CAP. 6 – INFERENZA STATISTICA BAYESIANA

Introduzione

Nei capitoli precedenti è stata affrontata, in modo quasi esclusivo, la problematica dell'inferenza statistica parametrica, presupponendo, cioè, nota la forma analitica del modello rappresentativo del fenomeno o dei fenomeni oggetto d'analisi mentre non sono noti i parametri che li caratterizzano. I soli dati campionari sono stati utilizzati per pervenire ad una stima (puntuale o di intervallo) o per sottoporre a verifica empirica ipotesi riguardanti tali parametri.

Dopo aver fissato ragionevoli criteri di ottimalità, sono state analizzate le procedure e le condizioni che consentono il perseguimento dei risultati che soddisfano uno o più criteri tra quelli elencati. Sono stati dunque i ***parametri*** (***costanti non note***) l'oggetto specifico della trattazione usualmente indicata come ***inferenza statistica classica o frequentista*** secondo l'impostazione di ***Fisher*** e ***Neyman-Pearson***.

Questo capitolo è dedicato alla trattazione, seppure molto sommaria di un modo diverso di risoluzione dei problemi di inferenza induttiva: l'***approccio bayesiano all'inferenza statistica***, basato su una filosofia di analisi dei dati alternativa a quella propria dell'approccio classico. Nell'approccio classico i dati campionari sono l'unica fonte utilizzata ed utilizzabile per pervenire ad una ***conoscenza "oggettiva"***[1] della realtà rispetto alla quale non si presuppone alcuna conoscenza pregressa, mentre nell'approccio bayesiano una tale conoscenza si presuppone e i dati campionari servono solo per procedere al suo aggiornamento. Poiché, come più volte sottolineato, per facilitare la comprensione della realtà caratterizzata dalla variabilità presente nelle manifestazioni dei fenomeni di interesse, la realtà stessa viene rappresentata attraverso

---

[1] *Giuseppe* ***Pompilj*** (nel volume sulla teoria dei campioni 1961) scrive: "*..Cercherò di illustrare il significato e la portata delle formule di Bayes riportando alcuni brani di un mio articolo della rivista Archimede (**Pompilj**, 1951a).*
L'esperienza quotidiana ci pone continuamente di fronte a contrasti apparentemente paradossali perché in essi le parti invocano, a sostegno delle opposte tesi, gli stessi fatti, su cui perfettamente concordano.
Come mai, …., le parti concordano sui fatti (e talvolta anche nei minimi particolari di questi fatti) ed arrivano poi a conclusioni contrastanti? ……
…..Attraverso quale meccanismo ciascuno di noi si persuade di certe interpretazioni? Qual è di questa persuasione la componente soggettiva e quella oggettiva? Si tratta di problemi assai vecchi ; e non può certo soddisfare la spiegazione dogmatica degli antichi sofisti: l'uomo è la misura di tutte le cose ……
Nei Sei personaggi in cerca di autore quando il Capocomico interrompe la tirata della figliastra esclamando: veniamo al fatto; veniamo al fatto, signori miei! Queste sono discussioni – Il padre, il personaggio padre - interviene chiarendo:

Ecco, sissignore! Ma un fatto è come un sacco: vuoto non si regge: perché si regga, bisogna prima farci entrar dentro la ragione e i sentimenti che lo han determinato.

Questa battuta del padre contiene la vera essenza del problema testé delineato; perché una volta riconosciuto, secondo l'immagine pirandelliana, che un fatto è come un sacco, possiamo facilmente capire come a seconda di quel che ci si mette dentro potrà assumere un aspetto piuttosto che un altro. ….."
Sullo stesso argomento si può utilmente consultare ***Corrado Gini*** che, oltre ad essere stato precursore (***Gini***, 1911) di quello che viene usualmente definito come *approccio bayesiano empirico all'inferenza statistica* (***Chiandotto***, 1978), in due contributi (1939 e 1943) anticipa gran parte delle critiche rivolte alla teoria dei test di significatività (inferenza statistica classica) negli anni successivi dai sostenitori dell'approccio bayesiano.

opportuni modelli analitici (***modelli probabilistici***), anche per rappresentare la conoscenza pregressa si procede all'introduzione di specifici modelli che in questo caso però non sono rappresentativi della ***variabilità oggettiva*** insita nei dati, in quanto i parametri che caratterizzano i modelli sono delle costanti, ma rappresentano invece una ***variabilità virtuale*** che dipende dalla mancanza di conoscenza o dalla conoscenza parziale di cui si dispone.

Tecnicamente il problema si risolve considerando i parametri non più delle costanti incognite ma delle variabili casuali governate da una propria legge di ***distribuzione delle probabilità*** (***probabilità a priori***).

L'approccio bayesiano viene rifiutato da una componente molto rilevante della comunità scientifica che ritiene l'approccio stesso troppo condizionato da possibili preconcetti che poco hanno a che vedere con l'oggettività del processo scientifico, e ciò vale in particolare nei casi in cui si perviene alla formulazione della legge di distribuzione a priori rifacendosi alla definizione soggettiva della probabilità[2].

Quest'ultima considerazione evidenzia un fatto su cui vale la pena richiamare l'attenzione del lettore, e cioè sulla presunta oggettività dell'approccio classico alla problematica dell'inferenza induttiva che assegna ai soli dati campionari il compito di fornire informazioni sul fenomeno oggetto d'indagine: se si presuppone nota la forma analitica del modello rappresentativo della realtà, risulta ovvio che non sono solo i dati campionari a giocare un ruolo rilevante nel processo cognitivo, ma anche la conoscenza pregressa che suggerisce la forma del modello. Una conoscenza pregressa che potrebbe comunque essere fondata esclusivamente su dati campionari (dati oggettivi), ma allora si riproporrebbe il dilemma dell'esistenza di un a priori della conoscenza pregressa in un processo del quale non si intravede il motore primo.

La conoscenza pregressa del processo generatore dei dati è l'elemento che suggerisce il modello probabilistico rappresentativo della realtà cui fare riferimento nell'analisi, modello che determina anche le conclusioni cui si perviene, che possono essere molto diverse, anche se basate sugli stessi dati campionari, se diversi sono i processi che hanno generato i dati.

Esempio 6.1

Si supponga che in n lanci di una moneta la faccia testa si sia presentata k volte; l'evidenza empirica disponibile è, quindi, rappresentata da k successi in n prove indipendenti. Si tratta di una evidenza la cui rappresentazione attraverso un modello probabilistico dipende strettamente dal processo che l'ha generata; infatti, se il numero dei lanci è prefissato, il modello cui fare riferimento è la distribuzione binomiale; se invece il numero n dei lanci è il risultato di un processo che richiede di effettuare tanti lanci quanti ne occorrono per il conseguimento di k teste il modello da considerare è la distribuzione binomiale negativa. Ora, se con p si indica la probabilità di testa, in presenza di uno stesso risultato campionario k

[2] Al riguardo conviene, comunque, sottolineare che in letteratura si ritrovano numerosi contributi di autori che propongono la derivazioni di distribuzioni a priori "***oggettive"*** a partire dalla distribuzione ***a priori uniforme*** (a priori non informativa) proposta Laplace, a quelle proposte da ***Jeffreys,*** da ***Jaynes***, da ***Bernardo*** e da altri autori.

successi in n prove indipendenti le conclusioni cui si perviene sono diverse: nel primo caso (numero di lanci prefissato) la variabile casuale X ha distribuzione binomiale con funzione di massa di probabilità

$$f(x) = f(x;n,p) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{n}{x} p^x q^{n-x}$$

la cui media e varianza sono, rispettivamente $E(X) = n\,p$ *e* $Var(X) = n\,p\,q$*; mentre, nel secondo caso la variabile casuale X ha distribuzione binomiale negativa (numero di insuccessi prima di ottenere k successi) con funzione di massa di probabilità (III[a] versione)*

$$P(X = x) = f(x;k,p) = \binom{k+x-1}{x} p^k \cdot (1-p)^x = \binom{k+x-1}{x} p^k \cdot q^x$$

dove $n = k + x$,

la cui media e varianza sono, rispettivamente $E(X) = \frac{k\,q}{p}$ *e* $Var(X) = \frac{k\,q}{p^2}$.

La verosimiglianza per i due diversi processi generatori dei dati è:

$$\pi(p / X = 10) = \binom{15}{10} p^{10} (1-p)^{15-10} \quad e \quad \pi(p / X = 10) = \binom{10+5-1}{5} p^{10} \cdot (1-p)^5.$$

Come si può rilevare le due espressioni sono identiche a meno della costante di normalizzazione (permutazioni con ripetizione)

$$\binom{15}{10} \neq \binom{10+5-1}{5} = \binom{14}{5} = \binom{14}{9}.$$

Le stime di massima verosimiglianza del parametro p (probabilità di successo) sono molto diverse, rispettivamente, $p = \frac{k}{n} = \frac{10}{15} = 0,67$ *nel primo caso e* $p = \frac{k}{k+n} = \frac{10}{25} = 0,4$ *nel secondo caso.*

Diverse sono anche le conclusioni cui si perviene quando si procede alla verifica di ipotesi statistiche.

L'esempio sottolinea la rilevanza delle "***conoscenze a priori***" nel condizionare sia la scelta della procedura di analisi statistica dei dati sia le conclusioni che dalle analisi stesse derivano. L'interpretazione restrittiva e (a parere dell'autore di queste note) scorretta dell'*oggettività della scienza* che esclude dal processo scientifico ogni elemento di soggettività non può giustificare il rifiuto dell'approccio bayesiano se basato sull'impiego di probabilità soggettive. Probabilità che derivano dal bagaglio conoscitivo posseduto dal soggetto che è chiamato ad esprimerle e che lo caratterizzano; il problema non risiede tanto nell'uso delle conoscenze a priori quanto nella natura e nel corretto impiego delle stesse; la natura dipende dalla "*caratteristiche"* del soggetto e un corretto impiego è rappresentato dalla formula di Bayes.

In letteratura sono stati proposti numerosi altri approcci all'inferenza statistica, oltre a quello classico (*frequentista*) e quello bayesiano (*soggettivista*), tra i più rilevanti si segnalano l'approccio[3]:

- **Fiduciale *(Fisher, 1930, 1935 e 1956)***
- **Della verosimiglianza *(Barnard, 1949, 1985; Birnbaum, 1962; Edwards, 1972; Azzalini, 1996 e Royall, 1997)***
- **Della plausibilità *(Barndorff-Nielsen, 1976)***
- **Strutturale *(Fraser, 1968)***
- **Pivotale *(Barnard, 1949, 1985)***
- **Prequentiale *(Dawid, 1984, 1997 e 2000)***
- **Predittivo *(Geisser, 1993)***
- **Bayesiano/verosimiglianza integrato *(Aitkin, 2010)***

6.1 La formula di Bayes

Nei capitoli precedenti sono stati illustrati i metodi che consentono la derivazione di risultati che soddisfano a certi criteri di ottimalità predefiniti per la risoluzione di problemi di stima (puntuale e di intervallo) o di verifica di ipotesi statistiche relative ai parametri (uno o più costanti non note) presupponendo la conoscenza della funzione di massa o di densità di probabilità della v.c. X

$$X \sim f\left(x;\theta_1,\theta_2,...,\theta_\eta\right) = f\left(x;\boldsymbol{\theta}\right)$$

e la disponibilità di un campione casuale semplice di osservazioni sulla v.c. X

$$\boldsymbol{X}' = \left(X_1, X_2, ..., X_i, ..., X_n\right)$$

con funzione di massa o di densità di probabilità

$$f\left(x_1, x_2, ..., x_n; \theta_1, \theta_2, ..., \theta_\eta\right) = f\left(\boldsymbol{x};\boldsymbol{\theta}\right) = \prod_{i=1}^{n} f\left(x_i;\boldsymbol{\theta}\right)$$

dove

$$f\left(x_i;\boldsymbol{\theta}\right) \equiv f\left(x;\boldsymbol{\theta}\right).$$

Nel contesto dell'inferenza statistica classica, un ruolo particolarmente rilevante è svolto dalla funzione di verosimiglianza. Al riguardo basta ricordare quanto detto a proposito del metodo di stima della massima verosimiglianza e del test del rapporto di massima verosimiglianza.

Se si osserva l'espressione analitica della funzione di massa o di densità di probabilità del campione e della funzione di verosimiglianza

$$\Rightarrow \text{funzione di verosimiglianza } \Rightarrow L\left(\boldsymbol{\theta}\right) = L\left(\boldsymbol{\theta};\boldsymbol{x}\right) = f\left(\boldsymbol{\theta}/\boldsymbol{x}\right) = \prod_{i=1}^{n} f\left(\boldsymbol{\theta};x_i\right)$$

$$\Rightarrow \text{funzione di massa o densità di probabilità } \Rightarrow f\left(\boldsymbol{x};\boldsymbol{\theta}\right) = f\left(\boldsymbol{x}/\boldsymbol{\theta}\right) = \prod_{i=1}^{n} f\left(x_i;\boldsymbol{\theta}\right)$$

[3] Sull'argomento si può consultare ***Barnett** (1999)*.

si rileva immediatamente come ad una apparente uguaglianza formale corrisponde una rilevante differenza sostanziale; infatti, si tratta di due probabilità condizionate, nel primo caso, della variabile $\boldsymbol{\theta}$ dato uno specifico risultato campionario $\left[L(\boldsymbol{\theta}) = f(\boldsymbol{\theta} / \boldsymbol{X} = \boldsymbol{x})\right]$, nel secondo caso della variabile $\boldsymbol{X}$ dato uno specifico valore di $\boldsymbol{\theta}$.

In altri termini, le due funzioni, di verosimiglianza e di probabilità (massa o densità), sono formalmente del tutto equivalenti ma è completamente diversa la loro interpretazione. Nel caso della funzione di verosimiglianza l'argomento è la variabile θ o il vettore di variabili $\boldsymbol{\theta}$ una volta acquisita l'informazione campionaria $\boldsymbol{X}$ che rappresenta l'elemento condizionante, mentre nella funzione di massa o di densità di probabilità è il vettore casuale delle osservazioni campionare $\boldsymbol{X}$ la cui distribuzione dipende dai valori assunti dal/i parametro/i $\theta / \boldsymbol{\theta}$.

Per risolvere i problemi inferenziali si è fatto riferimento, a seconda della tipologia di problema, a specifiche variabili casuali, verificandone il comportamento nell'universo di tutti i possibili campioni estraibili dalla popolazione rappresentata dal modello $f(x;\boldsymbol{\theta}) = f(x/\boldsymbol{\theta})$; in particolare, sono state considerate le funzioni degli elementi campionari:

- la v.c. ***stimatore*** $\Theta_i = T_i(X_1, X_2, ..., X_n) = T_i(\boldsymbol{X})$ per $i = 1, 2, ..., \eta$
- la v.c. ***elemento pivotale*** $Y_i = T_i(\boldsymbol{X}; \theta_i) = T_i(\boldsymbol{X} / \theta_i)$ per $1 = 1, 2, ..., \eta$
- la ***v.c. tes***t $V_i = T_i(\boldsymbol{X}; \theta_i) = T_i(\boldsymbol{X} / \theta_i)$ per $i = 1, 2, ..., \eta$.

Nota la legge di distribuzione nell'universo dei campioni delle variabili sopra elencate è possibile risolvere i problemi inferenziali verificando il soddisfacimento dei criteri di ottimalità predefiniti. Al riguardo si segnala che, nella generalità dei casi, quando il modello è caratterizzato da più parametri ma solo alcuni sono di interesse occorre intervenire sui cosi detti ***parametri di disturbo***, cioè sui parametri ai quali non si è interessati ma che sono presenti quali elementi caratterizzanti la distribuzione campionaria delle tre variabili sopra elencate e che spesso non consentono il perseguimento dell'obiettivo prefissato. In tali circostanze, se non si riesce ad ottenere i risultati d'interesse, qualunque sia il valore assunto dal/dai parametro/i di disturbo si procede sostituendo al/i valore/i incognito/i del parametro/i una sua/loro stima. Operazione quest'ultima non sempre consente il perseguimento dell'obiettivo desiderato.

Nelle pagine seguenti si avrà modo di evidenziare come il problema della presenza di parametri di disturbo trovi una immediata e soddisfacente soluzione nel contesto bayesiano. Inoltre, in tale contesto è possibile affrontare e risolvere in modo soddisfacente anche il problema della scelta della forma analitica del modello quale rappresentazione semplificata della realtà.

Nell'approccio bayesiano non si fa più riferimento ad un modello probabilistico $f(x;\boldsymbol{\theta}) = f(x/\boldsymbol{\theta})$ rappresentativo del fenomeno d'interesse noto a meno del valore

assunto dal/i parametro/i che lo caratterizzano ed individuano lo specifico modello quale/i elemento/i condizionante/i, si fa invece riferimento ad una distribuzione congiunta (di massa o di densità di probabilità)

$$f\left(x,\boldsymbol{\theta}\right)$$

Entrambi gli argomenti della funzione x e $\boldsymbol{\theta}$ hanno natura di variabili casuali, la prima dovuta alla naturale variabilità del fenomeno indagato (*variabilità aleatoria*) la seconda dovuta alla mancata conoscenza del suo valore numerico (*variabilità virtuale o epistemica*).

Riprendendo quanto detto a proposito delle probabilità condizionate di eventi valgono le uguaglianze

$$f\left(x,\boldsymbol{\theta}\right)=f\left(x/\boldsymbol{\theta}\right)\cdot\pi\left(\boldsymbol{\theta}\right)$$
$$f\left(x,\boldsymbol{\theta}\right)=\pi\left(\boldsymbol{\theta}/x\right)\cdot f\left(x\right)$$

dove $\pi\left(\boldsymbol{\theta}\right)$ rappresenta la forma analitica del modello rappresentativo del vettore casuale $\boldsymbol{\theta}$. Dalle due relazioni di uguaglianza si deriva l'espressione analitica della formula di Bayes

$$\pi\left(\boldsymbol{\theta}/x\right)=\frac{f\left(x/\boldsymbol{\theta}\right)\cdot\pi\left(\boldsymbol{\theta}\right)}{f\left(x\right)}=\frac{f\left(x/\boldsymbol{\theta}\right)\cdot\pi\left(\boldsymbol{\theta}\right)}{\int_{\boldsymbol{\theta}}f\left(x/\boldsymbol{\theta}\right)\cdot\pi\left(\boldsymbol{\theta}\right)\,d\left(\boldsymbol{\theta}\right)}$$

dove è stato ipotizzato un spazio di variabilità dei parametri continuo.

Se anziché fare riferimento alla variabile X si considera il vettore casuale campionario $\boldsymbol{X'}=\left(X_1,X_2,...,X_i,...,X_n\right)$ la formula di Bayes diventa

$$\pi\left(\boldsymbol{\theta}/\boldsymbol{x}\right)=\frac{f\left(\boldsymbol{x}/\boldsymbol{\theta}\right)\cdot\pi\left(\boldsymbol{\theta}\right)}{f\left(\boldsymbol{x}\right)}=\frac{f\left(\boldsymbol{x}/\boldsymbol{\theta}\right)\cdot\pi\left(\boldsymbol{\theta}\right)}{\int_{\boldsymbol{\theta}}f\left(\boldsymbol{x}/\boldsymbol{\theta}\right)\cdot\pi\left(\boldsymbol{\theta}\right)\,d\left(\boldsymbol{\theta}\right)}=$$
$$=\frac{L\left(\theta\right)\cdot\pi\left(\boldsymbol{\theta}\right)}{f\left(\boldsymbol{x}\right)}\propto L\left(\boldsymbol{\theta}\right)\cdot\pi\left(\boldsymbol{\theta}\right)$$

dove

$$f\left(\boldsymbol{x}\right)=\int_{\boldsymbol{\theta}}f\left(\boldsymbol{x}/\boldsymbol{\theta}\right)\cdot\pi\left(\boldsymbol{\theta}\right)\,d\left(\boldsymbol{\theta}\right)$$

definisce la distribuzione marginale di $\boldsymbol{X'}=\left(X_1,X_2,...,X_i,...,X_n\right)$, usualmente detta ***distribuzione predittiva a priori di*** $\boldsymbol{X}$, che rappresenta la ***costante di normalizzazione della distribuzione a posteriori di*** $\boldsymbol{\theta}$, il simbolo $\propto$ sta ad indicare la relazione di proporzionalità tra le due quantità poste a confronto, mentre $\left[L\left(\boldsymbol{\theta}\right)\cdot\pi\left(\boldsymbol{\theta}\right)\right]$ rappresenta ***nucleo*** (***kernel*** in inglese) ***della distribuzione a posteriori***[4].

[4] Il **nucleo** di una funzione di massa o di densità di probabilità è dato dalla rappresentazione analitica della stessa funzione dopo aver omesso tutti i termini che non sono funzioni della variabile casuale di riferimento, ad esempio alla funzione di densità di probabilità della v.c. normale

$$f\left(x,\mu/\sigma^2\right)=\frac{1}{\left(2\pi\sigma^2\right)^{n/2}}\cdot e^{-\frac{1}{2\sigma^2}\left(x-\mu\right)^2}$$

A fronte della distribuzione predittiva a priori si colloca la ***distribuzione predittiva a posteriori***

$$f\left(\tilde{\boldsymbol{x}}/\boldsymbol{x}\right)=\int_{\boldsymbol{\theta}} f\left(\tilde{\boldsymbol{x}}/\boldsymbol{\theta},\boldsymbol{x}\right)\cdot\pi\left(\boldsymbol{\theta}/\boldsymbol{x}\right)\,d\left(\boldsymbol{\theta}\right)$$

che fa riferimento ad un nuovo campione di possibili osservazioni $\tilde{\boldsymbol{X}}$ avendo già osservato *n* manifestazioni dello stesso fenomeno $\boldsymbol{X}=\boldsymbol{x}$.

Le funzioni sopra introdotte hanno la seguente interpretazione probabilistica

$f\left(\boldsymbol{x}/\boldsymbol{\theta}\right)$ $\Rightarrow$ probabilità condizionata del campione

$L\left(\boldsymbol{\theta}\right)=f\left(\boldsymbol{\theta};\boldsymbol{x}\right)$ $\Rightarrow$ verosimiglianza (*che non deve essere interpretata come distribuzione di probabilità*)

$\pi\left(\boldsymbol{\theta}\right)$ $\Rightarrow$ probabilità a priori del parametro/i

$\pi\left(\boldsymbol{\theta}/\boldsymbol{x}\right)$ $\Rightarrow$ probabilità a posteriori del parametro/i

$f\left(\boldsymbol{x}\right)$ $\Rightarrow$ probabilità predittiva a priori

$f\left(\tilde{\boldsymbol{x}}/\boldsymbol{x}\right)$ $\Rightarrow$ probabilità predittiva a posteriori

dove la probabilità va intesa come funzione di densità di probabilità nel caso continuo e come funzione di massa di probabilità nel caso discreto.

Le ragioni principali che hanno frenato lo sviluppo e l'impiego della teoria e dei metodi propri dell'inferenza statistica bayesiana sono da ricercare soprattutto i due problemi presenti nella formula di bayes. Il primo è rappresentato dal già segnalato rifiuto da parte di molti autori del modo soggettivo con cui si perviene alla misura della probabilità a priori $\pi\left(\boldsymbol{\theta}\right)$, anche a prescindere dalle difficoltà di traduzione, a volte molto rilevanti, delle conoscenze a priori in distribuzioni di probabilità significative. Il secondo problema risiede, invece, nella difficoltà di derivazione in forma chiusa (analiticamente) dell'espressione

$$f\left(\boldsymbol{x}\right)=\int_{\boldsymbol{\theta}} f\left(\boldsymbol{x}/\boldsymbol{\theta}\right)\cdot\pi\left(\boldsymbol{\theta}\right)\,d\left(\boldsymbol{\theta}\right).$$

Ad entrambi i problemi sono state proposte delle soluzioni che non sono però condivise dall'intera comunità scientifica soprattutto per ciò che concerne il problema della scelta della distribuzione a priori.

Una delle proposte di rilevanza non marginale, e che offre una soluzione relativamente soddisfacente ad entrambi i problemi, è rappresentata dall'impiego delle ***distribuzioni a priori coniugate*** introdotte nel paragrafo 14 del primo capitolo. Infatti, tale scelta, fornendo direttamente l'espressione analitica della distribuzione a posteriori, oltre a risultare ragionevole in molti contesti di ricerca non richiede il computo della distribuzione marginale $f\left(\boldsymbol{x}\right)$. In realtà, operativamente, il passaggio dalla

---

è associato il nucleo $e^{-2\frac{1}{2\sigma^2}(x-\mu)^2}$ che consente di scrivere

$$f\left(x,\mu/\sigma^2\right)\propto e^{-\frac{1}{2\sigma^2}(x-\mu)^2}.$$

distribuzione a priori coniugata alla distribuzione a posteriori avviene facendo riferimento al nucleo della distribuzione: dal ***nucleo della distribuzione a priori*** coniugata si passa al ***nucleo della distribuzione a posteriori*** alla cui espressione completa si perviene attraverso la facile derivazione della costante di normalizzazione.

Il mancato ricorso all'impiego delle distribuzioni a priori coniugate richiede necessariamente il computo dell'espressione a denominatore della formula di Bayes che, come già sottolineato, solo in rare occasioni può essere ottenuta per via analitica.

Le stesse difficoltà di derivazione analitica si presentano quando si vuol procedere ad una sintesi della distribuzione a posteriori attraverso il computo di indici caratteristici (ad esempio i momenti della v.c. θ). Se si considera una generica funzione $g(\theta)$ si deve procedere, cioè, al computo della relazione

$$E\left[g(\theta)\right] = \int_{\theta} g(\theta)\pi(\theta / \boldsymbol{x})\ d(\theta) = \frac{\int_{\theta} g(\theta)\, f(\boldsymbol{x}/\theta)\ \cdot \pi(\theta)\ d(\theta)}{\int_{\theta} f(\boldsymbol{x}/\theta)\ \cdot \pi(\theta)\ d(\theta)}$$

dove le difficoltà di derivazione analitica riguardano entrambi gli integrali, quello a denominatore e quello a numeratore dell'espressione.

I metodi classici di integrazione numerica, a ragione della complessità dei problemi, nella generalità dei casi, non portavano a soluzioni soddisfacenti cui si è invece pervenuti attraverso il ricorso ai cosidetti metodi Montecarlo (***M**arkov **C**hain **M**onte **C**arlo* **- *MCMC***).

All'introduzione dei metodi *MCMC* in ambito statistico si deve sostanzialmente attribuire l'enorme sviluppo, sia nel contesto teorico che in quello applicativo, dell'inferenza Bayesiana.

Il principio si cui sono basati i metodi *MCMC* è relativamente semplice, si tratta di effettuare operazioni ripetute di campionamento casuale da una popolazione di riferimento fino a pervenire ad una approssimazione della distribuzione desiderata attraverso l'impiego delle catene di Markov ricorrendo a specifici algoritmi proposti in letteratura. Tra i più noti e di più largo impiego si segnalano l'algoritmo di *Metropolis-Hastings*, il *Gibbs sampler*, lo *slice sampling* e il *perfect sampling*; al riguardo si segnala, in particolare, il software gratuito *WinBUGS*[5].

Le difficoltà di traduzione delle informazioni a disposizione in distribuzioni di probabilità a priori e, soprattutto, il rifiuto delle stesse in quanto caratterizzate da elevata soggettività (preconcetti) associate alla constatazione che in molte situazioni di ricerca non si ritiene sufficiente, o del tutto assente, il bagaglio informativo disponibile a priori, hanno suggerito l'introduzione delle cosidette *distribuzioni a priori oggettive*[6].

Al paradigma bayesiano fanno, pertanto, riferimento almeno due scuole di pensiero: da un lato si collocano i sostenitori della scelta soggettiva della probabilità a priori

[5] Si tratta di un software molto flessibile prodotto nell'ambito del progetto ***B**ayesian inference **U**sing **G**ibbs **S**ampling* (***BUGS***) che consente l'analisi bayesiana di modelli statistici complessi attraverso l'impiego di metodi ***M**arkov **C**hain **M**onte **C**arlo* (***MCMC***). Il progetto avviato nel 1989 dall'Unità Biostatistica MRC di Cambridge è stato successivamente sviluppato da questa Unità in collaborazione con l'*Imperial College School of Medicine* di Londra.

[6] Altri termini utilizzati per qualificare tali distribuzioni sono: *non informative, di default, convenzionali, di riferimento, non soggettive.*

(*bayesiani soggettivisti),* e in questa categoria si colloca anche l'autore di queste note, dall'altro lato si collocano coloro che ritengono sia possibile pervenire ad una misura "oggettiva" delle probabilità a priori (*bayesiani oggettivisti*). Nell'ambito di questa seconda categoria vanno però distinti almeno 2 diversi filoni:

- i bayesiani empirici che ritengono giustificato l'impiego del metodo bayesiano solo quando si dispone di un'evidenza empirica a priori adeguata[7];
- coloro che suggeriscono l'impiego di distribuzioni a priori usualmente, ma impropriamente, dette non informative[8].

In questa sede non si procederà all'approfondimento dell'argomento[9] limitando l'esposizione a brevi considerazioni su alcune tipologie di distribuzioni a priori e alla presentazione di alcuni esempi di derivazione della distribuzione a posteriori in dipendenza di una scelta acritica[10] della distribuzioni a priori. Verranno illustrati esempi di derivazione della distribuzione a posteriori per alcune variabili casuali considerando le distribuzioni a priori coniugate e alcune distribuzioni a priori non informative. In particolare, in questa sede non si procederà all'approfondimento dell'argomento[11] limitando l'esposizione a brevi considerazioni sulle alcune specifiche tipologie di distribuzioni a priori e alla presentazione di alcuni esempi di derivazione della distribuzione a posteriori in dipendenza di una scelta acritica[12] della distribuzioni a priori.

In particolare, non verrà trattato il tema della elicitazione delle probabilità a priori (derivazione soggettiva) che è del tutto simile a quello della elicitazione delle funzioni di utilità. Il lettore interessato ad un un'approfondimento sulla derivazione soggettiva delle distribuzioni di probabilità a priori può, tra gli altri, consultare i contributi di Jenkinson (2005), e quello di Garthwaite, Kadane e O'Hagan (2005).

Prima di procedere nelle esemplificazioni risulta conveniente anticipare alcuni concetti che verranno ripresi e meglio precisati nelle pagine successive.

Nell'introdurre il concetto di probabilità a priori è stata utilizzata la generica espressione $\pi(\boldsymbol{\theta})$, si tratta ovviamente di una rappresentazione che necessita di ulteriori elementi caratterizzanti. Trattandosi di una distribuzione di massa o di densità di probabilità, l'espressione analitica sarà generalmente caratterizzata da uno o più

[7] L'impiego del termine bayesino empirico qui utilizzato non corrisponde a quello impiegato nella letteratura corrente che prevede l'impiego dell'evidenza empirica corrente per inferire sia sulla verosimiglianza sia sulla distribuzione a priori. Alcuni autori ritengono che quest'ultima procedura non rispetti la filosofia base del ragionamento bayesiano che presuppone l'impiego di informazioni a priori.

[8] Uno dei più autorevoli sostenitori dell'approccio bayesiano oggettivo ***Bernardo*** (1997) al riguardo dichiara: "*Non-informative priors do not exist*". A dialogue with José M. Bernardo".

[9] Il lettore interessato può utilmente consultare, tra gli altri, i contributi di ***Berger*** (2006) e di ***Goldstein*** (2006). Al riguardo particolarmente interessanti sono anche i lavori di ***Joyce*** (2009) e quello di ***Robert*** *e al.,* (2009).

[10] Il lettore interessato al tema può utilmente consultare i contributi di ***Kass*** *e* ***Wasserman*** (1996) e quello di ***Berger***, ***Bernardo*** *e* ***Sun*** (2009). Per un'approfondimento sulla derivazione soggettiva delle distribuzioni di probabilità a priori si può, tra gli altri, consultare il lavoro di ***Jenkinson*** (2005).

[11] Il lettore interessato può utilmente consultare, tra gli altri, il contributo di ***Berger*** "(2006) e quello di ***Goldstein*** (2006). Al riguardo particolarmente interessanti sono anche i lavori di ***Joyce*** (2009) e quello di ***Robert*** *e al.,* (2009).

[12] Il lettore interessato al tema può utilmente consultare i contributi di ***Kass*** *e* ***Wasserman*** (1996) e quello di ***Berger***, ***Bernardo*** *e* ***Sun*** (2009).

parametri $\boldsymbol{\delta}' = (\delta_1, \delta_2, ..., \delta_s)$ usualmente detti *iperparametri*; pertanto, per esplicitare tale dipendenza si deve utilizzare la forma $\pi(\boldsymbol{\theta} / \boldsymbol{\delta})$ per rappresentare la probabilità a priori, mentre l'espressione della probabilità a posteriori diventa

$$\pi(\boldsymbol{\theta} / \boldsymbol{x}, \boldsymbol{\delta}) = \frac{f(\boldsymbol{x} / \boldsymbol{\theta}) \cdot \pi(\boldsymbol{\theta}, \boldsymbol{\delta})}{f(\boldsymbol{x})} = \frac{f(\boldsymbol{x} / \boldsymbol{\theta}) \cdot \pi(\boldsymbol{\theta}, \boldsymbol{\delta})}{\int_{\underline{\theta}} f(\boldsymbol{x} / \boldsymbol{\theta}) \cdot \pi(\boldsymbol{\theta}, \boldsymbol{\delta}) \, d(\boldsymbol{\theta})} =$$

$$= \frac{L(\boldsymbol{\theta}) \cdot \pi(\boldsymbol{\theta}, \boldsymbol{\delta})}{f(\boldsymbol{x})} \propto L(\boldsymbol{\theta}) \cdot \pi(\boldsymbol{\theta}, \boldsymbol{\delta}).$$

di conseguenza, le distribuzioni predittive a priori e a posteriori assumono la forma

$$f(\boldsymbol{x} / \boldsymbol{\delta}) = \int_{\boldsymbol{\theta}} f(\boldsymbol{x} / \boldsymbol{\theta}) \cdot \pi(\boldsymbol{\theta}, \boldsymbol{\delta}) \, d(\boldsymbol{\theta}) \text{ e } f(\tilde{\boldsymbol{x}} / \boldsymbol{x}, \boldsymbol{\delta}) = \int_{\boldsymbol{\theta}} f(\tilde{\boldsymbol{x}} / \boldsymbol{\theta}) \cdot \pi(\boldsymbol{\theta} / \boldsymbol{x}, \boldsymbol{\delta}) \, d(\boldsymbol{\theta}).$$

6.2 - Distribuzioni a priori coniugate

Si riporta la definizione di distribuzione coniugata introdotta nel paragrafo 14 del I° capitolo di queste Note: "***Quando la distribuzione di probabilità a posteriori appartiene alla stessa famiglia della distribuzione a priori, quest'ultima viene detta distribuzione di probabilità coniugata***".

Come si avrà modo di verificare scorrendo gli esempi di seguito riportati, il ricorso alle distribuzioni a priori coniugate presenta notevoli vantaggi; infatti, si tratta spesso di distribuzioni molto flessibili che proprio per questa loro caratteristica si rivelano adeguate in molte situazioni di ricerca. Comunque, al fine di evitare errate conclusioni, il ricorso ad una tale tipologia di distribuzioni non deve essere acritico ma deve essere limitato ai soli casi in cui si possiede un adeguato patrimonio informativo a priori che ne giustifichi l'impiego.

Esempio 6.2 (distribuzione di Bernoulli)

Per la distribuzione di Bernoulli $f(x, p) = p^x (1-p)^{1-x}$ per $x: 0,1$ *, la v.c. Beta*

$$\pi(p; \alpha, \beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{\int_0^1 p^{\alpha-1}(1-p)^{\beta-1} dp} = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha, \beta)} = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta) / \Gamma(\alpha+\beta)}$$

è distribuzione coniugata a priori, infatti

$$\pi(p / x) = \frac{L(p; x) \cdot \pi(p)}{f(x)} = \frac{p^x (1-p)^{1-x}}{f(x)} \frac{p^{\alpha-1}(1-p)^{\beta-1}}{\int_0^1 p^{\alpha-1}(1-p)^{\beta-1} dp} \propto$$

$$\propto p^{(x+\alpha)-1}(1-p)^{(1-x+\beta)-1} = p^{\alpha^*-1}(1-p)^{\beta^*-1} \quad \text{dove } (x+\alpha) = \alpha^*, (1-x+\beta) = \beta^*$$

che è una distribuzione Beta con parametri α^* e β^*. La costante di normalizzazione è quindi espressa da

$$f(x) = B(\alpha^*, \beta^*) = \Gamma(\alpha^*)\Gamma(\beta^*) / \Gamma(\alpha^* + \beta^*).$$

Esempio 6.3 (distribuzione Binomiale)
La v.c Beta è anche distribuzione coniugata a priori della distribuzione Binomiale, infatti, riprendendo la funzione di massa di probabilità della distribuzione binomiale

$$F(x,n,p) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{n}{x} p^x q^{n-x}$$

si ha

$$\pi(p/x=m) = \frac{L(p;x=m)\cdot\pi(p)}{f(x)} = \left[\binom{n}{m} p^m (1-p)^{n-m} / f(x)\right] \frac{p^{\alpha-1}(1-p)^{\beta-1}}{\int_0^1 p^{\alpha-1}(1-p)^{\beta-1}\,dp} =$$

$$= \frac{\frac{(\alpha+\beta-1)!}{(\alpha-1)!\ (\beta-1)!}\ p^{\alpha-1}\ (1-p)^{\beta-1}\ \ \frac{n!}{x!\ (n-x)!}\ p^x\ (1-p)^{n-x}}{\int_0^1 \frac{(\alpha+\beta-1)!}{(\alpha-1)!\ (\beta-1)!}\ p^{\alpha-1}\ (1-p)^{\beta-1}\ \ \frac{n!}{x!\ (n-x)!}\ p^x\ (1-p)^{n-x}\ dp} =$$

$$= \frac{(x+\alpha+n-x+\beta-1)!}{(x+\alpha-1)!\ (n-x+\beta-1)!}\ p^{x+\alpha-1}\ (1-p)^{n-x+\beta-1} =$$

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

$$\propto p^{(m+\alpha)-1}(1-p)^{(1-m+\beta)-1} = p^{\alpha^*-1}(1-p)^{\beta^*-1} \quad \text{dove } \alpha^* = (m+\alpha),\ \beta^* = (n-m+\beta)$$

che è una distribuzione Beta con parametri $\alpha^* = (m+\alpha)$ e $\beta^* = (n-m+\beta)$. La costante di normalizzazione è quindi espressa da

$$f(x) = B(\alpha^*, \beta^*) = \Gamma(\alpha^*)\Gamma(\beta^*) / \Gamma(\alpha^* + \beta^*).$$

Esempio 6.4 (distribuzione Multinomiale)
Come già sottolineato nel Cap. 1 la v.c di Dirichlet

$$\pi(\boldsymbol{p}) = \pi(\boldsymbol{p};\boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{i=1}^{k+1}\alpha_i\right)}{\prod_{i=1}^{k+1}\Gamma(\alpha_i)} \prod_{i=1}^{k+1} p_i^{\alpha_i-1} \propto \prod_{i=1}^{k+1} p_i^{\alpha_i-1}$$

è distribuzione a priori coniugata della v.c. multinomiale

$$f(\boldsymbol{x},\boldsymbol{p}) = \frac{n!}{x_1!\,x_2!\ldots x_k!\left(n-\sum_{i=1}^{k} x_i\right)!}\ p_1^{x_1}\cdot p_2^{x_2}\ldots p_k^{x_k}\ q^{n-\sum_{i=1}^{k} x_i}$$

infatti

$$\pi(\boldsymbol{p}/\boldsymbol{x}=\boldsymbol{m}) = \propto \prod_{i=1}^{k+1} p_i^{\alpha_i-1}\cdot\prod_{i=1}^{k+1} p_i^{m_i} = \prod_{i=1}^{k+1} p_i^{(\alpha_i+m_i)-1} = \prod_{i=1}^{k+1} p_i^{\alpha_i^*-1}$$

che è una distribuzione di Dirichlet con parametri $\alpha_i^* = \alpha_i + m_i$. La costante di normalizzazione è quindi espressa da

$$\frac{\Gamma\left(\sum_{i=1}^{k+1}\alpha_i^*\right)}{\prod_{i=1}^{k+1}\Gamma\left(\alpha_i^*\right)}.$$

Esempio 6.5 (distribuzione di Poisson)
La funzione di verosimiglianza della v.c. di Poisson è

$$L(\lambda;\boldsymbol{x}) = \prod_{i=1}^{n}\frac{e^{-\lambda}\cdot\lambda^{x_i}}{x_i!} \propto\ e^{-n\lambda}\cdot\lambda^{\sum_{i=1}^{n}x_i}$$

La v.c. Gamma

$$\pi(\lambda) = \frac{\lambda^{\alpha-1}e^{-\frac{\lambda}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}} \propto\ \lambda^{\alpha-1}e^{-\frac{\lambda}{\beta}}$$

è distribuzione a priori coniugata della v.c. di Poisson, infatti

$$\pi(\gamma/\boldsymbol{x}) = \frac{L(\lambda;\boldsymbol{x})\cdot\pi(\lambda)}{f(\boldsymbol{x})} = \frac{\prod_{i=1}^{n}\frac{e^{-n\lambda}\cdot\lambda^{x_i}}{x_i!}\cdot\frac{\lambda^{\alpha-1}e^{-\frac{\lambda}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}}}{f(\boldsymbol{x})} \propto \lambda^{\sum_{i=1}^{n}x_i+\alpha-1}\cdot e^{-\left(n\lambda+\frac{\lambda}{\beta}\right)} =$$

$$= \lambda^{\alpha^*-1}\cdot e^{-\frac{\lambda}{\beta^*}} = p^{\alpha^*-1}(1-p)^{\beta^*-1} \quad \text{dove} \sum_{i=1}^{n}x_i+\alpha=\alpha^*, \frac{\beta}{\beta\ n+\lambda}=\beta^*$$

che è una distribuzione Gamma con parametri $\alpha^* = \sum_{i=1}^{n}x_i + \alpha$ e $\beta^* = \frac{\beta}{\beta\ n+\lambda}$. La costante di normalizzazione è

$$f(x) = 1/\Gamma(\alpha)\beta^{\alpha}.$$

Esempio 6.6 (distribuzione Normale)
La funzione di verosimiglianza della v.c. Normale è

$$L\left(\mu,\sigma^2;\boldsymbol{x}\right) = \frac{1}{\left(2\pi\sigma^2\right)^{n/2}}\cdot e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} \propto\ e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} \propto e^{-\frac{n}{2\sigma^2}\left(-2\mu\bar{x}+\mu^2\right)} \propto\ e^{-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2}$$

Per μ ***nota****, la v.c.Gamma inversa*[13]

$$\pi(\sigma^2 / \mu;\alpha,\beta) = \frac{\left(\sigma^2\right)^{-\alpha-1}\beta^{\alpha}\ e^{-\frac{\beta}{\sigma^2}}}{\Gamma(\alpha)} \propto \left(\sigma^2\right)^{-\alpha-1} e^{-\frac{\beta}{\sigma^2}}$$

è distribuzione a priori coniugata della v.c. Gamma inversa, infatti dalla verosimiglianza

$$L\left(\mu,\sigma^2 / \boldsymbol{x}\right) = \frac{1}{\left(2\pi\sigma^2\right)^{n/2}} \cdot e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

si ha

$$\pi\left(\sigma^2 / \mu, \boldsymbol{x};\alpha,\beta\right) = \frac{L\left(\sigma^2 / \mu;\boldsymbol{x}\right)\cdot\pi\left(\sigma^2\right)}{f\left(\boldsymbol{x}\right)} = \frac{\frac{1}{\left(2\pi\sigma^2\right)^{n/2}} \cdot e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} \cdot \frac{\left(\sigma^2\right)^{-\alpha-1}\beta^{\alpha}\ e^{-\frac{\beta}{\sigma^2}}}{\Gamma(\alpha)}}{f\left(\boldsymbol{x}\right)} \propto$$

$$\propto e^{-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2} \left(\sigma^2\right)^{-\left(\alpha+\frac{n}{2}\right)-1} e^{-\frac{\beta}{\sigma^2}} \propto \left(\sigma^2\right)^{-\left(\alpha+\frac{n}{2}\right)-1} e^{-\frac{\left[\beta+\frac{n}{2}(\bar{x}-\mu)^2\right]}{\sigma^2}} = \left(\sigma^2\right)^{-\alpha^*-1} e^{-\frac{\beta^*}{\sigma^2}}$$

dove $\alpha^* = \alpha + \frac{n}{2}$ e $\beta^* = \beta + \frac{n}{2}\left(\bar{x}-\mu_0\right)^2$.

che è una distribuzione Gamma inversa con parametri $\alpha^* = \alpha + \frac{n}{2}$ e $\beta^* = \beta + \frac{n}{2}\left(\bar{x}-\mu\right)^2$

cioè : $\sigma^2 / \mu, \boldsymbol{x} \sim Inv\Gamma\left[\alpha + \frac{n}{2}\ , \beta + \frac{n}{2}\left(\bar{x}-\mu\right)^2\right]$.

Per σ^2 ***nota****, la v.c. normale*

$$\pi\left(\mu / \sigma^2;\mu_0,\sigma_0^2\right) = \frac{1}{\left(2\pi\sigma_0^2\right)^{n/2}} \cdot e^{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2}$$

è distribuzione a priori coniugata della v.c. Normale, infatti

$$\pi\left(\mu / \sigma, \underline{x};\mu_0,\sigma_0^2\right) = \frac{f\left(\boldsymbol{x} / \mu\right)\cdot\pi\left(\mu\right)}{f\left(\boldsymbol{x}\right)} = \frac{1}{\left(2\pi\sigma^2\right)^{n/2}} \cdot e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} \cdot \frac{1}{\left(2\pi\sigma_0^2\right)^{n/2}} \cdot e^{-\frac{1}{2\sigma_1^2}(\mu-\mu_0)^2} \ / \ f\left(\boldsymbol{x}\right) =$$

$$= \frac{1}{\left(2\pi\sigma^2\right)^{n/2}} \cdot e^{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i-\bar{x})^2 + n\ (\bar{x}-\mu)^2\right]} \cdot \frac{1}{\left(2\pi\sigma_0^2\right)^{n/2}} \cdot e^{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2} \ / \ f\left(\boldsymbol{x}\right) \propto e^{-\frac{1}{2}\left[\frac{n}{\sigma^2}(\bar{x}-\mu)^2 + \frac{1}{\sigma_0^2}(\mu-\mu_0)^2\right]}$$

---

[13] Se $Y \sim \Gamma(\alpha,\beta)$ la v.c $X = 1/Y$ è detta Gamma inversa ed ha funzione di densità $f(x;\alpha,\beta) = \frac{x^{-\alpha-1}\beta^{\alpha}\ e^{-\frac{\beta}{x}}}{\Gamma(\alpha)} \propto x^{-\alpha-1}e^{-\frac{\beta}{x}}$. Si segnala che a risultati analoghi si perviene anche se si considera la v.c. Gamma anziché la v.c. Gamma inversa.

Ma

$$\frac{n}{\sigma^2}\left(\bar{x}-\mu\right)^2+\frac{1}{\sigma_0^2}\left(\mu-\mu_0\right)^2=\left(\frac{n\ \sigma_0^2+\sigma^2}{\sigma^2\sigma_0^2}\right)\left[\mu-\frac{n\,\bar{x}\ \sigma_0^2+\mu_0\sigma^2}{\sigma^2+n\ \sigma_0^2}\right]^2+\frac{n}{\sigma^2+n\ \sigma_0^2}\left(\bar{x}-\mu_0\right)^2$$

da cui

$$\pi\left(\mu/\sigma^2,\underline{x};\mu_0,\sigma_0^2\right)\propto\ e^{-\frac{n}{2\sigma^2}\left(\bar{x}-\mu\right)^2}\ e^{-\frac{1}{2\sigma_0^2}\left(\mu-\mu_0\right)^2}=e^{-\frac{1}{2}\frac{\sigma^2+n\ \sigma_0^2}{\sigma^2\sigma_0^2}\left[\mu-\frac{n\,\bar{x}\ \sigma_0^2+\mu_1\sigma^2}{\sigma^2+n\ \sigma_0^2}\right]^2}=$$

$$=e^{-\frac{1}{2\ \sigma_*^2}\left[\mu-\mu_*\right]^2}$$

dove $\mu_*=\dfrac{n\,\bar{x}\ \sigma_0^2+\mu_1\sigma^2}{\sigma^2+n\ \sigma_0^2}$ *e* $\sigma_*^2=\dfrac{\sigma^2\sigma_0^2}{\sigma^2+n\ \sigma_0^2}$.

La distribuzione marginale a posteriori di $\mu/\boldsymbol{x}$ *è quindi normale*

$$\mu/\sigma^2,\boldsymbol{x}\sim N\left(\frac{n\,\bar{x}\ \sigma_0^2+\mu_0\sigma^2}{\sigma^2+n\ \sigma_0^2}\ ,\frac{\sigma^2\sigma_0^2}{\sigma^2+n\ \sigma_0^2}\right).$$

Si dimostra la relazione

$$\frac{n}{\sigma^2}\left(\bar{x}-\mu\right)^2+\frac{1}{\sigma_1^2}\left(\mu-\mu_1\right)^2=\left(\frac{n\ \sigma_1^2+\sigma^2}{\sigma^2\sigma_1^2}\right)\left[\mu-\frac{n\,\bar{x}\ \sigma_1^2+\mu_1\sigma^2}{\sigma^2+n\ \sigma_1^2}\right]^2+\frac{n}{\sigma^2+n\ \sigma_1^2}\left(\bar{x}-\mu_1\right)^2$$

$$\frac{n}{\sigma^2}\left(\bar{x}-\mu\right)^2+\frac{1}{\sigma_0^2}\left(\mu-\mu_0\right)^2=\frac{n}{\sigma^2}\bar{x}^2+\frac{n}{\sigma^2}\mu^2-\frac{n}{\sigma^2}2\,\bar{x}\,\mu+\frac{1}{\sigma_0^2}\mu^2+\frac{1}{\sigma_0^2}\mu_0^2-2\frac{1}{\sigma_0^2}\mu\,\mu_0=$$

$$=\mu^2\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)-2\,\mu\left(\frac{n}{\sigma^2}\,\bar{x}+\frac{1}{\sigma_0^2}\,\mu_0\right)+\frac{n}{\sigma^2}\bar{x}^2+\frac{1}{\sigma_0^2}\mu_0^2=$$

$$=\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\left[\mu^2+\frac{-2\,\mu\left(\frac{n}{\sigma^2}\,\bar{x}+\frac{1}{\sigma_0^2}\,\mu_0\right)+\frac{n}{\sigma^2}\bar{x}^2+\frac{1}{\sigma_0^2}\mu_0^2}{\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)}\right]=$$

$$=\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\left[\mu^2+\frac{-2\,\mu\left(\frac{n}{\sigma^2}\,\bar{x}+\frac{1}{\sigma_0^2}\,\mu_0\right)}{\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)}+\frac{\left(\frac{n}{\sigma^2}\,\bar{x}+\frac{1}{\sigma_0^2}\,\mu_0\right)^2}{\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)^2}\right]+$$

$$-\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\left[\frac{\left(\frac{n}{\sigma^2}\,\bar{x}+\frac{1}{\sigma_0^2}\,\mu_0\right)^2}{\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)^2}+\frac{\frac{n}{\sigma^2}\bar{x}^2+\frac{1}{\sigma_0^2}\mu_0^2}{\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)}\right]=$$

$$=\left(\frac{n\,\sigma_0^2+\sigma^2}{\sigma^2\sigma_0^2}\right)\left\{\left[\mu-\frac{\frac{n}{\sigma^2}\,\bar{x}+\frac{1}{\sigma_0^2}\,\mu_0}{\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)}\right]^2-\frac{\left(\frac{n}{\sigma^2}\,\bar{x}+\frac{1}{\sigma_0^2}\,\mu_0\right)^2}{\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)^2}+\frac{\frac{n}{\sigma^2}\bar{x}^2+\frac{1}{\sigma_0^2}\mu_0^2}{\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)}\right\}=$$

$$=\left(\frac{n\,\sigma_0^2+\sigma^2}{\sigma^2\sigma_0^2}\right)\left[\mu-\frac{n\,\bar{x}\,\sigma_0^2+\mu_0\sigma^2}{\sigma^2+n\,\sigma_0^2}\right]^2-\frac{\left(\frac{n}{\sigma^2}\,\bar{x}+\frac{1}{\sigma_0^2}\,\mu_0\right)^2}{\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)}+\frac{n}{\sigma^2}\bar{x}^2+\frac{1}{\sigma_0^2}\mu_0^2=$$

$$=\left(\frac{n\,\sigma_0^2+\sigma^2}{\sigma^2\sigma_0^2}\right)\left[\mu-\frac{n\,\bar{x}\,\sigma_0^2+\mu_0\sigma^2}{\sigma^2+n\,\sigma_0^2}\right]^2+\frac{n}{\sigma^2+n\,\sigma_0^2}\left(\bar{x}-\mu_0\right)^2 \qquad c.v.d.$$

Relativamente più onerosa è la derivazione della distribuzione a posteriori quando entrambi i parametri (media e varianza) sono incogniti.

Se si considerano le distribuzioni a priori sopra definite, si assume implicitamente l'indipendenza tra μ e σ^2 *ma in questo caso non è possibile ottenere una distribuzione a priori coniugata, cosa che risulta invece possibile se si assume una relazione di dipendenza tra le 2 variabili esplicitandola nella definizione della distribuzione a priori*

$$\pi\left(\mu,\sigma^2\right)=\pi\left(\mu/\sigma^2\right)\cdot\pi\left(\sigma^2\right).$$

Le due distribuzioni sotto definite (normale e Gamma inversa)

$$\pi\left(\mu/\sigma^2\right)=\frac{1}{\left(2\pi\sigma_1^2\right)^{n/2}}\cdot e^{-\frac{1}{2\sigma_1^2}\left(\mu-\mu_1\right)^2}=\frac{1}{\left(2\pi\ \sigma^2/n_0\right)^{n/2}}\cdot e^{-\frac{1}{2\ \sigma^2/n_0}\left(\mu-\mu_1\right)^2}\quad per\quad \sigma_1^2=\sigma^2/n_0$$

e

$$\pi\left(\sigma^2\right)=\frac{\left(\sigma^2\right)^{-\alpha-1}\beta^{\alpha}\ e^{-\frac{\beta}{\sigma^2}}}{\Gamma\left(\alpha\right)}$$

generano la distribuzione congiunta Normale-Gamma inversa

$$\pi\left(\mu,\sigma^2\right)=\pi\left(\mu/\sigma^2\right)\pi\left(\sigma^2\right)=\frac{e^{-\frac{1}{2\ \sigma^2/n_0}\left(\mu-\mu_1\right)^2}}{\sqrt{2\pi\ \sigma^2/n_0}}\cdot\frac{\left(\sigma^2\right)^{-\alpha-1}\beta^{\alpha}\ e^{-\frac{\beta}{\sigma^2}}}{\Gamma\left(\alpha\right)}\propto$$

$$\propto\left(\sigma^2\right)^{-\frac{1}{2}}e^{-\frac{1}{2\ \sigma^2/n_0}\left(\mu-\mu_1\right)^2}\left(\sigma^2\right)^{-\alpha-1}e^{-\frac{\beta}{\sigma^2}}$$

cioè $\left(\mu,\sigma^2\right)\sim NInvГ\left(\mu_1,\sigma^2;n_0;\alpha,\beta\right)$ *che è distribuzione a priori coniugata di una v.c. che appartiene alla stessa famiglia. Infatti, se si considera la verosimiglianza*

$$L\left(\mu,\sigma^2;\boldsymbol{x}\right)=\frac{1}{\left(2\pi\sigma^2\right)^{n/2}}\cdot e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i-\mu\right)^2}$$

tenendo presente che $x_i-\mu=x_i-\overline{x}+\overline{x}-\mu$ *si ha*

$$L\left(\mu,\sigma^2;\boldsymbol{x}\right)=\frac{1}{\left(2\pi\sigma^2\right)^{n/2}}\cdot e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i-\mu\right)^2}=\frac{1}{\left(2\pi\sigma^2\right)^{n/2}}\cdot e^{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2+n\ \left(\overline{x}-\mu\right)^2\right]}=$$

$$=\frac{1}{\left(2\pi\sigma^2\right)^{n/2}}\cdot e^{-\frac{1}{2\sigma^2}\left[(n-1)\ s^2+n\ \left(\overline{x}-\mu\right)^2\right]}$$

la distribuzione a posteriori congiunta è

$$\pi\left(\mu,\sigma^2/\boldsymbol{x}\right)=\frac{\pi\left(\mu/\sigma^2\right)\pi\left(\sigma^2\right)L\left(\mu,\sigma^2;\underline{x}\right)}{f\left(\boldsymbol{x}\right)}=$$

$$=\frac{e^{-\frac{1}{2\ \sigma^2/n_0}\left(\mu-\mu_1\right)^2}}{\sqrt{2\pi\ \sigma^2/n_0}}\cdot\frac{\left(\sigma^2\right)^{-\alpha-1}\beta^{\alpha}\ e^{-\frac{\beta}{\sigma^2}}}{\Gamma\left(\alpha\right)}\cdot\frac{1}{\left(2\pi\sigma^2\right)^{n/2}}\cdot e^{-\frac{1}{2\sigma^2}\left[(n-1)\ s^2+n\ \left(\overline{x}-\mu\right)^2\right]}/\ f\left(\boldsymbol{x}\right)=$$

$$=\frac{\left(\sigma^2\right)^{-\left(\frac{n}{2}+\alpha\right)-1}\left(\sigma^2\right)^{-\frac{1}{2}}\beta^{\alpha}\ e^{-\frac{1}{2\sigma^2}\left\{\left[n_0\left(\mu-\mu_1\right)^2+n\ \left(\overline{x}-\mu\right)^2\right]+\left[2\beta+(n-1)\ s^2\right]\right\}}}{\sqrt{2\pi\ \sigma^2/n_0}\,\Gamma\left(\alpha\right)\left(2\pi\sigma^2\right)^{n/2}}/\ f\left(\boldsymbol{x}\right)=$$

$$=\frac{\left(\sigma^2\right)^{-\left(\frac{n}{2}+\alpha\right)-1}e^{-\frac{\left(n+n_0\right)}{2\ \sigma^2}\left(\mu-\frac{n_0\ \mu_1}{n+n_0}-\frac{n\ \overline{x}}{n+n_0}\right)^2}}{\left(2\pi\right)^{(n+1)/2}\left(n_0\right)^{-1/2}}\cdot\frac{\left(\sigma^2\right)^{-\frac{1}{2}}\beta^{\alpha}\ e^{-\frac{1}{2\ \sigma^2}\left[2\beta+(n-1)\ s^2+\frac{n\ n_0}{(n+n_0)}\left(\overline{x}-\mu_1\right)^2\right]}}{\Gamma\left(\alpha\right)}/\ f\left(\boldsymbol{x}\right)$$

dove per derivare l'ultimo termine dell'ultima uguaglianza è stata utilizzata la relazione

$$n\ \left(\overline{x}-\mu\right)^2+\ n_0\left(\mu-\mu_1\right)^2=\frac{n_0\ n}{n_0+n}\left(\overline{x}-\mu_1\right)+\left(n_0+n\right)\left(\ \mu-\frac{n_0\ \mu_1}{n+n_0}-\frac{n\ \overline{x}}{n+n_0}\ \right)^2$$

la cui dimostrazione è la stessa svolta in precedenza dove i coefficienti che moltiplicano i due quadrati sono n e n_0 *anziché* $\frac{n}{\sigma^2}\ e\ \frac{1}{\sigma_1^2}$.

Se si pone :

$$\sigma_*^2=\sigma^2/\left(n+n_0\right),\ \mu^*=\frac{n_0\ \mu_1}{n+n_0}-\frac{n\ \overline{x}}{n+n_0},$$

$$\alpha^*=\left(\frac{n}{2}+\alpha\right),\ \ \beta^*=\beta+\frac{(n-1)}{2}\ s^2+\frac{n\ n_0}{2\ \left(n+n_0\right)}\left(\overline{x}-\mu_1\right)^2$$

si ha

$$\pi\left(\mu,\sigma^2\ /\ \boldsymbol{x}\right)=\frac{\pi\left(\mu/\sigma^2\right)\pi\left(\sigma^2\right)L\left(\mu,\sigma^2;\boldsymbol{x}\right)}{f\left(\boldsymbol{x}\right)}\propto$$

$$\propto\ \left(\sigma^2\right)^{-\frac{1}{2}}e^{-\frac{(n+n_0)}{2\ \sigma^2}\left(\mu-\frac{n_0\ \mu_1}{n+n_0}-\frac{n\ \overline{x}}{n+n_0}\right)^2}\left(\sigma^2\right)^{-\left(\frac{n}{2}+\alpha\right)-1}\ e^{-\frac{1}{\sigma^2}\left[\beta+\frac{(n-1)}{2}s^2+\frac{n\ n_0}{2\ (n+n_0)}\left(\overline{x}-\mu_1\right)^2\right]}$$

$$\propto\ \left(\sigma^2\right)^{-\frac{1}{2}}e^{-\frac{(n+n_0)}{2\ \sigma_*^2}\left(\mu-\mu^*\right)^2}\left(\sigma^2\right)^{-\alpha^*-1}\ e^{-\frac{\beta^*}{\sigma^2}}$$

che è una v.c. Normale-Gamma inversa, cioè

$$\left(\mu,\sigma^2\ /\ \boldsymbol{x}\right)\sim NInv\Gamma\left(\mu^*,\sigma_*^2;\alpha^*,\beta^*\right)$$

La distribuzione a posteriori marginale della v.c. $\sigma^2\ /\ \boldsymbol{x}$ *si deduce immediatamente dall'ultima relazione sopra scritta; infatti, se si integra rispetto a* μ *si ottiene* [14]

$$\pi\left(\sigma^2\ /\ \boldsymbol{x}\right)=\int_{-\infty}^{+\infty}\left(\sigma^2\right)^{-\frac{1}{2}}e^{-\frac{1}{2\ \sigma_*^2}\left(\mu-\mu^*\right)^2}\left(\sigma^2\right)^{-\alpha^*-1}\ e^{-\frac{\beta^*}{\sigma^2}}\ d\mu\propto$$

$$\propto\left(\sigma_*^2\right)^{1/2}\left(\sigma^2\right)^{-\frac{1}{2}}\left(\sigma^2\right)^{-\alpha^*-1}\ e^{-\frac{\beta^*}{\sigma^2}}\propto\ \left(\sigma^2\right)^{-\alpha^*-1}\ e^{-\frac{\beta^*}{\sigma^2}}$$

Che è una distribuzione Gamma inversa con parametri

$$\alpha^*=\left(\frac{n}{2}+\alpha\right),\ \ \beta^*=\beta+\frac{(n-1)}{2}\ s^2+\frac{n\ n_0}{2\ \left(n+n_0\right)}\left(\overline{x}-\mu_1\right)^2$$

quindi

$$\sigma^2\ /\ \boldsymbol{x}\sim Inv\Gamma\left[\alpha^*,\beta^*\right]\sim Inv\Gamma\left[\alpha+\frac{n}{2},\beta+\frac{(n-1)}{2}\ s^2+\frac{n\ n_0}{2\ \left(n+n_0\right)}\left(\overline{x}-\mu_1\right)^2\right]$$

La distribuzione condizionata a posteriori di μ *è*

[14] Questa operazione rappresenta un esempio di quanto affermato in precedenza riguardo al trattamento dei parametri di disturbo che possono essere spesso rimossi attraverso una semplice operazione di marginalizzazione.

$$\pi\left(\mu/\sigma^2,\boldsymbol{x}\right)=\frac{\pi\left(\mu,\sigma^2/\boldsymbol{x}\right)}{\pi\left(\sigma^2/\boldsymbol{x}\right)}\propto\frac{\left(\sigma^2\right)^{-\frac{1}{2}}e^{-\frac{1}{2\,\sigma_*^2}\left(\mu-\mu^*\right)^2}\left(\sigma^2\right)^{-\alpha^*-1}\;e^{-\frac{\beta^*}{\sigma^2}}}{\left(\sigma^2\right)^{-\alpha^*-1}\;e^{-\frac{\beta^*}{\sigma^2}}}\propto e^{-\frac{1}{2\,\sigma_*^2}\left(\mu-\mu^*\right)^2}$$

quindi

$$\mu/\sigma^2,\underline{x}\sim N\left(\frac{n_0\;\mu_1}{n+n_0}-\frac{n\;\bar{x}}{n+n_0},\sigma^2/\left(n+n_0\right)\right)$$

Per derivare la distribuzione marginale a posteriori di μ conviene considerare una sottofamiglia della v.c. gamma inversa attraverso una specificazione dei parametri caratteristici ponendo $\alpha=\frac{\nu}{2}$, $\beta=\frac{\nu\;\sigma_1^2}{2}$ *nella distribuzione a priori della varianza. La densità*

$$\pi(\sigma^2)=\frac{\left(\sigma^2\right)^{-\alpha-1}\beta^{\alpha}\;e^{-\frac{\beta}{\sigma^2}}}{\Gamma\left(\alpha\right)}\propto\;\left(\sigma^2\right)^{-\alpha-1}e^{-\frac{\beta}{\sigma^2}}$$

diventa

$$\pi(\sigma^2)=\frac{\left(\sigma^2\right)^{-\frac{\nu}{2}-1}\left(\nu\;\sigma_1^2\right)^{\frac{\nu}{2}}\;e^{-\frac{\nu\;\sigma_1^2}{2\,\sigma^2}}}{\Gamma\left(\nu/2\right)}\propto\;\left(\sigma^2\right)^{-\frac{\nu}{2}-1}e^{-\frac{\nu\;\sigma_1^2}{2\,\sigma^2}}$$

che è una v.c. chi quadro inversa scalata [15] *con* ν *gradi di libertà e parametro di scala* σ_1^2 *cioè*

$$\sigma^2\sim InvS\chi^2\left(\nu,\sigma_1^2\right).$$

Con tale specifica la distribuzione a priori congiunta assume la forma

$$\pi\left(\mu,\sigma^2\right)=\pi\left(\mu/\sigma^2\right)\pi\left(\sigma^2\right)=\frac{e^{-\frac{1}{2\,\sigma^2/n_0}(\mu-\mu_1)^2}}{\sqrt{2\pi\;\sigma^2/n_0}}\cdot\frac{\left(\sigma^2\right)^{-\frac{\nu}{2}-1}\left(\nu\;\sigma_1^2\right)^{\frac{\nu}{2}}\;e^{-\frac{\nu\;\sigma_1^2}{2\,\sigma^2}}}{\Gamma\left(\nu/2\right)}\propto$$

$$\propto\left(\sigma^2\right)^{-\frac{1}{2}}e^{-\frac{1}{2\,\sigma^2/n_0}(\mu-\mu_1)^2}\left(\sigma^2\right)^{-\alpha-1}\;e^{-\frac{\nu\;\sigma_1^2}{2\,\sigma^2}}$$

che è una v.c.normale-chi quadro inversa scalata con ν *gradi di libertà e parametro di scala* σ_1^2 *cioè*

$$\mu,\sigma^2\sim NInvS\chi^2\left(\mu_1,\frac{\sigma^2}{n_0};\nu,\sigma_1^2\right).$$

Con tale specifica la distribuzione a posteriori congiunta assume la forma

[15] La v.c. chi-quadro inversa è definita come sottofamiglia della v.c. gamma inversa mediante una specifica dei parametri. Nella v.c. chi-quadro inversa scalata, oltre al parametro che misura i gradi di libertà, è presente un ulteriore parametro di scala.

$$\pi\left(\mu,\sigma^2 / \boldsymbol{x}\right)=\frac{\pi\left(\mu / \sigma^2\right)\pi\left(\sigma^2\right)L\left(\mu,\sigma^2;\boldsymbol{x}\right)}{f\left(\boldsymbol{x}\right)}=$$

$$=\frac{\left(\sigma^2\right)^{-\left(\frac{n}{2}+\frac{\nu}{2}\right)-1} e^{-\frac{1}{2\,\sigma^2/(n+n_0)}\left(\mu-\frac{n_0\,\mu_1}{n+n_0}-\frac{n\,\bar{x}}{n+n_0}\right)^2}}{\left(2\pi\right)^{(n+1)/2}\left(n_0\right)^{-1/2}}\cdot\frac{\left(\sigma^2\right)^{-\frac{1}{2}}\left(2/\nu\;\sigma_1^2\right)^{-\frac{\nu}{2}} e^{-\frac{1}{2\,\sigma^2}\left[\nu\;\sigma_1^2+(n-1)\,s^2+\frac{n\,n_0}{(n+n_0)}(\bar{x}-\mu_1)^2\right]}}{\Gamma\left(\nu/2\right)}/\;f\left(\boldsymbol{x}\right)$$

da cui

$$\pi\left(\mu,\sigma^2 / \boldsymbol{x}\right)\propto\frac{e^{-\frac{(n+n_0)}{2\,\sigma^2}\left(\mu-\frac{n_0\,\mu_1}{n+n_0}-\frac{n\,\bar{x}}{n+n_0}\right)^2}}{\sqrt{2\pi\;\sigma^2/n_0}}\cdot\frac{\left(\sigma^2\right)^{-\left(\frac{n+\nu}{2}+\frac{3}{2}\right)} e^{-\frac{1}{2\,\sigma^2}\left[\nu\;\sigma_1^2+(n-1)\,s^2+\frac{n\,n_0}{(n+n_0)}(\bar{x}-\mu_1)^2\right]}}{\Gamma\left(\alpha\right)\left(2\pi\right)^{n/2}}\propto$$

$$\propto\left(\sigma^2\right)^{-\frac{1}{2}} e^{-\frac{1}{2\,\sigma^2/(n+n_0)}\left(\mu-\frac{n_0\,\mu_1}{n+n_0}-\frac{n\,\bar{x}}{n+n_0}\right)^2}\left(\sigma^2\right)^{-\frac{n+\nu}{2}-1} e^{-\frac{1}{2\,\sigma^2}\left[\nu\;\sigma_1^2+(n-1)\,s^2+\frac{n\,n_0}{(n+n_0)}(\bar{x}-\mu_1)^2\right]}=$$

$$=\left(\sigma^2\right)^{-\frac{1}{2}} e^{-\frac{1}{2\,\sigma_*^2}\left(\mu-\mu^*\right)^2}\left(\sigma^2\right)^{-\frac{\nu^*}{2}-1} e^{-\frac{\beta^*}{2\,\sigma^2}}$$

dove

$$\mu^*=\frac{n_0\,\mu_1}{n+n_0}-\frac{n\,\bar{x}}{n+n_0},\;\sigma_*^2=\frac{\sigma^2}{n+n_0},\;\nu^*=n+\nu\;e\;\beta^*=\nu\;\sigma_1^2+(n-1)\,s^2+\frac{n\,n_0}{\left(n+n_0\right)}\left(\bar{x}-\mu_1\right)^2.$$

Pertanto, la distribuzione a priori congiunta è data dal prodotto di una v.c. normale e una v.c. χ^2 *inversa scalata*

$$\mu,\sigma^2/\boldsymbol{x}\sim NInvS\chi^2\left[\frac{n_0\,\mu_1}{n+n_0}-\frac{n\,\bar{x}}{n+n_0},\;\frac{\sigma^2}{n+n_0};\;n+\nu,\;\nu\;\sigma_1^2+(n-1)\,s^2+\frac{n\,n_0}{\left(n+n_0\right)}\left(\bar{x}-\mu_1\right)^2\right]$$

cioè

$$\mu,\sigma^2\;/\;\boldsymbol{x}\sim NInvS\chi^2\left(\mu^*,\sigma_*^2;\nu^*,\beta^*\right).$$

quindi, la distribuzione marginale a posteriori di σ^2 *è*

$$\sigma^2\;/\;\boldsymbol{x}\sim InvS\chi^2\left(\nu+n,\;\nu\;\sigma_1^2+(n-1)\;s^2+\frac{n\,n_0}{\left(n+n_0\right)}\left(\bar{x}-\mu_1\right)^2\right)=InvS\chi^2\left(\nu^*,\beta*\right)$$

.

cioè, la distribuzione marginale a posteriori della varianza è una v.c. chi-quadro inversa scalata con $\nu^*=\nu+n$ *gradi di libertà e con parametro di scala*

$$\beta^*=\nu\;\sigma_1^2+(n-1)\;s^2+\frac{n\,n_0}{\left(n+n_0\right)}\left(\bar{x}-\mu_1\right)^2.$$

Mentre la distribuzione marginale a posteriori di $\mu\;/\;\boldsymbol{x}$ *si ottiene integrando rispetto a* σ^2 *la distribuzione a posteriori congiunta .*

$$\pi(\mu / \boldsymbol{x}) = \int_0^{+\infty} \pi(\mu, \sigma^2 / \boldsymbol{x})\, d\sigma^2 \propto$$

$$= \int_0^{+\infty} \left(\sigma^2\right)^{-\frac{1}{2}} e^{-\frac{1}{2\,\sigma^2/(n+n_0)}\left(\mu - \frac{n_0\,\mu_1}{n+n_0} - \frac{n\,\bar{x}}{n+n_0}\right)^2} \left(\sigma^2\right)^{-\frac{n+\nu}{2}-1} e^{-\frac{1}{2\,\sigma^2}\left[\nu\,\sigma_1^2 + (n-1)\,s^2 + \frac{n\,n_0}{(n+n_0)}(\bar{x}-\mu_1)^2\right]} d\sigma^2 =$$

$$= \int_0^{+\infty} \left(\sigma^2\right)^{-\frac{n+\nu+3}{2}} e^{-\frac{1}{2\,\sigma^2}\left[\nu\,\sigma_1^2 + (n-1)\,s^2 + \frac{n\,n_0}{(n+n_0)}(\bar{x}-\mu_1)^2 + (n+n_0)\left(\mu - \frac{n_0\,\mu_1}{n+n_0} - \frac{n\,\bar{x}}{n+n_0}\right)^2\right]} d\sigma^2$$

Se si pone

$$A = \nu\,\sigma_1^2 + (n-1)\,s^2 + \frac{n\,n_0}{(n+n_0)}(\bar{x}-\mu_1)^2 + (n+n_0)\left(\mu - \frac{n_0\,\mu_1}{n+n_0} - \frac{n\,\bar{x}}{n+n_0}\right)^2$$

$$e\;\; z = A/2\sigma^2 \;\Rightarrow\; \sigma^2 = A/2\,z \;\Rightarrow d\sigma^2 \;\Rightarrow\; d\sigma^2 = -\frac{A}{2z^2}dz$$

si ha

$$\pi(\mu / \boldsymbol{x}) \propto \int_0^{+\infty} \left(\frac{A}{2z}\right)^{-(n+\nu+3)/2} e^{-z} \left(-\frac{A}{2z^2}\right) dz \propto A^{-(n+\nu+1)/2} \int_0^{+\infty} (z)^{(n+\nu-1)/2}\, e^{-z}\, dz =$$

$$= A^{-\frac{n+\nu+1}{2}} = \left[\nu\,\sigma_1^2 + (n-1)\,s^2 + \frac{n\,n_0}{(n+n_0)}(\bar{x}-\mu_1)^2 + (n+n_0)\left(\mu - \frac{n_0\,\mu_1}{n+n_0} - \frac{n\,\bar{x}}{n+n_0}\right)^2\right]^{-(n+\nu-1)/2} \propto$$

$$\propto \left\{1 + (n+n_0)\left(\mu - \frac{n_0\,\mu_1}{n+n_0} - \frac{n\,\bar{x}}{n+n_0}\right)^2 \Big/ \left[\nu\,\sigma_1^2 + (n-1)\,s^2 + \frac{n\,n_0}{(n+n_0)}(\bar{x}-\mu_1)^2\right]\right\}^{-(n+\nu-1)/2} =$$

$$= \left[1 + \frac{(\mu-\mu^*)^2}{\delta/\nu^*}\right]^{-(n+\nu-1)/2} \qquad con\;\; \delta = \nu\,\sigma_1^2 + (n-1)\,s^2 + \frac{n\,n_0}{(n+n_0)}(\bar{x}-\mu_1)^2$$

che, a meno della costante di normalizzazione, rappresenta una v.c. t scalata non centrale con $\nu^* = n + \nu$ *gradi di libertà, parametro di non centralità* $\mu^* = \frac{n_0\,\mu_1}{n+n_0} - \frac{n\,\bar{x}}{n+n_0}$ *e parametro di*

scala $\beta^* = \left(\mu - \frac{n_0\,\mu_1}{n+n_0} - \frac{n\,\bar{x}}{n+n_0}\right)^2 / \nu\,\sigma_1^2 + (n-1)\,s^2 + \frac{n\,n_0}{(n+n_0)}(\bar{x}-\mu_1)^2$ *, cioè*

$$\mu / \underline{x} \sim NCSt_{\nu^*}(\mu^*, \delta) = NCSt_{\nu+n}\left[\frac{n_0\,\mu_1}{n+n_0} - \frac{n\,\bar{x}}{n+n_0}\;,\; \nu\,\sigma_1^2 + (n-1)\,s^2 + \frac{n\,n_0}{(n+n_0)}(\bar{x}-\mu_1)^2\right].$$

Esempio 6.7 (distribuzioni multidimensionali)

Nel Cap. 1 se è già avuto modo di considerare la v.c. di Dirichlet come distribuzione a priori coniugata della v.c multinomiale, in questo esempio si procederà all'esame della v.c. normale a k dimensioni.

Operando in modo analogo a quanto già fatto per la v.c. normale semplice e ricordando che la funzione di densità di probabilità della v.c. normale a k dimensioni è espressa da

$$f(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{k/2}\left|\boldsymbol{\Sigma}\right|^{1/2}} e^{-\frac{1}{2}\left[(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]}$$

dove

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_k \end{bmatrix};\quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \cdot \\ \cdot \\ \mu_k \end{bmatrix};\quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1k} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{k1} & \sigma_{k2} & \cdots & \sigma_{kk} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1k} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{k1} & \sigma_{k2} & \cdots & \sigma_k^2 \end{bmatrix}$$

Per $\boldsymbol{\mu}$ *noto, la v.c., se si ipotizza che la matrice di dispersione* $\boldsymbol{\Sigma}$ *(definita positiva) si distribuisce come una Wishart inversa con parametri* ν *(gradi di libertà) e* $\boldsymbol{\Sigma}_0$ *(matrice definita positiva) è facile verificare che la stessa è distribuzione a priori coniugata della v.c. multidimensionale Wishart inversa. Infatti, poiché la funzione a priori di densità di probabilità è*

$$\pi(\boldsymbol{\Sigma}/\boldsymbol{\mu};\nu,\boldsymbol{\Sigma}_0) = \frac{\left|\boldsymbol{\Sigma}_0\right|^{\frac{n}{2}}\left|\boldsymbol{\Sigma}\right|^{\frac{n-k-1}{2}}}{2^{\frac{n\,k}{2}}\boldsymbol{\Gamma}_k\left(n/2\right)} e^{-\frac{1}{2}tr\left(\boldsymbol{\Sigma}_0\,\boldsymbol{\Sigma}^{-1}\right)} \propto \left|\boldsymbol{\Sigma}\right|^{\frac{n-k-1}{2}} e^{-\frac{1}{2}tr\left(\boldsymbol{\Sigma}_0\,\boldsymbol{\Sigma}^{-1}\right)}$$

dove $\boldsymbol{\Gamma}_k\left(n/2\right)$ *è la funzione gamma multivariata*

$$\boldsymbol{\Gamma}_k\left(n/2\right) = \pi^{k(k-1)/4}\prod_{i=1}^{n}\Gamma\left[n/2+(1-i)/2\right]$$

mentre la funzione di verosimiglianza per un campione di dimensione n estratto da una v.c. normale a k dimensioni è

$$L(\boldsymbol{\mu},\boldsymbol{\Sigma}/\boldsymbol{X}) = \frac{1}{(2\pi)^{\frac{n\,k}{2}}\left|\boldsymbol{\Sigma}\right|^{\frac{n}{2}}} e^{-\frac{1}{2}\left[\sum_{i=1}^{n}(\boldsymbol{x}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})\right]}$$

si ha

$$\pi(\boldsymbol{\Sigma}/\boldsymbol{\mu},\boldsymbol{X};\nu,\boldsymbol{\Sigma}_0) = \frac{\pi(\boldsymbol{\Sigma}/\boldsymbol{\mu},\boldsymbol{X};\nu,\boldsymbol{\Sigma}_0)L(\boldsymbol{\mu},\boldsymbol{\Sigma}/\boldsymbol{X})}{f(\boldsymbol{X})} =$$

$$= \frac{1}{(2\pi)^{n\,k/2}\left|\boldsymbol{\Sigma}\right|^{n/2}} e^{-\frac{1}{2}\left[\sum_{i=1}^{n}(\boldsymbol{x}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})\right]} \frac{\left|\boldsymbol{\Sigma}_0\right|^{n/2}\left|\boldsymbol{\Sigma}\right|^{(n-k-1)/2}}{2^{n\,k/2}\boldsymbol{\Gamma}_k\left(n/2\right)} e^{-\frac{1}{2}tr\left(\boldsymbol{\Sigma}_0\,\boldsymbol{\Sigma}^{-1}\right)} \propto$$

$$\propto e^{-\frac{1}{2}\left[\sum_{i=1}^{n}(\boldsymbol{x}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})\right]} \frac{\left|\boldsymbol{\Sigma}_0\right|^{n/2}\left|\boldsymbol{\Sigma}\right|^{(n-k-1)/2}}{2^{\frac{n\,k}{2}}\boldsymbol{\Gamma}_k\left(n/2\right)} e^{-\frac{1}{2}tr\left(\boldsymbol{\Sigma}_0\,\boldsymbol{\Sigma}^{-1}\right)}$$

dove $\boldsymbol{x}_i = (x_{1i}, x_{2i}, \cdots, x_{ki})'$ è l'*i-esimo* vettore delle osservazione campionarie.

Esempio 6.8 (famiglia esponenziale)
Ai risultati sopra illustrati si poteva pervenire attraverso specificazioni delle conclusioni cui si perviene se si fa riferimento alla famiglia esponenziale regolare la cui distribuzione a priori coniugata è facilmente derivabile. Infatti, se si riprende in considerazione la funzione di verosimiglianza di una v.c. appartenente alla famiglia esponenziale regolare caratterizzata da un solo parametro θ (cfr. paragrafo 2 del secondo capitolo)

$$f(\theta; x_1,\ x_2, ..., x_n) = f(\theta; \boldsymbol{x}) = \prod_{i=1}^{n} f(\theta,\ x_i) =$$

$$= \left[a(\theta)\right]^n \cdot \prod_{i=1}^{n} h(x_i) \cdot e^{\varphi(\theta) \sum_{i=1}^{n} t(x_i)}$$

e si introduce una distribuzione a priori per il parametro θ appartenente alla stessa famiglia esponenziale

$$\pi(\theta / \alpha, \beta) \propto \left[a(\theta)\right]^{\alpha} \cdot e^{\varphi(\theta)\,\beta} \quad per \quad \alpha > 0,$$

si ottiene la distribuzione a posteriori

$$\pi(\theta / \boldsymbol{x}, \alpha, \beta) \propto \left[a(\theta)\right]^{\alpha+n} \cdot e^{\varphi(\theta)\left[\,\beta + t(\underline{x})\right]} = \left[a(\theta)\right]^{\alpha^*} \cdot e^{\varphi(\theta)\cdot\beta^*}$$

che appartiene alla stessa famiglia.

L'estensione al caso multi-parametrico è immediata. Infatti, se la distribuzione a priori appartiene alla famiglia esponenziale

$$\pi(\theta / \alpha, \boldsymbol{\beta}) \propto \left[a(\boldsymbol{\theta})\right]^{\alpha} \cdot e^{\sum_{i=1}^{r} \varphi_i(\boldsymbol{\theta})\,\beta_i}$$

dove, $\alpha > 0,\ \boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_r)$ *e* $\boldsymbol{\theta} = (\theta_1,\ \theta_2, ...,\ \theta_k)$,

tenendo presente che la verosimiglianza della famiglia esponenziale nel caso multi-parametrico è

$$L(\boldsymbol{\theta} / \boldsymbol{x}) = a(\boldsymbol{\theta})\ h(\boldsymbol{x}) \cdot e^{\sum_{i=1}^{r} \varphi_i(\boldsymbol{\theta})\ t_i(\boldsymbol{x})}$$

Si ottiene la distribuzione a posteriori del vettore dei parametri $\boldsymbol{\theta}$

$$\pi(\theta / \boldsymbol{x}, \alpha, \boldsymbol{\beta}) \propto \left[a(\boldsymbol{\theta})\right]^{\alpha+n} \cdot e^{\sum_{i=1}^{r} \varphi_i(\boldsymbol{\theta})\ t_i(\boldsymbol{x})\ \beta_i} = \left[a(\boldsymbol{\theta})\right]^{\alpha^*} \cdot e^{\varphi(\boldsymbol{\theta})\cdot\beta_i^+}$$

che appartiene alla famiglia esponenziale.

A conclusione di questo paragrafo si deve sottolineare che alla scelta della distribuzioni a priori coniugata si perviene, nella generalità dei casi, soggettivamente, mentre la sua specificazione completa può avere sia natura soggettiva che oggettiva; specificazione che riguarda in particolare la presenza di iperparametri, usualmente

incogniti, cui deve essere attribuito un valore numerico che può essere derivato soggettivamente dalla valutazione di esperti o oggettivamente da rilevazioni empiriche precedenti relative alle manifestazioni dello stesso fenomeno di fenomeni di natura analoga.
Si tratta in ogni caso di distribuzioni a priori che risulta ragionevole classificare come informative, la cui caratteristica principale è quella di contribuire in modo rilevante alla sintesi dei dati, al riguardo O'Hagan (2004) afferma: "*The most important consideration in the use of prior information is to ensure that the prior distribution honestly reflects genuine information, not personal bias, prejudice, superstition or other factors that are justly condemned in science as 'subectivity'*".

Per contro, a caratteristica principale delle distribuzioni a priori non informative è quella di essere dominate dalla verosimiglianza, nel senso che incidono in modo marginale sulla distribuzione a posteriori. Il paragrafo successivo è dedicato ad un sintetico richiamo di alcune tra le proposte più significative dedicate all'argomento.

6.3 - Distribuzioni a priori non informative

Un aspetto preliminare su cui richiamare l'attenzione quando si propone l'utilizzo di una distribuzione non informativa è la possibilità che si tratti di una ***distribuzione impropria***, cioè di una distribuzione per la quale vale la relazione $\int_{\theta} f(\theta) d\theta = \infty$ che può comportare come conseguenza una distribuzione a posteriori impropria, in questo caso non è possibile alcuna inferenza; non sorge nessun problema, invece, quando pur essendo impropria la distribuzione a priori la corrispondente distribuzione a posteriori è propria.

La prima regola per la determinazione di una distribuzione a priori non informativa è quella collegata al *principio della ragione insufficiente*, usualmente attribuita a ***Bayes*** e a ***Laplace***, che facendo riferimento alla distribuzione di Binomiale assegnano al parametro *p* un'uguale probabilità a tutte le possibili alternative (distribuzione uniforme nell'intervallo $[0,1]$).

Esempio 6.9 – Distribuzione binomiale e distribuzione a priori Uniforme
Nell'esempio 6.3 si è proceduto alla derivazione della distribuzione a posteriori della binomiale introducendo la v.c. Beta come a priori. La distribuzione a posteriori è espressa dalla formula

$$f(p/n,x) = \frac{p^{x+\alpha-1}\,(1-p)^{n-x+\beta-1}}{B(x+\alpha, n-x+\beta)}$$

che è la funzione di densità di probabilità di una v.c. di tipo Beta con parametri $x+\alpha$ e $n-x+\beta$.

Ovviamente, per poter utilizzare questa distribuzione occorre conoscere i valori dei parametri α e β *che identificano la specifica v.c. appartenente alla famiglia Beta; fissazione dei valori che può essere effettuata utilizzando il patrimonio informativo a disposizione o in modo completamente soggettivo. Se non si possiede alcuna informazione oggettiva e si ritiene*

opportuno evitare la valutazione puramente soggettiva si possono scegliere i valori $\alpha = 2$ e $\beta = 1$ *che definisce la funzione di densità a priori per il parametro* p

$$f(p) = \frac{1}{1-0}$$

che rappresenta la funzione di densità di un v.c. rettangolare, cioè una variabile casuale uniforme definita nell'intervallo unitario. La distribuzione a posteriore sopra definita diventa

$$f(p/n,x) = \frac{f(p)\, f(x/p)}{\int_0^1 f(p)\, f(x/p)\, dp} = \frac{p^{x+1}\,(1-p)^{n-x+}}{B(x+2, n-x+1)}\ .$$

Il ricorso alla distribuzione a priori uniforme ingenera due problemi, il primo è che la distribuzione uniforme non è invariante rispetto alla riparametrizzazione, il secondo problema è legato alla dimensione dello spazio parametrico, se tale spazio è infinito l'a priori uniforme è impropria.

Jeffreys nel 1946 propone come regola generale per la derivazione della distribuzione a priori la radice quadrata positiva del determinante della matrice dell'informazione di Fisher

$$\pi_{jef}(\boldsymbol{\theta}) \propto \det\left[\boldsymbol{I}_n(\theta)_{i,j}\right]^{1/2} = \det\left\{-E\left[\frac{\partial^2 \log f(\boldsymbol{x};\boldsymbol{\theta})}{\partial\theta_i\ \partial\theta_j}\right]\right\}^{1/2} =$$

$$= \det\left\{-E\left[\frac{\partial^2}{\partial\theta_i\ \partial\theta_j}\log\ \prod_{i=1}^{n} f\left(X_i; f(\boldsymbol{x};\boldsymbol{\theta})\right)\right]\right\}^{1/2} = Cov\left[\frac{\partial \log f(\boldsymbol{x};\boldsymbol{\theta})}{\partial\theta_i}, \frac{\partial \log f\left(\boldsymbol{x}; f(\boldsymbol{x};\boldsymbol{\theta})\right)}{\partial\theta_j}\right].$$

Nel caso di un solo parametro θ, la distribuzione a priori è

$$\pi_{jef}(\theta) \propto I(\theta)^{1/2} = \left\{-E\left[\frac{d^2 \log f(\boldsymbol{x};\theta)}{d\theta^2}\right]\right\}^{1/2}$$

La giustificazione di una tale scelta è duplice: l'invarianza rispetto alla riparametrizzazione e la constatazione che l'informazione di Fisher è un indicatore dell'ammontare di informazione fornite, tramite il modello, dalle osservazioni campionarie sul valore del parametro incognito $\boldsymbol{\theta}$. La proposta di Jeffreys è largamente accettata per modelli caratterizzati da un solo parametro, ad analoga conclusione non si perviene quando la distribuzione è caratterizzata da più paramentri , inoltre, per molte distribuzioni l'a priori di Jeffreys è impropria e viola il principio di verosimiglianza.

Esempio 6.10 – Distribuzione binomiale e distribuzione a priori di Jeffreys

La Jeffreys prior della distribuzione Binomiale $X \sim Bin(n,p)$ *è*

$$\pi_{jef}(p) \propto\ p^{-1/2}\ (1-p)^{-1/2}$$

infatti

$$I(p) = -E\left[\frac{d^2 \log f(\boldsymbol{x}; p)}{dp^2}\right] = \frac{np}{p^2} + \frac{n-np}{(1-p)^2} = \frac{n}{p(1-p)}$$

quindi

$$\pi_{jef}(p) = \left[I(p)\right]^{1/2} = p^{-\frac{1}{2}}\ (1-p)^{-\frac{1}{2}}$$

La distribuzionea priori di Jeffreys è, quindi, una variabile casuale di tipo Beta con parametri $\alpha = \frac{1}{2}$ *e* $\beta = \frac{1}{2}$*, distribuzione questa che, come già sottolineato, è distribuzione a priori coniugata della binomiale. Anche la distribuzione a priori uniforme è di tipo Beta con parametri* $\alpha = 1$ *e* $\beta = 1$.

Si sottolinea che a differenza di quanto verificato per la distribuzione binomiale, nella generalità dei casi la distribuzione a priori di Jeffreys non si risolve in una distribuzione a priori coniugata, come si avrà modo di vericare nel successivo esempio.

Esempio 6.11 – Distribuzione di Poisson e distribuzione a priori di Jeffreys

La Jeffreys prior della distribuzione di Poisson $X \sim P(\lambda)$ *è* $\frac{1}{\lambda}$*, infatti*

$$I(p) = -E\left[\frac{d^2 \log f(\boldsymbol{x}; \lambda)}{dp^2}\right] = \frac{1}{\lambda}$$

quindi

$$\pi_{jef}(\lambda) = \left[I(\lambda)\right]^{1/2} = \lambda^{-1/2}$$

che è una distribuzione Gamma impropria con parametri $\alpha = 0{,}5$ *e* $\beta = 0$.

Altre interessanti proposte di derivazione della distribuzione a priori, ma non esenti da critiche, sono state avanzate Bernardo e da Jaynes[16].

La ***reference prior***[17], proposta inizialmente da ***Bernardo*** e sviluppato successivamente soprattutto da questo stesso autore in collaborazione con ***Berger*** (1992, 2009) è basata sulla massimizzazione della divergenza attesa tra la distribuzione a posteriori e la distribuzione a priori.

Se $f(x,\theta)$ è la funzione di densità di probabilità della variabile casuale *X* caratterizzata da un solo parametro θ e *T(X)* una statistica sufficiente per θ , il che implica la corrispondenza biunivoca $f(x,\theta) \Leftrightarrow f\left[T(x),\theta\right]$, ***Bernardo*** (1979)

[16] Per altri esempi si veda ***Lisman*** e ***Zuylen*** (1972)

[17] Al riguardo si sottolinea che diversi autori hanno proposto di utilizzare la terminologia *reference prior* (distribuzione a priori di riferimento) anziché la terminologia *distribuzioni a priori non informativa* sostenendo, a ragione secondo l'autore di queste note, che qualunque distribuzione a priori contiene un qualche elemento informativo. Accettando tale proposta si potrebbe connotare, come avviene in altri casi, tale distribuzione rifacendosi all'autore che l'ha proposta: quindi distribuzione a priori di ***Bernardo***, o anche di ***Bernardo-Berger***, anziché reference prior. Si sottolinea, inoltre, che le tre proposte di ***Jeffreys***, ***Bernardo*** e ***Jaynes***, pur differenziandosi tra loro, hanno numerosi punti di contatto.

ipotizzando la disponibilità di un campione di osservazioni $\boldsymbol{X}' = \left(X_1, X_2, ..., X_n\right) = x_n^*$ propone di derivare la distribuzione a priori $\pi_{ref}\left(\theta\right)$ massimizzando il valore atteso della distanza di ***Kulback-Leibler*** definita dalla relazione

$$K_n\left[\pi\left(\theta, x_n^*\right), \pi\left(\theta\right)\right] = \int \pi\left(\theta, x_n^*\right) \; log \; \left[\pi\left(\theta, x_n^*\right) / \pi\left(\theta\right)\right] \; d\theta \, .$$

Indicando con K_n^{π} il valore atteso delle distanza rispetto a $\boldsymbol{X}$ si ottiene

$$K_n^{\pi} = E_{x_n^*}\left\{K_n\left[\pi\left(\theta, x_n^*\right), \pi\left(\theta\right)\right]\right\} = \int\int ... \int \; \left\{\int \pi\left(\theta, x_n^*\right) \; log \; \left[\pi\left(\theta, x_n^*\right) / \pi\left(\theta\right)\right] \; d\theta\right\} \; dx_1 \; dx_2 ... dx_n$$

.

la reference prior è quella che massimizza

$$K_{\infty}^{\pi} = \lim_{n \to \infty} K_n^{\pi} \, .$$

Nella generalità dei casi tale limite è infinito, per superare questa difficoltà si determina la distribuzione priori K_n^{π} che massimizza K_n^{π} e si cerca il limite della corrispondente sequenza di distribuzioni a posteriori, la reference prior è quella che corrisponde alla distribuzione limite a posteriori.

Per le distribuzioni caratterizzate da un solo parametro la reference prior e la Jeffrey's prior coincidono, differeiscono nel caso multiparametrico

Un'altra proposta di distribuzione non informativa è quella basata sulla massimizzazione dell'entropia, sviluppata soprattutto da ***Jaynes*** (1963, 1968).

Per variabili casuali semplici discrete caratterizzate da un solo parametro θ

$$P\left(X = x_i\right) = f\left(x_i, \theta\right) \text{ per } i = 1, 2, ..., k$$

l'entropia è definita da

$$H\left(X\right) = -\sum_{i=1}^{k} f\left(x_i, \theta\right) \; log \; f\left(x_i, \theta\right).$$

per variabili casuali continue con funzione di densità di probabilità $f\left(x\right)$ l'entropia è definita da

$$H\left(X\right) = -\int_{-\infty}^{\infty} f\left(x, \theta\right) \; log \; \; f\left(x, \theta\right) \; dx \, .$$

La distribuzione a priori $\pi_{ja}\left(\theta\right)$ del parametro θ deriva dalla massimizzazione dell'entropia soggetta ai vincoli derivanti dalle conoscenze disponibili sulla distribuzione.

Nel caso di variabili casuali discrete e di nessun vincolo, oltre a quello della normalizzazione, l'entropia è massimizzata dalla distribuzione uniforme $\pi\left(\theta\right) = \frac{1}{k}$.

Allo stesso risultato, distribuzione uniforme $\pi\left(\theta\right) = \frac{1}{b-a}$, si perviene per le variabili casuali continue definite in un intervallo finito $\left[a, b\right]$.

Nel caso in cui al vincolo di normalizzazione si aggiungono i vincoli della conoscenza del momento primo rispetto all'origine $E(\theta)=\beta$ e di valori non negativi della variabile la distribuzione risultante è l'esponenziale negativa

$$\pi(\theta)=\frac{1}{\beta}\, e^{-\frac{\theta}{\beta}}.$$

Il ricorso alle distribuzioni a priori, impropriamente dette non informative[18], viene usualmente connotato, impropriamente, come oggettivo; infatti, anche se le tre proposte di ***Jeffreys***, ***Bernardo*** e ***Jaynes*** hanno numerosi punti di contatto, i risultati cui si perviene non sono coincidenti in molte situazioni di ricerca. Pertanto, la scelta della distribuzione a priori non informativa, che nella generalità dei casi non può che essere basata su considerazioni di natura soggettiva, ingenera forti dubbi sulla presunta oggettività delle a priori non informative anche se la specifica caratteristica di queste distribuzioni è, come sopra sottolineato, di incidere in modo marginale sulle distribuzioni a posteriori.

Un ulteriore elemento di riflessione riguardo all'impiego delle distribuzioni a priori non informative è quanto affermato da ***Seidenfeld*** (1979): "*I claim the twin inductive principles which form the core of objective Bayesianism are unacceptable. Invariance (due to H. Jeffreys) and the rule of maximum entropy (due to E. Jaynes) are each incompatible with conditionalization (Bayes theorem). I argue that the former principle leads to inconsistent representations of "ignorance", i.e., so called informationless priors generated by invariance principle are at odds with Bayes theorem, I claim that Jaynes rule of maximizing the entropy of a distribution to represent 'partial information' is likewise unacceptable . It leads precise probability distributions that are excessively aprioristic, conteining more information than the evidence generating them allows. Again, the conflicts is with Bayes' theorem.*"

6.4 - Stima e verifica di ipotesi in ottica bayesiana

Nei capitoli precedenti son stati illustrati alcuni tra i metodi statistici proposti in letteratura per la risoluzione dei problemi di stima, puntuale e di intervallo, e di verifica di ipotesi relativamente all'entità incognita θ, cioè al parametro o ai parametri che caratterizzano il modello $f(x;\theta)$ la cui forma analitica si presume nota. Sono state discusse, come più volte sottolineato, le soluzioni proposte nel contesto del cosi detto approccio frequentista all'inferenza statistica (inferenza statistica classica). In questo paragrafo verranno illustrate molto sommariamente le soluzioni proposte nel contesto bayesiano.

[18] Uno dei più autorevoli sostenitori dell'approccio bayesiano oggettivo ***Bernardo*** (1997) al riguardo dichiara: "*Non-informative priors do not exist*".

6.4.1 - Stima puntuale e di intervallo di parametri caratteristici

Da quanto illustrato nei paragrafi precedenti dovrebbe emergere in modo evidente la logica di base del cosi detto paradigma bayesiano quando si vuol procedere all'analisi di un qualunque fenomeno attraverso l'impiego di un modello probabilistico capace di fornirne una rappresentazione soddisfacente.

Il modello, la cui forma analitica si è presupposta nota, è caratterizzato da uno o più parametri nei confronti dei quali si presume una conoscenza a priori incerta che viene espressa facendo ricorso ad uno specifico modello probabilistico. Il livello di conoscenza attuale del ricercatore si incrementa attraverso l'acquisizione di informazioni campionarie (oggettive) che consentono l'aggiornamento dello stato di conoscenza attraverso un passaggio dalla distribuzione di probabilità a priori alla distribuzione di probabilità a posteriori che costituirà l'a priori del gradino successivo nel processo di apprendimento dall'esperienza.

In questo contesto, l'utilizzazione dei dati campionari per derivare una stima puntuale di $\boldsymbol{\theta}$ risulta improprio, infatti, i dati devono servire esclusivamente per procedere all'aggiornamento della conoscenza, che sarà ancora una volta espressa attraverso una distribuzione di probabilità, solo quando la distribuzione a posteriori degenera e si riduce ad un solo punto, cui è associata una probabilità pari ad 1, si prefigura un uso dei dati campionari per la derivazione di un valore puntuale di $\boldsymbol{\theta}$.

Comunque, in diversi contesti operativi può risultare conveniente (o necessario) sintetizzare la distribuzione attraverso un unico indice, la scelta più ragionevole dovrebbe ricadere sul valor di $\boldsymbol{\theta}$ cui è associata la probabilità a posteriori più elevata (la moda della distribuzione), in realtà si ricorre, nella generalità dei casi, al calcolo della media aritmetica e, talvolta, alla mediana.

Ad esempio, nel caso mono-parametrico (un solo parametro caratteristico) per derivare una stima puntuale di θ si può procedere all'applicazione del metodo della massima verosimiglianza ottenendo come risultato la moda della distribuzione a posteriori

$$\tilde{\theta} = \tilde{M}_o = \underset{\theta}{argmax}\ \pi(\theta / \boldsymbol{x}).$$

Alternativamente si può procedere al calcolo della media aritmetica[19] $\hat{\theta}$ o della mediana $\bar{\theta}$.

$$\hat{\theta} = \hat{\mu} = \mathrm{E}\left[g(\theta)\right] = \int_{\theta} \theta \, d\left[\pi(\theta / \boldsymbol{x})\right]$$

$$\hat{\theta} = \hat{M}_e = \int_{\bar{\theta}}^{+\infty} d\left[\pi(\theta / \boldsymbol{x})\right] = \int_{-\infty}^{\bar{\theta}} d\left[\pi(\theta / \boldsymbol{x})\right] = \frac{1}{2}.$$

[19] Ovviamente, oltre alla media aritmetica, si può procedere al calcolo di tutti i momenti di interesse specificando in modo adeguato la funzione $g(\theta)$ nella relazione

$$\mathrm{E}\left[g(\theta)\right] = \int_{\theta} g(\theta) \, d\left[\pi(\theta / \boldsymbol{x})\right].$$

In ogni caso l'operazione di stima puntuale contraddice la logica bayesiana, logica che risulta invece interamente soddisfatta se si procede al computo di una stima per intervallo.

Gli intervalli bayesiani di confidenza, usualmente denominati ***intervalli o regioni*** (nel caso si considerino più parametri) ***di credibilità***, non solo sono coerenti con la logica bayesiana ma risolvono anche alcuni problemi interpretativi.

Come si è avuto modo di sottolineare, quando è stata trattata la stima di intervallo nel contesto classico la quantità $1-\alpha$ che, inizialmente, cioè quando è riferita all'elemento pivotale, è una probabilità, al termine del processo perde tale natura; infatti, non si parla più di livello di probabilità ma di livello di confidenza. Nella stima per intervallo, l'entità casuale è l'intervallo stesso che ha una probabilità pari all' $1-\alpha$ di contenere al suo interno il vero valore di θ (costante incognita), ma una volta ottenuto l'intervallo non ha più senso parlare di probabilità in quanto l'intervallo, o contiene al suo interno il vero valore di θ, allora la probabilità è pari ad 1, o non lo contiene, allora la probabilità e zero. Il termine ***confidenza*** sta ad indicare che si "confida" che l'intervallo ottenuto sia uno degli $(1-\alpha)\%$ degli intervalli che contengono al proprio interno il vero valore di θ.

Se si indica con $C_\alpha(\boldsymbol{x})$ la ***regione di credibilità a posteriori*** a livello $1-\alpha$ per θ si ha

$$\int_{C_\alpha(\boldsymbol{x})} \pi(\theta/\boldsymbol{x}) = 1-\alpha$$

Se in questa espressione si sostituisce alla probabilità a posteriori $\pi(\theta/\boldsymbol{x})$ la probabilità a priori $\pi(\theta)$ si ottiene la ***regione di credibilità a priori*** a livello $1-\alpha$

$$\int_{C_\alpha} \pi(\theta) = 1-\alpha.$$

Dalle considerazioni sopra svolte risulta in modo del tutto evidente che la regione (intervallo) di credibilità non è univocamente individuato, anche in questo caso come sottolineato a proposito degli intervalli di confidenza, l'obiettivo che si vuol perseguire è quello della derivazione della regione più informativa cioè della regione che, al prefissato livello di probabilità $(1-\alpha)$, ha la dimensione più piccola.

La conoscenza della distribuzione a posteriori del parametro θ consente, ovviamente, il calcolo immediato di intervalli di stima (intervalli di confidenza bayesiani o ***intervalli di credibilità***); ad esempio un intervallo al livello di credibilità $(1-\alpha)$ è espresso da qualunque intervallo (L_1, L_2) che soddisfa l'uguaglianza

$$\int_{L_1}^{L_2} d\left[\pi(\theta/\boldsymbol{x})\right] = 1-\alpha.$$

Come nel caso già trattato, tra tutti gli intervalli che soddisfano tale relazione si dovrà scegliere quello maggiormente informativo che nel caso di un solo parametro è rappresentato dall'intervallo di lunghezza minima.

La differenza tra gli intervalli di confidenza e gli intervalli di credibilità è sostanziale; infatti, come già chiarito in precedenza, mentre per gli intervalli di confidenza è l'intervallo stesso (L_1, L_2) che a priori, cioè prima di effettuare la rilevazione campionaria, ha una probabilità dell'*(1 – α) %* di contenere al suo interno il vero valore del parametro Θ, nel caso degli intervalli di credibilità la probabilità è riferita al parametro Θ *(variabile casuale)* che ha una probabilità dell'*(1 – α) %* di essere contenuto nell'intervallo (L_1, L_2).

Riprendendo in considerazione quanto detto nel Capitolo 3 riguardo agli intervalli di confidenza, si può procedere alla determinazione degli intervalli di credibilità senza alcuna difficoltà, infatti, basterà fare riferimento alla distribuzione a posteriori del parametro o dei parametri di interesse. Si sottolinea, ancora una volta, la superiorità degli intervalli di credibilità, rispetto agli intervalli di confidenza, sia dal punto di vista interpretativo che da quello operativo quando sono presenti parametri di disturbo.

Nel caso degli intervalli di confidenza il problema si può risolvere attraverso una stima puntuale del parametro di disturbo che richiede, però la derivazione di una diversa distribuzione campionaria degli estremi dell'intervallo stesso; derivazione che in alcuni casi non presenta alcuna difficoltà, come ad esempio la determinazione degli intervalli di confidenza per la media di una distribuzione normale semplice quando la varianza è incognita (si passa dalla distribuzione normale alla distribuzione t di Student), ma che in altri casi presenta notevoli difficoltà, al riguardo basta citare il caso della determinazione degli intervalli per la differenza tra medie di due distribuzioni normali quando le due corrispondenti varianze non sono note. Come segnalato più volte, nel contesto bayesiano il problema della presenza di parametri di disturbo si risolve attraverso una semplice operazione di marginalizzazione della distribuzione a posteriori.

6.4.2 - Test d'ipotesi

Nel contesto classico di verifica di ipotesi statistiche sono state introdotte due ipotesi, l'ipotesi nulla o ipotesi di lavoro $H_0 : \theta \in \Theta_0$ e l'ipotesi alternativa $H_1 : \theta \in \Theta_1$ dove $\Theta_0 \cup \Theta_1 = \Theta$ e $\Theta_0 \cap \Theta_1 = \varnothing$, fissato un livello di significatività α (probabilità dell'errore di I° tipo, cioè rifiutare un'ipotesi nulla vera) si procede al rifiuto o all'accettazione (non rifiuto) dell'ipotesi nulla a seconda che il punto campionario cada o meno nella regione critica o, alternativamente si procede alla determinazione del *p-value* (probabilità che la variabile casuale test assuma un valore "più estremo" di quello osservato se l'ipotesi nulla è vera) agendo di conseguenza. Nel contesto bayesiano il problema di verifica d'ipotesi diventa banale, infatti, avendo a disposizione la distribuzione a posteriori del parametro/i basterà procedere al computo delle probabilità a posteriori relative alle due ipotesi

$$\pi_{0/\mathbf{x}} = P(\Theta \in \Theta_0 / \mathbf{x})$$
$$\pi_{1/\mathbf{x}} = P(\Theta \in \Theta_1 / \mathbf{x})$$

e procedere alla scelta dell'ipotesi che presenta la probabilità più elevata, cioè scegliere $H_0 : \theta \in \Theta_0$ o $H_1 : \theta \in \Theta_1$ in funzione del valore assunto dal rapporto a posteriori (***posterior odds***)

$$\frac{\pi_{0/x}}{\pi_{1/x}} = \frac{P(\Theta \in \Theta_0 / x)}{P(\Theta \in \Theta_1 / x)}.$$

se favorevole (>1) o meno (< 1) all'ipotesi $H_0 : \theta \in \Theta_0$.

Analogamente al rapporto tra le probabilità a posteriori si può procedere al calcolo del rapporto tra le probabilità a priori (***prior odds***)

$$\frac{\pi_0}{\pi_1} = \frac{P(\Theta \in \Theta_0)}{P(\Theta \in \Theta_1)}.$$

Il rapporto tra gli odds

$$B_0 = \frac{\pi_{0/x}}{\pi_{1/x}} / \frac{\pi_0}{\pi_1} = \frac{\pi_{0/x}\ \pi_1}{\pi_{1/x}\ \pi_0} = \frac{P(\Theta \in \Theta_0 / x)\ P(\Theta \in \Theta_1)}{P(\Theta \in \Theta_1 / x)\ P(\Theta \in \Theta_0)}$$

viene detto ***fattore di Bayes*** in favore dell'ipotesi $H_0 : \theta \in \Theta_0$, ovviamente il fattore di Bayes in favore dell'ipotesi $H_1 : \theta \in \Theta_1$ è espresso da

$$B_1 = 1 / B_0 = \frac{\pi_{1/x}\ \pi_0}{\pi_{0/x}\ \pi_1} = \frac{P(\Theta \in \Theta_1 / x)\ P(\Theta \in \Theta_0)}{P(\Theta \in \Theta_0 / x)\ P(\Theta \in \Theta_1)}.$$

Procedere nell'accettazione o al rifiuto di una specifica ipotesi $H_0 : \theta \in \Theta_0$ contro l'ipotesi alternativa $H_1 : \theta \in \Theta_1$ in funzione del valore assunto dalle probabilità a posteriori delle due ipotesi appare del tutto ragionevole, in realtà tale approccio presenta degli inconvenienti di natura tutt'altro che marginale. Ad esempio, nel caso di un ipotesi nulla semplice $H_0 : \theta \in \theta_0$ contro l'ipotesi alternativa composita bidirezionale $H_0 : \theta \neq \theta_0$, la procedura è inapplicabile essendo pari a *0* la probabilità a posteriori dell'ipotesi nulla

$$\pi_{0/x} = P(\Theta \in \theta_0 / x) = \int_{\theta_0}^{\theta_0} f(x/\theta)\ \pi(\theta)\ d\theta = 0$$

in quanto $\pi_0 = P(\Theta = \theta_0) = 0$.

Il problema si può risolvere o tenendo presente la relazione che tra intervalli di stima e test delle ipotesi, procedendo all'accettazione se θ_0 ricade nell'intervallo di credibilità calcolato per la v.c. Θ, ma la procedura non ha più la natura di test d'ipotesi, oppure assegnando all'ipotesi nulla una probabilità a priori maggiore di *0* $\left[\pi(\theta_0) = P(\Theta = \theta_0) > 0\right]$, cioè inserendo una probabilità a priori mistura tra una v.c. discreta ed una v.c. continua.

Un modo alternativo per risolvere il problema di scelta dell'ipotesi è quello di fare riferimento al valore assunto dal fattore di bayes. Procedura questa che, pur non risolvendo il problema di scelta tra un'ipotesi nulla semplice ed un'ipotesi alternativa composita, presenta, come si avrà modo di chiarire nelle righe successive, indubbi vantaggi.

La procedura basata sul fattore di Bayes non presenta delle carenze anche quando entrambe le ipotesi sono semplici. Infatti, se entrambe le ipotesi sono semplici si ha $\pi_{0/\boldsymbol{x}} \propto \pi_0 \, f\left(x/\boldsymbol{\theta}_0\right)$ e $\pi_{1/\boldsymbol{x}} \propto \pi_1 \, f\left(x/\boldsymbol{\theta}_1\right)$ pertanto il fattore di Bayes

$$B_0 = \frac{\pi_{0/\boldsymbol{x}}}{\pi_{1/\boldsymbol{x}}} / \frac{\pi_0}{\pi_1} = \frac{\pi_{0/\boldsymbol{x}} \; \pi_1}{\pi_{1/\boldsymbol{x}} \; \pi_0} = \frac{f\left(x/\boldsymbol{\theta}_0\right) \; \pi_0 \; \pi_1}{f\left(x/\boldsymbol{\theta}_1\right) \; \pi_1 \; \pi_0} = \frac{f\left(x/\boldsymbol{\theta}_0\right)}{f\left(x/\boldsymbol{\theta}_1\right)}$$

si riduce al rapporto tra le due verosimiglianze. Risultato questo che, se per un verso può soddisfare i critici dell'approccio bayesiano, per altro verso non può soddisfare i fautori dell'approccio bayesiano soggettivo in quanto implica una eliminazione "meccanica" della conoscenza a priori[20].

L'utilità del ricorso al fattore di Bayes emerge in modo evidente quando al problema di scelta delle ipotesi viene attribuita la valenza di scelta tra modelli alternativi di rappresentazione della realtà fenomenica.

Se con $\boldsymbol{M}$ si indica un generico modello capace di rappresentare il fenomeno oggetto di analisi, l'ipotesi $H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ può essere interpretata anche come $H_0 : M = M_0$, cioè l'ipotesi che il modello rappresentativo della realtà sia proprio $M_0 \in \boldsymbol{\Xi}$, dove $\boldsymbol{\Xi}$ rappresenta lo spazio contenente tutti i possibili modelli rappresentativi del fenomeno oggetto d'analisi, mentre l'ipotesi $H_1 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_1$ resta specificata da $H_1 : M = M_1$ con $M_1 \in \boldsymbol{\Xi}$, il fattore di Bayes assume la forma

$$B_0 = \frac{P\left(M = M_0 / \boldsymbol{x}\right) \; P\left(M = M_1\right)}{P\left(M = M_1 / \boldsymbol{x}\right) \; P\left(M = M_0\right)} = \frac{\int_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} f\left(\boldsymbol{x}/\boldsymbol{\theta}\right) \; \pi_0\left(\boldsymbol{\theta}\right) \; d\boldsymbol{\theta}}{\int_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_1} f\left(\boldsymbol{x}/\boldsymbol{\theta}\right) \; \pi_1\left(\boldsymbol{\theta}\right) \; d\boldsymbol{\theta}}.$$

Il fattore di Bayes, che è definito dal rapporto ponderato delle verosimiglianze dei due modelli, misura la capacità relativa del modello M_0 rispetto al modello M_1, di rappresentare la realtà; proprietà, questa, indubbiamente apprezzabile. Per contro, la scelta del modello basata sul confronto tra le probabilità a posteriori solleva delle perplessità soprattutto se si tiene conto di quanto riportato nella premessa a queste Note: ***tutti i modelli sono sbagliati (hanno quindi probabilità 0 di essere veri) ma qualcuno è utile***; ovviamente, l'utilità è strettamente condizionata dalla sua capacità rappresentativa della realtà[21].

---

[20] Diversa è la situazione quando la conoscenza a priori perde di rilevanza a ragione dell'acquisizione di evidenza empirica (campionaria oggettiva) sempre più estesa. Al riguardo se segnala la convergenza tra risultati bayesiani e quelli classici al crescere della dimensione campionaria, si dimostra, infatti, l'equivalenza asintotica dei due approcci.

[21] In letteratura è stata proposta una regola pratica per interpretare il valore numerico assunto dal fattore di Bayes:

se $B_0 \geq 1$ l'evidenza (a priori e campionaria) supporta il modello M_0;

se $10^{-1/2} \leq B_0 < 1$ l'evidenza contro il modello M_0 è minima;

se $10^{-1} \leq B_0 < 10^{-1/2}$ l'evidenza contro il modello M_0 è sostanziale;

se $10^{-2} \leq B_0 < 10^{-1}$ l'evidenza contro il modello M_0 è molto elevata;

se $B_0 < 10^{-2}$ l'evidenza contro il modello M_0 è decisiva.

Se i modelli alternativi non sono 2 $(M_0$ e $M_1)$ ma s $(M_i$ per $i = 1, 2, ..s)$ il fattore di Bayes potrà essere calcolato per $s(s-1)/2$ confronti tra modelli. Calcolo questo non necessario per operare la scelta del modello, infatti, per perseguire tale finalità basterà operare $s - 1$ confronti: si calcola il fattore di Bayes per i modelli M_1 e M_2, il modello migliore viene confrontato co il modello M_3 e cosi via fino al confronto tra il modello M_s ed il modello risultante dal processo di selezione che ha evidenziato la maggiore capacità rappresentativa.

6.5 - Regressione bayesiana

Come illustrato nel capitolo precedente nel modello di regressione lineare multipla si studia la relazione tra una variabile spiegata (variabile dipendente) y e $k-1$ $(k \geq 2)$ variabili esplicative. Il modello è espresso dalla relazione

$$y_i = \beta_1 + \beta_2 \cdot x_{i2} + \beta_3 \cdot x_{i3} + \cdots + \beta_k \cdot x_{ik} + u_i \quad \text{per} \quad i = 1,2,....,n$$

che in forma matriciale diventa

$$\underset{n,1}{\boldsymbol{y}} = \underset{n,k}{\boldsymbol{X}}\,\underset{k,1}{\boldsymbol{\beta}} + \underset{n,1}{\boldsymbol{u}}$$

dove

$$\underset{n,1}{\boldsymbol{y}} = \begin{bmatrix} y_1 \\ y_2 \\ .... \\ .... \\ y_i \\ .... \\ .... \\ y_n \end{bmatrix} \quad ; \quad \underset{n,k}{\boldsymbol{X}} = \begin{bmatrix} 1 & x_{12} & x_{13} \cdots & x_{1j} \cdots & x_{1k} \\ 1 & x_{22} & x_{23} \cdots & x_{2j} \cdots & x_{2k} \\ & & ..... & & \\ & & ..... & & \\ 1 & x_{i2} & x_{i3} \cdots & x_{ij} \cdots & x_{ik} \\ & & ..... & & \\ & & ..... & & \\ 1 & x_{n2} & x_{n3} \cdots & x_{nj} \cdots & x_{nk} \end{bmatrix} \quad ; \quad \underset{k,1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_1 \\ \beta_2 \cdot \\ .... \\ .... \\ \beta_j \\ .... \\ .... \\ \beta_k \end{bmatrix} \quad ; \quad \underset{n,1}{\boldsymbol{u}} = \begin{bmatrix} u_1 \\ u_2 \cdot \\ .... \\ .... \\ u_i \\ .... \\ .... \\ u_n \end{bmatrix}$$

Le ipotesi di specificazione poste alla base del modello sono:

1. la matrice $\boldsymbol{X}_{nxk}$ è costituita da variabili matematiche o determinazioni di variabili casuali, sono cioè costanti in ripetuti campioni; in particolare il primo vettore colonna della matrice è costituito da 1, il coefficiente β_1 rappresenta, pertanto, l'intercetta dell'iperpiano di regressione;
2. la matrice $\boldsymbol{X}_{nxk}$ è di rango massimo = $k \leq n$;
3. il vettore $\boldsymbol{u}_{nx1}$ ha componenti aleatorie con valore atteso nullo ($E(\boldsymbol{u}_{n\times 1}) = \boldsymbol{0}_{n\times 1}$), varianza costante (omoschedasticità $Var(u_i) = E(u_i^2) = 0$) e risultano incorrelate ($E(u_i \cdot u_j) = 0$ per i ≠ j), in forma matriciale

$$Var\left(\underset{n,1}{\boldsymbol{u}}\right) = \underset{n,n}{\boldsymbol{\Sigma}_u} = E\left(\underset{n,1}{\boldsymbol{u}} \cdot \underset{1,n}{\boldsymbol{u}'}\right) = \sigma^2 \cdot \boldsymbol{I}_n$$

4. ipotesi di **normalità del vettore casuale**

$$\boldsymbol{u} \sim N\left(\boldsymbol{0}, \sigma^2 \boldsymbol{I}\right).$$

Se le prime tre ipotesi sono soddisfatte, si possono derivare le stime dei minimi quadrati $\hat{\boldsymbol{\beta}}$ del vettore $\boldsymbol{\beta}$ che sono date da:

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{y}$$

dove l'apice riportato ad esponente vuole indicare che si sta facendo riferimento alla matrice trasposta. Si ricorda che tali stime sono le migliori (minimizzano l'errore quadrato medio) nell'ambito delle stime lineari e corrette (***BLU- Best Linear Unbiased***).

La stima corretta della varianza σ^2 è data da:

$$\hat{\sigma}^2 = \left[\left(\boldsymbol{y} - \hat{\boldsymbol{y}}\right)'\left(\boldsymbol{y} - \hat{\boldsymbol{y}}\right)\cdot\right] / \left(n-k\right)) = \left[\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)\cdot\right] / \left(n-k\right)) =$$

$$= \left\{\boldsymbol{y}'\left[\boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right]'\boldsymbol{y}\right\} / \left(n-k\right)) = \left(\hat{\boldsymbol{u}}'\cdot\hat{\boldsymbol{u}}\right) / \left(n-k\right)) = S^2$$

Se si introduce l'ipotesi di normalità si può calcolare la verosimiglianza

$$L\left(\boldsymbol{\beta}, \sigma^2\right) = f\left(\boldsymbol{\beta}, \sigma^2 / \boldsymbol{y}, \boldsymbol{X}\right) = \prod_{i=1}^{n} f\left(y_i\right) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2\right] =$$

$$= \left(2\pi\sigma^2\right)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)'\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)\right]$$

La stima di massima verosimiglianza $\tilde{\boldsymbol{\beta}}$ del vettore $\boldsymbol{\beta}$ è identica alle stime dei minimi quadrati:

$$\tilde{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y} = \hat{\boldsymbol{\beta}}$$

ma, come già richiamato nel capitolo precedente le stime di massima verosimiglianza sono le migliori nell'ambito delle stime corrette (***BU- Best Unbiased***). Inoltre valgono le proprietà degli stimatori sotto elencate:

- l'ipotesi di incorrelazione tra le componenti accidentali $u_i (i = 1, 2, ..., n)$ implica l'indipendenza, ne consegue quindi l'indipedenza tra le componenti $y_i (i = 1, 2, ..., n)$ del vettore casuale $\boldsymbol{y}$, inoltre:
- $\tilde{\boldsymbol{\beta}} \sim N\left[\boldsymbol{\beta}, \sigma^2\left(\boldsymbol{X}'\cdot\boldsymbol{X}\right)^{-1}\right]$
- $\tilde{\boldsymbol{y}} \sim N\left(\boldsymbol{X}\cdot\boldsymbol{\beta}, \sigma^2\left(\boldsymbol{X}'\cdot\boldsymbol{X}\right)^{-1}\right)$
- $W = \left(n-k\right)\cdot\hat{\sigma}^2 / \sigma^2 \sim \chi^2_{n-k}$
- Il vettore casuale $\tilde{\boldsymbol{\beta}}$ e la v.c. *W* sono indipendenti.

Questi risultati consentono di procedere alla determinazione degli intervalli di confidenza per i parametri incogniti β_i *(i = 1, 2,……,k)* e σ^2, gli intervalli di previsione in corrispondenza ad una specifica determinazione del vettore delle variabili esplicative $\boldsymbol{x}_p$ e di procedere alla verifica di ipotesi statistiche. Al riguardo si ricorda

che se la varianza σ^2 non è nota basterà fare riferimento alla variabile t di Student, con n-k gradi di libertà, anziché alla variabile normale.

Quanto sopra richiamato fa riferimento all'impostazione classica dell'inferenza statistica nel cui contesto i parametri sono costanti incognite da stimare e/o sui quali verificare ipotesi statistiche utilizzando soltanto l'informazione campionaria a disposizione.

6.5.1 Regressione bayesiana con distribuzioni a priori non informative e coniugate

Nell'impostazione bayesiana, i parametri β_i *(i = 1, 2,……,k)* e σ^2, essendo entità incognite, assumono la natura di variabili casuali con una propria distribuzione di probabilità.

La verosimiglianza sopra introdotta soddisfa la relazione

$$L\left(\boldsymbol{\beta},\sigma^2\right) = f\left(\boldsymbol{\beta},\sigma^2 / \boldsymbol{y},\boldsymbol{X}\right) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\left(\boldsymbol{y}-\boldsymbol{X\beta}\right)'\left(\boldsymbol{y}-\boldsymbol{X\beta}\right)\right] =$$

$$= \left(2\pi\sigma^2\right)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\left(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{X\hat{\beta}}+\boldsymbol{X\hat{\beta}}\right)'\left(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{X\hat{\beta}}+\boldsymbol{X\hat{\beta}}\right)\right] =$$

$$= \left(2\pi\sigma^2\right)^{-n/2} \exp-\left\{\frac{1}{2\sigma^2}\left[\left(\boldsymbol{y}-\boldsymbol{X\hat{\beta}}\right)'\left(\boldsymbol{y}-\boldsymbol{X\hat{\beta}}\right)-2\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\boldsymbol{X}'\left(\boldsymbol{y}-\boldsymbol{X\hat{\beta}}\right)+\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right]\right\} =$$

$$= \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-k)S^2+\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right]\right\}$$

poiché

$$\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\boldsymbol{X}'\left(\boldsymbol{y}-\boldsymbol{X\hat{\beta}}\right) = \left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{X}'\boldsymbol{y}-\boldsymbol{X}'\boldsymbol{X\hat{\beta}}\right) = \left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{X}'\boldsymbol{X\hat{\beta}}-\boldsymbol{X}'\boldsymbol{X\hat{\beta}}\right) = \left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\boldsymbol{0} = 0$$

dove le statistiche $\hat{\boldsymbol{\beta}}$ e S^2, stime corrette di $\boldsymbol{\beta}$ e σ^2, sono congiuntamente sufficienti.

Distribuzioni a priori non informative

La procedura standard è per l'introduzione di distribuzioni a priori non informative prevede le seguenti distribuzioni

$$\pi\left(\boldsymbol{\beta}\right) \propto c_1 (\textit{costante})$$

inoltre, ponendo $\psi = log\ \sigma^2$ e $\pi\left(\psi\right) \propto c_2$ $(\textit{costante})$, tenendo conto che lo Jacobiano della trasformazione da ψ a σ^2 è pari a σ^{-2} si ha $\pi\left(\sigma^2\right) \propto \sigma^{-2}$, quindi

$$\pi\left(\boldsymbol{\beta},\sigma^2\right) \propto \sigma^{-2} \quad \text{per } \sigma^2 > 0$$ [22]

[22] Si tratta di una distribuzione a priori impropria che genera, comunque, una distribuzione a posteriori propria. Da sottolineare che anche se viene etichettata come non informativa, in realtà implica che la probabilità a priori associata a $\boldsymbol{\beta}$, qualunque sia il suo valore, possa essere anche molto elevata.

Se si procede al prodotto di questa quantità (probabilità a priori) con la verosimiglianza si ha la distribuzione a posteriori

$$\pi\left(\boldsymbol{\beta},\sigma^2 / \boldsymbol{y},\boldsymbol{X}\right) = L\left(\boldsymbol{\beta},\sigma^2\right)\ \pi\left(\boldsymbol{\beta},\sigma^2\right) \propto \left(\sigma^2\right)^{-(n+2)/2} \exp\left[-\frac{1}{2\sigma^2}\left(\boldsymbol{y}-\boldsymbol{X\beta}\right)'\left(\boldsymbol{y}-\boldsymbol{X\beta}\right)\right] =$$

ed anche

$$\pi\left(\boldsymbol{\beta},\sigma^2 / \boldsymbol{y},\boldsymbol{X}\right) \propto \left(\sigma^2\right)^{-(n+2)/2} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-k)S^2 + \left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right]\right\} =$$

$$= \left(\sigma^2\right)^{-[(n-k)/2]-1} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-k)S^2\right]\right\}\ \left(\sigma^2\right)^{-k/2} \exp\left\{-\frac{1}{2\sigma^2}\left[\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right]\right\}$$

Se si fa riferimento all'ultimo membro della relazione si evince immediatamente la forma della distribuzione condizionata a posteriori del vettore $\boldsymbol{\beta}$ e la distribuzione marginale a posteriori del parametro σ^2

$$\boldsymbol{\beta} / \sigma^2, \boldsymbol{y} \sim N\left[\hat{\boldsymbol{\beta}}, \sigma^2\left(\boldsymbol{X'X}\right)^{-1}\right]$$

$$\sigma^2 / \boldsymbol{y} \sim Inv\boldsymbol{\Gamma}\left[\frac{n-k}{2}, \frac{(n-k)S^2}{2}\right]$$

dove il simbolo $Inv\boldsymbol{\Gamma}$ sta ad indicare la variabile casuale Gamma inversa.

Senza eccessiva difficoltà si deriva anche la distribuzione marginale a posteriori di $\boldsymbol{\beta}$, infatti

$$\pi\left(\boldsymbol{\beta} / \boldsymbol{y},\boldsymbol{X}\right) = \int_0^\infty \pi\left(\boldsymbol{\beta},\sigma^2 / \boldsymbol{y}\right) d\sigma^2 \propto$$

$$\propto \int_0^\infty \left(\sigma^2\right)^{-(n+2)/2} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-k)S^2 + \left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right]\right\} d\sigma^2$$

ma l'espressione sotto il segno di integrale rappresenta, a meno della costante moltiplicativa l'espressione della funzione di densità di una variabile casuale Gamma inversa

$$Inv\boldsymbol{\Gamma}\left\{\frac{n}{2},\ 2\left[(n-k)S^2 + \left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right]\right\}$$

pertanto, il valore dell'integrale è, semplicemente, dato dal reciproco della costante di normalizzazione della densità di una $Inv\boldsymbol{\Gamma}(\alpha,\beta)$ che è pari a $\boldsymbol{\Gamma}(\alpha)\beta^\alpha$, dove $\alpha = n/2$ e

$$\beta = 2\left[(n-k)S^2 + \left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right].$$

Se si pone $\nu = n-k$ si ha

$$\pi\left(\boldsymbol{\beta} / \boldsymbol{y},\boldsymbol{X}\right) \propto \left[(n-k)S^2 + \left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right]^{-(n-k+k)/2} \propto$$

$$\propto \left[\nu + \left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'S^{-2}\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right]^{-(\nu+k)/2}$$

che rappresenta, a meno della costante moltiplicativa, l'espressione della funzione di

densità di una variabile casuale t di Student multivariata[23], cioè:

$$\beta / \boldsymbol{y} \sim t_k \left[\nu, \hat{\boldsymbol{\beta}}, S^2 \left(\boldsymbol{X}' \boldsymbol{X} \right)^{-1} \right].$$

Se si ricorre alla distribuzione a priori di Jeffreys definita da

$$\pi_R \left(\boldsymbol{\beta}, \sigma^2 \right) \propto \sigma^{-(k+2)/2}$$

si deriva la distribuzione a posteriori congiunta

$$\pi \left(\boldsymbol{\beta}, \sigma^2 / \boldsymbol{y}, \boldsymbol{X} \right) \propto \frac{1}{\left(\sigma^2 \right)^{(n+k+2)/2}} \exp \left\{ -\frac{1}{2\sigma^2} \left[(n-k) S^2 + \frac{c}{1+c} \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}} \right)' \boldsymbol{X}' \boldsymbol{X} \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}} \right) \right] \right\}$$

mentre le distribuzioni marginali a posteriori di $\boldsymbol{\beta}$ e σ^2 hanno la forma

$$\beta / \sigma^2, \boldsymbol{y}, \boldsymbol{X} \sim N \left[\hat{\boldsymbol{\beta}}, \sigma^2 \left(\boldsymbol{X}' \boldsymbol{X} \right)^{-1} \right]$$

$$\sigma^2 / \boldsymbol{y}, \boldsymbol{X} \sim Inv\Gamma \left[n/2, S^2 (n-k)/2 \right].$$

Distribuzione a priori informative (coniugate)

Riprendendo in considerazione e generalizzando quanto riportato nell'esempio 6.5 riguardo alla distribuzione coniugata a priori di una v.c. normale, una possibile specificazione della distribuzione a priori nel caso in esame

$$\pi \left(\boldsymbol{\beta}, \sigma^2 \right) = \pi \left(\boldsymbol{\beta} / \sigma^2 \right) \cdot \pi \left(\sigma^2 \right)$$

è la distribuzione coniugata congiunta di una normale e una gamma inversa

$$\left(\boldsymbol{\beta} / \sigma^2 \right) \sim N \left(\boldsymbol{\beta}_*, \sigma^2 \boldsymbol{\Sigma}_\beta \right) \text{ e } \sigma^2 \sim Inv\Gamma \left(\alpha, \delta \right)$$

pertanto la distribuzione a priori assume la forma

$$\pi \left(\boldsymbol{\beta}, \sigma^2 \right) = \pi \left(\boldsymbol{\beta} / \sigma^2 \right) \pi \left(\sigma^2 \right) = \frac{e^{-\frac{1}{2\sigma^2} (\boldsymbol{\beta} - \boldsymbol{\beta}_*)' \boldsymbol{\Sigma}_\beta^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_*)}}{(2\pi\sigma^2)^{1/2} \left| \boldsymbol{\Sigma}_\beta \right|^{1/2}} \cdot \frac{\left(\sigma^2 \right)^{-\alpha-1} \delta^\alpha \, e^{-\frac{\delta}{\sigma^2}}}{\Gamma(\alpha)}$$

cioè $\left(\boldsymbol{\beta}, \sigma^2 \right) \sim N \left(\boldsymbol{\beta}_*, \sigma^2 \boldsymbol{\Sigma}_\beta \right) \otimes Inv\Gamma(\alpha, \delta)$ che è distribuzione a priori coniugata di una v.c. che appartiene alla stessa famiglia. Infatti, se si considera la verosimiglianza

$$L \left(\boldsymbol{\beta}, \sigma^2 \right) = \left(2\pi\sigma^2 \right)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \left[(n-k) S^2 + \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}} \right)' \boldsymbol{X}' \boldsymbol{X} \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}} \right) \right] \right\}$$

[23] La funzione di densità di probabilità t di Student multivariata di un vettore casuale $\boldsymbol{V}$ di dimensione k è data da

$$f \left(\boldsymbol{w}; \nu, \boldsymbol{\mu}, \boldsymbol{\Sigma} \right) = \frac{\Gamma \left[(\nu + k)/2 \right] \left[1 + \frac{1}{\nu} (\boldsymbol{w} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{w} - \boldsymbol{\mu}) \right]^{-(\nu+k)/2}}{\Gamma(\nu/2) \; \nu^{k/2} \; \pi^{k/2} \left| \boldsymbol{\Sigma} \right|^{1/2}}$$

dove ν rappresentano i gradi di libertà, $\boldsymbol{\mu}$ è un vettore di dimensione k (parametri di locazione) e $\boldsymbol{\Sigma}$ è una matrice simmetrica definita positiva di dimensione k (parametri di scala).

la distribuzione a posteriori congiunta è data da

$$\pi\left(\boldsymbol{\beta},\sigma^2 / \boldsymbol{y}\right)=\frac{\pi\left(\boldsymbol{\beta},\sigma^2\right)L\left(\boldsymbol{\beta},\sigma^2 / \boldsymbol{x}\right)}{f\left(\boldsymbol{y}\right)}=\frac{\pi\left(\boldsymbol{\beta} / \sigma^2\right)\pi\left(\sigma^2\right)L\left(\boldsymbol{\beta},\sigma^2 / \boldsymbol{x}\right)}{f\left(\boldsymbol{y}\right)}=$$

$$=\frac{e^{-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}-\boldsymbol{\beta}_*\right)'\boldsymbol{\Sigma}_\beta^{-1}\left(\boldsymbol{\beta}-\boldsymbol{\beta}_*\right)}}{\left(2\pi\sigma^2\right)^{k/2}\left|\boldsymbol{\Sigma}_\beta\right|^{k/2}}\cdot\frac{\left(\sigma^2\right)^{-\alpha-1}\delta^\alpha\, e^{-\frac{\delta}{\sigma^2}}}{\Gamma\left(\alpha\right)}\cdot$$

$$\cdot\left(2\pi\sigma^2\right)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\left[\left(n-k\right)S^2+\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right]\right\}/\, f\left(\boldsymbol{y}\right)$$

ma

$$\left(\boldsymbol{\beta}-\boldsymbol{\beta}_*\right)'\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\left(\boldsymbol{\beta}-\boldsymbol{\beta}_*\right)+\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)=$$

$$=\left(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}}\right)'\bar{\boldsymbol{\Sigma}}^{-1}\left(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}}\right)+\left(\boldsymbol{\beta}_*-\hat{\boldsymbol{\beta}}\right)'\boldsymbol{\Sigma}_*^{-1}\left(\boldsymbol{\beta}_*-\hat{\boldsymbol{\beta}}\right)$$

dove

$$\bar{\boldsymbol{\beta}}=\left(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}+\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left[\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\beta}_*+\left(\boldsymbol{X}'\boldsymbol{X}\right)\hat{\boldsymbol{\beta}}\right]^{-1}$$

$$\bar{\boldsymbol{\Sigma}}=\left(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}+\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$$

$$\boldsymbol{\Sigma}_*=\boldsymbol{\Sigma}_{\boldsymbol{\beta}}+\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$$

si ha

$$\pi\left(\boldsymbol{\beta},\sigma^2 / \boldsymbol{y}\right)=\frac{\pi\left(\boldsymbol{\beta},\sigma^2\right)L\left(\boldsymbol{\beta},\sigma^2 / \boldsymbol{y}\right)}{f\left(\boldsymbol{y}\right)}=\frac{\pi\left(\boldsymbol{\beta} / \sigma^2\right)\pi\left(\sigma^2\right)L\left(\boldsymbol{\beta},\sigma^2 / \boldsymbol{y}\right)}{f\left(\boldsymbol{y}\right)}\propto$$

$$\propto exp\left[-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}}\right)'\bar{\boldsymbol{\Sigma}}^{-1}\left(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}}\right)\right]\cdot$$

$$\cdot\left(\sigma^2\right)^{-n/2-\alpha-1}exp\left\{-\frac{1}{\sigma^2}\frac{\delta}{2}\left[\left(n-k\right)S^2+\left(\boldsymbol{\beta}_*-\hat{\boldsymbol{\beta}}\right)'\boldsymbol{\Sigma}_*^{-1}\left(\boldsymbol{\beta}_*-\hat{\boldsymbol{\beta}}\right)\right]\right\}=$$

$$=exp\left[-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}}\right)'\bar{\boldsymbol{\Sigma}}^{-1}\left(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}}\right)\right]\cdot\left(\sigma^2\right)^{-\alpha_*-1}exp\left(-\frac{\delta_*}{\sigma^2}\right)$$

dove $\alpha_*=n/2-\alpha$ e $\delta_*=\frac{\delta}{2}\left[\left(n-k\right)S^2+\left(\boldsymbol{\beta}_*-\hat{\boldsymbol{\beta}}\right)'\boldsymbol{\Sigma}_*^{-1}\left(\boldsymbol{\beta}_*-\hat{\boldsymbol{\beta}}\right)\right]$, quindi

$$\left(\boldsymbol{\beta},\sigma^2 / y\right)\sim NInvI\left(\bar{\boldsymbol{\beta}},\sigma^2\bar{\boldsymbol{\Sigma}};\alpha_*,\delta_*\right)$$

che appartiene alla stessa famiglia della distribuzione a priori normale gamma inversa.

Una proposta alternativa di distribuzione a priori informativa, molto utilizzata nel contesto econometrico, è quella suggerita da ***Zellner*** nel 1986, usualmente denominata ***G-prior*** , proposta che si differenzia dalla a-priori non informativa sopra illustrata per l'a-priori su $\boldsymbol{\beta}$.

Le due distribuzioni a priori sono

$$\pi\left(\sigma^2\right) \propto 1/\sigma^2$$

$$\boldsymbol{\beta} \sim N\left[\boldsymbol{\beta}_0,\ g\ \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right].$$

Attraverso passaggi algebrici analoghi a quelli sopra riportati si deriva la distribuzione congiunta a posteriori

$$\pi\left(\boldsymbol{\beta},\sigma^2 / \boldsymbol{y}\right) \propto \frac{1}{\left(\sigma^2\right)^{(n+2)/2}} \exp\left\{-\frac{1}{2\sigma^2}\left[\left(n-k\right)S^2 + \frac{g}{1+g}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)' \boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right]\right\}$$

mentre la distribuzione marginale a posteriori del vettore $\boldsymbol{\beta}$ è

$$\left(\beta / \boldsymbol{y}\right) \sim t_k\left[\nu, \frac{1}{g+1}\left(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}\right), \frac{g\left[S^2 + \left(\boldsymbol{\beta}_0-\hat{\boldsymbol{\beta}}\right)' \boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{\beta}_0-\hat{\boldsymbol{\beta}}\right)/\left(g+1\right)\right]}{n\left(g+1\right)}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right]$$

cioè, una v.c t di Student multivariata di dimensione k .

I risultati riportati nelle righe precedenti consentono la risoluzione dei problemi di stima puntuale, stima d'intervallo e di test delle ipotesi seguendo la procedura già illustrata. Si sottolinea che nel contesto della regressione multipla assumono particolare rilevanza, sia i temi connessi alla scelta del modello più appropriato (quello che evidenzia la capacità rappresentativa più elevata della realtà sotto esame), nel cui ambito è ricompresa anche la problematica relativa alla selezione delle variabili esplicative da includere nel modello stesso, sia i temi collegati all'impiego del modello a fini previsionali, previsioni che potranno essere effettuate utilizzando la distribuzione predittiva a posteriori di Y.