special education classes were randomly selected and tested for a certain type of learning disability. In the pilot study, $y_1 = 2$ children tested positive for the disability.

a) Using a uniform prior distribution, find the posterior distribution of θ, the fraction of students in special education classes who have the disability. Find the posterior mean, mode and standard deviation of θ, and plot the posterior density.

   Researchers would like to recruit students with the disability to participate in a long-term study, but first they need to make sure they can recruit enough students. Let $n_2 = 278$ be the number of children in special education classes in this particular school disctrict, and let Y_2 be the number of students with the disability.

b) find $Pr(Y_2 = y_2|Y_1 = 2)$, the posterior predictive disctribution of Y_2, as follows:

   i. Discuss what assumptions are needed about the joint distribution of (Y_1, Y_2) such that the following is true:

$$Pr(Y_2 = y_2) = \int_0^1 Pr(Y_2 = y_2|\theta)p(\theta|Y_1 = 2)\ d\theta$$

   ii. Now plug in the forms for $Pr(Y_2 = y_2|\theta)$ and $p(\theta|Y_1 = 2)$ in the above integral.

iii. Figure out what the above integral must be by using the calculus result discussed in Section 3.1

c) Plot the function $Pr(Y_2 = y_2|Y_1 = 2)$ as a function of y_2. Obtain the mean and standard deviation of Y_2, given Y_1=2.

d) The posterior mode and the MLE (maximum likelihood estimate; see Exercise 3.14) of θ, based on data from the pilot study, are both $\hat{\theta} = 2/15$. Plot the distribution $Pr(Y_2 = y_2|\theta = \hat{\theta})$, and find the mean and standard deviation of Y_2 give $\theta = \hat{\theta}$. Compare these results to the plots and calculations in c) and discuss any differences. Which distribution for Y_2 would you use to make prediction, and why?

Svolgimento

a) Possiamo modellare lo studio trattato nel testo con una distribuzione binomiale ovvero

$$P(\mathbf{y}|\theta) = \theta^y(1-\theta)^{1-y}$$

e come indicato dal testo useremo come prior una uniforme

$$P(\theta) = \mathbb{1}[0,1]^\theta.$$

Procediamo adesso al calcolo della *a posteriori*:

$$\begin{aligned} P(\theta|\mathbf{y}) &= K \times \mathcal{L}(\theta;\mathbf{y}) \times P(\theta) \\ &= K \times \left(\prod_{i=1}^{15} \theta^{y_i}(1-\theta)^{1-y_i}\right) \times \mathbb{1}[0,1]^\theta \\ &\text{...avendo 2 “successi” e 13 “insuccessi” avremo...} \\ &= K \times \theta^2(1-\theta)^{13} \times \mathbb{1}[0,1]^\theta \end{aligned}$$

Per calcolare il valore K e volendo trovare una distribuzione di probabilità propria, sappiamo che l'integrale di quest'ultima dovrà essere uguale a 1 nello spazio ammissibile di θ, ovvero

$$\begin{aligned} 1 &= \int_0^1 K \times \theta^2(1-\theta)^{13} \times \mathbb{1}[0,1]^\theta \; d\theta \\ &= K \int_0^1 \theta^2(1-\theta)^{13} \; d\theta \end{aligned}$$

Riconosciamo il kernel di una distribuzione Beta(3,14), quindi moltiplicando e

dividendo per la sua costante di normalizzazione otterrò:

$$\begin{aligned} 1 &= K \int_0^1 \theta^2(1-\theta)^{13} \frac{\Gamma(17)}{\Gamma(3)\Gamma(14)} \frac{\Gamma(3)\Gamma(14)}{\Gamma(17)}\, d\theta \\ &= K \times \frac{\Gamma(3)\Gamma(14)}{\Gamma(17)} \underbrace{\int_0^1 \theta^2(1-\theta)^{13} \frac{\Gamma(17)}{\Gamma(3)\Gamma(14)}\, d\theta}_{=1} \\ \frac{1}{K} &= \frac{\Gamma(3)\Gamma(14)}{\Gamma(17)} \end{aligned}$$

Posso quindi concludere che

$$\begin{aligned} P(\theta|\mathbf{y}) &= \frac{\Gamma(17)}{\Gamma(3)\Gamma(14)} \theta^2(1-\theta)^{13} \\ &\sim Beta(3,14) \end{aligned}$$

dalla quale posso facilmente calcolare valore medio, moda e standard deviation

$$\begin{aligned} E(\theta|\mathbf{y}) &= \frac{\alpha}{\alpha+\beta} = \frac{3}{17} \\ Moda(\theta|\mathbf{y}) &= \frac{\alpha-1}{\alpha+\beta-2} = \frac{2}{15} \\ Var(\theta|\mathbf{y}) &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{42}{17^2 \times 18} \approx 8 * 10^{-3} \end{aligned}$$

In Figura 2.1 possiamo visualizzare la distribuzione appena calcolata

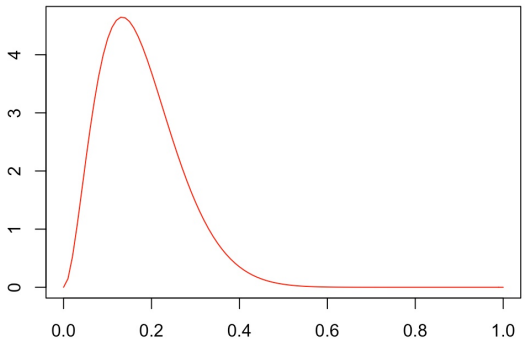


Figura 2.1: una Beta(3,14)

b) Affinchè

$$Pr(Y_2 = y_2) = \int_0^1 Pr(Y_2 = y_2|\theta) p(\theta|Y_1 = 2)\, d\theta \tag{2.1}$$

dobbiamo innanzitutto assumere che la distribuzione predittiva non dipenda da quantità incognite ma dipenda dai dati osservati (altrimenti sarebbe inutilizzabile ai fini predittivi), ma che il valore predetto sia indipendente dai dati osservati dato θ, ovvero:

$$Y_2 \perp\!\!\!\perp Y_1|\theta$$
$$(\text{ma non } Y_2 \perp\!\!\!\perp Y_1)$$

Con queste assunzioni otterremo che $Pr(Y_2|\theta, Y_1) = Pr(Y_2|\theta)$ e quindi la (2.1) risulterà valida.
Considerando le assunzioni del testo e svolgendo i calcoli avremo quanto segue:

$$\begin{aligned} Pr(Y_2 = y_2) &= \int_0^1 Pr(Y_2 = y_2|\theta)p(\theta|Y_1 = 2)\ d\theta \\ &= \int_0^1 \binom{278}{y_2}\theta^{y_2}(1-\theta)^{278-y_2}\frac{\Gamma(17)}{\Gamma(3)\Gamma(14)}\theta^2(1-\theta)^{13}\ d\theta \\ &= \binom{278}{y_2}\frac{\Gamma(17)}{\Gamma(3)\Gamma(14)}\int_0^1 \theta^{y_2+2}(1-\theta)^{291-y_2}\ d\theta \end{aligned}$$

Riconosco nell'integrale il kernel di una Beta$(y_2+3, 292-y_2)$ e quindi utilizzando la stessa tecnica vista al punto precedente avremo:

$$\begin{aligned} Pr(Y_2 = y_2) &= \binom{278}{y_2}\frac{\Gamma(17)}{\Gamma(3)\Gamma(14)}\frac{\Gamma(y_2+3)\Gamma(292-y_2)}{\Gamma(295)} \\ &= \frac{\Gamma(3+14)}{\Gamma(3)\Gamma(14)\Gamma(3+14+278)}\binom{278}{y_2}\Gamma(3+y_2)\Gamma(14+278-y_2) \end{aligned}$$

ovvero una *Beta-Binomiale*(3, 14, 278) (rappresentata in Figura 2.2) dalla quale posso facilmente calcolare media e standard deviation:

$$\begin{aligned} E(Y_2|Y_1 = 2) &= n\frac{\alpha}{\alpha+\beta} = 278\frac{3}{17} \approx 49,06 \\ Var(Y_2|Y_1 = 2) &= \frac{n\alpha\beta}{(\alpha+\beta)^2} \times \frac{(\alpha+\beta+n)}{(\alpha+\beta+1)} = \frac{278\times 3\times 14}{17^2} \times \frac{295}{18} \approx 662,13 \end{aligned}$$

Supponendo che la moda a posteriori e la MLE di θ siano entrambe $\hat{\theta} = 2/15$ allora avremo

$$\begin{aligned} P\left(Y_2 = y_2|\theta = \frac{2}{15}\right) &= \binom{278}{y_2}\left(\frac{2}{15}\right)^{y_2}\left(1-\frac{2}{15}\right)^{278-y_2} \\ &\sim Binomiale\left(\frac{2}{15}, 278\right) \end{aligned}$$

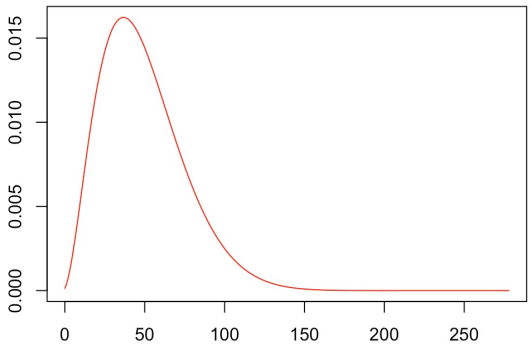


Figura 2.2: Beta-Binomiale(3, 14, 278)

Otteniamo quindi una distribuzione binomiale (rappresentata nella parte destra di Figura 2.3) con relative media e standard deviation:

$$E\left(Y_2 = y_2|\theta = \frac{2}{15}\right) = n\theta = 278 \times \frac{2}{15} \approx 37,07$$
$$Var\left(Y_2 = y_2|\theta = \frac{2}{15}\right) = n\theta(1-\theta) = 278 \times \frac{2}{15} \times \frac{13}{15} \approx 32,12$$

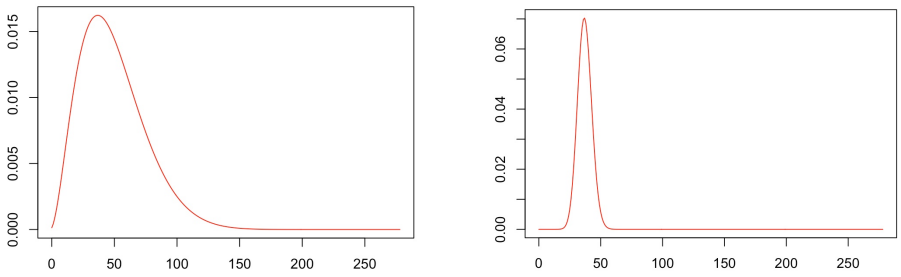


Figura 2.3: Una Betabinomiale(3, 14, 278) (a sinistra) e una Binomiale(2/15, 278) (a destra) a confronto.

Risulta evidente come media e moda delle due distribuzioni siano piuttosto simili a differenza della varianza che risulta notevolmente più bassa nella Binomiale. Questo risultato piuttosto intuitivo è dovuto all'aver fissato il parametro θ nel secondo punto, che ha come effetto una drastica riduzione della varianza nell'intervallo di confidenza rispetto alla media della nostra predizione. Concludendo, per la predizione di Y_2, preferiremo utilizzare quest'ultima variante ovvero assegnare a θ la MLE, che nonostante sia in leggera contraddizione con la definizione bayesiana di *a priori* (come vedremo nell'esercizio 3.14) ha come effetto una notevole riduzione della variabilità nella nostra predizione.

Capitolo 3

Unit information prior

3.1 Esercizio 3.14 - Hoff

(Testo dell'esercizio ripreso dal libro [1]).
Unit information prior: Let $Y_1, ..., Y_n \sim$ i.i.d.$p(y|\theta)$. Having observed the values $Y_1 = y_1, ..., Y_n = y_n$, the *log likelihood* is given by $l(\theta|\mathbf{y}) = \sum \log p(y_i|\theta)$, and the value $\hat{\theta}$ of θ that maximizes $l(\theta|\mathbf{y})$ is called the *maximum likelihood estimator*. The negative of the curvature of the loglikelihood, $J(\theta) = -\frac{\partial^2 l(\theta|\mathbf{y})}{\partial \theta^2}$, describes the precision of the MLE $\hat{\theta}$ and is called the *observed Fisher information*. For situations in which it is difficult to quantify prior information in terms of a probability distribution, some have suggested that the "prior" distribution be based on the likelihood, for example, by centering the prior distribution around the MLE $\hat{\theta}$. To deal with the fact that the MLE is not really prior information, the curvature of the prior is chosen so that it has only "one nth" as much information as the likelihood, so that $-\frac{\partial^2 \log p(\theta)}{\partial \theta^2} = \frac{J(\theta)}{n}$. Such a prior is called a *unit information prior* (Kass and Wasserman, 1995; Kass and Raftery, 1995), as it has as much information as the average amount of information from a single observation. The unit information prior is not really a prior distribution, as it is computed from the observed data. However, it can be roughly viewed as the prior information of someone with weak but accurate prior information.

a) Let $Y_1, ..., Y_n \sim$ i.i.d. binary(θ). Obtain the MLE $\hat{\theta}$ and J($\hat{\theta}$)/n.

b) Find a probability density $p_U(\theta)$ such that $\log p_U(\theta) = \frac{l(\theta|\mathbf{y})}{n} + c$, where c is a constant that does not depend on θ. Compute the information $-\frac{\partial^2 \log p(\theta)}{\partial \theta^2}$ of this density.

c) Obtain a probability density for θ that is proportional to $p_U(\theta) \times p(y_1, \cdots, y_n|\theta)$. Can this be considered a posterior distribution for θ?

d) Repeat a), b) and c) but with $p(y|\theta)$ being the Poisson distribution.

Svolgimento

a) Nel caso analizzato la funzione di densità sarà una Bernuoulli:

$$p(y|\theta) = \theta^y(1-\theta)^{1-y}$$

e la relativa funzione di verosimigliana per un campione di n osservazioni i.i.d. sarà

$$\mathcal{L}(\theta : \mathbf{y}) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i}$$

Calcolandonde il logaritmo otterremo la funzione di log-verosomiglianza seguente:

$$l(\theta : \mathbf{y}) = \left(\sum_{i=1}^{n} y_i\right) ln(\theta) + \left(n - \sum_{i=1}^{n} y_i\right) ln(1-\theta)$$

Calcoliamo adesso lo stimatore di massima verosomiglianza per θ che indicheremo con $\hat{\theta}$, ottenuto ponendo a zero la derivata prima della log-verosomiglianza, ovvero

$$\begin{aligned}\frac{\partial l(\theta;\mathbf{y})}{\partial\theta} &= \frac{\sum_{i=1}^n y_i}{\theta} - \frac{n - \sum_{i=1}^n y_i}{1-\theta} = 0 \\ &= \frac{\sum_{i=1}^n y_i - \cancel{\theta\sum_{i=1}^n y_i} - n\theta + \cancel{\theta\sum_{i=1}^n y_i}}{\theta(1-\theta)} = 0 \\ &= \frac{\sum_{i=1}^n y_i}{n} = \hat{\theta} = MLE\end{aligned}$$

Adesso dobbiamo calcolarci l'informazione osservata di Fisher, ovvero $J(\hat{\theta}) = -\frac{\partial^2 l(\theta;\mathbf{y})}{\partial\theta^2}$, dove al posto di θ sostituiamo $\hat{\theta}$ per ottenere una misura dell'informazione nel punto di massima verosomiglianza:

$$\begin{aligned}\frac{\partial^2 l(\theta;\mathbf{y})}{\partial\theta^2} &= -\frac{\sum_{i=1}^n y_i}{\theta^2} + \frac{n - \sum_{i=1}^n y_i}{(1-\theta)^2} \\ J(\theta) = -\frac{\partial^2 l(\theta;\mathbf{y})}{\partial\theta^2} &= \frac{\sum_{i=1}^n y_i}{\theta^2} - \frac{n - \sum_{i=1}^n y_i}{(1-\theta)^2}\end{aligned}$$

e sostituendo θ col nostro stimatore di massima verosomiglianza $\hat{\theta}$ avremo

$$J(\hat{\theta}) = n\left(\frac{\hat{\theta}}{\hat{\theta^2}} - \frac{1-\hat{\theta}}{(1-\hat{\theta})^2}\right) = n\left(\frac{1}{\hat{\theta^2}} - \frac{1}{1-\hat{\theta}}\right).$$

L'esercizio chiede di calcolare $\frac{J(\hat{\theta})}{n}$ ovvero l'informazione associata alla unit information prior. Otteremo quindi:

$$\frac{J(\hat{\theta})}{n} = \left(\frac{1}{\hat{\theta^2}} - \frac{1}{1-\hat{\theta}}\right).$$

b) Come chiesto dall'esercizio:

$$\log p_U(\theta) = \frac{l(\theta|y)}{n} + c$$

ovvero

$$\log p_U(\theta) = \frac{(\sum_{i=1}^n y_i)\, ln(\theta)}{n} + \frac{(n - \sum_{i=1}^n y_i)\, ln(1-\theta)}{n}$$

e riportandosi all'esponente della formula appena scritta otteremo la unit information prior:

$$\begin{aligned} p_U(\theta) &= \theta^{\frac{\sum_{i=1}^n y_i}{n}} (1-\theta)^{1-\frac{\sum_{i=1}^n y_i}{n}} e^c \\ &\sim Beta(\hat{\theta} + 1, 2 - \hat{\theta}). \end{aligned}$$

A questo punto per calcolare l'informazione di Fisher come richiesto dobbiamo come primo passo calcolare la derivata prima:

$$\frac{\partial p_U(\theta)}{\partial \theta} = \frac{\sum_{i=1}^n y_i}{n\theta} - \frac{n - \sum_{i=1}^n y_i}{n(1-\theta)}.$$

Poi la derivata seconda:

$$\frac{\partial^2 p_U(\theta)}{\partial \theta^2} = -\frac{\sum_{i=1}^n y_i}{n\theta^2} + \frac{n - \sum_{i=1}^n y_i}{n(1-\theta)^2}.$$

Cambiando di segno alla derivata seconda otteremo l'informazione di Fisher:

$$J_U(\theta) = -\frac{\partial^2 p_U(\theta)}{\partial \theta^2} = \frac{\sum_{i=1}^n y_i}{n\theta^2} - \frac{n - \sum_{i=1}^n y_i}{n(1-\theta)^2} = \frac{J(\theta)}{n}.$$

Notiamo che l'informazione di Fisher per la unit information prior $J_U(\theta)$ non è altro che un n-esimo dell'informazione di Fisher per l'intero campione $J(\theta)$, quindi la distribuzione ottenuta rispetta la proprietà desiderata.

c) L'esercizio chiede di trovare una densità di probabilità per θ che sia proporzionale a $p_U(\theta) \times p(y_1, \cdots, y_n|\theta)$.
Procediamo quindi a svolgere i calcoli richiesti nell'intento di riconoscere il kernel di una distribuzione nota:

$$\begin{aligned} p_U(\theta) \times \mathcal{L}(\theta; \mathbf{y}) &\propto \theta^{\frac{\sum_{i=1}^n y_i}{n}} (1-\theta)^{1-\frac{\sum_{i=1}^n y_i}{n}} \theta^{\sum_{i=1}^n y_i} (1-\theta)^{n-\sum_{i=1}^n y_i} \\ &\propto \theta^{\sum_{i=1}^n y_i (1+\frac{1}{n})} (1-\theta)^{(n+1)-\sum_{i=1}^n y_i (1+\frac{1}{n})} \end{aligned}$$

Riconosciamo il kernel di una Beta, ovvero:

$$p_U(\theta) \times \mathcal{L}(\theta; \mathbf{y}) \propto Beta\left(\sum_{i=1}^n y_i \left(1 + \frac{1}{n}\right) + 1, (n+2) - \sum_{i=1}^n y_i \left(1 + \frac{1}{n}\right)\right)$$

Se avessimo lavorato con una a priori classica potremmo concludere che:

$$p_U(\theta|\mathbf{y}) = Beta\left(\sum_{i=1}^{n} y_i\left(1+\frac{1}{n}\right)+1, (n+2) - \sum_{i=1}^{n} y_i\left(1+\frac{1}{n}\right)\right)$$

ma nel nostro caso non abbiamo usato una vera e propria *a priori*, bensì una *unit information prior* che per sua definizione viene ricavata dal campione e non da una pregressa conoscenza del fenomeno analizzato come impone la definizione di "a priori" della statistica bayesiana.
Possiamo però interpretare questa unit information prior, come l'informazione a priori di una persona che è riuscita a centrare la sua distribuzione sullo stimatore di massima verosimiglianza, ma che è molto insicuro di questa sua informazione (per questo dividiamo l'informazione di Fisher per n in modo da aumentarne la varianza). Solo se applichiamo questo ragionamento possiamo considerare la distribuzione trovata come una a posteriori.

d) Ripetiamo l'intero esercizio ma considerando $p(y|\theta)$ come una distribuzione di Poisson, ovvero:

$$p(y|\theta) = \frac{e^{-\theta}\theta^{y}}{y!}.$$

La relativa verosomiglianza per un campione i.i.d. sarà:

$$p(y|\theta) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{y_i}}{y_i!}$$

e la log-likelihood

$$l(\theta;\mathbf{y}) = -n\theta + \left(\sum_{i=1}^{n} y_i\right) ln(\theta) - ln\prod_{i=1}^{n} y_i!.$$

Calcoliamo come fatto in precedenza lo stimatore di massima verosomiglianza $\hat{\theta}$:

$$\frac{\partial l(\theta;\mathbf{y})}{\partial\theta} = -n + \frac{\sum_{i=1}^{n} y_i}{\theta}$$

Ponendolo uguale a zero avremo:

$$\begin{aligned} -n + \frac{\sum_{i=1}^{n} y_i}{\theta} &= 0 \\ \frac{\sum_{i=1}^{n} y_i}{\theta} &= \hat{\theta} = MLE. \end{aligned}$$

Procediamo adesso al calcolo dell'informazione osservata di FIsher $J(\hat{\theta})$ e quindi alla derivata seconda:

$$\frac{\partial^2 l(\theta;\mathbf{y})}{\partial\theta^2} = -\frac{\sum_{i=1}^{n} y_i}{\theta^2}.$$

Sostituendo a θ lo stimatore di massima verosomiglianza $\hat{\theta}$ avremo l'informazione osservata:

$$J(\hat{\theta}) = \frac{n}{\hat{\theta}}.$$

L'esercizio chiede di calcolare $\frac{J(\hat{\theta})}{n}$ ovvero un n-esimo dell'informazione osservata di Fisher relativa al campione, che corrisponde all'informazione associata alla unit information prior:

$$\frac{J(\hat{\theta})}{n} = \frac{1}{\hat{\theta}}.$$

Calcoliamo adesso come chiesto

$$\begin{aligned}\log p_U(\theta) &= \frac{l(\theta|y)}{n} + c \\ &= -\theta + \frac{\sum_{i=1}^n y_i}{n} ln(\theta) \underbrace{- \frac{ln(\prod_{i=1}^n y_i)}{n} + c}_{c^*}.\end{aligned}$$

Riportandosi all'esponente avremo:

$$\begin{aligned}p_U(\theta) &= e^{-\theta} \theta^{\frac{\sum_{i=1}^n y_i}{n}} e^{c^*} \\ &\sim Gamma(\hat{\theta} + 1, 1).\end{aligned}$$

Per calcolare l'informazione di Fisher dobbiamo calcolare come primo passo la derivata prima:

$$\frac{\partial p_U(\theta)}{\partial \theta} = -1 + \frac{\sum_{i=1}^n y_i}{n\theta}$$

Di conseguenza, la derivata seconda sara:

$$\frac{\partial^2 p_U(\theta)}{\partial \theta^2} = -\frac{\sum_{i=1}^n y_i}{n\theta^2}.$$

Possiamo adesso calcolare la unit information prior come richiesto

$$J_U(\theta) = -\frac{\partial^2 p_U(\theta)}{\partial \theta^2} = \frac{\sum_{i=1}^n y_i}{n\theta^2} = \frac{J(\theta)}{n}.$$

Procedendo adesso ad indentificare una distribuzione di densità nota per $p_U(\theta) \times \mathcal{L}(\theta; \mathbf{y})$ avremo:

$$\begin{aligned}p_U(\theta) \times \mathcal{L}(\theta; \mathbf{y}) &\propto e^{-\theta} \theta^{\frac{\sum_{i=1}^n y_i}{n}} e^{-n\theta} \theta^{\sum_{i=1}^n y_i} \\ &\propto e^{-\theta(1+n)} \theta^{\sum_{i=1}^n y_i (1+\frac{1}{n})} \\ &\sim Gamma\left(\sum_{i=1}^n y_i \left(1 + \frac{1}{n}\right) + 1, (n+1)\right).\end{aligned}$$

Possiamo quindi concludere che la "a posteriori" di $p_U(\theta|\mathbf{y})$ sia

$$p_U(\theta|\mathbf{y}) \sim Gamma\left(\sum_{i=1}^{n} y_i \left(1+\frac{1}{n}\right)+1, (n+1)\right).$$

Capitolo 4

Approssimazioni a posteriori con Gibbs sampling

4.1 Esercizio 6.1 - Hoff

(Testo dell'esercizio ripreso dal libro [1]).
Poisson population comparisons: let's reconsider the number of children data of Exercise 4.8. We'll assume Poisson sampling models for the two groups as before, but now we'll parameterize θ_A and θ_B as $\theta_A = \theta$, $\theta_B = \theta_A \times \gamma$. In this parameterization, γ represents the relative rate θ_B/θ_A. Let $\theta \sim$ Gamma (a_θ, b_θ) and let $\gamma \sim$ Gamma(a_γ, b_γ).

a) Are θ_A and θ_B independent or dependent under this prior distribution? In what situations is such a joint prior distribution justified?

b) Obtain the form of the full conditional distribution of θ given y_A, y_B and γ.

c) Obtain the form of the full conditional distribution of γ given y_A, y_B and θ.

d) Set $a_\theta = 2$ and $b_\theta = 1$. Let $a_\gamma = b_\gamma \in \{8, 16, 32, 64, 128\}$. For each of these five values, run a Gibbs sampler of at least 5,000 iterations and obtain $\mathrm{E}[\theta_B - \theta_A | y_A, y_B]$. Describe the effects of the prior distribution for γ on the results.

Svolgimento

A e B indicano due popolazioni di uomini di 30 anni con e senza laurea, rispettivamente, di cui vogliamo calcolare il numero medio di figli e sia Y il numero dei figli.
Per quanto riguarda la popolazione A, abbiamo

$$Y|\theta_A \sim Poisson(\theta_A) \quad \text{con} \quad \theta_A = \theta.$$

Similmente, per la popolazione B abbiamo

$$Y|\theta_B \sim Poisson(\theta_B) \quad \text{con} \quad \theta_B = \theta \times \gamma.$$

Dal momento che

- $\theta \sim Gamma(a_\theta, b_\theta)$
- $\gamma \sim Gamma(a_\gamma, b_\gamma)$
- $\theta \perp \gamma$

il modello delle osservazioni può essere riscritto come

$Y|\theta \sim$ Poisson(θ) per A
$Y|\theta, \gamma \sim$ Poisson($\theta\gamma$) per B.

Una volta ricavato θ, e di conseguenza θ_A, possiamo calcolare θ_B essendo uguale a $\theta \times \gamma$. In particolare se:

- $\theta = 1$, allora $\theta_A = \theta_B$
- $\theta > 1$, allora $\theta_A > \theta_B$
- $\theta < 1$, allora $\theta_A < \theta_B$

La riparametrizzazione $\theta_A = \theta$, $\theta_B = \theta \times \gamma$ è di conseguenza un'operazione utile per analizzare e confrontare le due distribuzioni.

a) Per poter valutare se θ_A e θ_B sono indipendenti, calcoliamo il valore della

covarianza $Cov(\theta_A, \theta_B)$.

$$\begin{aligned}
Cov(\theta_A, \theta_B) &= E(\theta_A\theta_B) - E(\theta_A)E(\theta_B) \\
&= E(\theta^2\gamma) - E(\theta)E(\theta\gamma) \\
&= \int_\theta \int_\gamma \theta^2 p(\theta, \gamma) d\theta d\gamma - E(\theta) \int_\theta \int_\gamma \theta p(\theta, \gamma) d\theta d\gamma \\
&= \int_\theta \theta^2 p(\theta) d\theta \int_\gamma p(\gamma) d\gamma - E(\theta) \int_\theta \theta p(\theta) d\theta \int_\gamma \gamma p(\gamma) d\gamma \\
&= E(\theta^2)E(\gamma) - E(\theta)^2 E(\gamma) \\
&= E(\gamma)(E(\theta^2) - E(\theta)^2) \\
&= E(\gamma)Var(\theta) \\
&= \frac{\alpha_\gamma}{b_\gamma} \frac{\alpha_\theta}{b_\theta^2} > 0
\end{aligned}$$

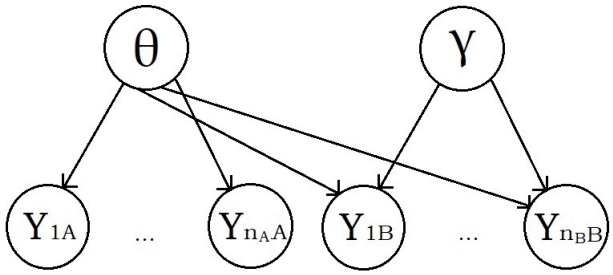
Tale quantità è maggiore di 0, quindi θ_A e θ_B sono linearmente dipendenti. Si può verificare tale dipendenza anche con un ragionamento meno formale, considerando la seguente uguaglianza:

$$p(\theta_B|\theta_A) = p(\theta\gamma|\theta).$$

θ compare non solo nella variabile, ma anche come parametro dato nella condizione. Per tale ragione, θ_A e θ_B sono dipendenti.

Adesso consideriamo la distribuzione di Poisson per analizzare eventi relativi a un certo intervallo di tempo. Esaminiamo due casi distinti, A e B, ai quali sono associati θ_A e θ_B per misurare gli eventi che si verificano. Sappiamo che $\theta_A = \theta$ e che $\theta_B = \theta\gamma$ e che il numero degli eventi di B è influenzato da γ: si tratta perciò di un parametro che differenzia B rispetto alle condizioni "standard" di A.
La distribuzione congiunta a priori può essere dunque utilizzata in situazioni analoghe a quella appena descritta, ossia quando abbiamo l'obiettivo di esaminare i cambiamenti su una determinata popolazione rispetto a un'altra di partenza.

b) Per ricavare la distribuzione full conditional di θ, come primo passo dobbiamo definire il Markov blanket, poi applicare la proprietà locale di Markov.



Come si può osservare dallo schema, il Markov blanket di θ comprende le osservazioni di A, le osservazini di B e γ (questo perchè le osservazioni di B hanno come nodo padre anche γ). In formula risulterà quindi:

$$bl(\theta) = \{Y_{1A}, Y_{2A} \dots Y_{n_A A}, Y_{1B}, Y_{2B} \dots Y_{n_B B}, \gamma\}.$$

Dunque:

$$\begin{aligned} p(\theta|y_{1A}, y_{2A} \dots y_{n_A A}, y_{1B}, y_{2B} \dots y_{n_B B}, \gamma) &\propto p(\theta|a_\theta, b_\theta) \prod_{i=1}^{n_A} p(y_{iA}|\theta) \prod_{j=1}^{n_B} p(y_{jB}|\theta, \gamma) \\ &= \frac{b_\theta^{a_\theta}}{\Gamma(a_\theta)} \theta^{a_\theta - 1} e^{-b_\theta \theta} \prod_{i=1}^{n_A} \frac{\theta^{y_{iA}} e^{-\theta}}{y_{iA}!} \prod_{j=1}^{n_B} \frac{(\theta\gamma)^{y_{jB}} e^{-\theta\gamma}}{y_{jB}!} \\ &\propto \theta^{a_\theta - 1} e^{-b_\theta \theta} \theta^{\sum_{i=1}^{n_A} y_{iA}} e^{-n_A \theta} (\theta\gamma)^{\sum_{j=1}^{n_B} y_{jB}} e^{-n_B \theta\gamma} \\ &= \theta^{a_\theta + \sum_{i=1}^{n_A} y_{iA} + \sum_{j=1}^{n_B} y_{jB} - 1} e^{-\theta(b_\theta + n_A + n_B \gamma)} \end{aligned}$$

Dal risultato ottenuto riconosciamo il kernel di una Gamma:

$$Gamma\left(a_\theta + \sum_{i=1}^{n_A} y_{iA} + \sum_{j=1}^{n_B} y_{jB}, b_\theta + n_A + n_B \gamma\right).$$

c) Procediamo in modo analogo al punto precedente, ricavando il Markov blanket di γ che risulta essere $bl(\gamma) = \{Y_{1B}, Y_{2B} \dots Y_{n_B B}, \theta\}$.

Quindi:

$$\begin{aligned}
p(\gamma|y_{1A}, y_{2A} \dots y_{n_A A}, y_{1B}, y_{2B} \dots y_{n_B B}, \theta) &= p(\gamma|y_{1B}, y_{2B} \dots y_{n_B B}, \theta) \\
&\propto p(\gamma|a_\gamma, b_\gamma) \prod_{i=1}^{n_B} p(y_{iB}|\theta, \gamma) \\
&= \frac{b_\gamma^{a_\gamma}}{\Gamma(a_\gamma)} \gamma^{a_\gamma - 1} e^{-b_\gamma \gamma} \prod_{i=1}^{n_B} \frac{(\theta\gamma)^{y_{iB}} e^{-\theta\gamma}}{y_{iB}!} \\
&\propto \gamma^{a_\gamma - 1} e^{-b_\gamma \gamma} (\theta\gamma)^{\sum_{i=1}^{n_B} y_{iB}} e^{-n_B \theta_\gamma} \\
&\propto \gamma^{a_\gamma + \sum_{i=1}^{n_B} y_{iB} - 1} e^{-\gamma(b_\gamma + n_B \theta)}
\end{aligned}$$

Anche in questo caso riconosciamo il kernel di una Gamma:

$$Gamma\left(a_\gamma + \sum_{i=1}^{n_B} y_{iB}, b_\gamma + n_B\theta\right).$$

d) Riportiamo di seguito il codice R relativo all'algoritmo di Gibbs.

Come indicato nel testo dell'esercizio 4.8, i dataset utilizzati sono presenti nei file menchild30bach.dat e menchild30nobach.dat.

Come primo passo, carichiamo i dati:

```
1
2  A<-c(1, 0, 0, 1, 2, 2, 1, 5, 2, 0, 0, 0, 0, 0,
3      0, 1, 1, 1, 0, 0, 0, 1, 1, 2, 1, 3, 2, 0, 0,
4      3, 0, 0, 0, 2, 1, 0, 2, 1, 0, 0, 1, 3, 0, 1,
5      1, 0, 2, 0, 0, 2, 2, 1, 3, 0, 0, 0, 1, 1)
6
7  B<-c(2, 2, 1, 1, 2, 2, 1, 2, 1, 0, 2, 1, 1, 2,
8      0, 2, 2, 0, 2, 1, 0, 0, 3, 6, 1, 6, 4, 0, 3, 2, 0, 1,
9      0, 0, 0, 3, 0, 0, 0, 0, 0, 1, 0, 4, 2, 1, 0, 0, 1, 0,
10     3, 2, 5, 0, 1, 1, 2, 1, 2, 1, 2, 0, 0, 0, 2, 1,
11     0, 2, 0, 2, 4, 1, 1, 1, 2, 0, 1, 1, 1, 1, 0, 2, 3, 2,
12     0, 2, 1, 3, 1, 3, 2, 2, 3, 2, 0, 0, 0, 1, 0, 0,
13     0, 1, 2, 0, 3, 3, 0, 1, 2, 2, 2, 0, 6, 0, 0, 0, 2, 0,
14     1, 1, 1, 3, 3, 2, 1, 1, 0, 1, 0, 0, 2, 0, 2, 0,
15     1, 0, 2, 0, 0, 2, 2, 4, 1, 2, 3, 2, 0, 0, 0, 1, 0, 0, 1,
16     5, 2, 1, 3, 2, 0, 2, 1, 1, 3, 0, 5, 0, 0, 2,
17     4, 3, 4, 0, 0, 0, 0, 0, 0, 2, 2, 0, 0, 2, 0, 0, 1, 1, 0,
18     2, 1, 3, 3, 2, 2, 0, 0, 2, 3, 2, 4, 3, 3, 4,
19     0, 3, 0, 1, 0, 1, 2, 3, 4, 1, 2, 6, 2, 1, 2, 2)
```

Procediamo con l'inizializzazione dei parametri delle varie distribuzioni, del numero delle iterazioni da eseguire e del seed:

```
1  a_theta <- 2
2  b_theta <- 1
```

```r
v_gamma <- c(8, 16, 32, 64, 128)
n <- 5000
set.seed(150)
```

Per la distribuzione gamma facciamo uso di un vettore dato che dobbiamo valutare i risultati per $a_\gamma = b_\gamma \in \{8, 16, 32, 64, 128\}$.
Inizializziamo inoltre il vettore delle medie, il cui uso verrà chiarito a breve

```r
valori_medie <- NULL
```

Nell'algoritmo di Gibbs viene eseguito un doppio ciclo: nel for più interno, per ognuna delle 5 coppie di iperparametri, vengono simulate le distribuzioni full conditionall (con 5000 valori sia per theta che per gamma); in ognuna delle iterazioni del for più esterno, vengono invece calcolati i valori attesi.

Per comodità, facciamo uso di un array in 3 dimensioni: le prime due fanno riferimento a una matrice (5000×2) in cui vengono memorizzati i valori campionati di theta e gamma per ogni iterazione; la terza dimensione serve per rieseguire questi passi, variando però gli iperparametri dell'a priori della distribuzione gamma (per questo a dimensione 5).

```r
gibbs <- array(NA, c(n, 2, 5))
```

Prima di eseguire l'algoritmo, calcoliamo, per ogni gruppo, le numerosità e i totali.

```r
nA <- length(A)
nB <- length(B)
ytotA <- sum(A)
ytotB <- sum(B)
ytot <- ytotA + ytotB
```

Tali valori caratterizzano i parametri delle full conditional.

Adesso inizializziamo i parametri gamma, theta e k:

```r
gamma <- 1
theta <- 1
k <- 1
```

Come passo successivo, eseguiamo l'algoritmo:

```r
for (i in v_gamma) {
  a_gamma <- b_gamma <- i
  for (j in 1:n) {
    theta <- rgamma(1, a_theta + ytot, b_theta + nA + nB * gamma)
    gamma <- rgamma(1, a_gamma + ytotB, b_gamma + nB * theta)
    gibbs[j, , k] <- cbind(theta, gamma)
  }
  k <- k + 1
}
```

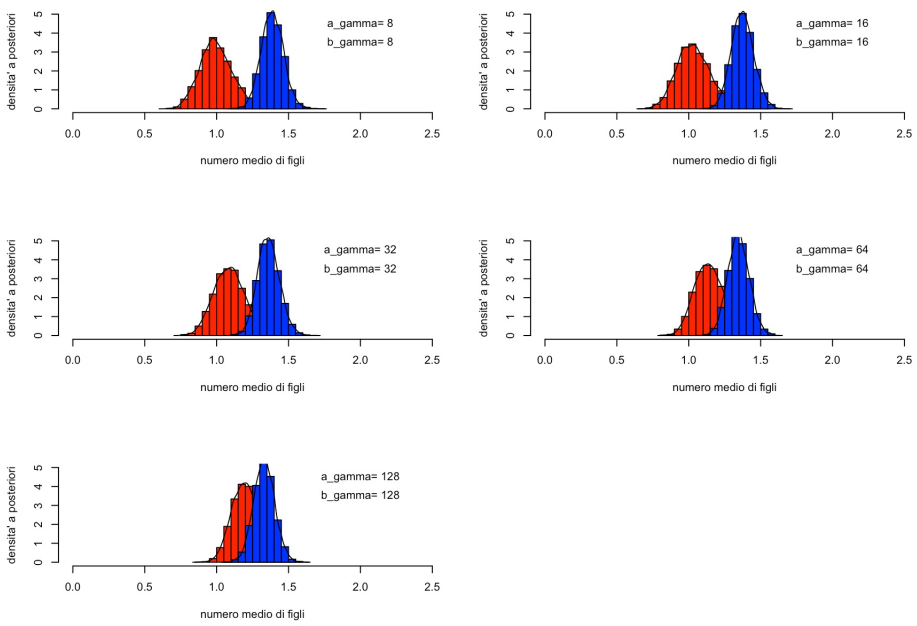
Considerato che vogliamo ricavare il valore atteso della differenza tra θ_B e θ_A, dobbiamo calcolare la trasformazione theta*gamma-theta per ogni estrazione di theta e gamma ed eseguire la media su tutte le simulazioni. Questo è possibile perchè abbiamo a disposizione il campione della congiunta a posteriori sia di theta che di gamma.
Ovviamente il procedimento deve essere ripetuto per ognuna delle 5 possibili scelte degli iperparametri della a priori di gamma, perciò memorizziamo i vari risultati nell'array "valori_medie".

```
for (i in 1:5) {
  valori_medie <-
    c(valori_medie, mean(gibbs[, 1, i] * gibbs[, 2, i] - gibbs[, 1,
        i]))
}
valori_medie
```

```
[1] 0.3910311 0.3388716 0.2708131 0.2026279 0.1333113
```

Come si può notare nell'output, il valore atteso della differenza tra θ_B e θ_A diminuisce progressivamente all'aumentare dei parametri dell'a priori di gamma, quindi il numero di figli delle due popolazioni prese in esame, decresce.

```
leg <- v_gamma
leg.txt <- rep("ab=bb", 5)
par(mfrow = c(3, 2))
for (i in 1:5) {
  hist(
    gibbs[, 1, i],
    prob = T,
    col = "red",
    ylim = c(0, 5),
    xlim = c(0, 2.5),
    ylab = "densita' a posteriori",
    xlab = "numero medio di figli",
    main = ""
  )
  lines(density(gibbs[, 1, i]))
  hist(gibbs[, 1, i] * gibbs[, 2, i],
       prob = T,
       col = "blue",
       add = T)
  lines(density(gibbs[, 1, i] * gibbs[, 2, i]))
  text(2, 4.5, paste("a_gamma=", leg[i]))
  text(2, 3.5, paste("b_gamma=", leg[i]))
}
```

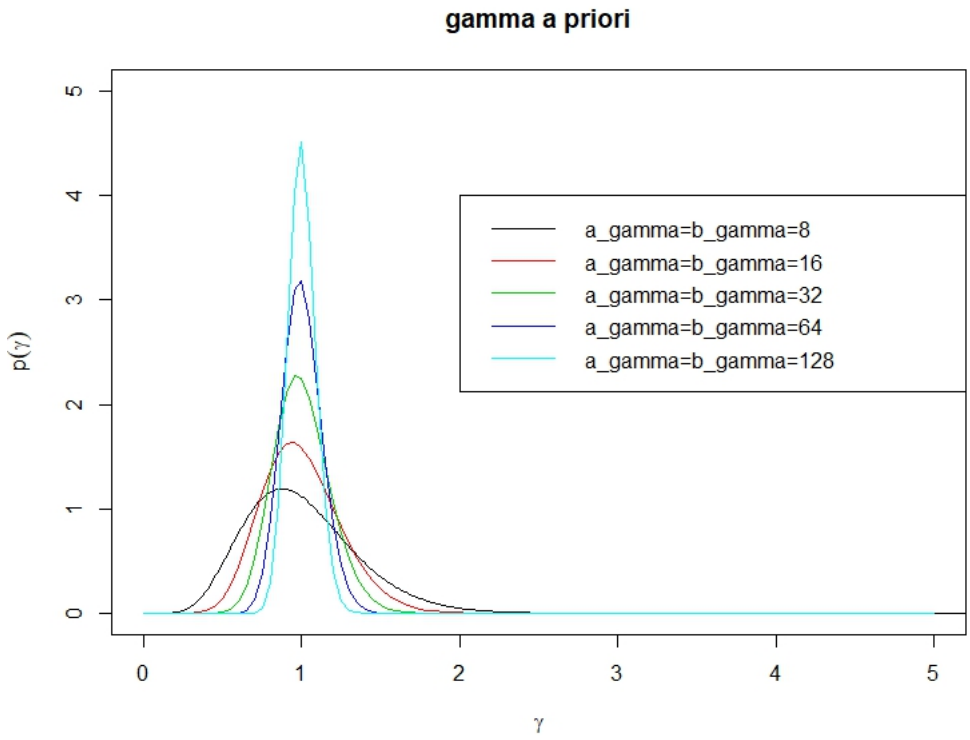


Quanto detto precedentemente, è evidente anche osservando i grafici. La parte in rosso rappresenta θ_A, quella blu θ_B. Come si può notare, le due distribuzioni sono sempre più vicine l'una all'altra all'aumentare degli iperparametri dell'a priori.

```
x <- seq(0, 10, by = 0.01)
par(mfrow = c(1, 1))
plot(
  x,
  dgamma(x, 8, 8),
  type = "l",
  xlim = c(0, 5),
  ylim = c(0, 5),
  xlab = expression(gamma),
  ylab = expression(p(gamma)),
  main = "gamma a priori",
  col = 1
)


for (j in 2:5) {
  curve(dgamma(x, v_gamma[j], v_gamma[j]),
        add = T,
        col = j)
}
legend(
```

```
  2,
  4,
  c(
    "a_gamma=b_gamma=8",
    "a_gamma=b_gamma=16",
    "a_gamma=b_gamma=32",
    "a_gamma=b_gamma=64",
    "a_gamma=b_gamma=128"
  ),
  col = c(1, 2, 3, 4, 5),
  lty = 1
)
```



Osserviamo questo grafico: al crescere degli iperparametri, l'a priori si "schiaccia" attorno a 1, che è proprio il valore atteso. Per quanto riguanrda la varianza, essa è invece inversamente proporzionale al valore dei due iperparametri. Andando avanti quindi al caso limite in cui i due iperparametri crescono all'infinito ci troviamo pertanto nella situazione in cui non si vuole imparare nulla dalle osservazioni e si ha un'opinione forte e sicura sulla prior di gamma che non si vuole cambiare (notiamo però che in ottica

bayesiana di aggiornamento dell'informazione tramite i dati, è una situazione priva di senso e utilità). Tutto questo comunque ci fa notare come l'inferenza cambia notevolmente in base alla scelta della distribuzione a priori dei parametri.

Sarebbe opportuno valutare dettagliatamente la convergenza dell'algoritmo di Gibbs ma in questo caso non è stato richiesto dall'esercizio. Ci limitiamo quindi a calcolare l'effective sample size e ad accertarsi che sia sufficientemente elevata per essere sicuri di aver raggiunto la distribuzione di equilibrio e di averla ben approssimata.

```
library(coda)
effectiveSize(gibbs[, , 1])
```

```
    var1     var2
609.1041 605.4248
```

Capitolo 5

La normale multivariata

5.1 Esercizio 7.2 - Hoff

(Testo dell'esercizio ripreso dal libro [1]).
Unit information prior: Letting $\Psi = \Sigma^{-1}$, show that a unit information prior for $(\boldsymbol{\theta}, \Psi)$ is given by $\boldsymbol{\theta}|\Psi \sim$ multivariate normal $(\bar{\mathbf{y}}, \Psi^{-1})$ and $\Psi \sim Wishart(p+1, \mathbf{S}^{-1})$, where $\mathbf{S} = \sum(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^T/n$. This can be done by mimicking the procedure outlined in Exercise 5.6 as follows:

a) Reparameterize the multivariate normal model in terms of the precision matrix $\Psi = \Sigma^{-1}$. Write out the resulting log likelihood, and find a probability density $p_U(\boldsymbol{\theta}, \Psi) = p_U(\boldsymbol{\theta}|\Psi)p_U(\Psi)$ such that $\log p(\boldsymbol{\theta}, \Psi) = l(\boldsymbol{\theta}, \Psi|\mathbf{Y})/n + c$, where c does not depend on $\boldsymbol{\theta}$ or Ψ.
Hint: Write $(\mathbf{y}_i - \boldsymbol{\theta})$ as $(\mathbf{y}_i - \bar{\mathbf{y}} + \bar{\mathbf{y}} - \boldsymbol{\theta})$, and note that $\sum \mathbf{a}_i^T \mathbf{B} \mathbf{a}_i$ can be written as $\text{tr}(\mathbf{AB})$, where $\mathbf{A} = \sum \mathbf{a}_i \mathbf{a}_i^T$.

b) Let $p_U(\Sigma)$ be the inverse-Wishart density induced by $p_U(\Psi)$. Obtain a density $p_U(\boldsymbol{\theta}, \Sigma|\mathbf{y}_i, ..., \mathbf{y}_n)) \propto p_U(\boldsymbol{\theta}|\Sigma)p_U(\boldsymbol{\Sigma})p(\mathbf{y}_i, ..., \mathbf{y}_n|\boldsymbol{\theta}, \Sigma)$. Can this be interpreted as a posterior distribution for θ and Σ?

Svolgimento

a) La distribuzione di probabilità di una normale multivariata è

$$p(\mathbf{Y}|\Sigma, \theta) = (2\pi)^{-1/2}|\Sigma|^{-1/2} \exp\left\{-1/2(\mathbf{y} - \boldsymbol{\theta})^T \Sigma^{-1}(\mathbf{y} - \boldsymbol{\theta})\right\}$$

e riparametrizzando con $\Psi = \Sigma^{-1}$ avremo

$$p(\mathbf{Y}|\Psi, \theta) = (2\pi)^{-p/2}|\Psi|^{1/2} \exp\left\{-1/2(\mathbf{y} - \boldsymbol{\theta})^T \Psi(\mathbf{y} - \boldsymbol{\theta})\right\}$$

con relativa likelihood

$$\mathcal{L}(\mathbf{Y}|\Psi,\theta) = \prod_{i=1}^{n}(2\pi)^{-p/2}|\Psi|^{1/2}\exp\left\{-\frac{1}{2}(\mathbf{y}_i-\boldsymbol{\theta})^T\Psi(\mathbf{y}_i-\boldsymbol{\theta})\right\}$$
$$\propto |\Psi|^{n/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i-\boldsymbol{\theta})^T\Psi(\mathbf{y}_i-\boldsymbol{\theta})\right\}$$

come suggerito dall'esercizio ci calcoliamo $\log p(\boldsymbol{\theta},\Psi) = l(\boldsymbol{\theta},\Psi|\mathbf{Y})/n + c$

$$\log(\mathcal{L}(\mathbf{Y}|\Psi,\theta)) = \frac{n}{2}\log|\Psi|\left\{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i-\boldsymbol{\theta})^T\Psi(\mathbf{y}_i-\boldsymbol{\theta})\right\}$$
$$\log(\mathcal{L}(\mathbf{Y}|\Psi,\theta))/n + c = \frac{1}{2}\log|\Psi|\left\{-\frac{1}{2n}\sum_{i=1}^{n}(\mathbf{y}_i-\boldsymbol{\theta})^T\Psi(\mathbf{y}_i-\boldsymbol{\theta})\right\} + c$$

Usando il suggerimento proposto dal testo avremo

$$-\frac{1}{2n}\sum_{i=1}^{n}(\mathbf{y}_i-\boldsymbol{\theta})^T\Psi(\mathbf{y}_i-\boldsymbol{\theta}) = -\frac{1}{2n}\sum_{i=1}^{n}(\mathbf{y}_i-\bar{\mathbf{y}}+\bar{\mathbf{y}}-\boldsymbol{\theta})^T\Psi(\mathbf{y}_i-\bar{\mathbf{y}}+\bar{\mathbf{y}}-\boldsymbol{\theta})$$
$$= -\frac{1}{2n}\sum_{i=1}^{n}\left[(\mathbf{y}_i-\bar{\mathbf{y}})^T\Psi(\mathbf{y}_i-\bar{\mathbf{y}})\right] - \frac{\cancel{n}}{2\cancel{n}}(\theta-\bar{\mathbf{y}})^T\Psi(\theta-\bar{\mathbf{y}})$$
$$= -\frac{1}{2}tr(\mathbf{S}\Psi) - \frac{1}{2}tr(\mathbf{S}_\theta\Psi)$$

Quindi la log likelihood calcolata al passo precedente diventerà

$$\log(\mathcal{L}(\mathbf{Y}|\Psi,\theta))/n + c = \frac{1}{2}\log|\Psi| - \frac{1}{2}tr(\mathbf{S}\Psi) - \frac{1}{2}tr(\mathbf{S}_\theta\Psi) + c$$

e tornando all'esponente avremo

$$\mathcal{L}(\mathbf{Y}|\Psi,\theta)/n + c \propto \underbrace{\exp\{|\Psi|\} - \exp\left\{\frac{1}{2}tr(\mathbf{S}\Psi)\right\}}_{\sim \text{ Wishart}(\Psi|\ \text{k+2},\ \mathbf{S}^{-1})} \underbrace{- \exp\left\{\frac{1}{2}tr(\mathbf{S}_\theta\Psi))\right\}}_{\sim NM(\bar{\mathbf{y}},\,\Psi^{-1})}$$

Abbiamo quindi trovato che $p_U(\boldsymbol{\theta},\Psi) = p_U(\boldsymbol{\theta}|\Psi)p_U(\Psi)$ dove

$$p_U(\Psi) = Wishart(\Psi|k+2,\mathbf{S}^{-1})$$
$$p_U(\boldsymbol{\theta}|\Psi) = NM(\bar{\mathbf{y}},\Psi^{-1})$$

b) Da ciò che abbiamo appena calcolato al punto a), se volessimo tornare a $p_U(\Sigma)$ come suggerito dal testo avremo un Inverse Wishart e le relative distribuzioni

di probabilità saranno

$$\begin{aligned} p_U(\Sigma) &\propto |\Sigma|^{-(p+k)/2} \exp\Big\{ -\frac{1}{2} tr(S\Sigma^{-1}) \Big\} \\ p_U(\boldsymbol{\theta}|\Sigma) &\propto \exp\Big\{ -\frac{1}{2} (\boldsymbol{\theta} - \bar{\mathbf{y}})^T \Sigma^{-1} (\boldsymbol{\theta} - \bar{\mathbf{y}}) \Big\} \\ p(\mathbf{y}_i, ..., \mathbf{y}_n|\boldsymbol{\theta}, \Sigma) &\propto |\Sigma|^{-n/2} \exp\Big\{ -\frac{1}{2} \sum_{i=1}^{n} (\mathbf{y}_i - \boldsymbol{\theta})^T \Sigma^{-1} (\mathbf{y}_i - \boldsymbol{\theta}) \Big\} \\ &\propto |\Sigma|^{-n/2} \exp\Big\{ -\frac{1}{2} tr(S_1\Sigma^{-1}) \Big\} \end{aligned}$$

volendo a questo punto trovare la densità chiesta, $p_U(\boldsymbol{\theta}, \Sigma|\mathbf{y}_i, ..., \mathbf{y}_n)) \propto p_U(\boldsymbol{\theta}|\Sigma) p_U(\boldsymbol{\Sigma}) p(\mathbf{y}_i, ..., \mathbf{y}_n|\boldsymbol{\theta}, \Sigma)$, avremo

$$\begin{aligned} p_U(\boldsymbol{\theta}, \Sigma|\mathbf{y}_i, ..., \mathbf{y}_n)) &\propto |\Sigma|^{-n/2} |\Sigma|^{-(p+k)/2} \exp\Big\{ -\frac{1}{2} tr[(S_1 + S)\Sigma^{-1}] \Big\} \exp\Big\{ -\frac{1}{2} (\boldsymbol{\theta} - \bar{\mathbf{y}})^T \Sigma^{-1} (\boldsymbol{\theta} - \bar{\mathbf{y}}) \Big\} \\ &\propto \underbrace{|\Sigma|^{-\frac{n+p+k+1}{2}} \exp\Big\{ -\frac{1}{2} tr[(S_1 + S)\Sigma^{-1}] \Big\}}_{\sim\, Inv - Wishart(n+k, (S+S_1)^{-1})} \underbrace{|\Sigma|^{-1/2} \exp\Big\{ -\frac{1}{2} (\boldsymbol{\theta} - \bar{\mathbf{y}})^T \Sigma^{-1} (\boldsymbol{\theta} - \bar{\mathbf{y}}) \Big\}}_{\sim\, N(\bar{y}, \Sigma)} \end{aligned}$$

che può essere interpretata come la distribuzioni a posteriori di $\boldsymbol{\theta}$ e Σ.

Capitolo 6

Importance sampling

6.1 Esercizio Importance Sampling

Realizza un importance sampling per stimare il valore atteso della mistura di due beta $(0.3 \times \beta(5, 2) + 0.7 \times \beta(2, 8))$. Valuta altresì utilizzando il campione ottenuto la probabilità di questa mistura nell'intervallo [0.45 - 0.55].

Svolgimento

La funzione $g(x)$ usata campionare gli x_i è una uniforme tra 0 ed 1.
Il codice R scritto per effettuare l'*importance sampling* è il seguente:

```
importanceSampling = function(samples, min_unif = 0, max_unif = 1) {
  beta_mixture = function(x) {
    value = 0.3 * (x ** 4) * (1 - x) * (factorial(6) /
           (factorial(4))) + 0.7 * (x * (1 - x) ** 7) * factorial(9) /
               factorial(7)
    return(value)
  }

  #x_i
  y_unif = runif(samples, min_unif, max_unif)


  #f(x_i)
  f_x = c()
  #g(x_i)
  g_x = c()
  for (i in y_unif) {
    f_x = c(f_x, beta_mixture(i))
    g_x = c(g_x, dunif(i, min_unif, max_unif))
  }
```

```
20
21   #w_i = f(x_i)/g(x_i)
22   w = f_x / g_x
23   expected_value = sum(y_unif * w) / samples
24
25
26   in_interval = c()
27   for (f_x_i in f_x) {
28     in_interval = c(in_interval, f_x_i >= 0.45 && 0.55 >= f_x_i)
29   }
30   pr = sum(in_interval) / length(in_interval)
31   return(c(expected_value, pr))
32 }
33
34 importanceSampling(10000)
35 importanceSampling(50000)
36 importanceSampling(100000)
```

Il codice è una funzione che, data in input la dimensione del campione, effettua l' *importance sampling* nella mistura restituendo il valore atteso sul campione e la probabilità che si trovi nell'intervallo $[0.45 - 0.55]$.
Il valore atteso calcolato analiticamente dalla mistura è il seguente:

$$\begin{aligned} E[0.3 \times \beta(5,2) + 0.7 \times \beta(2,8)] &= 0.3 \times E[\beta(5,2)] + 0.7 \times E[\beta(2,8)] \\ &= 0.3 \times \frac{5}{7} + 0.7 \times \frac{2}{10} \\ &= 0.3542857 \end{aligned}$$

Invece qui di seguito sono presentati in forma tabellare i risultati ottenuti dall'esecuzione dello script al variare della dimensione del campione:

Dimensione campione	Valore atteso	Probabilità intervallo [0.45 - 0.55]
10000	0.3560484	0.1990000
50000	0.3552311	0.2028800
100000	0.3552139	0.2034500

Come possiamo notare il valore atteso calcolato tramite *importance sampling* è abbastanza preciso in tutte e tre righe con un errore dell'ordine di 10^{-3} mentre la probabilità di finire nell'intervallo [0.45 - 0.55] è sempre intorno allo 0.2 e ciò suggerisce che la densità in questo intervallo è piuttosto alta.

Capitolo 7

Modello di regressione lineare

7.1 Esercizio su regressione lineare

Dimostrare che SSR_g (come definita a pagina 158 del libro di P. Hoff) tende a $SSR_{ols} = \sum(y_i - \hat{\beta}_{ols}\mathbf{x}_i)^2$ per $g \to \infty$.

Svolgimento

Ricordiamo che SSR_g per come è definita a pag 158 del libro di P. Hoff risulta:

$$\begin{aligned} SSR_g &= \mathbf{y}^T\mathbf{y} - \mathbf{m}^T\mathbf{V}^{-1}\mathbf{m} \\ &= \mathbf{y}^T(\mathbf{I} - \frac{g}{g+1}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{y} \end{aligned}$$

Noi vogliamo dismotrare che per $g \to \infty$ quanto appena scritto tende a

$$\begin{aligned} SSR_{ols} &= \sum(y_i - \hat{\beta}_{ols}\mathbf{x}_i)^2 \\ &= \mathbf{y}^T\mathbf{y} - 2\hat{\beta}_{ols}^T\mathbf{X}^T\mathbf{y} + \hat{\beta}_{ols}^T\mathbf{X}^T\mathbf{X}\hat{\beta}_{ols} \end{aligned}$$

dove

$$\hat{\beta}_{ols} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

Un modo per dimostrare quanto chiesto è quello di svolgere semplicemente i calcoli sostituendo $\hat{\beta}_{ols}$ alla formula di SSR_{ols}, ovvero:

$$\begin{aligned} SSR_{ols} &= \mathbf{y}^T\mathbf{y} - 2\hat{\beta}_{ols}^T\mathbf{X}^T\mathbf{y} + \hat{\beta}_{ols}^T\mathbf{X}^T\mathbf{X}\hat{\beta}_{ols} \\ &= \mathbf{y}^T\mathbf{y} - 2\underbrace{[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}]^T}_{\hat{\beta}_{ols}^T}\mathbf{X}^T\mathbf{y} + \underbrace{[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}]^T}_{\hat{\beta}_{ols}^T}\mathbf{X}^T\mathbf{X}\underbrace{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}}_{\hat{\beta}_{ols}} \end{aligned}$$

Sapendo che

$$\hat{\beta}_{ols}^T = [(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}]^T = \mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$$

allora possiamo riscrivere la formula come segue:

$$\begin{aligned} SSR_{ols} &= \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} + \mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\underbrace{\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}}_{=\mathbf{I}}\mathbf{X}^T\mathbf{y} \\ &= \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} + \mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ &= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \end{aligned}$$

Riprendendo la definizione di SSR_g per $g \to \infty$ avremo che $\frac{g}{g+1} \to 1$ e quindi

$$\begin{aligned} SSR_g &= \mathbf{y}^T(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{y} \\ &= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ &= SSR_{ols} \end{aligned}$$

Abbiamo quindi appena dimostrato che per $g \to \infty$ avremo che $SSR_g = SSR_{ols}$

Un secondo modo di dimostrare quanto chiesto è quello di utilizzare la proprietà di *idempotenza* delle matrici, ovvero una matrice A si dice *idempotente* se $A^r = A, \forall\, r \geqslant 1$.
Sappiamo infatti che la matrice $\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$ è idempodente dato che

$$\begin{aligned} \left(\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)^2 &= \left(\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)\cdot\left(\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)^T \\ &= \left(\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)\cdot\left(\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right) \\ &= \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\underbrace{\mathbf{X}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}}_{=\mathbf{I}}\mathbf{X}^T \\ &= \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T \end{aligned}$$

Procediamo adesso alla dimostrazione ricordando che:

$$\begin{aligned} \lim_{g\to+\infty} SSR_g &= \lim_{g\to+\infty} \mathbf{y}^T\left(I - \frac{g}{g+1}\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)\mathbf{y} \\ &= \mathbf{y}^T\left(I - \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)\mathbf{y} \end{aligned}$$

Possiamo perciò dimostrare quanto chiesto partendo dal limite appena calcolato

$$\begin{aligned}
\lim_{g\to+\infty} SSR_g &= \mathbf{y}^T\left(I - \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)\mathbf{y} \\
&= \mathbf{y}^T\left(I - \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T \underbrace{+\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T - \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T}_{\text{sommiamo e sottraiamo } \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T}\right)\mathbf{y} \\
&= \mathbf{y}^T\left(I - 2\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T + \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)\mathbf{y} \\
&= \mathbf{y}^T\left(I - 2\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T + \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)\mathbf{y} \\
&= \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} + \mathbf{y}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \\
&= \mathbf{y}^T\mathbf{y} - 2\beta_{ols}^T\mathbf{X}^T\mathbf{y} + \beta_{ols}^T\mathbf{X}^T\mathbf{X}\beta_{ols} \\
&= \sum_{i=1}^{n}(y_i - \beta_{ols}^T x_i)^2 \\
&= SSR_{ols}
\end{aligned}$$

che è esattaemente ciò che volevamo dimostrare.

Capitolo 8

Modelli gerarchici

8.1 Esercizio 8.2 - Hoff

(Testo dell'esercizio ripreso dal libro [1]).
Sensitivity analysis: In this exercise we will revisit the study from Exercise 5.2, in which 32 students in a science classroom were randomly assigned to one of two study methods, A and B, with $n_A = n_B = 16$. After several weeks of study were examined on hte course material, and the scores are summarized by $\{\bar{y}_A = 75.2, s_A = 7.3\}$, $\{\bar{y}_B = 77.5, s_B = 8.1\}$. We will estimate $\theta_A = \mu + \delta$ and $\theta_B = \mu - \delta$ using the two-sampled model and prior distributions of Section 8.1.

a) Let μ *sim* normal(75,100), $1/\sigma^2 \sim$ gamma(1,100) and $\delta \sim$ normal(δ_0, τ_0^2). For each combination of $\delta_0 \in \{-4, -2, 0, 2, 4\}$ and $\tau_0^2 \in \{10, 50, 100, 500\}$, obtain the posterior distribution of μ, δ and σ^2 and compute

   i. $Pr(\delta < 0|\mathbf{Y})$;

   ii. a 95% posterior confidence interval for δ;

   iii. the prior and posterior correlation of θ_A and θ_B.

b) Describe how you might use there results to convey evidence that $\theta_A < \theta_B$ to people of variety of prior opinions

Svolgimento

L'obiettivo di questo esercizio è quello di condurre un'analisi di sensitività. Ossia, si vuole valutare il cambiamento dell'inferenza al variare della specificazione

della prior. Dato il modello delle osservazioni

$$
\begin{aligned}
Y_{iA} &= \mu + \delta + \varepsilon_{iA} \\
Y_{iB} &= \mu - \delta + \varepsilon_{iB} = \varepsilon_{iB} \\
\varepsilon_{ij}|\sigma^2 &\sim i.i.d \quad N(0, \sigma^2) \quad j = A, B \\
\mu|\gamma_0^2, \mu_0 &\sim N(\mu_0, \gamma_0^2) \\
\mu|\tau_0^2, \delta_0 &\sim N(\delta_0, \tau_0^2) \\
\sigma^2|\nu_0, \sigma_0^2 &\sim \text{Inverse-Gamma}\Big(\frac{\nu_0}{2}, \frac{\nu_0\sigma^2}{2}\Big) \\
p(\mu, \delta, \sigma^2) &= p(u)p(\delta)p(\sigma^2)
\end{aligned}
$$

Possiamo rappresentare il modello col DAG in Figura 8.1.

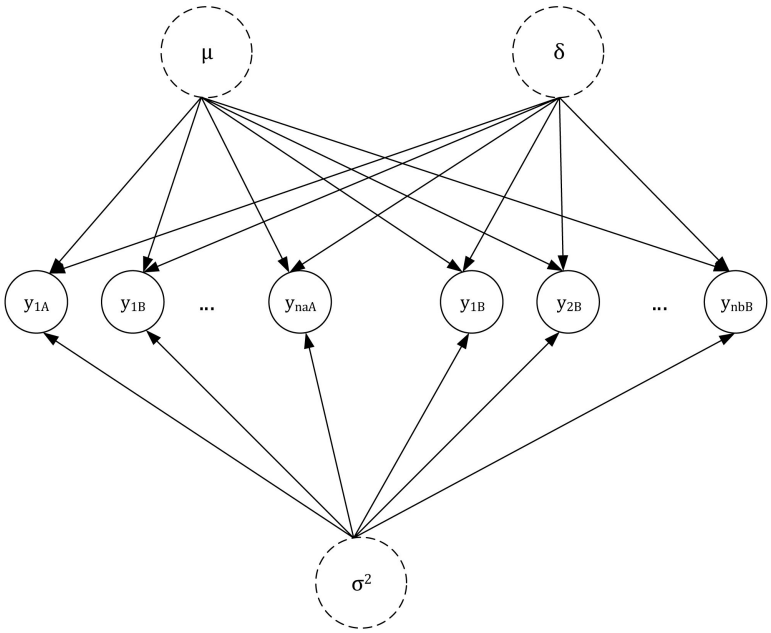


Figura 8.1: DAG: confronto tra due popolazioni normali

Assegniamo agli iperparametri i valori che sono stati richiesti

$$\begin{aligned}
\mu_0 &= 75;\\
\gamma_0^2 &= 100;\\
\frac{\nu_0}{2} &= 1 \implies \nu_0 = 2;\\
\frac{\nu_0\sigma_0^2}{2} &= 100 \implies \sigma_0^2 = 100;\\
\delta_0 &\in \{-4,-2,0,2,4\};\\
\tau_0^2 &\in \{10,50,100,500\}
\end{aligned}$$

Ricordiamo inoltre d'avere i seguenti dati campionari:

$$\left.\begin{array}{llll} \bar{y_A} = 75.2, & s_A = 7.3, & n_A = 16; \\ \bar{y_B} = 77.5, & s_B = 8.1, & n_B = 16; \end{array}\right\} \quad \Longrightarrow \quad \left\{\begin{array}{l} j = A, B \\ n_j = n = 16; \\ \bar{y_j} = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}; \\ s_j = \sqrt{\frac{1}{n_j - 1}\sum_{i=1}^{n_j}(y_{ij} - \bar{y_j})^2} \end{array}\right.$$

Prima di ottenere quanto richiesto è necessario determinare la distribuzione a posteriori delle variabili μ, δ, σ^2, ossia

$$p(\mu, \delta, \sigma^2 | y_{1A}, \ldots, y_{n_A A}, y_{1B}, \ldots, y_{n_B B}).$$

Approssimando tale distribuzione mediante campionamento MCMC, secondo l'ottica Gibbs, le distribuzioni *full conditional* dei parametri sono quindi le seguenti:

a) $(\sigma^2 | y_1 A, \ldots, y_{n_A} A, y_1 B, \ldots, y_{n_B} B, \mu, \delta) \sim \text{Inverse-Gamma}\left(\frac{\nu_n}{2}, \frac{\nu_n \sigma_n^2}{2}\right)$

dove

$$\begin{aligned}
\nu_n &= \nu_0 + n_A + n_B;\\
\nu_n\sigma_n^2 &= \nu_0\sigma_0^2 + \sum_{i=1}^{n_A}[y_{iA} - (\mu+\delta)]^2 + \sum_{k=1}^{n_B}[y_{kB} - (\mu-\delta)]^2\\
&= \nu_0\sigma_0^2 + \sum_{i=1}^{n_A}[y_{iA}^2 - 2(\mu+\delta)y_{iA} + (\mu+\delta)^2] + \sum_{k=1}^{n_B}[y_{kB}^2 - 2(\mu-\delta) + (\mu-\delta)^2]\\
&= \nu_0\sigma_0^2 + \sum_{i=1}^{n_A} y_{iA}^2 - 2(\mu+\delta)\sum_{i=1}^{n_A} y_{iA} + n_A(\mu+\delta)^2 + \sum_{k=1}^{n_B} y_{kB}^2 - 2(\mu-\delta)\sum_{k=1}^{n_B} y_{kB} + n_B(\mu-\delta)^2\\
&= \nu_0\sigma_0^2 + (n_A-1)s_A^2 + n_A\bar{y}_A^2 - 2(\mu+\delta)n_A\bar{y}_A + n_A(\mu+\delta)^2\\
&+ (n_B-1)s_B^2 + n_B\bar{y}_B^2 - 2(\mu-\delta)n_B\bar{y}_B + n_B(\mu-\delta)^2\\
&= \nu_0\sigma_0^2 + (n-1)(s_A^2 + s_B^2) + n(\bar{y}_A^2 + \bar{y}_B^2) + n[-2(\mu+\delta)\bar{y}_A + (\mu+\delta)^2 - 2(\mu-\delta)\bar{y}_B + (\mu-\delta)^2\\
&= \nu_0\sigma_0^2 + (n-1)(s_A^2 + s_B^2) + n(\bar{y}_A^2 + \bar{y}_B^2) + +2n[-\mu(\bar{y}_A + \bar{y}_B) + \delta(\bar{y}_B - \bar{y}_A) + \mu^2 + \delta^2].
\end{aligned}$$

b) $(\mu|y_1A,\ldots,y_{nA}A,y_1B,\ldots,y_{nB}B,\delta,\sigma^2)\sim N(\mu_n,\gamma_n^2)$

dove:

$$\begin{aligned}
\gamma_n^2 &= \left[\frac{1}{\gamma_0^2}+\frac{(n_A+n_B)}{\sigma^2}\right]^{-1}=\frac{\gamma_0\sigma^2}{\sigma^2+(n_A+n_B)\gamma_0^2};\\
\mu_n &= \gamma_n^2\left[\frac{\mu_0}{\gamma_0^2}+\frac{1}{\sigma^2}\sum_{i=1}^{n_A}(y_iA-\delta)+\frac{1}{\sigma^2}\sum_{k=1}^{n_B}(y_kB+\delta)\right]\\
&= \gamma_n^2\frac{\mu_0}{\gamma_0^2}+\frac{1}{\sigma^2}\left[\left(\sum_{i=1}^{n_A}y_iA-n_A\delta\right)+\left(\sum_{k=1}^{n_B}y_kB+n_B\delta\right)\right]\\
&= \gamma_n^2\left[\frac{\mu_0}{\gamma_0^2}+\frac{1}{\sigma^2}(n_A\bar{y}_A-n_A\delta+n_B\bar{y}_B+n_B\delta)\right]\\
&= \gamma_n^2\left[\frac{\mu_0}{\gamma_0^2}+\frac{n}{\sigma^2}(\bar{y}_A+\bar{y}_B)\right].
\end{aligned}$$

c) $(\delta|y_1A,\ldots,y_{n_A}A,y_1B,\ldots,y_{n_B}B,\mu,\sigma^2)\sim N(\delta_n,\tau_n^2)$

dove:

$$\begin{aligned}
\tau_n^2 &= \left(\frac{1}{\tau_0^2}+\frac{n_A+n_B}{\sigma^2}\right)^{-1}=\frac{\sigma^2\tau_0^2}{\sigma^2+(n_A+n_B)\tau_0^2};\\
\delta_n &= \tau_n^2\left[\frac{\delta_0}{\tau_0^2}+\frac{1}{\sigma^2}\sum_{i=1}^{n_A}(y_{iA}-\mu)-\frac{1}{\sigma^2}\sum_{k=1}^{n_B}(y_{kB}-\mu)\right]\\
&= \tau_n^2\left[\frac{\delta_0}{\tau_0^2}+\frac{1}{\sigma^2}\left[\sum_{i=1}^{n_A}(y_{iA}-n_A\mu)-\sum_{k=1}^{n_B}(y_{kB}-n_B\mu)\right]\right]\\
&= \tau_n^2\left[\frac{\delta_0}{\tau_0^2}+\frac{1}{\sigma^2}(n_A\bar{y_A}-n_A\mu-n_B\bar{y_B}+n_B\mu)\right]\\
&= \tau_n^2\left[\frac{\delta_0}{\tau_0^2}+\frac{n}{\sigma^2}(\bar{y_A}-\bar{y_B})\right].
\end{aligned}$$

Prima di mostrare il codice R per rispondere ai punti i), ii) e iii) calcoliamo la

correlazione a priori tra θ_A e θ_B utile al punto iii):

$$\begin{aligned} Corr(\theta_A, \theta_B) &= \frac{Cov(\theta_a, \theta_B)}{\sqrt{Var(\theta_A)Var(\theta_B)}} \\ &= \frac{Cov(\mu+\delta, \mu-\delta)}{\sqrt{Var(\mu+\delta)Var(\mu-\delta)}} \\ &= \frac{Cov(\mu,\mu) + Cov(\mu,\delta) - Cov(\mu,\delta) - Cov(\delta,\delta)}{\sqrt{[Var(\mu)+Var(\delta)][Var(\mu)+Var(\delta)]}} \\ &= \frac{Var(\mu) - Var(\delta)}{\sqrt{[Var(\mu)+Var(\delta)]^2}} \\ &= \frac{\gamma_0^2 - \tau_0^2}{\gamma_0^2 + \tau_0^2} \end{aligned}$$

e quindi

τ_0^2	10	50	100	500
$Corr(\theta_A, \theta_B)$	0.81	0.33	0	-0.67

Codice R:

```
#Quantita' campionarie a disposizione per i due gruppi (medie, deviazioni standard e numerosita'):
y.barA <- 75.2
y.barB <- 77.5
sA <- 7.3
sB <- 8.1
nA <- nB <- n_ <- 16

#Setting iperparametri delle prior: gli iperparametri media e varianza( rispettivamente chiamati delta0 e tau20) sono due vettori dal momento che l'obbiettivo dell' esercizio e' quello di valutare se e come cambia l'inferenza a seconda della prior specificata proprio su delta.

delta0 <- c(-4 , -2 , 0 , 2 , 4)
tau20 <- c(10, 50, 100, 500)
mu0 <- 75
gamma20 <- 100
v0 <- 2
sigma20 <- 100

#Valori iniziali dei parametri + numero di simulazioni:

delta <- (y.barA - y.barB) / 2
mu <- (y.barA + y.barB) / 2

#NB Si sono scelte ragionevolmente la semi differenza e la media delle medie campionarie come starting values dell'algoritmo
```

```
    rispettivamente per i parametri delta e mu. Nulla vieta pero' di
    scegliere un altro setting: se il numero di iterazioni e'
    sufficientemente elevato l'inferenza non cambia.

nsimul <- 1000

#Creiamo un array dove andremo ad immagazzinare i valori campionati
    durante l'algoritmo:
gibbs <- array(NA, c(nsimul, 3, length(delta0) * length(tau20)))

#Osservazione: e' stato creato un array di dimensioni 1000x3x20 dove 1000
     corrisponde al numero di righe (estrazioni dalla congiunta a
    posteriori), 3 corrisponde al numero di colonne (numero di variabili
     casuali della distribuzione a posteriori congiunta in esame) e 20
    corrisponde alla lunghezza della terza dimensione (tutti i possibili
     modi di specificare la coppia degli iperparametri di delta secondo
    i valori richiesti).

#Ciclo Gibbs:

v <- 1
for (j in delta0) {
  for (k in tau20) {
    for (i in 1:nsimul) {
      #Aggiorniamo sigma:

      vn <- v0 + nA + nB
       vnsigma2n <-
        v0 * sigma20 + (n_ - 1) * (sA ^ 2 + sB ^ 2) +
                            n_ * (y.barA ^ 2 + y.barB ^2) +
                        2 * n_ * (mu ^ 2 + delta ^ 2 - mu * (y.barA + y.
                            barB) + delta * (y.barB - y.barA))
      sigma2 <- 1 / rgamma(1, vn / 2, vnsigma2n / 2)

      #Aggiorniamo mu:

      gamma2n <- gamma20 * sigma2 / (sigma2 + (nA + nB) * gamma20)
      mun <- gamma2n * (mu0 / gamma20 + n_ * (y.barA + y.barB) / sigma2)
      mu <- rnorm(1, mun, sqrt(gamma2n))

      #Aggiorniamo delta:

      tau2n <- k * sigma2 / (sigma2 + (nA + nB) * k)
      deltan <- tau2n * (j / k + n_ * (y.barA - y.barB) / sigma2)
      delta <- rnorm(1, deltan, sqrt(tau2n))

      gibbs[i, , v] <- c(mu, delta, sigma2)
    }
    v <- v + 1
  }
```

```
}

colnames(gibbs) <- c("mu", "delta", "sigma2")

#a)

# Per ognuna delle possibili prior definite su delta:
#-)Probabilita' a posteriori che la semi-differenza tra le medie sia
    negativa
#-)Intervallo di credibilita' a posteriori per la semi-differenza tra le
    medie
#-)Correlazione a priori e a posteriori tra la media del primo gruppo e
    quella del secondo

probabilita.post <- correlazione.post <- matrix (NA, 5, 4)
quantili.post <- NULL
v = 1
for (i in 1:5) {
  for (j in 1:4) {
    probabilita.post[i, j] <- mean(gibbs[, 2, v] < 0)
    quantili.post <-
      rbind(quantili.post, quantile(gibbs[, 2, v], c(0.025, 0.975)))
    correlazione.post[i, j] <- cor(gibbs[, 1, v] + gibbs[, 2, v],
                                   gibbs[, 1, v] - gibbs[, 2, v])
    v <- v + 1
  }
}

rownames(probabilita.post) <-
  c("delta0=-4", "delta0=-2", "delta0=0", "delta0=2", "delta0=4")
rownames(correlazione.post) <-
  c("delta0=-4", "delta0=-2", "delta0=0", "delta0=2", "delta0=4")
colnames(probabilita.post) <-
  c("tau20=10", " tau20=50", " tau20=100", " tau20=500")
colnames(correlazione.post) <-
  c("tau20=10", " tau20=50", " tau20=100", " tau20=500")
rownames(quantili.post) <-
  c(
    "delta0=-4 tau20=10",
    " delta0=-4 tau20=50",
    "delta0=-4 tau20=100",
    "delta0=-4 tau20=500",
    "delta0=-2 tau20=10",
    "delta0=-2 tau20=50",
    "delta0=-2 tau20=100",
    "delta0=-2 tau20=500",
    "delta0=0 tau20=10",
    "delta0=0 tau20=50",
    "delta 0=0 tau20=100",
    "delta0=0 tau20=500",
```

```
    "delta0=2 tau20=10",
    "delta0=2 tau20=50",
    "delta0=2 tau20=100",
    "delta0=2 tau20=500",
    "delta0=4 tau20=10",
    "delta0=4 tau20=50",
    "delta0=4 tau20=100",
    "delta0=4 tau20=500"
  )

#i)

probabilita.post
```

```
          tau20=10 tau20=50 tau20=100 tau20=500
delta0=-4 0.894    0.818    0.811     0.804
delta0=-2 0.846    0.798    0.780     0.780
delta0=0  0.784    0.808    0.791     0.807
delta0=2  0.688    0.799    0.792     0.778
delta0=4  0.604    0.783    0.765     0.767
```

```
 #Commentiamo la matrice di probabilita' appena calcolata. Come gia'
     detto, ognuna corrisponde alla probabilita' a posteriori che la
     semi-differenza tra le medie dei due gruppi sia negativa per una
     delle 20 possibili specificazioni di delta0 e tau20, iperparametri
      del parametro semi-differenza delta. Facciamo notare come queste
     combinazioni portino a distribuzioni priori molto diverse tra loro
     : fissando tau2 e facendo variare delta0 si rappresentano opinioni
      che cambiano dal ritenere la media del gruppo A sia minore di
     quella del gruppo B (delta0=-4)al ritenere le due medie uguali (
     delta0=0) al ritenere che la media del gruppo B sia di gran lunga
     minore di quella del gruppo A (delta0=4); fissando delta0 e
     facendo variare tau20 si rappresentano distribuzioni che cambiano
     in base all' essere molto informative(tau20=10) o all' esserlo il
     meno possibile (tau20=500).
#Come ci aspettavamo, i risultati cambiano a seconda della prior
    specificata: in particolare la probabilita' che la semi-differenza
    delta sia minore di zero e' molto piu' elevata quanto piu' l'
    informazione a priori su tale parametro lo centra su valori negativi
     ed e' molto accurata (si notino casi estremi come quello in cui si
    ha il valore atteso e la varianza di delta come delta0=-4 e tau2=10
    e quello in cui, a parita' di varianza, si ha la forte opinione a
    priori che la semi-differenza tra le due medie di gruppo sia
    positiva: la probabilita' in esame scende da circa 0.9 a circa 0.78)
    .

#Rappresentiamo tutte queste probabilita' anche graficamente:
par(mfrow = c(3, 2))
labY <- expression(paste("p(", delta < 0, " | ", bold(Y), ")"))
for (i in 1:5) {
```

```
  plot(
    tau20,
    probabilita.post[i, ],
    pch = 20,
    xlab = expression(tau[0] ^ 2),
    ylab = labY,
    type = "l"
  )
  title(main = paste("delta0=", delta0[i]))
}
```

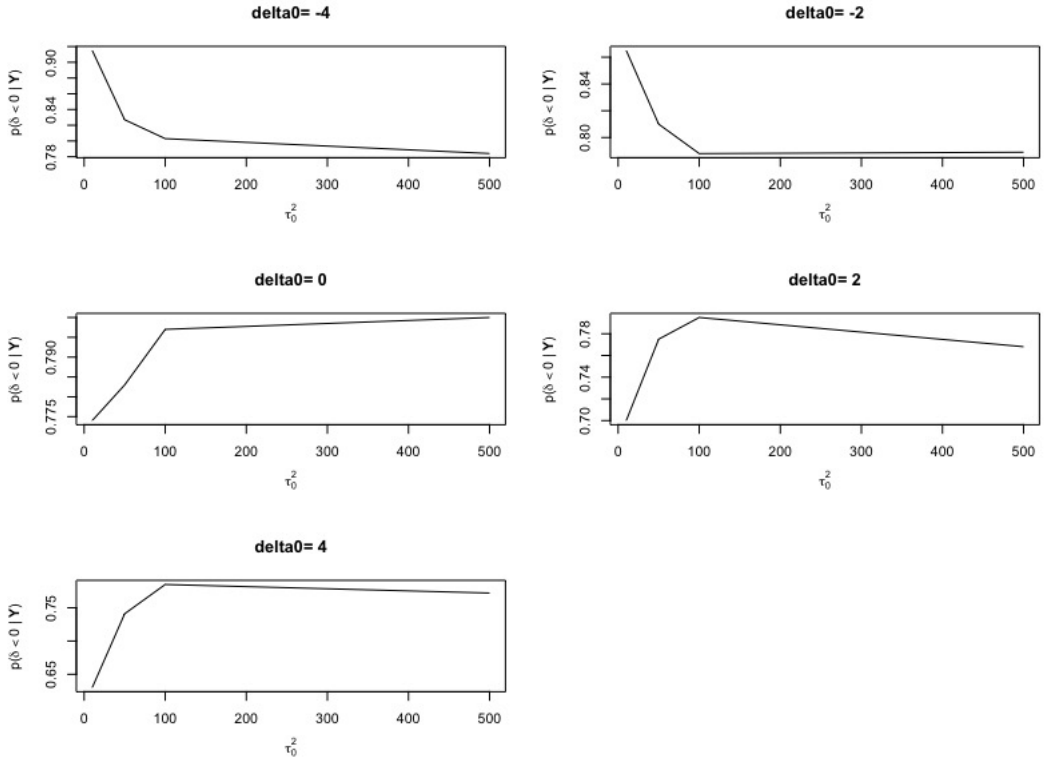


Figura 8.2: Probabilità a posteriori al variare delle prior

```
   #a)ii
   quantili.post
```

```
                          2.5%       97.5%
delta0=-4 tau20=10  -4.212939   0.8176838
delta0=-4 tau20=50  -3.958839   1.5604882
delta0=-4 tau20=100 -3.932212   1.5266037
delta0=-4 tau20=500 -3.935497   1.5035893
delta0=-2 tau20=10  -3.874308   1.1002857
delta0=-2 tau20=50  -4.032136   1.6492710
delta0=-2 tau20=100 -3.865491   1.8102229
delta0=-2 tau20=500 -4.156178   1.5988746
delta0=0 tau20=10   -3.658038   1.5326800
delta0=0 tau20=50   -4.035698   1.5638912
delta0=0 tau20=100  -4.012275   1.7541019
delta0=0 tau20=500  -4.150441   1.5241721
delta0=2 tau20=10   -3.295866   1.9141349
delta0=2 tau20=50   -3.722581   1.6005384
delta0=2 tau20=100  -3.743812   1.7210320
delta0=2 tau20=500  -3.915782   1.5726794
delta0=4 tau20=10   -2.967477   2.3844463
delta0=4 tau20=50   -3.680657   1.5563948
delta0=4 tau20=100  -3.567708   2.0593004
delta0=4 tau20=500  -3.959294   1.6453291
```

```
#Per quanto riguarda gli intervalli di credibilita' per delta si osserva
    quanto appena notato per la precedente probabilita': la
    distribuzione a priori influisce piu' dei dati sull'inferenza a
    posteriori quanto piu' e' informativa.

#a)iii

#Confrontiamo ora le correlazioni a priori e quelle a posteriori fra le
    medie dei due gruppi thetaA=mu+delta e thetaB=mu-delta. La
    correlazione a priori si puo' calcolare analiticamente e rimandiamo
    per questo a i passaggi precedenti al codice e riportiamo di seguito
     i valori per ogni combinazione degli iperparametri della prior su
    delta):

correlazione.prior <- c(0.81, 0.33, 0,-0.67)

#Le correlazioni a posteriori precedentemente calcolate invece sono:

correlazione.post
```

```
             tau20=10    tau20=50   tau20=100     tau20=500
delta0=-4  0.06158233 0.004727415  0.03245450 -0.006088312
```

```
delta0=-2   0.09445041 0.030069664 -0.02193223  -0.062654646
delta0=0    0.07866538 0.018216734 -0.06297014   0.020164257
delta0=2    0.09412201 0.041561563  0.02180072   0.034959824
delta0=4    0.08710102 0.091829449 -0.01360880   0.040456734
```

```
#Si osserva che le correlazioni a posteriori tra le medie dei due gruppi
    sono pressoche' nulle per tutte le possibili specificazioni della
    prior su delta; a priori invece si osserva che la correlazione e'
    molto alta in direzione positiva per specificazioni di prior molto
    informative e viceversa decresce fino a valori negativi per prior
    sempre piu' diffuse.


#b)

#Il confronto tra le due medie di gruppo thetaA e thetaB puo' essere
    ricondotto al parametro delta, semi-differenza tra le due medie, per
     come e' stato formulato il modello gerarchico: in questo caso
    thetaA e' minore di thetaB se delta e' minore di z e ro. Possiamo
    pertanto usare i risultati del punto precedente per descrivere come
    cambia l'evidenza che thetaA sia minore di thetaB tra persone che
    hanno opinioni molto diverse tra loro.
#Per quanto prima osservato risulta che l'inferenza a posteriori segue la
     direzione che ci aspettiamo in base alla distribuzione a priori
    specificata: le prior della semi differenza piu' informative e
    centrate su valori negativi hanno un peso maggiore dei dati e
    viceversa. Vediamolo di seguito riportando alcuni plot che mettono a
     confronto la distribuzione a priori e quella a posteriori di delta,
     in particolare quelli per delta0=-4 in cui varia tau2 e quelli per
    delta0=4 in cui varia tau2 (situazioni estreme)

#win.graph()
dev.new()
par(mfrow = c(2, 2))
x <- seq(-50, 50, by = 0.01)
v <- 0
for (j in 1:4) {
  plot(
    x,
    dnorm(x, delta0[1], sqrt(tau20[j])),
    xlim = c(-10, 10),
    ylim = c(0, 0.3),
    xlab = expression(delta),
    ylab = "density",
    type = "l",
    col = "grey"
  )
  lines(density(gibbs[, 2, j]))
  legend(
    "topright",
```

```
    legend = c("posterior", "prior"),
    lwd = c(2, 2),
    col = c("black",
            "gray"),
    bty = "n"
  )
  text(5.5, 0.15, paste("tau20=", tau20[j]))
}

#win.graph()
dev.new()
par(mfrow = c(2, 2))
x <- seq(-50, 50, by = 0.01)
v <- 0
for (j in 1:4) {
  plot(
    x,
    dnorm(x, delta0[5], sqrt(tau20[j])),
    xlim = c(-10 , 10),
    ylim = c(0, 0.3),
    xlab = expression(delta),
    ylab = "density",
    type = "l",
    col = "grey"
  )
  lines(density(gibbs[, 2, j + 16]))
  legend(
    "topright",
    legend = c("posterior", "prior"),
    lwd = c(2, 2),
    col = c("black",
            "gray"),
    bty = "n"
  )
  text(5.5, 0.15, paste("tau20=", tau20[j]))
}
```

```
#Come sempre facciamo una veloce verifica della convergenza dell'
    algoritmo:
#library(coda)
effectiveSize(gibbs[,,1])
```

```
   mu  delta  sigma2
 1000   1000    1000
```

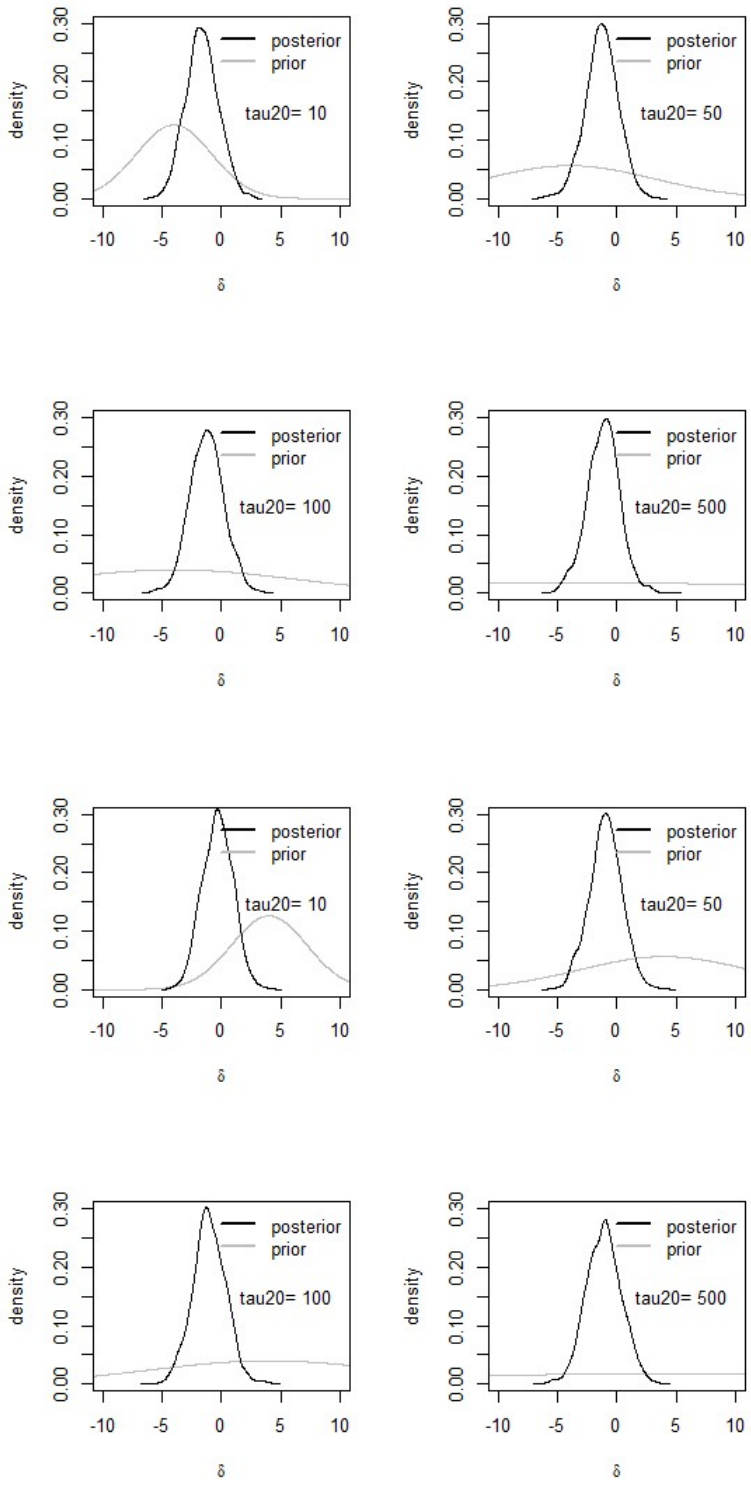


Figura 8.3: Densità a posteriori al variare della prior

Capitolo 9

Algoritmo Metropolis-Hasting

9.1 Esercizio 10.2 Hoff

(Testo dell'esercizio ripreso dal libro [1]).
Nesting success: younger male sparrows may or may not nest during a mating season, perhaps depending on their physical characteristics. Researchers have recorded the nesting success of 43 young male sparrows of the same age, as well as their wingspan, and the data appear in the file `msparrownest.dat`. Let Y_i be the binary indicator that sparrow i successfully nests, and let x_i denote their wingspan. Our model for Y_i is $logit\theta(Y_i = 1|\alpha, \beta, x_i)) = \alpha + \beta x_i$, where the logit function is given by $logit\theta = log\left[\frac{\theta}{1-\theta}\right]$.

a) Write out the joint sampling distribution $\prod_{i=1}^{n} p(y_i|\alpha, \beta, x_i)$ and simplify as much as possible.

b) Formulate a prior probability distribution over α and β by considering the range of $Pr(Y = 1|\alpha, \beta, x)$ as x ranges over 10 to 15, the approximate range of the observed wingspans.

c) Implement a Metropolis algorithm that approximates $p(\alpha, \beta|\mathbf{y}, \mathbf{x})$. Adjust the proposal distribution to achieve a reasonable acceptance rate, and run the algorithm long enough so that the effective sample size is at least $1,000$ for each parameter.

d) Compare the posterior densities of α and β to their prior densities.

e) Using output from the Metropolis algorithm, come up with a way to make a confidence band for the following function $f_{\alpha\beta}(x)$ of wingspan:

$$f_{\alpha\beta}(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

where α and β are the parameters in your sampling model. Make a plot of such a band.

Svolgimento

Possiamo modellare il problema proposto come segue:

$$Y_i = \begin{cases} 1, & \text{se l'uccellino } i \text{ nidifica} \\ 0, & \text{altrimenti} \end{cases}; \quad x_i = \text{ampiezza dell'uccellino } i; \quad i = 1, \dots, 43$$

Quindi ogni osservazione è una Bernoulli:

$$p(y_i|p_i) = p_i^{y_i}(1-p_i)^{1-y_i}$$

Ricordiamo che l'esercizio ci propone il seguente modello:

$$g(p_i) = logit(p_i) = log\left(\frac{p_i}{1-p_i}\right) = \eta_i = \alpha + \beta x_i$$

pertanto

$$g^{-1}(\eta_i) = p_i = \frac{e^{\eta_i}}{1+e^{\eta_i}} = \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}$$

a)

Scriviamo la verosomiglianza secondo il modello appena descritto, in funzione dei parameteri (indipendenti condizionatamnte) α e β:

$$\mathcal{L}(\alpha;\beta;\mathbf{y};\mathbf{X}) = p(\mathbf{y}|\alpha,\beta,\mathbf{X}) = \prod_{i=1}^{n} p(y_i|\alpha,\beta,\mathbf{x_i}) = \prod_{i=1}^{n}\left[\left(\frac{e^{\eta_i}}{1+e^{\eta_i}}\right)^{y_i}\left(\frac{1}{1+e^{\eta_i}}\right)^{1-y_i}\right]$$

$$= \prod_{i=1}^{n}\frac{e^{y_i\eta_i}}{1+e^{\eta_i}} = \prod_{i=1}^{n}(e^{y_i\eta_i} - log(1-e^{\eta_i})) = e^{\sum_{i=1}^{n}[y_i\eta_i - log(1+e^{\eta_i})]}$$

$$= e^{\sum_{i=1}^{n}[y_i(\alpha+\beta x_i) - log(1+e^{\alpha+\beta x_i})]}$$

b)

Possiamo formulare la a priori per α e β pensando che la probabilità di nidificare sia alta e che vari tra $[0.5, 0.9]$; sapendo inoltre che il campo di variazione della covariata è $[10, 15]$, troviamo il range di α e β che sia compatibile

conq uello della probabilità e in base ad esso formuliamo la prior sui parametri. In dettaglio: Pensiamo che $p = Pr(Y = 1|\alpha, \beta, \mathbf{x}) \in [0.5, 0.9]$ e quindi che $logit(p) = log\left(\frac{p}{1-p}\right) = \alpha + \beta x \in [0, 2.2]$. Si ha il seguente sistema di disequazioni:

$$\begin{cases} \alpha + \beta x \geq 0 \\ \alpha + \beta x \leq 2.2 \end{cases}$$

Troviamo il range di α e di β risolvendo il sistema per il valore minimo e per quello massimo di x:

$$\begin{cases} \alpha + 10\beta = 0 \\ \alpha + 15\beta = 2.2 \end{cases} \quad \begin{cases} \beta = 0.44 \\ \alpha = -4.4 \end{cases} \quad \begin{cases} \alpha + 15\beta = 0 \\ \alpha + 10\beta = 2.2 \end{cases} \quad \begin{cases} \beta = -0.44 \\ \alpha = -4.4 \end{cases}$$

Quindi $\beta \in [-0.44, 0.44]$ e $\alpha \in [-4.46.6]$. Ipotizzando come prior per α e β una normale (soluzione più naturale dal momento che in ogni caso non è possibile fare inferenza ne in forma chiusa ne tramite Gibbs sampler ma con un algoritmo Metropolis-Hastings), specifichiamo come vettore delle medie il centroide $(\alpha_0, \beta_o)^T = (1.1, 0)^T$. Resta da specificare la matrice di varianza e covarianza. Innanzitutto, dal momento che i valori di α e β che sono contemporaneamente massimi e contemporaneamente minimi generano valori del logit fuori dal range a priori, ipotizziamo covarianza nulla tra i due parametri in modo che i valori appena citati siano meno probabili: $\sigma_{\alpha\beta} = 0$. Per specificare le varianze σ^2_α e σ^2_β seguiamo la logica degli intervalli di confidenza: date le distribuzioni normali di α e β, sappiamo che:

$$P(\alpha_0 - 2\sigma_\alpha \leq x \leq \alpha_0 + 2\sigma_\alpha) \simeq 0.95; \qquad P(\beta_0 - 2\sigma_\beta \leq x \leq \beta_0 + 2\sigma_\beta) \simeq 0.95$$

Quindi cerchiamo le deviazioni standard in modo che

$$2\sigma_\alpha = \frac{6.6 - (-4.4)}{2} = 5.5; \qquad 2\sigma_\beta = 0.44$$

e si ha che

$$\sigma_\alpha 2.75; \qquad \sigma_\beta = 0.22$$

Per tutto quanto detto, la prior formulata considerando il range di $P(Y = 1|\alpha, \beta, x)$ al variare di x in $[10, 15]$ è:

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} \sim N_2\left(\begin{pmatrix} \alpha_0 \\ \beta_0 \end{pmatrix}, \begin{pmatrix} \sigma^2_\alpha & \sigma^2_{\alpha\beta} \\ \sigma^2_{\alpha\beta} & \sigma^2_\beta \end{pmatrix}\right) \equiv N_2\left(\begin{pmatrix} 1.1 \\ 0 \end{pmatrix} \begin{pmatrix} 2.75^2 & 0 \\ 0 & 0.22^2 \end{pmatrix}\right)$$

Rispondiamo adesso alle altre richieste dell'esercizio in R.
Di seguito il codice con output e commenti.

```
#Funzione per campionare da una distribuzione normale multivariata
rmvnorm <- function(n, mu, Sigma)
{
  E<-matrix(rnorm(n*length(mu)), n, length(mu))
```

```
  t( t(E%*%chol(Sigma)) + c(mu))
}
#Lettura dei dati:
dati<-as.matrix(dati<-
read.table(url('http://www.stat.washington.edu/people/pdhoff/Book/Data/
    hwdata/msparrownest.dat')),
  col.names=c("Y", "X"))
head(dati)
```

```
     Y     X
[1,] 0 13.03
[2,] 1 13.69
[3,] 1 12.62
[4,] 0 11.70
[5,] 0 12.39
[6,] 0 12.44
```

```
#Matrice del modello e vettore delle osservazioni:
X = cbind(1, dati[,2])
head(X)
```

```
     [,1]  [,2]
[1,]    1 13.03
[2,]    1 13.69
[3,]    1 12.62
[4,]    1 11.70
[5,]    1 12.39
[6,]    1 12.44
```

```
y = dati[,1]
head(y)
```

```
[1] 0 1 1 0 0 0
```

```
#Numero di osservazioni e numero dei parametri su cui fare inferenza:
n<-length(y)
p<-dim(X)[2]
#b)
#Si veda il setting per la formulazione della prior:
pmn.beta<-c(1.1, 0)
psd.beta<-c(2.75, 0.22)
#c)
#Creiamo una funzione che approssima la distribuzione a posteriori dei
    coefficienti di regressione mediante algoritmo di Metropolis-
    Hastings. E' richiesto di aggiustare la distribuzione proposta in
    modo che il tasso di accettazione sia ragionevole e di considerare
    un numero di iterazioni che portano ad una effective sample size di
    circa 1000.
#Osservazione:
```

```
#La matrice di varianza e covarianza della distribuzione proposal e' inserita come input della funzione in modo da vedere come cambiano i risultati al variare di questa e poterla cosi' scegliere in maniera adeguata a rispondere alle rischieste. Ricordiamo pero' che essa e' fissa per ogni catena, ovvero ogni catena ha la sua. In verita' in alcuni casi si puo' aggiustare, quando per esempio vorremmo poter fare a volte piccoli passi e a volte grandi: quindi e' possibile fare un certo numero di iterazioni con varianza piccola e un altro con varianza grande, sempre sotto la condizione che queste due varianze siano prespecificate o comunque prese random e non dipendenti dai valori estratti durante la catena. In questo modo la distribuzione proposal e' piu' flessibile e riesce ad esplorare piu' facilmente la distribuzione a posteriori. Si aumenta cosi' la cosiddetta capacita' di mixing dell'algoritmo.
metropolis <- function(tuning, nsimul) {
  #setting della distribuzione a priori:
  pmn.beta <- c(1.1, 0)
  psd.beta <- c(2.75, 0.22)
  #setting della distribuzione proposal: si sceglie una distribuzione normale multivariata con media vettore nullo e matrice di varianza e covarianza quella specificata in ingresso nella funzione.
  var.prop <- tuning
  #valore iniziale del vettore dei coefficenti:
  beta <- rep(0, p)
  #numero di simulazioni
  S <- nsimul
  #Vettore in cui immagazzino i valori campionati della distribuzione a posteriori:
  BETA <- matrix(0, nrow = S, ncol = p)
  #contatore del numero di accetazioni:
  ac <- 0
  set.seed(1)
  library(coda)
  #algoritmo metropolis (dal momento che la proposal e' simmetrica siamo in questo caso particolare di Metropolis-Hastings):
  for (s in 1:S) {
    #Proposta dei coefficenti di regressione:
    beta.p <- t(rmvnorm(1, beta, var.prop))
    #Rapporto di metropolis:
    #-) Nulla viete di lavorare con le verosogiglianze, ma usiamo le log-verosogmilianze perche' migliori da pun punto di vista computazionale.
    lhr <-
      sum(log(dbinom(y, 1, exp(X %*% beta.p) / (
        1 + exp(X %*% beta.p)
      )))) +
      dnorm(beta.p[1], pmn.beta[1], psd.beta[1], log = TRUE) +
      dnorm(beta.p[2], pmn.beta[2], psd.beta[2], log = TRUE) -
      sum(log(dbinom(y, 1, exp(X %*% beta) / (
        1 + exp(X %*% beta)
```

```
     )))) -
     dnorm(beta[1], pmn.beta[1], psd.beta[1], log = TRUE) -
     dnorm(beta[2], pmn.beta[2], psd.beta[2], log = TRUE)

   if (log(runif(1)) < lhr) {
     beta <- beta.p

     ac <- ac + 1
   }
   BETA[s, ] <- beta
 }
 Ef <- effectiveSize(BETA)
 #Tasso di accettazione ed effective sample size:
 cat("acceptance rate=", ac / S, "\n")
 cat("effective sample size=", Ef, "\n")
 return (BETA)
}
#cerchiamo adesso un numero di iterazioni una matrice di varianze e
    covarianza dalla distribuzione proposal in modo da avere un buon
    tasso di accettazione ed una effective sample size di circa 1000
    come richiesto dall'esercizio.
#Cominciamo con 1000 iterazioni (sicuramente un numero troppo ottimistico
    ) e con una matrice di varianza e covarianza simile a quella della g
    -prior:
nsimul <- 1000
var.prop <- var(log(y + 1 / 2)) * solve(t(X) %*% X)
var.prop

beta.post <- metropolis(var.prop, nsimul)
```

```
                [,1]         [,2]
[1,]        1.11413865   -0.085406852
[2,]       -0.085406852   0.006588971
```

```
beta.post <- metropolis(var.prop, nsimul)
```

```
Acceptance rate = 0.764
Effective sample size = 51.65653
```

```
#Commento ai risultati della prima catena di Metropolis :
#il tasso di accettazione e' molto alto e la numerosita' campionaria
    effettiva troppo bassa; di conseguenza si accettano molte volte i
    valori proposti ma stiamo facendo passi troppo piccoli all' interno
    della distribuzione a posteriori . Per raggiungere il nostro
    obiettivo dobbiamo allo stesso tempo aumentare il numero di
    simulazioni e aumentare i valori della matrice di varianza e
    covarianza della distribuzione proposta: in questo modo faremo passi
     piu' grandi (meno valori proposti saranno accettati) e allo stesso
    tempo l'autocorrelazione diminuira'.

```

```r
#Proviamo per quanto detto a prendere ad aumentare di 3 volte la varianza della proposal e di 5 volte il numero di simulazioni.
var.prop <- var.prop*3
nsimul <- 5000
beta.post <- metropolis( var.prop, nsimul)
```

```
Acceptance rate= 0.6474
Effective sample size= 604.3077 599.3595
```

```r
#Commento al risultato della seconda catena di Metropolis :
#L'effective sample size e' aumentata ma l'acceptance rate e' ancora abbastanza alto. Quindi possiamo mantenere lo stesso numero di simuazioni e aumentare ancora la varianza della proposal. Prendiamo 9 volte la varianza originaria
var.prop<-var.prop*3
beta.post<-metropolis(var.prop, nsimul)
```

```
Acceptance rate= 0.458
Effective sample size= 915.514 912.21
```

```r
#Commento ai risultati della terza catena di Metropolis :
#ci fermiamo dal momento che il tasso di accettazione e' sceso al di sotto del 50% e la numerosita' campionaria effettiva e' quasi pari a 1000.

#Ossevazione :
#La scelta della matrice di varianza e covarianza della distribuzione proposal e dell'effective sample size e' una questione molto euristica: in questo caso, piu' che a cambiare il numero di simulazioni, si e' pensato di cambiare il tuning parameter della distribuzione proposal ma nulla vieta di procedere nella direzione contraria.

#Per completezza, dal momento che abbiamo scelto un numero di simulazioni che e' 5 volte la numerosita' campionaria effettiva, facciamo di seguito un'operazione di thinning che corrisponde a selezionare un solo valore ogni 5 (in questo caso quindi solo 1000 valori della catenza e non 5000). In questo modo con un numero minore di valori riusciamo a dare all'incirca le stesse informazioni:

skips<-seq(5,nsimul,by=10)
plot(skips ,beta.post[skips ,1], type="l", xlab="iteration", ylab=expression(alpha))
plot(skips, beta.post[skips ,2] ,type="l",xlab="iteration", ylab= expression ( beta ) )
```

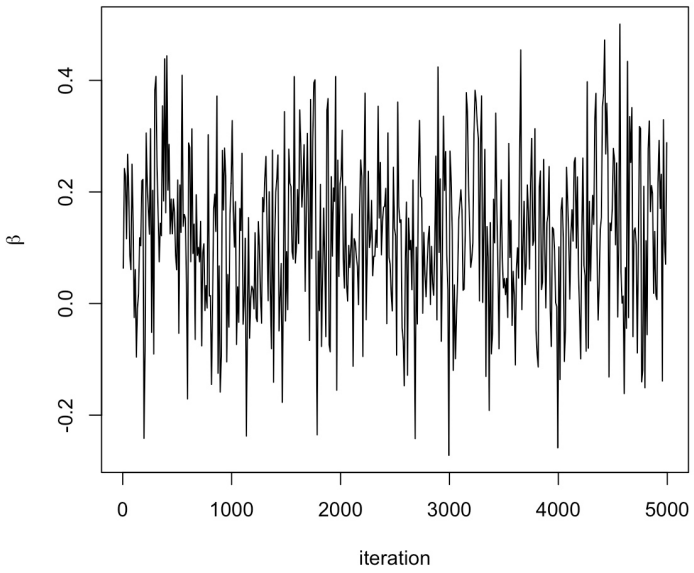


Figura 9.1: Valori β durante l'esecuzione dell'algoritmo

```
#d)
#Confronto delle distribuzioni a priori e a posteriori dei coefficienti
    di regressione:
par(mfrow=c(1,2))
x<-seq(-10, 10, by=0.1)
plot(x, dnorm(x, pmn.beta[1], psd.beta[1]) ,ylim=c(0,0.25), type="l",lwd
    =1, lty=2, xlab=expression(alpha), ylab="density")
lines(density(beta.post[ ,1]) ,lty=1)
legend("topright",legend=c("Prior","Posterior"),cex=0.7,lty=c(2,1), bty="
    n",seg.len=1.5)
y<-seq(-2,2,by=0.01)
plot(y,dnorm(y,pmn.beta[2],psd.beta[2]), ylim=c(0,3), type="l", lwd=1,lty
    =2, xlab=expression(beta), ylab="density")
lines(density(beta.post[ ,2]) ,lty=1)
legend("topright",legend=c("Prior","Posterior"),cex=0.7,lty=c(2,1),bty="n
    ", seg.len=1.5)
```

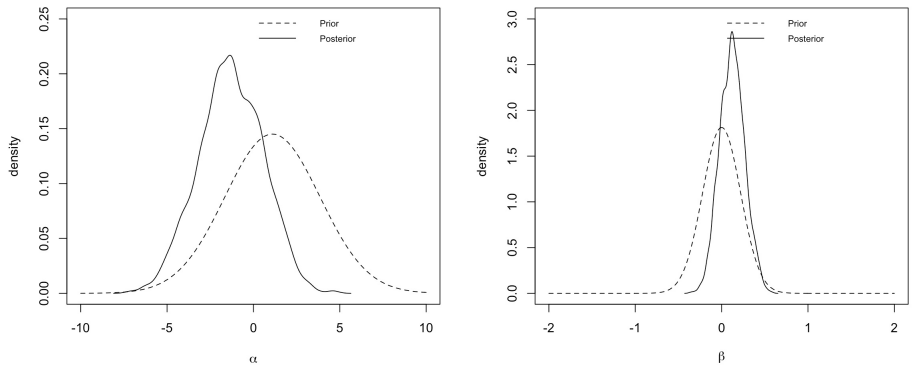


Figura 9.2: Distribuzioni a priori e a posteriori a confronto

```
#Commento ai risultati:
#Il range ipotizzato per la probabilita' di nidificare (che ricordiamo
    era [0.5, 0.9]) era troppo ottimistico rispetto ai dati e quindi
    risulta che la a posteriori da' piu' probabilita' a valori piu'
    bassi per alpha e a valori piu' alti per beta.

#e)

#Dall'output dell'algoritmo di Metropolis abbiamo una campione di coppie
    di coefficienti di regressione che approssimano la distribuzione
    corrispondente a posteriori. E' possibile pertanto approssimare una
    qualsiasi funzione di tali parametri, come la funzione f nel punto
    in esame che corrisponde alla probabilita' di nidificare.
f<-exp(t(X%*%t(beta.post))) / (1+exp(t(X%*%t(beta.post))))
#Per ogni colonna abbiamo una distribuzione condizionata ai valori delle
    x, della quale consideriamo i quantili di ordine 0.025 e 0.975:
qE<-apply(f, 2, quantile, probs=c(0.025,0.975))
#Plottiamo la banda di confidenza al livello del 95%:
par(mfrow=c(1,1))
plot( c(10,15) ,range(c(0,qE)) , type="n", xlab="wingspan", ylab="f")
lines(qE[1,], col="black", lwd=1)
lines(qE[2,], col="black", lwd=2)
```

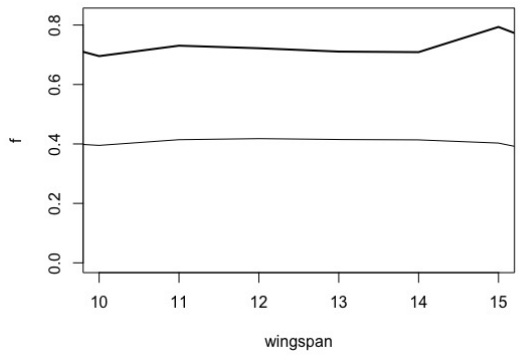


Figura 9.3: Bande di confidenza a posteriori

```
#Commento al grafico:
#Risulta non esserci nessun effetto dell'ampiezza delle ali sulla
    probabilita' di nidificare: la banda di confidenza infatti si muove
    quasi parallelamente (un po' meno nel tratto finale) all'asse delle
    ascisse.
#Riportiamo un plot della densita' empirica dei due coefficienti di
    regressione:
par(mfrow=c(1,2))
hist(beta.post[ ,2], prob=T, xlab="beta", main="Regression coefficent (
    empirical density)")
lines(density(beta.post[ ,2]))
hist(beta.post[ ,1] , prob=T, xlab="alpha", main="Intercept (empirical
    density)")
lines(density(beta.post[ ,1]))
```

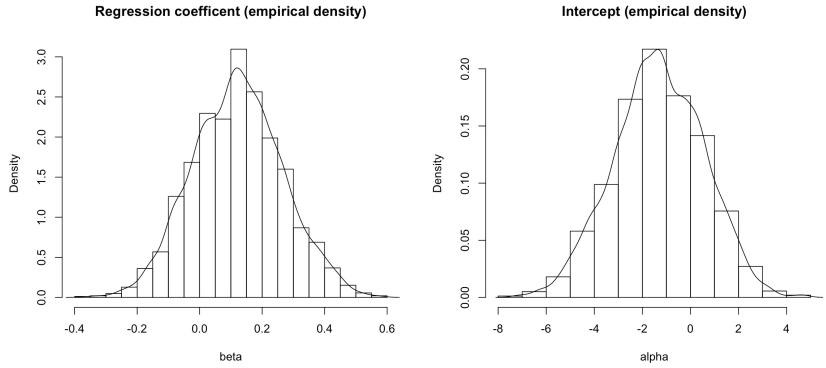


Figura 9.4: Densità empirica a posteriori

Capitolo 10

GLMM - Generalized Linear Mixed Models

10.1 Esercizio 11.2 pag. 246 Hoff

(Testo dell'esercizio ripreso dal libro [1]).
Randomized block design: researchers interested in identifying the optimal planting density for a type of perennial grass performed the following randomized experiment: ten different plots of land were each divided into eight subplots, and planting densities of 2, 4, 6 and 8 plants per square meter were randomly assigned to the subplots, so that there are two subplots at each density in each plot. At the end of the growing season the amount of plant matter yield was recorded in metric tons per hectare. These data appear in the file pdensity.dat. The researchers want to fit a model like $y = \beta_1 + \beta_2 x + \beta_3 x^2 + \epsilon$, where y is yield and x is planting density, but worry that since soil conditions vary across plots they should allow for some across-plot heterogeneity in this relationship. To accommodate this possibility we will analyze these data using the hierarchical linear model described in Section 11.1.

1. Before we do a Bayesian analysis we will get some ad hoc estimates of these parameters via least squares regression. Fit the model $y = \beta_1 + \beta_2 x + \beta_3 x^2 + \epsilon$ using OLS for each group, and make a plot showing the heterogeneity of the least squares regression lines. From the least squares coefficients find ad hoc estimates of θ and Σ. Also obtain an estimate of σ 2 by combining the information from the residuals across the groups.

2. Now we will perform an analysis of the data using the following distributions as prior distributions:

$$\Sigma^{-1} \sim \text{Wishart}(4, \hat{\Sigma}^{-1})$$

$$\theta \sim \text{multivariate normal}(\hat{\theta}, \hat{\Sigma})$$
$$\sigma^2 \sim \text{inverse-gamma}(1, \hat{\sigma}^2)$$

where $\hat{\theta}, |\hat{\Sigma}, \sigma^2$ are the estimates you obtained in a). Note that this analysis is not combining prior information with information from the data, as the "prior" distribution is based on the observed data. However, such an analysis can be roughly interpreted as the Bayesian analysis of an individual who has weak but unbiased prior information.

3. Use a Gibbs sampler to approximate posterior expectations of β

   for each group j, and plot the resulting regression lines. Compare to the regression lines in a) above and describe why you see any differences between the two sets of regression lines.

4. From your posterior samples, plot marginal posterior and prior densities of θ and the elements of Σ. Discuss the evidence that the slopes or intercepts vary across groups.

5. Suppose we want to identify the planting density that maximizes average yield over a random sample of plots. Find the value x_max of x that maximizes expected yield, and provide a 95% posterior predictive interval for the yield of a randomly sampled plot having planting density x_max.

Svolgimento

L'esercizio ha l'obiettivo di valutare la relazione tra raccolto e densità di piante relativamente a 10 lotti di terra su cui sono stati osservati i dati secondo il seguente modello:

- 10 lotti di terra.
- Ogni lotto è diviso a sua volta in 8 sottolotti.
- Densità di piantagione pari a 2, 4, 6 e 8 sono assegnate in maniera casuale tra gli 8 sottolotti di ogni lotto: in questo modo ogni lotto ha due sottolotti di ognuna delle quattro densità.

È richiesta l'analisi della relazione tra raccolto e densità mediante un modello di regressione lineare quadratica e tenendo conto della variabilità tra gruppi in termini di condizioni di suolo. Per come è descritto l'esercizio procediamo all'analisi attraverso un modello di regressione lineare gerarchico.
Il setting del modello è il seguente:

$$\underset{1*pp*1}{Y_{ij}} = \beta_j^T x_{ij} + \epsilon_{ij} \qquad \epsilon_{ij}|\sigma^2 \sim i.i.d.N(0, \sigma^2)$$

o equivalentemente

$$\left.\begin{array}{l} \underset{n_j*1}{Y_j} \,|\, \underset{p*1}{\beta_j}, \underset{n_j*p_j}{X}, \sigma^2 \sim N_{n_j}(\underset{n_j*pp*1}{X_j\beta_j}, \sigma^2 \underset{n_jnj}{I}) \\ \end{array}\right\} \quad \Longrightarrow \quad \begin{array}{r} i = 1, \ldots, n; \quad j = 1, \ldots, m. \\ \text{Modello che descrive la variabilità} \\ \text{all'interno di ogni gruppo.} \end{array}$$

$$\left.\begin{array}{r} Y_i \perp Y_j | \beta_1, \ldots, \beta_m, \sigma^2 i \neq j \\ \underset{n_j * 1}{Y_j} \; | \underset{p*1}{\beta_j} | \underset{p*1}{\theta}, \underset{p*p}{\Sigma} \sim i.i.d. N_p(\underset{p*1}{\theta}, \underset{p*p}{\Sigma}) \end{array}\right\} \Longrightarrow$$

Modello che descrive la variabilità tra gruppi.

$$\left.\begin{array}{r} \underset{p*1}{\theta} | \underset{p*p}{\mu_0}, \underset{p*p}{\Lambda_0} \sim N_p(\underset{p*1}{\mu_0}, \underset{p*p}{\Lambda_0}) \\ \underset{p*p}{\Sigma} \sim \text{Inverse - Wishart}(\eta_0, \underset{p*1}{S_0 - 1}) \\ \sigma^2 | v_0, \sigma_0^2 \sim \text{Inverse - Gamma}(\frac{v_0}{2}, \frac{v_0 \sigma_0^2}{2}) \end{array}\right\} \Longrightarrow$$

Prior semiconiugate: è possibile fare inferenza approssimando la distribuzione congiunta a posteriori $p(\sigma^2, \theta, \beta_1, \ldots, \beta_m, \Sigma | X_1, \ldots, X_m, y_1, \ldots, y_m)$ mediante Gibbs sampler

Riportiamo adesso il codice R con output e commenti necessari per rispondere ai quesiti dell'esercizio.

```r
#Funzione per campionare da una normale multivariata:
rmvnorm <-
  function(n, mu, Sigma) {
    p <- length(mu)
    res <- matrix (0, nrow = n, ncol = p)
    if (n > 0 & p > 0) {
      E <- matrix(rnorm(n * p), n, p)
      res <- t (t (E %*% chol (Sigma)) + c (mu))
      #R <- chol(A)
    }
    res
  }

#Funzione per campionare da una Wishart:
rwish <- function (n, nu0, S0)
{
  sS0 <- chol(S0)
  S <- array (dim = c (dim(S0), n))
  for (i in 1:n)
  {
    Z <-
      matrix(rnorm(nu0 * dim(S0) [1]), nu0, dim(S0) [1]) %*% sS0
    S [, , i] <- t (Z) %*% Z
  }
  S [, , 1:n]
}

#Funzione per campionare da una Inverse-Wishart:

rinvwish <- function(n, nu0, iS0)
{
  sL0 <- chol(iS0)
  S <- array (dim = c (dim(iS0), n))
  for (i in 1:n)
  {
    Z <-
```

```
    matrix (rnorm(nu0 * dim(iS0) [1]), nu0, dim(iS0) [1]) %*% sL0
  S [, , i] <- solve (t (Z) %*% Z)
 }
 S [, , 1:n]
}

#Lettura dei dati:

dati <-
 read.table(
  'https://www.stat.washington.edu/people/pdhoff/Book/Data/hwdata/
      pdensity.dat',
  header = TRUE
 )
head (dati)
```

```
plot density yield
1       2     8.25
1       2     5.81
1       4     8.69
1       4     8.03
1       6     7.96
1       6     8.89
```

```
#Calcolo di quantita' utili per ogni gruppo (numerosita', vettore delle
    osservazioni, matrice del modello e numero di parametri ):
ids <- unique (dati$plot)
m <- length (ids)
Y <- list()
X <- list()
N <- NULL
for (j in 1:m)
{
 Y[[j]] <- dati[dati[, 1] == ids[j], 3]
 N[j] <- sum(dati$plot == ids [j])
 xj <- dati [dati [, 1] == ids [j], 2]
 X[[j]] <- cbind(rep(1, N[j]), xj, xj ^ 2)
}
p <- dim(X[[1]])[2]
N
```

```
[1] 8 8 8 8 8 8 8 8 8 8
```

```
Y[[1]]
```

```
[1] 8.25 5.81 8.69 8.03 7.96 8.89 6.13 9.40
```

```
X[[1]]
```

```
     xj
[1,] 1 2 4
[2,] 1 2 4
[3,] 1 4 16
[4,] 1 4 16
[5,] 1 6 36
[6,] 1 6 36
[7,] 1 8 64
[8,] 1 8 64
```

```
#a)
#Stimiamo i coefficienti di regressione secondo il metodo OLS
    indipendentemente in ciascuno dei 10 lotti:
S2.OLS <- BETA.OLS <- NULL
for (j in 1:m) {
  fit <- lm(Y[[j]] ~ -1 + X[[j]])
  BETA.OLS <- rbind (BETA.OLS, c(fit$coef))
  S2.OLS <- c(S2.OLS, summary(fit)$sigma ^ 2)
}

colnames (BETA.OLS) <- c ("1", " x ", " x^2")
BETA.OLS
```

```
            1       x         x^2
[1,]    4.84000 1.357250 -0.1243750
[2,]    4.53375 1.193375 -0.1290625
[3,]    2.07750 2.128250 -0.1643750
[4,]    2.60375 2.114875 -0.1928125
[5,]    3.57000 1.540500 -0.1500000
[6,]    1.47375 1.930875 -0.1215625
[7,]    3.96375 1.424875 -0.1278125
[8,]    0.52375 2.941875 -0.2653125
[9,]    3.36250 1.675500 -0.1400000
[10,]   1.73875 2.241125 -0.1771875
```

```
S2.OLS
```

```
[1] 1.8005320 1.0760545 0.8134580 0.5019505 0.5886680 0.8074545 0.9575905
[8] 0.3965025 0.1328380 0.8030505
```

```
#Rappresentiamo graficamente le 10 linee di regressione per valutare la
    variabilita' tra gruppi. Riportiamo sul grafico anche la loro media.
par(mfrow = c (1, 2))
plot(
  range(dati[, 2]),
  range(dati [, 3]),
  type = "n",
  xlab = "planting density ",
  ylab = "Expected yield ",
```

```
  main = "Regression lines OLS"
)

for (j in 1:m) {
  curve (BETA.OLS[j, 1] + BETA.OLS[j, 2] * x + BETA.OLS[j, 3] * x ^ 2,
         col = "gray ",
         add = T)
}
BETA.OLS.MEAN <- apply (BETA.OLS, 2, mean)
curve (
  BETA.OLS.MEAN[1] + BETA.OLS.MEAN[2] * x + BETA.OLS.MEAN[3] * x ^ 2,
  lwd = 2,
  add = T
)

BETA.OLS.MEAN
```

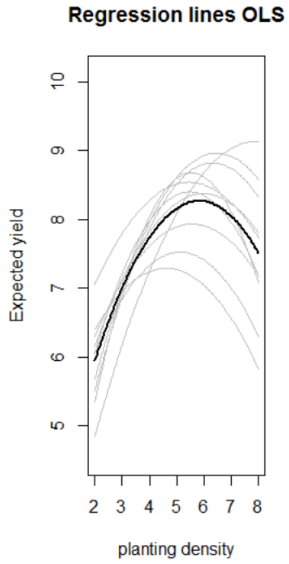


Figura 10.1: Curve di regressione con stime OLS.

```
    1       x       x^2
2.86875 1.85485 -0.15925
```

```
cov(BETA.OLS)
```

```
          1           x           x^2
1 2.00120764  -0.69321313  0.044309549
x -0.69321313   0.27555421 -0.020742679
x 2 0.04430955 -0.02074268  0.001968451
```

```r
mean(S2.OLS)
```

```
[1] 0.7878099
```

```r
#Si osserva che:
#-)Tutte le curve hanno lo stesso andamento quadratico: crescente fino ad una densita' di piante pari circa a 6 e poi decrescente.
#-)Alcuni gruppi si discostano particolarmente dalla media generale: il raccolto medio risulta o molto inferiore o molto superiore, soprattutto per i valori piu' grandi del range della covariata. Nel caso in cui i lotti avessero diverse numerosita' al loro interno potremmo immaginare che questo si verifica per i lotti meno numerosi ; in realta' per come e' strutturato l'esperimento questa variabilita' osservata e' dovuta ad altro, forse alle diverse condizioni del terreno dei 10 appezzamenti.

#b)

#Setting delle prior: impostiamo le prior semiconiugate per theta, Sigma e sigma2 come richiesto.
theta <- BETA.OLS.MEAN
Sigma <- cov(BETA.OLS)
sigma2 <- mean(S2.OLS)
eta0 <- 4
S0 <- Sigma
mu0 <- theta
L0 <- Sigma
v0 <- 2
sigma20 <- sigma2

#c)

#Impostiamo come richiesto un Gibbs sampler

#Numero di simulazioni:
nsimul = 10000

#Valori iniziali dei parametri:

beta <- BETA.OLS
Sigma <- cov(BETA.OLS)
sigma2 <- mean(S2.OLS)

#valori delle full-conditional (durante l'algoritmo)
Sigma.post <- matrix (0, p, p)
BETA.pp <- THETA.POST <- S2.POST <- NULL
BETA.POST <- BETA.OLS * 0
SIGMA.POST <- array (0, c(p, p, nsimul))
set.seed (1)

```

```
#Algoritmo Gibbs per nsimul iterazioni
for (s in 1:nsimul) {
  for (j in 1:m)
  {
    #Per ogni gruppo campioniamo i coefficienti di regressione:
    Vj <- solve (solve(Sigma) + t(X[[j]]) %*% X[[j]] / sigma2)
    Ej <-
      Vj %*% (solve(Sigma) %*% theta + t(X[[j]]) %*% Y[[j]] / sigma2)
    beta[j,] <- rmvnorm(1, Ej, Vj)
  }

  #Campioniamo la supermedia theta dei coefficienti di regressione:
  Lm <- solve(solve(L0) + m * solve(Sigma))
  mum <-
    Lm %*% (solve(L0) %*% mu0 + solve(Sigma) %*% apply(beta, 2, sum))
  theta <- t(rmvnorm(1, mum, Lm))

  #Campioniamo la matrice di varianza e covarianza Sigma dei coefficienti
       di regressione:
  mtheta <- matrix(theta, m, p, byrow = TRUE)
  Sigma <-
    solve(rwish(1, eta0 + m, solve(S0 + t(beta - mtheta) %*% (beta -
        mtheta))))
  #Campioniamo la varianza residua:
  RSS <- 0
  for (j in 1:m) {
    RSS <- RSS + sum((Y[[j]] - X[[j]] %*% beta[j,]) ^ 2)
  }
  sigma2 <-
    1 / rgamma(1, (v0 + sum(N)) / 2, (v0 * sigma20 + RSS) / 2)

  #Immagazziniamo i valori appena campionati:
  S2.POST <-
    c(S2.POST, sigma2)
  THETA.POST <- rbind (THETA.POST, t (theta))
  Sigma.post <- Sigma.post + Sigma
  BETA.POST <- BETA.POST + beta
  SIGMA.POST[, , s] <- Sigma
  #Campioniamo dalla posterior predictive dei coefficienti di regressione
       che ci servira' per il punto e).
  BETA.pp <- rbind (BETA.pp, rmvnorm(1, theta, Sigma))
}
colnames (THETA.POST) <- c(" theta1 ", " theta2 ", " theta3 ")
colnames (BETA.POST) <-
  colnames (BETA.pp) <- c(" beta1 ", " beta2 ", " beta3 ")

#Plottiamo adesso le curve di regressione con le stime dei coefficienti
    derivanti dal Gibbs sampler per poi confrontarle con quelle
    precedenti in caso di stima OLS. Anche in questo caso la media delle
     curve e' di colore nero.
```

```
82
83 BETA.PM <- BETA.POST / nsimul
84 plot (
85   range (c (0, 10)),
86   range (c (0, 10)),
87   type = "n",
88   xlab = "planting density ",
89   ylab = "Yield ",
90   main = "Bayesian regression lines "
91 )
92 for (j in 1:m) {
93   curve (BETA.PM[j, 1] + BETA.PM[j, 2] * x + BETA.PM[j, 3] * x ^ 2,
94          col = "gray ",
95          add = T)
96 }
97 curve (
98   mean(THETA.POST[, 1]) + mean(THETA.POST[, 2]) * x +
99     mean(THETA.POST[, 3]) * x ^ 2,
100  lwd = 2,
101  add = T
102 )
```

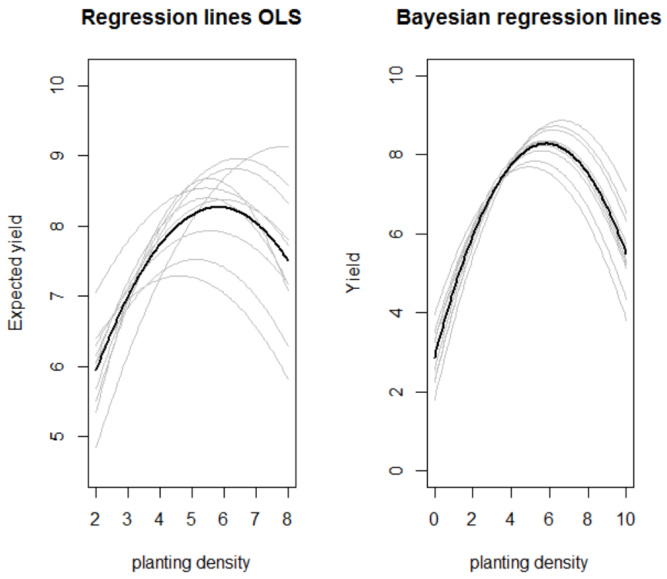


Figura 10.2: GLMM: stime dei minimi quadrati e stime bayesiane a confronto.

```
1 #Osservando i due grafici a confronto si nota che il modello gerarchico
      permette di trarre informazioni dai gruppi, riportando le curve di
      regressione lungo la media. In particolare vediamo adesso che l'
      andamento delle curve e' ancora piu' simile rispetto al caso OLS e
```

```
    che per valori piu' grandi del range di x il valore atteso del
    raccolto si discosta molto dalla media. Dal momento che lavoriamo
    con gruppi di piccola numerosita' c' una grande variabilit nelle
    stime OLS mentre nel caso del modello gerarchico i gruppi si
    influenzano (in termini di informazioni) e per l' effetto di
    shrinkage la stima OLS viene portata verso la stima media.

#Controlliamo la convergenza dell' algoritmo:

library (coda)
effectiveSize (THETA.POST)
```

```
theta1      theta2   theta3
1009.3091 742.6614 656.8365
```

```
#d)

#Approssimiamo le distribuzioni a priori tramite simulazione Monte Carlo:

n <- 10000
THETA.PRIOR <- rmvnorm(n, mu0, L0)
colnames (THETA.PRIOR) <- colnames (THETA.POST)
SIGMA.PRIOR <- rinvwish(n, eta0, solve(S0))
head(THETA.PRIOR)
```

```
           theta1      theta2     theta3
[1,]  -1.77780314    3.529232 -0.2613000
[2,]   3.47065885    1.752018 -0.1395134
[3,]   3.60717709    1.583506 -0.1350985
[4,]   0.07323521    2.928965 -0.2604753
[5,]   3.05388848    1.832459 -0.1571405
[6,]   2.27382048    1.786574 -0.1183835
```

```
head(SIGMA.PRIOR)
```

```
[1] 3.63026418 -0.93710237 0.04650011 -0.93710237 0.28263581 -0.01833983
```

```
#A priori e a posteriori di theta a confronto (plot):

par (mfrow = c(2, 2))
for (i in 1:3) {
  plot (density (THETA.POST[, i]),
        xlab = paste (" theta ", i),
        main = "")
  lines (density (THETA.PRIOR[, i]), col = "grey ")
  legend (
    "topright",
    legend = c(" Prior ", " Posterior "),
    col = c(" grey ", " black "),
    lty = 1,
```

```
    bty = "n",
    cex = 0.7
  )
}
```

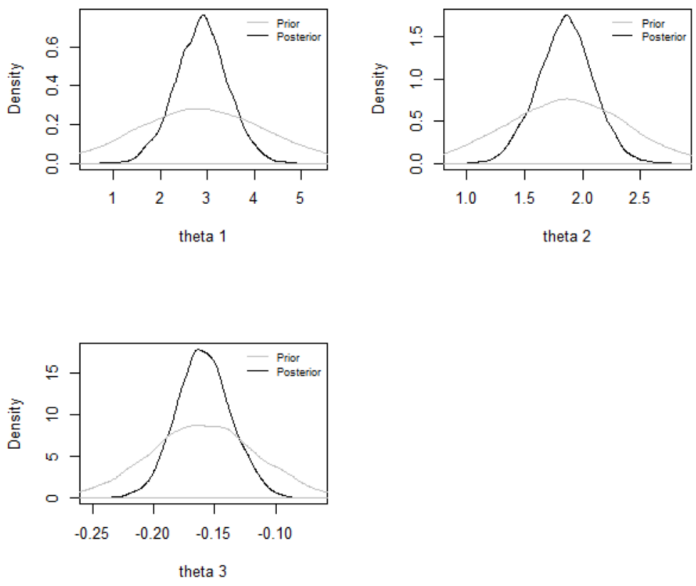


Figura 10.3: Densità a priori e a posteriori a confronto (media).

```
#Si osserva che le distribuzioni a posteriori, pur essendo centrate sulla
     stessa media delle a priori (come ci aspettavamo dal momento che
    abbiamo centrato le prior nelle stime di massima verosimiglianza),
    sono meno diffuse: in questo modo l'informazione a posteriori cambia
     nel senso che si da' maggiore probabilita' ai valori che cadono
    attorno ad essi.

#Prior e posterior di Sigma a confronto (plot):

par (mfrow = c(2, 2))
for (i in 1:3) {
  plot (
    density (log (SIGMA.POST[i, i,])),
    type = "l",
    xlab = paste ("Sigma", i),
    ylab = "density ",
    main = ""
  )
  lines (density (log (SIGMA.PRIOR[i, i,])), col = "gray ", lwd = 2)
  legend (
```

```
    "topright",
    legend = c(" Prior ", " Posterior "),
    col = c(" gray ", " black "),
    bty = "n",
    lty = 1
  )
}

Abbiamo preso il logaritmo degli elementi di Sigma dal momento che sono
    valori molto bassi e visto che e' richiesto di commentare la
    variabilita' relativa all' intercetta e alle pendenze nei gruppi,
    riportiamo i plot delle distribuzioni dei soli elementi della
    diagonale principale di Sigma.
```

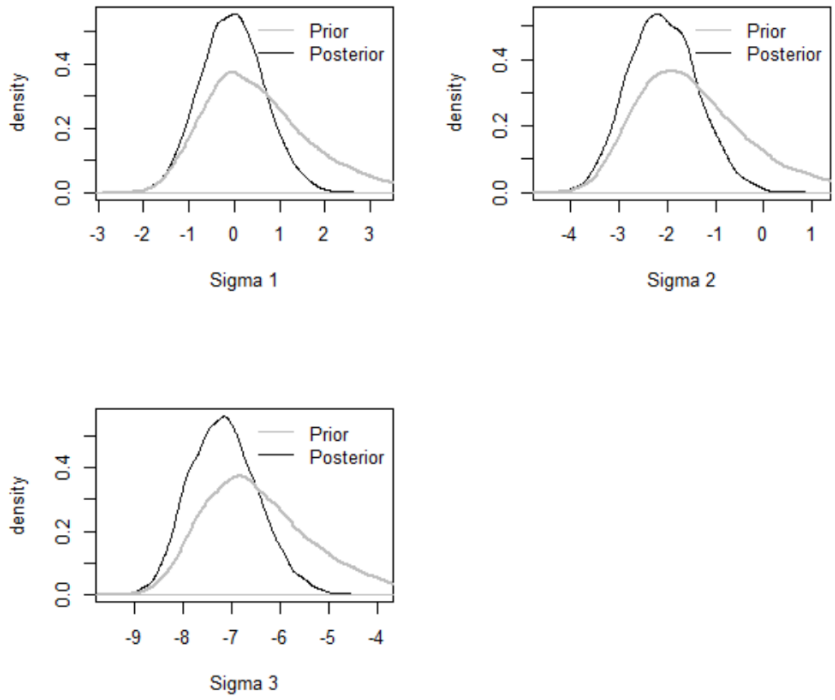


Figura 10.4: Densità a priori e a posteriori a confronto (variabilità e covariabilità).

```
#Commento al plot: C'e' evidenza che la variabilita' tra i gruppi delle
    intercette non e' molto elevata, ma diminuisce ancora per il primo e
     per il secondo coefficiente (ricordiamo che consideriamo il
    logaritmo delle varianze).

#e)

```

```
#Si vuole trovare la densita' di piante che massimizza il raccolto atteso per un campione di lotti. Ricordiamo che stiamo lavorando con un modello gerarchico e quindi vogliamo tenere in considerazione anche la variabilita' tra gruppi: ci serviamo a questo scopo della distribuzione predittiva a posteriori dei coefficienti di regressione (una normale multivariata con vettore delle medie e matrice di varianza e covarianza pari a quelli estratti ad ogni passo dell' algoritmo) per cui ogni vettore di coefficienti estratto rappresenta il vettore dei coefficienti di regressione per un gruppo futuro.
#Disponiamo gia' di tale distribuzione dal momento che e' stata calcolata durante l'algoritmo.
#E' necessario adesso confrontare 4 distribuzioni: quelle dei valori attesi del raccolto di un generico lotto, una per ogni valore osservato della covariata x. Anche in questo caso vogliamo tenere in considerazione la variabilita' tra lotti e per questo approssimiamo le distribuzioni usando la predittiva a posteriori dei coefficienti di regressione appena discussa. Cerchiamo il valore atteso di y date le y passate e le x. Un modo per farlo e' campionare i beta dalla loro predittive a posteriori, ovvero:
# 1) Campioniamo le y dati i valori della x
# 2) Facciamo la media delle distribuzioni dei valori attesi
# 3) Campioniamo i beta tilde: i beta per un gruppo futuro (per ogni valore di questo beta campionato abbiamo un valore del valore atteso per un y futuro)
# 4) Per ogni x si genera una distribuzione del valore atteso della y futuro (4 vettori).
# 5) Confrontiamo le 4 distribuzioni (per esempio con la media o vedendo la probabilita' che una sia maggiore dell'altra).
# 6) Prendendo il valore medio per ogni valore di beta tilde e per l'x max e la varianza a posteriori, campiono un valore y (cosi' si incorpora anche l'incertezza derivante dai gruppi)

x0 <- c (2, 4, 6, 8)
raccolto <- NULL
for (i in x0) {
  x <- c (1, i, i ^ 2)
  raccolto <- cbind (raccolto, BETA.pp %*% x)
}
colnames (raccolto) <- x0
head(raccolto)
```

```
          2        4        6        8
[1,] 5.989406 7.758386 8.317531 7.666840
[2,] 6.033526 7.921516 8.605911 8.086712
[3,] 6.350655 7.597301 7.712147 6.695192
[4,] 6.888414 7.776763 7.907158 7.279599
[5,] 5.973396 7.669427 8.482208 8.411740
[6,] 6.247427 7.685203 7.955526 7.058398
```

```r
#Confrontiamo adesso le quattro distribuzioni dei valori attesi,
    graficamente e con un indice sintetico, la media:

par (mfrow = c (1, 1))
hist (
  raccolto [, 1],
  prob = T,
  main = "Raccolto atteso a posteriori ",
  xlab = "Raccolto ",
  xlim = c (2, 13),
  ylim = c (0, 1.5)
)
lines (density (raccolto [, 1]))
hist (raccolto [, 2],
      prob = T,
      col = "red ",
      add = T)
lines (density (raccolto [, 2]), col = "red ")
hist (raccolto [, 3],
      prob = T,
      col = "green ",
      add = T)
lines (density (raccolto [, 3]), col = "green ")
hist (raccolto [, 4],
      prob = T,
      col = "blue ",
      add = T)
lines (density (raccolto [, 4]), col = "blue ")
legend (
  "topright",
  legend = c("x=2", "x=4", "x=6", "x=8"),
  col = c(" black ", " red ",
          "green ", " blue "),
  lty = 1
)
apply (raccolto, 2, mean)
```

```
       2        4        6        8
5.942984 7.739188 8.264142 7.517844
```

```r
#Si osserva che il valore della covariata che massimizza il raccolto
    atteso per un generico lotto e' x=6.
#Procediamo quindi considerando tale valore per il predittore lineare,
    ovvero aprrossimiamo la distribuzione predittiva a posteriori per un
     generico lotto avendo una densita' di piantagioni pari a 6. La
    logica e' la stessa di quella usata per la predittiva a posteriori
    dei coefficienti di regressione, con la differenza che adesso siamo
    al livello piu' basso della gerarchia e quindi va aggiunta un parte
    di variabilita' dovuta alle osservazioni campionarie.
```

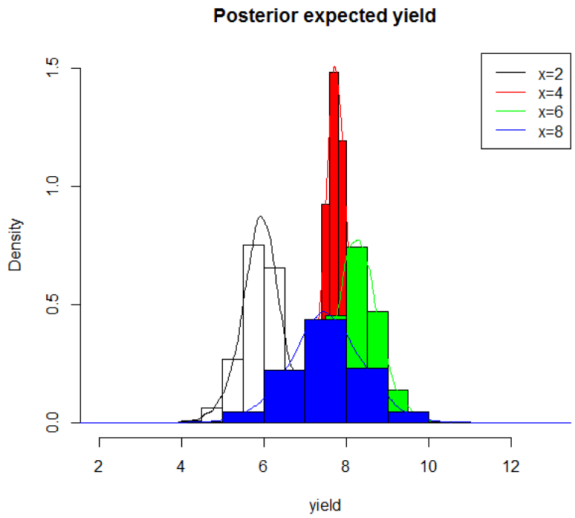


Figura 10.5: Valori attesi a posteriori per più valori di x a confronto.

```
#Per ogni vettore dei coefficienti estratto dalla predittiva a posteriori
     e per ogni elemento estratto dalla a posteriori della varianza
    residua (anche questo gia' fatto nell' algoritmo) campioniamo un
    valore del raccolto da una normale con media pari al predittore
    lineare e varianza pari alla varianza residua.
#Riportiamo infine un plot della densita' e l'intervallo di confidenza
    richiesto per tale distribuzione:

y.pred <- rnorm(nsimul, BETA.pp %*% x, S2.POST)
quantile(y.pred, c(0.025, 0.975))
```

```
    2.5%    97.5%
5.213170 9.855576
```

```
hist(y.pred,
    prob = T,
    main = "Predittiva a posteriori ",
    xlab = "raccolto ")
lines(y.pred)
```

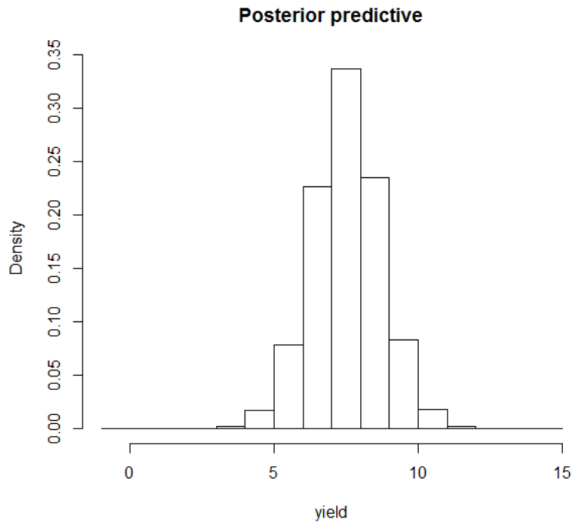


Figura 10.6: Predittiva a posteriori.

Bibliografia

[1] Peter D. Hoff. *A First Course in Bayesian Statistical Methods.* Springer, 2009.

[2] Attilio Wedlin Luciano Daboni. *Statistica.* UTET, 1982.