Μπεϋζιανή Στατιστική και MCMC
Μέρος 2ο : MCMC

Δημήτρης Φουσκάκης,
Επίκουρος Καθηγητής,
Σχολή Εφαρμοσμένων Μαθηματικών
και Φυσικών Επιστημών,
Τομέας Μαθηματικών,
Τηλέφωνο: (210) 772-1702,
Φαξ: (210) 772-1775.
fouskakis@math.ntua.gr.
Κτίριο Ε, Γραφείο 205.

Περιεχόμενα Μαθήματος

- Εισαγωγή στο Πρόβλημα.
- Monte Carlo Εκτιμητές.
- Προσομοίωση.
- Αλυσίδες Markov.
- Αλγόριθμοι MCMC (Metropolis – Hastings & Gibbs Sampling).
- WinBugs.
- Διαγνωστικοί Έλεγχοι.
- MCMC στα Γενικευμένα Γραμμικά Μοντέλα.

Διδασκαλία

- Σελίδα μαθήματος: http://www.math.ntua.gr/~fouskakis/
- Ώρες διδασκαλίας: Πέμπτη 4μμ – 7μμ, PC – LAB , Τομέα Μαθηματικών.
- Από την σελίδα του μαθήματος κατεβάστε τις σημειώσεις του στατιστικού πακέτου R και εξοικειωθείτε με το λογισμικό.

Βιβλιογραφία

- Gamerman, D. and Lopes, H.F. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman and Hall, New York and London.
- Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1995). *Markov Chain Monte Carlo in Practice*. Chapman and Hall, New York and London.
- Chen, M., Shao, Q. and Ibrahim, J.G. (2000). *Monte Carlo Methods in Bayesian Computation.* Springer Verlag, New York.
- Ntzoufras, N. (2009). *Bayesian Modeling using Winbugs*. Wiley, John & Sons, Incorporated, New York.

Μπεϋζιανή Στατιστική

- Δεδομένα $\mathbf{y} = (y_1, \ldots, y_n)$
- Παράμετρος $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k) \in \Theta$
- Πιθανοφάνεια (Likelihood) $L(\mathbf{y} \mid \boldsymbol{\theta})$
- Εκ των προτέρων (Prior) $p(\boldsymbol{\theta})$
- Εκ των υστέρων (Posterior)

$$p(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{L(\mathbf{y} \mid \boldsymbol{\theta}) \times p(\boldsymbol{\theta})}{\int_{\Theta} L(\mathbf{y} \mid \boldsymbol{\theta}) \times p(\boldsymbol{\theta}) d\boldsymbol{\theta}}$$

Σταθερά κανονικοποίησης

Posterior $\propto$ Likelihood x Prior

Παραδείγματα

$$Y_1,\ldots,Y_n \sim N(\theta,1)$$

$$p(\theta) = \frac{1}{\pi(1+\theta^2)}$$

$$p(\theta \mid \mathbf{y}) \propto \exp\left\{-\frac{\sum_{i=1}^{n}(y_i-\theta)^2}{2}\right\} \times \frac{1}{1+\theta^2}$$

$$\propto \exp\left\{-\frac{n(\theta-\bar{y})^2}{2}\right\} \times \frac{1}{1+\theta^2}$$

Παραδείγματα

$$Y_1,\ldots,Y_n \sim N(\mu,\sigma^2)$$

$$p(\mu,\sigma^2) \propto \frac{1}{\sigma^2}$$

$$p(\theta \mid \mathbf{y}) \propto \exp\left\{-\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2\sigma^2}\right\} \times \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+1}$$

Εισαγωγή

Παρατηρούμε από τα δύο προηγούμενα παραδείγματα ότι ο πλήρης υπολογισμός της εκ των υστέρων κατανομής είναι δύσκολος σε κάποιες περιπτώσεις. Το πρόβλημα γίνεται ακόμα πιο περίπλοκο αν η παράμετρος είναι πολυδιάστατη.

Λύσεις:

- Συζυγείς εκ των προτέρων (conjugate priors).
- Ασυμπτωτικές Προσεγγίσεις.
- MCMC.

Εισαγωγή

MCMC

- 1949 Metropolis –Ulam (αρχική ιδέα).
- 1954 Metropolis et al. (Metropolis Algorithm).
- 1970 Hastings (Metropolis – Hastings Algorithm).
- 1984 Geman & Geman (Gibbs Sampling).
- 1990 Gelfand & Smith (Εφαρμογη MCMC σε Μπεϋζιανή Στατιστική).
- 1995 Green (Reversible Jump MCMC).

Εισαγωγή

- **ΙΔΕΑ**

  Ότι θέλεις να μάθεις για μια κατανομή μπορεί να επιτευχθεί απλά **προσομοιώνοντας** τυχαίες τιμές από αυτή (Metropolis – Ulam 1949).

Monte Carlo Εκτιμητές

Ας υποθέσουμε ότι ενδιαφερόμαστε για την $p(\boldsymbol{\theta}|\mathbf{y})$ την οποία δεν μπορούμε να υπολογίσουμε αναλυτικά και ας υποθέσουμε χάριν ευκολίας ότι το θ είναι μονοδιάστατο. Στην Μπεϋζιανή συμπερασματολογία συνήθως ενδιαφερόμαστε για διάφορα "χαρακτηριστικά" της εκ των υστέρων κατανομής όπως:

Monte Carlo Εκτιμητές

- Μέσος $\mu = \mathrm{E}(\theta \mid \mathbf{y}).$
- Τυπική Απόκλιση $\sigma = \sqrt{\mathrm{V}(\theta \mid \mathbf{y})}.$
- Γράφημα.
- Τεταρτημόρια (π.χ. για την κατασκευή ενός 95% εκ των υστέρων διάστημα εμπιστοσύνης για το θ πρέπει να γνωρίζεις τα 2.5% και 97.5% τεταρτημόρια της εκ των υστέρων κατανομής.

Ας υποθέσουμε ότι με κάποιον τρόπο προσομοιώσαμε τιμές $\theta_1^*, \ldots, \theta_m^*$ από την $p(\theta \mid \mathbf{y}).$

Τότε μπορούμε εύκολα να εκτιμήσουμε όλα τα παραπάνω:

Monte Carlo Εκτιμητές

- Μέσος $\hat{\mu} = \hat{E}(\theta \mid \mathbf{y}) = \bar{\theta}^* = \frac{1}{m}\sum_{i=1}^{m} \theta_i^*.$
- Τυπική Απόκλιση $\hat{\sigma} = \sqrt{\hat{V}(\theta \mid \mathbf{y})} = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}\left(\theta_i^* - \bar{\theta}^*\right)^2}.$
- Γράφημα $\longrightarrow$ Ιστόγραμμα των θ_i^*.
- Τεταρτημόρια $\longrightarrow$ Μετράμε πόσες από τις τιμές των θ_i^* είναι μικρότερες από μια σειρά καθορισμένων τιμών, π.χ. για να εκτιμήσουμε το 2.5% τεταρτημόριο, λύνουμε ως προς t την εξίσωση

$$\hat{F}_\theta(t) = \frac{1}{m}\sum_{i=1}^{m} I(\theta_i^* \le t) = 0.025, \text{ I: δείκτρια συνάρτηση.}$$

Monte Carlo Εκτιμητές

Αποδεικνύεται ότι για μεγάλο m, οι συγκεκριμένες **Monte Carlo εκτιμήτριες**, συγκλίνουν στην ως προς εκτίμηση ποσότητα με αρκετά μεγάλη πιθανότητα, υπό την προϋπόθεση το δείγμα των $\boldsymbol{\theta}^*$ να είναι τυχαίο. Ένας τρόπος επιλογής τέτοιου δείγματος είναι τα θ_i^* να είναι ανεξάρτητα και ισόνομα (IID), αλλά όπως τελικά θα δούμε παρακάτω αυτό δεν είναι αναγκαίο.

Monte Carlo Εκτιμητές

Αν για παράδειγμα $\bar{\theta}^* = \frac{1}{m}\sum_{i=1}^{m}\theta_i^*$ προέρχεται από IID δείγμα μεγέθους m από την $p(\theta|\mathbf{y})$ τότε χρησιμοποιώντας κλασική στατιστική έχουμε ότι $V(\bar{\theta}^*) = \sigma^2/m$ με σ^2 η διασπορά της $p(\theta|\mathbf{y})$.

Άρα μπορούμε να δημιουργήσουμε ένα Monte Carlo τυπικό σφάλμα για το $\bar{\theta}^*$

$\widehat{SE}(\hat{\theta}^*) = \hat{\sigma}/\sqrt{m}$ ⟶ Χρησιμοποιείται συνήθως για να αποφασίσουμε πόσο μεγάλο πρέπει να είναι το m

IID Παράδειγμα

Έστω η εκ των υστέρων κατανομή

$$p(\lambda \mid \mathbf{y}) = \Gamma(29.001, 14.001)$$

Τότε $\mu = E(\lambda \mid \mathbf{y}) = 29.001/14.001 = 2.071$

Ας δούμε πόσο καλά εκτιμά αυτή την ποσότητα ο Monte Carlo εκτιμητής

IID Παράδειγμα

```
gamma.sim<- function(m,alpha,beta,n.sim,seed)  {
set.seed(seed)
theta.out <- matrix(0,n.sim,2)
for(i in 1:n.sim) {
  theta.sample<-rgamma(m,alpha,beta)
  theta.out[i,1]<-mean(theta.sample)
  theta.out[i,2]<-sqrt(var(theta.sample)/m)
}
return(theta.out)
}
```

IID Παράδειγμα

- Η παραπάνω R συνάρτηση προσομοιώνει (n.sim φορές) m τυχαίες τιμές, ανεξάρτητες και ισόνομες από την κατανομή Γ(α,β) και επιστρέφει τον μέσο τους μαζί με το τυπικό σφάλμα.

IID Παράδειγμα

```
> m <- 1000

> alpha <- 29.001

> beta <- 14.001

> n.sim <- 500

> seed <- c( 6425451, 9626954 )
```

IID Παράδειγμα

```
> theta.out <- gamma.sim( m, alpha, beta, n.sim, seed )

# This took about 1 second at 550 Unix MHz.

> theta.out[ 1:10, ]

          [,1]       [,2]
 [1,] 2.082105 0.01166379
 [2,] 2.072183 0.01200723
 [3,] 2.066756 0.01247277
 [4,] 2.060785 0.01200449
 [5,] 2.078591 0.01212440
 [6,] 2.050640 0.01228875
 [7,] 2.071706 0.01182579
 [8,] 2.063158 0.01176577
 [9,] 2.058440 0.01186379
[10,] 2.068976 0.01220723
```

Άρα οι τιμές του δειγματικού μέσου είναι πολύ κοντά στις πραγματικές τιμές συν-πλην περίπου 0.012 που είναι επίσης πολύ κοντά στο πραγματικό τυπικό σφάλμα $\sigma/\sqrt{m} = \sqrt{\alpha}/\beta\sqrt{m} = 0.01216.$