

UNIVERSITÀ DEGLI STUDI DI FIRENZE

CURRICULUM DATA SCIENCE

INFERENZA STATISTICA BAYESIANA

Quaderno degli esercizi

Studente

FILIPPO MAMELI

`filippo.mameli@stud.unifi.it`

Anno accademico 2017-2018

1 Prima Parte

1.1 Esercizio 1 Daboni Wedlin

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:

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

Ciò che dobbiamo calcolare sarà quindi:

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

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:

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

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

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

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:

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

1.2 Esercizio 2.6 Hoff

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:

$A \perp B|C$ significa che

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

Partendo da questa assunzione dobbiamo dimostrare che:

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

Per tutti e 3 i casi partiamo dalla parte destra dell'equazione e dimostriamo che è equivalente alla sinistra, utilizzando l'ipotesi.

1.2.1

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

1.2.2

è analogo al precedente:

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

1.2.3

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

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

1.3 Esercizio 3.7 - Hoff

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 district, 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 distribution 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*.

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

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

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

Riconosco il kernel di una distribuzione Beta(3,14) e quindi posso moltiplicare e dividere per la sua costante di normalizzazione

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

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 1 possiamo visualizzare la distribuzione appena calcolata

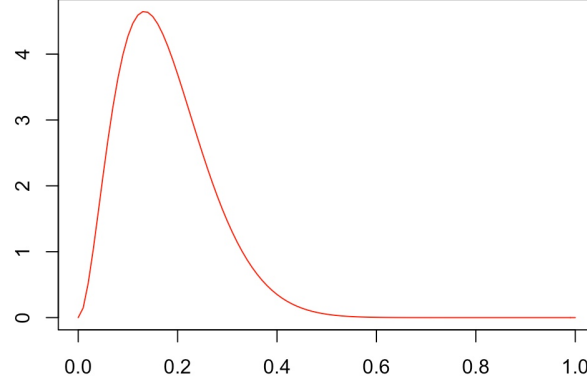


Figura 1: una Beta(3,14)

b) Affinchè

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

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

$$Y_2 \perp\!\!\!\perp Y_1|\theta$$

(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 l'equazione (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 $\text{Beta}(y_2+3, 292-y_2)$ 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) dalla quale posso facilmente calcolare media e standard deviation:

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

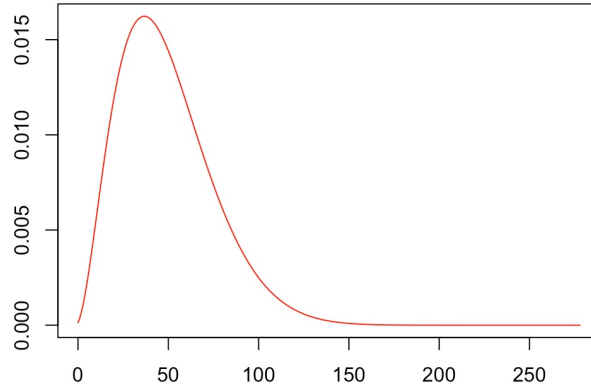


Figura 2: Beta-Binomiale(3, 14, 278)

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

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

Otteniamo quindi una distribuzione binomiale (rappresentata nella parte destra di Figura 3) con relative media e standard deviation:

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

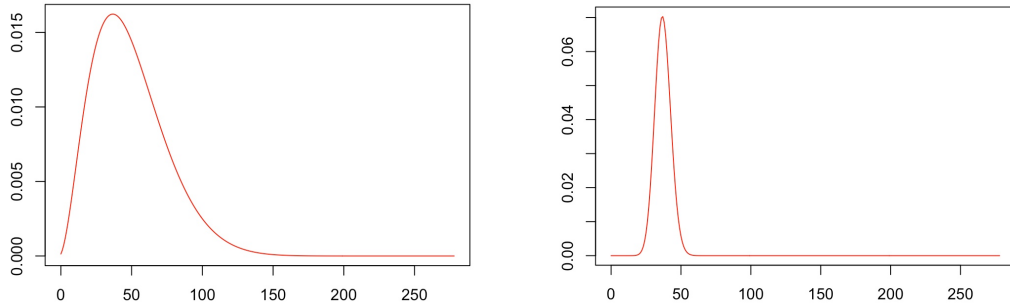


Figura 3: Una Betabinomiale(3, 14, 278) (a sinistra) e una Binomiale(2/15, 278) (a destra) a confronto

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 metodo, che ha come effetto una drastica riduzione dell'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 della nostra predizione.

1.4 Esercizio 3.14 - Hoff

Unit information prior: Let $Y_1, \dots, Y_n \sim \text{i.i.d. } p(y|\theta)$. Having observed the values $Y_1 = y_1, \dots, Y_n = y_n$, 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 $Y_1, \dots, Y_n \sim \text{i.i.d. binary}(\theta)$. 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 $-\frac{\partial^2 \log p_U(\theta)}{\partial \theta^2}$ of this density.
- c) Obtain a probability density for θ that is proportional to $p_U(\theta) \times p(y_1, \dots, y_n|\theta)$. 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 verosimiglianza per un campione di n osservazioni i.i.d. sarà

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

Calcolandone 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

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

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) L'esercizio richiede che

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

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 otterremo la unit information prior:

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

A questo punto per calcolare l'informazione di Fisher come richiesto dall'esercizio 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 otterremo 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 $p_U(\theta) \times p(y_1, \dots, y_n | \theta)$. 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:

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

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

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

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 $\hat{\theta}$:

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

Ponendola 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:

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

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 sarà:

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

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

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

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

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

1.5 Esercizio 28 ottobre informatici

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:

```
1 esercizio_informatici_19_oct = function(samples, min_unif = 0, max_
  unif = 1){
3   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)
5     return(value)
  }
7
  #x_i
9  y_unif= runif(samples, min_unif, max_unif)
  print(y_unif)
11
  #f(x_i)
13  f_x = c()
  print(f_x)
15  #g(x_i)
  g_x = c()
17  for(i in y_unif){
      f_x = c(f_x, beta_mixture(i))
19    g_x = c(g_x, dunif(i, min_unif, max_unif))
  }
21  print(f_x)
  print(g_x)
23  #w_i = f(x_i)/g(x_i)
  w = f_x / g_x
25  expected_value = sum(y_unif*w)/samples
27
  in_interval = c()
```

```

29  for(f_x_i in f_x){
      in_interval = c(in_interval, f_x_i >= 0.45 && 0.55 >= f_x_i)
31  }
      print(in_interval)
33  pr = sum(in_interval) / length(in_interval)
      return(c(expected_value, pr))
35 }

```

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 alta.

1.6 Esercizio 6 pag.237 Hoff

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 $\theta \sim \text{Gamma}(a_\theta, b_\theta)$ and let $\gamma \sim \text{Gamma}(a_\gamma, b_\gamma)$.

- 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 $E[\theta_B - \theta_A | y_A, y_B]$. 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.

Sia Y il numero dei figli.

Per quanto riguarda la popolazione A, abbiamo $Y | \theta_A \sim \text{Poisson}(\theta_A)$ con $\theta_A = \theta$.

Similmente, per la popolazione B abbiamo $Y | \theta_B \sim \text{Poisson}(\theta_B)$ con $\theta_B = \theta \times \gamma$.

Dal momento che

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

il modello delle osservazioni può essere riscritto come

$Y | \theta \sim \text{Poisson}(\theta)$ per A

$Y|\theta, \gamma \sim \text{Poisson}(\theta\gamma)$ 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 $Cov(\theta_A, \theta_B)$.

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

Tale quantità è maggiore di 0, quindi θ_A e θ_B sono linearmente dipendenti.

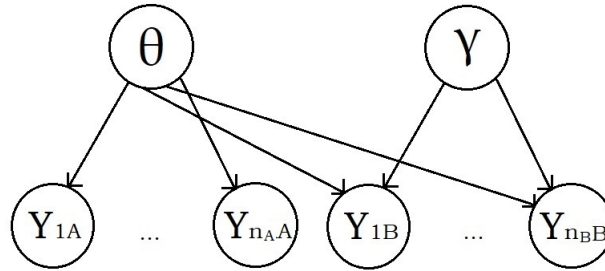
Si può verificare tale dipendenza anche con un ragionamento meno formale, considerando la seguente uguaglianza:

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

θ compare non solo nella variabile, ma anche nel "lato" della condizione. Per tale ragione, θ_A e θ_B sono dipendenti.

Adesso consideriamo la distribuzione di Poisson per analizzare degli 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. In particolare, sappiamo che $\theta_A = \theta$ e che $\theta_B = \theta\gamma$. Il numero degli eventi di B è influenzato 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, i figli di θ sono le osservazioni di

A, quelle di B e γ (questo perchè le osservazioni di B hanno come nodo padre anche γ). Inoltre θ non ha alcun nodo padre.

Quindi il Markov blanket risulta:

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

Dunque

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

Da cui

$$\theta|y_{1A}, \dots, y_{n_A A}, y_{1B}, \dots, y_{n_B B}, \gamma \sim \text{Gamma}(a_\theta + \sum_{i=1}^{n_A} y_{iA} + \sum_{j=1}^{n_B} y_{jB}, b_\theta + n_A + n_B \gamma).$$

c) Procediamo in modo simile al punto precedente, ricavando il Markov blanket di γ .

γ non ha padre, e i suoi figli sono le osservazioni di B, che a loro volta hanno come genitore anche θ (come già detto). Dunque il Markov blanket di γ è $bl(\gamma) = \{Y_{1B}, Y_{2B} \dots Y_{n_B B}, \theta\}$.

$$\begin{aligned} p(\gamma|y_{1A}, y_{2A} \dots y_{n_A A}, y_{1B}, y_{2B} \dots y_{n_B B}, \theta) &= \\ &= p(\gamma|y_{1B}, y_{2B} \dots y_{n_B B}, \theta) \propto \\ &\propto p(\gamma|a_\gamma, b_\gamma) \prod_{i=1}^{n_B} p(y_{iB}|\theta, \gamma) = \end{aligned}$$

$$\begin{aligned}
&= \frac{b_\gamma^{a_\gamma}}{\Gamma(a_\gamma)} \gamma^{a_\gamma-1} e^{-b_\gamma \gamma} \prod_{i=1}^{n_B} \frac{(\theta \gamma)^{y_{iB}} e^{-\theta \gamma}}{y_{iB}!} \propto \\
&\propto \gamma^{a_\gamma-1} e^{-b_\gamma \gamma} (\theta \gamma)^{\sum_{i=1}^{n_B} y_{iB}} e^{-n_B \theta \gamma} \propto \\
&\propto \gamma^{a_\gamma + \sum_{i=1}^{n_B} y_{iB} - 1} e^{-\gamma(b_\gamma + n_B \theta)}
\end{aligned}$$

Anche in questo caso possiamo riconoscere il kernel di una Gamma, perciò

$$\gamma | y_{1A}, \dots, y_{n_A A}, y_{1B}, \dots, y_{n_B B}, \theta \sim \text{Gamma}(a_\gamma + \sum_{i=1}^{n_B} y_{iB} + b_\gamma + n_B \theta).$$

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`.

Procediamo con l'inizializzazione dei dati:

```

>A<-c(1, 0, 0, 1, 2, 2, 1, 5, 2, 0, 0, 0, 0, 0,
+ 0, 1, 1, 1, 0, 0, 0, 1, 1, 2, 1, 3, 2, 0, 0,
+ 3, 0, 0, 0, 2, 1, 0, 2, 1, 0, 0, 1, 3, 0, 1,
+ 1, 0, 2, 0, 0, 2, 2, 1, 3, 0, 0, 0, 1, 1)

> B<-c(2, 2, 1, 1, 2, 2, 1, 2, 1, 0, 2, 1, 1, 2,
+ 0, 2, 2, 0, 2, 1, 0, 0, 3, 6, 1, 6, 4, 0, 3, 2, 0, 1,
+ 0, 0, 0, 3, 0, 0, 0, 0, 0, 1, 0, 4, 2, 1, 0, 0, 1, 0,
+ 3, 2, 5, 0, 1, 1, 2, 1, 2, 1, 2, 0, 0, 0, 2, 1,
+ 0, 2, 0, 2, 4, 1, 1, 1, 2, 0, 1, 1, 1, 1, 0, 2, 3, 2,
+ 0, 2, 1, 3, 1, 3, 2, 2, 3, 2, 0, 0, 0, 1, 0, 0,
+ 0, 1, 2, 0, 3, 3, 0, 1, 2, 2, 2, 0, 6, 0, 0, 0, 2, 0,
+ 1, 1, 1, 3, 3, 2, 1, 1, 0, 1, 0, 0, 2, 0, 2, 0,
+ 1, 0, 2, 0, 0, 2, 2, 4, 1, 2, 3, 2, 0, 0, 0, 1, 0, 0, 1,
+ 5, 2, 1, 3, 2, 0, 2, 1, 1, 3, 0, 5, 0, 0, 2,
+ 4, 3, 4, 0, 0, 0, 0, 0, 0, 2, 2, 0, 0, 2, 0, 0, 1, 1, 0,
+ 2, 1, 3, 3, 2, 2, 0, 0, 2, 3, 2, 4, 3, 3, 4,
+ 0, 3, 0, 1, 0, 1, 2, 3, 4, 1, 2, 6, 2, 1, 2, 2)

```

Adesso effettuiamo l'inizializzazione dei parametri delle varie distribuzioni, poi settiamo il numero delle iterazioni da eseguire e il seed:

```
> a_theta<-2
> b_theta<-1
> v_gamma<-c(8, 16, 32, 64, 128)
>
> n<-5000
> set.seed(150)
```

Per la distribuzione gamma facciamo uso di un vettore perchè 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

```
> 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 (5) serve per rieseguire questi passi, variando però gli iperparametri dell'a priori della distribuzione gamma.

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

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

```
> 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:

```
> gamma<-1
> theta<-1
> k<-1
```

Come passo successivo, eseguiamo l'algoritmo:

```
> 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 thetaB e thetaA, dobbiamo calcolare la trasformazione $\theta \cdot \gamma - \theta$ 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.

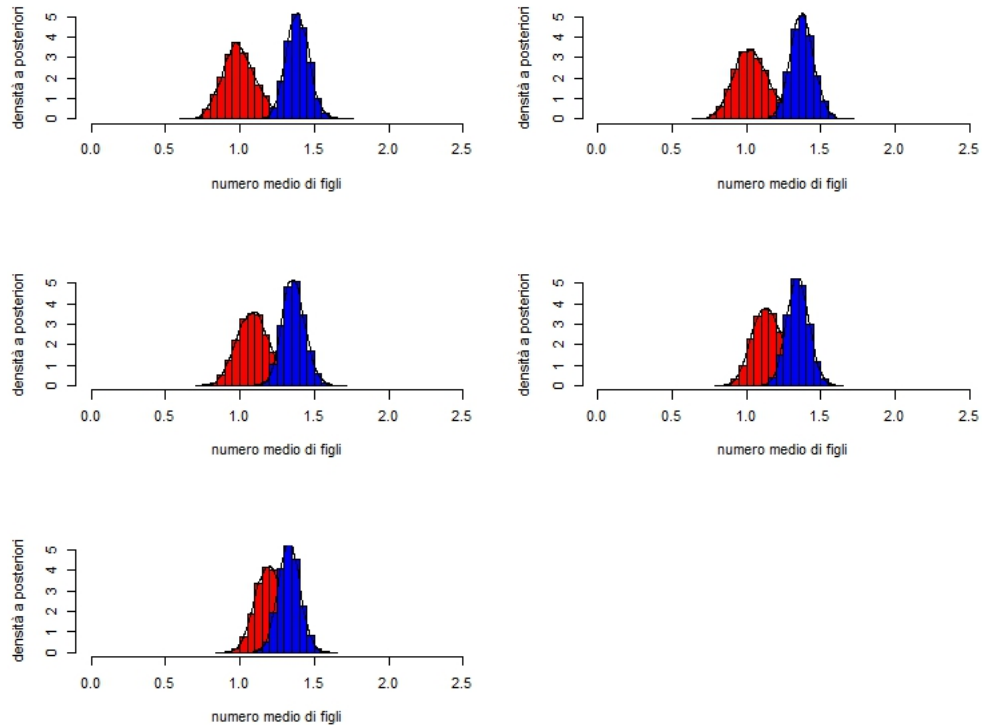
```
> for (i in 1:5){
+   valori_medie<-c(valori_medie, mean(gibbs[,1,i]*gibbs[,2,i]
+   -gibbs[,1,i]))
+ }
> valori_medie
[1] 0.3910311 0.3388716 0.2708131 0.2026279 0.1333113
```

Come si può notare nell'output, il valore atteso della differenza tra thetaB e thetaA diminuisce progressivamente all'aumentare dei parametri

dell'a priori di gamma.

Quindi, il numero di figli tra le 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="densità 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]))
+ }
> for(i in 1:5){
+ hist(gibbs[,1,i],prob=T,col="red",ylim=c(0,5),xlim=c(0,2.5),
+ ylab="densità a posteriori",xlab="numero medio di figli",main="")
+ lines(density(gibbs[,1,i]*gibbs[,2,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=",lege[i]))
+ text(2, 3.5, paste("b_gamma=",lege[i]))
+ }
>
```

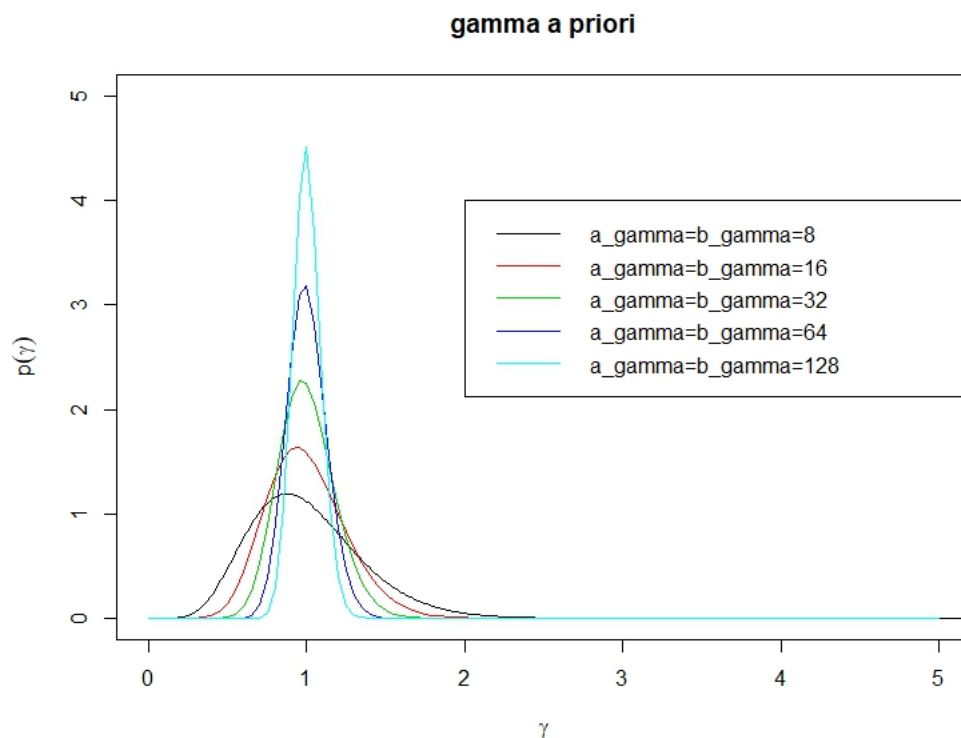


Quanto detto precedentemente, è evidente anche osservando i grafici. La parte in rosso rappresenta θ_A , quella blu θ_B . Come si può notare, le due distribuzioni sono sempre più vicine l'una all'altra all'aumentare degli iperparametri dell'a priori.

```
> 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 del numero degli iperparametri, l'apriori si "allunga" attorno a 1, ossia il valore atteso. Per quanto riguarda invece la varianza, questa risulta inversamente proporzionale al valore dei parametri. Di conseguenza, la scelta degli iperparametri influisce molto sull'inferenza. Più aumentano i valori di quest'ultimi, maggiore informazione è racchiusa nell'a priori: possiamo affermare ciò se consideriamo sia il cambiamento di tale distribuzione, sia il progressivo avvicinarsi di θ_A e θ_B . Tenderemo quindi a ignorare θ_A e θ_B , e a fare affidamento solo sull'a priori.

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

```
var1      var2
```

609.1041 605.4248

Con quest'ultime istruzioni, attraverso le quali possiamo calcolare l'effective size, siamo in grado di affermare che è stata raggiunta una condizione di equilibrio.

1.7 Esercizio 7.2 - Hoff

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 $\Psi \sim \text{Wishart}(p+1, \mathbf{S}^{-1})$, 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:

- 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 $\text{tr}(\mathbf{A}\mathbf{B})$, where $\mathbf{A} = \sum \mathbf{a}_i \mathbf{a}_i^T$.
- Let $p_U(\Sigma)$ be the inverse-Wishart density induced by $p_U(\Psi)$. Obtain a density $p_U(\boldsymbol{\theta}, \Sigma|\mathbf{y}_i, \dots, \mathbf{y}_n) \propto p_U(\boldsymbol{\theta}|\Sigma)p_U(\Sigma)p(\mathbf{y}_i, \dots, \mathbf{y}_n|\boldsymbol{\theta}, \Sigma)$. Can this be interpreted as a posterior distribution for $\boldsymbol{\theta}$ and Σ ?

Svolgimento

- La distribuzione di probabilità di una normale multivariata è

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

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

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

con relativa likelihood

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

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

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

Usando il suggerimento proposto dal testo avremo

$$\begin{aligned}-\frac{1}{2n} \sum_{i=1}^n (\mathbf{y}_i - \boldsymbol{\theta})^T \Psi (\mathbf{y}_i - \boldsymbol{\theta}) &= -\frac{1}{2n} \sum_{i=1}^n (\mathbf{y}_i - \bar{\mathbf{y}} + \bar{\mathbf{y}} - \boldsymbol{\theta})^T \Psi (\mathbf{y}_i - \bar{\mathbf{y}} + \bar{\mathbf{y}} - \boldsymbol{\theta}) \\ &= -\frac{1}{2n} \sum_{i=1}^n \left[(\mathbf{y}_i - \bar{\mathbf{y}})^T \Psi (\mathbf{y}_i - \bar{\mathbf{y}}) \right] - \frac{\mathcal{N}}{2\mathcal{N}} (\boldsymbol{\theta} - \bar{\mathbf{y}})^T \Psi (\boldsymbol{\theta} - \bar{\mathbf{y}}) \\ &= -\frac{1}{2} \text{tr}(\mathbf{S}\Psi) - \frac{1}{2} \text{tr}(\mathbf{S}_\theta \Psi)\end{aligned}$$

Quindi la log likelihood calcolata al passo precedente diventerà

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

e tornando all'esponente avremo

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

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

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

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

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

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

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

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

2 Seconda Parte

2.1 Esercizio 13 Novembre 2017

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

Svolgimento:

Per poter dimostrare l'enunciato è necessaria la proprietà di idempotenza ovvero una matrice A è idempotente se $A^r = A \forall r \geq 1$ e sapere che $\hat{\beta}_{ols} = X (X^T X)^{-1} X^T y$.

Innanzitutto risolviamo il limite:

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

Possiamo notare che la proprietà di idempotenza vale per $X (X^T X)^{-1} X^T$ dato che

$$\begin{aligned} \left[X (X^T X)^{-1} X^T \right]^2 &= X (X^T X)^{-1} X^T \cdot X (X^T X)^{-1} X^T = \\ &= X (X^T X)^{-1} \left[X^T X (X^T X)^{-1} \right] X^T = X (X^T X)^{-1} X^T \end{aligned}$$

Perciò possiamo concludere la dimostrazione con i seguenti passaggi algebrici

$$\begin{aligned} y^T \left(I - X (X^T X)^{-1} X^T \right) y &= y^T \left(I - X (X^T X)^{-1} X^T + X (X^T X)^{-1} X^T - X (X^T X)^{-1} X^T \right) y \\ &= y^T \left(I - 2X (X^T X)^{-1} X^T + X (X^T X)^{-1} X^T \right) y \\ &= y^T \left(I - X (X^T X)^{-1} X^T \right) y \\ &= y^T y - 2y^T X (X^T X)^{-1} X^T y + y^T X (X^T X)^{-1} X^T X (X^T X)^{-1} X^T y \\ &= y^T y - 2\hat{\beta}_{ols}^T X^T y + \hat{\beta}_{ols}^T X^T X \hat{\beta}_{ols} \\ &= \sum_{i=1}^n (y_i - \hat{\beta}_{ols}^T x_i)^2 \end{aligned}$$

che è ciò che volevamo dimostrare

2.2 Soluzione esercizio 8.2 Hoff

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

$$Y_{iA} = \mu + \delta + \varepsilon_{iA};$$

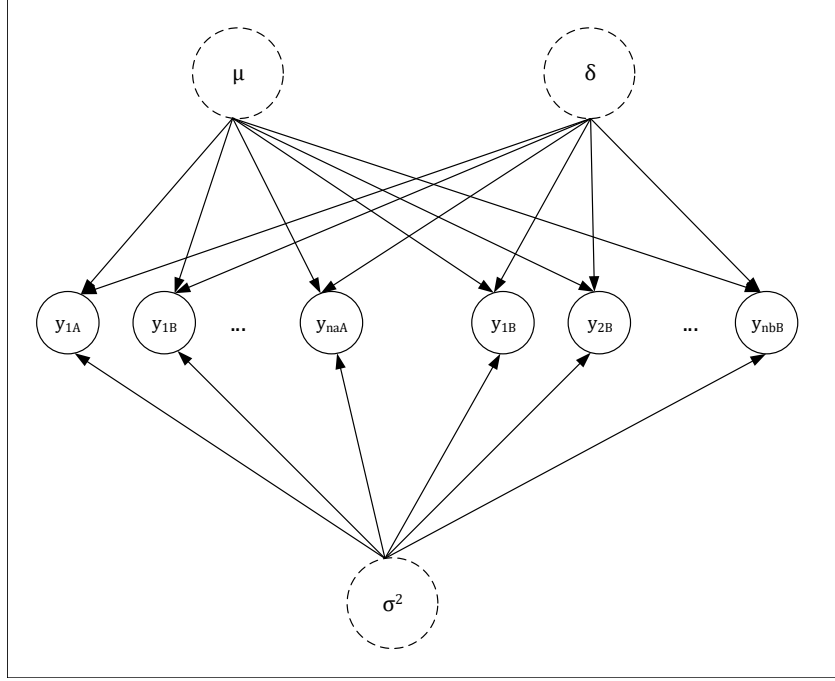


Figura 4: DAG: confronto tra due popolazioni normali

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

e la sua rappresentazione tramite DAG in Figura 4

- Assegniamo agli iperparametri i valori che sono stati richiesti

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

avendo i seguenti dati campionari

$$\bar{y}_A = 75.2, \quad s_A = 7.3, n_A = 16;$$

$$\bar{y}_B = 77.5, \quad s_B = 8.1, n_B = 16;$$

$$n_j = n = 16; \quad \bar{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}; \quad s_j = \sqrt{\frac{1}{n_j - 1} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2}$$

per fare l'inferenza richiesta e' necessario determinare la distribuzione a posteriori delle variabili μ, δ, σ^2 , ossia

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

Per far cio' approssimiamo tale distribuzione mediante campionamento MCMC, secondo l'ottica Gibbs. Le distribuzioni full conditional dei parametri sono quindi le seguenti:

$$\sigma^2 | y_{1A}, \dots, y_{n_A A}, y_{1B}, \dots, y_{n_B B}, \mu, \delta \sim \text{inverse-Gamma}\left(\frac{\nu_n}{2}, \frac{\nu_n \sigma_n^2}{2}\right)$$

in cui

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

$$\begin{aligned}
& +(n_B - 1)s_B^2 + n_B \bar{y}_B^2 - 2(\mu - \delta)n_B \bar{y}_B + n_B(\mu - \delta)^2 = \\
& +n_B(\mu - \delta)^2 = \nu_0 \sigma_0^2 + (n - 1)(s_A^2 + s_B^2) + n(\bar{y}_A^2 + \bar{y}_B^2) + \\
& +n[-2(\mu + \delta)\bar{y}_A + (\mu + \delta)^2 - 2(\mu - \delta)\bar{y}_B + (\mu - \delta)^2] = \\
& \quad \nu_0 \sigma_0^2 + (n - 1)(s_A^2 + s_B^2) + n(\bar{y}_A^2 + \bar{y}_B^2) + \\
& \quad +2n[-\mu(\bar{y}_A + \bar{y}_B) + \delta(\bar{y}_B - \bar{y}_A) + \mu^2 + \delta^2]. \\
& \mu|y_{1A}, \dots, y_{nA}, y_{1B}, \dots, y_{nB}, \delta, \sigma^2 \sim N(\mu_n, \gamma_n^2)
\end{aligned}$$

in cui

$$\begin{aligned}
\gamma_n^2 &= \left[\frac{1}{\gamma_0^2} + \frac{(n_A + n_B)}{\sigma^2} \right]^{-1} = \frac{\gamma_0 \sigma^2}{\sigma^2 + (n_A + n_B) \sigma_0^2}; \\
\mu_n &= \gamma_n^2 \left[\frac{\mu_0}{\gamma_0^2} + \frac{1}{\sigma^2} \sum_{i=1}^{n_A} (y_{iA} - \delta) + \frac{1}{\sigma^2} \sum_{k=1}^{n_B} (y_{kB} + \delta) \right] = \\
&= \gamma_n^2 \left[\frac{\mu_0}{\gamma_0^2} + \frac{1}{\sigma^2} \left[\left(\sum_{i=1}^{n_A} y_{iA} \right) - n_A \delta + \left(\sum_{k=1}^{n_B} y_{kB} + n_B \delta \right) \right] \right] = \\
&= \gamma_n^2 \left[\frac{\mu_0}{\gamma_0^2} + \frac{1}{\gamma^2} (n_A \bar{y}_A - n_A \delta + n_B \bar{y}_B + n_B \delta) \right] = \\
&= \gamma_n^2 \left[\frac{\mu_0}{\gamma_0^2} + \frac{n}{\sigma^2} (\bar{y}_A + \bar{y}_B) \right].
\end{aligned}$$

$$\delta|y_{1A}, \dots, y_{n_A A}, y_{1B}, \dots, y_{n_B B}, \mu, \delta^2 \sim N(\delta_n, \tau_n^2)$$

in cui

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

Di seguito viene mostrato il codice R che svolge i punti a i), a ii) e a iii). Prima però viene calcolata la correlazione a priori tra θ_A e θ_B per il punto a iii):

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

e quindi

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

```
#Quantita' campionarie a disposizione per i due
2 #gruppi (medie, deviazioni standard e numerosita'):
  y.barA<-75.2
4 y.barB<-77.5
  sA<-7.3
6 sB<-8.1
  nA<-nB<-n_<-16
8
#Setting iperparametri delle prior: gli iperparametri
10 #media e varianza(rispettivamente chiamati delta0 e tau20)
  #per la a priori di delta, la semidifferenza tra
12 #le medie dei due gruppi, sono due vettori dal momento
  #che l'obiettivo dell' esercizio e' quello di valutare se
14 #e come cambia l'inferenza a seconda della prior specificata
  #proprio su delta.
16
  delta0<-c(-4 , -2 , 0 , 2 , 4)
18 tau20<-c(10, 50, 100, 500)
  mu0<-75
20 gamma20<-100
  v0<-2
22 sigma20<-100
24
#Valori iniziali dei parametri + numero di simulazioni:
```

```

26 delta <-(y.barA-y.barB)/2
   mu<-(y.barA+y.barB) /2
28
   #NB Si sono scelte ragionevolmente 1 a semi differenza e la
       media delle medie
30 #campionarie come starting values dell' algoritmo
       rispettivamente per i
       #parametri delta e mu. Nulla vieta pero' di scegliere un altro
       setting: se
32 #il numero di iterazioni e' sufficientemente elevato l'
       inferenza non
       #cambia.
34
   nsimul<-1000
36
       #Creiamo un array dove andremo ad immagazzinare i valori
       campionati durante
38 #l'algoritmo:
40 gibbs<-array(NA,c(nsimul,3,length(delta0)*length(tau20)))
42 #Osservazione: e' stato creato un array di dimensioni 1000x3x20
       dove 1000
       #corrisponde al numero di righe (estrazioni dalla congiunta a
       posteriori),
44 #3 corrisponde al numero di colonne (numero di variabili
       casuali della
       #distribuzione a posteriori congiunta in esame) e 20
       corrisponde alla
46 #lunghezza della terza dimensione (tutti i possibili modi di
       specificare la
       #coppia degli iperparametri di delta secondo i valori richiesti
       ).
48 #Un altro modo di procedere poteva essere quello di creare una
       funzione che
       #prende in ingresso due valori corrispondenti ai valori degli
50 #iperparametri che assegnamo alla prior su delta e che svolge l
       ' algoritmo
       #Gibbs. Questo modo di procedere sarebbe piu' gene rale, ma

```

```

    decidiamo di
52 #lavorare con l'array dal momento che conosciamo le
    combinazioni richieste.

54 #Ciclo Gibbs:

56 v<-1
47
58 for(j in delta0){
    for(k in tau20){
60     for(i in 1:nsimul) {
        #Aggiorniamo sigma:
62
        vn<-v0+nA+nB
64 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))
66 sigma2<-1/rgamma(1,vn /2,vnsigma2n/2)

68     #Aggiorniamo mu:

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

74     #Aggiorniamo delta:

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

80     gibbs[i,,v]<-c(mu,delta,sigma2)
    }
82     v<-v+1
    }
84 }

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

88 #a )

```



```

90 #Probabilita' a posteriori che la semi-differenza tra le medie
    sia negativa
    #+ intervallo di credibilita' a posteriori per la semi-
        differenza tra le
92 #medie + correlazione a priori e a posteriori tra la media del
    primo gruppo
    #e quella del secondo (tutto per ognuna delle possibili prior
        definite su
94 #delta):

96 probabilita.post<-correlazione. post<-matrix (NA,5,4)
    quantili.post<-NULL
98 v=1
    for(i in 1:5){
100     for(j in 1:4){
        probabilita.post[i,j]<-mean(gibbs[,2,v]<0)
102     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],
104     gibbs[,1,v]-gibbs[,2,v])
        v<-v+1
106     }
    }
108
    rownames(probabilita.post)<-rownames(correlazione.post)<-
110 c("delta0=-4","delta0=-2","delta0=0","delta0=2","delta0=4")
    colnames(probabilita.post)<-colnames(correlazione.post)<-
112 c("tau20=10"," tau20=50"," tau20=100"," tau20=500")
    rownames(quantili.post)<-c("delta0=-4 tau20=10"," delta0=-4 tau
        20=50",
114 "delta0=-4 tau20=100","delta0=-4 tau20=500","delta0=-2 tau20=10
        ",
        "delta0=-2 tau20=50","delta0=-2 tau20=100","delta0=-2 tau20=500
        ",
116 "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",
118 "delta0=2 tau20=100","delta0=2 tau20=500","delta0=4 tau20=10",
        "delta0=4 tau20=50","delta0=4 tau20=100","delta0=4 tau20=500")
120

```

```

122 #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
1 #Commentiamo la matrice di probabilita' appena calcolata. Come
  gia' detto,
  #ognuna corrisponde alla probabilita' a posteriori che la semi-
  differenza
3 #tra le medie dei due gruppi sia negativa per una delle 20
  possibili
  #specificazioni di delta0 e tau20, iperparametri del parametro
5 #semi-differenza delta. Facciamo notare come queste
  combinazioni portino a
  #distribuzioni priori molto diverse tra loro: fissando tau2 e
  facendo
7 #variare delta0 si rappresentano opinioni che cambiano dal
  ritenere la
  #media del gruppo A sia minore di quella del gruppo B (delta0=-
  4)al
9 #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);
11 #fissando delta0 e facendo variare tau20 si rappresentano
  distribuzioni che
  #cambiano in base all' essere molto informative(tau20=10) o all
  ' esserlo il
13 #meno possibile (tau20=500).
  #Come ci aspettavamo, i risultati cambiano a seconda della
  prior
15 #specificata: in particolare la probabilita' che la semi-
  differenza delta
  #sia minore di zero e' molto piu' elevata quanto piu' l'

```

```

    informazione a
17 #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
19 #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
21 #due medie di gruppo sia positiva: la probabilita' in esame
    scende da circa
    #0.9 a circa 0.6).
23
    #Rappresentiamo tutte queste probabilita' anche graficamente:
25
    par(mfrow=c(3,2))
27 for(i in 1:5){
    plot(tau20,probabilita.post[i,],pch=20, xlab=expression(tau^2
    [0]),
29 ylab=expression(p(delta<0 given bold(Y)), type="l")
    legend("right",paste("delta0=",delta0[i]), bty="n")
31 }

```

```

2 #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

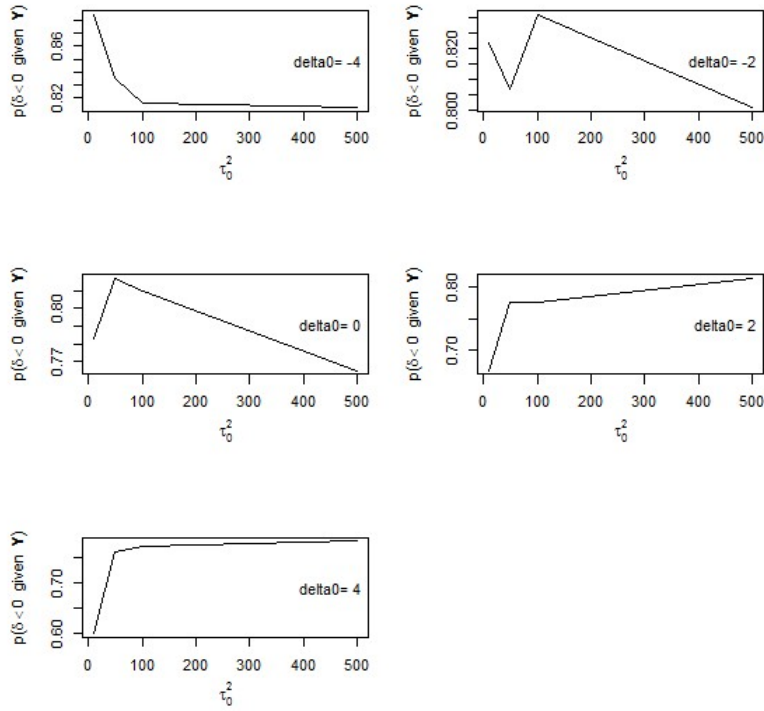


Figura 5: probabilità a posteriori al variare della prior

```

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

```

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

```

    quanto piu' e'
#informativa.
5
#a)iii
7
#Confrontiamo ora le correlazioni a priori e quelle #a
posteriori fra le
9 #medie dei due gruppi thetaA=mu+delta e thetaB=mu-delta.
#La correlazione a priori si puo' calcolare #analiticamente e
rimandiamo per
11 #questo a i passaggi precedenti al codice e riportiamo di
seguito i valori
#per ogni combinazione degli iperparametri della prior su delta
):
13
correlazione.prior<-c(0.81,0.33,0,-0.67)
15
#Le correlazioni a posteriori precedentemente calcolate invece
sono:
17 }
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
2 #sono pressoché nulle per tutte le possibili specificazioni
della prior
#su delta; a priori invece si osserva che la correlazione e'
molto alta in
4 #direzione positiva per specificazioni di prior molto
informative e
#viceversa decresce fino a valori negativi per prior sempre piu
' diffuse.
6

```

```

8 #b)

10 #Il confronto tra le due medie di gruppo thetaA e thetaB puo'
    essere
    #ricondotto al parametro delta, semi-differenza tra le due
    medie,
12 #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
14 #risultati del punto precedente per descrivere come cambia l'
    evidenza che
    #thetaA sia minore di thetaB tra persone che hanno opinioni
    molto diverse
16 #tra loro.
    #Per quanto prima osservato risulta che l'inferenza a
    posteriori segue
18 #la direzione che ci aspettiamo in base alla distribuzione a
    priori
    #specificata: le prior della semi differenza piu' informative e
    centrate su
20 #valori negativi hanno un peso maggiore dei dati e viceversa.
    Vediamolo di
    #seguito riportando alcuni plot che mettono a confronto la
    distribuzione a
22 #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
24 #estreme)

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

```

```

34 lines(density(gibbs[,2,j]))
    legend("topright",legend=c("posterior","prior"),lwd=c(2,2),
          col=c("black",
                "gray"),bty="n")
36 text(5.5,0.15,paste("tau20=",tau20[j]))
    }
38
win.graph()
40 par(mfrow=c(2,2))
    x<-seq(-50,50,by=0.01)
42 v<-0
    for(j in 1:4){
44 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")
46 lines(density(gibbs[,2,j+16]))
        legend("topright",legend=c("posterior","prior"),lwd=c(2,2),
              col=c("black",
                    "gray"),bty="n")
48 text(5.5,0.15, paste("tau20=",tau20[j]))
50 }

```

```

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

```

```

      mu  delta  sigma2
1000   1000   1000

```

Esercizio 10.2 Hoff

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

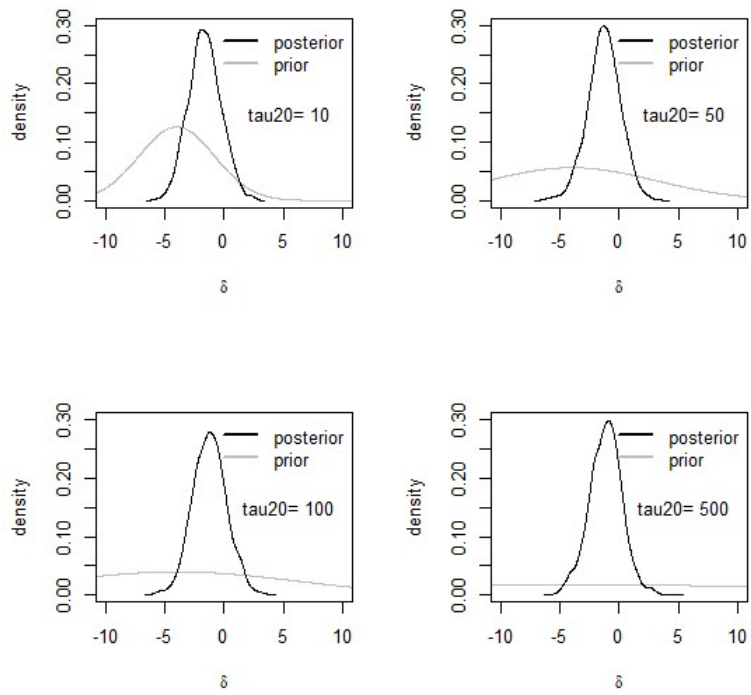


Figura 6: densita' a posteriori al variare della prior

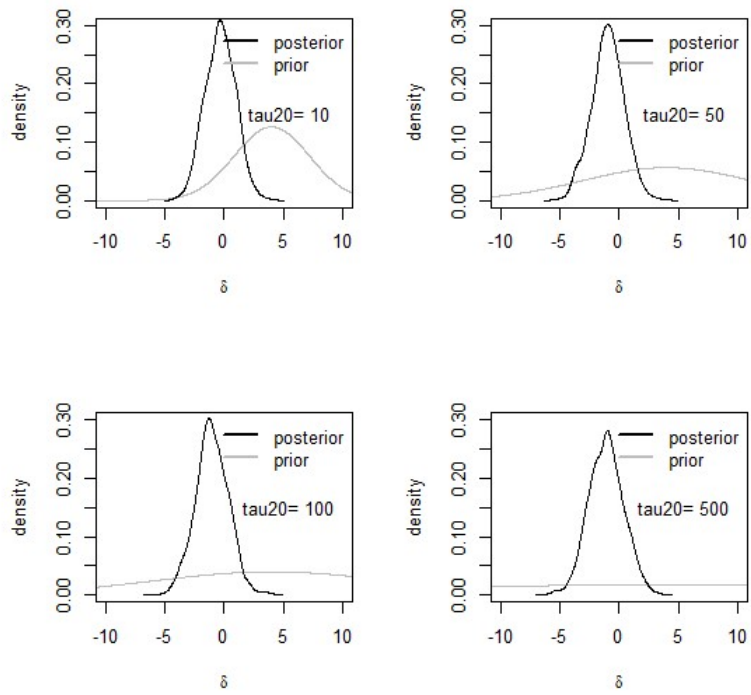


Figura 7: densita' a posteriori al variare della prior

denote their wingspan. Our model for Y_i is $\text{logit}\theta(Y_i = 1|\alpha, \beta, x_i) = \alpha + \beta x_i$, where the logit function is given by $\text{logit}\theta = \log \left[\frac{\theta}{1-\theta} \right]$.

1. Write out the joint sampling distribution $\prod_{i=1}^n p(y_i|\alpha, \beta, x_i)$ and simplify as much as possible.
2. 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.
3. 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.
4. Compare the posterior densities of α and β to their prior densities.
5. 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:

L'esercizio ha come obiettivo principale quello di studiare la relazione tra la probabilità di nidificare e l'ampiezza delle ali per un gruppo di 43 uccellini maschi della stessa età. Il setting del modello è il seguente:

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

La verosomiglianza per ogni singola osservazione è pertanto una Bernoulli:

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

Studiamo la relazione tra la probabilità di nidificare e l'ampiezza delle ali con il modello logistico (siamo quindi nell'ambito dei modelli lineari generalizzati):

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

pertanto

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

Parte a

Scriviamo la verosomiglianza secondo il modello appena descritto, quindi come funzione di α e β (ricordiamo l'indipendenza condizionata a tali parametri):

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

Potevamo in maniera analoga procedere passando direttamente attraverso la scrittura delle singole verosomiglianze nella forma della famiglia esponenziale in questo modo:

$$p(y_i|p_i) = p_i^{y_i} (1 - p_i)^{1-y_i} = e^{y_i \log(\frac{p_i}{1-p_i}) + \log(1-p_i)}$$

pertanto

$$\begin{aligned} \prod_{i=1}^n p(y_i|p_i) &= \prod_{i=1}^n e^{\sum_{i=1}^n [y_i \log(\frac{p_i}{1-p_i}) + \log(1-p_i)]} = e^{\sum_{i=1}^n [y_i \log(p_i) - \log(1-p_i)]} \\ &= e^{\sum_{i=1}^n [y_i (\alpha + \beta x_i) - \log(1 + e^{\alpha + \beta x_i})]} \end{aligned}$$

Parte b

Possiamo formulare la a priori per α e β in molti modi, due dei quali sono i seguenti:

1. **soggettivamente:** a priori pensiamo 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 con quello 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 $\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \alpha + \beta x \in [0, 2.2]$. 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 $n+$ in forma chiusa nè tramite Gibbs sampler ma con un algoritmo Metropolis-Hastings), specifichiamo come vettore delle medie il centroide $(\alpha_0, \beta_0)^T = (1.1, 0)^T$. 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_α^2 e σ_β^2 seguiamo la logica degli intervalli di confidenza: date le distribuzioni normali di α e β , sappiamo che:

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

Quindi cerchiamo le deviazioni standard in modo che

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

e si ha che

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

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:

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

2. **In maniera non informativa:** ricaviamo il range di p in base alla proporzione osservata e all'approssimazione alla distribuzione normale della proporzione campionaria. In questo caso $\hat{p} = 0.55$; poiché $\hat{p} \approx N(p, \frac{p(1-p)}{n})$, ragionando sempre secondo la logica degli intervalli di confidenza, ipotizziamo che il campo di variazione di p sia tra $\hat{p} - 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.55 - 2\sqrt{\frac{0.55 \cdot 0.45}{43}} \approx 0.47$ e $\hat{p} + 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.55 + 2\sqrt{\frac{0.55 \cdot 0.45}{43}} \approx 0.62$. Ponendo quindi $p \in [0.47, 0.62]$, impostiamo la prior per α e β secondo la logica seguita al punto precedente. Non svogliamo tutti i calcoli perché si è scelto di lavorare con la prior individuata al punto 1.

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
2 rmvnorm <- function(n, mu, Sigma)
{
4   #samples form the multivariate normal distribution
   E<-matrix(rnorm(n*length(mu)), n, length(mu))
6   t( t(E%*%chol(Sigma)) +c(mu))
}
8 #Lettura dei dati :
dati<-as.matrix(dati<-read.table(
10   "C:\\Documents\\Bene\\Desktop\\msparrownest.dat",
   col.names=c("Y" , "X"+)))
12 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 :
2 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
1 y = dati[,1]
  head(y)

[1] 0 1 1 0 0 0
#Numero di osservazioni e numero dei parametri si cui fare inferenza
:
2 n<-length(y)
  p<-dim(X)[2]
4 #b)
  #Si veda il setting per la formulazione della prior:
6 pmn.beta<-c(1.1, 0)
  psd.beta<-c(2.75, 0.22)
8 #c)
  #Creiamo una funzione che approssima la distribuzione a posteriori
    dei
10 #coefficienti di regressione mediante algoritmo di Metropolis-
    Hastings . E'
    #richiesto di aggiustare la distribuzione proposta in modo che il
      tasso di
12 #accettazione sia ragionevole e di considerare un numero di
      iterazioni che
      #portano ad una effective sample size di circa 1000: pertanto la
        funzione
14 #creata prende in ingresso questi due parametri
      #Osservazione :
16 #abbiamo appena detto che la matrice di varianza e covarianza della
      #distribuzione proposal e ' inserita come input della funzione in
        modo da
18 #vedere come cambiano i risultati al variare di questa e poterla
      cosi '
      #scegliere in maniera adeguata a rispondere alle rischieste .
        Ricordiamo
20 #pero ' che essa e ' fissa per ogni catena , ovvero ogni catena ha
      la sua . In

```

```

#verita ' in alcuni casi si puo ' anche aggiustare durante l '
  algoritmo a
22 #patto che siano soddisfatte certe condizioni : in teoria si puo'
    quindi
    #estrarre la varianza ogni volta ma a patto che la distribuzione da
        cui
24 #viene estratta non dipenda dai valori dei parametri estratti nella
    catena
    #(a parte quelli dell 'iterazione precedente).
26 #Questo succede in situazioni piu' complicate in cui a volte
    vorremmo poter > #fare 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 dipendere
    dai valori estratti durante la >#catena). In questo modo la
    distribuzione proposal e' piu' flessibile e
28 #riesce ad esplorare piu ' facilmente la distribuzione a posteriori
    . Come >#esempio concreto si puo' considerare il caso in cui la
    distribuzione a
    #posteriori di interesse e ' bimodale e quindi per approssimarla
        campionando
30 #secondo una algoritmo di Metropolis c'e' bisogno a volte di grandi
    apssi
    #per passare da una moda all'altra, altre di piccoli passi in modo
        che il
32 #tasso di accettazione non sia basso. Si aumenta cosi' la cosiddetta
    #capacita' di mixing dell'algoritmo.
34 metropolis<-function(tuning, nsimul) {
    #setting della distribuzione a priori:
36   pmn.beta<-c(1.1, 0)
    psd.beta<-c(2.75, 0.22)
38   #setting della distribuzione proposal: si sceglie come spesso si
        fa una
        #distribuzione normale multivariata con media vettore nullo e
            matrice
40   #di varianza e covarianza quella specificata in ingresso nella
        funzione.
    var.prop<-tuning
42   #valore inicial del vettore dei coefficienti:

```

```

beta<-rep(0, p)
44 #numero di simulazioni
S<-nsimul
46 #Vettore in cui immagazzino i valori campionati della
    distribuzione a posteriori:
BETA<-matrix(0, nrow=S, ncol=p)
48 #contatore del numero di accetazioni:
ac<-0
50 set.seed(1)
library(coda)
52 #algoritmo metropolis (dal momento che la proposal e' simmetrica
    siamo in
#questo caso particolare di Metropolis-Hastings):
54 for(s in 1:S){
    #proposta dei coefficienti di regressione:
56 beta.p<-t(rmvnorm(1, beta, var.prop))
    #Rapporto di metropolis:
58 #puo' essere calcolato in piu' modi tra cui usare la forma
    funzionale
    #della verosomiglianza trovata al punto a (e' scritto per
    completezza in
60 #commento) o usando la funzione dbinom di R. Facciamo
    relativamente a
    #questa alcune considerazioni:
62 #-) la funzione prende in ingresso tre elementi: il vettore y
    delle
    #osservazioni, la dimensione n e il vettore della probabilita'
    p. In
64 #questo caso, dal momento che la dimensione specificata e' 1,
    la funzione
    #restituisce un vettore della stessa dimensione del vettore
    risposta il cui
66 #i-mo elemento e' il valore della densita' Bernoulliana
    calcolata
    #nell'elemento i-mo di y con probabiita' pari all'i-mo elemento
    del
68 #vettore delle probabilita' p. La somma dei logaritmi di tali
    densita' e' la
    #log- verosomiglianza.
70 #-) Il vettore di probabilita' specificato nella funzione

```



```

    dbinom e'
    #calcolato secondo la formulazione del modello logistico come
    xpit del
72    #predittore lineare. Il rapporto di metropolis contiene due
    #log-verosomiglianze: una calcolata considerando come vettore
    di
74    #probabilita' quello calcolato con il vettore dei coefficienti
    #all'iterazione corrente e una considerando quello calcolato
76    #all'iterazione precedente.
    #-) Il rapporto di metropolis, oltre al rapporto tra
    verosomiglianze,
78    #contenga nche il rapporto tra la densita' della prior
    specificata per
    #i coefficienti di regressione valutata all'iterazione corrente
    e quella
80    #valutata all'iterazione precedente. Usiamo l'opzione log= TRUE
    perche'
    #lavoriamo con log-verosomiglianze.
82    #-) Nulla viete di lavorare con le verosomiglianze, ma usiamo
    le
    #log-verosomiglianze perche' migliori da un punto di vista
84    #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)
86  }
    if(log(runif(1))<lhr){beta<-beta.p; ac<-ac+1}BETA[s,<-beta
88  Ef<-effectiveSize(BETA)
    #Tasso di accettazione ed effective sample size:
90  cat("acceptance rate=", ac/S, "\n")
    cat("effective sample size=", Ef, "\n")
92  return (BETA)
}
94  #cerchiamo adesso un numero di iterazioni una matrice di varianze
    e
    #covarianza dalla distribuzione proposal in modo da avere un buon
    tasso di

```

```

96  #accettazione ed una effective sample size di circa 1000 come
    richiesto
    #dall'esercizio
98  #cominciamo con 1000 iterazioni (sicuramente un numero troppo
    ottimistico)
    #e con una matrice di varianza e covarianza simile a quella della
    g-prior:
100  nsimul<-1000
    var.prop <-var(log(y+1/2))*solve(t(X)%*%X)
102  var.prop
104
}

[,1] [,2]
[1,] 1.11413865 -0.085406852
[2,] -0.085406852 0.006588971
1 beta.post <- metropolis(var.prop, nsimul)

Acceptance rate = 0.764
Effective sample size = 51.65653

```

Esercizio 11.2 pag. 246 Hoff

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 $\hat{\theta}, |\hat{\Sigma}, \hat{\sigma}^2$ 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} .

Premessa e indicazioni generali

L'esercizio ha in generale 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 disegno a blocchi randomizzato:

- 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 costruito il disegno e per come è formulato il modello procediamo nell'analisi attraverso un modello di regressione lineare gerarchico. Il setting del modello è:

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

$1 \times pp \times 1$

o equivalentemente

$$Y_j | \beta_j, X_j, \sigma^2 \sim N_{n_j} \left(X_j \beta_j, \sigma^2 I \right) \quad i = 1, \dots, n; j = 1, \dots, m.$$

$n_j \times 1 \quad p \times 1 \quad n_j \times p_j \quad n_j \times pp \times 1 \quad n_j \times n_j$

Modello che descrive la variabilità all'interno di ogni gruppo.

$$Y_i \perp Y_j | \beta_1, \dots, \beta_m, \sigma^2, i \neq j$$

$$\beta_j | \theta, \Sigma \sim i.i.d. N_p \left(\theta, \Sigma \right) \quad \text{Modello che descrive la variabilità tra gruppi.}$$

$p \times 1 \quad p \times 1 \quad p \times p \quad p \times 1 \quad p \times p$

$$\theta | \mu_0, \Lambda_0 \sim N_p(\mu_0, \Lambda_0)$$

$p \times 1 \quad p \times p \quad p \times p \quad p \times 1 \quad p \times p$

$$\Sigma \sim \text{Inverse - Wishart}(\eta_0, S_0 - 1) \quad \text{Prior semiconiugate: è possibile fare inferenza}$$

$p \times p \quad p \times 1$

approssimando la distribuzione congiunta a posteriori

$p(\sigma^2, \theta, \beta_1, \dots, \beta_m, \Sigma | X_1, \dots, X_m, y_1, \dots, y_m)$
mediante Gibbs sampler.

$\sigma^2|v_0, \sigma_0^2 \sim \text{Inverse - Gamma}(\frac{v_0}{2}, \frac{v_0\sigma_0^2}{2})$

Per il DAG si faccia riferimento al grafo dell'esercizio precedente. Riportiamo adesso il codice R con output e commenti necessari per rispondere ai quesiti dell'esercizio.

```

1 #Funzione per campionare da una normale multivariata:
  rmvnorm<-
3 function(n,mu, Sigma ) {
  p<-length(mu)
5   res<-matrix( 0, nrow=n, ncol=p)
  if ( n>0 & p>0 ) {
7     E<-matrix( rnorm(n*p),n, p)
     res<-t ( t (E%%chol ( Sigma ) ) +c (mu) )
9     #R <- chol(A)
  }
11 res
  }
13
  #Funzione per campionare da una Wishart:
15 rwish<-function (n, nu0, SO )
  {
17   sSO <- chol ( SO )
   S<-array ( dim=c ( dim( SO ),n ) )
19   for ( i in 1: n)
   {
21     Z <-matrix ( rnorm( nu0 * dim( SO ) [ 1 ] ), nu0, dim( SO ) [
        1 ] ) %% sSO
        + S [,, i ]<-t (Z)%%Z
23   }
   S [,, 1: n ]
25 }

27 #Funzione per campionare da una Inverse-Wishart:

29 rinvwish<-function(n, nu0, iSO )
  {
31   sLO <- chol ( iSO )
   S<-array ( dim=c ( dim( iSO ),n ) )
33   for ( i in 1: n)
   {

```

```

35   Z <- matrix ( rnorm( nu0 * dim( iS0 ) [ 1 ] ), nu0, dim( iS0 ) [
      1 ] ) %*% sL0
      S [,, i ]<- solve ( t (Z)%*%Z)
37 }
      S [,, 1: n ]
39 }

41 #Lettura dei dati:

43 dati<-read.table("C:\\Desktop\\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 ):
2 ids<-unique (dati$plot )
  m<-length ( ids )
4 Y<-list() ; X<-list() ; N<-NULL
  for ( j in 1:m)
6 {
      Y[ [ j ] ]<-dati [ dati [,1]== ids [ j ], 3]
8      N[ j]<- sum( dati$plot==ids [ j ])
      xj<-dati [ dati [,1]== ids [ j ], 2]
10     X[ [ j ] ]<-cbind ( rep (1,N[ j ]), xj, xj^2 )
  }
12 p<-dim(X[ [ 1 ] ] ) [2]
  N

```

```

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

```

```

1 Y[ [ 1 ] ]

```

```

[1] 8.25 5.81 8.69 8.03 7.96 8.89 6.13 9.40

```

```

1 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
1 #a )
  #Stimiamo i coefficienti di regressione secondo il metodo OLS
3 #indipendentemente in ciascuno de i 10 lotti:
  S2.OLS<-BETA.OLS<-NULL
5 for( j in 1:m) {
    fit<-lm(Y[ [ j ] ]~1+X[ [ j ] ] )
7    BETA.OLS<-rbind (BETA.OLS, c ( fit$coef ) )
    S2.OLS<-c ( S2.OLS, summary( fit )$sigma ^2)
9 }
11 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
```

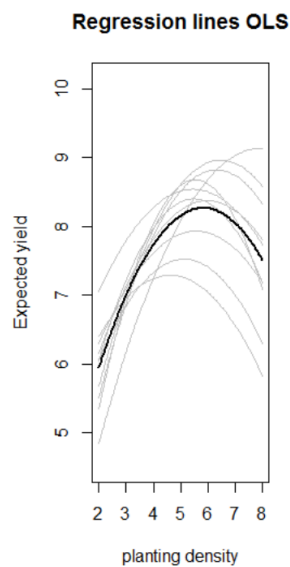


Figura 8: Curve di regressione con stime OLS.

```
[8] 0.3965025 0.1328380 0.8030505

1 #Rappresentiamo graficamente le 10 linee di regressione per valutare
  la variabilita' tra gruppi. Riportiamo sul grafico anche la loro
  media.

3 par (mfrow=c (1,2) )
  plot ( range ( dati [,2]), range ( dati [,3]), type="n", xlab="
    planting density ", +ylab="Expected yield ",main="Regression
    lines OLS")
5 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)}
7 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)
9
BETA.OLS.MEAN
```

```
      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
1 mean(S2.OLS)

[1] 0.7878099

1 #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.
3 #-)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' disegnato l' esperimento
    questa variabilita' osservata e' dovuta ad altro, forse alle
    diverse condizioni del terreno dei 10 appezzamenti.

5 #b)

7 #Setting delle prior: impostiamo le prior semiconiugate per theta,
  Sigma e sigma2 come richiesto. Ricordiamo che le prior sono
  specificate in ottica empirico-bayesiana e che quindi l' analisi
  non combina le informazioni a priori con quelle a posteriori ma
  puo' essere all' incirca interpretata come quella di un individuo
  con informazione debole ma non distorta.
  theta<-BETA.OLS.MEAN; Sigma<-cov(BETA.OLS) ; sigma2<-mean(S2.OLS)
9 eta0 <-4; S0<-Sigma
  mu0<-theta ; L0<-Sigma
11 v0<-2; sigma20<-sigma2

13 #c)

15 #Impostiamo come richiesto un Gibbs sampler: infatti e' possibile
    approssimare in questo modo la distribuzione a posteriori dal

```

```

        momento che usiamo prior semiconiugate. Per le distribuzioni full
        conditional si veda il quaderno.
#Numero di simulazioni:
17 nsimul=10000
19
21 #Valori iniziali dei parametri:
21 beta<-BETA.OLS
23 Sigma<-cov(BETA.OLS)
    sigma2<-mean(S2.OLS)
25
    #Oggetti in cui inseriamo durante l' algoritmo i valori campionati
    dalle full conditional:
27 Sigma.post<-matrix (0,p,p)
29 BETA.pp<-THETA.POST<-S2.POST<-NULL
    BETA.POST<-BETA.OLS*0
31 SIGMA.POST<-array (0, c(p,p, nsimul ) )
    set.seed (1)
33
    #Algoritmo Gibbs per nsimul iterazioni
35 for ( s in 1: nsimul ) {
    for ( j in 1:m)
37 {
        #Per ogni gruppo campioniamo i coefficienti di regressione:
39
        Vj<-solve ( solve (Sigma) + t (X[ [ j ] ] )%*%X[ [ j ] ]/
            sigma2 )
41 Ej<-Vj%*%( solve (Sigma)%*%theta + t (X[ [ j ] ] )%*%Y[ [ j
            ] ]/ sigma2 )
        beta [ j,]<-rmvnorm(1,Ej, Vj)
43 }

45 #Campioniamo la supermedia theta dei coefficienti di regressione:
    Lm<- solve ( solve (L0) + m* solve (Sigma) )
47 mum<- Lm%*%( solve (L0)%*%mu0 + solve (Sigma)%*%apply ( beta,2,
        sum) )
    theta<-t (rmvnorm(1,mum,Lm) )
49

```

```

#Campioniamo la matrice di varianza e covarianza Sigma dei
  coefficienti di regressione:
51 mtheta<-matrix ( theta,m,p, byrow=TRUE)
  Sigma<-solve ( rwish (1, eta0+m, solve ( S0+t ( beta-mtheta)%%(
    beta-mtheta) ) ) )
53 #Campioniamo la varianza residua:

55 RSS<-0
  for ( j in 1:m) { RSS<-RSS+sum( (Y[ [ j ]]-X[ [ j ]]*beta [ j,
    ] )^2 ) }
57 sigma2<-1/rgamma(1,( v0+sum(N) ) /2, (v0*sigma20+RSS)/2 )
  #Immagazziniamo i valori appena campionati:
59 S2.POST<-c(S2.POST, sigma2 ) ;THETA.POST<-rbind (THETA.POST, t (
    theta ) )
  Sigma.post<-Sigma.post+Sigma ; BETA.POST<-BETA.POST+beta
61 SIGMA.POST[, , s]<-Sigma
  #Campioniamo dalla posterior predictive dei coefficienti di
    regressione che ci servira' per il punto e).
63 BETA.pp<-rbind (BETA.pp,rmvnorm(1, theta, Sigma) )
}
65 colnames (THETA.POST)<-c(" theta1 "," theta2 "," theta3 ")
  colnames (BETA.POST)<-colnames (BETA. pp)<-c(" beta1 "," beta2 ","
    beta3 ")
67
  #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.
69
  BETA.PM<-BETA.POST/nsimul
71 plot ( range (c (0,10) ), range (c (0,10) ), type="n",
  xlab="planting density ", ylab="Yield ",main="Bayesian regression
    lines ")
73 for ( j in 1:m) { curve (BETA.PM[ j,1]+BETA.PM[ j,2]* x+BETA.PM[ j,3
    ]* x^2,
    col="gray ",add=T)}
75 curve ( mean(THETA.POST[,1])+mean(THETA.POST[,2]) *x+
  mean(THETA.POST[,3]) *x^2,lwd=2,add=T )
77 #dev.off ()

```

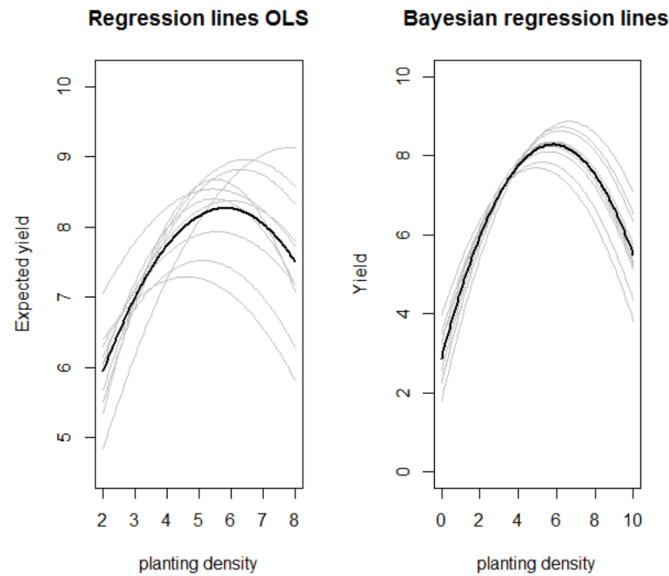


Figura 9: GLMM: stime dei minimi quadrati e stime bayesiane a confronto.

```

79 #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 e' si' sempre piu' basso in alcuni
    casi e sempre piu' alto in a l t r i rispetto alla media, ma con
    un' intensita' minore. Dal momento che lavoriamo con gruppi tutti
    di piccola numerosita' c' e' una grande variabilita' 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 in modo
    uguale per mtutti i gruppi perche' hanno la stessa numerosita'
    campionaria.

81 #Controlliamo la convergenza dell' algoritmo:

83 library (coda)
    effectiveSize (THETA.POST)

```

```

theta1      theta2    theta3
1009.3091 742.6614 656.8365
#d)
2
#Approssimazione delle distribuzioni a priori tramite simulazione
  Monte Carlo ( facciamo notare che potremmo derivarle anche
    analiticamente ):
4
n<-10000
6
THETA.PRIOR<-rmvnorm(n,mu0,L0) ; colnames (THETA.PRIOR)<-colnames (
  THETA.POST)
8 SIGMA.PRIOR<-rinv
  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

```

1 head(SIGMA.PRIOR)

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

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

3 par (mfrow=c (2,2) )
for ( i in 1:3) {
5   plot ( density (THETA.POST[, i ]), xlab=paste (" theta ", i ),
        main="")
        lines ( density (THETA.PRIOR[, i ]), col="grey ")
7   legend (" topright ", legend=c(" Prior "," Posterior "), col=c("
        grey "," black "),
        lty =1,bty="n", cex =0.7)
9 }

1 #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.

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

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

```

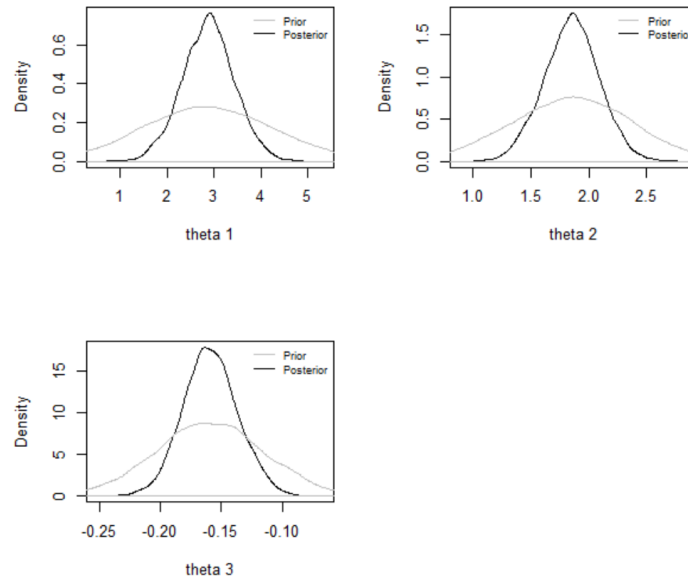


Figura 10: Densità a priori e a posteriori a confronto (media).

```

11 bty="n", lty=1)
12 }
13
14 #NB Abbiamo preso il logaritmo degli elementi di Sigma dal momento
15   che sono valori molto bassi. NB Dal momento che e' richiesto di
16   commentare la variabilita' relativa all' intercetta e alle
17   pendenze nei gruppi, riportiamo i plot delle distribuzioni dei
18   soli elementi della diagonale principale di Sigma.Commento al
19   plot: ' e' evidenza che la variabilita' tra i gruppi delle
20   intercette non e' molto elevata, ma diminuisce ancora per il
21   primo e per il secondo coefficiente ( ricordiamo che dal momento
22   che consideriamo il logaritmo delle varianze la variabilita' e'
23   minore e' indicata da una densita' concentrata su valori sempre
24   piu' negativi.
25
26 #e)
27
28 #Si vuole trovare la densita' di piante che massimizza il raccolto
29   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 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. ' 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 posterior predictive e campionare le y dati i valori della x e quindi per ogni valore della x potevamo avere un valore atteso. Facciamo la media delle distribuzioni dei valori attesi. 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. Alla fine per ogni x si genera una distribuzione del valore atteso della y futuro (4 vettori). Confrontiamo le 4 distribuzioni (per es con la media o vedendo la prob. che una sia maggiore dell' altra). Prendendo il valore medio per ogni valore di beta tilde e per l' ixmax e la varianza a posteriori campiono un valore y. Così' si incorpora anche l' incertezza derivante dai gruppi.

```

19 x0<-c (2,4,6,8)
21 raccolto<-NULL
   for ( i in x0){
23 x<-c (1, i, i ^2)
   raccolto<-cbind ( raccolto,BETA. pp%*%x)
25 }
   colnames ( raccolto )<-x0
27 head( raccolto )

```

```

           2           4           6           8
[1,] 5.989406 7.758386 8.317531 7.666840

```

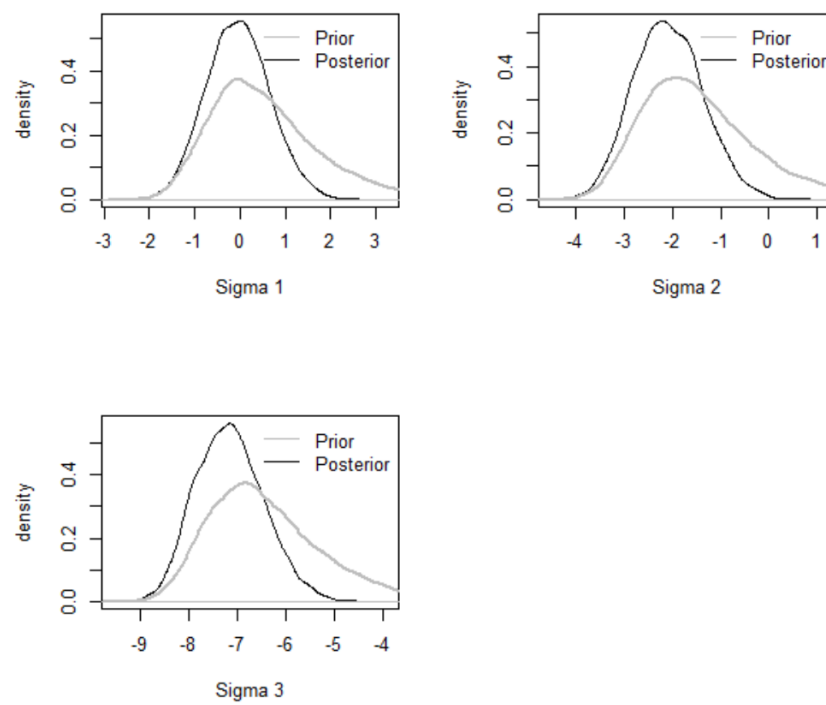



Figura 11: Densità a priori e a posteriori a confronto (variabilità e covariabilità).

```
[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
```

```
1 #Confrontiamo adesso le quattro distribuzioni dei valori attesi,
  #graficamente e con un indice sintetico, la media:
3
4 par (mfrow=c (1,1) )
5 hist ( raccolto [,1], prob=T, main="Raccolto atteso a posteriori ",
      xlab="Raccolto ",
      xlim=c (2,13), ylim=c (0,1.5) )
7 lines ( density ( raccolto [,1] )
8 hist ( raccolto [,2], prob=T, col="red ",add=T)
9 lines ( density ( raccolto [,2]), col="red ")
10 hist ( raccolto [,3], prob=T, col="green ",add=T)
11 lines ( density ( raccolto [,3]), col="green ")
12 hist ( raccolto [,4], prob=T, col="blue ",add=T)
13 lines ( density ( raccolto [,4]), col="blue ")
14 legend (" topright ", legend=c("x=2","x=4","x=6","x=8"), col=c("
    black "," red ",
15 "green "," blue "), lty=1)
16 apply ( raccolto,2,mean)
```

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

#Si osserva che il valore della covariata che masimizza il raccolto atteso per un generico lotto e' x=6. Procediamo quindi considerando tale valore per il predittore lineare. Si vuole infine approssimare 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 si aggiunge un parte di variabilita' dovuta alle osservazioni campionarie. Per ogni vettore dei coefficienti estratto dalla predittiva a posteriori e per ogni elemento estratto dalla a posteriori della varianza residua (anche questo fatto gia' fatto nell' algoritmo) campioniamo un valore del raccolto da una

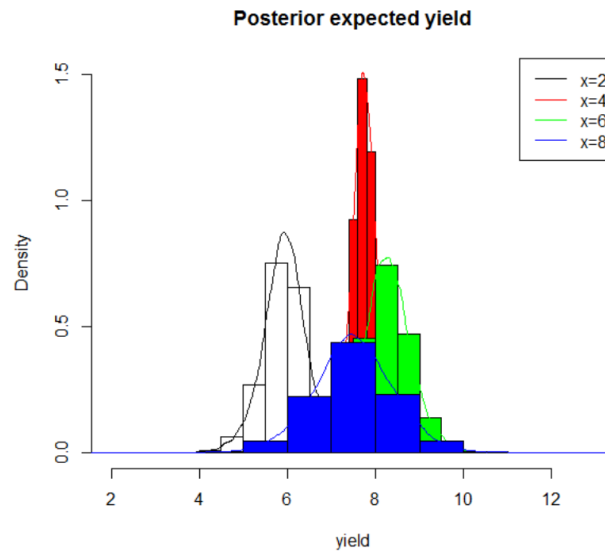


Figura 12: Valori attesi a posteriori per più valori di x a confronto.

```

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:
2
y. pred<-rnorm( nsimul,BETA. pp%*%x, S2.POST)
4 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 ")
2 lines (y. pred )

```

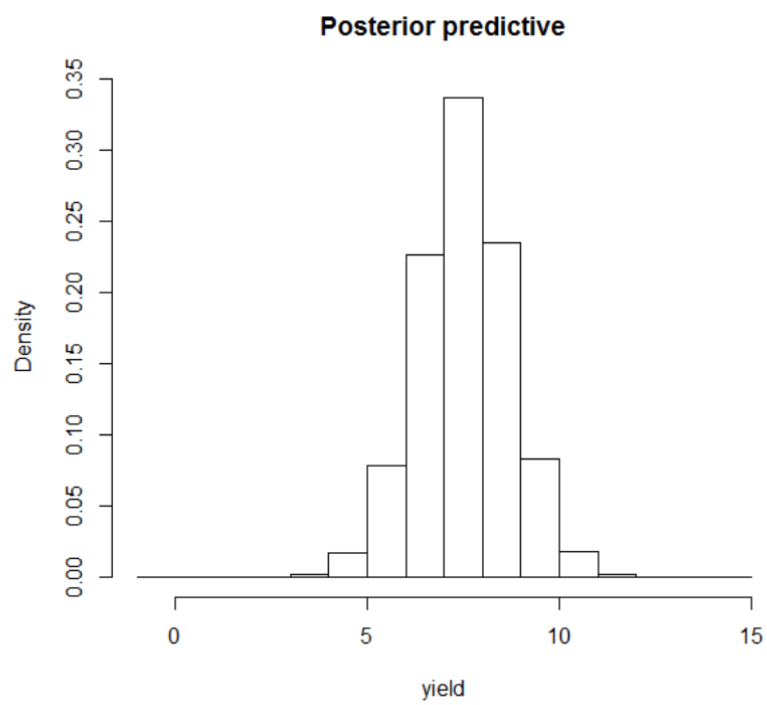


Figura 13: Predittiva a posteriori.