http://www.math.nyu.edu/faculty/goodman/teaching/MonteCarlo20/

Class notes: Monte Carlo methods
Week 8, Collective Modes and Samplers

Jonathan Goodman

November 11, 2020

1 Bayesian example

sec:Be

This section has philosophical and technical components. The philosophical component is the reasoning behind Bayesian model identification, *uncertainty quantification* (called *UQ*) and *uncertainty propagation* (*UP*). I will advocate a particular view of these activities that I call *Zen Bayesian* because it is simple, and because it takes considerable discipline and effort to maintain this level of simplicity. May advocates of Bayesian statistics have more complicated strategies, many of which I consider to be mistaken.

On the technical side, I want to motivate the fact that many sampling problems are multi-modal. I also want to motivate the interest in calculating the *Bayesian evidence*, which is the normalization constant in the posterior distribution.

Bayesian statistics and posterior sampling are a source of hard Monte Carlo problems, starting with sampling, but not limited to sampling. Recall the framework of Bayesian statistics. You have a dataset $Y = (Y_1, \ldots, Y_N)$ and a model with parameters $x = (x_1, \ldots, x_d)$ that describes how the data were generated.

For example, suppose a signal is a superposition of simple oscillations with unknown frequency and amplitude and phase offset. The exact signal at time t_j would be

$$y_j = \sum_{k=1}^{m} A_k \cos(\omega_k(t_j - T_k)) \; . \qquad (1)$$

eom

The A_k are unknown amplitudes. The ω_k are unknown frequencies. The T_k are unknown phase offsets. The number of modes in the model is M. Suppose measurements have been made at times t_j for $j = 1, \ldots, N$. We model the data values Y_j as the true values y_j given by (1) together with measurement errors, which are modeled as independent Gaussians with a common variance σ^2. Altogether, there are $d = 3m + 1$ parameters, the $3m$ model parameters and σ^2.

The Bayesian picture is that the model parameters are chosen from a prior distribution, which we describe by a PDF $\pi(x)$. If the distribution π depends on parameters (a typical case), whose are called *hyper-parameters.* If the hyper-parameters themselves are random, their distribution is the *hyper-prior.* Once the parameters are chosen, the data are drawn from a PDF called the *likelihood function* $Y \sim L(\cdot|X)$. The joint distribution of the parameters and the data is given by Bayes' rule

$$(X, Y) \sim L(y|x)\pi(x) \; .$$

The *posterior* distribution is the conditional distribution of the parameters given the data

$$X \sim \rho(x|Y) = \frac{1}{Z(Y)}\, L(Y|x)\pi(x) \; .$$

The normalization factor

$$Z(y) = \int L(y|x)\pi(x)\, dx \; . \qquad (2) \quad \boxed{\texttt{ei}}$$

is the *evidence integral.* This terminology is explained in Exercise 2.

For our specific model (1), the observations are taken to have mean y_j and variance σ^2. The N measurement errors are assumed to be independent. This makes the PDF for the observations, conditional on the model parameters,

$$L(y|A,\omega,T,\sigma) = \frac{1}{Z_0}\sigma^{-N} e^{-\frac{1}{2\sigma^2}\sum_{j=1}^{N}(Y_j - y_j)^2} \; .$$

In this case, we know that the normalization constant is $Z_0 = (2\pi)^{\frac{N}{2}}$, but this is irrelevant for most forms of MCMC sampling. The parameters $A = (A_1,\ldots,A_m)$, and $\omega = (\omega_1,\ldots,\omega_m)$, etc, enter L through y_j and (1). It is helpful to write this in the familiar exponential form

$$L(y|A,\omega,T,\sigma) = e^{l(y|A,\omega,T,\sigma)} \; .$$

The *log likelihood* function is

$$l(y|A,\omega,T,\sigma) = -\frac{1}{2\sigma^2}\sum_{j=1}^{N}(Y_j - y_j)^2 - N\sigma \; . \qquad (3) \quad \boxed{\texttt{ll}}$$

In practical computation, is is usually best to use the log likelihood function (3) rather than the likelihood function itself. This is because the likelihood function is likely to be outside the range of floating point arithmetic. The smallest 64 bit (double precision) floating point value is about 10^{-300}. It is easy to make values this small if you have 1000 data points. However, the log is well inside the range of double precision arithmetic. In the language of statistical mechanics, the target distribution is

$$\rho(x) = e^{-\beta\phi(0)} \; , \;\; \phi(x) = -l(Y|x) \; , \;\; \beta = 1 \; .$$

The better fits have larger log likelihood and smaller potential.

I advocate what I call the *Zen Bayesian* philosophy. Instead of the frequentist approach of producing a single sample, see Exercise 2, the statistician should produce a collection of samples of the posterior distribution. The consumer of the statistician's analysis will be able to compare samples to see how much parameters vary from sample to sample. There are visualization tools to uncover correlations in the posterior distribution. This is common for models that are *ill conditioned.* A model is ill conditioned if there are large changes in parameters that make little change in the prediction. It may be, for example, that increasing x_4 and decreasing x_5 by the same amount has little influence in the predicted values y_j.

In order to produce the samples, I suggest running MCMC on the posterior, estimating the auto-correlation time, and reporting samples from the Markov chain X_k with the k values separated by some multiple of τ. It would be ideal to return independent samples, but this is not possible with MCMC.

The collection of posterior samples is a *Zen* solution of *uncertainty quantification*, or *UQ*. Uncertainty quantification means telling the consumer of a computation how accurate the computation is likely to be. In parametric statistics, you can interpret this as the uncertainty in the parameter estimates. Frequentist statisticians are taught to report confidence intervals for each of the estimated parameters. The problem with that is that this may not represent the true shape of the posterior distribution. In the ill-conditioned model, for example, the individual parameters x_4 and x_5 may have large error bars (confidence intervals), but the sum may be much more accurately determined. Said another way, a point estimate and a collection of confidence intervals determine a box in parameter space. The true posterior may have a different shape. For example, it may be multi-modal.

The need to specify a prior is a weakness of Bayesian statistics. The prior is supposed to represent your prior understanding of the world before receiving the data. In practice, our understanding may be sketchy and the prior somewhat arbitrary. For example, the uncertainty in the amplitudes in the oscillator model eom (1) might span several orders of magnitude. This might mean that $A = 1$ or $A = 10$, or $A = 1000$ are all plausible. You might think of taking $\pi(A)$ to be uniform in a range such as $[1, 1000]$. This has the drawback that $1 < A < 10$ is very unlikely. One solution to this is the *Jeffries prior*, which is uniform in log space:

$$\Pr(\alpha \leq \log(A) \leq \alpha + d\alpha) = \begin{cases} \dfrac{d\alpha}{\log(A_{\max}) - \log(A_{\min})} & \text{if } A_{\min} \leq \alpha \leq A_{\max} \\ 0 & \text{otherwise} \ . \end{cases}$$

The problem with this is that it excludes the range $0 < A < A_{\min}$. We might take $A_{\min}$ to be the smallest amplitude that is detectable, which would make it dependent on the measurement error, σ. This has the drawback that we have to change the model of the outside world (the range of A) if we get a more sensitive detector (reducing σ). The parameters $A_{\max}$ and $A_{\min}$ are hyper-parameters.

It is also hard, when making up a prior, to model prior correlations between the amplitudes. Most often, people take a prior in which the amplitudes are independent:

$$\pi(A_1, \ldots, A_m) = \prod_{j=1}^{m} \pi_j(A_j) \ .$$

Creating a prior distribution could be described as "replacing ignorance with fiction". You don't know what it should be, but you are forced to create it. One hopes that this fiction does not have a large impact on the results.

A *flat prior* is $\pi(x) = \textit{const.}$ This is not a true probability density, but it has the effect of taking away arbitrary choices in true priors.

2 Bayesian example

sec:es

In Bayesian statistics, one wants a sampler that applies "out of the box" (with lots of problem specific tuning) to a generic problem. Generic sampling problems often are "ill conditioned", which means that the target density has some directions where it varies rapidly and some where it varies more slowly. Even Gaussian distributions can go this. Suppose

$$\rho(x) = \frac{1}{Z} e^{-\frac{1}{2}x^t Hx} \ .$$

Then ρ is narrow in directions of eigenvectors of H with large eigenvalue.

Gradient descent optimization has a similar conditioning problem. Suppose we want to minimize $\phi(x)$ The gradient descent algorithm with *learning rate* s is

$$x_{k+1} = x_k - s\nabla\phi(x) \ .$$

It is convenient to write this in terms of a *search direction* $p_k = -\nabla\phi(x)$, so $x_{k+1} = x_k + sp_k$.

3 Exercises

sec:ex

ex:ge

1. Consider the linear Gaussian process

$$X_{k+1} = AX_k + Z_k \ , \quad Z_k \sim \mathcal{N}(0, C) \ .$$

Suppose that A is symmetric and C is symmetric and positive definite. Let $\rho_k(x)$ be the PDF of X_k. Show that there is recurrence formula of the form

$$\rho_{k+1}(x) = \int \rho_k(y) L(y, x)\, dy \ .$$

Find an explicit Gaussian like formula for L. A *left eigenfunction* of L with eigenvalue λ is a function that satisfies

$$\int v(y) L(y, x)\, dy = \lambda v(x) \ .$$

Show that the invariant distribution ρ is an eigenfunction with eigenvalue $\lambda = 1$. Let r_j be the eigenvectors of A and suppose

$$Ar_j = \mu_j r_j \ .$$

These are real because A is symmetric. Show that

$$v_j(x) = r_k \cdot \nabla\rho = \sum_{i=1}^{d} r_{ji} \partial_i \rho$$

is an eigenfunction with eigenvalue μ_j. More generally, show that if μ_{j_m} is a family of eigenvalues of A, then

$$v = \prod_m r_m \cdot \nabla\rho$$

is an eigenfunction of L with eigenvalue $\lambda = \prod \mu_{j_m}$. Assuming that these are all the eigenfunctions, show that the spectral gap for L is the spectral radius of A.

ex:ei

2. One use of the evidence integral is *model selection.* We want to find which of a family of models best fits the data. The *maximum likelihood* estimate of *frequentist* statists, given model i, is the parameter combination that gives the best fit:

$$\widehat{X}_i = \arg\ \max_x\ L_i(Y, x_i)\ . \qquad (4)$$

ml

This is a *point estimate* because you are giving a single "best guess" of the parameter, which is a point in parameter space. One model selection idea would be to take the model that best fits the data

$$\widehat{i} = \arg\ \max_i\ L_i(Y, \widehat{X}_i)\ . \qquad (5)$$

bf

This can lead to *over-fitting*, which means that the model has so many parameters that it can fit any data.

Suppose the models are M_i with parameters $x_i \in \mathbb{R}^{d_i}$, priors $\pi_i(x)$, and likelihood functions $L_i(y|x_i)$. Let $Z_i(Y)$ be the evidence integral for model i and data Y. In a Bayesian approach, we would specify priors for the model, which means model i is chosen with probability r_i. Once the model is chosen, the Bayesian picture says we choose the parameters $X_I \sim \pi_i(x_i)$ and then the data $Y \sim L_i(y|X_i)$. Derive the following formula for the posterior probabilities

$$s_i(Y) = \Pr(\text{ model } i\ |Y) = \frac{Z_i(Y) r_i}{\sum_j Z_j(Y) r_j}\ .$$

The posterior probability is the prior probability amplified or reduced by the *evidence* for the model.

ex:dbs

3. We have at least two samplers for a distribution $\rho(x) = \frac{1}{Z} e^{-\beta\phi(x)}$. One is the Hamiltonian sampler together with Verlet time stepping and resampling of the momentum. Another is *MALA*, which stands for *Metropolis adjusted Langevin.* The Langevin sampler, in continuous time (which doesn't exist in the computer but can exist in our thoughts) is the stochastic differential equation, which is (not entirely properly) called the *Langevin* equation

$$dX = -\nabla\phi(X)dt + \sqrt{2\beta}\, dW_t\ .$$

The dW_t is the increment of Brownian motion in a time dt. For Δt small but not zero, the *Euler Maruyama* approximation calculates an approximate trajectory $X_k \approx X_{k\Delta t}$ using

$$X_{k+1} = X_k - \Delta t \, \nabla\phi(X_k) \, \Delta t + \sqrt{2\beta\Delta t} Z_k \ .$$

Here $Z_k \sim \mathcal{N}(0, I)$ are independent Gaussians with mean zero and covariance I in d dimensions. The continuous time Langevin process preserves the target distribution exactly. For small Δt, the Euler approximation preserves a distribution that is close to ρ. The disadvantage of small Δt is that many steps are required to get an effectively new sample (large autocorrelation time). For large Δt, the discrete time process has an invariant distribution that is far from ρ. This can be fixed by a Metropolis rejection step.

Take as target distribution a multi-variate double-well distribution that is a *Gaussian mixture*

$$\rho = p_1 \mathcal{N}(\mu_1, C_1) + p_2 \mathcal{N}(\mu_2, C_2) \ .$$

The *mixture coefficients* p_1 and p_2 should be positive and have $p_1 + p_2 = 1$. A simple case in d dimensions has $\mu_= 0$, $\mu_2 = (r, 0, \ldots, 0)$, and $C_1 = C_2 = \frac{1}{s^2} I$. The wells are distinct if $r \gg s$. Samplers are slow to go from well to well if $\frac{r}{s}$ is large. You can make many variations on this model, by varying the mixture coefficients and the covariance matrices. It may be that anisotropic (C_1 or C_2 not a multiple of I) covariance is harder than isotropic covariance.

This is another research project. Please spend some time playing with your codes and report some findings. But, to start,

(a) To *Metropolize* the discrete Langevin move, take the move to be a proposal

$$Y \sim X_k - \nabla\phi(X_k)\Delta t + \sqrt{2\beta\Delta t} Z_k \ .$$

Figure out the Metropolis Hastings acceptance probability and apply it to make an algorithm that samples exactly.

(b) Write a Hamiltonian sampler with parameters Δt, n (the number of Verlet steps between the acceptance/rejection test), and a (the damping factor in the momentum resampler).

(c) Verify that your samplers are correct for the one dimensional case by making a histogram of the samples and checking that they agree with the target. Then try $d = 2$ and check that the mean and covariance of the samples is correct.