Minicourse on:

Markov Chain Monte Carlo: Simulation Techniques in Statistics

Eric Slud, Statistics Program

Lecture 2: The 'Gibbs Sampler', via motivation from Metropolis-Hastings. Statistical applications in hierarchical-model inference, with computational examples.

Outline

(I) Begin with re-cap of Gibbs-Sampler motivation from 1st Lecture and ideas of checking for convergence in Example of generating uniform random 10-vector within unit simplex further restricted by another linear constraint. Compare behavior of Gibbs-sampler version.

(II) General definition of Gibbs-Sampler. Relation to Metropolis-Hastings. First examples.

(III) Relation of Gibbs-Sampler to Bayesian statistical analysis. Example of random-intercept logistic regression inference.

Geometric-Prob. Example. Define unit simplex

$$B = \{(x_1, \ldots, x_d) : \; x_i \geq 0, \; \sum_{i=1}^{d} x_i \leq 1\}$$

and for fixed $\mathbf{a} \in (\mathbf{R}^+)^d$, $b > 0$, objective was to simulate uniform random point in

$$D \;=\; \{\mathbf{x} \in B : \;\; \mathbf{x} \cdot \mathbf{a} \leq b\}$$

Fixed $d = 10$, and (random, but fixed) choice $\mathbf{a} =$

```
0.513 0.944 0.960 0.116 0.032 0.944
0.691 0.489 0.020 0.710
```

and $b = \mathbf{a} \cdot \mathbf{1}/20 = .271$.

Metropolis-Hastings Algorithm

We defined *Proposal Markov Chain* which, starting from point $\mathbf{x} \in \mathbf{R}^d$ had transition step with conditional density $q(\mathbf{x}, \cdot)$ consisting of multiplication of the coordinates x_i by independent r.v.'s e^{Z_i} with $Z_i \sim \mathcal{N}(\mu(\mathbf{x}, \alpha)$. M-H Algorithm using this chain takes the form: if $\mathbf{X}_1, \ldots, \mathbf{X}_k$ have already been generated, $Y_k \sim q(X_k, \cdot)$ and $U_{k+1} \sim \text{Unif}[0,1]$ are simulated and then:

$$X_{k+1} \;=\; Y_k \qquad \text{if} \quad U_{k+1} \;\leq\; \frac{\pi(Y_k)\, q(Y_k, X_k)}{\pi(X_k)\, q(X_k, Y_k)}$$

and $= X_k$ otherwise.

Another Proposal Chain

Transition affecting only i'th coordinate of $\mathbf{x} \in D$ is to replace x_i by conditional distribution for i'th coordinate of random D point given coord's $x_1, \ldots, x_{i-1}$, $x_{i+1}, \ldots, x_d$, or

$$\mathrm{Uniform}(0,\ \min(1 - \sum_{j:j \neq i} x_j,\ (b - \sum_{j:j \neq i} a_j x_j)/a_j))$$

New 'proposal-chain' step is to do these replacements for all $i \in \{1, \ldots, 10\}$. (In practice, we do them in random order!) This is the **Gibbs Sampler** for the present example.

Plotted picture shows that the blocks of successive smoothed-histograms for quantities $x_1 + \cdots + x_6$ by this method behave very stably!

Here is another indicator of convergence: tallied numbers of $x_1 + \cdots + x_6$ values in blocks of 1000 which fall in bins defined by breakpoints (0,.2,.3,.4,.45,.5,.55,.6,.7,.8,1); then tallied same for another block of 1000 occurring 10000 iterates later.

Interval	1	2	3	4	5	6	7	8	9	10
Count1	35	132	208	105	94	96	99	124	76	31
Count2	34	104	189	112	103	113	101	121	90	33

Two-sample χ^2_9 value is 7.566, which is OK!

Simple Gibbs-Sampler Example

Consider the problem of sampling bivariate r.v.'s from the joint density on the positive quadrant:

$$f(x, y) \ = \ c \, \exp(-x - y \ - \ 4xy)$$

Exact joint dist. fcn is messy, but *conditionals* are not:

$$f_{X|Y}(x|y) = (1 + 4y) \, e^{-x(1+4y)} \sim \text{Expon}(1 + 4y)$$

(by symmetry, conditional for Y given X has same form). Simulating exponentials is easy:

$$U \sim \text{Unif}(0, 1) \quad \Rightarrow \quad \frac{-\log U}{\lambda} \sim \text{Expon}(\lambda)$$

So begin with (X_0, Y_0) arbitrary (say independent Expon(1) coord's). Next

$$X_{t+1} \sim \text{Expon}(1 + 4Y_t) \ , \quad Y_{t+1} \sim \text{Expon}(1 + 4X_{t+1})$$

Generated 10,000 successive pairs (X_t, Y_t) this way:

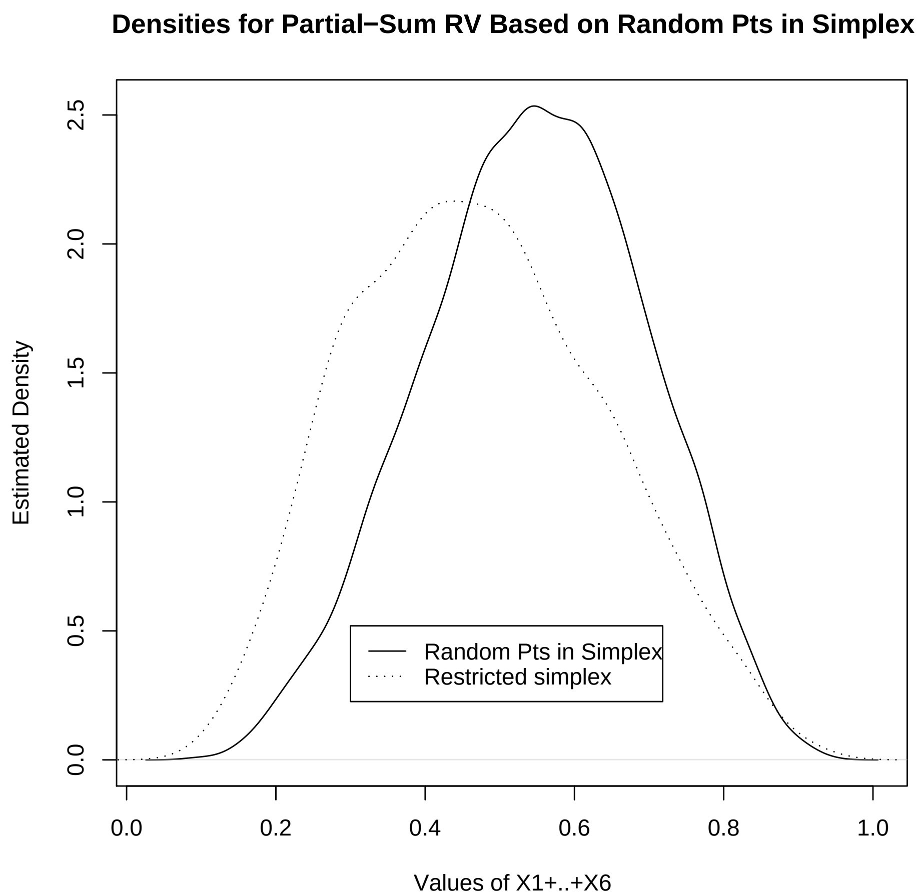

Figure 7: Plot of true density (hollow points) and 5 smoothed-histogram (density-estimate) pictures based on 5 successive blocks of 1000 x-values in bivariate exponential Gibbs-sampler example. Five thousand pre-iterates ($M = 5000$) preceded the first block.

General Gibbs-Sampler Step

So what characterizes the *Gibbs Sampler* as an MCMC technique is (primarily) that sampling transition-steps are done from the **full conditionals** and (usually) that the M-H acceptance probabilities are always 1.

Full conditionals means simulation of a random vector $\mathbf{X} = (X_1, \ldots, X_K)$ in a setting where all

$$f_{X_i|(X_j,\, j\neq i)}(x_i \,|\, \mathbf{x}^{(i)})) \quad , \qquad i = 1, \ldots, K$$

are simple to simulate from.

A single complete transition-step consists of a complete pass $\mathbf{X} \mapsto \mathbf{X}'$ through all components, say

$$X_i' \sim f_{X_i|(X_j,\, j\neq i)}(\cdot \,|\, (X_j',\, j < i;\; X_j,\, j > i)) \quad , i = 1, \ldots, K$$

If the actual conditional densities for the desired joint density are used, *and the order of stepping through coordinates is randomized*, then this is a Metropolis-Hastings step with all acceptance-probabilities = 1. This was the case in the previous examples with random point from simplex, and with bivariate exponential ($K = 2$).

Resulting chain is $f_{\mathbf{X}}$ **irreducible** under the *Positivity condition* saying:

$$f_{X_i}(x_i) > 0 \quad \text{for} \quad i = 1, \ldots, K \qquad \Longrightarrow \qquad f_{\mathbf{X}}(\mathbf{x}) > 0$$

General Gibbs-Sampler, continued

Note: the positivity condition is satisfied in both of the previous examples.

Hammersley-Clifford Thm, 1970. Under the positivity condition, $f_{\mathbf{X}}$ is uniquely determined by the full conditionals, satisfying $\forall\ \mathbf{x}'$

$$f_{\mathbf{X}}(\mathbf{x}) \ \propto \ \prod_{i=1}^{K} \frac{f_{X_i|X_j,\,j\neq i}(x_i\,|\,x_j,\ j<i;x'_j,\ j>i)}{f_{X_i|X_j,\,j\neq i}(x'_i\,|\,x_j,\ j<i;x'_j,\ j>i)}$$

Proposition. Under the positivity condition, if the Gibbs-Sampler Markov Chain is aperiodic, then for a probability-1 set of initial values $\mathbf{X}_0$, as $t \to \infty$, the probability law of $\mathbf{X}_t$ converges in total variation to the unique limiting distribution with density $f_{\mathbf{X}}$.

Bayesian vs. Frequentist Applications

Most statistical applications of MCMC involve likelihood-based estimation of parameters from data. Paradoxically, the Gibbs Sampler is applied to simulate not data $(Z_1, \ldots, Z_n)$ but parameters $\vartheta \in \mathbf{R}^p$!

Suppose for fixed but unknown parameter value $\vartheta = \theta_0$ the data are *iid* $Z_i \sim f(z|\vartheta)$. The observed data $(Z_i, \ 1 \le \imath \le n)$ are regarded as fixed, and statements about parameters ϑ *compatible* with the data are generally based on the **Likelihood**

$$L(\vartheta, \ \underline{Z}) \ = \ \prod_{i=1}^{n} \ f(Z_i \,|\, \vartheta)$$

as function of ϑ.

Frequentist statisticians often calculate:

(1) (**MLE**:) maximize $L(\cdot, \ \underline{Z})$ at $\hat{\vartheta}$, or

(2) (**Test-based CI**:) $\{\vartheta : \ \frac{L(\hat{\vartheta}, \underline{Z})}{L(\vartheta, \underline{Z})} \ \le \ \exp(\frac{1}{2} \chi^2_{p,\alpha})\}$.

Bayesian statisticians treat ϑ as random, distributed with *prior density* π, and calculate:

(3) (**Posterior density**:) $f_{\vartheta|\underline{Z}}(\vartheta \,|\, \underline{Z}) \ = \ \frac{\pi(\vartheta)\, L(\vartheta, \underline{Z})}{\int \ L(a, \underline{Z})\, \pi(a)\, da}$

Note: if we can fix prior π to be uniform over some large fixed region in $\mathbf{R}^p$ containing θ_0, then (1)-(2) can be viewed as resp. the *mode* (maximizer) and level-exceedance region for the *posterior density* (1).

So we simulate the parameter ϑ as a random variable with the posterior density, and derive quantities (1)-(3) **empirically**.

Hierarchical Models

Certain Bayesian-motivated models allow factorizations that make Gibbs Sampling particularly handy:

Hierarchy is:

$$X \sim f(x, \vartheta)\ , \quad \vartheta \sim g(\theta, \eta)\ , \quad \eta \sim h(\eta, b_0), \quad \text{etc.}$$

Additional structure used in simplifying conditionals:

Exponential families: $f(x, \vartheta) = k(x) \exp(T(x) \cdot \vartheta - \psi(\vartheta))$

Conjugate priors: if $\eta = (\mu, \lambda)$ and prior density for ϑ parameter is

$$\pi(\theta) = g(\theta, \eta) = K(\eta) = \exp(\theta \cdot \mu - \lambda\, \psi(\theta))$$

then posterior $f_{\vartheta|x}(\theta|x) = g(\theta,\ (\mu + T(x), \lambda + 1))$.

Example – Nuclear Pump Failures

Consider the following data (example pp. 301-2 in Robert & Casella 1999, from earlier paper by other authors)

	1	2	3	4	5	6	7	8	9	10
F	5	1	5	14	3	19	1	1	4	22
T	94.3	15.7	62.9	125.8	5.2	31.4	1.1	1.0	2.1	10.5

The model is that the numbers n_i of failures (F) for pump i in time T=t_i are Poisson$(\lambda_i t_i)$ r.v.'s, with

$$\lambda_i \sim \text{Gamma}(1.8, \beta) \;\; , \qquad \beta \sim \text{Gamma}(.01, 1)$$

Recall that

$$f_{\text{Gamma}(a,b)}(y) \; = \; \frac{b^a y^{a-1}}{\Gamma(a)} \, e^{-by} \; , \qquad p_{\text{Poiss}(\mu)}(k) \; = \; \frac{\mu^k}{k!} \, e^{-\mu k}$$

Then the posterior density (regarded as a joint density for the unknown parameters β and $\lambda_1, \ldots, \lambda_{10}$) is

$$\propto \prod_{i=1}^{10} \left\{ (\lambda_i t_i)^{n_i} \, e^{-\lambda_i t_i} \, \beta^{1.8} \, \lambda_i^{.8} \, e^{-\beta \lambda_i} \right\} \beta^{-.99} e^{-\beta}$$

so the conditionals are:

$$\lambda_i \sim \text{Gamma}(n_i + 1.8, t_i + \beta) \quad \text{given} \quad \beta, \lambda_j \; : \; j \neq i$$

$$\beta \sim \text{Gamma}(18.01, \; 1 + \sum_{i=1}^{10} \lambda_i) \quad \text{given} \quad \underline{\lambda}$$

Example, continued. Simulated successively from these conditionals, starting from $\beta_0, \underline{\lambda}_0$ from prior. Generated 10,000 Gibbs-Sampler iterations $(\beta_t, \underline{\lambda}_t)$.

Note that we are really interested primarily in β, although λ_i would be useful in forecasting future failures, since they are the pumpwise rates. (Even frequentists would include the λ_i if only to simplify the likelihood which is otherwise a mess involving Gamma functions !)

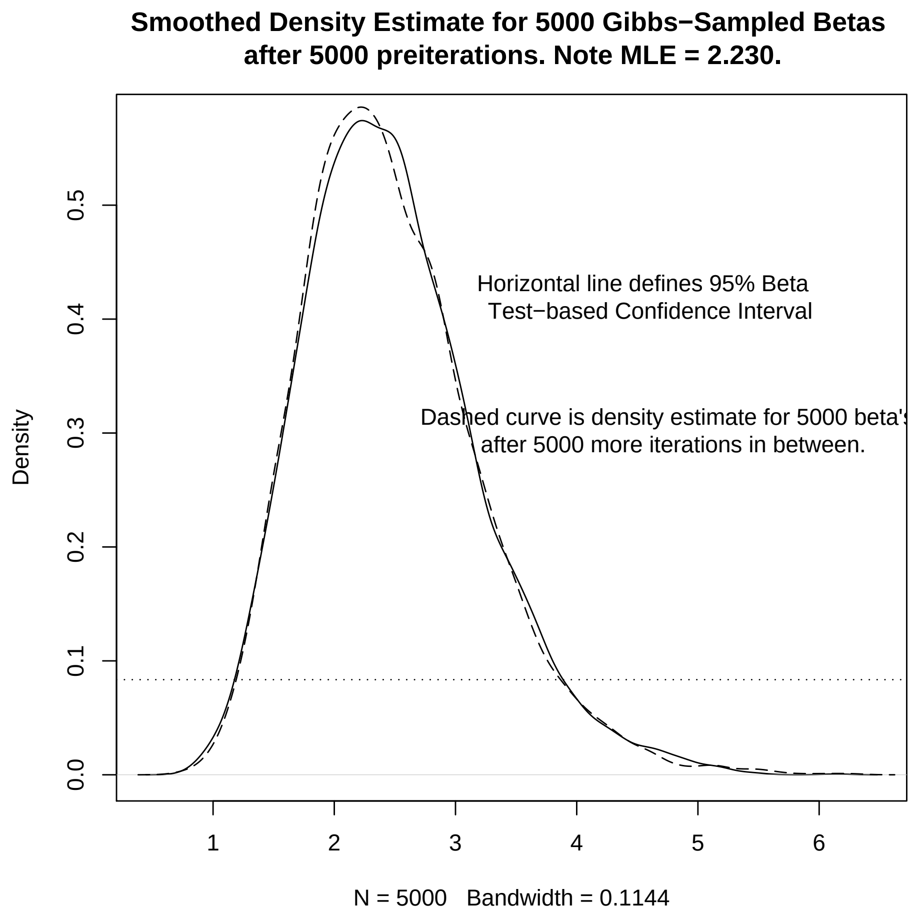

Figure 8: Smoothed density estimate for 5000 Gibbs-Sampled beta values, after 5000 burn-in iterations (solid curve). Dashed curve is density estimate for 5000 beta values after 5000 more intermediate iterations. Maximized posterior density (or likelihood) gave MLE for beta of 2.23, and test-based confidence interval for beta approximately (1.2, 3.9).

Remark about MCMC algorithms.

The choice of stopping-criterion is still not well understood: Jones and Hobert (2001) following Meyn & Tweedie (1993) and others show how to find computable theoretical bounds for rates of *geometric ergodicity*, but these may not accurately reflect algorithms' success in practice. There is room for a lot of computational experience **and** theoretical research here !

Random-Intercept Logistic Regression

An interesting class of statistical applications can be handled by either Metropolis-Hastings, MCMC, or missing-data (EM) techniques. These are statistical models with **random effects**. A good example is *random-intercept logistic regression*: suppose for experimental units $i = 1, \ldots, m$, we observe data on n_i potential occurrences and see R_i occurrences, with explanatory or *predictor* vector variables $\mathbf{W}_i$ assumed to affect the outcomes according to a model

$$R_i \sim \text{Binom}(n_i, \pi_i) \quad , \quad \log \frac{\pi_i}{1 - \pi_i} = a + \mathbf{b} \cdot \mathbf{W}_i + u_i$$

where $u_i \sim \mathcal{N}(0, \sigma^2)$ are unobservable and independent *random effects* related to unmodelled random differences between the experimental units, and $\vartheta = (a, \mathbf{b}, \sigma^2)$ are unknown statistical parameters which must be estimated (say by Maximum Likelihood). Because of the unobserved (integrated-out) variables u_i, the likelihood is complicated. An extended comparative discussion of how to calculate and maximize this likelihood is given on the Lecture 2 website

`http://www.math.umd.edu/~evs/Mini.MCMC/Lec2Figs`

or at

`http://www.math.umd.edu/~evs/s798c/Lec03Pt6.pdf`

References

Casella, G. and George, E. (1992) Explaining the Gibbs Sampler. *Amer. Statistician* **46** 167-174.

Chen, M.-H., Shao, Q-M., and Ibrahim, J. (2000) **Monte Carlo Methods in Bayesian Computation**. Springer.

Jones, G. and Hobert, J. (2001) Honest exploration of intractable probability distributions via Markov Chain Monte Carlo. *Statist. Sci.* **16** 312-334.

Robert, C. and Casella, G. (1999) **Monte Carlo Statistical Methods**. Springer.