UNIVERSITÀ DEGLI STUDI DI ROMA TRE
FACOLTÀ DI SCIENZE M.F.N.

Studio computazionale dell'algoritmo IPF Bayesiano

Sintesi della tesi di Laurea in Matematica
di Sandra Rizzello
Relatore: Prof. Mauro Piccioni
Correlatore: Prof. Gianpaolo Scalia Tomba

Metodi Monte Carlo nella statistica bayesiana

Metodi Monte Carlo nella statistica bayesiana

L'IPF (*iterative proportional fitting*) Bayesiano è un metodo **MCMC** (Monte Carlo per catene di Markov): i metodi Monte Carlo vengono usati, nella statistica bayesiana, per calcolare valori attesi rispetto a distribuzioni a posteriori su spazi potenzialmente grandi.

L'idea di base dei metodi **MCMC** é chiara e semplice: si desidera generare variabili aleatorie da una distribuzione π su E, ma non lo si puó fare direttamente, o perché la dimensione dello spazio è grande, o perché π è complicata; allora si costruisce una catena di Markov avente come unica distribuzione invariante proprio π e si simula l'andamento della catena. Da

ció deriva che

$$\frac{\sum_{i=1} nf(X_i)}{n} \to \int f\, d\pi,\ \forall f \in L^1(\pi),$$

quando n tende all'infinito, con probabilitá uno, per $\pi-$ quasi ogni punto iniziale X_0.
La variante di **MCMC** da noi studiata, detta Gibbs sampler, si basa sulla simulazione dalla distribuzione condizionata, infatti, se X é una variabile aleatoria a valori in E con distribuzione π, definendo $Y = f(X)$ e $Q(y, A) = P\{X \in A \mid Y = y\}$, la quantitá $P(x, A) = Q(f(x), A)$ é un nucleo di transizione su E con distribuzione invariante π.
Se $f_i(x)$, $i = 1, \cdots m$ è una generica funzione, allora $P_i(x, A) = Q_i(f_i(x), A)$, ha distribuzione invariante π, per ogni i.
Chiameremo, allora, Gibbs sampler sequenziale (**GSS**) una catena di Markov il cui nucleo é $P = P_k...P_1$, Gibbs sampler reversibilizzato (**GSR**) una catena in cui $P = P_1...P_{k-1}P_kP_{k-1}...P_2P_1$ e Gibbs sampler con random scan (**GSRS**) se $P = \frac{1}{k}\sum_{i=1}^{k} P_i$
In particolare, se E é un sottoinsieme di uno spazio prodotto $E_1 \times \cdots \times E_m$ e se $A_1, \cdots, A_k$ sono dei blocchi di indici tali che

$$A_1 \cup \cdots \cup A_k = \{1, \cdots, m\},$$

poniamo

$$f_i(x) = x_{A_i^c} \quad i = 1, ..., k.$$

Indicando con π_i la legge di X_{A_i} condizionata a $X_{A_i^c}$, possiamo scrivere

$$Q_i(x_{A_i^c}, A) = \int_{x_{A_i}:\,(x_{A_i}, x_{A_i^c}) \in A} \pi_i(dx_{A_i} \mid x_{A_i^c}).$$

Dalla definizione di Gibbs sampler, segue, ad esempio, che se partiamo da un valore iniziale $X^0 = (x^0_{A_1}, \cdots, x^0_{A_k})$ e generiamo
$x^1_{A_1}$ da $\pi_1(dx_{A_1} \mid x^0_{A_1^c})$
$x^1_{A_2}$ da $\pi_2(dx_{A_2} \mid x^1_{A_1}, x^0_{A_3}, \cdots, x^0_{A_k})$
$\vdots$
$x^1_{A_k}$ da $\pi_k(dx_{A_k} \mid x^1_{A_1}, x^1_{A_2}, \cdots, x^1_{A_{k-1}})$,
definiamo una transizione da $X^0 = (x^0_{A_1}, \cdots, x^0_{A_k})$ a $X^1 = (x^1_{A_1}, \cdots, x^1_{A_k})$ e

iterando questo procedimento generiamo una sequenza X^0, X^1, $\cdots$, $X^t, \cdots$ che ha come nucleo di transizione $P = P_t \cdots P_1$, dove

$$P_i(x, dy) = \pi_i \left(dy_{A_i} \mid x_{A_j} \left(j > i\right), y_{A_j} \left(j < i\right)\right) \delta_{x_{A_i^c}}(dy_{A_i^c}), \quad \forall\, i = 1, \cdots, m.$$

(Abbiamo indicato con X le variabili ancora da aggiornare e con Y quelle giá aggiornate), cioé

$$P(x, dy) = \prod_{i=1}^{k} \pi_i(y_{A_i} \mid x_{A_j} \left(j > i\right), y_{A_j} \left(j < i\right)) \delta_{x_{A_i^c}}(dy_{A_i^c}).$$

Per la simulazione non basta che π sia invariante, ma é essenziale che sia unica come distribuzione invariante.
Noi faremo delle assunzioni:

1. π ha densità positiva ovunque rispetto alla misura di Lebesgue che denotiamo ancora con $\pi(x_1, x_2, \cdots, x_m)$.

2. Se A è un blocco di indici, scriveremo $\pi(x_A, x_{-A})$ e per ogni valore fissato di $x_{-A} = y_{-A}$ si ha $\pi(x_A, y_{-A})$ integrabile con integrale positivo e quindi $\pi_A(x_A \mid y_{-A}) = \frac{\pi(x_A, y_{-A})}{\int \pi(x_A, y_{-A})\, dx_A}$ è una ben definita densità positiva su $\mathbb{R}^{|A|}$ (per il teorema di Fubini, ció è garantito quasi ovunque, ma noi lo richiediamo ovunque).

Da queste due assunzioni, seguono l'unicitá della misura invariante, l'aperiodicitá e la convergenza con probabilitá uno alla misura stazionaria. In particolare, analizzeremo il caso del *Gibbs sampler con random scan*, poiché sappiamo che le autocovarianze sono monotone decrescenti e positive.

Per presentare la distribuzione a cui siamo interessati, dobbiamo introdurre le tabelle di contingenza e l'espansione log-lineare.

Tabelle di contingenza

Una tabella di contingenza, in astratto, si puó rappresentare nel seguente modo: sia Δ un insieme finito di attributi e $\forall \delta \in \Delta$ sia I_δ l'insieme, finito, dei suoi possibili livelli. Gli elementi $i = (i_\delta)_{\delta \in \Delta}$ di $\mathcal{I} = \prod_{\delta \in \Delta} I_\delta$ sono

le caselle di una *tabella di contingenza* cui vengono assegnati dei conteggi $n = \{n(i)\}_{i\in\mathcal{I}}$; più in generale, abbiamo bisogno di considerare tabelle di contingenza con dati $\lambda(i), i \in \mathcal{I}$ reali positivi. La dimensione della tabella è il numero $|\Delta|$ di attributi e $\mid n \mid = \sum_i n(i)$. Se $a \subset \Delta$, la tabella $a-$*marginale* si ottiene considerando solo le frequenze che si riferiscono agli attributi in a, quindi le caselle della $a-$tabella marginale sono $i_a \in \mathcal{I}_a = \prod_{\delta\in a} \mathcal{I}_\delta$ e l'elemento corrispondente, detto $\lambda_a(i_a)$, é dato da

$$\lambda_a(i_a) = \sum_{j\in\mathcal{I}: j_a = i_a} \lambda(j).$$

La tabella condizionata è la tabella

$$\lambda_{a^c|a}(i_{a^c} \mid i_a) = \frac{\lambda(i_{a^c}, i_a)}{\lambda_a(i_a)}.$$

Sono stati considerati due differenti piani di campionamento per le tabelle di contingenza: quello multinomiale, che, essenzialmente, si puó vedere come l'estrazione di unità da una stessa urna contenente palline classificate secondo tutti gli attributi della tabella , con assegnate probabilità $p(i) > 0$ per ogni pallina e quello poissoniano. Il primo si puó ottenere dal secondo condizionando al numero di esperimenti.

L'obiettivo dell'inferenza è trovare il legame statistico tra i vari attributi (cfr.[**?**]), quindi abbiamo studiato la cosiddetta *espansione log-lineare* di una funzione su una tabella di contingenza. Avendo scelto un livello di riferimento, che indichiamo con 0, l'espansione log-lineare di una tabella $f(i) > 0$ è

$$log\, f(i) = \sum_{a\subseteq\Delta} u_a(i_a)\,, \tag{1}$$

dove la somma è fatta su tutti i possibili sottoinsiemi a di $\Delta = \{1, 2, ...k\}$ e dove gli $u - termini\, \{u_a\}$ sono funzioni delle coordinate di proiezione i_a, tali che $u_a(i_a) = 0$ se $i_k = 0$ e $k \in a$. I modelli a cui siamo interessati non dipendono da questo livello di riferimento.

Ad ogni famiglia $N_{\mathcal{M}}$ di sottoinsiemi di Δ associamo un modello che chiamiamo *log-lineare*, in cui gli $u-$termini u_a sono nulli per ogni $a \in N_{\mathcal{M}}$.

Un modello log-lineare si dice *gerarchico* se, quando un $u-$termine viene vincolato a zero, gli $u-$termini di ordine più alto che contengono lo stesso insieme di indici sono vincolati a zero, cioè se

$$u_a = 0 \Rightarrow u_t = 0 \, \forall \, t \supseteq a.$$

I modelli gerarchici possono essere specificati tramite le interazioni massimali, cioè tramite quei sottoinsiemi b tali che

$$b \notin N_{\mathcal{M}} \, \forall \, c \supset b \, e \, c \in N_{\mathcal{M}}.$$

Dato un modello gerarchico, associamo ad esso un *grafo* ponendo un vertice per ogni attributo e degli spigoli tra tutti gli elementi contenuti in uno stesso sottoinsieme di interazioni massimali; se le *clique* del grafo coincidono con le interazioni massimali, il modello si dice *modello grafico*.

I modelli grafici rappresentano una distribuzione del vettore degli attributi $(X_1, \cdots X_m)$, sotto la quale, se due blocchi di attributi a e b sono separati da un terzo blocco c (cioé se ogni cammino da a a b passa per c),

$$X_a \amalg X_b | X_c.$$

I modelli gerarchici sono parametrizzati da un numero ridotto di $u-$termini e l'obiettivo dell'inferenza statistica é di stimare i valori di questi parametri a partire da una tabella di conteggi osservata.

In particolare, l'approccio bayesiano cerca di quantificare l'incertezza su tali parametri mediante una distribuzione di probabilitá sui loro valori possibili (*a priori*). Attraverso la formula di Bayes, questa viene aggiornata mediante la verosimiglianza dei dati osservati (*a posteriori*) Avendo giá definito la forma della verosimiglianza (il modello statistico delle osservazioni), particolarmente opportune come distribuzioni a priori, sono le distribuzioni coniugate, poichè la distribuzione a posteriori è ancora dello stesso tipo.
Ad esempio, una famiglia coniugata al modello poissoniano di campionamento delle tabelle di contingenza, prescrive che le medie $\lambda(i)$ siano distribuite secondo $Gamma(\alpha(i), \beta(i))$ indipendenti, su tutte le entrate della tabella. Se $\beta(i) = \beta = 1$ (senza perdita di generalitá) per ogni $i \in \mathcal{I}$, allora

$p(i) = \dfrac{\lambda(i)}{\sum_i(\lambda(i))}$ con $i \neq$ dall'ultima casella (o da un'altra qualsiasi), ha distribuzione Dirichlet con parametri $\alpha(i)$, indipendente da $\sum_i \lambda(i)$. Da ció si puó dedurre che la distribuzione Dirichlet é coniugata alla distribuzione multinomiale.

Sia C un sottoinsieme degli attributi e consideriamo il nucleo di Gibbs sampler per il prodotto di gamma, corrispondente a $f_C(\lambda) = \lambda_{C^c|C}$, con nucleo $P(\lambda, dy) = \delta_{\lambda_{C^c|C}}(dy_{C^c|C})Gamma(dy)$; poichè la tabella λ si puó ottenere dalla tabella marginale λ_C e dalla tabella condizionata $\lambda_{C^c|C}$ e poichè queste due sono indipendenti, l'intera tabella viene aggiornata estraendo la tabella marginale dalla sua legge marginale, che è una gamma (somma di gamma).
Siamo quindi pronti per presentare la distribuzione a cui siamo interessati: data una distribuzione Gamma sui parametri $\lambda(i)$, siamo interessati alla distribuzione indotta sulle interazioni, condizionando a zero quelle indicate da un modello gerarchico. Per arrivare a questa distribuzione, abbiamo implementato un Gibbs sampler che avesse come distribuzione stazionaria quella di nostro interesse. Questo Gibbs sampler è proprio l'IPF Bayesiano, che andremo subito a presentare. Siano $C_1, \cdots, C_s$ sottoinsiemi dei sottoinsiemi massimali di un modello gerarchico relativo a una tabella di contingenza $\mathcal{I}$, con elementi $\alpha(i) > 0$. Sia $\mu^{(t)}(i)$ il valore simulato della variabile aleatoria $\mu(i)$ al passo t. Ad ogni passo t dell'algoritmo, generiamo una tabella marginale $g^t_{C_k}$ corrispondente a C_k, le cui caselle sono realizzazioni indipendenti di variabili aleatorie *gamma* con parametri $\alpha_{C_k}(i_{C_k}) = \sum_{j:j_{C_k}=i_{C_k}} \alpha(j)$ (ció verrà fatto ciclicamente per ogni C_k). Allora, l'algoritmo in questione aggiorna la tabella nel modo seguente per ogni C_k:

$$\mu^{(t+\frac{k}{s})}(i) = \mu^{(t+\frac{k-1}{s})}(i)\Big(\frac{g^t_{C_k}(i_{C_k})}{\mu^{(t+\frac{k-1}{s})}_{C_k}(i_{C_k})}\Big), \quad \forall\, C_k,\, k = 1, \cdots, s. \tag{2}$$

Se q_s e q sono, rispettivamente, una densità prodotto di *Gamma* sulle caselle della tabella e la densità ottenuta condizionando a zero le interazioni indicate dal modello gerarchico (ovvero la distribuzione di nostro interesse), allora, se aggiorniamo la tabella marginale su uno dei C_i estraendo da q_s e lasciando

fissa la tabella su C_i^c condizionata a C_i, aggiorniamo la tabella operando un passo di **GS** sulle interazioni contenute in C_i lasciando fisse le altre (osserviamo che q_s e q si possono vedere come distribuzioni sia sullo spazio delle tabelle, sia su quello delle interazioni, che é uno spazio prodotto).

Inoltre, la tabella condizionata $\lambda_{C_i^c|C_i}$ è in corrispondenza biunivoca con $\{u_A, A \not\subset C_i\}$ sia sotto q_s, sia sotto q, quindi aggiornare la tabella marginale λ_{C_i} estraendola dalla sua marginale, equivale ad aggiornare la tabella marginale λ_{C_i} estraendola dalla sua legge condizionata a $\{u_A,\, A \not\subset C_i\}$. Ma ció significa fare un passo di Gibbs sampler lasciando fisse le interazioni non contenute interamente in C_i.
Quindi, se aggiorniamo C_i dalla sua distribuzione marginale e teniamo fissata la tabella su C_i^c condizionata a C_i ($\lambda_{C_i^c|C_i}$ è in corrispondenza uno a uno con $\{u_A, A \not\subset C_i, \tilde{u}_A, A \not\subset C_i\}$, dove le $\tilde{u}$ sono quelle nulle indicate dal modello), facciamo un passo di Gibbs sampler sul modello *saturato* (q_s); allora, se aggiorniamo C_i dalla sua distribuzione marginale e teniamo fissate le interazioni non interamente contenute in C_i, facciamo un passo di Gibbs sampler sul modello di nostro interesse (λ_{C_i} è indipendente da $\lambda_{C_i^c|C_i} \leftrightarrow \{u_A, A \not\subset C_i, \tilde{u}_A, A \not\subset C_i\}$, quindi λ_{C_i} è indipendente da $\{u_A, A \not\subset C_i\} \mid \{\tilde{u}_A, A \not\subset C_i\}$); da ció segue che $P(\lambda, dy)$ è un nucleo anche per la distribuzione di interesse.

Notiamo che possiamo rappresentare questo Gibbs sampler in termini di coordinate sullo spazio prodotto delle interazioni presenti nel modello, infatti (cfr.[**?**]) dato che $\{u_A,\, A \not\subset C_i\}$ sono in corrispondenza biunivoca con la tabella condizionata $m_{C_i^c|C_i}(i_{C_i^c}, i_{C_i})$, deve essere possibile recuperare $\{u_A,\, A \subset C_i\}$ come funzione di $(m_{C_i},\, \{u_A,\, A \not\subset C_i\})$, cioè deve essere possibile scrivere $\{u_A,\, A \subset C_i\} = \psi_i(m_{C_i},\, \{u_A,\, A \not\subset C_i\})$. Quindi la legge di $\{u_A,\, A \subset C_i\}$ condizionata a $\{u_A,\, A \not\subset C_i\}$, è la legge di $\psi_i(\cdot,\, \{u_A,\, A \not\subset C_i\})$ quando m_{C_i} è gamma con determinati parametri α_{C_i}.

Allora, data una tabella iniziale $\mu^{(0)}$ che soddisfi i vincoli di un modello log-lineare gerarchico, questo algoritmo definisce una catena di Markov $\{\mu(i)^{(t)},\, i \in \mathcal{I},\, t = 1, 2, \cdots\}$ la cui legge converge una *Gamma* con parametri $\alpha_{C_k}(i_{C_k}) = \sum_{j:j_{C_k}=i_{C_k}} \alpha(j)$, condizionata all'azzeramento delle interazioni corrispondenti al modello gerarchico.

L'algoritmo in questione trae origini da un altro algoritmo, l'IPF, che non estrae la tabella marginale ma, al posto di questa, inserisce la tabella marginale di α. L'IPF è un algoritmo di massimizzazione iterata che converge alla moda della distribuzione a posteriori.

Studio dell'algoritmo IPF Bayesiano

Le prestazioni di un algoritmo **MCMC** sono tanto migliori quanto più la varianza di $\frac{\sum_{i=1}^{n} f(X_i)}{n}$ decresce rapidamente a zero; se le X_i fossero indipendenti, allora, in equilibrio, $Var\frac{\sum_i f(X_i)}{n} = \frac{f(X_1)}{n}$ e vale il teorema del limite centrale. Se invece $nVar\sum_i f(X_i)$ tende a una costante $c < \infty$, nel caso di catene reversibili vale sempre il teorema del limite centrale.
Nel caso di grafi *triangolati* è assicurata la convergenza alla distribuzione stazionaria in un solo passo di Gibbs sampler, quindi vale il teorema del limite centrale poichè siamo in equilibrio e le Xi sono indipendenti. Dato che nel caso di grafi non triangolati non è stata dimostrata l'esistenza di una *buca spettrale*, abbiamo effettuato esperimenti per calcolare empiricamente le prestazioni dell'algoritmo. Abbiamo cercato di stimare delle medie in equilibrio e abbiamo usato tecniche di analisi di output per ricavare intervalli di confidenza e per verificare il raggiungimento approssimato dell'equilibrio. Abbiamo usato una tecnica proposta dalla letteratura sui metodi **MCMC**, detta *tecnica dei batch means*; essa raggruppa le realizzazioni di una catena di Markov in l gruppi disgiunti e si basa essenzialmente sulla loro incorrelazione. Abbiamo quindi implementato la cosiddetta procedura $LBATCH$ per una singola catena. Questa procedura fornisce intervalli di confidenza asintoticamente validi per la grandezza studiata. Abbiamo poi implementato la tecnica di Gelman e Rubin: essi proposero di far partire più catene da punti diversi e di vedere ognuna di queste come un singolo batch. Essendo catene che partono da punti diversi, abbiamo assicurata la loro indipendenza e quindi la loro incorrelazione: possiamo, allora, usare la tecnica dei batch means per ottenere altri intervalli di confidenza e confrontarli con quelli ottenuti secondo la procedura $LBATCH$.
Queste tecniche diagnostiche, tuttavia, possono essere, a loro volta rigorosamente dimostrate se ha validità il teorema del limite centrale. Questa osser-

vazione permette di interpretare eventuali anomalie nei risultati sperimentali, come indizi di una convergenza all'equilibrio più lenta di quella implicata dal teorema del limite centrale.
Fissata la tabella delle gamma $\alpha(i)$, è stato fatto girare l'algoritmo per calcolare $\frac{\sum_i f(X_i)}{n}$ e i corrispondenti intervalli di confidenza con la tecnica $LBATCH$. La simulazione è stata interrotta quando questo intervallo risultava più piccolo di una grandezza fissata a priori.

La tabella iniziale è stata prodotta applicando l'IPF a una tabella ottenuta dividendo per un numero maggiore di uno la tabella delle $\alpha(i)$ (in questo modo abbiamo aumentato la varianza), quindi corrisponde alla moda della distribuzione di interesse, che è un punto rappresentativo della distribuzione di equilibrio.
Sono, poi, state effettuate simulazioni indipendenti, partendo da tabelle iniziali sufficientemente distanti dalla precedente ed è stata applicata a queste catene una tecnica diagnostica proposta da Gelman e Rubin, per valutare se la lunghezza della catena fosse sufficiente per *raggiungere* l'equilibrio. Questi test sono stati positivi, nel senso che i valori ottenuti sono stati molto al di sotto di quelli proposti da Gelman e Rubin (1.2-1.3). Un ulteriore controllo che abbiamo effettuato è stato ottenuto confrontando la stima della varianza prodotta dal metodo dei batch means e quella ottenuta tramite le autocovarianze su tutta la catena. Anche questi risultati sono stati confrontabili.

Bibliografia

[BN] Barndorff-Nielsen, *Informations and exponential families in statistical theory*, Wiley, Chichester,1978.

[DY] Diaconis,Ylvisaker, *Conjugate priors for exponential families* , The Annals of statistics, 1979 vol.7,No 2,pag.269-281.

[W] J. Whittaker, *Graphical Models in applied multivariate statistics.*

[MP] Mauro Piccioni, *Independence structure of natural conjugate densities to exponential families and the Gibbs sampler.*

[DG] Dani Gamerman, *Markov Chain Monte Carlo Stochastic simulation for bayesian inference.*

[LT] Luke Tierney, *Introduction to general state space Markov Chain Teory.*

[WWK] J.S.Liu, Wing H., Wong e Augustine Kong, *Covariance structure and convergence rate of the Gibbs sampler with various scan* , J.R.Statist. Soc. B (1995), **57**, No, 1, pp. 157-169.

[RS] G.O.Roberts, A.F.M.Smith *Simple conditions for the convergence of the Gibbs sampler and Metropolis Hastings algorithms* , Stochastic Processes and their Application 49 (1994) 207-216, North-Holland.

[G] Gareth O.Roberts, *Markov Chain concepts related to sampling algorithms.*

[F] G.S. Fishman, *Monte Carlo, concepts, algorithms and applications*, Springer, New York, 1996.

[S] M. Reed, B Simon, *Vol 4: Analysis of operators*, Academic Press, Boston, 1978.

[JS] J.L.Shaffer, *Analisys of incomplete multivariate data.*

[GR] A.Gelman, D.B.Rubin, *Inference from iterative simulation using multiple sequences* , Statistics Science 1992, Vol. 7, No.4, 457-511.