Sec 7.1.1 pg 186

Independence Chains

If the proposal in M-H is $q(x^*|x^{(t)}) = q(x^*)$ then we have an ind. chain in which each candidate value is drawn ind. of the path. and r becomes

$$r = \frac{f(x^*)g(x^{(t)})}{f(x^{(t)})g(x^*)}$$

The resulting M.C. is ergodic. if $g(x)>0$ whenever $f(x)>0$.
Also, r can be rewritten as the ratio of importance ratios $w = f(x)/g(x)$. So that $r = w^*/w^{(t)}$. Here if $w^{(t)}$ is much larger than typical w^* chain gets stuck.
- also choosing g like choosing e.
- works well if g is a good imitator of f.

pg 187 Ex 7.2
- observed $y_1, \ldots, y_{100}$ from mixture $\delta N(7, 0.5^2) + (1-\delta) N(10, .5^2)$
- see histogram.
- want posterior density of δ when $U(0,1)$ prior. (ex. $\delta = .7$) so post. density should concentrate here

- Choose 2 proposal
  - $\beta(1,1)$ sim to $U(0,1)$ mean $\frac{1}{2}$.
  - $\beta(2,10)$ skewed right w/ mean .167 so values near .7 unlikely.

Histograms pg 189 Sample plots pg 189.
(1) moves quickly near .7 & seems to support all points in post
(2) gets stuck a lot. has not converged.

Sec 7.1.2 Random Walk Chains.

Another variant of M-H. Let X^* be generated by drawing $\varepsilon \sim h(\varepsilon)$ for some density h & setting $X^* = x^{(t)} + \varepsilon$.

Here $g(x^*|x^{(t)}) = h(x^* - x^{(t)})$.

Common h: - Uniform ball centered at origin
- scaled normal
- scaled Student's t.

If the support of f is connected & h is pos in a neighborhood of 0 $\Rightarrow$ ergodic

Ex/ Fig 7.4. pg 190 -
Note at each step proposal centered at X_t.

Other variations discussed.

These ideas do extend to multivariate r.v. However, often acceptance rate drops as dim goes up.

There are better ways to extend these MC / conditional dist. ideas to higher dimensional random variables.

Sec 7.2 Gibbs Sampling

Suppose we want to sample from a multi-dim target distribution. For multivariate dist. with a large number of variables, the standard acceptance/rejection method is difficult to apply because it is difficult to usable proposal distribution. Often, it is not very effecient because of the high rejection rate.

The Gibbs sampler is specifically designed for multi-dim target distributions. The goal (like in M-H) is to develop a Markov chain whose stationary dist equals the target f. However, the basic idea is different than prior methods (i.e. we don't draw the vector X from g as a whole).

Basic idea: Build up the random vector, element by element, by sequentially sampling from univariate conditional distributions.

Sec 7.2.1 Basic Gibbs Sampler.

Let $\underline{X} = (X_1, X_2, \ldots, X_p)^T$ and denote

$$\underline{X}_{-i} = (X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_p)^T$$

Suppose we want information about target f. but cannot sample easily from it. However, suppose we can sample the univariate conditional density of $X_i | \underline{X}_{-i} = x_{-i}$ denoted $f(x_i | x_{-i})$. $\forall i = 1, \ldots, p$

Gibbs Sampling Alg. (Geman & Geman)

1. Select starting value $\underline{X}^{(0)} = (X_1^{(0)}, X_2^{(0)}, \ldots, X_p^{(0)})$ and set $t=0$.

2. Generate, in turn,

called a cycle

$X_1^{(t+1)}$ drawn from $f_1(X_1 \mid X_2^{(t)}, X_3^{(t)}, \ldots, X_p^{(t)})$
$X_2^{(t+1)}$ drawn from $f_2(X_2 \mid X_1^{(t+1)}, X_3^{(t)}, \ldots, X_p^{(t)})$
$\vdots$
$X_i^{(t+1)}$ drawn from $f_i(X_i \mid X_1^{(t+1)}, \ldots, X_{i-1}^{(t+1)}, X_{i+1}^{(t)}, \ldots, X_p^{(t)})$
$\vdots$
$X_p^{(t+1)}$ drawn from $f_p(X_p \mid X_1^{(t+1)}, \ldots, X_{p-1}^{(t+1)})$

3. Increment t and go to step 2.

Comments:

① At each draw in step 2. we are conditioning on the most recent update to all other elements.

② The densities $f_1, f_2, \ldots, f_p$ are called the full conditionals. Gibbs sampling only requires the full conditionals.

③ Even for high-dim problems, all of the simulations are univariate. Obviously advantageous, but can be slow.

Chain is Markov

④ Once convergence of the MC is achieved $X^{(t)} = (X_1^{(t)}, \ldots, X_p^{(t)})$ is from f, with each component individualy converging. (it converges in distribution). Also each X_i drawn from marginalization of f.

Simple Example

- Let's create a gibbs sampler for drawing samples from a bivariate normal dist. with

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \text{ & } \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

want

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N(\mu, \Sigma)$$

We need the conditionals. (HW?)

$$[X_1 | X_2] \sim N\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(X_2 - \mu_2), \sigma_1^2(1-\rho^2)\right)$$

and

$$[X_2 | X_1] \sim N\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(X_1 - \mu_1), \sigma_2^2(1-\rho^2)\right)$$

So the gibbs sampler is

$X_1^{(t+1)}$ is drawn from $f(X_1 | X_2^{(t)}) = \dfrac{e^{-\frac{1}{2}\left\{\frac{x - (\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(X_2^{(t)} - \mu_2))}{\sigma_1^2(1-\rho^2)}\right\}}}{\sqrt{2\pi\sigma_1^2(1-\rho^2)}}$

and

$X_2^{(t+1)}$ is drawn from $f(X_2 | X_1^{(t+1)}) = \dfrac{e^{-\frac{1}{2}\frac{(x - (\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(X_1^{(t+1)} - \mu_1)))^2}{\sigma_2^2(1-\rho^2)}}}{\sqrt{2\pi\sigma_2^2(1-\rho^2)}}$.

Ex 7.4 pg 196.

$Y = (Y_1, \ldots, Y_c)$ denote the counts of insects of diff classes.

$P = (P_1, \ldots, P_c)$ denote prob of each class & depend on $\alpha_1, \ldots, \alpha_c$

N total # of insects collected & depend on λ.

Want to compare $T_1(Y)$ & $T_2(Y)$.

Need to simulate Y so that T_1 & T_2 are able to be calculated. → Use Markov Chain.

Let $c = 3$

Model

$$\begin{cases} (Y_1, Y_2, Y_3) \mid (N=n, P_1=p_1, P_2=p_2, P_3=p_3) \sim \text{Multinomial}(n; p_1, p_2, p_3) \\ (P_1, P_2, P_3) \sim \text{Dirichlet}(\alpha_1, \alpha_2, \alpha_3) \\ N \sim \text{Poisson}(\lambda) \end{cases}$$

How to sample (Y_1, Y_2, Y_3). It is impossible to get a closed form expression for this marginal given $\lambda, \alpha_1, \alpha_2$ & α_3.

could use direct sim. approach

We will use a Gibbs sampler.

- $Y_1 + Y_2 + Y_3 = N$
  $P_1 + P_2 + P_3 = 1$

$$\underline{X} = (Y_1, Y_2, P_1, P_2, N)$$

We need the conditionals. (in 2 steps)

(HW?)

First note:

$$(Y_1, Y_2, Y_3) \mid (N=n, P_1=p_1, P_2=p_2, P_3=p_3) \sim \text{Multinomial}(n; p_1, p_2, p_3)$$

$$(P_1, P_2, P_3) \mid (Y_1=y_1, Y_2=y_2, Y_3=n-y_1-y_2, N=n) \sim \text{Dirichlet}(y_1+\alpha_1, y_2+\alpha_2, n-y_1-y_2+\alpha_3).$$

$$N - y_1 - y_2 \mid (Y_1=y_1, Y_2=y_2, Y_3=y_3, P_1=p_1, P_2=p_2, P_3=p_3) \sim \text{Poisson}(\lambda(1-p_1-p_2)).$$

∴ The gibbs sampler is.

$$Y_1^{(t+1)} \text{ from } \text{Bin}\left(n^{(t)} - y_2^{(t)}, \frac{p_1^{(t)}}{1-p_2^{(t)}}\right)$$

$$Y_2^{(t+1)} \text{ from } \text{Bin}\left(n^{(t)} - y_1^{(t+1)}, \frac{p_1^{(t)}}{1-p_2^{(t)}}\right)$$

$\frac{p_1^{(t+1)}}{1-p_2^{(t)}}$ from Beta$(y_1^{(t+1)}+\alpha_1, n^{(t)}-y_1^{(t+1)}-y_2^{(t+1)}+\alpha_3)$

$\frac{p_2^{(t+1)}}{1-p_1^{(t+1)}}$ from Beta$(y_2^{(t+1)}+\alpha_2, n^{(t)}-y_1^{(t+1)}-y_2^{(t+1)}+\alpha_3)$

and

$N^{(t+1)}-y_1^{(t+1)}-y_2^{(t+1)}$ from Poisson$(\lambda(1-p_1^{(t+1)}-p_2^{(t+1)}))$.

Further comments on Gibbs sampler.

7.2.2. We can relate Gibbs sampler to M-H. in which the proposal is allowed to vary over time. → Then each Gibbs cycle consists of p M-H steps.

For each i Gibbs proposes

$$X^* = (x_1^{(t+1)}, \ldots, x_{i-1}^{(t+1)}, x_i^*, x_{i+1}^{(t)}, \ldots, x_p^{(t)})$$

ie drawing X^* from $f(x_i^* | x_1^{(t+1)}, \ldots, x_{i-1}^{(t+1)}, x_{i+1}^{(t)} \ldots x_p^{(t)})$.

Then the M-H ratio can be shown to be 1

$$\frac{f(\text{new})\, g(\text{old}|\text{new})}{f(\text{old})\, g(\text{new}|\text{old})} =$$

and so X^* is always accepted.

OMIT ??

you may find these in the literature.

Variations / Generalizations

7.2.3 – The ordering of the updates can change from cycle to cycle. Sometimes a random ordering is used. R.scan G.S. may not update each i every time.

7.2.4 Blocking

You don't need to each element of X individually.

For example 7.4 it was natural to/easy to generate each of (Y_1, Y_2, Y_3), (P_1, P_2, P_3), $\vdots N$ conditionally as a group.

Ex

$$X_1^{(t+1)} \sim f(X_1 \mid X_2^{(t)}, X_3^{(t)}, X_4^{(t)})$$

$$X_2^{(t+1)}, X_3^{(t+1)} \sim f(X_2, X_3 \mid X_1^{(t+1)}, X_4^{(t)})$$

$$X_4^{(t+1)} \sim f(X_4 \mid X_1^{(t+1)}, X_2^{(t+1)}, X_3^{(t+1)})$$

Blocking is useful when some elements of X are correlated. It can converge faster.

7.2.5 Hybrid Gibbs Sampling.
- add Metropolis Hasting steps were convenient.

* This is particularly useful if the univariate conditional density for one or more elements of X is not available in closed form. → you still get a MC.

Sec 7.2.6 discusses other methods.

These alg. are rel. simple, ~~also~~ in theory so now we turn to implementation topics.

Sec 7.3 Implementation.

pg 200 Recall Goal of mcmc: To estimate features of f

$$u = \int h(x) f(x) dx.$$

How good the estimators are depends on how reliably the sample (mc) averages corres. to their expectation under the limiting dist of the mc. Now, the mcmc methods we have discussed all converge (in theory) to the correct limiting dist. However, in practice, we need to know how long to run the chain so that it adequately represents the target dist f and will have reliable estimates. Some times convergence is slow or misleading.

So, We need to ask:

- Has the chain run long enough?
- In the 1st part influenced by the starting value?
- Should the chain be run from several starting values?
- Are are sampled θ values approx draws from f?
- How shall we use the chain output to produce estimates & asses their precision?

7.3.1 pg 201

Ensuring good mixing & convergence.

There are two main concerns w/ a MCMC algorithm

① The mixing prop. of the chain - how quickly it forgets its starting value, how quickly it fully explores the support of the target dist, how far apart observat. need to be before they are approx independent.

② The convergence of the chain - when has it approx. reached its stationary dist.

These topics overlap & we will briefly discuss a variety of (diagnostic) techniques

7.3.1.1 — Choice of Proposal: Clearly mixing is strongly affected by features of the proposal.

For M-H we want:
- nice if g approx f well so the acceptance rate is high
- prefer f/g to be bounded → yielding faster convergence of MC
- this ⇒ g must be more diffuse than f.
- they suggest an iterative process where you adjust the variance of your proposal to acheive a desired acceptance rate. ~25% to 45%.

For Gibbs we want:
- components of X to be as ind. as possible
- they suggest reparameterization. to reduce dependence.

7.3.1.2. — Number of chains: How can you tell if your chain has become stuck in one or more modes. → hard. since other diag may indicate convergence.

Partial soln: run multiple chains from diverse starting values and compare.

This is a point of disagreement among stat. many of whom argue for a single longer run. They recommend several short "diag" runs & then 1 long run from a "good" starting value.

7.3.3 — Simple graphs to assess mixing & conv.

(1) Sample path: A plot of the iteration number t versus the realizations $X^{(t)}$.
- also called trace or history plots.
- if the chain is mixing poorly it will remain at or near the same value for many iter.
- good mixing quickly moves away from starting values and "wiggles about vigorously" in the region supported by f.

(2) Cusum (cumulative sum) diagnostic plot:
- asses the convergence of a one dim para $\Theta = E\{h(X)\}$.
- After discarding initial iterations: calculate the estimator $\hat{\Theta}_n = \frac{1}{n}\sum_{j=1}^{n} h(X^{(j)})$.

then plot

$$\sum_{i=1}^{t}\left[h(X^{(i)}) - \hat{\Theta}_n\right] \text{ v/s } t$$

- good mixing has wiggly plot w/ small excursions from zero v/s smooth w/ large.
- be careful if stuck in one of the modes.

Example pg 204 (3) Autocorrelation plot: plot i v/s correl. between interations that are i apart. Want steady decay. Slow decay indicates Poor mixing.

7.3.1.4

- Reparameterization to improve mixing.

Ex $[X_1, X_2]$ bivariate Normal.

Gibbs sampler explores f slowly. So try $Y = [X_1 + X_2, X_1 - X_2]$ → uniquely det. $[X_1, X_2]$ but explores faster.

7.3.1.5

- Burn-in & run length.

$X^{(t)} \sim f$ only in the limit. Usually throw away initial D iterations to reduce dependence on the starting value, but still need enough iterations, L.

They suggest the following calculations to determine if D was large enough.

Run J chains of length L from varied starting values, & discard 1st D.

Chain 1 $X_1^{(0)}, \ldots, X_1^{(D)}, X_1^{(D+1)}, \ldots, X_1^{(D+L-1)}$

⋮

Chain J $X_J^{(0)}, \ldots, X_J^{(D)}, X_J^{(D+1)}, \ldots, X_J^{(D+L-1)}$

Let $\bar{X}_j = \frac{1}{L} \sum_{t=D}^{D+L-1} X_j^{(t)}$ & $\bar{\bar{X}} = \frac{1}{J} \sum_{j=1}^{J} \bar{X}_j$

mean of jth chain mean overall.

Define Between-chain variance.

$$B = \frac{L}{J-1} \sum_{j=1}^{J} (\bar{X}_j - \bar{\bar{X}})^2$$

and within-chain variance as

$$s_j^2 = \frac{1}{L-1} \sum_{t=D}^{D+L-1} (x_j^{(t)} - \bar{x}_j)^2$$

with overall ave

$$W = \frac{1}{J} \sum_{j=1}^{J} s_j^2 .$$

The let $$R = \frac{\frac{L-1}{L} \cdot W + \frac{1}{L} B}{W} .$$

If good mixing is occuring then both num & denom should estimate the marginal variance of X and $\sqrt{R} \to 1$ as $L \to \infty$.

Many suggest $\sqrt{R} < 1.2$ as acceptable $\Rightarrow$ that D was large enough & L was large enough.

Authors overall advice:

If you have the computing power
- run several chains
- carry out diag. to determine one that is behaving well
- restart the chain for a final long run.

They give a very detailed fur-pup example in Sec 7.4.