Some notes on Markov Chain Monte Carlo: Metropolis-Hastings and Gibbs Steps for Exploring Posterior Distributions

1 Practical Bayesian Inference

In the Bayesian paradigm we build a *joint model* for *data* $\underline{y} = (y_1, \ldots, y_N)$, and *parameters* τ:

$$p(\underline{y}, \tau) \;\; = \;\; p(\underline{y}|\tau) \cdot p(\tau)$$

and we always want to know something about the *posterior density*

$$\begin{aligned} p(\tau|\underline{y}) \;\; &= \;\; \frac{p(\underline{y}|\tau) \cdot p(\tau)}{\int p(\underline{y}|t) \cdot p(t)\; dt} \\ &\propto \;\; p(\underline{y}|\tau) \cdot p(\tau) \;\; = \;\; p(\underline{y}, \tau) \end{aligned}$$

For example,

- *Posterior mode*: $argmax_\tau \, p(\tau|\underline{y})$
- *Posterior mean*: $E[\tau|\underline{y}] = \int \tau \, p(\tau|\underline{y}) \, d\tau$ (or some other risk-minimizer)
- *Credible interval*: Find a set A such that $p(\tau \in A|\underline{y}) = 0.95$
- *Graph or "shape"* of $p(\tau|\underline{y})$

<u>All</u> of these operations require that you know a lot about the shape of $p(\tau|\underline{y})$. *How do you learn enough about the shape of* $p(\tau|\underline{y})$*?*

- If $p(\tau)$ is *conjugate* to $p(\underline{y}|\tau)$, then $p(\tau|\underline{y})$ will belong to the same family as $p(\tau)$.
- If not, then you have a numerical integration problem. These problems are difficult because they depend on the "interesting" features of $p(\tau|\underline{y})$.
  - Difficult but do-able if $p(\tau|\underline{y})$ is *smooth* and *low-dimensional*.
  - Even if $p(\tau|\underline{y})$ is smooth, as $d = \dim(\tau)$ grows, space becomes too large for deterministic methods. *Curse of dimensionality:* For n "evaluations" of $p(\tau|\underline{y})$, learn at rate $O(1/n^{k/d})$.
- Monte Carlo (MC) methods have long been used to help with the curse. With a random sample $\tau^{(1)}, \tau^{(2)}, \ldots, \tau^{(n)}$, can plot histograms (shape), compute averages (integration), etc.
  - Improves convergence to $O(1/\sqrt{n})$
  - Less smoothing of interesting features as n grows
  - Curse returns: how do you sample from $p(\tau|\underline{y})$?

2 MCMC: The Cliff's Notes

The problem:

- Learn about $\pi(\tau) = \pi(\theta, \beta, \lambda)$, where $\tau = (\theta, \beta, \lambda)$ is some high-dimensional set of variables (parameters).

The essential idea:

- Define a (stationary) Markov chain $\mathcal{M}_0, \mathcal{M}_1, \mathcal{M}_2, \ldots$ with states $\mathcal{M}_k = (\theta^k, \beta^k, \lambda^k)$.
- Simulate the $\mathcal{M}_k$'s; under regularity conditions (e.g., Tierney, 1994), $\mathcal{M}_k$ will converge in distribution to a *stationary distribution* $\pi(\theta, \beta, \lambda)$:

$$\pi(\theta^1, \beta^1, \lambda^1) \;=\; \int P[\mathcal{M}_{k+1} = (\theta^1, \beta^1, \lambda^1) \mid \mathcal{M}_k = (\theta^0, \beta^0, \lambda^0)] \;\times\; \pi(\theta^0, \beta^0, \lambda^0)\, d(\theta^0, \beta^0, \lambda^0)$$

- For Bayes, design chain so that $\pi(\theta, \beta, \lambda)$ turns out to be the posterior, $p(\theta, \beta, \lambda | \underline{y})$.

Typical MCMC paradigm:

- Write $p(\theta, \beta, \lambda | \underline{y}) = p(\tau_1, \tau_2, \ldots, \tau_M | \underline{y})$, where $(\tau_1, \ldots, \tau_M)$ is a disjoint partition of the original model parameters (θ, β, λ).
- General theory of MCMC (e.g., Tierney, 1994; Chib and Greenberg, 1995): construct state $\mathcal{M}_k = (\tau_1^{(k)}, \ldots, \tau_M^{(k)})$, by sampling each τ_m from its "complete conditional" distribution:

  To step from $\mathcal{M}_{k-1} = (\tau_1^{(k-1)}, \ldots, \tau_M^{(k-1)})$ to $\mathcal{M}_k = (\tau_1^{(k)}, \ldots, \tau_M^{(k)})$:

  1. $\tau_1^{(k)} \sim p(\tau_1 | \tau_2^{(k-1)}, \ldots, \tau_M^{(k-1)}, \underline{y})$;

  2. $\tau_2^{(k)} \sim p(\tau_2 | \tau_1^{(k)}, \tau_3^{(k-1)}, \ldots, \tau_M^{(k-1)}, \underline{y})$;

  3. $\tau_3^{(k)} \sim p(\tau_3 | \tau_1^{(k)}, \tau_2^{(k)}, \tau_4^{(k-1)}, \ldots, \tau_M^{(k-1)}, \underline{y})$;

  $\vdots \qquad \vdots \qquad \vdots \qquad \vdots$

  M. $\tau_M^{(k)} \sim p(\tau_M | \tau_1^{(k)}, \tau_2^{(k)}, \tau_3^{(k)}, \ldots, \tau_{M-1}^{(k)}, \underline{y},)$;

  This is the basic idea behind all so-called "Gibbs samplers."

Markov Chain Monte Carlo (MCMC) vs. MC:

- Replace draws from $p(\tau | \underline{y})$ with a sequence of low-dimensional draws from $p_1(\tau_1)$, $p_2(\tau_2)$, etc.
- For Bayes, often can exploit "partial" conjugacy to simplify problem of drawing from $p(\tau_{(k)})$.
- *Costs:* dependent rather than iid samples, approximate rather than exact sampling of $p(\tau | \underline{y})$.

3 Implementation Details

Sampling schemes

Once a set of complete conditionals is decided upon:

- *If you can* sample τ_m directly from its complete conditional distribution, do so—this is a *Gibbs step*;
- *Otherwise* perform a *Metropolis step*:
  - Sample a "proposal value" τ_m^* from any convenient proposal distribution $q_k(\tau_m|\tau_m^{(k-1)})$
  - Compute the "acceptance probability" α_m^* from $q_k(\tau_m|\tau_m^{(k-1)})$ and the complete conditional $p(\tau_m|rest) \equiv p(\tau_m \mid \tau_1^{(k)}, \ldots, \tau_{m-1}^{(k)}, \tau_{m+1}^{(k-1)}, \ldots, \tau_M^{(k-1)}, \underline{y})$:

$$\alpha_n^* = \min\left\{ \frac{p(\tau_n^*|rest)q_k(\tau_n^{(k-1)}|\tau_n^*)}{p(\tau_n^{(k-1)}|rest)q_k(\tau_n^*|\tau_n^{(k-1)})}, 1 \right\};$$

  - Accept $\tau_n^{(k)} = \tau_n^*$ with probability α_n^*; otherwise set $\tau_n^{(k)} = \tau_n^{(k-1)}$.

This is often called "Metropolis-Hastings within Gibbs".

Identifying the complete conditionals

Suppose the θ's, β's and λ's have prior distributions $p(\theta|\lambda)$, $p(\beta)$ and $p(\lambda)$. Then the "complete conditional" for θ, for example, is

$$\begin{aligned} p(\theta|rest) = p(\theta|\underline{y},\beta,\lambda) &= \frac{p(\underline{y},\theta,\beta,\lambda)}{\int p(\underline{y},t,\beta,\lambda)\,dt} \\ &= \frac{p(\underline{y}|\theta,\beta)\,p(\theta|\lambda)\;p(\beta)\;p(\lambda)}{\int p(\underline{y}|t,\beta)\,p(t|\lambda)\,p(\beta)\,p(\lambda)\,dt} \\ &\propto p(\underline{y}|\theta,\beta)\,p(\theta|\lambda) \end{aligned}$$

Key observation: *The shape of $p(\theta|\underline{y},\beta,\lambda)$ is determined by just the parts of the likelihood that depend explicitly on θ.*

Convergence

$$\underbrace{\mathcal{M}_0, \mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_k,}_{\text{“Burn-in” segment}} \qquad \underbrace{\mathcal{M}_{k+1}, \ldots, \mathcal{M}_{k+n}}_{\text{Usable MCMC sample}}$$

How large should k be?

- Use `acf()` in Splus/R to check when $Corr(\mathcal{M}_\ell, \mathcal{M}_{\ell+k})$ small;
- Use time series plots to see when the chain has "stabilized".
- Run several chains from different starting points; let k be so large that *(within-chain variation) < (between chain variation).*
- `CODA` subroutine package for Splus/R offers a menu of such checks (www.mrc-bsu.cam.ac.uk/bugs/).

4 Example: The hierarchical beta-binomial model

Non-hierarchical version

Basic beta-binomial model, for examinees $i = 1, \ldots, N$, each getting y_i of n_i questions right. [$\underline{y} = (y_1, \ldots, y_N)$]

1. $y_i|\theta_i \sim Bin(\theta_i, n_i)$;
2. $\theta_i|\alpha, \beta \sim Beta(\alpha, \beta)$

Full model $p(\underline{y}, \underline{\theta}|\alpha, \beta)$ is

$$\prod_{i=1}^{N} \binom{n_i}{y_i} \theta_i^{y_i}(1-\theta_i)^{n_i-y_i} \prod_{i=1}^{N} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_i^{\alpha-1}(1-\theta_i)^{\beta-1}$$

Use the "key observation" to calculate

$$p(\theta_i|y_i, \alpha, \beta) \propto \theta_i^{\alpha+y_i-1}(1-\theta_i)^{\beta+n_i-y_i-1}$$

i.e. $\theta_i|y_i \sim Beta(\alpha + y_i, \beta + n_i - y_i)$.

Choose fixed values for α, β, then use $Beta(\alpha + y_i, \beta + n_i - y_i)$ for posterior mode, posterior mean, credible interval for θ, etc.

No need for anything fancy yet.

Hierarchical version

What if we want to estimate α, β (estimate the shape of the population / latent distribution)? Add a third modeling assumption:

3. $\alpha, \beta \sim p(\alpha, \beta)$

Now the full model $p(\underline{y}, \underline{\theta}, \alpha, \beta)$ is

$$\prod_{i=1}^{N} \binom{n_i}{y_i} \theta_i^{y_i} (1-\theta_i)^{n_i - y_i} \prod_{i=1}^{N} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_i^{\alpha-1} (1-\theta_i)^{\beta-1} p(\alpha, \beta)$$

Can again use the "key observation" to calculate

$$\begin{aligned} p(\theta_i | y_i, \alpha, \beta) &= Beta(\theta_i | \alpha + y_i, \beta + n_i - y_i) \\ p(\alpha, \beta | \underline{y}) &\propto p(\alpha, \beta) \prod_{i=1}^{N} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha + y_i)\Gamma(\beta + n_i - y_i)}{\Gamma(\alpha + \beta + n_i)} \end{aligned}$$

An MCMC solution

From the full model $p(\underline{y}, \underline{\theta}, \alpha, \beta)$ is

$$\prod_{i=1}^{N} \binom{n_i}{y_i} \theta_i^{y_i} (1-\theta_i)^{n_i - y_i} \prod_{i=1}^{N} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_i^{\alpha-1} (1-\theta_i)^{\beta-1} p(\alpha, \beta)$$

the "key observation" gives the complete conditionals

$$\begin{aligned} p(\theta_i | rest) &= Beta(\theta_i | \alpha + y_i, \beta + n_i - y_i) \\ p(\alpha | rest) &\propto \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)} \right]^N \prod_{i=1}^{N} \theta_i^{\alpha} p(\alpha, \beta) \\ p(\beta | rest) &\propto \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\beta)} \right]^N \prod_{i=1}^{N} (1-\theta_i)^{\beta} p(\alpha, \beta) \end{aligned}$$

- Usually take $p(\alpha, \beta) = p(\alpha)p(\beta)$; in this case we could take independent Gamma or Inv-Gamma densities, e.g.
- Canned routines sample $\theta_i \sim Beta(\cdots)$
- Can set up Metropolis steps for α and β using Normal proposal draws ("random walk M-H"). Normal variances are "tuning parameters".