MCMC

February 5, 2021

Posterior simulation

- For the situation where we know $\pi(\theta)p(Y \mid \theta) = f(y, \theta)$, in the sense that we can write an expression or program that gives its value for each possible value of θ, but we do not know how to draw randomly from the pdf in θ defined by $f(Y, \theta) / \int f(Y, \theta) d\theta$, because it defines no standard distribution.

- Examples in models we have discussed before: linear regression with t-distributed errors; mixed-normal regression for a heterogeneous population.

- Two main approaches: Importance sampling and Markov chain Monte Carlo (MCMC).

Importance Sampling

- Suppose θ has pdf $p(\theta)$, defining a non-standard distribution. We would like to calculate the expectation of a function $g(\theta)$ under the distribution defined by p.

$$E_p[g(\theta)] = \int g(\theta)p(\theta)d\theta = \int \frac{g(\theta)p(\theta)}{q(\theta)}q(\theta)d\theta = E_q\left[\frac{g(\theta)p(\theta)}{q(\theta)}\right].$$

  for any pdf $q(\theta)$.

- So estimate the expectation of θ under p by drawing randomly from the distribution defined by q and weighting the draws of $g(\theta)$ by $p(\theta)/q(\theta)$.

- requires that $p(\theta) > 0 \Rightarrow q(\theta) > 0$. In fact there are problems even if p/q just becomes very large in parts of the parameter space, because this tends to make a few, rare draws completely dominate the weighted average.

MCMC: General principles

- Given a draw θ_j, one generates a new draw θ_{j+1} from a distribution that may depend on θ_j (but not on earlier draws). The draws are generally serially correlated across j (unlike the importance sampling draws), but eventually their sample distribution function converges to that of the target distribution.

- Need to have the target a fixed point. Often proving this can proceed by showing that, when θ_j is drawn from the target p_0 pdf, the transition mechanism implies that the joint pdf $p(\theta_{j+1}, \theta_j)$ satisfies $p(\theta_{j+1}, \theta_j) = p(\theta_j, \theta_{j+1})$.

More principles

- But then need also to insure that the algorithm will not get stuck. This will depend on the particular algorithm, on the shape of the boundaries of the parameter space, and on the nature of the target pdf.

- Can't even get the target to be a fixed point if the target is not integrable. Note that we do not need to know the target's scale in order to implement these algorithms, so failing to detect non-integrability is a real possibility.

- These methods really do require the Markov property. One can be tempted to systematically tune up the algorithm based on what has been learned about the target distribution from previous draws. If this is done systematically and repeatedly, it makes the algorithm deliver wrong answers.

Checking convergence and accuracy

- Accuracy: *assuming* current sample is representative, do we have enough accuracy for our purposes?
- Accuracy can be different for different functions of β in the same sample.
- Convergence: Can we treat this sample as "representative", i.e. as having visited all relevant regions and displayed all relevant modes of variation?

Effective sample size

- One approach is based on modeling the serial dependence in the draws. The autocovariance function (ACF) is

$$R_\theta(k) = \mathrm{Cov}(\theta_j, \theta_{j-}) \, .$$

- Assuming convergence, we can estimate the ACF of the draws, and compute

$$\mathrm{Var}\left(\frac{1}{T}\sum \theta^j\right) = \frac{1}{T}\sum_{j=-T}^{T} \frac{|T-j|}{T} R_\theta(j) \doteq \frac{1}{T}\sum_{-\infty}^{\infty} R_\theta(j)$$

- This can be compared to what the variance of the mean θ_j would have been if the sample were i.i.d.:

$$\frac{1}{T} R_\theta(0) \, .$$

- Then we can use as a measure of "effective sample size"

$$\frac{T R_\theta(0)}{\sum R_\theta(j)} \, .$$

- It is not uncommon for MCMC to produce effective sample sizes of only a few hundred when the number of draws is in the hundreds of thousands. This implies such great dependence across j that even the effective sample size numbers, not to mention the mean of θ^j, is unreliable. It is likely to be a symptom of non-convergence.

Another approach to effective sample size

- Break the sample into k pieces indexed by i. Calculate sample average $\bar{g}_i$ of $g(\beta_j)$ for each piece. $1/\sqrt{k}$ times sample standard deviation of the $\bar{g}_i$'s is an estimate of the standard error of the sample mean from the overall sample.

- This is accurate to the extent that the pieces are long enough so that dependence between them is weak. Span of dependence among β_j draws must be considerably shorter than the length of the pieces.

- Variance from the whole sample, if had i.i.d. sample, would be N/k times the variance of sample means across pieces. So if s_k^2 is the sample variance of the means of the pieces and s_N^2 the sample variance from the whole sample, effective sample size is $ks_N^2/(Ns_k^2)$. This could in principle be larger than N, but in practice is usually much smaller than N.

Convergence checks based on subsample blocks

- That effective sample size is similar with different choices of k and is growing more or less linearly with N is a criterion for convergence.

- We are more suspicious of non-convergence in the beginning blocks of an MCMC chain. So some convergence checks (e.g. one suggested by Geweke) compare behavior of draws in the first part of the sample to behavior in the latter part.

Convergence checks based on multiple chains

- Start from different places.
- After one or two, start from a place that is fairly unlikely according to initial runs. Variation across runs from different starting points can be treated like variation across pieces of the sample.
- Often this leads to different conclusions about accuracy and convergence than working with pieces of a single run.

Is accuracy sufficient?

- If convergence is ok, the sample size may or not be big enough: That depends on whether the estimated accuracy of your estimate of $E[\beta]$ is within tolerances based on substantive considerations.

- Thinning. If effective sample size is running at about $N/10$, why not throw out all but every 10'th draw? This will make the result look more like an i.i.d. sample, but will not improve, and may harm, accuracy of estimates of $E[g(\beta_j)]$. However, it is often done, because the cost of moderate thinning (more like every 3rd or 4th draw, here) in reduced accuracy will be relatively small compared to the savings in disk space, if all the draws are being saved.

- Accuracy may be adequate for some g's and not others. Effective sample size may differ across g's. But if convergence looks bad for one g, it should not be very comforting that for other g's it looks ok.

Trace plots

- These simply plot elements of θ^j, or functions $g(\theta^j)$, against j. They will show clearly trending behavior, or slow oscillations, or switches between regions with different characteristics. The coda package (there is an R version, but also an Octave/Matlab version) generates these with a single command.

- They are not foolproof. If there are large high-frequency oscillations, they may obscure trends and low-frequency oscillations. In large MCMC samples the plot may even look like a black smear.

- If effective sample size is small, but the trace plots are black, it may help to do trace plots of thinned samples.

Pitfalls of high-dimensional models

- If the posterior has a $N(0, I)$ shape and the parameter vector is of length n, $E[||\beta - \hat{\beta})||]^2 = n$. the ratio of the posterior density at the peak to the posterior density at a given β is $\exp(-\frac{1}{2}||\beta - \hat{\beta}||^2)$. Thus draws from the posterior at a "typical" distance from the peak have posterior density about $e^{-n/2}$ times smaller than the posterior density at the peak.

- An MCMC run started at or near the peak may take a long time to reach "typical" levels of $\|\beta - \beta_{MLE}\|$.

- All the draws will look nice, but they will give a misleadingly optimistic picture of uncertainty about the parameters.

- So: Check trace plots of the log posterior density as well as of parameters.

Metropolis algorithm

Target kernel $p(\theta)$

Proposal density $q(\theta' \mid \theta)$

Procedure

1. Generate a proposed θ^*_{j+1} from θ_j using the $q(\theta^*_{j+1} \mid \theta_j)$ distribution.
2. Calculate $\rho = p(\theta^*_{j+1})/p(\theta_j)$.
3. Draw u from a $U(0,1)$ distribution.
4. If $\rho \geq u$, set $\theta_{j+1} = \theta^*_{j+1}$. Otherwise $\theta_{j+1} = \theta_j$.

Proof of fixed point property for Metropolis

We want to show that if θ is drawn from the target density $p(\theta)$, and θ' is then drawn by the Metropolis algorithm conditioned on θ as the previous draw using a q that is symmetric, i.e. with $q(\theta' \mid \theta) = q(\theta \mid \theta')$, the resulting joint distribution is symmetric in θ and θ'. This implies that θ' has the same marginal distribution as θ, i.e. that given by the target density p.

The joint distribution of θ and θ' consists of a certain amount of probability concentrated on the $\theta = \theta'$ subspace, plus a density over the rest of the parameter space. The part on the $\theta = \theta'$ subspace is obviously symmetric in θ and θ', so we can confine our attention to the density over the rest.

Symmetry of the density

Consider the region in which $p(\theta) < p(\theta')$. In this region, draws are never rejected. Since the draws are not rejected in this region, in this region the joint density is just $q(\theta' \mid \theta)p(\theta)$.

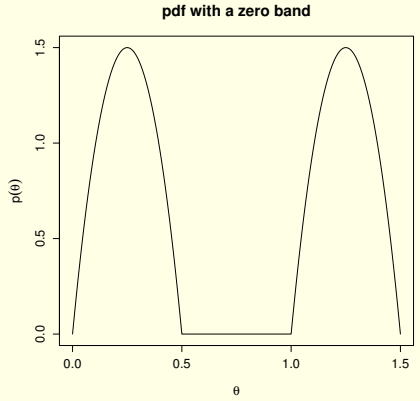
Now consider the part of the parameter space, in which $p(\theta') < p(\theta)$. In this region, a proposal draw θ' is accepted with probability $p(\theta')/p(\theta)$, so the joint density is

$$\big(p(\theta')/p(\theta)\big)q(\theta' \mid \theta)p(\theta) = p(\theta')q(\theta' \mid \theta) = p(\theta')q(\theta \mid \theta')$$

where the last equality invokes the symmetry condition. Since $p(\theta') < p(\theta)$, this last expression is exactly the density at the symmetric opposite point — where θ' is the previous point and θ is the proposal. □

How to pick jump distributions, pitfalls

If the parameter space Θ has pieces separated by bands with zero probability under the target density, the proposal density must have large enough support to jump across the bands.



Obviously for the pictured target density, the jump density has to have support spanning an interval of length greater than .5.

More on jump distributions

- Even with a jump density with unbounded support, like a normal density, if the jump probability is concentrated mainly in a small region around 0, it may take an extremely long time before the Metropolis iterations actually start sampling the second lobe of the distribution.

- In a multivariate distribution that is approximately Gaussian in shape in its high-probability region, practice suggests that finding the local Gaussian approximation in the neighborhood of the peak $\hat{\theta}$, i.e. a $N(\hat{\theta}, \Sigma)$ with $\Sigma = -(\partial^2 \log(f(Y, \theta))/\partial\theta\partial\theta')^{-1}$, and then taking the jump distribution to be $N(0, k\Sigma)$, with k about .3, is a practical starting point.

Picking k

- By making k very small, we can make the acceptance rate for draws very high, but since the steps the algorithm takes will be very small, serial correlation will be high and convergence slow.

- By making k very large, we can make the steps *proposed* large, but this may make the rate at which proposals are rejected very high, which again will create serial correlation and slow convergence.

- And this strategy of using the normal approximation around the peak only makes sense if the target density is in fact somewhat Gaussian in shape.

- For a bimodal distribution, one might use a jump distribution with a $k\Sigma$ that varied with θ, matching the local expansion at one mode when the last draw was near that mode and matching the local expansion at the other node when near the other one.

Metropolis-Hastings

It is possible to use a jump distribution that does not satisfy $q(\theta' \mid \theta) = q(\theta \mid \theta'))$, by modifying the rule for accepting candidate draws. Instead of $\rho = p(\theta^*_{j+1}/p(\theta_j)$, one uses

$$\rho = \frac{p(\theta^*_{j+1})q(\theta_j \mid \theta^*_{j+1})}{p(\theta_j)q(\theta^*_{j+1} \mid \theta_j)} \, .$$

Independence Metropolis-Hastings

This more general setup allows using a jump distribution that doesn't depend on θ_j at all, which is known as **independence Metropolis-Hastings**. If one can make draws from a pdf $q()$ that is very similar to the target $p()$, then independence M-H will make most ρ values very near one, allow keeping nearly all draws, and will be very efficient. But as with importance sampling, if there are values of theta where $q(\theta)$ is nearly zero, while $p(\theta)$ is not, the algorithm will not perform well. (Why not?)

Gibbs sampling

- Suppose our parameter vector has two blocks, i.e. $\theta = (\theta_1, \theta_2)$, and that we use as our proposal distribution $q(\theta' \mid \theta) = p(\theta_2' \mid \theta_1)$, holding $\theta_1' = \theta_1$.

- Since this makes the ratio of the proposal density to the target density constant, it satisfies the conditions for an M-H algorithm if we accept every draw.

- If on the next draw we draw from $p(\theta_1' \mid \theta_2)$ and then continue to alternate, the algorithm is likely (subject to regularity conditions) to converge.

- Obviously this is a useful idea only if, despite p being non-standard, the conditional densities for the blocks are standard. This happens more than you might expect.

Metropolis-within-Gibbs, etc.

- The fixed-point theorems for these algorithms concern a single step.
- Therefore the fixed point property holds even if we change the algorithm, say alternating Metropolis with Gibbs, with M-H, etc.
- Also Gibbs can obviously be done for more than two blocks.
- A common situation: Gibbs can be done for k of n blocks, but there is a small number $n - k$ of blocks for which the conditional distributions are non-standard.
- One can then do straight Gibbs for the k blocks, and for the other $n - k$ use Metropolis steps with the conditional densities as targets.

The heterogeneous linear regression example

- We suppose that the individuals in our sample are random draws from one of two sub-populations, in each of which there is a linear regression model connecting X to y.

- The regression for observation j, conditional on the sub-population $k(j)$, is

$$y_j \mid \left\{ X_j, \beta_{k(j)}, \sigma^2_{k(j)}, k(j) \right\} \sim N(X_j \beta_{k(j)}, \sigma^2_{k(j)}) \; .$$

- The probability of the observation being drawn from population $k = 1$ is p, and of course the probability of a draw from population $k(j) = 2$ is $1 - p$.

Likelihood, posterior

- It is natural to assume the $k(j)$ values are independent across j, but this is only reasonable when we think of their distribution conditional on p.

- This is an example where we want to make the sequence of $k(j)$ values exchangeable: independent conditional on p, but in our prior joint distribution for $\{k(j)\}, p$, there is dependence because of the dependence of all k's on p.

Likelihood including $k(j)$'s as "parameters"

With a flat prior on p and the other parameters *including the* $k(j)$ *sequence*, the joint distribution of parameters and data, conditional on the X_j's, is

$$\prod_{j=1}^{N} p^{k(j)}(1-p)^{2-k(j)}\phi(y_j; X_j\beta_{k(j)}, \sigma^2_{k(j)}) \ ,$$

where $\phi(x; \mu, \sigma^2)$ is the normal density function with mean μ and variance σ^2.

Integrating over $k(j)$

Since the unobserved $k(j)$ for each j can take on just the two values 1 or 2, we can integrate it out of the posterior distribution by just summing over its two possible values for each j, producing the usual form of a "mixture model" likelihood:

$$\prod_{j=1}^{N} \left(p\phi(y_j; X_j\beta_1, \sigma_1^2) + (1-p)\phi(y_j, X_j\beta_2, \sigma_2^2) \right) .$$

This is a simplification of the form of the likelihood, as we have eliminated $2N$ objects that were being treated as unknown parameters. But retaining the $k(j)$'s allows an MCMC strategy known as **data augmentation** that allows Gibbs sampling. (Data augmentation gets its name because in a sense we treat $k(j)$ as observable "data".)

Gibbs sampling for this model

- If we knew the $k(j)$ values for each observation j, the posterior density would just factor into two pieces, one for each sub-population, with the likelihood of a standard normal linear regression model.

- So conditional on the k's and p, we can easily draw from the conditional distribution of the other parameters.

- Also, if we know the $k(j)$ sequence, the conditional distribution of p given the k sequence is simple to draw from: we have N i.i.d. draws from a two-point distribution with probability p on point 1 and $1-p$ on point 2.

- Then how to draw the $k(j)$'s? Each $k(j)$, conditional on p, enters the joint pdf only via the likelihood element for observation j. $k(j)$ has a value in the set $\{1, 2\}$. This is exactly the situation of model choice we discussed in the 517 lecture on model choice. The relative probabilities of $k = 1$ or $k = 2$ are determined by the mdd's of the two models — but here from a single observation for each j.