Abstract

Examples of Applications of MCMC: Statistical Inference and Combinatorial Optimization. Reconstruction of a binary Image. Nonparametric Denstity Estimation.

Putting MCMC to Work

So what can be done with a Markov chain monte carlo method such us Metropolis? The simple answer is: A Lot! Computation of integrals, simulation of complex probabilistic models, and optimization, are just three general areas with practical applications to virtually all domains of current science and technology. During the last ten years the subject has been continuously picking up momentum at an accelerated rate and there is no sign that this trend will change much, at least in the near future. Check the online preprint service at Bristol

(*http://omega.stat.psu.edu:8008/summer99/lecture5/http://www.stats.bris.ac.uk/MCMC/*) or search for "Monte Carlo" at Los Alamos

(*http://omega.stat.psu.edu:8008/summer99/lecture5/http://xxx.lanl.gov*) xxx.lanl.gov archives. The recent discovery of the Propp and Wilson algorithm, that allows to rigorously determine convergence in some cases, has contributed to raise the expectations even further about the utility of these techniques. The
optimistic futurists

(*http://omega.stat.psu.edu:8008/summer99/lecture5/http://www.amazon.com/exec/obidos/ts/book-contents/0670882178/jonradelA/*) are even claiming that we'll be able to achieve *immortality* (no kidding) in about 40 years!... that's hype.

As simple illustrations of the applicability of MCMC we show general statistical inference and combinatorial optimization.

Statistical Inference

Recall that the subject of inference deals fundamentally with three kinds of propositions:

Data Denoted by D or x (the facts or things)

Hypotheses Denoted by H, θ, (the theories)

Prior information Denoted by O, (Everything else)

If we want to represent the amount of truth in a proposition A in the domain of discourse of another proposition B (i.e. given B) with real numbers then these numbers must (modulus re graduation) satisfy the usual rules of probability, i.e. if $P(A|B)$ denotes the truth in A given B:

Product $P(AB|C) = P(A|BC)P(B|C)$

Addition $P(A|B) + P(\text{not } A|B) = 1$

From the commutativity of conjunctions (in all domains of discourse) we have,

$$P(HD|O) = P(DH|O)$$

and using the product rule on both sides,

$$P(H|DO)P(D|O) = P(D|HO)P(H|O)$$

from where, the indispensable, trivial, and always under appreciated,

$$P(H|DO) = \frac{P(D|HO)P(H|O)}{P(D|O)}$$

rev. Bayes' magic rule is conjured. All of the quantities appearing in the holy formula have names:

$$(\text{posterior}) = \frac{(likelihood)(prior)}{(evidence)}$$

The evidence, is independent of the hypotheses and can always be recovered (in theory) as a normalization constant so it is usually dropped from the sacred formula and the equality is replaced by a proportionality sign to obtain,

$$\textit{posterior} \propto \textit{likelihood} \ \times \ \textit{prior}$$

Given the likelihood, the prior and the data, the problem of inference reduces to a problem in computing. The computation of the posterior. When we know how to sample from the posterior we know all there is to know about the inference problem (with that likelihood, prior and data). All the quantities necessary for making decisions can be computed (in principle) by MCMC methods. To be able to formalize the problem of making decisions in the face of uncertainty it is convenient to postulate the existence of a loss function that quantifies how much it is assumed to be lost when the true scenario is θ and we take action a, i.e.

$$L(\theta, a) = \text{ how much is lost when we do } a \text{ in scenario } \theta$$

The loss function provides the criteria for choosing the best action. The best action is that which minimizes the expected loss. The MCMC methods come to rescue us here also for we need to solve,

$$\min_a \int L(\theta, a)\pi(\theta|D)d\theta$$

A combination of Monte Carlo integration by sampling from the posterior $\pi(\theta|D)$ and simulated annealing for the minimization.

In estimation problems θ and a belong to the same vector space (with norm $|\cdot|$) and loss functions such as the quadratic $|\theta - a|^2$, the linear $|\theta - a|$ and the all-nothing $-\delta(\theta - a)$ (Dirac's delta) are used. Notice, that when these loss functions are used the best estimators become the posterior mean (for the quadratic), the posterior median (for the linear in dimension one) and the posterior mode. It is also worth seeing that maximum likelihood estimators are best for the all-nothing loss with $\pi(\theta) \propto 1$ (flat prior). So maximum likelihood does use a prior. It is only that sometimes it is the wrong prior!

When, the components of the data vector $D = (x_1, x_2, \ldots, x_n)$ are iid with pdf $f(x|\theta)$ the predictive distribution for a next observation $y = x_{n+1}$ is given by,

$$f(y|D) = \int f(y|\theta)\pi(\theta|D)d\theta$$

This is again done with MCMC by computing empirical averages of the likelihood by sampling from the posterior.

Combinatorial Optimization

Combinatorial optimization is another important application of MCMC methods. The problem is that of finding the global minimum of a cost function defined on a finite set of many elements. Here is a typical example.

Reconstructing a Binary Image

Consider the following problem. On a square with M pixels (e.g. the terminal screen) a few rectangles of random lengths and widths are allocated at random (e.g. by turning **on** the pixels of the rectangles). The image is then destroyed as follows: each of the M pixels is independently reversed with probability $0 < q < 0.5$. We are required to recover as close as possible the original image from these data.

The unknown true image (the theory) can be denoted by $\theta = (\theta_1, \ldots, \theta_M)$ where $\theta_j = 1$ if the j-th pixel is **on** otherwise it is 0. The hypothesis space, i.e. the space of all possible images is finite but big. There are $2^{10,000} \sim 10^{3,000}$ possible images. We can go superlative here with the usual...

> *not even if all the atoms of the universe*
> *were personal computers*
> *displaying images from this space*
> *operating day and night*
> *since the time of the big bang!*
> *not even in that case...*

not even there you would be able to see a sizable portion of the image space. That is an example of a combinatorial optimization problem.

Let us denote by $z = (z_1, \ldots, z_M)$ the corrupted image which, together with the prior knowledge of the definition of the problem, is the only thing that we have. A good looking image is obtained by solving,

$$\min_{\theta} \left\{ \sum_j \mathbf{1}(\theta_j \neq z_j) + \beta N(\theta) \right\}$$

where within the parentheses is the cost function when we guess θ but we see z. The first term counts the number of disagreements between guess and data and the second term is a constant (β) times the number of rectangles in the guessed image. The second term is a penalty to encode the prior information that the true image has only a few rectangles. It is worth noticing that this cost function can be actually deduced

(*http://omega.stat.psu.edu:8008/summer99/lecture5/http://omega.albany.edu:8008/entpriors.html*) from more or less first principles. Notice that we are trying to find the minimum of a function of $10,000$ binary variables.

Simulated annealing can be implemented by thinking of the cost function as the energy and running the Metropolis algorithm by picking at random between the following proposal moves:

1. enlarge a rectangle
2. shrink a rectangle
3. create a rectangle
4. delete a rectangle

all the moves are local, affecting only a few pixels, so the change in energy needed for the acceptance probability in the Metropolis algorithm is cheap to compute...

Homework Do it (and get a big A+)! you won't be able to do it in JavaScript though.... I don't think (but please prove me wrong) you may need plain Java for this one.

Non-parametric Density Estimation

Until recently nonparametric inference was the shame of the church of Bayes' and its latter day Saints. Good old plain asymptotics kept producing more and more theorems for the kernel estimator and its variants but the Bayesians just sat there unable to put useful priors on infinite dimensional spaces. Yes, OK they all took refuge in the Dirichlet proccess but obtaining results that were no competition for the prior-less flatlanders. The problem for the believers were many. First, subjectivity went to hell in spaces of more than a few meaningful variables. Second, they didn't dare to move too far away from the Dirichlet process for long ago it was shown that in nonparametric settings the great majority of priors produced inconsistent posteriors. That was a serious problem, and still is. For a believer with a wrong prior will get more and more convinced of the wrong thing as she collects more and more data. All of that changed with the proliferation of MCMC methods. In fact flatlanders' density estimators can be turned into non-regular likelihoods by cross-validation, use priors for the smoothness parameters (usually only a few, often only one) and let MCMC to sample from posteriors and predictive distributions.

For example consider the one dimensional kernel estimator with Gaussian kernel and scale parameter h based on iid observations of an unknown pdf f.

$$f_n(x|h, x_1, \ldots, x_n) = (2\pi h)^{-1/2} \sum_{j=1}^{n} \frac{1}{h} \exp\left(\frac{-(x-x_j)^2}{2h^2} \right)$$

Postulate a likelihood for h as,

$$\Phi(x_1, \ldots, x_n|h) \propto \prod_{j=1}^{n} f_{-j,h}(x_j)$$

where,

$$f_{-j,h}(x_j) = \frac{1}{n-1} \sum_{i \neq j} K_h(x_j - x_i)$$

is the kernel estimator of $f(x_j)$ based on all the data except the j-th observation. We denoted by K_h the $N(0, h^2)$ pdf. It is known that the h that maximizes this non-regular likelihood produces a consistent density estimator when replaced in the original kernel.

It is possible to write down a formula for the posterior mean of h as follows. First, write the likelihood as a sum of products instead of as a product of sums to get,

$$\Phi(x_1, \ldots, x_n|h) \propto \sum_{\text{all paths}} h^{-n} \exp\left(\frac{-s^2}{2h^2} \right)$$

where,

$$s^2 = s^2(\underline{i}) = \sum_j (x_j - x_{i_j})^2$$

and the sum above is over all the set of indices,

$$I = \{\underline{i} = (i_1, \ldots, i_n) : i_k \in \{1, 2, \ldots, n\} \setminus \{k\}\}$$

we call the elements $\underline{i}$ of this set, paths.

Notice that there are a total of $(n-1)^n$ paths. Typically $n \sim 100$ so the size of I is of the order of 10^{200}. Huge.

The posterior mean for h, for priors of the form $\pi(h) \propto h^{-(\delta+1)}$, can be written explicitly as,

$$\hat{h} = C_{n,\delta} \frac{\sum_{\text{all paths}} \alpha(\underline{i}) s(\underline{i})}{\sum_{\text{all paths}} \alpha(\underline{i})}$$

where $\alpha = s^{-(n+\delta)}$ and C is an explicitly known function of n and δ,

$$C_{n,\delta} = 2^{-1/2} \frac{\Gamma((n+\delta-1)/2)}{\Gamma((n+\delta)/2)}$$

Posterior means for h can be computed very accurately by MCMC for this problem. By using Metropolis to sample path $\underline{i}$ with probability proportional to $\alpha(\underline{i})$ and then computing empirical averages for $s(\underline{i})$. We know it works since the smootheness parameters produced by the method give remarkably good reconstructions of the density, much better than the ones obtained by plain cross-validation. Check it out

(*http://omega.stat.psu.edu:8008/summer99/lecture5/http://xxx.lanl.gov/abs/physics/9712041*)

It works!