# MCMC Diagnosis

Gwangsu Kim

JBNU

Second semester of 2024

# Stationary distribution of MCMC I

- Stationary distribution of MCMC is the key, and the ergodicity means that

$$\sum_{i=1}^{N} \theta_i / N \xrightarrow{p} \mathbb{E}[\theta]$$

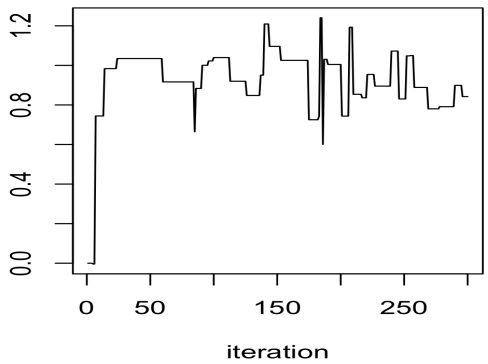
as $N \to \infty$ where $\{\theta_i\}_{i=1}^{N}$ is the MCMC samples.

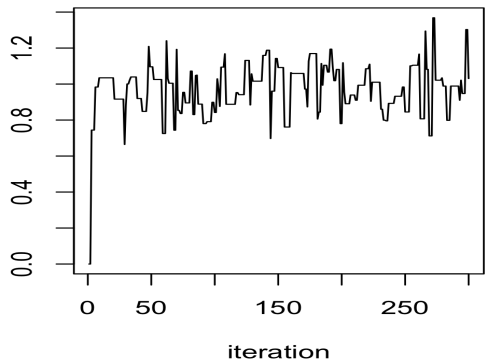
# Stationary distribution of MCMC II

- In practice we can have the following issues:
  1. How long do we need to run the Markov Chain to approximate the posterior distribution?
  2. Mixing is a concept from the ergodicity, which means the moving around the full distribution. Fast mixing indicates the fast convergence to the posteriors.

# Stationary distribution of MCMC III

- Examples of slow mixing and fast mixing



# Brief Sketch of Diagnosis I

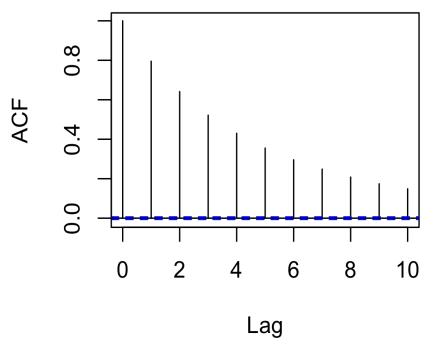
- Basic tools is the trace plot. Also the reason of ergodicity is the weak correlation between chain.

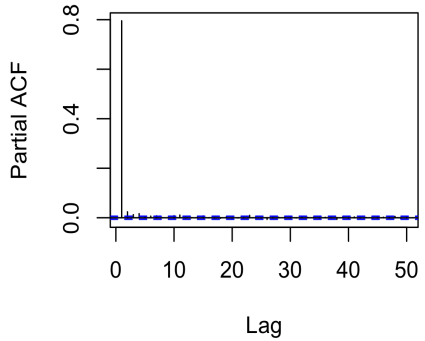
$$\frac{\sum_i \mathrm{Var}(X_i)}{\sum_{i \neq j} \mathrm{Cov}(X_i, X_j)} = O(1).$$

- This implies that the correlation within chain is the key points to show the speed of convergence. Note that the iid samples have the ratio of 0.

# Brief Sketch of Diagnosis II

- Simple measure: ACF (auto correlation function) and PACF (partial auto correlation plot)
- Example (ACF and PACF)



# Mathematical Aspects I

- When we have the following notations:

$$\hat{g}_{MC} = \frac{1}{N}\sum_i g(\theta^{(i)}) \text{ MC mean}$$

$$\hat{g}_{MCMC} = \frac{1}{N}\sum_i g(\theta^{(i)}) \text{ MCMC mean}$$

- Here, the distribution and the stationary distribution are same.

# Mathematical Aspects II

- Then we have the following:

$$\begin{aligned}\text{Var}(\hat{g}_{MCMC}) &= \text{Var}(\hat{g}_{MC}) + \sum_{i \neq j} \mathbb{E}\Big[ \Big(g(\theta^{(i)}) - \mathbb{E}[g(\theta^{(i)})]\Big) \\ &\quad \times \Big(g(\theta^{(j)}) - \mathbb{E}[g(\theta^{(j)})]\Big) \Big]\end{aligned}$$

  - The second term depends on the autocorrelation of samples within the Markov chain.

# Mathematical Aspects III

- Often positive so MCMC variance is larger than MC variance.
- Often positive so MCMC variance is larger than MC variance, and high correlation is an indicator of poor mixing.
- Effective sample size: how times samples are required for the equivalent to the variance of iid mean.

$$s_{eff} = \frac{\text{Var}(\hat{g}_{MCMC})}{\text{Var}(\hat{g}_{MC})}$$

# Mathematical Aspects IV

- When we know the $\rho_k$ of the lag $k$ autocorrelation, we have

$$S_{eff} = \frac{S}{1 + 2\sum_k \rho_k}$$

where $S$ is the length of Markov chain (sample size).

# Measures to be studied I

- Geweke (1992)
  1. Test for equality of the means of the first and last part of a Markov chain (by default the first 10% and the last 50%).
  2. If the samples are drawn from the stationary distribution, the Geweke's statistic has an asymptotically standard normal distribution.

# Measures to be studied II

- Gelman and Rubin (1992): using $m > 1$ chains
  1. Use different starting values.
  2. The chains have ”forgotten” their initial values, and the output from all chains should be indistinguishable.

# Measures to be studied III

3 Based on a comparison of within-chain and between-chain variances:

$$R = \frac{\frac{L-1}{L}W + \frac{1}{L}B}{W}$$

where $B : L \times$ variance between chains, $W$ : average over variances of chains, and $L$ is the length of chain.

# Measures to be studied IV

- Heidelberg-Welch (1981)
  1. After discarding the first 10%, 20%, of the chain until either the null hypothesis is accepted, or 50% of the chain has been discarded.
  2. The test is for the stationary distribution.
  3. The halfwidth test calculates half the width of the $(1`\alpha)\%$ credible interval around the mean.
  4. If the ratio of the halfwidth and the mean is lower than some $\epsilon$, then the chain passes the test. Otherwise, the chain must be run out longer.

# Measures to be studied V

- Raftery-Lewis (1992): burn-in and iterations
  1. Select a posterior quantile of interest $q$.
  2. Select an acceptable tolerance $r$ for this quantile
  3. Select a probability $s$, which is the desired probability of being within $(q - r, q + r)$.
  4. Run a "pilot" sampler to generate a Markov chain of minimum length given by rounding up $nmin = \left[\Phi^{-1}\left(\frac{s+1}{2}\frac{\sqrt{q(1-q)}}{r}\right)\right]^2$
  5. Compare the required number of burin-in and iteration with $nmim$. (Inflation factor)

# Measures to be studied VI

- Example of Geweke

```
library(coda)
rs = geweke.diag(x, frac1=0.1, frac2=0.5)
2*(1-pnorm(abs(unlist(rs[[1]]))))
 var1
0.6886054
```

- First 10% and the last 50%, the p-value if greater than 0.05 (not rejection of stationarity).

# Measures to be studied VII

- Example of Heidelberg-Welch

```
heidel.diag(x, eps=0.1, pvalue=0.05)
Stationarity start     p-value
     test         iteration
[,1] passed       1           0.768
     Halfwidth Mean Halfwidth
     test
[,1] passed    1.92 0.00467
```

- The first step is passes with high p-value greater than 0.05, and the second step is passes with a small halfwidth less than 0.1.

# Measures to be studied VIII

- Example of Raftery-Lewis

```
raftery.diag(x, q=0.025, r=0.005, s=0.95, converge.eps=0.001
Quantile (q) = 0.025
Accuracy (r) = +/- 0.005
Probability (s) = 0.95
 Burn-in  Total Lower bound  Dependence
 (M)      (N)   (Nmin)       factor (I)
 8        8515  3746         2.27
```

- The 25% quantile ± 0.005 with prob. 0.95 requires the 8 burn-in and 8515 samples compare to $nmin$=3746, IF 2.27.

# Measures to be studied IX

- Example of Gelman-Rubin

```
library(Bolstad2) ; theta0 = c(0,1) ; theta1 = c(3,2)
p = 0.6 ;  candidate = c(0, 3)
v1 = normMixMH(theta0, theta1, p, candidate, steps = 200)
v2 = normMixMH(theta0, theta1, p, candidate, steps = 200)
v3 = normMixMH(theta0, theta1, p, candidate, steps = 200)
v4 = normMixMH(theta0, theta1, p, candidate, steps = 200)
theta = cbind(v1,v2,v3,v4) ;
GelmanRubin(theta)
1.0007757715663 3
```

# Model Selection I

- Amon various metrics, we consider three.
- 
- Bayes factor

$$\frac{\int p_k(y \mid \theta_k)\pi(\theta_k)d\theta_k}{\int p_l(y \mid \theta_l)\pi(\theta_l)d\theta_l}$$

# Model Selection II

- DIC (Deviance Information Criteria)

$$-2\log p(y \mid \hat{\theta}) + 2p_d$$

where $\hat{\theta}$ is Bayes estimator and

$$p_d = 2\left(\log p(y \mid \hat{\theta}) - \mathbb{E}_{post}[\log p(y \mid \theta)]\right).$$

- Key idea is to assess the uncertainty of posteriors by $p_d$. Pratical in various models, and comparison possible with difference models.

# Model Selection III

- It is assumed that the specified parametric family of probability distributions that generate future observations encompasses the true model. Overfitted properties are observed.

# Model Selection IV

- Pseudo-marginal likelihood (LPML): (psuedo) predictive distribution is the measure.

$$\text{LPML}_m = \sum_{i=1}^{m} \log p(y_i \mid y^{-i}, \text{Data})$$

where

$$p(y_i \mid y^{-i}, \text{Data}) \approx \left( \frac{1}{S} \frac{1}{p(y_i \mid \theta_s)} \right)^{-1}$$

and $\theta_s$ are posterior samples.