Markov Chain Monte Carlo

Recall: To compute the expectation $\mathbb{E}\big(h(Y)\big)$ we use the approximation

$$\mathbb{E}(h(Y)) \approx \frac{1}{n}\sum_{t=1}^{n} h(Y^{(t)}) \qquad \text{with } Y^{(1)}, \ldots, Y^{(n)} \sim h(y).$$

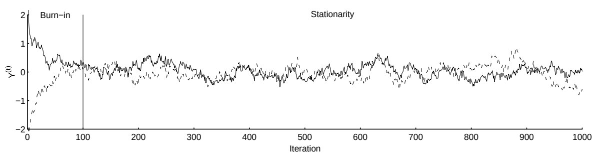
Thus our *aim* is to sample $Y^{(1)}, \ldots, Y^{(n)}$ from $f(y)$.

PROBLEM: Independent sampling from $f(y)$ may be difficult.

Markov chain Monte Carlo (MCMC) approach

- Generate Markov chain $\{Y^{(t)}\}$ with stationary distribution $f(y)$.
- Early iterations $Y^{(1)}, \ldots, Y^{(m)}$ reflect starting value $Y^{(0)}$.
- These iterations are called burn-in.
- After the burn-in, we say the chain has "converged".
- Omit the burn-in from averages:

$$\frac{1}{n-m}\sum_{t=m+1}^{n} h(Y^{(t)})$$



How do we construct a Markov chain $\{Y^{(t)}\}$ which has stationary distribution $f(y)$?

- Gibbs sampler
- Metropolis-Hastings algorithm (Metropolis *et al* 1953; Hastings 1970)

Gibbs Sampler

Let $Y = (Y_1, \ldots, Y_d)$ be d dimensional with $d \geq 2$ and distribution $f(y)$.

The full conditional distribution of Y_i is given by

$$f(y_i|y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_d) = \frac{f(y_1, \ldots, y_{i-1}, y_i, y_{i+1}, \ldots, y_d)}{\int f(y_1, \ldots, y_{i-1}, y_i, y_{i+1}, \ldots, y_d)\, dy_i}$$

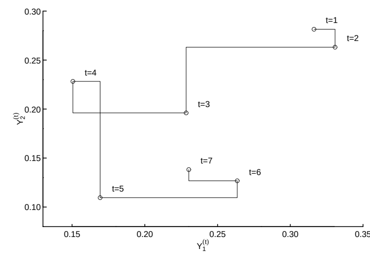
Gibbs sampling

Sample or update in turn:

$$\begin{array}{lll} Y_1^{(t+1)} & \sim & f(y_1|Y_2^{(t)}, Y_3^{(t)}, \ldots, Y_d^{(t)}) \\ Y_2^{(t+1)} & \sim & f(y_2|Y_1^{(t+1)}, Y_3^{(t)}, \ldots, Y_d^{(t)}) \\ Y_3^{(t+1)} & \sim & f(y_3|Y_1^{(t+1)}, Y_2^{(t+1)}, Y_4^{(t)}, \ldots, Y_d^{(t)}) \\ \vdots & \vdots & \vdots \\ Y_d^{(t+1)} & \sim & f(y_d|Y_1^{(t+1)}, Y_2^{(t+1)}, \ldots, Y_{d-1}^{(t+1)}) \end{array}$$

Always use most recent values.

In two dimensions, the sample path of the Gibbs sampler looks like this:



Gibbs Sampler

Detailed balance for Gibbs sampler: For simplicity, let $Y = (Y_1, Y_2)^\mathsf{T}$. Then the update $Y^{(t+1)}$ at time $t+1$ is obtained from the previous $Y^{(t)}$ in two steps:

$$\begin{array}{lll} Y_1^{(t+1)} & \sim & p(y_1|Y_2^{(t)}) \\ Y_2^{(t+1)} & \sim & p(y_2|Y_1^{(t+1)}) \end{array}$$

Accordingly the transition matrix $P(y, y') = \mathbb{P}(Y^{(t+1)} = y'|Y^{(t)} = y)$ can be factorized into two separate transition matrices

$$P(y, y') = P_1(y, \tilde{y})\, P_2(\tilde{y}, y')$$

where $\tilde{y} = (y_1', y_2)^\mathsf{T}$ is the intermediate result after the first step. Obviously we have

$$P_1(y, \tilde{y}) = p(y_1'|y_2) \qquad \text{and} \qquad P_2(\tilde{y}, y') = p(y_2'|y_1').$$

Note that for any y, y', we have $P_1(y, y') = 0$ if $y_2 \neq y_2'$ and $P_2(y, y') = 0$ if $y_1 \neq y_1'$.

According to the detailed balance for time-dependent Markov chains, it suffices to show detailed balance for each of the transition matrices: For any states y, y' such that $y_2 = y_2'$

$$\begin{aligned} p(y)\, P_1(y, y') &= p(y_1, y_2)\, p(y_1'|y_2) = p(y_1|y_2)\, p(y_1', y_2) \\ &= p(y_1|y_2')\, p(y_1', y_2') = P_1(y', y)\, p(y'), \end{aligned}$$

while for y, y' with $y_2 \neq y_2'$ the equation is trivially fulfilled.
Similarly we obtain for y, y' such that $y_1 = y_1'$

$$\begin{aligned} p(y)\, P_2(y, y') &= p(y_1, y_2)\, p(y_2'|y_1) = p(y_2|y_1)\, p(y_2', y_1) \\ &= p(y_2|y_1')\, p(y_1', y_2') = P_2(y', y)\, p(y'), \end{aligned}$$

while for y, y' with $y_1 \neq y_1'$ the equation trivially holds. Altogether this shows that $p(y)$ is indeed the stationary distribution of the Gibbs sampler. Note that combined we get

$$p(y)\, P(y, y') = p(y)\, P_1(y, \tilde{y})\, P_1(\tilde{y}, y') = p(y')\, P_2(y', \tilde{y})\, P_1(\tilde{y}, y) \neq p(y')\, P(y', y).$$

Explanation: Markov chains $\{Y_t\}$ which satisfy the detailed balance equation are called time-reversible since it can be shown that

$$\mathbb{P}(Y_{t+1} = y'|Y_t = y) = \mathbb{P}(Y_t = y|Y_{t+1} = y').$$

For the above Gibbs sampler, to go back in time we have to update the two components in reverse order - first $Y_2^{(t+1)}$ and then $Y_1^{(t+1)}$.

Gibbs Sampler

Example: Bayes inference for a univariate normal sample

Consider normally distributed observations $Y = (Y_1, \ldots, Y_n)^\mathsf{T}$

$$Y_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2).$$

Likelihood function:

$$f(Y|\mu, \sigma^2) \sim \Big(\frac{1}{\sigma^2}\Big)^{\frac{n}{2}} \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \mu)^2\Big)$$

Prior distribution (noninformative prior):

$$\pi(\mu, \sigma^2) \sim \frac{1}{\sigma^2}$$

Posterior distribution:

$$\pi(\mu, \sigma^2|Y) \sim \Big(\frac{1}{\sigma^2}\Big)^{\frac{n}{2}+1} \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \mu)^2\Big)$$

Define $\tau = 1/\sigma^2$. Then we can show that

$$\begin{aligned} \pi(\mu|\sigma^2, Y) &= \mathcal{N}\big(\bar{Y}, \sigma^2/n\big) \\ \pi(\tau|\mu, Y) &= \Gamma\Big(\frac{n}{2}, \frac{1}{2}\sum_{i=1}^{n}(Y_i - \mu)^2\Big) \end{aligned}$$

Gibbs sampler:

$$\begin{aligned} \mu^{(t+1)} &\sim \mathcal{N}\big(\bar{Y}, (n \cdot \tau^{(t)})^{-1}\big) \\ \tau^{(t+1)} &\sim \Gamma\Big(\frac{n}{2}, \frac{1}{2}\sum_{i=1}^{n}(Y_i - \mu^{(t+1)})^2\Big) \end{aligned}$$

with $\sigma^{2\,(t+1)} = 1/\tau^{(t+1)}$

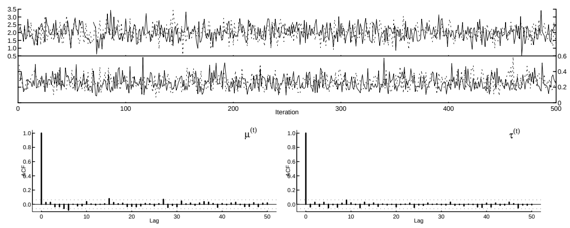
Gibbs Sampler

Implementation in *R*

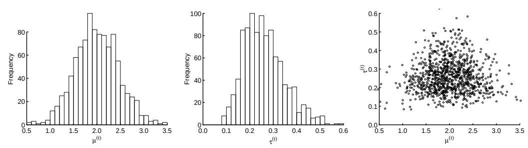
```
n<-20                                        #Data
Y<-rnorm(n,2,2)
MC<-2;N<-1000                                #Run MC=2 chains of length N=1000
p<-rep(0,2*MC*N)                             #Allocate memory for results
dim(p)<-c(2,MC,N)
for (j in (1:MC)) {                          #Loop over chains
  p2<-rgamma(1,n/2,1/2)                      #Starting value for tau
  for (i in (1:N)) {                         #Gibbs iterations
    p1<-rnorm(1,mean(Y),sqrt(1/(p2*n)))      #Update mu
    p2<-rgamma(1,n/2,sum((Y-p1)^2)/2)        #Update tau
    p[1,j,i]<-p1                             #Save results
    p[2,j,i]<-p2
  }
}
```

Results: Bayes inference for a univariate normal sample

Two runs of Gibbs sampler (N=500):



Marginal and joint posterior distributions (based on 1000 draws):



Markov Chain Monte Carlo

Example: Bivariate normal distribtution

Let $Y = (Y_1, Y_2)^\top$ be normally distributed with mean $\mu = (0,0)^\top$ and covariance matrix

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

The conditional distributions are

$$Y_1|Y_2 = \mathcal{N}(\rho\, Y_2, 1-\rho^2)$$
$$Y_2|Y_1 = \mathcal{N}(\rho\, Y_1, 1-\rho^2)$$

Thus the steps of the Gibbs sampler are

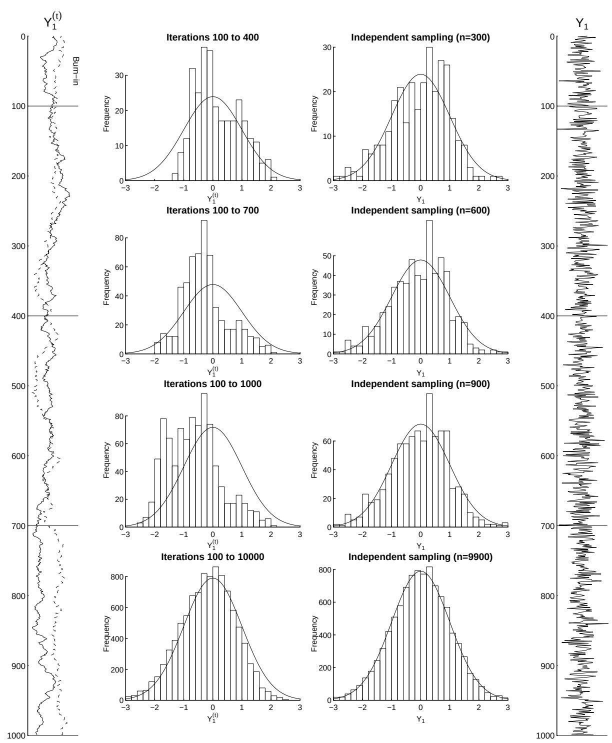
$$Y_1^{(t+1)} \sim \mathcal{N}(\rho\, Y_2^{(t)}, 1-\rho^2),$$
$$Y_2^{(t+1)} \sim \mathcal{N}(\rho\, Y_1^{(t+1)}, 1-\rho^2).$$

NOTE: We can obtain an independent sample $Y^{(t)} = (Y_1^{(t)}, Y_2^{(t)})^\top$ by

$$Y_1^{(t+1)} \sim \mathcal{N}(0,1),$$
$$Y_2^{(t+1)} \sim \mathcal{N}(\rho\, Y_1^{(t+1)}, 1-\rho^2).$$

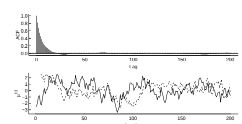
Markov Chain Monte Carlo

Comparison of MCMC and independent draws



Markov Chain Monte Carlo

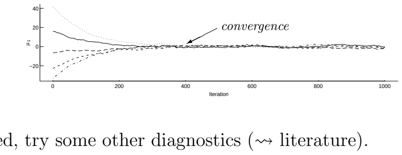
Convergence diagnostics



- Plot chain for each quantity of interest.
- Plot auto-correlation function (ACF)

  $$\rho_i(h) = \operatorname{corr}\left(Y_i^{(t)}, Y_i^{(t+h)}\right).$$

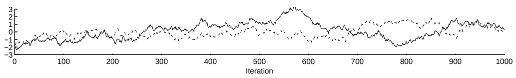
  measures the correlation of values h lags apart.
  - ◦ Slow decay of ACF indicates slow convergence and bad mixing.
  - ◦ Can be used to find independent subsample.
- Run multiple, independent chains (e.g. 3-10).
  - ◦ Several long runs (Gelman and Rubin 1992)
    - · gives indication of convergence
    - · a sense of statistical security
  - ◦ one very long run (Geyer, 1992)
    - · reaches parts other schemes cannot reach.
- Widely dispersed starting values are particularly helpful to detect slow convergence.

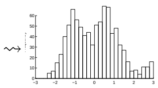


If not satisfied, try some other diagnostics (⇝ literature).

Markov Chain Monte Carlo

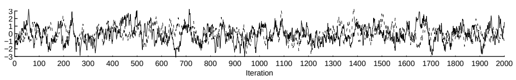
Note: Even after the chain reached convergence, it might not yet good enough for estimating $\mathbb{E}(h(Y))$.

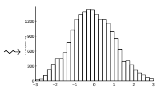


$\rightsquigarrow$


Problem: Chain should show good mixing (transition between states)

$\rightsquigarrow$ run the chain for a longer period



$\rightsquigarrow$


Monte Carlo error

Suppose we want to estimate $\mathbb{E}\big(g(Y)\big)$ by

$$\hat{h} = \frac{1}{N}\sum_{t=1}^{N} h(Y^{(t)}) \qquad \text{with } Y^{(t)} \sim f(y).$$

The error of the approximation (*Monte Carlo error*) is $\sqrt{\mathrm{var}(\hat{h})}$.

Estimation of Monte Carlo error:

Let $\{Y^{(i,t)}\}$ be I Markov chains. Then $\mathrm{var}(\hat{h})$ can be estimated by

$$\frac{1}{I(I-1)}\sum_{i=1}^{I}(\hat{h}^{(i)} - \hat{h})^2$$

where
- ∘ $\hat{h}^{(i)}$ is the MCMC estimate based in the ith chain
- ∘ $\hat{h}$ is the average of the $\hat{h}^{(i)}$ (overall estimate)