Advanced Statistical Computing

Fall 2012
Lecture 3

Steve Qin

Outline

- Slice sampler
- Reversible jump
- Parallel tempering
- Collapsing, predictive updating
- Sequential Monte Carlo
- Convergence checking

Fundamental theorem of simulation

- Simulating $X \sim f(x)$ is equivalent to simulating

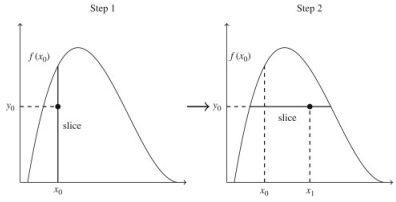
  $(X,U) \sim Uniform\{(x,u): 0 < u < f(x)\}$.

  f is the marginal density of the joint distribution.

3

Slice sampler

2D slice sampler



- At iteration t, simulate
  1. $u^{(t+1)} \sim Uniform(0, f(x^{(t)}))$
  2. $x^{(t+1)} \sim Uniform(A^{(t+1)})$, with $A^{(t+1)} = \{x: f(x) \geq u^{(t+1)}\}$.
- Neal (1997), Damien et al. (1999)

4

Examples

- Simple slice sampler.
  density function $f(x) = \frac{1}{2} e^{-\sqrt{x}}$ for $x > 0$.
  - $U \mid x \sim Uniform(0, \frac{1}{2} e^{-\sqrt{x}})$,
  - $X \mid u \sim Uniform(0, [\log(2u)]^2)$,
- Truncated normal distribution.

$$f(x) \propto e^{-x^2/2} I_{(a,b)}(x).$$

  - $U \mid x \sim Uniform(0, e^{-x^2/2})$,
  - $X \mid u \sim Uniform\left(\max\left\{a, -\sqrt{-2\log y}\right\}, \min\left\{b, \sqrt{-2\log_5 y}\right\}\right)$.

The general slice sampler

Suppose $f(x) = \prod_{i=1}^{k} f_i(x)$,

Introduce auxiliary variables ω_i, such that

$$f_i(x) = \int I_{[0, f_i(x)]}(\omega_i) d\omega_i$$

f is the marginal distribution of the joint dist

$$(x, \omega_1, \omega_2, ..., \omega_k) \sim p(x, \omega_1, \omega_2, ..., \omega_k) \propto \prod_{i=1}^{k} I_{[0, f_i(x)]}(\omega_i),$$

6

Slice sampler algorithm

At iteration t+1, simulate

1. $\omega_1^{(t+1)} \sim Uniform(0, f_1(x^{(t)}))$,
2. $\omega_2^{(t+1)} \sim Uniform(0, f_2(x^{(t)}))$,

…

k. $\omega_k^{(t+1)} \sim Uniform(0, f_k(x^{(t)}))$,

k+1. $x^{(t+1)} \sim Uniform(A^{(t+1)})$, where

$$A^{(t+1)} = \left\{ y : f_i(y) \geq \omega_k^{(t+1)}, \quad i = 1, \ldots, k \right\}$$

7

Examples

- Truncated exponential $E(\beta, a, b)$

  $f(x) \propto e^{-\beta x} I(a < x < b), \quad \beta > 0, a, b > 0.$

  Sample $U \sim Uniform(1 - e^{-a\beta}, 1 - e^{-b\beta})$, and set

  $$X = -\frac{1}{\beta} \log(1 - U).$$

- $B(\beta, a, b)$

  $f(x) \propto (1-x)^{\beta - 1} I(a < x < b), \quad \beta > 0, 0 \leq a < b \leq 1.$

  Sample Y from the above distribution, and set $X = 1 - e^{-Y}$.

8

Examples

- Standard normal $f(x) \propto e^{-x^2/2}$.
  introduce latent variable $Y > 0$,

$$f(x,y) \propto e^{-y/2} I(y > x^2).$$
$$Y \mid X = x \sim E(0.5, x^2, \infty),$$
$$X \mid Y = y \sim Uniform(-\sqrt{y}, +\sqrt{y}).$$

- *Gamma* $f(x) \propto x^{\alpha-1} e^{-x} I(x > 0), \quad \alpha > 0.$
  introduce latent variable $Y > 0$,

$$f(x,y) \propto e^{-x} I(y < x^{\alpha-1}, x > 0).$$
$$Y \mid X = x \sim Uniform(0, x^{\alpha-1}),$$
$$X \mid Y = y \sim E(1, y^{1/(\alpha-1)}, \infty), \quad \text{if } \alpha > 1.$$

9

Scale mixture of uniforms

- Normal
  If $X \mid V = v \sim Uniform(\mu - \sigma\sqrt{v}, \mu + \sigma\sqrt{v}),$
  $V \sim Gamma(3/2, 1/2),$
  then $X \sim N(\mu, \sigma^2).$
- student-*t*

$$X \mid V = v \sim Uniform(\mu - \sigma\sqrt{v}, \mu + \sigma\sqrt{v}),$$
$$V \mid \xi \sim Gamma(3/2, \xi/2),$$
$$\xi \sim Gamma(\alpha/2, \alpha/2).$$

10

Related algorithms

- Auxiliary variable algorithm
  - Edwards and Sokal (1988)
- Swendson and Wang
  - Swendson and Wang (1987)

11

Reversible jump

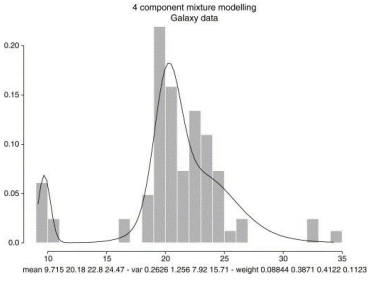
- Motivation: variable dimension models

  a "model where one of the things you do not know is the number of things you do not know." –Peter Green.
- Bayesian model comparison and model selection.

12

An example

- Mixture modeling
  Number of components
- Order of an AR (p) model



13

Green's algorithm

- At iteration t, if $x^{(t)} = (m,\theta_m^{(t)})$,
  - Select model M_n with probability π_{mn}
  - Generate $u_{mn} \sim \varphi_{mn}(u)$
  - Set $(\theta_n, v_{nm}) \sim T_{mn}(\theta_m^{(t)}, u_{mn})$
  - Take $\theta_n^{(t)} = \theta_n$ with probability

$$\min\left(\frac{\pi(n,\theta_n)}{\pi(m,\theta_m^{(t)})}\frac{\pi_{nm}\varphi_{nm}(v_{nm})}{\pi_{mn}\varphi_{mn}(u_{nm})}\left|\frac{\partial T_{mn}(\theta_m^{(t)},u_{mn})}{\partial(\theta_n^{(t)},u_{mn})}\right|,1\right)$$

14

Remarks

- The clever idea of RJMCMC is to supplement "smaller" space with artificial space.
- Dimension matching transform T_{mn} is flexible, but quite difficult to create and optimize, a serious drawback of the method.
- Methodologically brilliant but difficult to implement.

15

Convergence check

- Trace plot
- Autocorrelation plot
- Gelman and Rubin convergence measure r.

16

Convergence Diagnostics

Patrick Lam

Within Chain Variance

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

where

$$s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\theta_{ij} - \bar{\theta}_j)^2$$

s_j^2 is just the formula for the variance of the jth chain. W is then just the mean of the variances of each chain.

W likely underestimates the true variance of the stationary distribution since our chains have probably not reached all the points of the stationary distribution.

Between Chain Variance

$$B = \frac{n}{m-1}\sum_{j=1}^{m}(\bar{\theta}_j - \bar{\bar{\theta}})^2$$

where

$$\bar{\bar{\theta}} = \frac{1}{m}\sum_{j=1}^{m}\bar{\theta}_j$$

This is the variance of the chain means multiplied by n because each chain is based on n draws.

Estimated Variance

We can then estimate the variance of the stationary distribution as a weighted average of W and B.

$$\hat{\mathrm{Var}}(\theta) = (1 - \frac{1}{n})W + \frac{1}{n}B$$

Because of overdispersion of the starting values, this overestimates the true variance, but is unbiased if the starting distribution equals the stationary distribution (if starting values were not overdispersed).

Gelman and Rubin Multiple Sequence Diagnostic

Steps (for each parameter):

1. Run $m \geq 2$ chains of length $2n$ from overdispersed starting values.
2. Discard the first n draws in each chain.
3. Calculate the within-chain and between-chain variance.
4. Calculate the estimated variance of the parameter as a weighted sum of the within-chain and between-chain variance.
5. Calculate the potential scale reduction factor.

Potential Scale Reduction Factor

The potential scale reduction factor is

$$\hat{R} = \sqrt{\frac{\hat{\text{Var}}(\theta)}{W}}$$

When $\hat{R}$ is high (perhaps greater than 1.1 or 1.2), then we should run our chains out longer to improve convergence to the stationary distribution.

If we have more than one parameter, then we need to calculate the potential scale reduction factor for each parameter.

We should run our chains out long enough so that all the potential scale reduction factors are small enough.

We can then combine the mn total draws from our chains to produce one chain from the stationary distribution.