Monte Carlo sampling

- Monte Carlo (MC) sampling is the predominant method of Bayesian inference because it can be used for high-dimensional models (i.e., with many parameters).

- The main idea is to approximate posterior summaries by drawing samples from the posterior distribution, and then using these samples to approximation posterior summaries of interest.

- For example, if $\theta^{(1)}, \ldots, \theta^{(S)}$ are samples from $p(\theta | Y)$, then the mean of the $S$ samples can be used to approximate the posterior mean.

- This only provides approximations of the posterior summaries of interest.

- But how to draw samples from some arbitrary distribution $p(\theta | y)$?
**Gibbs sampling**

- Gibbs sampling was proposed in the early 1990s (Geman and Geman, 1984; Gelfand and Smith, 1990) and fundamentally changed Bayesian computing.

- Gibbs sampling is attractive because it can sample from high-dimensional posteriors.

- The main idea is to break the problem of sampling from the high-dimensional joint distribution into a series of samples from low-dimensional conditional distributions.

- Updates can also be done in blocks (groups of parameters).

- Because the low-dimensional updates are done in a loop, samples are not independent.

- The dependence of the samples turns out to follow a Markov distribution, leading to the name Markov chain Monte Carlo (MCMC).
Gibbs sampling

• The algorithm begins by setting initial values for all parameters, $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_p^{(0)})$.

• Variables are then sampled one at a time from their full conditional distributions,

$$p(\theta_j | \theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p, y).$$

• Rather than a $p$-dimensional joint sample, we make $p$ 1-dimensional samples.

• The process is repeated until the required number of samples have been generated.

• Formally, the algorithm is:
Gibbs sampling

- Consider the normal model: \( Y_i \sim N(\mu, \sigma^2) \) with priors \( \mu \sim N(\mu_0, \sigma_0^2) \) and \( \sigma^2 \sim \text{InvGamma}(a, b) \).

- The \( p = 2 \) model parameters are \( \theta = (\mu, \sigma^2) \).

- We want samples from the bivariate posterior \( f(\mu, \sigma^2 | Y) \).

- **Initial values**: A reasonable choice is to set \( \mu \) and \( \sigma^2 \) to the sample mean and variance.

- The **full conditionals** are:

  \[ - \mu | \sigma^2, y \sim \]

  \[ - \sigma^2 | \mu, y \sim \]

---

Code is online at

Gibbs sampling

• Why does this work?
  • $\theta^{(0)}$ isn’t a sample from the posterior, it is an arbitrarily chosen initial value.
  • $\theta^{(1)}$ likely isn’t from the posterior either. Its distribution depends on $\theta^{(0)}$.
  • $\theta^{(2)}$ likely isn’t from the posterior either. Its distribution depends on $\theta^{(0)}$ and $\theta^{(1)}$.
  • **Theorem**: For any initial values, the chain will eventually converge to the posterior.
  • **Theorem**: If $\theta^{(s)}$ is a sample from the posterior, then $\theta^{(s+1)}$ is too.
  • Ideally, the chains will look like this:

• From this output we need to decide (1) when has it converged? and (2) when have we taken enough samples to approximate the posterior?

• Once we decide the chain has converged at iteration $t$, we discard the first $t$ samples as “burn-in” and use the remaining $S - t$ to approximate the posterior.

• For example, the posterior mean and variance are approximated as:
Convergence diagnostics

(1) When have the chains converged?

- The most common way to tell when the chains have converged is visual inspection:

- Convergence is much easier to diagnose if we start multiple chains (3-5) with very different initial values.

- Therefore, there is a trade-off at play when picking initial values:
Convergence diagnostics

(1) When have the chains converged?

- There are many many formal convergence diagnostics.
- The most widely-used is probably the Gelman-Rubin R statistic:
Convergence diagnostics

(2) When have we taken enough samples:

- Autocorrelation is a major factor in determining the number of samples needed.
- The lag-1 autocorrelation of the chain is

- The autocorrelation function is

- With all else being equal, independent chains provide far more information about the posterior than highly-correlated chains.
Convergence diagnostics

(2) When have we taken enough samples:

- The sample mean of the MCMC iterations is an approximation to the posterior mean.
- How much error in this estimate if samples are independent?
- How much error in this estimate if samples are dependent?
- These standard errors give a way to determine when to stop sampling:
- The effective sample size is also a good convergence diagnostic.
Recommendations
What to do if the chains don’t converge?