STAT 4224 Final Review
Carlyle Morgan UNI: scm2195
December 12, 2021

1 Numerical integration and optimization

Our objective is typically to compute $p(\theta|y)$ and $p(\tilde{y}|y)$, the posterior and posterior predictive distributions respectively.

We have a few go-to methods for accomplishing this:

- Analytic Calculation (the subject of the first half of the semester)
- Discrete approximation on a dense grid
- Monte Carlo Simulation

1.1 Finding Posterior Models

Define $\hat{\theta}_{MAP} = argmax\{p(\theta|y)\} = argmax\{logp(\theta) + logp(y|\theta)\}$. This is essentially the estimation of theta using the posterior mode. This can also be accomplished by solving $\frac{d}{d\theta} logp(y|theta) = 0$

But what if there is not a "analytic solution"? Use Newton's method:

1. Choose initial value θ^0
2. For $t = 1, 2, 3, ...$ with $L(\theta)$ as the log of the posterior of theta, calculate $L'(\theta^{t-1})$ and $L''(\theta^{t-1})$ (which is a Hessian matrix if theta is multi-dimensional).
3. Calculate $\theta^t = \theta^{t-1} - (L''(\theta^{t-1}))^{-1} L'(\theta^{t-1})$

   θ^t is selected to be the maximizer of the quadratic approximation to $L(\theta)$ at θ^{t-1}

   Repeat all steps until $\theta^t \approx \theta^{t-1}$. This can be accomplished in R via the NLL function.

1.2 Numerical Integration

Say we wish to find $E[h(\theta)|y] = \int h(\theta p(\theta|y) d\theta$.

We can use the deterministic or stochastic methods:

- Deterministic: $E[h(\theta)|y] \approx \sum_{i=1}^{s} w_i h(\theta^2) p(\theta^2|y)$ where w_i is the width of the ith rectangle. Use this if sampling of θ^2 is structured.
- Stochastic: $E[h(\theta)|y] \approx \frac{1}{s} \sum_{i=1}^{s} h(\theta^i)$. Use this if sampling is random. By law of large numbers, $lim_{s\to\infty} \frac{1}{S} \sum_{i=1}^{s} h(\theta^i) = E[h(\theta)|y]$.

  The stochastic and deterministic methods achieve similar results because $P(\theta^i \in I_k) \approx w_i P(\theta^2|y) \approx \frac{countof\theta^i \in I_k}{S}$

1.3 Distributional Approximation

One can approximate the posterior distribution by some simple parametric distribution from which integrals can be computed directly.

A really common practice is using the normal distribution to approximate posteriors. Why? The Taylor approximation to $L(\theta)$ centered at $\hat{\theta}_{MAP}$ is:

$logp(\theta|y) \approx logp(\hat{\theta}_{MAP}) + \frac{1}{2}(\theta - \hat{\theta}_{MAP})^T [\frac{d^2}{d\theta^2} logp(\theta|y)]_{\theta=\hat{\theta}_{MAP}} (\theta - \hat{\theta}_{MAP})$, which means that $p(\theta|y) \sim N(\hat{\theta}_{MAP}, [I(\hat{\theta}_{MAP})]^{-1}])$, where I is the observed information matrix: $I(\theta) = \frac{d^2}{d\theta^2} logp(\theta|y)$. This is an extension of the Bayesian CLT.

2 Monte Carlo Sampling

There is a hierarchy of ideal situations when doing MC sampling. From most to least ideal:

1. We know the posterior distribution
2. We can calculate the posterior on a discrete grid of points. Suppose for **evenly spaced** $(\theta_1, ..., \theta_n)$, these thetas are reasonably dense, span the entire range of nontrivial posterior probabilities, and that n is not too large. Any function that is proportional to the prior times the likelihood is called an **unnormalized posterior**. If approximation doesn't work with an unnormalized target density, it's probably not worth considering.
3. **Rejection Sampling**:
   - Inputs:

- $q(\theta|y)$, a possibly unnormalized target density
- $g(\theta)$, a proposal density that satisfies:
  - $q(\theta|y) > 0 \Rightarrow g(\theta) > 0$
  - $\frac{q(\theta|y)}{g(\theta)} \leq M$ with M known.
  - It is possible to sample $\theta \sim g(\theta)$ directly.

- Outputs: $(\theta_i)_{i \in S} \sim p(\theta|y)$
- Algorithm:
  - Sample $\theta^* \sim g(\theta)$ and $U^* \sim U(0,1)$
  - If $U^* \leq \frac{q(\theta^*|y)}{Mg(\theta^*)}$, then $\theta^i = \theta^*$. Else, reject this sample and repeat from the first step.

4. **Importance Sampling**:

- Motivation: Suppose we want to evaluate the posterior expectation $E[h(\theta)|y] = \int h(\theta p(\theta|y) d\theta = \frac{\int h(\theta) q(\theta|y) d\theta}{\int q(\theta|y) d\theta}$.
- Inputs:
  - $q(\theta|y)$, a possibly unnormalized target density
  - $g(\theta)$, a proposal density that satisfies:
    - $q(\theta|y) > 0 \Rightarrow g(\theta) > 0$
    - It is possible to sample $\theta \sim g(\theta)$ directly.
- Outputs: S_{eff}
- Algorithm:
  - Sample $\theta^* \sim g(\theta)$ and calculate $w(\theta^*) = \frac{q(\theta^*|y)}{g(\theta^*)}$ and repeat S times.
  - Calculate $\tilde{w}(\theta_i) = \frac{w(\theta_i)}{\sum_{j=1}^{s} w(\theta_j}$ for all samples. These w are the "importance weights".
  - Calculate $S_{eff} = (\sum_{i=1}^{S} (\tilde{w}(\theta_i))^2)^{-1}$
    The more wildly the importance weights vary, the smaller the effective sample size. If $\tilde{w}(\theta_i) = \frac{1}{S}$, $S_{eff} = S$. S_{eff} means roughly that an IS approximation based on S draws from $g(\theta)$ has the same precision as would S_{eff} draws from $p(\theta|y)$

5. **Markov Chain Monte Carlo(MCMC):**

   MCMC is a partial realization of an ergodic markov chain for which the posterior is the unique stationary distribution.

- Inputs:
  - $q(\theta|y)$, a possibly unnormalized target density
  - θ^0 starting value
  - A "jump proposal distribution" $J(\theta^*|\theta)$, where if the chain is at $\theta^{t-1}, \theta^* \sim J(\theta^*|\theta^{t-1})$ is the "proposed jump" for θ^t
- Outputs: $(\theta_i)_{i \in S} \sim p(\theta|y)$
- Algorithm:
  - Sample $\theta^* \sim J(\theta^*|\theta^{t-1})$
  - Compute $r = \frac{q(\theta^*)p(y|\theta^*)}{q(\theta^{t-1})p(y|\theta^{t-1})}$
  - Sample $U^* \sim Unif(0,1)$. If $U \leq r$, set $\theta^t = \theta^*$, else $\theta^t = \theta^{t-1}$

2.1 FAQs about MCMC Sampling

2.1.1 Why does MCMC work?

The ergodic theorem guarantees that θ^s converge in distribution to the posterior and that the mean value of function h evaluated over the entirety of the MC samples is that function given y.

2.1.2 Why use MCMC?

Direct sampling from the posterior isn't possible.

2.1.3 How do we choose the jump distribution?

There are two popular methods: Metropolis-Hastings Independence Sampler(MHIS) and Metropolis Random Walk(MRW):

1. **Metropolis-Hastings Independence Sampler(MHIS)**:
   $J(\theta^*|\theta) = g(\theta^*)$ for some pdf g, where g is nonzero when the posterior is nonzero.
2. **Metropolis Random Walk(MRW)**: For $J(\theta^*|\theta) = J(\theta|\theta^*)$(reversibility)
   (a) $\theta^* \sim unif(\theta^{t-1} - \delta, \theta^{t-1} + \delta)$
   (b) $\theta^* \sim N(\theta^{t-1}, \delta^2)$

We should be choosing J such to minimize the autocorrelation in the resulting chain. For MHIS, we can do this by choosing a good $g(\theta)$. For MRW, we want an ideally sized δ, as if δ were large, there would be a lot of rejections and the sampler would be inefficent, but if δ were too small, there would be high autocorrelation again.

2.1.4 What does the posterior is the unique stationary distribution mean?

Once you're sampling from the stationary distribution, you're always sampling from the stationary distribution.

2.1.5 How do we choose starting values for MCMC?

Ideally, we'd like $\theta^0 \sim p(\theta|y)$, but this is kind of cheesing it. We'd also like θ^0 close to a posterior mode if possible. Lastly, the tried and true strategy is to do a "burn-in", where one picks some arbitrary θ^0, runs the algo for S^0 updates, and then sets $\theta^0 = \theta^{S_0}$. To go even stronger, do a burn-in with 2-5 different starting values to obtain 2-5 different chains. Combine the post-burn in samples from each chain to make the final samples.

2.1.6 Effective Sample Size for MCMC

The higher the autocorrelation in the chain, the lower the effective sample size.

2.1.7 How many samples do we really need in OMC?

Choose S large enough such that the margin of error is less than the precision to which you want to report $E(h(\theta)|y))$:

$$S \geq \frac{4\hat{V}}{(marginoferror)^2}$$

where $\hat{V}$ is $\frac{1}{S-1}\sum_{i=1}^{S}(h(\theta)^2 - \bar{h}(\theta))^2$.

The "posterior uncertainty" about $h(\theta)$ is never going away no matter how big S is. MC error adds almost nothing to the uncertainty coming from the posterior variance in most large samples.

2.1.8 Convergence in MCMC vs OMC

1. In OMC, $Pr(\psi^i \in A) = \int_A p(\psi|y)d\psi$, whereas in MCMC, $Pr(\psi^i \in A) = \int_A p(\psi|y)d\psi$.

2. In both MCMC and OMC, $\frac{1}{S}\sum_{i=1}^{S}\psi^i \to E(\psi|y) = \int \psi p(\psi|y)$ as $S \to \infty$

3. In OMC, $corr(\psi^i, \psi^{i+1}) = 0$ since draws are i.i.d. In MCMC $= corr(\psi^i, \psi^{i+1}) = P_t \neq 0$(generally).

4. The OMC standard error is the square root of $var_{mc}(\bar{\psi}) = E[(\bar{\psi} - E(\psi|y))^2] = \frac{var(\psi^i)}{S} = \frac{var(\psi|y)}{S}$. Meanwhile,

$$var_{mcmc}(\bar{\psi}) = var_{mc}(\bar{\psi}) + \frac{1}{S^2}\sum_{i\neq j}^{S}\sum_{i\neq j}^{S} E[(\psi^i - \psi^0)(\psi^j - \psi^0)]$$

, whose rightmost term depends on correlation within the chain.

2.1.9 Effective Sample Size for MCMC

$var_{mcmc}(\bar{\psi}) = \frac{var(\psi^i)}{S_{eff}}$. It can be proved that

$$\lim_{S\to\infty} S var_{mcmc}(\bar{\psi}) = var(\psi|y)[1 + 2\sum_{t=1}^{\infty} P_t]$$

and that $S_{eff} = \frac{S}{1+2\sum P_t}$. Where P_t is the lag t autocorrelation in the chain.

We don't use this formula in practice because calculation of S_{eff} can be unstable with IS and MCMC. Instead we calculate MCMC Standard Error by another method, such as batch means, then

$$S_{eff} = \frac{\hat{V}_\psi}{var_{mcmc}(\bar{\psi})}$$

.

2.1.10 Calculating Batch Means to estimate MCMC standard errors

Split the output into α batches of size β each, such that $S = \alpha * \beta$. Let $\bar{\psi}_i$ be the ith batch mean. If β is large enough, $\bar{\psi}_i$ are approximately independent, with $\bar{\psi} = \frac{1}{\alpha}\sum_{i=1}^{\alpha}\bar{\psi}_i$.

Thus, approximate the square of MCMC standard error via

$$MCMCSE^2 \approx \frac{1}{\alpha(\alpha-1)}\sum_{i=1}^{\alpha}(\bar{\psi}_i - \bar{\psi})^2$$

.

Thus we successfully computed MCMC standard errors without computing any sample correlations.

3 The Gibbs Sampler

The Gibbs Sampler is a special form of the MCMC algorithm.

Suppose $\Theta = (\theta_1, \theta_2, ..., \theta_d)$ where θ_i can be a vector.

Given $\Theta^0 = (\theta_1^0, ..., \theta_d^0)$ and $p(\theta_1, ..., \theta_d|y)$ for $t = 1, ...T$:

(a) $\theta_1^t \sim p(\theta_1|\theta_2^{t-1}, ..., \theta_d^{t-1}, y)$
(b) $\theta_1^t \sim p(\theta_2|\theta_1^t, ..., \theta_d^{t-1}, y)$
(c) Keep updating with general form $\theta_j^t \sim p(\theta_j|\theta_1^t, \theta_2^t, ..., \theta_{j-1}^t, \theta_{j+1}^{t-1}, ..., \theta_d^{t-1}, y)$

Note that if we could do $\theta_1 \sim p(\theta_1|y)$ or $\theta_2 \sim p(\theta_2|\theta_1, y)$ (and so on) we could do i.i.d sampling. The Gibbs Sampler is useful for problems in which the marginal distributions cannot be solved but the full conditional distributions can be.

3.1 Properties of the Gibbs Sampler

(a) $p(\Theta|y)$ is the stationary distribution if $\Theta^{t-1} \sim p(\Theta|y) \Rightarrow \Theta^t \sim p(\Theta|y)$
(b) Θ^{t-1} and Θ^t are not independent
(c) As $t \to \infty$, Θ^t converges in distribution to $p(\Theta|y)$ for all starting values Θ^0. Likewise, as $T \to \infty$, $\frac{1}{T}\sum_{t=1}^{T} h(\theta^t) \to E[h(\theta)|y] = \int h(\theta)p(\theta|y)d\theta$.

Definition 1. *The set of conditional distributions* $p(\theta_j|\theta_1, ...\theta_{j-1}, \theta_{j+1}, ..., \theta_d, y)$ *are the* ***full conditional distributions****.*

3.2 Gibbs Sampler Tricks

(a) If the plot shows a drift, your chain probably started in the tail of the target distribution. Throw away some of the early drawn samples.
(b) If the autocorrelation is poor, only use every 10th or so of the samples. We call this process "thinning".

4 MVN Inference

Inference about normal distributions can be extended to multivariate cases. But now, instead of a μ value we need a μ vector, $\mu_\circ$.

5 Bayesian Linear Regression

6 Hierarchical Normal Models with Informative Priors

7 Missing Data