SAMPLING METHODS

Lecture Notes

Ahmet Ademoglu, *PhD*
Bogazici University
Institute of Biomedical Engineering

Monte Carlo Method

Find the volume of the sphere with radius r using *Acceptance-Rejection* Sampling.

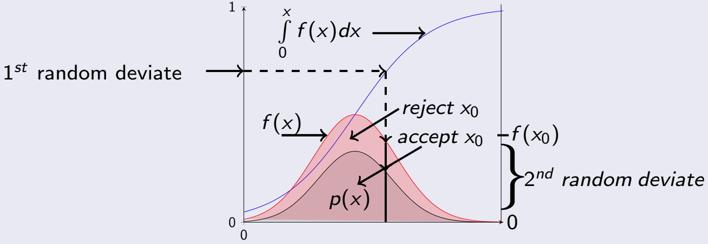
We encapsulate the volume of the sphere having a radius r, within a cube with edge $2r$.

1. Generate uniform random number triplets $\{x_i, y_i, z_i\}$ in $[0, 1]$ and convert them into $[-r, r]$ by using the transformation $X \rightarrow (X - 0.5)r$.
2. Accept those if $\sqrt{x_i^2 + y_i^2 + z_i^2} \leq r$.

If N_0 is the number of points falling within the sphere then

$$V_{sphere} \approx \frac{8r^3 N_0}{N}$$

Acceptance-Rejection Sampling



1. Generate a random number α from a distribution uniform in $[0, 1]$.
2. Determine $x_0 = F^{-1}(\alpha)$ where $F(x) = \int_0^x f(x)dx$.
3. Generate a second random variable β from a distribution uniform in $[0, f(x_0)]$.
4. Accept β if $\beta \leq p(x_0)$ or reject if otherwise.

Bayes Rule

Let $\mathbf{y}$ denote a $n \times 1$ vector of a system and let θ denote some unobserved parameter(s) for the model of the system, a joint model can be factorized

$$p(\mathbf{y}, \theta) = p(\mathbf{y}|\theta)p(\theta)$$

where $p(\mathbf{y}|\theta)$ is the sampling distribution and $p(\theta)$ is your prior. We are interested in conditioning on observable, $\theta|\mathbf{y}$ which follows,

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}, \theta)}{p(\mathbf{y})}$$

Three Quantities of Interest

1. *Prior Predictive* : The normalizing constant in Bayes Theorem $p(y) = \int p(\mathbf{y}, \theta) d\theta$,
   It represents the evidence for a particular model, defined by $p(\theta)$.
   It represents the probability of observing the data that was observed, before it was observed.
   It is also called *evidence* or *marginal likelihood*.
2. *Marginal Effects* : A subset of parameters in a multivariate model .
   Let $\theta = (\theta_1, \ldots, \theta_p)$ denote a p dimensional model, we can determine $p(\theta_i|\mathbf{y}) = \int p(\theta_i, \theta_{-i}|\mathbf{y}) d\theta_{-i}$
   $= \int p(\theta_i|\theta_{-i}, \mathbf{y}) p(\theta_{-i}|\mathbf{y}) d\theta_{-i}$ where $\theta_{-i} = \theta \backslash \theta_i$.

3. *Posterior Predictions*
   Let $\tilde{\mathbf{y}}$ denote some future unobserved response of the system.
   Then the posterior predictive $p(\tilde{\mathbf{y}}|\mathbf{y}) = \int p(\tilde{\mathbf{y}}|\mathbf{y})p(\theta|\mathbf{y})d\theta$
   follows,
   $$p(\tilde{\mathbf{y}}|\mathbf{y}) = \int p(\tilde{\mathbf{y}}|\theta,\mathbf{y})p(\theta|\mathbf{y})d\theta = \int p(\tilde{\mathbf{y}}|\theta)p(\theta|\mathbf{y})d\theta$$
   $\tilde{\mathbf{y}}$ and $\mathbf{y}$ are conditionally independent given θ; though clearly $p(\tilde{\mathbf{y}},\mathbf{y})$ are dependent.

All 3 quantities above are defined via probability statements on the unknown variable of interest.

Consider a normal linear regression,

$$\mathbf{y} = \mathbf{x}\beta + \epsilon$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$.
Alternatively,

$$\mathbf{y} \sim \mathcal{N}(\mathbf{y}|\mathbf{x}\beta, \sigma^2 \mathbf{I})$$

assuming that σ is known.
ML estimation of regression coefficients, β, given a data set, $(\mathbf{y}, \mathbf{x})$,

$$\beta = (\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T \mathbf{y}$$

Bayesian Estimation of β

$$p(\mathbf{y}, \beta | \mathbf{x}, \sigma^2) = p(\mathbf{y} | \mathbf{x}, \beta\sigma^2) p(\beta | \mathbf{x}, \sigma^2)$$

where $p(\beta) = \mathcal{N}(\beta | 0, v)$

$p(\beta|\mathbf{y}) \propto p(\mathbf{y}|\beta)p(\beta) \propto \sigma^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{x}\beta)^T(\mathbf{y}-\mathbf{x}\beta)} v^{-\frac{1}{2}} e^{-\beta^T v \beta}$

which can be written

$p(\beta|\mathbf{y}) = \mathcal{N}(\beta|\hat{\beta}, \hat{v})$

$\hat{\beta} = (\mathbf{x}^T\mathbf{x} + v^{-1})^{-1}\mathbf{x}^T\mathbf{y}$

$\hat{v} = (\mathbf{x}^T\mathbf{x} + v^{-1})^{-1}$

Estimating the predictive density

Using the new data, y_0, x_0,

$$\begin{aligned} p(y_0|x_0, \mathbf{y}) &= \int p(y_0|x_0, \beta, \mathbf{y}) p(\beta|\mathbf{y}) d\beta \\ &= \int \mathcal{N}(y_0|x_0\beta, \sigma^2)\mathcal{N}(\beta|\hat{\beta}, \hat{v}) d\beta \\ &= \mathcal{N}(y_0, |x_0\hat{\beta}, \sigma^2(1 + x_0\hat{v}x_0^T)) \end{aligned}$$

Prior: where does $p(\theta)$ come from?

1. Formally it represents our subjective beliefs, via a probability statement, about likely values of unobserved θ before we have observed $\mathbf{y}$
2. Practically, there are often standard and well used forms for the set $\{p(y|\theta), p(\theta)\}$
3. When the prior and the posterior are from the same family of distributions then the prior is termed conjugate which make these calculations easy.

Markov Chain Monte Carlo (MCMC) Method

MCMC is a sampling method that simulataneously solves inference of $\{p(\theta|\mathbf{y}), p(\theta_i|\mathbf{y}), p(\tilde{\mathbf{y}}|\mathbf{y})\}$ *i.e.* prior predictive, marginal effects and posterior predictions, respectively.

It requires the evaluation of

$$p(\mathbf{y}, \theta) \propto p(\mathbf{y}|\theta)p(\theta)$$

up to a proportionality, for $\theta \in \Theta$.

MCMC simulation approximates the true posterior density, $p(\theta|\mathbf{y})$, using a bag of samples drawn from the density.
MCMC is especially useful

1. when we have non-conjugacy;
2. when the combination of prior and sampling distribution do not lead to a standard form for the posterior, $p(\theta|\mathbf{y})$.

Normal Linear Regression

Choosing $p(\beta) = \mathcal{N}(0, v)$, the posterior is,

$$
\begin{aligned}
p(\beta|\mathbf{y}) &= \mathcal{N}(\beta|\hat{\beta}, \hat{v}) \\
\hat{\beta} &= \hat{v}\mathbf{x}^T\mathbf{y} \\
\hat{v} &= (\mathbf{x}^T\mathbf{x} + v^{-1})^{-1}
\end{aligned}
$$

MCMC approximates this distribution with M samples drawn from the posterior,

$$\{\beta^{(1)}, \ldots, \beta^{(M)}\} \sim \mathcal{N}(\hat{\beta}, \hat{v})$$

Simulation based inference

If $p(\theta|\mathbf{y})$ is not quantifiable as a standard distribution;

1. We draw samples, $\theta^{(1)}, \ldots, \theta^{(M)}$, from $p(\theta|\mathbf{y})$, so that,

$$\theta^{i} \sim p(\theta|\mathbf{y})$$

2. Then inferential quantities of interest can be estimated using the bag of samples, $\{\theta^{(i)}|_{i=1}^{M}\}$, as a proxy for $p(\theta|\mathbf{y})$.

Suppose we are interested in $Pr(\theta < a|\mathbf{y})$:

$$Pr(\theta < a|\mathbf{y}) \approx \frac{1}{M}\sum_{i=1}^{M} I(\theta^{(i)} < a)$$

where $I(\cdot)$ is the logical indicator function. More generaly, for a set $A \in \Theta$

$$Pr(\theta \in A|y) \approx \frac{1}{M}\sum_{i=1}^{M} I(\theta^{(i)} < a)$$

Prediction

Suppose we are interested in $p(\tilde{\mathbf{y}}|\mathbf{y})$, for some future $\tilde{\mathbf{y}}$. Then,

$$\begin{aligned} p(\tilde{\mathbf{y}}|\mathbf{y}) &\approx \frac{1}{M}\sum_{i=1}^{M} p(\tilde{\mathbf{y}}|\theta^{(i)}, \mathbf{y}) \\ &\approx \frac{1}{M}\sum_{i=1}^{M} p(\tilde{\mathbf{y}}|\theta^{(i)}) \end{aligned}$$

Inference of marginal effects: Suppose, θ is multivariate and we are interested in $\theta_j \in \theta$ (*i.e.* a particular parameter in a normal linear regression model).

$$F_{\theta_j} \approx \frac{1}{M}\sum_{i=1}^{M} I(\theta^{(i)_j} \leq a)$$

where $F(\cdot)$ denotes the distribution function; More generaly for any set $A_j \in \Theta_j$, the lower dimensional parameter space,

$$Pr(\theta_j \in A_j|\mathbf{y}) \approx \frac{1}{M}\sum_{i=1}^{M} I(\theta^{(i)_j} \leq a)$$

MCMC Algorithms

Metropolis-Hastings Algorithm

1. Let the current state of the chain be $\theta^{(i)}$,
2. Consider a conditional density $q(\tilde{\theta}|theta^{(i)})$, defined on $\theta \in \Theta$,
3. We call $q(\cdot|\theta^{(i)})$ the proposal density and use it to update the chain.

Metropolis-Hastings Algorithm

$\theta^{(0)} \longleftarrow x$

For i=0 to M

Draw $\tilde{\theta} \sim q(\tilde{\theta}|\theta^{(i)})$

Set $\theta^{(i+1)} \longleftarrow$ with probability $\alpha(\theta^{(i)}, \tilde{\theta})$

Set $\theta^{(i+1)} \longleftarrow \theta^{(i)}$ where $\alpha(a, b) = \min\left\{1, \frac{p(b|\mathbf{y})q(a|b)}{p(a|\mathbf{y})q(b|a)}\right\}$

- To accept with probability $\frac{p(b|y)q(a|b)}{p(a|y)q(b|a)}$,

    First, draw a uniform random variable, say U, on $[0, 1]$.
    IF $U < \alpha(\theta^{(i)}, \tilde{\theta})$;
    THEN accept $\tilde{\theta}$;
    ELSE reject and chain stays in $\theta^{(i)}$
- The ratio of densities means that the normalising constant $p(y) = \int p(\mathbf{y}|\theta)p(\theta)d\theta$ cancels, top and bottom.
  Hence, we can use MCMC when this is unknown (as is often the case).

Tips for choosing the proposal density $q(a, b)$.

- In the special case of a symmetric proposal density (Hastings algorithm), $q(a|b) = q(b|a)$.
  For example, if $q(a|b) = \mathcal{N}(a|b, 1)$, then the ratio reduces to that of the probabilities $\alpha(a, b) = \min\{1, \frac{p(b|\mathbf{y})}{p(a|\mathbf{y})}\}$
- The proposal density, $q(a|b)$, is user defined.
  It is more of an art than a science.
- $q(a|b)$ should be easy to simulate and evaluate.
  A popular choice when θ is real valued is to take $q(a|b) = b + \mathcal{N}(a|0, V)$ where V is user specified.
  That is, a normal density centred at the current state b.
- It is usual to centre the proposal density around the current state and make local moves.

Gibbs Sampler

- As alternative method for MCMC, suppose that the multidimensional θ can be partitioned into p subvectors, $\theta = \{\theta_1, \ldots, \theta_p\}$, such that the conditional distribution,

$$p(\theta_j | \theta_{-j}, \mathbf{y})$$

  is easy to sample from; where $\theta_{-j} = \theta \backslash \theta_j$
- Iterating over the p subvectors and updating each subvector in turn using $p(\theta_j | \theta_{-j}, \mathbf{y})$ leads to a valid MCMC scheme known as the Gibbs Sampler.

Gibbs Sampler

$\theta^{(0)} \longleftarrow x$

For i=0 to M

Set $\tilde{\theta} \longleftarrow \theta^{(i)}$

For j=1 to p

Draw $X \sim p(\theta_j|\tilde{\theta}_{-j}, \mathbf{y})$

Set $\tilde{\theta}_j \longleftarrow X$

End

Set $\theta^{i+1} \longleftarrow \tilde{\theta}$

- The Gibbs Sampler is a special case of the Metropolis-Hastings algorithm using the ordered sub-updates, $q(\cdot) = p(\theta_j|\theta_{-j}, \mathbf{y})$
- All proposed updates are accepted (there is no accept-reject step),
- θ_j may be multidimensional or univariate,
- Often, $p(\theta_j|\theta_{-j}, \mathbf{y})$ will have standard form even if $p(\theta|\mathbf{y})$ does not.

Considering the normal linear regression model $\mathbf{y} = \mathbf{x}\beta + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$. Alternately, $y \sim \mathcal{N}(\mathbf{y}|\mathbf{x}\beta, \sigma^2\mathbf{I})$ assume that σ is unknown.

Construct a joint model for the data and unknown parameters,

$$\begin{aligned} p(\mathbf{y}, \beta, \sigma^2|\mathbf{x}) &= p(\mathbf{y}|\mathbf{x}, \beta, \sigma^2)p(\beta, \sigma^2|\mathbf{x}) \\ &= \mathcal{N}(\mathbf{y}|\mathbf{x}\beta, \sigma^2\mathbf{I})p(\beta)p(\sigma^2) \end{aligned}$$

where we have assumed that the joint prior $p(\beta, \sigma^2)$ is independent.
Suppose we take,

$$p(\beta) = \mathcal{N}(\beta|0, v) \text{ and } p(\sigma^2) = IG(\sigma^2|a, b)$$

where $IG(\mathbf{x}|a, b) \propto x^{-\frac{(a-2)}{2}} e^{-\frac{b}{2x}}$ (Inverse-Gamma Distribution).

The joint posterior is not a standard distribution *i.e.*
$p(\beta, \sigma^2|\mathbf{y}) \propto p(\mathbf{y}|\beta)p(\beta)p(\sigma^2)$
$= \sigma^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{x}\beta)^T(\mathbf{y}-\mathbf{x}\beta)} |v|^{-\frac{1}{2}} e^{-\beta^T v \beta} (\sigma^2)^{-\frac{a-2}{2}} e^{-\frac{b}{2\sigma^2}}$

However, the conditionals are known;

$$p(\beta|\mathbf{y}, \sigma^2) = \mathcal{N}(\beta|\hat{\beta}, \hat{v})$$
$$\hat{\beta} = (\sigma^{-2}\mathbf{x}^T\mathbf{x} + v^{-1})^{-1}\sigma^{-2}\mathbf{x}^T\mathbf{y}$$
$$\hat{v} = (\sigma^{-2}\mathbf{x}^T\mathbf{x} + v^{-1})^{-1}$$

and $$p(\sigma^2|\beta, \mathbf{y}) = IG(\sigma^2|a + n, b + (\mathbf{y} - \mathbf{x}\beta)^T(\mathbf{y} - \mathbf{x}\beta))$$

Hence, we can use Gibbs Sampler.

Gibbs Sampler for normal linear regression.

$(\beta, \sigma^2)^{(0)} \longleftarrow x$

For i=0 to M

Set $(\tilde{\beta}, \tilde{\sigma}^2) \longleftarrow (\beta, \sigma^2)^{(i)}$

Draw $\tilde{\beta}|\sigma^2 \sim \mathcal{N}(\beta|\hat{\beta}, \hat{V})$

Draw $\tilde{\sigma}^2|\tilde{\beta} \sim IG(\sigma^2|a + n, b + (\mathbf{y} - \mathbf{x}\beta)^T(\mathbf{y} - \mathbf{x}\beta))$

Set $(\beta, \sigma^2)^{(i)} \longleftarrow (\tilde{\beta}, \tilde{\sigma}^2)$

End

Considering the hierarchical normal linear regression model $\mathbf{y} = \mathbf{x}\beta + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$ we assume that both σ and prior variance v of $p(\beta)$ are unknown.

Hierarchy determines the dependence among the variables;

$$\mathbf{y} \sim \mathcal{N}(\mathbf{y}|\mathbf{x}\beta, \sigma^2 \mathbf{I})$$
$$\beta \sim \mathcal{N}(\beta|0, v\mathbf{I})$$
$$\sigma^2 \sim IG(\sigma^2|a, b)$$
$$v \sim IG(v|c, d)$$

Then the joint posterior is not a standard distribution!

$$p(\beta, \sigma^2, v|\mathbf{y}) \propto p(\mathbf{y}|\beta, \sigma^2)p(\beta|v)p(v)p(\sigma^2)$$
$$\propto \sigma^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{x}\beta)^T(\mathbf{y}-\mathbf{x}\beta)} |v|^{-\frac{1}{2}} e^{-\beta^T v \beta} (\sigma^2)^{-\frac{a-2}{2}} e^{-\frac{b}{2\sigma^2}} v^{-\frac{c-2}{2}} e^{-\frac{d}{2v}}$$

However, the conditionals are known and

$$p(\beta|\mathbf{y}, \sigma^2, v) = \mathcal{N}(\beta|\hat{\beta}, \hat{v})$$
$$\hat{\beta} = (\sigma^{-1}\mathbf{x}^T\mathbf{x} + v^{-1})^{-1}\sigma^{-2}\mathbf{x}^T\mathbf{y}$$
$$\hat{v} = (\sigma^{-2}\mathbf{x}^T\mathbf{x} + v^{-1})^{-1}$$

$$p(\sigma^2|\beta, \mathbf{y}) = IG(\sigma^2|a + n, b + (\mathbf{y} - \mathbf{x}\beta)^T(\mathbf{y} - \mathbf{x}\beta))$$

and

$$p(v|\beta) = IG(v|a + p, b + \beta^T\beta)$$

where p : lenght of β vector.

Gibbs Sampler for hierarchical normal linear regression.

$(\beta, \sigma^2, v)^{(0)} \longleftarrow x$

For i=0 to M

Set $(\tilde{\beta}, \tilde{\sigma}^2, \tilde{v}) \longleftarrow (\beta, \sigma^2, v)^{(i)}$

Draw $\tilde{\beta} | \sigma^2, v \sim \mathcal{N}(\beta | \hat{\beta}, \hat{V})$

Draw $\tilde{\sigma}^2 | \tilde{\beta} \sim IG(\sigma^2 | a + n, b + (\mathbf{y} - \mathbf{x}\beta)^T (\mathbf{y} - \mathbf{x}\beta))$

Draw $\tilde{v} | \tilde{\beta} \sim IG(v | c + p, d + \beta^T \beta))$

Set $(\beta, \sigma^2, v)^{(i)} \longleftarrow (\tilde{\beta}, \tilde{\sigma}^2, \tilde{v})$

End