Sampling using Auxiliary variables †

Soma S Dhavala

† Based on Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables, Stephen Walker et al JRSS(B), 61(2), pp.331-344. 1999

Outline

Introduction

- MCMC requires/assumes efficient sampling algorithms
- Gibbs sampling is common but full conditionals may not always be available in standard forms
- Metropolis and its variants require proposal densities/ tuning
- Black-box techniques like Rejection sampling/ Adaptive Rejection sampling may be inefficient (too generic)
- Slice sampling
  - problem-specific approach, no tuning required
  - can be drawn from standard densities (uniform, truncated standard distributions)

Slice Sampling: Basic Idea

- Suppose that X is a r.v with density $f(x)$ and you want to draw x from this distribution
- Introduce another r.v U and construct a joint density $f(x, u)$ such that its marginal density is $f(x)$
- Now sample from the conditionals $f(x|u)$ and $f(u/x)$
- The new artificial r.v U should be elicited in such a way that $f(x|u)$ and $f(u/x)$ are easy to sample from

Slice sampling: Bayesian Setting

Suppose that we wish to generate r.v from

$$f(x) \propto \pi(x) \prod_{i=1}^{N} l_i(x)$$

If

1. π is a density in known form
2. l_i are non-negative invertible functions (not necessarily densities)
3. $l_i(x) > u$

Then

1. A Gibbs sampler exists where conditional distributions are either uniform or truncated version of $\pi(x)$
2. It is possible to obtain the set $A_u^i = \{x : l_i(x) > u\}$ that determines the truncation boundaries

Slice sampling: Example

Suppose that we wish to sample a r.v from $f(x)$ with

$$l(x) = \exp[\exp(-x)]$$

Introduce a latent variable U s.t $f(u/x) = u(0, l(x))$

- Now $f(x, u) \propto I(0 < u < l(x))\pi(x)$ (obvious case)
- $f(u/x) \propto I(0 < \exp[\exp(-x)])$ (uniform)
- $f(x/u) \propto \pi(x)I(x > l^{-1}(u))$ (truncated π)
- Note that the marginal distribution is $\int f(x, u)du = l(x)\pi(x)$

Slice sampling: Example (contd.)

Choosing a latent variable is not unique. For the same example
Alternatively, define V s.t

- $f(x, v) \propto \exp[-v]\, I(v > \exp[x])\, \pi(x)$
- $f(v/x) \propto \exp[-v]\, I(v > \exp[x])\, \pi$
- $f(x/v) \propto \pi(x)\, I(x < \log[v])\, \pi$

Gibbs sampling can be implemented as:

- $e \sim \exp(1)$
- $v = exp(\tilde{x}) + e$, ($\tilde{x}$ is the previous draw)
- $x \sim \pi(x)\, I(x < \log[v])$

Note that the conditional distribution of the latent variable has exponential distribution in this case

Applications

- Choosing latent variables is usually self-evident
- If not, one can always introduce them via $I(0, l(x))$
- Can be tailored to the problem at hand (like the example)
- Can write a general algorithm for distributions in exponential family
- We mention some applications in non-conjugate models, hierarchical models

Applications: Non-conjugate models

Consider Poisson likelihood and log-Normal prior for the Poisson means($\exp[x]$). Specifically:

$$\begin{aligned} y &\sim \text{Poisson}(\exp[x]) \\ x &\sim \text{N}(0,1) \\ f(x) &\propto \exp[yx - \exp(x)]\exp[-0.5x^2] \\ &= \exp[-\exp(x)]\exp[-0.5(x^2 - 2yx)] \end{aligned}$$

and recall previous example and introduce latent variable U s.t

- $f(x,u) \propto \exp[-u]I(u > \exp[x])\exp[-0.5(x^2 - 2yx)]$ So that
- $f(u/x) \propto \exp[-u]I(u > \exp[x])$ and (shifted exponential)
- $f(x/u) \propto \exp[-0.5(x^2 - 2yx)]I(x < \log[u])$ (truncated normal)

Applications: Bayesian Hierarchical Model

Consider a radom effects Poisson model:

$$
\begin{aligned}
y_i|\theta_i &\sim \text{Poisson}(\exp[\theta_i]) \\
\theta_i &= w_i\beta + b_i \\
b_i &\sim N(0, \lambda^{-1}) \\
\beta &\sim N(\mu, \Sigma) \\
\lambda^{-1} &\sim \text{Gamma}(a, b)
\end{aligned}
$$

The posterior is:

$$
f(\beta, b, \lambda) \propto \lambda^{N/2}\pi(\lambda, \beta)\prod_{i=1}^{N}\exp\left[y_i\theta_i - \exp[\theta_i] - 0.5b_i^2\lambda\right]
$$

Applications: Bayesian Hierarchical Model (contd.)

Rewrite the posterior as:

$$f(\beta, b, \lambda) \quad \propto \quad \lambda^{N/2}\pi(\lambda, \beta)\prod_{i=1}^{N} \exp[y_i\theta_i]\exp[-\exp(\theta_i)]\exp\left[-0.5b_i^2\lambda\right]$$

Introduce $U = (U_1, ...U_N), V = (V_1, ..., V_N)$ latent variables such that

$$\begin{aligned} f(\beta, b, \lambda, u, v) \quad \propto \quad & \lambda^{N/2}\pi(\lambda, \beta) \times \\ & \prod_{i=1}^{N} e^{-v_i} I(v_i > e^{\theta_i}) I(u_i < e^{y_i\theta_i}) e^{-0.5b_i^2\lambda} \end{aligned}$$

Applications: Bayesian Hierarchical Model (contd.)

The conditionals are:

- $f(u_i|.) \propto I(u_i < e^{y_i \theta_i})$
- $f(v_i|.) \propto e^{-v_i} I(v_i > e^{\theta_i})$
- $f(b_i|) \propto \exp[-0.5 b_i^2 \lambda] I(b_i \in A_i{}^{\dagger})$
- $f(\beta_k|) \propto \pi(\beta_k|) I(\beta_i \in B_k)$
- $f(\lambda|) \propto \pi \lambda^{N/2} e^{-0.5\lambda \sum_i b_i^2} \pi(\lambda)$

They all are either uniform distributions or truncated versions of priors or in standard form

† Refer the paper for exact details

Discussion

Advantages:

- can always introduce latent variables
- generic algorithm given for Generalized Linear models, non-linear models
- can be better than independent chain M-H
- Gibbs sampler always uses (truncated) standard distributions

Disadvantages:

- Number of latent variables grows with data-points
- Induces correlation among samples generated
- Computing truncation sets could be cumbersome
- Many other flavors of slice sampling exist which could be much better [†]

[†] Radford Neal, Slice Sampling, Annals of Statistics, 31(3):705-767, 2003 for a discussion