MCMC

1. Posterior simulation

- For the situation where we know $f(Y\,|\,\theta)$, 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.
- Two main approaches: Importance sampling and Markov chain Monte Carlo (MCMC).

2. 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.

3. 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 $p(\theta_{j+1},\theta_j) = p(\theta_j,\theta_{j+1})$.
- 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.

4. 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?
- Both are likely to involve fitting models of dependence to the $\{\beta_j\}$ sequence. Sophisticated models are time series models, beyond our scope.
- Breaking the sample into pieces, checking for homogeneity
  - 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.
  - Results should not be drastically different when the number of pieces is changed. This is a convergence check.
  - Are results from first part of the sample similar to later parts (after, usually, discarding some initial fraction of the sample)? This is a convergence check.
  - "Effective sample size": 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/s_k^2. This could in principle be larger than N, but in practice is usually much smaller than N.
  - 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.
  - If convergence looks ok, accuracy estimates can be based either on fitting a model of dependence to the whole artificial sample, or simply on the between-group variance estimates.

- 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.
- 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.
- 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.

5. Metropolis algorithm

6. How to pick jump distributions, pitfalls

7. Gibbs sampling

8. Metropolis-within-Gibbs, etc.

9. Metropolis-Hastings

- Removes the requirement for symmetry in the Metropolis jump distribution.
- (1) Draw θ^*_{j+1} from $q(\cdot \mid \theta_j)$.
  (2) Construct
  $$\rho = \frac{p(\theta_{j+1})q(\theta_j \mid \theta_{j+1})}{p(\theta_j)q(\theta_{j+1}/\theta_j)}.$$
  (3) If $\rho > 1$, $\theta_{j+1} = \theta^*_{j+1}$.
  (4) Otherwise $P[\theta_{j+1} = \theta^*_{j+1}] = \rho$; $P[\theta_{j+1} = \theta_j] = 1 - \rho$.

10. Rejection Sampling