Stat314 /461 Term 4: Gibbs sampling

Patrick Graham

October, 2021

A reminder about our computation problem

Let $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$ We are interested in the posterior distribution

$$p(\boldsymbol{\theta}|\text{data}) = \frac{p(\text{data}|\boldsymbol{\theta} p(\boldsymbol{\theta})}{\int p(\text{data}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}}. \tag{1}$$

We should always be able to write down the numerator of (1). But the integral in the denominator may not be friendly (it is also K-dimensional). We also need to to be able integrate $p(\boldsymbol{\theta}|\text{data})$ to compute useful things, e.g. posterior mean, variance, marginal tail probabilities etc.

Some methods for sampling from the posterior in multi-parameter problems

$\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \ldots, \theta_K)$

- We have studied the Metropolis-Hastings algorithm for sampling from $p(\boldsymbol{\theta}|\mathrm{data})$ by jumping through the k dimensional space in a manner guided by the posterior - sample more points from areas of high posterior density.
- For small dimensional problems we may be able to apply rejection or importance sampling.

Sampling $p(\boldsymbol{\theta}|\text{data})$ by breaking $\boldsymbol{\theta}$ into components

- Recall that

$$p(\boldsymbol{\theta}|\text{data}) = p(\theta_1|\text{data})p(\theta_2|\theta_1, \text{data}) \times \ldots \times p(\theta_q|\theta_{K-1}, \ldots, \theta_1|\text{data}) \qquad (2)$$

  In complex problems some components of this decomposition may be difficult to obtain or simulate. Metropolis-Hastings or Rejection sampling may help with some components.

- Gibbs sampling is another method of posterior simulation that involves breaking $\boldsymbol{\theta}$ into components but instead of sampling from the sequence of conditionals in (2) Gibbs Sampling proceeds over a series of iterations and samples from the posterior of each component conditionally on the most-recently sampled value of all other components.

Gibbs sampler - background

- Gelfand et al (1990) two JASA articles, introduced the Gibbs sampler to a general statistical audience.
- Tanner and Wong (1987) (missing data problems)
- Geman and Geman (1984) - Image analysis,
- A special case of Metropolis-Hastings

The Gibbs sampler: General Statement

- The Gibbs sampler proceeds by
  1. assigning (K-1) component of $\boldsymbol{\theta} = \theta_1, \ldots, \theta_K$ an initial value;
  2. alternately sampling from the "full conditional posterior" distribution of each component given not only the data but all other components of $\boldsymbol{\theta}$
  3. repeating step (2) for some number of draws until the sampling process converges to the desired joint distribution ("burn-in")
  4. repeating step (2) a further M times until to obtain M simulations from the desired joint distribution.

If each of the "full-conditionals" is easy to sample from we have a readily implemented algorithm. Sometimes the full-conditionals correspond to conjugate models so sampling from the full conditionals is straightforward.

Simple Example: Posterior for mean and precision of a normal distribution

$$Y_i|\mu,\tau \overset{\text{indep}}{\sim} \text{Normal}(\mu,\tau), i = 1,\ldots,n$$

where τ is the precision (inverse of the variance). $\mathbf{Y} = Y,\ldots,Y_n$ We adopt a prior of the form

$$p(\mu,\tau) = p(\mu)p(\tau) \tag{3}$$

where

$$\begin{aligned}\mu &\sim \text{Normal}(m_{prior},\kappa_{prior})\\ \tau &\sim \text{Gamma}(a,b)\end{aligned} \tag{4}$$

A Gibbs sampler for this problem alternates between the following steps

(i) draw μ from $p(\mu|\tau,\mathbf{Y})$

(ii) draw τ from $p(\tau|\mu,\mathbf{Y})$

Gibbs for Normal cont'd; full conditionals

From Elena's notes we know that given our prior (normal for μ, Gamma for τ, $p(\mu, \tau) = p(\mu)p(\tau)$

$$[\mu|\tau, \mathbf{Y}] \sim \text{Normal}\left(\frac{\tau n\overline{Y} + \kappa_{prior} m_{prior}}{\tau n + \kappa_{prior}}, \tau n + \kappa_{prior}\right)$$

$$[\tau|\mu, \mathbf{Y}] \sim \text{Gamma}\left(a + n/2, b + 0.5\sum_{i=1}^{i=n}(Y_i - \mu)^2\right) \tag{5}$$

where $\overline{Y} = \frac{1}{n}\sum_{i=1}^{i=n} Y_i$.

Gibbs for Normal cont'd - The sampling algorithm

1. set $\tau = \tau^{(0)}$ (usually by drawing from an approximation)
2. For t in $1 : T$ {
   2.1 draw $\mu^{(t)}$ from
   $$\text{Normal}\left(\frac{\tau^{(t-1)} n\overline{Y} + \kappa_{prior} m_{prior}}{\tau^{(t-1)} n + \kappa_{prior}}, \tau^{(t-1)} n + \kappa_{prior}\right)$$
   2.2 draw $\tau^{(t)}$ from
   $$\text{Gamma}\left(a + n/2, b + 0.5 \sum_{i=1}^{i=n} (Y_i - \mu^{(t)})^2\right) \tag{6}$$
   }
3. discard first L draws (burn-in)

Example see `Gibbs_example1_normalmodel.Rmd`

Aside about the Normal distribution (1)

Term 3: Normal density parameterised with mean μ and precision (as in WinBugs)

$$p(Y = y|\mu, \tau) = \sqrt{\frac{\tau}{2\pi}} \exp\left(-\frac{\tau(y-\mu)^2}{2}\right) \tag{7}$$

Term 4: (Mostly) Normal density parameterised with mean and standard deviation (as in R and Gelman et al BDA)

$$p(Y = y|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \tag{8}$$

Conditionally conjugate prior for this parameterisation is:

$$p(\mu, \sigma) = p(\mu)p(\sigma)$$

$$\mu \sim \mathrm{Normal}(m_0, s_0^2) \tag{9}$$

$$\sigma^2 \sim \mathrm{Inv}\chi^2(c, d) \tag{10}$$

No time (or need) to derive the corresponding "full-conditionals" here – see Gelman et al BDA)

Asides about the normal distribution (2)

- It is probably mildly annoying that the two parameterisations of the Normal are used in the course.
- . But do not worry about this. Just think of the normal distribution parameterised by mean and precision as one model; the normal distribution parameterised by mean and standard deviation or variance as another model.
- We try to make clear which parameterisation we are using.

Asides about the normal distribution (3)

- Thinking in terms of the normal distribution parameterised by mean and precision: No prior of the form $p(\mu, \tau) = p(\mu)p(\tau)$, can be conjugate for the normal model. - Consider the form of the normal density.
- Similarly, if working with the normal distribution parameterised by mean and standard deviation, σ, no prior of the form $p(\mu, \sigma) = p(\mu)p(\sigma)$ can be conjugate
- The priors $p(\mu, \tau) = p(\mu)p(\tau)$, $p(\mu, \sigma) = p(\mu)p(\sigma)$, where $p(\mu)$ is Normal, $p(\tau)$ is Gamma and $p(\sigma^2)$ is inverse-Gamma are only *conditionally* conjugate.
- The Gibbs sampler takes advantage of *conditional* conjugacy.
- However in more complex problems conditional conjugacy is not guaranteed and more advanced methods are required to simulate the full conditionals (e.g. Metropolis-Hastings)

More general statement of the Gibbs sampler

1. initialise $\theta_1, \theta_2, \ldots, \theta_K$ to $\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_K^{(0)}$
2. for (t in 1 : T) {
    - draw $\theta_1^{(t)}$ from $p(\theta_1|\theta_2^{(t-1)}, \ldots, \theta_K^{(t-1)}, \mathrm{data})$
    - draw $\theta_2^{(t)}$ from $p(\theta_2|\theta_1^{(t)}, \theta_3^{(t-1)} \ldots, \theta_K^{(t-1)}, \mathrm{data})$

    $$\vdots$$

    - draw $\theta_K^{(t)}$ from $p(\theta_1|\theta_1^{(t)}, \theta_2^{t)} \ldots, \theta_{K-1}^{(t)}, \mathrm{data})$, }
3. discard the first L iterations

General comments on the Gibbs sampler

- It is a MCMC procedure, in fact it can be shown to be a special-case of Metropolis-Hastings.
- Usual MCMC good practices therefore apply
    1. Discard burn-in sample
    2. run multiple chains from over-dispersed starting points
    3. Use Gelman-Rubin Rhat to check for convergence; Increase burn-in period if Rhat too big (e.g $Rhat > 1.1$ for important analyses).

Gibbs sampler for problems of the “missing data type”

The Gibbs sampler deals easily with problems of the “missing data” type e.g

- non-response
- mis-measured variables
- latent variables - a relevant variable is not directly observable

The general idea is to alternate between

i simulating the “missing” data from their conditional posterior (predictive) distribution given the current value of the model parameters

ii drawing from the conditional posterior of the parameters given the observed data and most recent imputations of the missing data.

This is the idea behind the “data-augmentation” approach developed by Tanner and Wong (1987).

Gibbs for missing data: Theory (1)

- Suppose $\boldsymbol{\theta}$ is the model parameter of interest, $\mathbf{D}^{obs}$ the data actually observed and $\mathbf{D}^{mis}$ the missing data. We define the full data, that we would have liked to observed, by $\mathrm{data} = (\mathbf{D}^{obs}, \mathbf{D}^{mis})$
- Assume we know how to compute $p(\boldsymbol{\theta}|\mathrm{data})$, the posterior given the full data- a standard Bayesian inference problem.
- Since we only observe $\mathbf{D}^{obs}$ the posterior we need to compute is $p(\boldsymbol{\theta}|\mathbf{D}^{obs})$ - we can only condition on the data actually observed.
- However

$$p(\boldsymbol{\theta}|\mathbf{D}^{obs}) = \int p(\boldsymbol{\theta}, \mathbf{D}^{mis}|\mathbf{D}^{obs}) \, \mathrm{d}\mathbf{D}^{mis} \tag{11}$$

Gibbs for missing data: Theory (2)

- want: $p(\boldsymbol{\theta}|\mathbf{D}^{obs}) = \int p(\boldsymbol{\theta}, \mathbf{D}^{mis}|\mathbf{D}^{obs})\,\mathrm{d}\mathbf{D}^{mis}$
- We can use the Gibbs Sampling algorithm to, effectively, do the integration for us by simulating $p(\boldsymbol{\theta}, \mathbf{D}^{mis}|\mathbf{D}^{obs})$ by sampling alternately from:
    - i $p(\boldsymbol{\theta}|\mathbf{D}^{mis}, \mathbf{D}^{obs}) = p(\boldsymbol{\theta}|\mathrm{data})$ (standard)
    - ii $p(\mathbf{D}^{mis}|\boldsymbol{\theta}, \mathbf{D}^{obs})$
- For inference we can ignore the generated $\mathbf{D}^{mis}$ values and treat the generated θ values as a sample from $p(\boldsymbol{\theta}|\mathbf{D}^{obs})$

Gibbs sampler for missing data problems: Zero-inflated models

see .Rmd file `Gibbs_example2_ZIPmodel.Rmd`

Gibbs sampler: Example 3 random rounding

For example suppose $\mathbf{Y} = (Y1, \dots, Y_n)$ is a vector of counts and we adopt the model

$$
\begin{aligned}
Y_i &\overset{\text{indep}}{\sim} \text{Poisson}(\theta) \\
\theta &\sim \text{Gamma}(\alpha, \beta)
\end{aligned} \tag{12}
$$

for fixed α, β. Instead of observing the counts we see only a randomly rounded version of the counts, $\mathbf{R} = R_1, R_2, \dots, R_n$ Our inferential task is then to compute $p(\theta|\mathbf{R})$ since $\mathbf{R}$ is the observed data. Since

$$
p(\theta|\mathbf{R}) = \int p(\theta, \mathbf{Y}|\mathbf{R}) \mathrm{d}\mathbf{Y} \tag{13}
$$

if we can compute the joint posterior $p(\theta, \mathbf{Y}|\mathbf{R})$ we are done. Given a sample from the *joint* posterior we can just focus on the generated θ values for inference.

Gibbs sampler for inference under random rounding

1 initialise θ to $\theta^{(0)}$

2 for t in 1 to K

   i draw $\mathbf{Y}^{(t)}$ from

$$
\begin{aligned}
p(\mathbf{Y}|\mathbf{R},\theta^{(t-1)}) &\propto p(\mathbf{R}|\mathbf{Y})p(\mathbf{Y}|\theta = \theta^{(t-1)}) && (14)\\
&= \prod_i p(R_i|Y_i) \prod_i poisson(Y_i|\theta = \theta^{(t-1)}) && (15)\\
&= \prod_i p(R_i|Y_i) poisson(Y_i|\theta = \theta^{(t-1)}) && (16)
\end{aligned}
$$

   ii

$$\text{draw } \theta^t \sim \text{Gamma}(\alpha + \sum_i Y_i^{(t)}, \beta + n) \quad (17)$$

(16) can be simulated easily by direct simulation or (rejection sampling); (17) follows from the conjugate Poisson-Gamma model.

Random Rounding

For details see: `Gibbs_example3_RR3.Rmd`

The Gibbs sampler is a special case of the Metropolis-Hastings algorithm

- recognise the full conditional posterior distributions as a special jumping distribution in which the only jumps allowed are to values of $\boldsymbol{\theta}$ which match the current value of $\boldsymbol{\theta}$ wrt to elements except the one currently being updated.

$$J_{j,t}^{Gibbs}(\boldsymbol{\theta}^{new}|\boldsymbol{\theta}^{(t-1)}) = \begin{cases} p(\theta_j^{new}|\boldsymbol{\theta}_{-j}^{(t-1)*}, \mathrm{data}) \text{ if } \boldsymbol{\theta}_{-j}^{new} = \boldsymbol{\theta}_{-j}^{(t-1)*} \\ 0 \text{ otherwise} \end{cases} \tag{18}$$

- plug $J_{j,t}^{Gibbs}$ into the M-H acceptance ratio formula

$$r_{MH,j}(\boldsymbol{\theta}^{new}, \boldsymbol{\theta}^{(t-1)*}) = \frac{p(\boldsymbol{\theta}^{new}|\mathrm{data})/J_{j,t}^{Gibbs}(\boldsymbol{\theta}^{new}|\boldsymbol{\theta}^{(t-1)*})}{p(\boldsymbol{\theta}^{(t-1)*}|\mathrm{data})/J_{j,t}^{Gibbs}(\boldsymbol{\theta}^{(t-1)*}|\boldsymbol{\theta}^{new})}$$

see notes for explanation of $(t-1)*$ superscript (all elements except θ_j set to their most recently updated value.

Acceptance probabilities for Gibbs sampler viewed as Metropolis-Hastings

The acceptance ratio at the j^{th} step of the t^{th} iteration is therefore

$$r_{MH,j}(\boldsymbol{\theta}^{new}, \boldsymbol{\theta}^{(t-1)*}) = \frac{p(\boldsymbol{\theta}^{new}|\mathrm{data})/J_{j,t}(\boldsymbol{\theta}^{new}|\boldsymbol{\theta}^{(t-1)*})}{p(\boldsymbol{\theta}^{(t-1)*}|\mathrm{data})/J_{j,t}(\boldsymbol{\theta}^{(t-1)*}|\boldsymbol{\theta}^{new})} \tag{19}$$

$$= \frac{p(\boldsymbol{\theta}^{new}|\mathrm{data})/p(\theta_j^{new}|\boldsymbol{\theta}_{-j}^{(t-1)*}, \mathrm{data})}{p(\boldsymbol{\theta}^{(t-1)*}|\mathrm{data})/p(\theta_j^{(t-1)*}|\boldsymbol{\theta}_{-j}^{(t-1)*}, \mathrm{data})} \tag{20}$$

$$= \frac{p(\boldsymbol{\theta}_{-j}^{(t-1)*}|\mathrm{data})}{p(\boldsymbol{\theta}_{-j}^{(t-1)*}|\mathrm{data})} \tag{21}$$

$$= 1. \tag{22}$$

see notes for explanation of steps.

Gibbs sampler when not all full conditionals can be directly simulated

- For difficult full conditionals we can use a Metropolis-Hastings step
- Suppose the difficult full conditional is for component j, and that this is the last component to be updated on each Gibbs iteration
  - (i) draw a proposal $\theta_j^{(t)}$ from $J_{j,t}(\theta_j|\theta_j^{(t-1)*})$
  - (ii) evaluate

$$r_{MH,j}(\theta^{(t)},\theta^{(t-1)}) = \frac{q(\theta_j^{(t)}|\mathrm{data},\boldsymbol{\theta}_{-j}^{(t-1)*})/J_{j,t}(\theta_j^{(t)}|\theta_j^{(t-1)*})}{q(\theta_j^{(t-1)*}|\mathrm{data},\boldsymbol{\theta}_{-j}^{(t)})/J_{j,t}(\theta_j^{(t-1)*}|\theta_j^{(t)})}$$

  - (iii) accept $\theta_j^{(t)}$ with probability $\min(1, r_{MH,j}^{(t)})$.
    If $\theta_j^{(t)}$ is not accepted set $\theta_j^{(t)} = \theta_j^{(t-1)*}$, i.e stay at $\theta_j^{(t-1)*}$.

Hybrid Gibbs/Metropolis-Hastings samplers arise frequently in practice

- Reconsider our fishing example: Suppose instead of a just focussing on the expected number of fish caught (given that a party fished) we were interested in a Poisson regression relating catch to covariates.

$$
\begin{aligned}
[Y_i|Z_i = 1, \mathbf{X}_i, \boldsymbol{\beta}] &\overset{\text{indep}}{\sim} \text{Poisson}(\theta_i), i = 1, \ldots, n \\
\log(\theta_i) &= \beta_0 + X_{1i}\beta_1 + X_{2i}\beta_2 + \ldots, i = 1, \ldots, n
\end{aligned}
$$

- There is no obvious conditionally conjugate prior for $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots)$.
- A hybrid Gibbs / Metropolis-Hastings sampler would alternate between
  - i simulating $\mathbf{Z}$ from $p(\mathbf{Z}|\mathbf{Y}, \mathbf{X}, \phi, \boldsymbol{\beta})$
  - ii updating $\boldsymbol{\beta}$ using a Metropolis-Hastings step
  - iii simulating ϕ from $p(\phi|\mathbf{Z})$

Comment on Gibbs sampler with Metropolis-Hastings steps

- Used to be referred to as "Metropolis-Hastings within Gibbs"
- Given that Gibbs itself is a special case of Metropolis-Hastings, it is possibly easier to think of hybrid Gibbs / Metropolis-Hastings procedures simply as versions of Metropolis-Hastings.
- In problems with many parameters it is difficult to apply Metropolis-Hastings to the full parameter vector; some chunking of parameters into sub-groups seems inevitable.
- Hybrid Gibbs / Metropolis-Hastings algorithms provide a practical way to apply MCMC in problems without conditional conjugacy and/or many parameters.