Gibbs sampling - another look

Scott C. Schmidler

Stat 863: Advanced Statistical Computing
Duke University
Fall 2018

Gibbs sampling

We've seen the effects of *dependence* on Gibbs sampler efficiency:

- Slow mixing of Gibbs sampler for Ising model
- Strong autocorrelation/slow convergence for for multivariate normal

Let's take a closer look at what's going on.

Gibbs sampling

The Gibbs sampler for $\theta = (\theta_1, \ldots, \theta_d)$ proceeds by cycling through conditional distributions:

$$\theta_i^{(n)} \sim \pi(\theta_i \mid \theta_1^{(n)}, \ldots, \theta_{i-1}^{(n)}, \theta_{i+1}^{(n-1)}, \ldots, \theta_d^{(n-1)})$$

An alternative is the *random-scan* Gibbs sampler, which iteratively chooses $i \in \{1, \ldots, d\}$ at random (according to $P(i)$ say), and sets

$$\theta^{(n+1)} = (\theta_1^{(n)}, \ldots, \theta_{i-1}^{(n)}, \theta_i^*, \theta_{i+1}^{(n)}, \ldots, \theta_d^{(n)})$$

with

$$\theta_i^* \sim \pi(\theta_i \mid \theta_1^{(n)}, \ldots, \theta_{i-1}^{(n)}, \theta_{i+1}^{(n)}, \ldots, \theta_d^{(n)})$$

Component-wise MH

Special case: What if we can draw from $\pi(x_1 \mid x_2)$ exactly?
As in above example: conditional densities are "nice".

Then

$$\begin{aligned}\alpha(x_1, y_1 \mid x_2) &= \min\left(1 \frac{\pi(y_1 \mid x_2) q_1(y_1, x_1 \mid x_2)}{\pi(x_1 \mid x_2) q_1(x_1, y_1 \mid x_2)}\right) \\ &= \min\left(1 \frac{\pi(y_1 \mid x_2) \pi(x_1 \mid x_2)}{\pi(x_1 \mid x_2) \pi(y_1 \mid x_2)}\right) \\ &\equiv 1\end{aligned}$$

so moves are *always accepted*.

If can do this for all the conditionals, call this a *Gibbs sampler*.

Gibbs sampler

Let $x = (x_1, \ldots, x_p)$. x_i may be uni- or multi-dimensional.

Suppose we can draw from *conditional* distributions $\pi(x_i \mid x_{[-i]})$ where $x_{[-i]} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p)$.

Given $x^{(t)} = (x_1^{(t)}, \ldots, x_p^{(t)})$, Gibbs sampling proceeds by:

$$\begin{aligned}&\text{Draw } x_1^{(t+1)} \sim \pi(x_1 \mid x_2^{(t)}, \ldots, x_p^{(t)}) \\ &\text{Draw } x_2^{(t+1)} \sim \pi(x_2 \mid x_1^{(t+1)}, x_3^{(t)}, \ldots, x_p^{(t)}) \\ &\qquad \vdots \\ &\text{Draw } x_i^{(t+1)} \sim \pi(x_i \mid x_1^{(t+1)}, \ldots, x_{i-1}^{(t+1)}, x_{i+1}^{(t)}, \ldots, x_p^{(t)}) \\ &\qquad \vdots \\ &\text{Draw } x_p^{(t+1)} \sim \pi(x_p \mid x_1^{(t+1)}, \ldots, x_{p-1}^{(t+1)})\end{aligned}$$

and iterate. *(successive substitution sampling)*

Gibbs sampling

Note: ordering may be fixed (systematic scan) or random (random scan).

Example: Bivariate Normal $x \sim N_2(\mu, \Sigma)$ with $\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$

$$\begin{aligned}\text{Draw } x_1^{(t+1)} &\sim N(\rho x_2^{(t)}, 1 - \rho^2) \\ x_2^{(t+1)} &\sim N(\rho x_1^{(t+1)}, 1 - \rho^2)\end{aligned}$$

Iterate

Gibbs sampling

Example: 2D Ising model

$\sigma_i \in \{-1, 1\} \quad \sigma = (\sigma_1, \ldots, \sigma_n)$.

$\pi(\sigma) = Z^{-1} \exp(-H(\sigma)) \quad \text{for} \quad H(\sigma) = -J \sum_{i \sim j} \sigma_i \sigma_j$

Then as we've seen $\pi(\sigma_i = 1 \mid \sigma_{j \neq i}) = \frac{1}{1+\exp(2J \sum_{i \sim j} \sigma_j)}$[1]

So *easy* to draw from conditionals.

Compare this to original Metropolis alg: Gibbs sampler has no rejection.

[1] Becomes $J(\sigma_{i-1} + \sigma_{i+1})$ in 1D

A note on hybrid chains

Composition of kernels need not inherity irreducibility and aperiodicity of parts.

E.g. P_1, P_2 ϕ-irreduc, aperiod, π-invariant; $P_1 \circ P_2$ may not even be irreduc.

Example: (Roberts & Rosenthal, 1997)

Let $\mathcal{X} = \{1, 2, 3\}$.

$$P_1 = \begin{array}{c c} & \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \end{array} & \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix} \end{array} \qquad P_2 = \begin{array}{c c} & \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \end{array} & \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \end{array}$$

Is P_1 aperiod? Irreduc? P_2? Stationary distn? $\pi = (\frac{1}{4}, \frac{1}{4}, \frac{1}{2})$.
But define $P = P1 \circ P_2$, and note $P(1, 1) = 1$, so reducible!

A note on hybrid chains

Note that *random*-scan combinations of ϕ-irreducible chains are *always* ϕ-irreducible. (Why?)

(Note for *Gibbs samplers*, the component chains act on *subsets* of components so are typically reducible. Must verify irreducibility of combined chain directly.)

(figure)

So should we always use random scan?

Convergence rates of Markov kernels

Recall that MC is geometrically ergodic if $\exists 0 < \lambda < 1$ and $V : \mathcal{X} \to \mathbb{R}^+$ s.t.

$$\|P^n(x, \cdot) - \pi(\cdot)\|_{\text{TV}} \leq V(x)\lambda^n \qquad \forall n \in \mathbb{N}, \forall x \in \mathcal{X}$$

If $V(x)$ is bounded, then chain is uniformly ergodic[1].

The smallest such λ^* for which such a V exists is called the *rate of convergence*.

We will come back later and relate this to eigenvalues of P.

[1] All finite-state MCs are uniformly ergodic.

Dependence and convergence

Consider a *product* target distribution:

$$\pi(X_1, \ldots, X_d) = \prod_{i=1}^{d} \pi_i(X_i)$$

Q: How fast does the Gibb sampler converge on such a target? A: It matters which one!

- Deterministic scan?

Dependence and convergence

Consider a *product* target distribution:

$$\pi(X_1, \ldots, X_d) = \prod_{i=1}^{d} \pi_i(X_i)$$

Q: How fast does the Gibb sampler converge on such a target? A: It matters which one!

- Deterministic scan: d steps suffice

Dependence and convergence

Consider a *product* target distribution:

$$\pi(X_1, \ldots, X_d) = \prod_{i=1}^{d} \pi_i(X_i)$$

Q: How fast does the Gibb sampler converge on such a target? A: It matters which one!

- Deterministic scan: d steps suffice
- Random scan?

Dependence and convergence

Consider a *product* target distribution:

$$\pi(X_1, \ldots, X_d) = \prod_{i=1}^{d} \pi_i(X_i)$$

Q: How fast does the Gibb sampler converge on such a target? A: It matters which one!

- Deterministic scan: d steps suffice
- Random scan? $O(d \log d)$[1]

So deterministic scan *can* be significantly faster.

$\Rightarrow$ Neither R.S. or D.S. is always preferrable.

[1] Coupon collectors problem

A results about Gibbs sampler convegence

Consider a two-state Gibbs sampler. Then we have

Theorem (Liu 1991)

For $d = 2$, the spectral radius is given by

$$\lambda^* = \sup_{f,g} \mathrm{corr}_\pi(f(x), g(y))$$

over all functions f, g.

(Note in irreducible example above, taking $f(x) = g(x) = \mathbf{1}(X \geq 0)$ yields $\rho = 1$. So fails to be geometrically ergodic, or indeed ergodic at all.)

Example: Multivariate normal

For multivariate normal distribution $X \sim N(\mu, \Sigma)$, supremum is always obtained by linear functions.

So for $d = 2$, we have

$$\lambda^* = \rho^2$$

where ρ is the correlation of the bivariate normal.

Example: Bayesian linear regression

Simple linear regression model:

$$y_i = \alpha + \beta x_i + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma^2)$$

Bayesian analysis with flat priors $p_0(\alpha, \beta) \propto 1$, σ^2 known.

Then we have

$$\mathrm{corr}_\pi(\alpha, \beta) = \rho^2_{\alpha,\beta} = -\frac{\bar{x}^2}{\bar{x}^2 + \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Note: if $|\bar{x}|$ large relative to sample s.d., $\rho_{\alpha,\beta}$ near ± 1.

Example: Bayesian linear regression

Solution: reparameterize.

Centering of covariate:

$$x_i' = x_i - \bar{x}$$

So model becomes

$$y_i = \alpha' + \beta' x_i' + \epsilon_i$$

where

$$\alpha' = \alpha + \beta\bar{x}$$
$$\beta' = \beta$$

Now $\rho_{\alpha',\beta'} = 0$, and Gibbs sampler yields *iid* !!

Example: Bayesian linear regression

More generally, let $y_i = \sum_{j=0}^{p} \beta_j x_{ij} + \epsilon_i$ so

$$y = X\beta + \epsilon$$

for design matrix $X = (x_1, x_2, \ldots, x_n)^T$ and $x_{i0} \equiv 1$.

Again consider flat prior $p_o(\beta) \propto 1$ and σ^2 known. Then

$$\mathrm{cov}_\pi(\beta) = \sigma^2 (X^T X)^{-1}$$

Reparameterization: To remove all correlations, columns of X must be *orthogonal*.
(Centering yields orthogonalization wrt 1st column.)
Can achieve by PCA.

Don't really need complete orthogonality, rather need to avoid near-collinearity of covariates.

Example: Hierarchical/random effects models

Consider

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij} \qquad i = 1, \ldots, m; j = 1, \ldots, n$$

with $\alpha_i \sim N(0, \sigma_\alpha^2)$ and $\epsilon_{ij} \sim N(0, \sigma_y^2)$

For simplicity, again take σ_α, σ_y known and flat priors.

Gelfand *et al* (1995) show

$$\rho_{\mu,\alpha_i} = -\Big(1 + \frac{m\sigma_y^2}{n\sigma_\alpha^2}\Big)^{-\frac{1}{2}} \qquad \rho_{\alpha_i,\alpha_j} = \Big(1 + \frac{m\sigma_y^2}{n\sigma_\alpha^2}\Big)^{-1} \quad i \neq j$$

Correlations (hence convergence rate) depend on relative sizes of variance components.

Example: Hierarchical/random effects models

$$\rho_{\mu,\alpha_i} = -\Big(1 + \frac{m\sigma_y^2}{n\sigma_\alpha^2}\Big)^{-\frac{1}{2}} \qquad \rho_{\alpha_i,\alpha_j} = \Big(1 + \frac{m\sigma_y^2}{n\sigma_\alpha^2}\Big)^{-1} \quad i \neq j$$

More specifically:

If # random effects large or σ_α small, faster mixing.

But if # observations *per random effect* large, or observation noise σ_y^2 small, mixing worse.

This is exactly when data are most informative!

Example: Hierarchical/random effects models

Reparameterization:

One approach: "hierarchical centering"[1]

$$y_{ij} = \eta_i + \epsilon_{ij} \qquad \eta_i \sim N(\mu, \sigma_\alpha^2)$$

So $\eta_i = \mu + \alpha_i$.

Then (Gelfand *et al*):

$$\rho_{\mu,\eta_i} = -\Big(1 + \frac{mn\sigma_\alpha^2}{\sigma_y^2}\Big)^{-\frac{1}{2}} \qquad \rho_{\eta_i,\eta_j} = \Big(1 + \frac{mn\sigma_\alpha^2}{\sigma_y^2}\Big)^{-1} \quad i \neq j$$

Now both large m and large n improve mixing.
(Updating μ moves all η_i's *simultaneously*.)

[1] note here we're centering parameters, not covariates, unlike before

Example: Hierarchical/random effects models

For nested models

$$\begin{aligned} y_{ijk} &= \mu + \alpha_i + \beta_{ij} + \epsilon_{ijk} \\ \beta_{ij} &\sim N(0, \sigma_\beta^2) \\ \alpha_I &\sim N(0, \sigma_\alpha^2) \end{aligned}$$

we can use a "hierarchically centered" parameterization[1]:

$$\begin{aligned} y_{ijk} &= \gamma_{ij} + \epsilon_{ijk} \\ \gamma_{ij} &\sim N(\eta_{ij}, \sigma_\beta^2) \\ \eta_i &\sim N(\mu, \sigma_\alpha^2) \end{aligned}$$

However, this is not the only way to reparameterize, and may not be the best.

[1] Taking $\gamma_{ij} = \mu + \alpha_i + \beta_{ij}$ and $\eta_i = \mu + \alpha_i$

Some conclusions

Parameterization is important!

Can go from non-geometrically ergodic ($\rho = 1$) to *iid* ($\rho = 0$).

Reparameterization is *model-specific* and can be painful/time-consuming.

Motivation for *adaptive* MCMC.

But first some other ways to improve Gibbs samplers.

Interweaving (Yu & Meng, 2011)

Consider a simple 2-level hierachical normal model:

$$\begin{aligned} y \mid (\theta, \mu) &\sim N(\mu, 1) \\ \mu \mid \theta &\sim N(\theta, \sigma^2) \end{aligned} \qquad (1)$$

with σ^2 known and $p_0(\theta) \propto 1$.

The corresponding Gibbs sampler becomes:

$$\begin{aligned} &(1) \quad \mu \mid (\theta, y) \sim N\left(\frac{\theta + \sigma^2 y}{1+\sigma^2}, \frac{\sigma^2}{1+\sigma^2}\right) \\ &(2) \quad \theta \mid (\mu, y) \sim N(\mu, \sigma^2) \end{aligned}$$

Note: rhs of (1) does not depend on θ.
Call this parameterization *sufficient* augmentation[1]

[1] since μ a sufficient statistic for θ

Interweaving (Yu & Meng, 2011)

If reparameterize using $\tilde{\mu} = \mu - \theta$, we have

$$\begin{aligned} y \mid (\theta, \tilde{\mu}) &\sim N(\tilde{\mu} + \theta, 1) \\ \tilde{\mu} \mid \theta &\sim N(0, \sigma^2) \end{aligned} \qquad (2)$$

This gives Gibbs sampler:

$$\begin{aligned} &(1') \quad \tilde{\mu} \mid (\theta, y) \sim N\left(\frac{\sigma^2(y-\theta)}{1+\sigma^2}, \frac{\sigma^2}{1+\sigma^2}\right) \\ &(2') \quad \theta \mid (\tilde{\mu}, y) \sim N(y - \tilde{\mu}, 1) \end{aligned}$$

Note: rhs of (2) does not depend on θ.
Call this parameterization *ancillary* augmentation[1]

[1] since μ an ancillary statistic for θ; for Bayesians, $\tilde{\mu}$ and θ indpt *a priori*

Interweaving (Yu & Meng, 2011)

But these two Gibbs samplers have *different* convergence rates:

$$\lambda_{\mathrm{SA}} = \frac{1}{1+\sigma^2} \qquad \lambda_{\mathrm{AA}} = \frac{\sigma^2}{1+\sigma^2}$$

Notice $\lambda_{\mathrm{SA}} + \lambda_{\mathrm{AA}} = 1$, so when one fast, other slow: When σ^2 small, SA slow but AA fast.
When σ^2 large, AA slow but SA fast.

Possible solution: alternate steps of both:

$$(1) \to (2) \to (1') \to (2') \to (1) \ldots$$

Yields convergence rate: $\lambda_{\mathrm{Alt}} = \lambda_{\mathrm{SA}} \cdot \lambda_{\mathrm{AA}}$

Interweaving (Yu & Meng, 2011)

Better strategy (Yu & Meng): Interweaving

Replace (2),(1') steps by a single step drawing $\tilde{\mu} \mid \mu$
(not conditioning on θ)

$$[\mu \mid \theta^{(t)}, y] \to [\tilde{\mu} \mid \mu, y] \to [\theta^{(t+1)} \mid \tilde{\mu}, y]$$

How to draw $\tilde{\mu} \mid \mu$?

Draw $[\theta \mid \mu, y]$ then $[\tilde{\mu} \mid \mu, \theta]$.

So we get

$$[\mu \mid \theta^{(t)}, y] \to [\theta \mid \mu] \to [\tilde{\mu} \mid \mu, \theta] \to [\theta^{(t+1)} \mid \tilde{\mu}, y]$$

Theory a bit complicated but in practice can show significant speedups.
E.g. can be geometric even when both AA and SA are not.