

Corso Intensivo di Metodi Bayesiani: obiettivi, metodi e requisiti.

Docente: S. Andreon, stefano.andreon@brera.inaf.it

Obiettivi e metodi

Il corso si articola in 5 lezioni di 3 ore, di solito divise in un'ora di chiacchiere e due ore di lavoro al computer (dove le chiacchiere diventano numeri pubblicabili), tranne per la prima lezione, dove la ripartizione sarà presumibilmente opposta. Obiettivo del corso é insegnare a fare cose di diretta pertinenza con la propria attività di ricerca (la sezione "Analisi" dell'articolo), ben diverso dall'ascoltare cose interessantissime di un argomento diverso dal proprio (ossia ascoltare un seminario). La natura del corso (fare, non ascoltare) e il poco tempo a disposizione ci impedirà di affrontare casi molto difficili e iper-specializzati (e di interesse, di solito, solo di alcuni). Chiunque abbia un problema statistico che pensa essere di interesse generale e desidera sia trattato durante il corso, é invitato a sottopormelo (ben prima dell'inizio del corso, mandandomi una succinta descrizione del problema, senza tralasciare gli aspetti critici, e i dati). Tutti i problemi sono benvenuti.

Il programma preliminare del corso é all'URL:

http://www.brera.mi.astro.it/~andreon/corso_metodi_bayesiani/

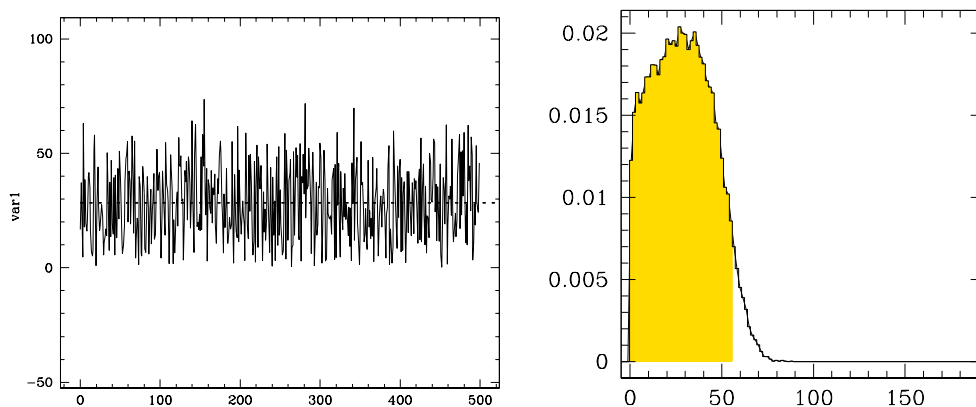


Figure 1: Left panel: Trace plot. Right panel: Distribuzione Marginale

Requisiti

A causa del poco tempo a disposizione, é richiesto del lavoro preliminare e una frequenza costante.

Le lezioni di laboratorio richiedono l'uso di un computer con un sistema operativo Linux-like o Mac. I calcoli statistici saranno effettuati con l'ausilio di JAGS. Prima dell'inizio del corso occorre installare JAGS, disponibile all' URL:

<http://www-ice.iarc.fr/~martyn/software/jags/>

Si noti che JAGS richiede l'installazione di BLAS and LAPACK, nel caso queste librerie non fossero già installate. Tutto il software in questione é free ed é già stato installato con successo negli anni passati in numerosi sistemi operativi. Mi preme sottolineare che frequentare il corso senza poter usare JAGS é, in gran misura, una perdita di tempo (vostro).

Non si assume che sappiate usare JAGS, ma si assume che sappiate usare un ambiente grafico (di vostra scelta) ove poter plottare grafici e fare semplici conti. In considerazione della breve

durata del corso si richiede che chi frequenta il corso abbia già acquisito la capacità di

- a) importare nell'ambiente grafico files i files di output di JAGS. Prendete come esempio quelli all'URL:

http://www.brera.mi.astro.it/~andreon/corso_metodi_bayesiani/CODAindex.txt

http://www.brera.mi.astro.it/~andreon/corso_metodi_bayesiani/CODAchain1.txt

CODAindex.txt indica cosa contiene CODAchain.txt (piú variabili, in coda una dietro l'altra) e dove queste iniziano e finiscono. Per esempio *s* inizia alla riga 1 e finisce alla 50000 del file CODAchain, mentre la variabile *bkg* inizia alla riga 50001 e finisce alla 100000. Si preveda già una certa flessibilità, il prossimo CODAindex.txt conterrà un numero diverso di righe (p.e. 30).

- b) calcolare la media e deviazione standard di una sequenza di valori (p.e. *s* del file allegato ha media 28.55 e standard deviation 16.5)
- c) calcolare l'intervallo piú corto che racchiude l'*x* % dei valori (p.e. *s* del file allegato ha un intervallo al 95 % che va da 0 a 56)
- d) plottare l'andamento di una variabile con il suo indice di sequenza (trace plot), per una variabile sola, e 8 alla volta. Si veda la fig 1, pannello di sinistra. Suggerimento: basta leggere la prima e la seconda colonna, e plottare una contro l'altra.
- e) plottare la distribuzione delle frequenze di una variabile (p.e. si veda il pannello di destra della Fig. 1 per la variabile *s*). Si noti che:

- l'integrale della distribuzione é uno per definizione. Suggerimento: farsi l'integrale a occhio sulla figura, se non torna meglio verificare.
- la forma della distribuzione é indipendente dal bin size (o kernel) usato, per bin sizes ragionevoli. Suggerimento: se cambiando il bin size, la distribuzione si muove ... Huston, c'è un problema.

In SuperMango basta usare il comando histogram e normalizzare, per esempio:

```
set myhisto=histogram(intr.scatt:mycent)/step/dimen(intr.scatt).
```

- f) plottare il classico grafico con i contorni di confidenza per due parametri (si usi come esempio *s* e *bkg*). I contorni devono essere di forma smooth e non fissata (no a ellissi o cerchi a priori. Prevedete già che vi possano essere due isole), e devono includere il 68 % e il 95 % dei punti. Non é necessario che il conto sia esatto, é accettabile una certa approssimazione, se la curva che dovrebbe contenere il 68 % dei punti ne contiene invece il 65 %, amen.

Si noti che non é possibile settare una threshold pari al massimo - "magic numer" (come suggerito da Avni 1976, o Numerical Recipes), perché questa contiene la percentuale voluta dei punti solo nel caso gaussiano (piú un certo numero di condizioni, non vi tedio). Va invece trovato il contorno che racchiude la percentuale voluta dei punti.

In SuperMango si ottiene per esempio creando una matrice/immagine (io uso 33x33 pixels) in cui ogni pixel ha un valore pari al numero di punti cascano nell'area del pixel. Usando il comando contour dopo aver settato i livelli si ottiene il plot in Fig 2. I livelli sono tali per cui per cui la somma dei valori dei pixels con valore maggiore del livello é pari al 68 % del numero totale di punti. I maghi tra voi potranno anche tener conto dell'effetto di smoothing indotto dalla pixellizzazione, mentre gli altri possono ignorare la complicazione.

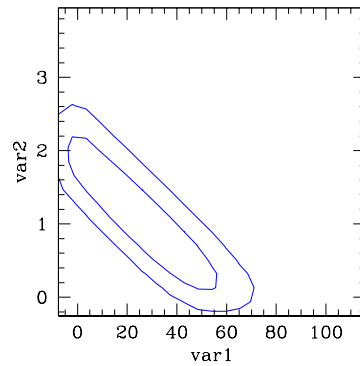


Figure 2: Contorni di confidenza (Bayesiani). Si noti che i contorni sono imprecisi vicino a $var1 = 0$ e $var2 = 0$ a causa della sharpness della distribuzione, e non corretti per gli effetti di smoothing.

- g) nella prima lezione del corso calcoleremo i contorni di confidenza su un parametro secondo la ricetta frequentista. A tale scopo, lo studente legga la matrice

http://www.brera.inaf.it/utenti/andreon/JKCS041_nH-T.grid

e si prepari un programmino che calcola, e plotta, somme o max lungo le righe e/o colonne e che plotti la matrice (ispiratevi alla Fig 2). La matrice é scritta nel formato $x_i, y_i, valore$.

Se avete fatto correttamente, dovrete ottenere le figure allegate. Se non vi tornano, cercate l'errore, magari chiedendo aiuto ai vostri vicini di ufficio (o scrivendo a me, una volta per errore ha incluso la figura sbagliata!). Il tutto deve essere scritto in modo tale che con un altro CODAindex (con un diverso numero di righe) riproducete i diversi plot senza dover riscrivere le macro grafiche. Meglio ancora che ad ogni lettura di un file CODAindex tutti i sopradetti plot e numeri sono generati automaticamente (un trace plot e una distribuzione marginale per variabile e un contour plot per ogni coppia di variabili).

L'esperienza degli anni passati mi suggerisce di stressare l'importanza di avere i programmi grafici pronti *prima* del corso: frequentare il corso senza poter effettuare le operazioni sopradette in pochi secondi implica che userete il corso di statistica per fare della computer graphics, ossia l'impossibilità di potersi dedicare agli aspetti "statistici" del problema e perdere così inutilmente 15h del proprio tempo. Considerate invece il tempo dedicato alla scrittura di queste routines grafiche come un'investimento per gli articoli che scriverete (non vedo come potreste farne a meno!). Quindi, non abbiate paura a dedicarci del tempo. Nulla impedisce che vi dividiate la scrittura delle routines di plot tra di voi, purché ciascuno abbia tutte le routines e le sappia usare/modificare per i casi specifici. Sconsiglio vivamente, per esempio, di approfittare delle routines IDL dell'amico, se poi non sapete usare IDL per fare semplici operazioni e modificare il .pro originario (come già successo l'anno scorso!).

A presto.

stefano.andreon@brera.inaf.it