Modelli Bayesiani Gerarchici

Un'introduzione e qualche applicazione

Matilde Trevisani

Inferenza statistica Bayesiana

Trieste, 2022-2024

Schema della lezione I

Un esempio didattico

Idea fondamentale: scambiabilità

Stima dei modelli gerarchici Bayesiani

(Gibbs sampling e DAG)

BUGS

MCMC tricks

Convergence diagnosis

Schema della lezione II

Alcuni riferimenti bibliografici

Indice della sezione I

Indice della sezione II

Un modello lineare a 2 stadi

Da Robert (2007) pag. 145 sec. 5.6.

Un modello lineare a 2 stadi

Una meta-analisi in medicina

Un modello lineare a 2 stadi

Una meta-analisi in medicina

- Data pertain *beta-blockers* treatment effect on mortality (after heart attack) in 22 clinical trials (data from Yusuf et al., 1985; Table 5.4 in Robert (2007)).

▸ data ▸ each study

Un modello lineare a 2 stadi

Una meta-analisi in medicina

- Data pertain *beta-blockers* treatment effect on mortality (after heart attack) in 22 clinical trials (data from Yusuf et al., 1985; Table 5.4 in Robert (2007)).

  ▸ data ▸ each study

- Under the assumption that the studies are someway comparable, it is possible to combine the findings of single studies to infer on meta-population characteristics.

Un modello lineare a 2 stadi

Una meta-analisi in medicina

- Data pertain *beta-blockers* treatment effect on mortality (after heart attack) in 22 clinical trials (data from Yusuf et al., 1985; Table 5.4 in Robert (2007)).

  ▸ data ▸ each study

- Under the assumption that the studies are someway comparable, it is possible to combine the findings of single studies to infer on meta-population characteristics.
- Hypothesis of exchangeability

Un modello lineare a 2 stadi

Una meta-analisi in medicina

- Data pertain *beta-blockers* treatment effect on mortality (after heart attack) in 22 clinical trials (data from Yusuf et al., 1985; Table 5.4 in Robert (2007)).

  ▸ data ▸ each study

- Under the assumption that the studies are someway comparable, it is possible to combine the findings of single studies to infer on meta-population characteristics.
- Hypothesis of exchangeability
- A HBM for meta-analysis

Multiple studies

Each study consists of 2 groups of patients randomly allocated to receive or not receive beta-blockers treatment

study	control: dead/ total	treated: dead/ total
1	3/ 39	3/ 38
2	14/ 116	7/ 114
3	11/ 93	5/ 69
4	127/1520	102/1533
5	27/ 365	28/ 355
6	6/ 52	4/ 59
7	152/ 939	98/ 945
8	48/ 471	60/ 632
9	37/ 282	25/ 278
10	188/1921	138/1916
⋮	⋮/ ⋮	⋮/ ⋮
16	38/ 213	33/ 207
17	12/ 122	28/ 251
18	6/ 154	8/ 151
19	3/ 134	6/ 174
20	40/ 218	32/ 209
21	43/ 364	27/ 391
22	39/ 674	22/ 680

◂ es. beta-blockers ▸ graph 1 ▸ graph 2 ▸ graph 3 ▸ μ, τ posterior ▸ θ posterior

Da ogni studio: dati grezzi e parametro θ_j

I dati grezzi sono realizzazioni di 2 modelli binomiali indipendenti

$$y_{1j} \sim \text{Bin}(n_{1j}, p_{1j}) \qquad y_{0j} \sim \text{Bin}(n_{0j}, p_{0j})$$

dove

- $j = 1, \ldots, 22$ indicizza ciascun singolo studio

Da ogni studio: dati grezzi e parametro θ_j

I dati grezzi sono realizzazioni di 2 modelli binomiali indipendenti

$$y_{1j} \sim \text{Bin}(n_{1j}, p_{1j}) \qquad y_{0j} \sim \text{Bin}(n_{0j}, p_{0j})$$

dove

- $j = 1, \ldots, 22$ indicizza ciascun singolo studio
- $(y_{1j},\, n_{1j},\, p_{1j})$, $(y_{0j},\, n_{0j},\, p_{0j})$ sono morti, pazienti e mortalità nei gruppi di trattamento e controllo, rispettivamente, nello studio j.

◂ es. beta-blockers

Da ogni studio: dati grezzi e parametro θ_j

I dati grezzi sono realizzazioni di 2 modelli binomiali indipendenti

$$y_{1j} \sim \text{Bin}(n_{1j}, p_{1j}) \qquad y_{0j} \sim \text{Bin}(n_{0j}, p_{0j})$$

dove

- $j = 1, \ldots, 22$ indicizza ciascun singolo studio
- $(y_{1j},\, n_{1j},\, p_{1j})$, $(y_{0j},\, n_{0j},\, p_{0j})$ sono morti, pazienti e mortalità nei gruppi di trattamento e controllo, rispettivamente, nello studio j.

◂ es. beta-blockers

Funzioni di stima di interesse comune sono la differenza nelle probabilità, $p_{1j} - p_{0j}$, il rapporto di probabilità o tasso di rischio, p_{1j}/p_{0j}, e l'*odds ratio*, $\rho_j = (p_{1j}/(1 - p_{1j}))/(p_{0j}/(1 - p_{0j}))$. Per varie ragioni, e per il fatto che la sua distribuzione a posteriori sia prossima alla normalità anche per grandezzze campionarie relativamente piccole, ci concentriamo sull'inferenza per il logaritmo (naturale) dell'odds ratio, che denotiamo come $\theta_j = \log \rho_j$.

Da ogni studio: dati grezzi e parametro θ_j

I dati grezzi sono realizzazioni di 2 modelli binomiali indipendenti

$$y_{1j} \sim \text{Bin}(n_{1j}, p_{1j}) \qquad y_{0j} \sim \text{Bin}(n_{0j}, p_{0j})$$

dove

- $j = 1, \ldots, 22$ indicizza ciascun singolo studio
- $(y_{1j},\, n_{1j},\, p_{1j})$, $(y_{0j},\, n_{0j},\, p_{0j})$ sono morti, pazienti e mortalità nei gruppi di trattamento e controllo, rispettivamente, nello studio j.

◂ es. beta-blockers

Per stimare l'effetto del trattamento sulla mortalità si consideri il log-*odds ratio* (lor)

$$\theta_j = log(p_{1j}/(1-p_{1j}))/(p_{0j}/(1-p_{0j}))$$

come parametro θ_j di interesse per ogni studio.

Da ogni studio: dati grezzi e parametro θ_j

I dati grezzi sono realizzazioni di 2 modelli binomiali indipendenti

$$y_{1j} \sim \text{Bin}(n_{1j}, p_{1j}) \qquad y_{0j} \sim \text{Bin}(n_{0j}, p_{0j})$$

dove

- $j = 1, \ldots, 22$ indicizza ciascun singolo studio
- $(y_{1j},\, n_{1j},\, p_{1j})$, $(y_{0j},\, n_{0j},\, p_{0j})$ sono morti, pazienti e mortalità nei gruppi di trattamento e controllo, rispettivamente, nello studio j.

◂ es. beta-blockers

Per stimare l'effetto del trattamento sulla mortalità si consideri il log-*odds ratio* (lor)

$$\theta_j = log(p_{1j}/(1-p_{1j}))/(p_{0j}/(1-p_{0j}))$$

come parametro θ_j di interesse per ogni studio.
Si sa infatti come la distribuzione a posteriori di lor sia prossima alla normalità anche per numerosità campionarie relativamente piccole.

Un'approssimazione normale della verosimiglianza

Ci sono diversi approcci analitici std che producono una statistica dei dati per ogni studio che possa essere considerato come approssimazione di una variabile normale.

Un'approssimazione normale della verosimiglianza

Ci sono diversi approcci analitici std che producono una statistica dei dati per ogni studio che possa essere considerato come approssimazione di una variabile normale.
E.g. si consideri i *empirical logits*

$$y_j = log\left(\frac{y_{1j}}{n_{1j} - y_{1j}}\right) - log\left(\frac{y_{0j}}{n_{0j} - y_{0j}}\right)$$

Un'approssimazione normale della verosimiglianza

Ci sono diversi approcci analitici std che producono una statistica dei dati per ogni studio che possa essere considerato come approssimazione di una variabile normale.
E.g. si consideri i *empirical logits*

$$y_j = log\left(\frac{y_{1j}}{n_{1j} - y_{1j}}\right) - log\left(\frac{y_{0j}}{n_{0j} - y_{0j}}\right)$$

con varianza campionaria approssimata

$$\sigma_j^2 = 1/y_{1j} + 1/(n_{1j} - y_{1j}) + 1/y_{0j} + 1/(n_{0j} - y_{0j})$$

Un'approssimazione normale della verosimiglianza

Ci sono diversi approcci analitici std che producono una statistica dei dati per ogni studio che possa essere considerato come approssimazione di una variabile normale.
E.g. si consideri i *empirical logits*

$$y_j = log\left(\frac{y_{1j}}{n_{1j} - y_{1j}}\right) - log\left(\frac{y_{0j}}{n_{0j} - y_{0j}}\right)$$

con varianza campionaria approssimata

$$\sigma_j^2 = 1/y_{1j} + 1/(n_{1j} - y_{1j}) + 1/y_{0j} + 1/(n_{0j} - y_{0j})$$

Per campioni grandi, tramite il *delta method* (Bishop, Fienberg, and Holland, 1975, Section 14.6), lo stimatore Y_j di θ_j è approssimativamente normale con standard error asintotico che può essere stimato con la formula sopra.

Un'approssimazione normale della verosimiglianza

Ci sono diversi approcci analitici std che producono una statistica dei dati per ogni studio che possa essere considerato come approssimazione di una variabile normale.
E.g. si consideri i *empirical logits*

$$y_j = log\left(\frac{y_{1j}}{n_{1j} - y_{1j}}\right) - log\left(\frac{y_{0j}}{n_{0j} - y_{0j}}\right)$$

con varianza campionaria approssimata

$$\sigma_j^2 = 1/y_{1j} + 1/(n_{1j} - y_{1j}) + 1/y_{0j} + 1/(n_{0j} - y_{0j})$$

Siamo quindi pronti a stimare un modello normale a due stadi.

Nella sezione

Un esempio didattico

Il modello Normale-Normale

Idea fondamentale: scambiabilità

Il modello Normale-Normale

Per $j = 1, \ldots, J$

$$y_j|\theta_j \overset{ind}{\sim} N(\theta_j, \sigma_j^2) \quad (\sigma^2 \text{ nota}) \tag{1}$$

1. il modello per le unità individuali

Il modello Normale-Normale

Per $j = 1, \ldots, J$

$$y_j|\theta_j \overset{ind}{\sim} N(\theta_j, \sigma_j^2) \quad (\sigma^2 \text{ nota}) \tag{1}$$

$$\theta_j|\mu,\, \tau^2 \overset{iid}{\sim} N(\mu, \tau^2) \tag{2}$$

1. il modello per le unità individuali
2. il modello per la meta-popolazione (ipotesi di scambiabilità)

Il modello Normale-Normale

Per $j = 1, \ldots, J$

$$y_j|\theta_j \overset{ind}{\sim} N(\theta_j, \sigma_j^2) \quad (\sigma^2 \text{ nota}) \tag{1}$$

$$\theta_j|\mu, \tau^2 \overset{iid}{\sim} N(\mu, \tau^2) \tag{2}$$

$$(\mu, \tau^2) \sim p(\mu, \tau^2) \tag{3}$$

1. il modello per le unità individuali
2. il modello per la meta-popolazione (ipotesi di scambiabilità)
3. il modello a priori per gli iperparametri (*full HB*)

Il modello Normale-Normale

Per $j = 1, \ldots, J$

$$y_j|\theta_j \overset{ind}{\sim} N(\theta_j, \sigma_j^2) \quad (\sigma^2 \text{ nota}) \tag{1}$$

$$\theta_j|\mu,\, \tau^2 \overset{iid}{\sim} N(\mu, \tau^2) \tag{2}$$

$$(\mu,\, \tau^2) \sim p(\mu, \tau^2) \tag{3}$$

1. il modello per le unità individuali
2. il modello per la meta-popolazione (ipotesi di scambiabilità)
3. il modello a priori per gli iperparametri (*full HB*)

Questo modello viene anche detto *one-way normal random-effects model* con varianza del modello di campionamento dei dati nota, ed è un caso speciale del più generale modello lineare normale gerarchico.

Note: Nel livello

1. La semplificazione della varianza σ^2 nota è irrilevante in questo esempio dato che le varianze in ogni studio sono stimate con precisione per la numerosità elevata (più di 50 individui) in quasi la totalità degli studi.
2. Assumiamo un modello normale per gli effetti θ_j per ragioni di comodo, ma è importante fare un check di tale ipotesi.

1.&2. Se altre informazioni (oltre y, n) fossero disponibili per distinguere tra i J studi della meta-analisi, è possibile espandere il modello in modo che sia scambiabile nei dati osservati e nelle covariate (eg usando un modello di regressione gerarchico) in modo da stimare una *response surface* sulla base di caratteristiche note della popolazione e della sua esposizione al rischio.

3. Assumiamo una a priori noninformativa o localmente uniforme per μ dato che anche con un numero piccolo di studi $(5\text{-}10)$ i dati combinati sono relativam informativi riguardo al centro della distribuzione della popolazione delle dimensioni degli effetti θ_j. Allo stesso modo assumiamo per convenienza una a priori localmente uniforme per τ, anche se si può modificare facilmente l'analisi per includere un'informazione a priori.

H(B)M media tra due estremi

- $\tau^2 \to \infty$: analisi separate

  N-N si riduce a

$$y_j|\theta_j \overset{ind}{\sim} N(\theta_j, \sigma_j^2), \quad \theta_j \sim p_j(\theta_j)$$

  $\to$ i parametri di interesse sono gli effetti θ_j

H(B)M media tra due estremi

- $\tau^2 \to \infty$: analisi separate

  N-N si riduce a

$$y_j|\theta_j \overset{ind}{\sim} N(\theta_j, \sigma_j^2), \quad \theta_j \sim p_j(\theta_j)$$

  $\to$ i parametri di interesse sono gli effetti θ_j

- $\tau^2 = 0$: pooling completo

  N-N si riduce a

$$y_j|\mu, \sigma_j^2 \overset{iid}{\sim} N(\mu, \sigma_j^2), \quad \mu \sim p(\mu)$$

  $\to$ parametro di interesse è μ

H(B)M media tra due estremi

- $\tau^2 \to \infty$: analisi separate

  N-N si riduce a

$$y_j|\theta_j \overset{ind}{\sim} N(\theta_j, \sigma_j^2), \quad \theta_j \sim p_j(\theta_j)$$

  $\to$ i parametri di interesse sono gli effetti θ_j

- $\tau^2 = 0$: pooling completo

  N-N si riduce a

$$y_j|\mu, \sigma_j^2 \overset{iid}{\sim} N(\mu, \sigma_j^2), \quad \mu \sim p(\mu)$$

  $\to$ parametro di interesse è μ

- $0 < \tau^2 < \infty$: scambiabilità

L'analisi Bayesiana in un modello gerarchico fornisce un compromesso che combina le informazioni da tutti gli esperimenti senza assumere che tutti i θ_j siano uguali.

La struttura gerarchica del modello è strumento essenziale per ottenere un *pooling* parziale delle stime e fare un compromesso in modo scientifico tra fonti alternative di informazione.

In particolare, consideriamo gli studi scambiabili, ie non necessariamente identici o completamente scollegati, o, in altri termini, ammettiamo che ci siano differenze da studio a studio ma in modo che non ci aspettiamo a priori che tali differenze abbiano effetti prevedibili che influenzino uno studio piuttosto che un altro.

Stime a posteriori

- analisi separate

  Nell'ipotesi $p_j(\theta_j)$ uniforme su $(-\infty, \infty)$

Stime a posteriori

- analisi separate

  Nell'ipotesi $p_j(\theta_j)$ uniforme su $(-\infty, \infty)$

  $\hat{\theta}_j = y_j$, intervallo a posteriori (CI) al 95% è $\hat{\theta}_j \pm 1.96 * \sigma_j$

  ▸ separate analyses graph ▸ difficulties

Stime a posteriori

- analisi separate

Nell'ipotesi $p_j(\theta_j)$ uniforme su $(-\infty, \infty)$

$\hat{\theta}_j = y_j$, intervallo a posteriori (CI) al 95% è $\hat{\theta}_j \pm 1.96 * \sigma_j$

▸ separate analyses graph ▸ difficulties

Si Consideri un'osservazione scalare singola y da una distribuzione normale parametrizzata da media θ e varianza σ^2, ove assumiamo che σ^2 sia noto. La distribuzione campionaria è

$$p(y|\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\theta)^2}.$$

Se $p(\theta) = 1$, allora $p(\theta|y) \propto p(y|\theta) \cdot 1$, pertanto

$$p(\theta|y) = N(y, \sigma^2).$$

Si noti che nonostante la distribuzione a priori sia impropria (una densità a priori è propria se non dipende dai dati e integra a 1), la distribuzione a posteriori è propria ($\int p(\theta|y)d\theta$ è finita per tutti gli y), purchè ci sia almeno un dato puntuale.

Stime a posteriori

- pooling completo

  Nell'ipotesi $p(\mu)$ uniforme su $(-\infty, \infty)$

Stime a posteriori

- pooling completo

  Nell'ipotesi $p(\mu)$ uniforme su $(-\infty, \infty)$

  $\hat{\mu} = \frac{\sum_j \frac{1}{\sigma_j^2} y_j}{\sum_j \frac{1}{\sigma_j^2}}$, 95%CI$= \hat{\mu} \pm 1.96 * V_\mu^{1/2}$, dove $V_\mu^{-1} = \sum_j \frac{1}{\sigma_j^2}$

Stime a posteriori

- pooling completo

Nell'ipotesi $p(\mu)$ uniforme su $(-\infty, \infty)$

$\hat{\mu} = \frac{\sum_j \frac{1}{\sigma_j^2} y_j}{\sum_j \frac{1}{\sigma_j^2}}$, 95%CI$= \hat{\mu} \pm 1.96 * V_\mu^{1/2}$, dove $V_\mu^{-1} = \sum_j \frac{1}{\sigma_j^2}$

Si consideri un campione di osservazioni iid $y = (y_1, \ldots, y_m)$ da una distribuzione normale parametrizzata da media μ e varianze individuali σ_j^2 (assunte note). Se $p(\mu) = 1$, allora

$$\begin{aligned} p(\mu|y) &\propto \prod_j N(\mu, \sigma_j^2) \cdot 1 \\ &\propto \exp\left(-\tfrac{1}{2} \sum_j \tfrac{1}{\sigma_j^2} (y_j - \mu)^2\right) \end{aligned}$$

che dopo opportuni calcoli porta a

$$p(\mu|y) = N\left(\frac{\sum_j \frac{1}{\sigma_j^2} y_j}{\sum_j \frac{1}{\sigma_j^2}}, \left(\sum_j \frac{1}{\sigma_j^2}\right)^{-1}\right) .$$

Stime a posteriori

- pooling completo

  Nell'ipotesi $p(\mu)$ uniforme su $(-\infty, \infty)$

  $\hat{\mu} = \frac{\sum_j \frac{1}{\sigma_j^2} y_j}{\sum_j \frac{1}{\sigma_j^2}}$, 95%CI$= \hat{\mu} \pm 1.96 * V_\mu^{1/2}$, dove $V_\mu^{-1} = \sum_j \frac{1}{\sigma_j^2}$

  Inoltre, 95%CI$= \hat{\mu} \pm 1.96 * \sqrt{\bar{\sigma_j^2}}$ è sovrapposto, per comparare i dati osservati con la variabilità attesa se l'ip. $\tau^2 = 0$ dovesse essere quella "vera". ▸ complete pooling graph ▸ difficulties

Stime a posteriori

- scambiabilità

Invece che essere forzati a scegliere tra

1. *complete pooling*, $\hat{\theta}_j = \hat{\mu}\ \forall j$, o
2. nessun *pooling*, $\hat{\theta}_j = y_j$,
3. possiamo pensare ad una media pesata

$$\hat{\theta}_j = \lambda_j \cdot \hat{\mu} + (1 - \lambda_j) \cdot y_j$$

con λ_j tra 0 e 1

Stime a posteriori

- scambiabilità

Si consideri un'osservazione scalare singola y da una distribuzione normale parametrizzata da media θ e varianza σ^2, ove assumiamo che σ^2 sia noto. La distribuzione campionaria è

$$p(y|\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\theta)^2} .$$

se $p(\theta) = N(\mu, \tau^2)$, allora $p(\theta|y) \propto p(y|\theta)p(\theta)$, pertanto

$$p(\theta|y) = N\left(\frac{\frac{1}{\sigma_j^2}y_j + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}, \frac{1}{\sigma_j^2} + \frac{1}{\tau^2}\right) .$$

Stime a posteriori

- scambiabilità

Si consideri un'osservazione scalare singola y da una distribuzione normale parametrizzata da media θ e varianza σ^2, ove assumiamo che σ^2 sia noto. La distribuzione campionaria è

$$p(y|\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\theta)^2} .$$

se $p(\theta) = N(\mu, \tau^2)$, allora $p(\theta|y) \propto p(y|\theta)p(\theta)$, pertanto

$$p(\theta|y) = N\left(\frac{\frac{1}{\sigma_j^2}y_j + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}, \frac{1}{\sigma_j^2} + \frac{1}{\tau^2}\right) .$$

La precisione a posteriori eguaglia la precisione a priori più la precisione dei dati.

Stime a posteriori

- scambiabilità

Si consideri un'osservazione scalare singola y da una distribuzione normale parametrizzata da media θ e varianza σ^2, ove assumiamo che σ^2 sia noto. La distribuzione campionaria è

$$p(y|\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\theta)^2} .$$

se $p(\theta) = N(\mu, \tau^2)$, allora $p(\theta|y) \propto p(y|\theta)p(\theta)$, pertanto

$$p(\theta|y) = N\left(\frac{\frac{1}{\sigma_j^2}y_j + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}, \frac{1}{\sigma_j^2} + \frac{1}{\tau^2}\right) .$$

La media a posteriori viene espressa come media pesata della media a priori e del valore osservato con pesi proporzionali alle precisioni.

Stime a posteriori

- scambiabilità

Si consideri un'osservazione scalare singola y da una distribuzione normale parametrizzata da media θ e varianza σ^2, ove assumiamo che σ^2 sia noto. La distribuzione campionaria è

$$p(y|\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\theta)^2} .$$

se $p(\theta) = N(\mu, \tau^2)$, allora $p(\theta|y) \propto p(y|\theta)p(\theta)$, pertanto

$$p(\theta|y) = N\left(\frac{\frac{1}{\sigma_j^2}y_j + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}, \frac{1}{\sigma_j^2} + \frac{1}{\tau^2}\right) .$$

Alternativamente, possiamo esprimere la media a posteriori come la media a priori aggiustata verso il valore osservato, o come il dato osservato "shrunk" verso la media a priori.

Stime a posteriori

- scambiabilità

Si consideri un'osservazione scalare singola y da una distribuzione normale parametrizzata da media θ e varianza σ^2, ove assumiamo che σ^2 sia noto. La distribuzione campionaria è

$$p(y|\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\theta)^2} .$$

se $p(\theta) = N(\mu, \tau^2)$, allora $p(\theta|y) \propto p(y|\theta)p(\theta)$, pertanto

$$p(\theta|y) = N\left(\frac{\frac{1}{\sigma_j^2}y_j + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}, \frac{1}{\sigma_j^2} + \frac{1}{\tau^2}\right) .$$

Ognuna delle formulazioni rappresenta la media a posteriori come un compromesso tra la media a priori e il valore osservato.

Stime a posteriori

- scambiabilità

Si consideri un'osservazione scalare singola y da una distribuzione normale parametrizzata da media θ e varianza σ^2, ove assumiamo che σ^2 sia noto. La distribuzione campionaria è

$$p(y|\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\theta)^2} .$$

se $p(\theta) = N(\mu, \tau^2)$, allora $p(\theta|y) \propto p(y|\theta)p(\theta)$, pertanto

$$p(\theta|y) = N\left(\frac{\frac{1}{\sigma_j^2}y_j + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}, \frac{1}{\sigma_j^2} + \frac{1}{\tau^2}\right) .$$

Il calcolo della a posteriori per il modello gerarchico (completo di iper*a*priori) verrà mostrato più avanti.

Stime a posteriori

- scambiabilità

  $2.5\%, 50\%, 97.5\%$ quantili a posteriori di θ_j e di μ.

Stime a posteriori

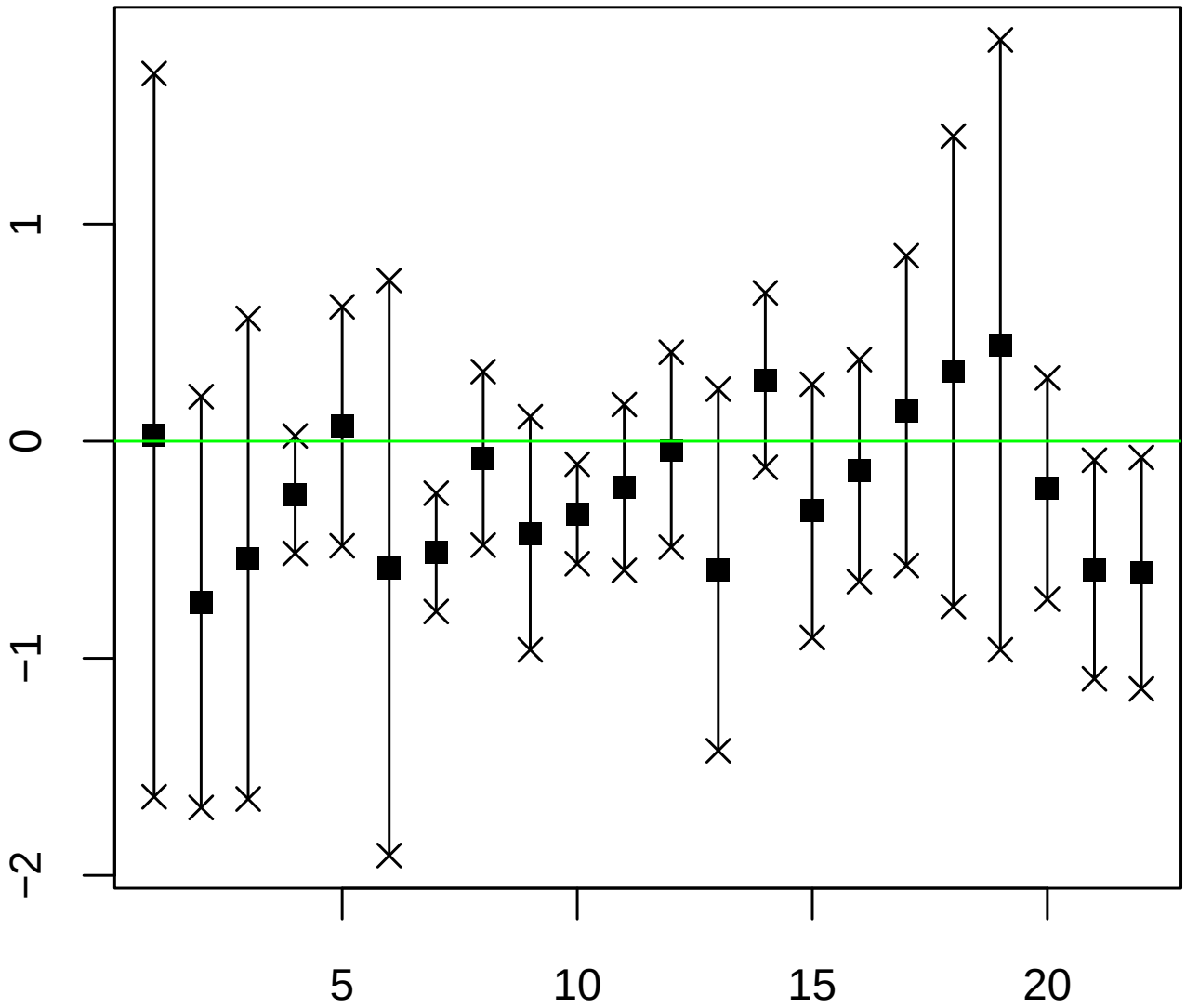
- scambiabilità

  $2.5\%, 50\%, 97.5\%$ quantili a posteriori di θ_j e di μ.

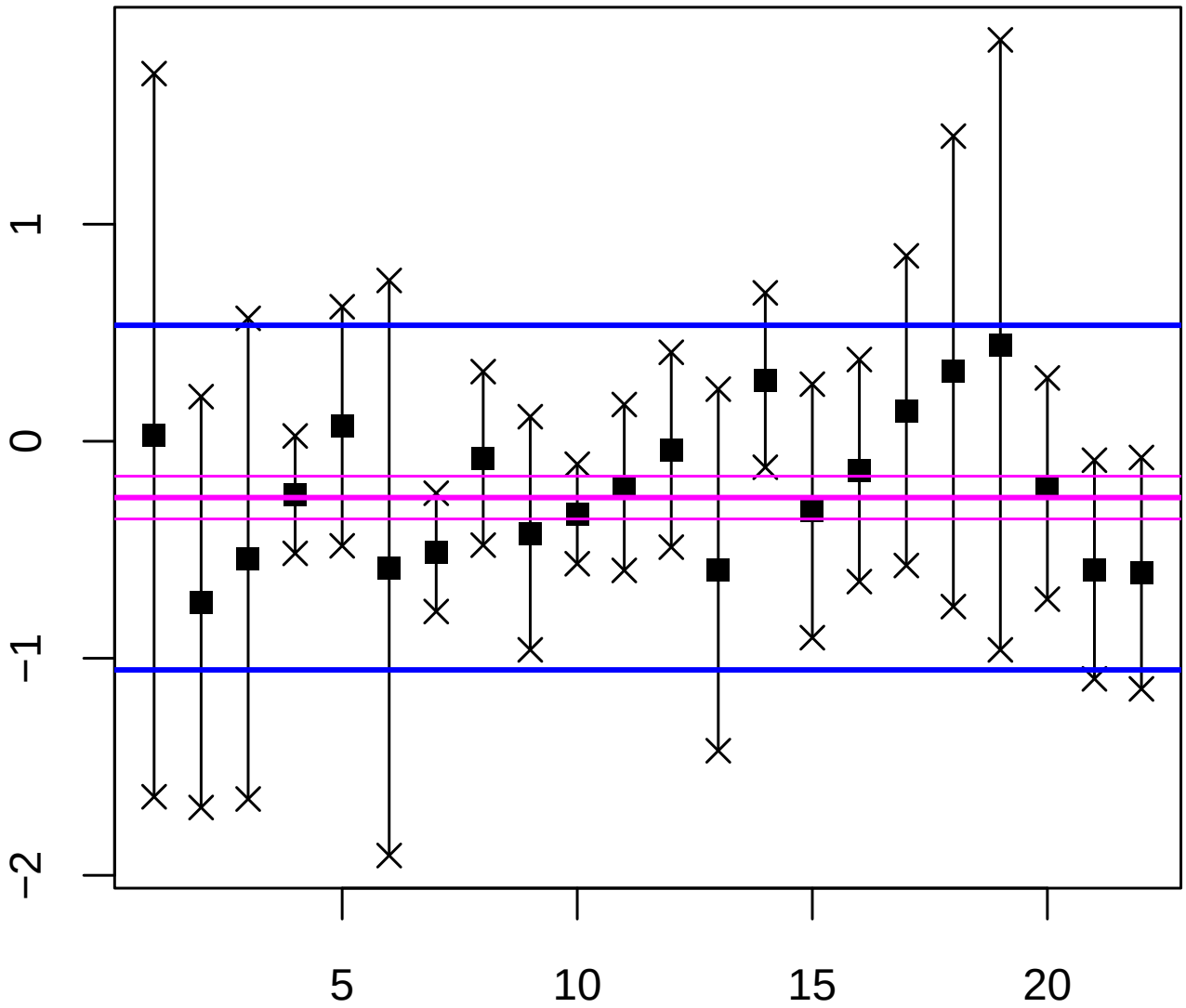
  Inoltre, $\hat{\mu}$ e 95%CI del pooling completo sono sovrapposti.

  ▸ HBM graph

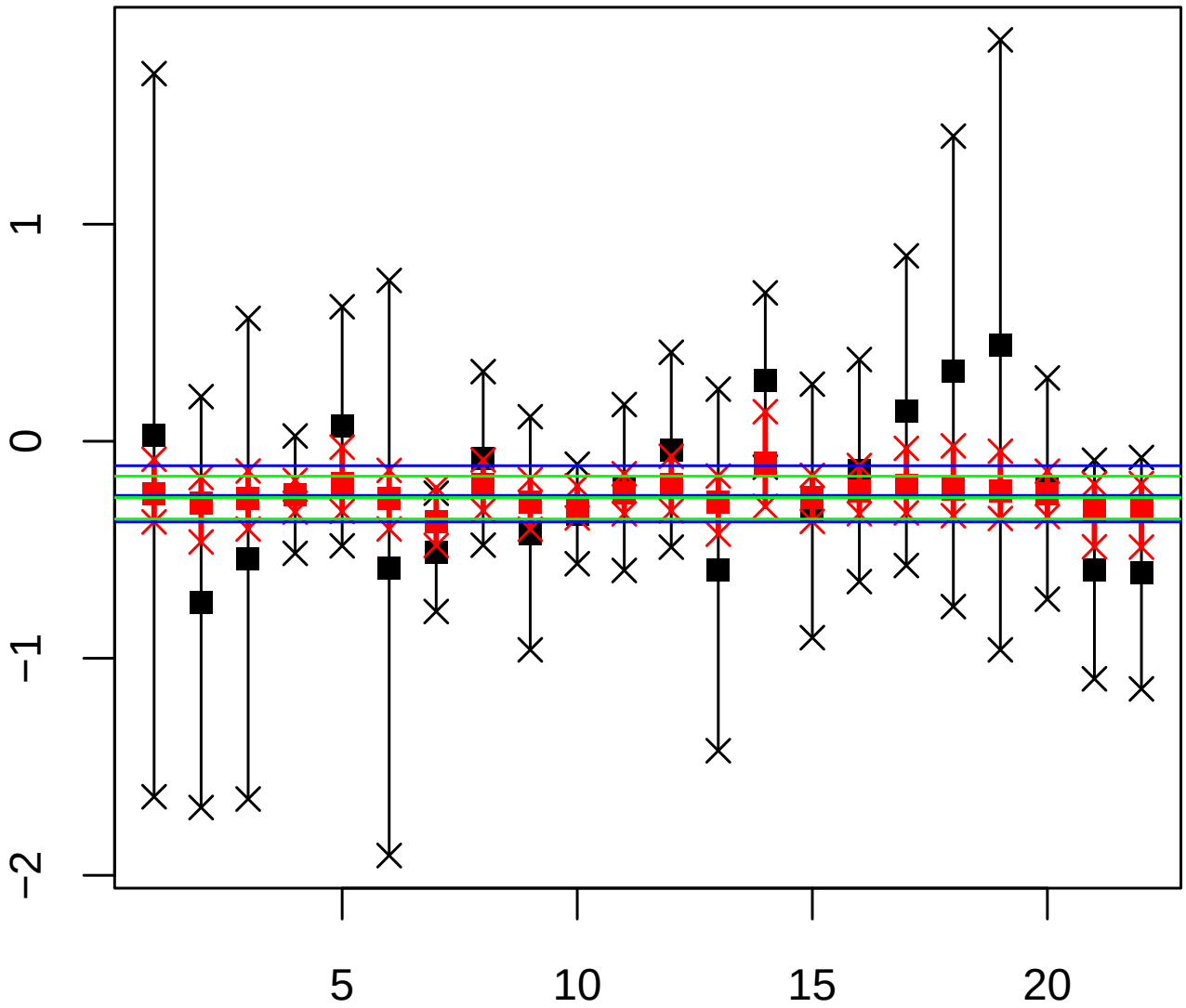
Separate analyses



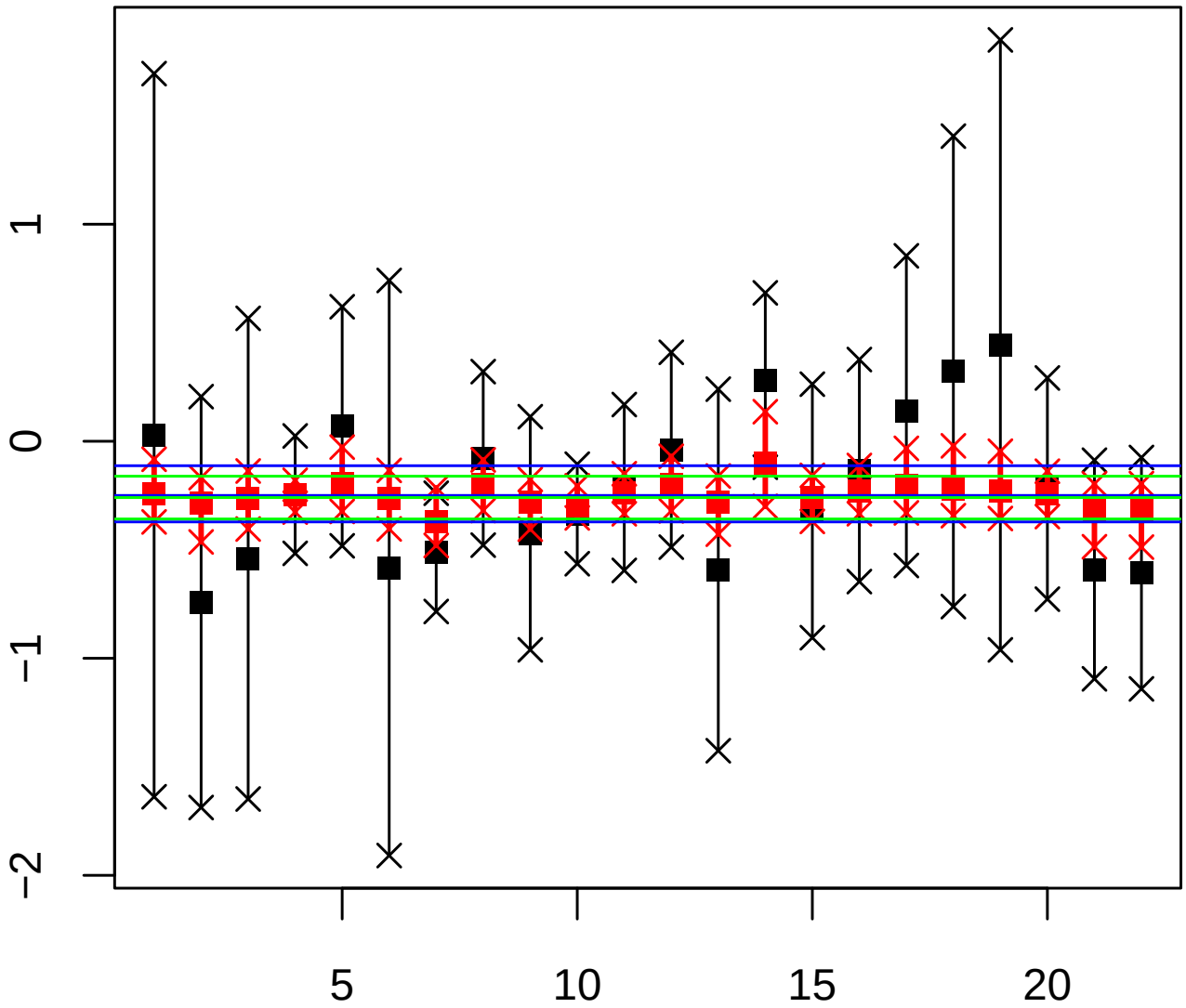
or complete pooling?



Exchangeable studies: pooling strength



Exchangeable studies: pooling strength



Difficoltà con le analisi 'estreme'

◂ separate analyses ◂ graph

analisi separate eg, per lo studio 19: $\theta_{19}|y_{19} \sim N(0.44, 0.72^2)$

the probability is 1/2 that the true effect in study-19 is more than 0.44

a doubtful statement, considering the results for the other 21 studies.

Difficoltà con le analisi 'estreme'

◂ separate analyses ◂ graph ◂ complete pooling ◂ graph

analisi separate eg, per lo studio 19: $\theta_{19}|y_{19} \sim N(0.44, 0.72^2)$

the probability is 1/2 that the true effect in study-19 is more than 0.44

a doubtful statement, considering the results for the other 21 studies.

analisi pooled $\mu|y \sim N(-0.26, 0.05^2)$

the probability is 1/2 that θ_{19}*, the true effect in study-19 (*$y_{19} = 0.44$ *con sd* 0.72*), is less than* -0.26 *or, even less than* θ_2*, the true effect in study-2 (*$y_2 = -0.74$ *con sd* 0.48*)*

given that $P(\theta_{19} - \theta_2 < 0|y) = 1/2$

Distribuzioni a posteriori

1. Densità (marginali) a posteriori di τ e μ ▸ graph

Distribuzioni a posteriori

1. Densità (marginali) a posteriori di τ e μ ▸ graph

- La densità marginale a posteriori di τ ha un picco su un valore non-zero, sebbene valori vicini a zero sono chiaramente plausibili, con zero che ha una densità a posteriori solo circa 25% più bassa che sulla moda.

Distribuzioni a posteriori

1. Densità (marginali) a posteriori di τ e μ ▸ graph

- La densità marginale a posteriori di τ ha un picco su un valore non-zero, sebbene valori vicini a zero sono chiaramente plausibili, con zero che ha una densità a posteriori solo circa 25% più bassa che sulla moda.
- Poichè τ si concentra su valori piccoli relativamente a σ_j,

Distribuzioni a posteriori

1. Densità (marginali) a posteriori di τ e μ ▸ graph

- La densità marginale a posteriori di τ ha un picco su un valore non-zero, sebbene valori vicini a zero sono chiaramente plausibili, con zero che ha una densità a posteriori solo circa 25% più bassa che sulla moda.
- Poichè τ si concentra su valori piccoli relativamente a σ_j,

quantili a posteriori per τ sono

2.5%	25%	*Mediana*	75%	97.5%
0.02	0.08	0.13	0.18	0.31

mentre i valori per σ_j's sono considerabilmente più grandi (due volte in media).

Distribuzioni a posteriori

1. Densità (marginali) a posteriori di τ e μ ▸ graph

- La densità marginale a posteriori di τ ha un picco su un valore non-zero, sebbene valori vicini a zero sono chiaramente plausibili, con zero che ha una densità a posteriori solo circa 25% più bassa che sulla moda.
- Poichè τ si concentra su valori piccoli relativamente a σ_j, uno shrinkage marcato di θ_j risulta evidente specialmente per gli studi meno precisi.

Distribuzioni a posteriori

1. Densità (marginali) a posteriori di τ e μ ▸ graph

- La densità marginale a posteriori di τ ha un picco su un valore non-zero, sebbene valori vicini a zero sono chiaramente plausibili, con zero che ha una densità a posteriori solo circa 25% più bassa che sulla moda.
- Poichè τ si concentra su valori piccoli relativamente a σ_j, uno shrinkage marcato di θ_j risulta evidente specialmente per gli studi meno precisi.

Si esplori l'effetto su e.g. gli studi 1, 3, 6, 18, 19.

◂ data ◂ HBM graph

Distribuzioni a posteriori

1. Densità (marginali) a posteriori di τ e μ ▸ graph

- La densità marginale a posteriori di τ ha un picco su un valore non-zero, sebbene valori vicini a zero sono chiaramente plausibili, con zero che ha una densità a posteriori solo circa 25% più bassa che sulla moda.
- Poichè τ si concentra su valori piccoli relativamente a σ_j, uno shrinkage marcato di θ_j risulta evidente specialmente per gli studi meno precisi.
- Inoltre, questo fatto (i.e. una sostanziale omogeneità degli studi) produce anche un sostanziale decremento della variabilità a posteriori di θ_j che riflette il fatto che ogni singolo studio borrows strength da tutti gli altri.

Distribuzioni a posteriori

1. Densità (marginali) a posteriori di τ e μ graph

- La densità marginale a posteriori di τ ha un picco su un valore non-zero, sebbene valori vicini a zero sono chiaramente plausibili, con zero che ha una densità a posteriori solo circa 25% più bassa che sulla moda.
- Poichè τ si concentra su valori piccoli relativamente a σ_j, uno shrinkage marcato di θ_j risulta evidente specialmente per gli studi meno precisi.
- Inoltre, questo fatto (i.e. una sostanziale omogeneità degli studi) produce anche un sostanziale decremento della variabilità a posteriori di θ_j che riflette il fatto che ogni singolo studio borrows strength da tutti gli altri (si veda la notevole riduzione della lunghezza dell'intervallo di credibilità in HBM graph).

Distribuzioni a posteriori

1. Densità (marginali) a posteriori di τ e μ ▸ graph
2. Densità (istogrammi) a posteriori degli effetti θ_j ▸ graph

Distribuzioni a posteriori

1. Densità (marginali) a posteriori di τ e μ ▸ graph
2. Densità (istogrammi) a posteriori degli effetti θ_j ▸ graph

- Gli istogrammi delle densità a posteriori simulate per ognuno degli effetti individuali mostrano un'asimmetria lontano dal valore centrale della media globale, mentre la distribuzione della media globale ha maggiore simmetria.

Distribuzioni a posteriori

1. Densità (marginali) a posteriori di τ e μ ▸ graph
2. Densità (istogrammi) a posteriori degli effetti θ_j ▸ graph

- Gli istogrammi delle densità a posteriori simulate per ognuno degli effetti individuali mostrano un'asimmetria lontano dal valore centrale della media globale, mentre la distribuzione della media globale ha maggiore simmetria.

quantili a posteriori per μ sono

2.5%	25%	*Mediana*	75%	97.5%
-0.37	-0.29	-0.25	-0.20	-0.11

Distribuzioni a posteriori

1. Densità (marginali) a posteriori di τ e μ ▸ graph
2. Densità (istogrammi) a posteriori degli effetti θ_j ▸ graph

- Gli istogrammi delle densità a posteriori simulate per ognuno degli effetti individuali mostrano un'asimmetria lontano dal valore centrale della media globale, mentre la distribuzione della media globale ha maggiore simmetria.

quantili a posteriori per μ sono

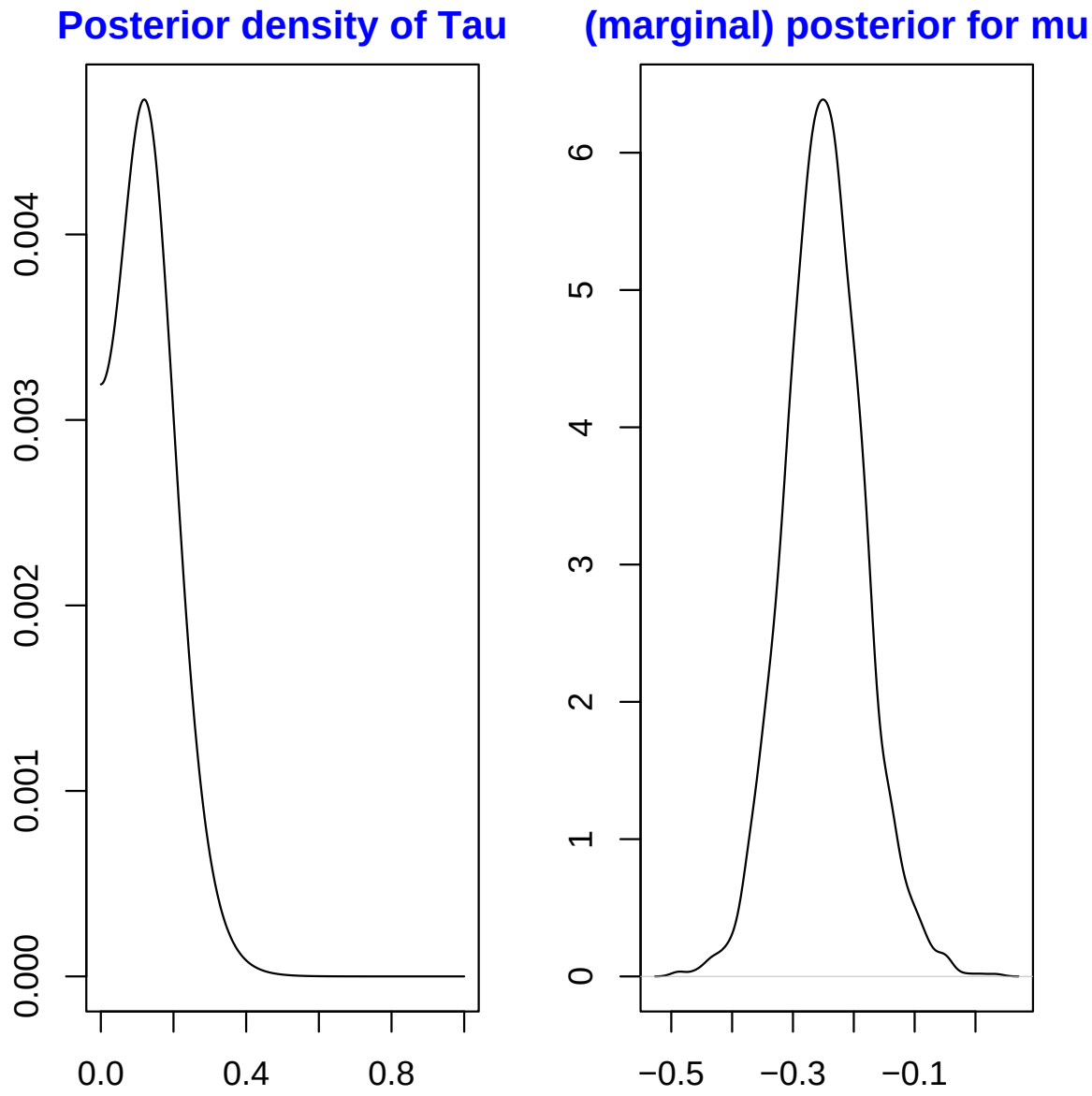
2.5%	25%	*Mediana*	75%	97.5%
-0.37	-0.29	-0.25	-0.20	-0.11

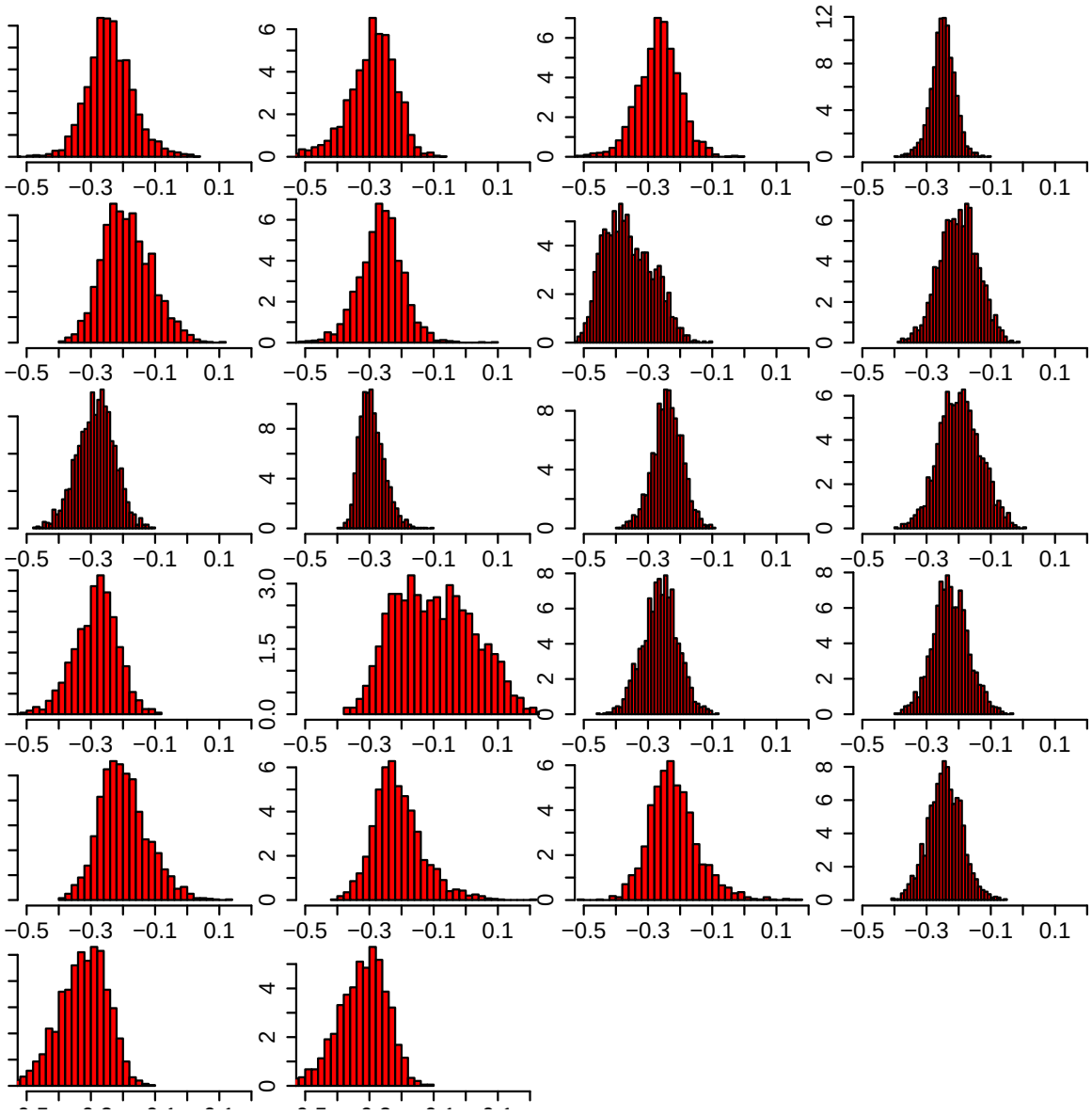
L'intervallo 95%HPD approssimato per μ è $[0.37, 0.11]$, o $[0.69, 0.90]$ se convertito nella scala dell'odds ratio. Vice versa, l'intervallo a posteriori del 95% che risulta da un pooling completo, $[0.70, 0.85]$, è meno ampio.

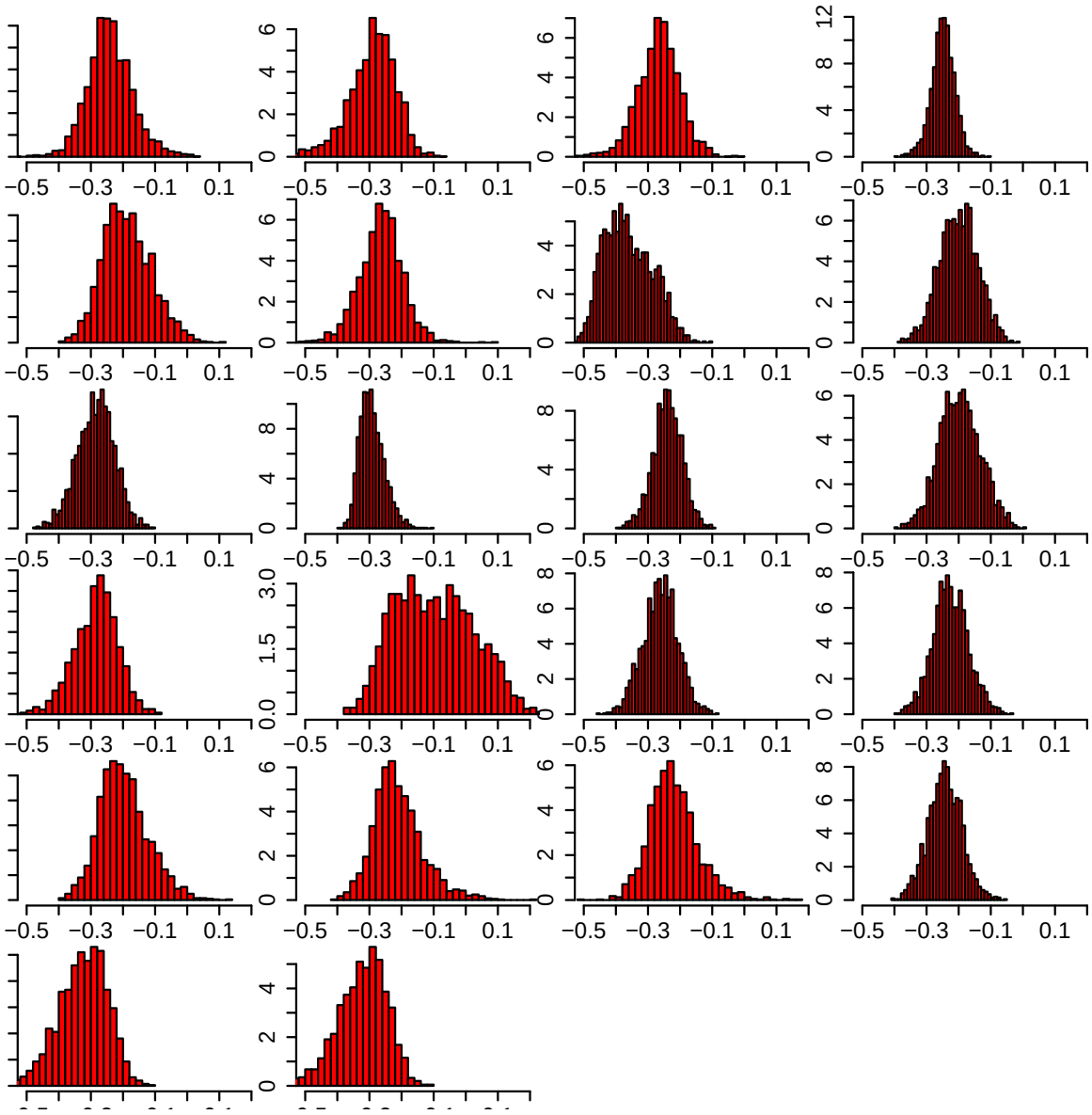
Distribuzioni a posteriori

1. Densità (marginali) a posteriori di τ e μ ▸ graph
2. Densità (istogrammi) a posteriori degli effetti θ_j ▸ graph

- Gli istogrammi delle densità a posteriori simulate per ognuno degli effetti individuali mostrano un'asimmetria lontano dal valore centrale della media globale, mentre la distribuzione della media globale ha maggiore simmetria.
- Si noti che studi maggiormente precisi sono asimmetrici rispetto al valore centrale della media della popolazione e hanno code più pesanti (si veda e.g. gli studi precisi 7 e 14).


Posterior density of Tau
0.000
0.001
0.002
0.003
0.004
0.0
0.4
0.8
(marginal) posterior for mu
0
1
2
3
4
5
6
−0.5
−0.3
−0.1


−0.5 −0.3 −0.1 0.1
0 2 4 6
0 2 4 6
0 4 8 12
0 2 4 6
0 2 4 6
0 2 4
0 2 4 6
0 4 8
0 4 8
0 2 4 6
0.0 1.5 3.0
0 2 4 6 8
0 2 4 6 8
0 2 4 6
0 2 4 6
0 2 4 6
0 2 4 6 8
0 2 4


−0.5 −0.3 −0.1 0.1
0 2 4 6
0 2 4 6
0 4 8 12
0 2 4 6
0 2 4 6
0 2 4
0 2 4 6
0 4 8
0 4 8
0 2 4 6
0.0 1.5 3.0
0 2 4 6 8
0 2 4 6 8
0 2 4 6
0 2 4 6
0 2 4 6
0 2 4

Fully HB

Sotto, per ogni valore di τ (preso nel range plausibile di valori di τ, secondo la sua marginale a posteriori), $E(\mu|\tau, y)$ e $sd(\mu|\tau, y)$ sono dati per riga.

```
tau: 0.00 0.05 0.10 0.15 0.20 0.25 0.30

mu|tau -0.260 0.05
       -0.256 0.05
       -0.245 0.06
       -0.244 0.07
       -0.240 0.07
       -0.239 0.08
       -0.238 0.09
```

Fully HB

Sotto, per ogni valore di τ (preso nel range plausibile di valori di τ, secondo la sua marginale a posteriori), $E(\mu|\tau, y)$ e $sd(\mu|\tau, y)$ sono dati per riga.

```
tau: 0.00 0.05 0.10 0.15 0.20 0.25 0.30

mu|tau -0.260 0.05
       -0.256 0.05
       -0.245 0.06
       -0.244 0.07
       -0.240 0.07
       -0.239 0.08
       -0.238 0.09
```

Si noti il quasi raddoppiamento degli sd. Ciò mostra l'importanza di mediare su τ (invece di fissare τ come in un approccio non full Bayes) in modo da adeguatamente tenere conto dell'incertezza nella sua stima.

Fully HB

Sotto, per ogni valore di τ (preso nel range plausibile di valori di τ, secondo la sua marginale a posteriori), $E(\mu|\tau, y)$ e $sd(\mu|\tau, y)$ sono dati per riga.

```
tau: 0.00 0.05 0.10 0.15 0.20 0.25 0.30

mu|tau -0.260 0.05
       -0.256 0.05
       -0.245 0.06
       -0.244 0.07
       -0.240 0.07
       -0.239 0.08
       -0.238 0.09
```

Infatti, la standard deviation condizionale a posteriori, $sd(\mu|\tau, y)$, ha il valore 0.060 per $\tau = 0.13$ (mediana a posteriori), mentre mediando rispetto alla distribuzione a posteriori per τ otteniamo un valore di $sd(\mu|y) = 0.071$.

Indice della sezione I

Indice della sezione II

Nella sezione

Simulazione diretta

Nei HBM semplici (come il modello N-N), il calcolo (stima) viene effettuato attraverso la simulazione diretta della distribuzione a posteriori congiunta

$$p(\boldsymbol{\theta}, \mu, \tau|y) \propto p(y|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mu, \tau)p(\mu, \tau)$$

Simulazione diretta

Nei HBM semplici (come il modello N-N), il calcolo (stima) viene effettuato attraverso la simulazione diretta della distribuzione a posteriori congiunta

$$p(\boldsymbol{\theta}, \mu, \tau|y) \propto p(y|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mu, \tau)p(\mu, \tau)$$

scomponendola in più parti:

Simulazione diretta

Nei HBM semplici (come il modello N-N), il calcolo (stima) viene effettuato attraverso la simulazione diretta della distribuzione a posteriori congiunta

$$p(\boldsymbol{\theta}, \mu, \tau|y) \propto p(y|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mu, \tau)p(\mu, \tau)$$

scomponendola in più parti:

I/ simula dalla distribuzione condizionale a posteriori

$$p(\boldsymbol{\theta}|\mu, \tau, y)$$

Simulazione diretta

Nei HBM semplici (come il modello N-N), il calcolo (stima) viene effettuato attraverso la simulazione diretta della distribuzione a posteriori congiunta

$$p(\boldsymbol{\theta}, \mu, \tau|y) \propto p(y|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mu, \tau)p(\mu, \tau)$$

scomponendola in più parti:

I/ simula dalla distribuzione condizionale a posteriori

$$p(\boldsymbol{\theta}|\mu, \tau, y)$$

II/ simula dalla distribuzione marginale a posteriori

$$p(\mu, \tau|y).$$

Simulazione per passi

a. Inizia dalla distribuzione congiunta a posteriori non-normalizzata

$$p(\boldsymbol{\theta}, \mu, \tau|y) \quad \propto \quad p(y|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mu, \tau)p(\mu, \tau)$$

Simulazione per passi

a. Inizia dalla distribuzione congiunta a posteriori non-normalizzata

$$p(\boldsymbol{\theta}, \mu, \tau|y) \propto p(y|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mu, \tau)p(\mu, \tau)$$

che altrimenti scritta è $p(\boldsymbol{\theta}|\mu, \tau, y)p(\mu|\tau, y)p(\tau|y)$

Simulazione per passi

a. Inizia dalla distribuzione congiunta a posteriori non-normalizzata

$$p(\boldsymbol{\theta}, \mu, \tau|y) \propto p(y|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mu, \tau)p(\mu, \tau)$$

che altrimenti scritta è $p(\boldsymbol{\theta}|\mu, \tau, y)p(\mu|\tau, y)p(\tau|y)$

b. In modelli semplici è possibile calcolare la forma analitica di $p(\boldsymbol{\theta}|\mu, \tau, y)$.

Simulazione per passi

a. Inizia dalla distribuzione congiunta a posteriori non-normalizzata

$$p(\boldsymbol{\theta}, \mu, \tau|y) \propto p(y|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mu, \tau)p(\mu, \tau)$$

che altrimenti scritta è $p(\boldsymbol{\theta}|\mu, \tau, y)p(\mu|\tau, y)p(\tau|y)$

b. In modelli semplici è possibile calcolare la forma analitica di $p(\boldsymbol{\theta}|\mu, \tau, y)$.
In N-N, con ipotesi di scambiabilità, $p(\boldsymbol{\theta}|\mu, \tau, y)$ si fattorizza in J componenti ove ciascuna risulta essere $\theta_j|\mu, \tau, y \sim N(\hat{\theta}_j, V_j)$ con

$$\hat{\theta}_j = \frac{1/\sigma_j^2 \cdot y_j + 1/\tau^2 \cdot \mu}{1/\sigma_j^2 + 1/\tau^2} \qquad V_j = \frac{1}{1/\sigma_j^2 + 1/\tau^2}$$

Simulazione per passi

a. Inizia dalla distribuzione congiunta a posteriori non-normalizzata

$$p(\boldsymbol{\theta}, \mu, \tau|y) \propto p(y|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mu, \tau)p(\mu, \tau)$$

che altrimenti scritta è

$$p(\boldsymbol{\theta}|\mu, \tau, y)p(\mu|\tau, y)p(\tau|y)$$

b. In modelli semplici è possibile calcolare la forma analitica di $p(\boldsymbol{\theta}|\mu, \tau, y)$.

c. Con riguardo a $p(\mu, \tau|y)$, il calcolo dipende da caso a caso.

Per iniziare assegniamo una a priori noninformativa per $\mu|\tau$, ie

$$p(\mu, \tau) = p(\mu|\tau)p(\tau) \propto p(\tau)$$

> *The uniform prior density for μ is generally reasonable for this problem; because the combined data from all J experiments are generally highly informative about μ, we can afford to be vague about its prior distribution.*

Simulazione per passi

c. Con riguardo a $p(\mu, \tau|y)$, il calcolo dipende da caso a caso.

Nell'esempio binomiale avevamo menzionato la via dell'integrazione o del calcolo analitico per derivare $p(\mu, \tau|y)$ da $p(\boldsymbol{\theta}, \mu, \tau|y)$

Simulazione per passi

c. Con riguardo a $p(\mu, \tau|y)$, il calcolo dipende da caso a caso.

c1. In N-N è possibile calcolare la verosimiglianza marginale $p(y|\mu, \tau)$, ad es. essa è $\forall j\ y_j \sim N(\mu, \sigma_j^2 + \tau^2)$ pertanto

$$p(\mu, \tau|y) \propto p(\mu, \tau) \prod_j N(y_j|\mu, \sigma_j^2 + \tau^2)$$

Potremmo quindi calcolare direttamente $p(\mu, \tau|y)$ come funzione di due variabili e procedere come nel caso binomiale. Invece, possiamo semplificare ulteriormente, fattorizzandola,

$$p(\mu, \tau|y) = p(\mu|\tau, y)p(\tau|y),$$

e calcolare prima $p(\mu|\tau, y)$ e poi la marginale $p(\tau|y)$. Infatti, sotto l'ipotesi $p(\mu|\tau) \propto 1$, i.e. uniforme, su $(-\infty, \infty)$, otteniamo $\mu|\tau, y \sim N(\hat{\mu}, V_\mu)$ con

$$\hat{\mu} = \frac{\sum \frac{1}{\sigma_j^2+\tau^2} y_j}{\sum \frac{1}{\sigma_j^2+\tau^2}} \qquad V_\mu = \frac{1}{\sum \frac{1}{\sigma_j^2+\tau^2}}$$

Simulazione per passi

c. Con riguardo a $p(\mu, \tau|y)$, il calcolo dipende da caso a caso.

c2. Infine, $p(\tau|y)$

Simulazione per passi

c. Con riguardo a $p(\mu, \tau|y)$, il calcolo dipende da caso a caso.

c2. Infine, $p(\tau|y)$

- può essere calcolato analiticamente come

$$\frac{p(\mu, \tau|y)}{p(\mu|\tau, y)} = \text{una forma complicata ... (vd. BDA3 p.117, eq. 5.21)}$$

Simulazione per passi

c. Con riguardo a $p(\mu, \tau|y)$, il calcolo dipende da caso a caso.

c2. Infine, $p(\tau|y)$

- può essere calcolato analiticamente come

$$\frac{p(\mu, \tau|y)}{p(\mu|\tau, y)} = \text{una forma complicata ... (vd. BDA3 p.117, eq. 5.21)}$$

- Per completare dobbiamo assegnare una a priori per τ
  - $p(\tau) \propto 1$ porta ad una densità a post propria
  - $p(\log \tau) \propto 1$ porta ad una densità a post impropria
  - $p(\tau) \sim \text{Scaled} - \text{Inv}\chi^2$ data una *best guess* e un *upper bound* (a priori informativa, più realistica)

Simulazione per passi

a. Inizia dalla distribuzione congiunta a posteriori non-normalizzata

$$p(\boldsymbol{\theta}, \mu, \tau|y) \quad \propto \quad p(y|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mu, \tau)p(\mu, \tau)$$

b. In modelli semplici è possibile calcolare la forma analitica di $p(\boldsymbol{\theta}|\mu, \tau, y)$.

c. Con riguardo a $p(\mu, \tau|y)$, il calcolo dipende da caso a caso.

Nella sezione

Stima dei modelli gerarchici Bayesiani

- Simulazione diretta
- Simulazione iterativa

Simulazione MCMC

La stima di un HBM può essere effettuata attraverso una

Simulazione MCMC

La stima di un HBM può essere effettuata attraverso una

1. simulazione diretta

Simulazione MCMC

La stima di un HBM può essere effettuata attraverso una

1. simulazione diretta come con Il modello N-N dove la distribuzione a posteriori

$$p(\boldsymbol{\theta}, \mu, \tau|y) \propto p(\tau|y)p(\mu|\tau, y)p(\boldsymbol{\theta}|\mu, \tau, y)$$

consiste in forme chiuse $p(\mu|\tau, y)$ e $p(\boldsymbol{\theta}|\mu, \tau, y)$ (per a priori coniugate), e in $p(\tau|y)$ che, a seconda della specificazione a priori, può anche essere calcolabile analiticamente.

Simulazione MCMC

La stima di un HBM può essere effettuata attraverso una

1. simulazione diretta
2. simulazione iterativa o MCMC,

Simulazione MCMC

La stima di un HBM può essere effettuata attraverso una

1. simulazione diretta
2. simulazione iterativa o MCMC, i.e. integrazione Monte Carlo di estrazioni simulate da una Markov Chain che ha la densità a posteriori $p(\boldsymbol{\theta}|y)$ (dove $\boldsymbol{\theta}$ sono il complesso dei parametri) come distribuzione target.

Simulazione MCMC

La stima di un HBM può essere effettuata attraverso una

1. simulazione diretta
2. simulazione iterativa o MCMC,
   - in particolare, via Gibbs sampling.

Indice della sezione I

Indice della sezione II

Gibbs sampling

MCMC methods allow for sampling from complicated, high dimensional models.

Gibbs sampling

MCMC methods allow for sampling from complicated, high dimensional models.

In particular, Gibbs sampler can be straightforwardly implemented when there is

Gibbs sampling

MCMC methods allow for sampling from complicated, high dimensional models.

In particular, Gibbs sampler can be straightforwardly implemented when there is

- a conditional independence structure (like in HM);

Gibbs sampling

MCMC methods allow for sampling from complicated, high dimensional models.

In particular, Gibbs sampler can be straightforwardly implemented when there is

- a conditional independence structure (like in HM);
- (conditionally) conjugate priors for each hierarchical stage.

Gibbs sampling

MCMC methods allow for sampling from complicated, high dimensional models.

In particular, Gibbs sampler can be straightforwardly implemented when there is

- a conditional independence structure (like in HM);
- (conditionally) conjugate priors for each hierarchical stage. E.g., every parameter is assigned a prior to be conjugate to the likelihood associated with its own 'child' in the hierarchy.

Full conditionals

Specification of conditionally conjugate priors makes the full conditional distributions (required for implementing the Gibbs sampler) be in closed forms.

Full conditionals

Specification of conditionally conjugate priors makes the full conditional distributions (required for implementing the Gibbs sampler) be in closed forms.

The full conditional for each variable can be derived from

Full conditionals

Specification of conditionally conjugate priors makes the full conditional distributions (required for implementing the Gibbs sampler) be in closed forms.

The full conditional for each variable can be derived from

- the links

Full conditionals

Specification of conditionally conjugate priors makes the full conditional distributions (required for implementing the Gibbs sampler) be in closed forms.

The full conditional for each variable can be derived from

- the links
- the conditional distributions

Full conditionals

Specification of conditionally conjugate priors makes the full conditional distributions (required for implementing the Gibbs sampler) be in closed forms.

The full conditional for each variable can be derived from

- the links
- the conditional distributions

of the Directed Acyclic Graph (DAG).

DAG

A directed acyclic graph (DAG) is an alternative representation of a model.

- it is ‘directed’ because each link is an arrow

DAG

A directed acyclic graph (DAG) is an alternative representation of a model.

- it is ‘directed’ because each link is an arrow
- it is ‘acyclic’ because by following the arrows it is not possible to return to a node after leaving it

DAG

- The graph model displays the causal structure of the problem:

▸ a DAG example

DAG

- The graph model displays the causal structure of the problem:
  - direct arrows (stochastic or logical dependences) towards nodes (stochastic or deterministic variables or constants) represent a direct influence of (direct) parents on children.

▸ a DAG example

DAG

- The graph model displays the causal structure of the problem:
  - direct arrows (stochastic or logical dependences) towards nodes (stochastic or deterministic variables or constants) represent a direct influence of (direct) parents on children.
  - missing links represent non-relevance between nodes.

▸ a DAG example

DAG

- The graph model displays the causal structure of the problem:
  - direct arrows (stochastic or logical dependences) towards nodes (stochastic or deterministic variables or constants) represent a direct influence of (direct) parents on children.
  - missing links represent non-relevance between nodes.
- The probabilistic structure is then added to the model. Let denote

▸ a DAG example

DAG

- The graph model displays the causal structure of the problem:
  - direct arrows (stochastic or logical dependences) towards nodes (stochastic or deterministic variables or constants) represent a direct influence of (direct) parents on children.
  - missing links represent non-relevance between nodes.
- The probabilistic structure is then added to the model. Let denote
  - $V = \{v\}$ the set of nodes

▸ a DAG example

DAG

- The graph model displays the causal structure of the problem:
  - direct arrows (stochastic or logical dependences) towards nodes (stochastic or deterministic variables or constants) represent a direct influence of (direct) parents on children.
  - missing links represent non-relevance between nodes.
- The probabilistic structure is then added to the model. Let denote
  - $V = \{v\}$ the set of nodes
  - $p(v|pa(v))$ the conditional distribution of v (child) given its (stochastic) parents $pa(v)$.

▸ a DAG example

DAG

- The graph model displays the causal structure of the problem:
  - direct arrows (stochastic or logical dependences) towards nodes (stochastic or deterministic variables or constants) represent a direct influence of (direct) parents on children.
  - missing links represent non-relevance between nodes.
- The probabilistic structure is then added to the model. Let denote
  - $V = \{v\}$ the set of nodes
  - $p(v|pa(v))$ the conditional distribution of v (child) given its (stochastic) parents $pa(v)$.
  - Hence, the probability structure results $p(V) = \prod_{v \in V} p(v|pa(v))$ by assuming conditional independence wherever there is non-relevance between nodes.

▸ a DAG example

DAG

- The graph model displays the causal structure of the problem:
  - direct arrows (stochastic or logical dependences) towards nodes (stochastic or deterministic variables or constants) represent a direct influence of (direct) parents on children.
  - missing links represent non-relevance between nodes.
- The probabilistic structure is then added to the model. Let denote
  - $V = \{v\}$ the set of nodes
  - $p(v|pa(v))$ the conditional distribution of v (child) given its (stochastic) parents $pa(v)$.
  - Hence, the probability structure results $p(V) = \prod_{v \in V} p(v|pa(v))$ by assuming conditional independence wherever there is non-relevance between nodes.
    Thereby, $p(V)$ consists of the product of priors for nodes without parents and of conditional distributions for the remaining children given their (stochastic) parents.

▸ a DAG example

DAG

- The graph model displays the causal structure of the problem:
  - direct arrows (stochastic or logical dependences) towards nodes (stochastic or deterministic variables or constants) represent a direct influence of (direct) parents on children.
  - missing links represent non-relevance between nodes.
- The probabilistic structure is then added to the model. Let denote
  - $V = \{v\}$ the set of nodes
  - $p(v|pa(v))$ the conditional distribution of v (child) given its (stochastic) parents $pa(v)$.
  - Hence, the probability structure results $p(V) = \prod_{v \in V} p(v|pa(v))$ by assuming conditional independence wherever there is non-relevance between nodes.
    Thereby, $p(V)$ consists of the product of priors for nodes without parents and of conditional distributions for the remaining children given their (stochastic) parents.
    The marginal distribution for each node (given the observed data) can be calculated via Bayes rule.

▸ a DAG example

A DAG example

Consider a simple linear regression problem given by

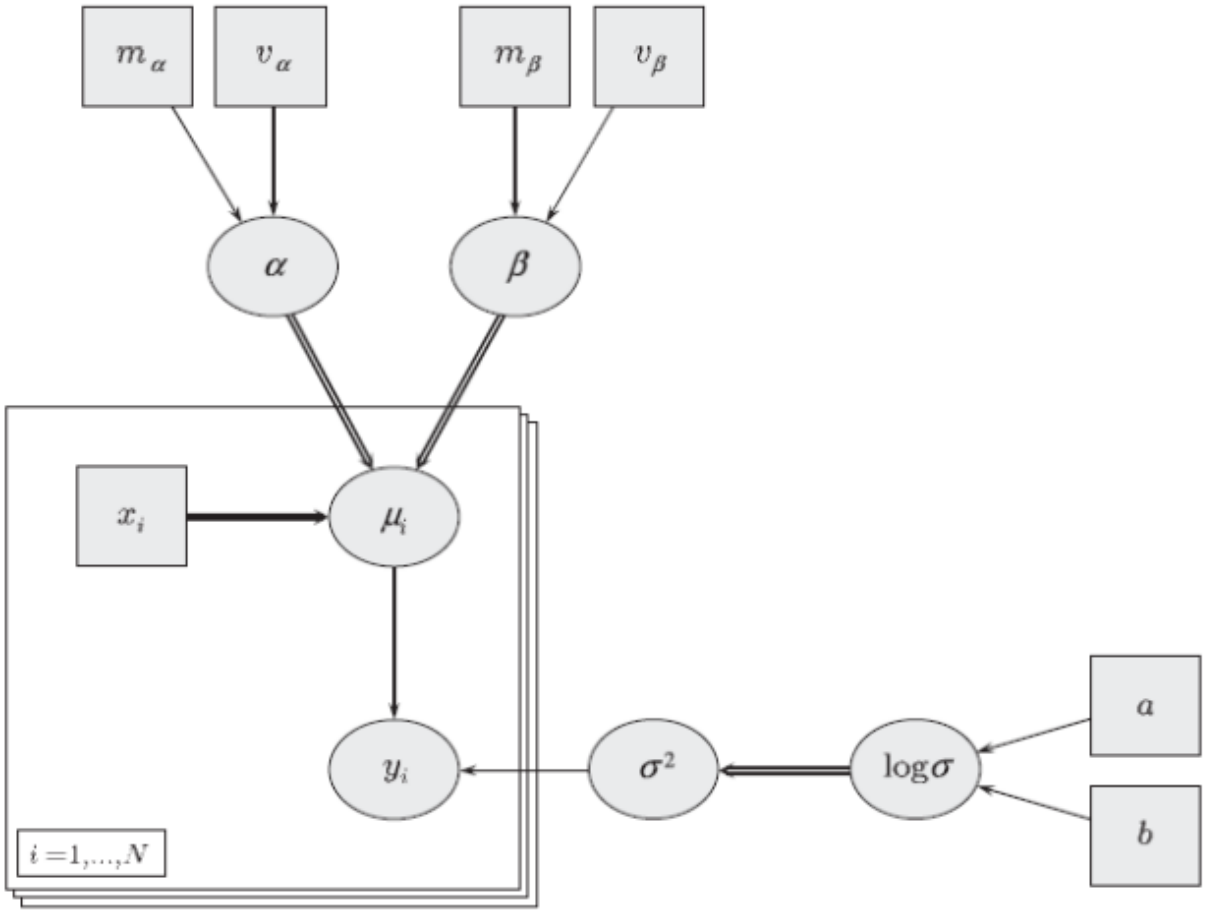
$$y_i \sim N(\mu_i, \sigma^2) \qquad \mu_i = \alpha + \beta\, x_i \quad i = 1, \ldots, N$$
$$\alpha \sim N(m_\alpha, v_\alpha) \quad \beta \sim N(m_\beta, v_\beta) \quad \log \sigma \sim U(a, b)$$

for fixed constants $m_\alpha, v_\alpha, m_\beta, v_\beta, a$ and b.
An alternative representation of this model is the directed acyclic graph shown below.

▸ a 2-stage linear model

A DAG example

◂ DAG description ▸ DAG interpretation



DAG interpretation

◂ a DAG example The notation is defined as follows.

- Rectangular nodes denote known constants.

DAG interpretation

◂ a DAG example The notation is defined as follows.

- Rectangular nodes denote known constants.
- Elliptical nodes represent either deterministic relationships (i.e. functions) or stochastic quantities, i.e. quantities that require a distributional assumption.

DAG interpretation

◂ a DAG example The notation is defined as follows.

- Rectangular nodes denote known constants.
- Elliptical nodes represent either deterministic relationships (i.e. functions) or stochastic quantities, i.e. quantities that require a distributional assumption.
- Stochastic dependence and functional dependence are denoted by single-edged arrows and double-edged arrows, respectively.

DAG interpretation

◂ a DAG example The notation is defined as follows.

- Rectangular nodes denote known constants.
- Elliptical nodes represent either deterministic relationships (i.e. functions) or stochastic quantities, i.e. quantities that require a distributional assumption.
- Stochastic dependence and functional dependence are denoted by single-edged arrows and double-edged arrows, respectively.
- Repetitive structures, such as the 'loop' from i =1 to N, are represented by 'plates', which may be nested if the model is hierarchical.

Full conditional

- The full conditional for v is calculated as

$$p(v|V-v) \propto p(v, V-v) = p(V).$$

That is, it is proportional to the product of the terms in $p(V)$ containing v, i.e.,

$$p(v|V-v) \propto p(v|pa(v)) \prod_{v \in pa(w)} p(w|pa(w)),$$

where $\prod_{v \in pa(w)} p(w|pa(w))$ is the product of distributions where v is a parent itself.

Full conditional

- The full conditional for v is calculated as

$$p(v|V-v) \propto p(v, V-v) = p(V).$$

That is, it is proportional to the product of the terms in $p(V)$ containing v, i.e.,

$$p(v|V-v) \propto p(v|pa(v)) \prod_{v \in pa(w)} p(w|pa(w)),$$

where $\prod_{v \in pa(w)} p(w|pa(w))$ is the product of distributions where v is a parent itself.

- Thus, by exploiting the DAG as well as the 'conjugancy', the full conditionals of the linear hierarchical model are $\rightarrow$

Model specification within BUGS

◂ a simple regression

A customary GLMM model in HB form is specified as

▸ prior specification ▸ full conditionals

Model specification within BUGS

◂ a simple regression

A customary GLMM model in HB form is specified as

1. $\mathbf{Y}_{ij} \overset{ind}{\sim} N(\alpha_i + \beta_i x_{ij},\, \sigma^2) \qquad j = 1, \ldots, J$

▸ prior specification ▸ full conditionals

Model specification within BUGS

◂ a simple regression

A customary GLMM model in HB form is specified as

1. $\mathbf{Y}_{ij} \overset{ind}{\sim} N(\alpha_i + \beta_i x_{ij},\, \sigma^2) \qquad j = 1, \ldots, J$

2. $\boldsymbol{\theta}_i \equiv \begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \overset{i.i.d.}{\sim} N\left(\boldsymbol{\theta}_0 \equiv \begin{pmatrix} \alpha_0 \\ \beta_0 \end{pmatrix}, \Sigma\right) \qquad i = 1, \ldots, I$

▸ prior specification ▸ full conditionals

Model specification within BUGS

◂ a simple regression

A customary GLMM model in HB form is specified as

1. $\mathbf{Y}_{ij} \overset{ind}{\sim} N(\alpha_i + \beta_i x_{ij},\ \sigma^2) \qquad j = 1, \ldots, J$

2. $\boldsymbol{\theta}_i \equiv \begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \overset{i.i.d.}{\sim} N\left(\boldsymbol{\theta}_0 \equiv \begin{pmatrix} \alpha_0 \\ \beta_0 \end{pmatrix}, \Sigma\right) \qquad i = 1, \ldots, I$

   $\sigma^{-2} \sim G(a, b)$ (according to a full HB)

▸ prior specification ▸ full conditionals

Model specification within BUGS

◂ a simple regression

A customary GLMM model in HB form is specified as

1. $\mathbf{Y}_{ij} \overset{ind}{\sim} N(\alpha_i + \beta_i x_{ij},\ \sigma^2) \qquad j = 1, \ldots, J$

2. $\boldsymbol{\theta}_i \equiv \begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \overset{i.i.d.}{\sim} N\left(\boldsymbol{\theta}_0 \equiv \begin{pmatrix} \alpha_0 \\ \beta_0 \end{pmatrix}, \Sigma\right) \qquad i = 1, \ldots, I$

   $\sigma^{-2} \sim G(a, b)$ (according to a full HB)

3. $\boldsymbol{\theta}_0 \sim N(\eta,\ C) \qquad \Sigma^{-1} \sim W((\rho R)^{-1},\ \rho)$ (according to a full HB)

▸ prior specification ▸ full conditionals

Model specification within BUGS

◂ a simple regression

A customary GLMM model in HB form is specified as

1. $\mathbf{Y}_{ij} \overset{ind}{\sim} N(\alpha_i + \beta_i x_{ij},\ \sigma^2) \qquad j = 1, \ldots, J$

2. $\boldsymbol{\theta}_i \equiv \begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \overset{i.i.d.}{\sim} N\left(\boldsymbol{\theta}_0 \equiv \begin{pmatrix} \alpha_0 \\ \beta_0 \end{pmatrix}, \Sigma\right) \qquad i = 1, \ldots, I$

   $\sigma^{-2} \sim G(a, b)$ (according to a full HB)

3. $\boldsymbol{\theta}_0 \sim N(\eta,\ C) \qquad \Sigma^{-1} \sim W((\rho R)^{-1},\ \rho)$ (according to a full HB)

   $\sigma^2,\ \boldsymbol{\theta}_0,\ \Sigma^{-1}$ are the "stochastic founder nodes" in the Directed Acyclic Graph (DAG).

▸ prior specification ▸ full conditionals

Prior specification within BUGS

◂ a 2-stage linear model ▸ full condirionals

If one is not willing/able to specify an “influent” prior, a non-informative prior has to be chosen.

Prior specification within BUGS

◂ a 2-stage linear model ▸ full condirionals

If one is not willing/able to specify an “influent” prior, a non-informative prior has to be chosen. Yet, BUGS requires a proper probability model.

Prior specification within BUGS

◂ a 2-stage linear model ▸ full condirionals

If one is not willing/able to specify an “influent” prior, a non-informative prior has to be chosen. Yet, BUGS requires a proper probability model.

- A locally uniform prior (on likelihood support region) is typically chosen for location parameters (e.g. regression coefs)

Prior specification within BUGS

If one is not willing/able to specify an “influent” prior, a non-informative prior has to be chosen. Yet, BUGS requires a proper probability model.

- A locally uniform prior (on likelihood support region) is typically chosen for location parameters (e.g. regression coefs)

  e.g. a “vague” Normal

$$\boldsymbol{\theta}_0 \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1.0E-4 & 0 \\ 0 & 1.0E-4 \end{pmatrix}\right)$$

Prior specification within BUGS

◂ a 2-stage linear model ▸ full condirionals

If one is not willing/able to specify an “influent” prior, a non-informative prior has to be chosen. Yet, BUGS requires a proper probability model.

- A locally uniform prior (on likelihood support region) is typically chosen for location parameters (e.g. regression coefs)
- A “barely proper” prior is chosen for the variance parameters (in precision form)

Prior specification within BUGS

◂ a 2-stage linear model ▸ full condirionals

If one is not willing/able to specify an “influent” prior, a non-informative prior has to be chosen. Yet, BUGS requires a proper probability model.

- A locally uniform prior (on likelihood support region) is typically chosen for location parameters (e.g. regression coefs)
- A “barely proper” prior is chosen for the variance parameters (in precision form)
  e.g. an “almost improper” Gamma for (univariate) scale parameters

$$\sigma^{-2} \sim G(\epsilon, \epsilon)$$

Prior specification within BUGS

◂ a 2-stage linear model ▸ full condirionals

If one is not willing/able to specify an “influent” prior, a non-informative prior has to be chosen. Yet, BUGS requires a proper probability model.

- A locally uniform prior (on likelihood support region) is typically chosen for location parameters (e.g. regression coefs)
- A “barely proper” prior is chosen for the variance parameters (in precision form)
  e.g. a Wishart for multivariate scale parameters

$$\Sigma^{-1} \sim W\left(\begin{pmatrix} S_1 & 0 \\ 0 & S_2 \end{pmatrix}, \rho\right)$$

Prior specification within BUGS

◂ a 2-stage linear model ▸ full condirionals

If one is not willing/able to specify an "influent" prior, a non-informative prior has to be chosen. Yet, BUGS requires a proper probability model.

- A locally uniform prior (on likelihood support region) is typically chosen for location parameters (e.g. regression coefs)
- A "barely proper" prior is chosen for the variance parameters (in precision form)
  e.g. a Wishart for multivariate scale parameters

$$\Sigma^{-1} \sim W\left(\begin{pmatrix} S_1 & 0 \\ 0 & S_2 \end{pmatrix}, \rho\right)$$

The prior for the precision matrix Σ^{-1} is made 'vague' by choosing the minimum value for the number of df ρ, i.e., $\rho = \text{rank}(\Sigma)$ (in this case minimum is 2).
The scale matrix divided by df is our prior opinion on the order of magnitude of the covariance matrix Σ.

Prior specification within BUGS

◂ a 2-stage linear model ▸ full condirionals

If one is not willing/able to specify an “influent” prior, a non-informative prior has to be chosen. Yet, BUGS requires a proper probability model.

- A locally uniform prior (on likelihood support region) is typically chosen for location parameters (e.g. regression coefs)
- A “barely proper” prior is chosen for the variance parameters (in precision form)

In the above examples all the priors for stochastic founder nodes have been chosen as conjugate priors.

Prior specification within BUGS

◂ a 2-stage linear model ▸ full condirionals

If one is not willing/able to specify an “influent” prior, a non-informative prior has to be chosen. Yet, BUGS requires a proper probability model.

- A locally uniform prior (on likelihood support region) is typically chosen for location parameters (e.g. regression coefs)
- A “barely proper” prior is chosen for the variance parameters (in precision form)

In the above examples all the priors for stochastic founder nodes have been chosen as conjugate priors. Their own hyperparameters were assumed known.

Full conditionals of 2-stage linear model

◂ a 2-stage linear model ◂ prior specification

- $\boldsymbol{\theta}_i|\boldsymbol{\theta}_0, \Sigma, \mathbf{y}, \sigma \ \propto \ N(\boldsymbol{\theta}_0, \Sigma) \cdot \prod_j N(\alpha_i + \beta_i x_{ij}, \ \sigma^2)$
  $\sim N(\ldots),$

Full conditionals of 2-stage linear model

◂ a 2-stage linear model ◂ prior specification

- $\boldsymbol{\theta}_i|\boldsymbol{\theta}_0, \Sigma, \mathbf{y}, \sigma \;\propto\; N(\boldsymbol{\theta}_0, \Sigma) \cdot \prod_j N(\alpha_i + \beta_i x_{ij},\, \sigma^2)$
  $\sim N(\ldots),$
- $\boldsymbol{\theta}_0|\{\boldsymbol{\theta}_i\}, \Sigma^{-1} \;\propto\; N(\eta,\, C) \cdot \prod_i N(\boldsymbol{\theta}_0, \Sigma)$
  $\sim N(\ldots),$

Full conditionals of 2-stage linear model

◂ a 2-stage linear model ◂ prior specification

- $\boldsymbol{\theta}_i|\boldsymbol{\theta}_0, \Sigma, \mathbf{y}, \sigma \;\propto\; N(\boldsymbol{\theta}_0, \Sigma) \cdot \prod_j N(\alpha_i + \beta_i x_{ij},\, \sigma^2)$
  $\sim N(\ldots)$,
- $\boldsymbol{\theta}_0|\{\boldsymbol{\theta}_i\}, \Sigma^{-1} \;\propto\; N(\eta,\, C) \cdot \prod_i N(\boldsymbol{\theta}_0, \Sigma)$
  $\sim N(\ldots)$,
- $\Sigma^{-1}|\{\boldsymbol{\theta}_i\}, \boldsymbol{\theta}_0 \;\propto\; W((\rho R)^{-1},\, \rho) \cdot \prod_i N(\boldsymbol{\theta}_0, \Sigma)$
  $\sim W(\ldots)$

Full conditionals of 2-stage linear model

◂ a 2-stage linear model ◂ prior specification

- $\boldsymbol{\theta}_i|\boldsymbol{\theta}_0, \Sigma, \mathbf{y}, \sigma \;\propto\; N(\boldsymbol{\theta}_0, \Sigma) \cdot \prod_j N(\alpha_i + \beta_i x_{ij},\, \sigma^2)$
  $\sim N(\ldots),$
- $\boldsymbol{\theta}_0|\{\boldsymbol{\theta}_i\}, \Sigma^{-1} \;\propto\; N(\eta,\, C) \cdot \prod_i N(\boldsymbol{\theta}_0, \Sigma)$
  $\sim N(\ldots),$
- $\Sigma^{-1}|\{\boldsymbol{\theta}_i\}, \boldsymbol{\theta}_0 \;\propto\; W((\rho R)^{-1},\, \rho) \cdot \prod_i N(\boldsymbol{\theta}_0, \Sigma)$
  $\sim W(\ldots)$
- $\sigma^{-2}|\mathbf{y}, \{\boldsymbol{\theta}_i\} \;\propto\; G(a, b) \cdot \prod_{i,j} N(\alpha_i + \beta_i x_{ij},\, \sigma^2)$
  $\sim G(\ldots)$

Full conditionals of 2-stage linear model

◂ a 2-stage linear model ◂ prior specification

- $\boldsymbol{\theta}_i|\boldsymbol{\theta}_0, \Sigma, \mathbf{y}, \sigma \;\propto\; N(\boldsymbol{\theta}_0, \Sigma) \cdot \prod_j N(\alpha_i + \beta_i x_{ij},\, \sigma^2)$
  $\sim N(\ldots)$,
- $\boldsymbol{\theta}_0|\{\boldsymbol{\theta}_i\}, \Sigma^{-1} \;\propto\; N(\eta,\, C) \cdot \prod_i N(\boldsymbol{\theta}_0, \Sigma)$
  $\sim N(\ldots)$,
- $\Sigma^{-1}|\{\boldsymbol{\theta}_i\}, \boldsymbol{\theta}_0 \;\propto\; W((\rho R)^{-1},\, \rho) \cdot \prod_i N(\boldsymbol{\theta}_0, \Sigma)$
  $\sim W(\ldots)$
- $\sigma^{-2}|\mathbf{y}, \{\boldsymbol{\theta}_i\} \;\propto\; G(a, b) \cdot \prod_{i,j} N(\alpha_i + \beta_i x_{ij},\, \sigma^2)$
  $\sim G(\ldots)$

now we have all we need for simulating the parameters from each full conditional distribution.

Indice della sezione I

Indice della sezione II

Software

BUGS is a software package (a high-level language) for performing Bayesian inference Using Gibbs Sampling.

Software

BUGS is a software package (a high-level language) for performing Bayesian inference Using Gibbs Sampling.

1. The user specifies a statistical model, of (almost) arbitrary complexity (by simply stating the relationships between related variables).

Software

BUGS is a software package (a high-level language) for performing Bayesian inference Using Gibbs Sampling.

1. The user specifies a statistical model, of (almost) arbitrary complexity (by simply stating the relationships between related variables).
2. The software includes an 'expert system', which determines an appropriate MCMC scheme (based on the Gibbs sampler) for analysing the specified model.

Software

BUGS is a software package (a high-level language) for performing Bayesian inference Using Gibbs Sampling.

1. The user specifies a statistical model, of (almost) arbitrary complexity (by simply stating the relationships between related variables).
2. The software includes an 'expert system', which determines an appropriate MCMC scheme (based on the Gibbs sampler) for analysing the specified model.
3. The user then controls the execution of the scheme and is free to choose from a wide range of output types.

Software

BUGS is a software package (a high-level language) for performing Bayesian inference Using Gibbs Sampling.

1. The user specifies a statistical model, of (almost) arbitrary complexity (by simply stating the relationships between related variables).
2. The software includes an 'expert system', which determines an appropriate MCMC scheme (based on the Gibbs sampler) for analysing the specified model.
3. The user then controls the execution of the scheme and is free to choose from a wide range of output types.

All you must care about is essentially the model specification!

Software

BUGS is a software package (a high-level language) for performing Bayesian inference Using Gibbs Sampling.

1. The user specifies a statistical model, of (almost) arbitrary complexity (by simply stating the relationships between related variables).
2. The software includes an 'expert system', which determines an appropriate MCMC scheme (based on the Gibbs sampler) for analysing the specified model.
3. The user then controls the execution of the scheme and is free to choose from a wide range of output types.

All you must care about is essentially the model specification!
There are several versions of Bugs.
See the BUGS project at
http://www.mrc-bsu.cam.ac.uk/software/bugs/.

Software

WinBUGS

http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml
It is an established and stable, stand-alone version of the software, which will remain available but not further developed.

Software

WinBUGS

http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml
It is an established and stable, stand-alone version of the software, which will remain available but not further developed.

OpenBUGS

http://www.openbugs.info/w/
It is an open-source version of the package, on which all future development work will be focused.
For reference, see *The BUGS project: Evolution, critique and future directions*, by David Lunn, David Spiegelhalter, Andrew Thomas and Nicky Best, *Statistics in Medicine*, 2009.

JAGS

http://www-fis.iarc.fr/~martyn/software/jags/
JAGS is Just Another Gibbs Sampler.
It is a program for analysis of Bayesian hierarchical models using MCMC simulation not wholly unlike BUGS. JAGS was written with three aims in mind:

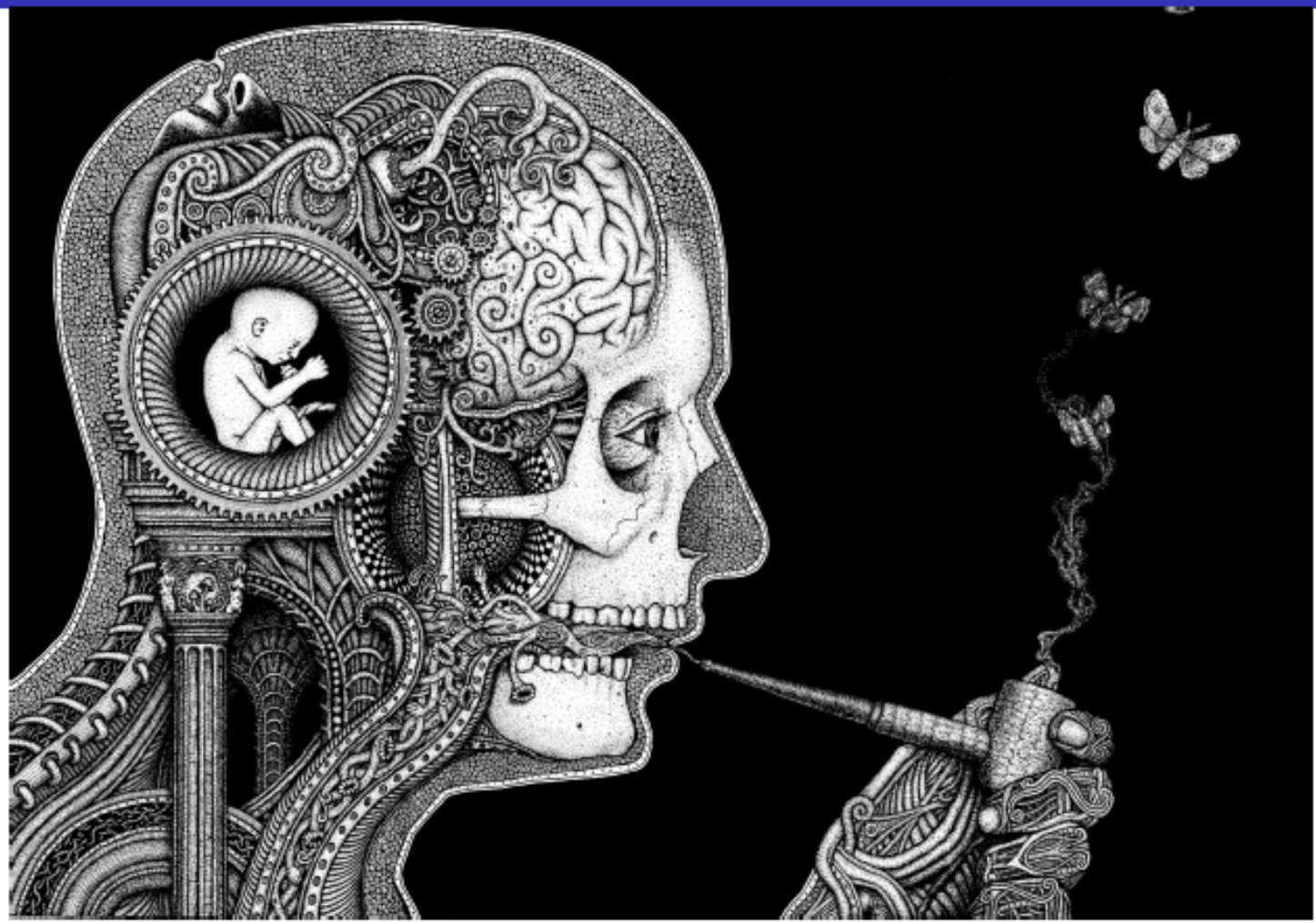
- To have an engine for the BUGS language that runs on Unix
- To be extensible, allowing users to write their own functions, distributions and samplers.
- To be a plaftorm for experimentation with ideas in Bayesian modelling.

To this last end, JAGS is licensed under the GNU General Public License.

Software

It is possible to set up models and run them entirely within Bugs, but in practice it is almost always necessary to process data before entering them into a model, and to process the inferences after the model is fitted, and so (after learned how Bugs works!) you can opt for running Bugs by calling it from R (using the bugs() function in R, you can find the instructions at http://www.stat.columbia.edu/~gelman/bugsR/).

A possible model of Bayesian data analysis



The computer does step 2 (Bayesian inference), the homunculus does step 1 (model building) and step 3 (model checking)

Indice della sezione I

Indice della sezione II

Improving Gibbs sampling computation

1. Variable transformation: centering the covariates around the grand mean.

Improving Gibbs sampling computation

1. Variable transformation: centering the covariates around the grand mean.

Improving Gibbs sampling computation

1. Variable transformation: centering the covariates around the grand mean.
   Here, variable transformation is strategically used for improving MCMC. In fact, cross-correlation between intercept and slope is reduced with covariates centering.

Improving Gibbs sampling computation

1. Variable transformation: centering the covariates around the grand mean.
   Here, variable transformation is strategically used for improving MCMC. In fact, cross-correlation between intercept and slope is reduced with covariates centering.
   The 'univariate' version of the Normal hierarchical model is:

Improving Gibbs sampling computation

1. Variable transformation: centering the covariates around the grand mean.
   Here, variable transformation is strategically used for improving MCMC. In fact, cross-correlation between intercept and slope is reduced with covariates centering.
   The 'univariate' version of the Normal hierarchical model is:
   (1) $\mathbf{Y}_{ij} \overset{ind}{\sim} N(\alpha_i + \beta_i(x_{ij} - \bar{x}), \sigma^2)$,

Improving Gibbs sampling computation

1. Variable transformation: centering the covariates around the grand mean.
   Here, variable transformation is strategically used for improving MCMC. In fact, cross-correlation between intercept and slope is reduced with covariates centering.
   The 'univariate' version of the Normal hierarchical model is:
   (1) $\mathbf{Y}_{ij} \overset{ind}{\sim} N(\alpha_i + \beta_i(x_{ij} - \bar{x}), \sigma^2)$,
   so that α_i and β_i are a priori independent (full balanced data).

Improving Gibbs sampling computation

1. Variable transformation: centering the covariates around the grand mean.
   Here, variable transformation is strategically used for improving MCMC. In fact, cross-correlation between intercept and slope is reduced with covariates centering.
   The 'univariate' version of the Normal hierarchical model is:
   (1) $\mathbf{Y}_{ij} \overset{ind}{\sim} N(\alpha_i + \beta_i(x_{ij} - \bar{x}), \sigma^2)$,
   so that α_i and β_i are a priori independent (full balanced data).
   (2) $\alpha_i \overset{i.i.d.}{\sim} N(\alpha_0, \sigma^2_\alpha)$, $\quad \beta_i \overset{i.i.d.}{\sim} N(\beta_0, \sigma^2_\beta)$

Improving Gibbs sampling computation

1. Variable transformation: centering the covariates around the grand mean.
   Here, variable transformation is strategically used for improving MCMC. In fact, cross-correlation between intercept and slope is reduced with covariates centering.
   The 'univariate' version of the Normal hierarchical model is:
   (1) $\mathbf{Y}_{ij} \overset{ind}{\sim} N(\alpha_i + \beta_i(x_{ij} - \bar{x}), \sigma^2)$,
   so that α_i and β_i are a priori independent (full balanced data).
   (2) $\alpha_i \overset{i.i.d.}{\sim} N(\alpha_0, \sigma^2_\alpha)$, $\quad \beta_i \overset{i.i.d.}{\sim} N(\beta_0, \sigma^2_\beta)$
   (3) $\alpha_0 \sim N(0, .00001)$, $\quad \beta_0 \sim N(0, .00001)$,
   $\sigma^{-2}_\alpha \sim G(\epsilon, \epsilon)$, $\quad \sigma^{-2}_\beta \sim G(\epsilon, \epsilon)$

Improving Gibbs sampling computation

1. Variable transformation: centering the covariates around the grand mean.
2. Hierarchically centered parameterization: each stage-k is centered about stage-k+1.

Improving Gibbs sampling computation

1. Variable transformation: centering the covariates around the grand mean.
2. Hierarchically centered parameterization: each stage-k is centered about stage-k+1.

   It is the parameterization that we have always set in this lecture.

Improving Gibbs sampling computation

1. Variable transformation: centering the covariates around the grand mean.
2. Hierarchically centered parameterization: each stage-k is centered about stage-k+1.

   It is the parameterization that we have always set in this lecture.

   Whereas random coefficients (/mixed/multilevel) models are typically specified by a classical statistician in an "horizontal" fashion.

Indice della sezione I

Indice della sezione II

Diagnostic strategy

- Run 3 or 5 parallel chains, with starting points drawn from a distribution believed to be overdispersed wrt the stationary distribution;

Diagnostic strategy

- Run 3 or 5 parallel chains, with starting points drawn from a distribution believed to be overdispersed wrt the stationary distribution;
- visually inspect these chains by overlaying them;

Diagnostic strategy

- Run 3 or 5 parallel chains, with starting points drawn from a distribution believed to be overdispersed wrt the stationary distribution;
- visually inspect these chains by overlaying them;
- Gelman-Rubin statistic;

Diagnostic strategy

- Run 3 or 5 parallel chains, with starting points drawn from a distribution believed to be overdispersed wrt the stationary distribution;
- visually inspect these chains by overlaying them;
- Gelman-Rubin statistic;
- check autocorrelation;

Diagnostic strategy

- Run 3 or 5 parallel chains, with starting points drawn from a distribution believed to be overdispersed wrt the stationary distribution;
- visually inspect these chains by overlaying them;
- Gelman-Rubin statistic;
- check autocorrelation;
- investigate crosscorrelation

CODA

One can summarize the Gibbs samples generated by BUGS using CODA: *Convergence Diagnostics and Output Analysis*

CODA

One can summarize the Gibbs samples generated by BUGS using CODA: *Convergence Diagnostics and Output Analysis*

It is avalaible for R the R-CODA version: Suite of S-functions to analyse output from BUGS and any other MCMC algorithm.

CODA

One can summarize the Gibbs samples generated by BUGS using CODA: *Convergence Diagnostics and Output Analysis*

It is avalaible for R the R-CODA version: Suite of S-functions to analyse output from BUGS and any other MCMC algorithm.

- Output Analysis:
  Plots, Printing graphical output from CODA, Multiple pages of plots, Displaying multiple graphics windows on screen,

CODA

One can summarize the Gibbs samples generated by BUGS using CODA: *Convergence Diagnostics and Output Analysis*

It is avalaible for R the R-CODA version: Suite of S-functions to analyse output from BUGS and any other MCMC algorithm.

- Output Analysis:
  Plots, Printing graphical output from CODA, Multiple pages of plots, Displaying multiple graphics windows on screen,
- Statistics

CODA

One can summarize the Gibbs samples generated by BUGS using CODA: *Convergence Diagnostics and Output Analysis*

It is avalaible for R the R-CODA version: Suite of S-functions to analyse output from BUGS and any other MCMC algorithm.

- Output Analysis:
  Plots, Printing graphical output from CODA, Multiple pages of plots, Displaying multiple graphics windows on screen,
- Statistics
- Convergence Diagnostics:
  Geweke Plotting, Geweke's diagnostic, Gelman & Rubin Plotting, Gelman & Rubin's diagnostic, Raftery & Lewis, Heidelberger and Welch,

CODA

One can summarize the Gibbs samples generated by BUGS using CODA: *Convergence Diagnostics and Output Analysis*

It is avalaible for R the R-CODA version: Suite of S-functions to analyse output from BUGS and any other MCMC algorithm.

- Output Analysis:
  Plots, Printing graphical output from CODA, Multiple pages of plots, Displaying multiple graphics windows on screen,
- Statistics
- Convergence Diagnostics:
  Geweke Plotting, Geweke's diagnostic, Gelman & Rubin Plotting, Gelman & Rubin's diagnostic, Raftery & Lewis, Heidelberger and Welch,
- Autocorrelations plots

CODA

One can summarize the Gibbs samples generated by BUGS using CODA: *Convergence Diagnostics and Output Analysis*

It is avalaible for R the R-CODA version: Suite of S-functions to analyse output from BUGS and any other MCMC algorithm.

- Output Analysis:
  Plots, Printing graphical output from CODA, Multiple pages of plots, Displaying multiple graphics windows on screen,
- Statistics
- Convergence Diagnostics:
  Geweke Plotting, Geweke's diagnostic, Gelman & Rubin Plotting, Gelman & Rubin's diagnostic, Raftery & Lewis, Heidelberger and Welch,
- Autocorrelations plots
- Cross-correlations plots

Indice della sezione I

Indice della sezione II

Alcuni riferimenti bibliografici

Alcuni riferimenti

Robert C. (2007) *The Bayesian Choice* 2nd ed. Springer ◂ def. di HBM

Alcuni riferimenti

Robert C. (2007) *The Bayesian Choice* 2nd ed. Springer ◂ def. di HBM

Gelman *et al.* (2004) *Bayesian Data Analysis* 2nd ed. London: Chapman & Hall ◂ esempio NN

Alcuni riferimenti

Robert C. (2007) *The Bayesian Choice* 2nd ed. Springer ◂ def. di HBM

Gelman *et al.* (2004) *Bayesian Data Analysis* 2nd ed. London: Chapman & Hall ◂ esempio NN

Carlin B. and Louis T. (2000) *Bayes and Empirical Bayes Methods for Data Analysis* 2nd ed. London: Chapman & Hall

Alcuni riferimenti

Robert C. (2007) *The Bayesian Choice* 2nd ed. Springer ◂ def. di HBM

Gelman *et al.* (2004) *Bayesian Data Analysis* 2nd ed. London: Chapman & Hall ◂ esempio NN

Carlin B. and Louis T. (2000) *Bayes and Empirical Bayes Methods for Data Analysis* 2nd ed. London: Chapman & Hall

Carlin B. and Louis T. (2008) *Bayesian Methods for Data Analysis* 3rd ed. Chapman & Hall/CRC Texts in Statistical Science

Alcuni riferimenti

Robert C. (2007) *The Bayesian Choice* 2nd ed. Springer ◂ def. di HBM

Gelman *et al.* (2004) *Bayesian Data Analysis* 2nd ed. London: Chapman & Hall ◂ esempio NN

Carlin B. and Louis T. (2000) *Bayes and Empirical Bayes Methods for Data Analysis* 2nd ed. London: Chapman & Hall

Carlin B. and Louis T. (2008) *Bayesian Methods for Data Analysis* 3rd ed. Chapman & Hall/CRC Texts in Statistical Science

Gelman A. and Hill J. (2007) *Data Analysis Using Regression and Multilevel/Hierarchical Models* Cambridge University Press

Alcuni riferimenti

Robert C. (2007) *The Bayesian Choice* 2nd ed. Springer ◂ def. di HBM

Gelman *et al.* (2004) *Bayesian Data Analysis* 2nd ed. London: Chapman & Hall ◂ esempio NN

Carlin B. and Louis T. (2000) *Bayes and Empirical Bayes Methods for Data Analysis* 2nd ed. London: Chapman & Hall

Carlin B. and Louis T. (2008) *Bayesian Methods for Data Analysis* 3rd ed. Chapman & Hall/CRC Texts in Statistical Science

Gelman A. and Hill J. (2007) *Data Analysis Using Regression and Multilevel/Hierarchical Models* Cambridge University Press

Gilks, Richardson, and Spiegelhalter (1995) *Markov Chain Monte Carlo in Practice*

Alcuni riferimenti

Robert C. (2007) *The Bayesian Choice* 2nd ed. Springer ◂ def. di HBM

Gelman *et al.* (2004) *Bayesian Data Analysis* 2nd ed. London: Chapman & Hall ◂ esempio NN

Carlin B. and Louis T. (2000) *Bayes and Empirical Bayes Methods for Data Analysis* 2nd ed. London: Chapman & Hall

Carlin B. and Louis T. (2008) *Bayesian Methods for Data Analysis* 3rd ed. Chapman & Hall/CRC Texts in Statistical Science

Gelman A. and Hill J. (2007) *Data Analysis Using Regression and Multilevel/Hierarchical Models* Cambridge University Press

Gilks, Richardson, and Spiegelhalter (1995) *Markov Chain Monte Carlo in Practice*

Congdon P. (2006) *Bayesian Statistical Modelling* 2nd ed. Wiley

Alcuni riferimenti

Robert C. (2007) *The Bayesian Choice* 2nd ed. Springer ◂ def. di HBM

Gelman *et al.* (2004) *Bayesian Data Analysis* 2nd ed. London: Chapman & Hall ◂ esempio NN

Carlin B. and Louis T. (2000) *Bayes and Empirical Bayes Methods for Data Analysis* 2nd ed. London: Chapman & Hall

Carlin B. and Louis T. (2008) *Bayesian Methods for Data Analysis* 3rd ed. Chapman & Hall/CRC Texts in Statistical Science

Gelman A. and Hill J. (2007) *Data Analysis Using Regression and Multilevel/Hierarchical Models* Cambridge University Press

Gilks, Richardson, and Spiegelhalter (1995) *Markov Chain Monte Carlo in Practice*

Congdon P. (2006) *Bayesian Statistical Modelling* 2nd ed. Wiley

Congdon P. (2003) *Applied Bayesian Modelling* Wiley

Alcuni riferimenti

Robert C. (2007) *The Bayesian Choice* 2nd ed. Springer ◂ def. di HBM

Gelman *et al.* (2004) *Bayesian Data Analysis* 2nd ed. London: Chapman & Hall ◂ esempio NN

Carlin B. and Louis T. (2000) *Bayes and Empirical Bayes Methods for Data Analysis* 2nd ed. London: Chapman & Hall

Carlin B. and Louis T. (2008) *Bayesian Methods for Data Analysis* 3rd ed. Chapman & Hall/CRC Texts in Statistical Science

Gelman A. and Hill J. (2007) *Data Analysis Using Regression and Multilevel/Hierarchical Models* Cambridge University Press

Gilks, Richardson, and Spiegelhalter (1995) *Markov Chain Monte Carlo in Practice*

Congdon P. (2006) *Bayesian Statistical Modelling* 2nd ed. Wiley

Congdon P. (2003) *Applied Bayesian Modelling* Wiley

Congdon P. (2005) *Bayesian Models for Categorical Data* Wiley