ST4231 Computer Intensive

Lingjie, December 14, 2021

Fundamental Theorem of Simulation

If X is a random variable with pdf $f(x)$, then simulating X is equivalent to simulating a pair of random variable (X, U) jointly from

$$(X, U) \sim unif\,\{(x, u) : 0 < u < f(x)\}$$

Explanation

Let $S = \{(x, u) : 0 < u < f(x)\}$

want: generate $x \sim F(x)$ (area)

method: (1) generate $x \in D_F$, domain of distribution

(2) generate $u \sim unif(0, f(x))$

(3) resultant area is $F(x)$

Generating directly from S may be difficult, use rejection sampling to generate from a proxy distribution instead.

Misc

Finding distribution following an algorithm

The goal is to deduce the distribution of a random variable X, following a given algorithm.

1. Determine individual distributions in the algo

   e.g. $y_1 \sim exp(1), v \sim unif(0, 1)$

2. Determine marginal or joint distribution in final step

   [a] independent joint probability

   $$f(x, y) = f(x)f(y)$$

   [b] constrained joint probability

   $$\tilde{f}(x, y) = \frac{1}{c} f(x, y), x < y$$

   [c] marginal distribution

   $$f(y) = \int_D^{x<y} \tilde{f}(x, y)dx$$

   Remember to find normalizing constant C

3. Determine distribution of final RV X

   [a] if $X = Y \Rightarrow f_X(x) = f_Y(y)$

   [b] if $X = \frac{1}{2}Y + \frac{1}{2}(-Y)$

   $$\begin{aligned}\Rightarrow P(X \le x) &= \frac{1}{2}P(Y \le x) + \frac{1}{2}P(Y \ge -x) \\ &= \begin{cases} \frac{1}{2}(0) + \frac{1}{2}P(Y \ge -x), & x < 0 \\ \frac{1}{2}P(Y \le x) + \frac{1}{2}(1), & x \ge 0 \end{cases}\end{aligned}$$

   deduce $f_X(x)$ based on $P(X = x) = \frac{d}{dx}P(X \le x)$

Determine mixture distribution

Given a mixture distribution

$$f(x) \propto f_1(x) + f_2(x)$$

Trick: $f_1(x), f_2(x)$ must be pdf $\Rightarrow \int_D f_i(x) = 1$

e.g.

$$\begin{aligned} f(x) &\propto cf_1(x) + \frac{c}{2}2f_2(x) \\ &\Rightarrow c + \frac{c}{2} = 1 \\ &\Rightarrow c = \frac{2}{3} \end{aligned}$$

Beta ordered statistics

Given $X_1, X_2, \cdots, X_n \sim unif(0, 1)$

the ordered statistics $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$

has property

$$X_{(k)} \sim Beta(k, n + 1 - k)$$

Monte Carlo Methods

Monte Carlo Integration

Key: with $\mathbf{U} \sim unif(a, b)$, identify

- $g(\mathbf{U})$
- $\theta = E[g(\mathbf{U})]$

LLN Condition must be met: $E[g(X)] < \infty$

Procedure:

1. Generate rectangle enclosing function: $\mathbf{U} \sim (a, b)$
2. Calculate area of interest: $g(\mathbf{U})$
3. Calculate percentage of sample within area: θ
4. Multiply area of rectangle

Example: finding $ln(3) = \int_1^3 (1/t)dt$

1. Generate $U_1 \sim unif(1, 3), U_2 \sim unif(0, 1)$
2. $g(\mathbf{U}) = I\left(U_1 \le (1/U_2)\right)$
3. $\hat{\theta} = (1/M)\sum_{i \in M} g(\mathbf{U_i})$
4. $\hat{ln(3)} = 2 \times \hat{\theta}$

where M is number of sample generated and $\left\{(x, y) : 1 < x < 3, 0 < y < \frac{1}{x}\right\}$

Code:

```r
M <- 10^6
U1 <- runif(M, min=1, max=3)
U2 <- runif(M, min=0, max=1)
g <- U1 <= (1/U2)
theta <- mean(g)
ln3.est <- 2 * theta
```

Generating Random Variable

Inversion Method

Limitation:

- Discrete: time-consuming (but default for this mod)
- Continuous: explicit and invertible cdf F

Discrete Random Number Generators

Let

$$P(X = x_j) = p_j, j = 0, 1, \cdots, \sum_j p_j = 1$$

Sequential Inversion

1. generate $U \sim unif(0, 1)$
2. set $X = 0, S = p_0$
3. while $U > S$:

   [3.1] $X = X + 1$

   [3.2] $S = S + p_x$

4. return X

Continuous Random Variable

Assume we know an invertible cdf

$$F(x) = \int_D f(x)$$

Inverse Transform Algorithm

1. generate $U \sim unif(0, 1)$
2. return $X = F^{-1}(U)$

Rejection Sampling

Theorem

If X is generated via rejection sampling method, X has pdf $f(X)$

Algorithm

Based on Fundamental Theorem of Simulation
Let

$$g := \text{proposal distribution}$$
$$f := \text{target distribution}$$
$$M := \text{scaling parameter}, M > 1$$

Rejection sampling

1. generate $Y \sim g$
2. generate $U \sim unif(0,1)$
3. if $U \leq \frac{f(Y)}{Mg(Y)}$:
   set $X = Y$, exit
4. else: return to step 1

Efficiency = finding optimal M

Optimal M = smallest M possible

$$M^* = \sup_{x\in\mathbb{R}^d} \frac{f(x)}{g(x)}$$
$$\Leftrightarrow \sup_{x\in\mathbb{R}^d} log(f(x)) - log(g(x))$$

where

$$\sup = \max \text{ but allow } \infty$$
$$P\{(Y,U) \text{ is accepted}\} = \frac{1}{M}$$
$$c := E(N) = M \sim geometric(1/M)$$

Condition

Must check for the following conditions

- Domain of $g(x)$ must include domain of $f(x)$
- Tail of $g(x)$ must be heavier than $f(x)$, check

$$\lim_{x\to|\infty|} \frac{f(x)}{g(x)} < \infty$$

[edge case 1] $x \to |\infty|$
[edge case 2] $f(x) \to \infty$
[checking case] need to check if parameter for $g(x)$ will not violate this condition

Unknown Normalizing Constant

Suppose $f(x) = c\tilde{f(x)}$, where $\tilde{f(x)}$ is known and c is unknown.
We can find $\tilde{M}$ satisfies that $\tilde{f(x)} \leq \tilde{M}g(x), \ \forall x$
Useful to ignore the normalising constant of $f(x)$, even normalizing constant for $g(x)$ can be ignored.

Polar Method for Bivariate Normal

Box-Muller Algorithm v1

1. Generate $U_1 \sim Unif(0,1)$, $U_2 \sim Unif(0,1)$
2. Set $R = \sqrt{-2log(U_1)}$, $\theta = 2\pi U_2$
3. Set

$$X = \sqrt{-2log(U_1)}cos(2\pi U_2)$$
$$Y = \sqrt{-2log(U_1}sin(2\pi U_2)$$

Box-Muller Algorithm v2

1. Generate $U_1 \sim Unif(0,1)$, $U_2 \sim Unif(0,1)$
2. Set $V_1 = 2U_1 - 1$, $V_2 = 2U_2 - 1$, $S = V_1^2 + V_2^2$
3. If $S > 1$ return to step 1 (rejection sampling)
4. Return the independent unit normals

$$X = \sqrt{-2log(S)/S}V_1$$
$$Y = \sqrt{-2log(S)/S}V_2$$

General Multivariate Normal

$d-$dimensional normal with mean μ, covariance matrix $\sum$

1. Generate

$$Z = \begin{pmatrix} Z_1 \\ \vdots \\ Z_d \end{pmatrix}, Z_1, \cdots, Z_d \text{ i.i.d } N(0,1)$$

2. Set

$$X = LZ + \mu$$

where L satisfies $LL^T = \sum$
usually L is taken as the Cholesky factor, a lower triangular matrix with positive diagonal entries

Variance Reduction Techniques

Goal: estimate

$$\theta = E[\varphi(x)] = \int_S \varphi(x)f(x)dx$$

$S :=$ support
$f(x) :=$ pdf
Explain the manner of uncertainty/CI: smaller asymptotic variance.

Simple Sampling

$X_i \sim f(x)$

$$\hat{\theta}_{SS} = \frac{1}{n}\sum_{i=1}^{n} \varphi(X_i)$$

By SLLN, $\hat{\theta} \to \theta$ as $n \to \infty$ with probability 1
Potential Issues

- Variance $\sigma^2 = Var(\varphi(X))$ can be infinity
- We can find an estimator with smaller variance than $Var(\hat{\theta}) = \sigma^2/n$
- Might not be possible to sample from $f(x)$

Variance

Asymptotic variance

$$Var(\varphi(X)) = \left(\int_S \varphi^2(x)f(x)dx - \theta^2\right)$$
$$= \sigma^2$$

Exact/Approximate (with CLT) variance

$$Var(\hat{\theta}) = \frac{1}{n}Var(\varphi(X))$$
$$= \frac{1}{n}\sigma^2$$

Estimated asymptotic variance

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \varphi^2(X_i) - \hat{\theta}^2$$

Asymptotic Confidence Interval

asymptotic 95% confidence interval for θ

$$\hat{\theta} \pm 1.96 \frac{\hat{\sigma}}{\sqrt{n}}$$

Importance Sampling

Instead, we sample from the important part of the sample space and re-weight
$Y_i \sim g(x)$

$$\hat{\theta}_{IS} = \frac{1}{n} \sum_{i=1}^{n} \varphi(Y_i) w(Y_i)$$
$$w(Y_i) = \frac{f(Y_i)}{g(Y_i)}$$

Note: $\hat{\theta}_{IS}$ is unbiased
Arise from

$$\begin{aligned}\theta = E_f(\varphi(X)) &= \int_S \varphi(x) f(x) dx \\ &= \int_S \frac{\varphi(x) f(x)}{g(x)} g(x) dx \\ &= E_g[\varphi(Y) w(Y)]\end{aligned}$$

Importance Sampling Algorithm

1. Draw $X_1, \cdots, X_n$ from proposal density g
2. Calculate importance weight $w(X_i) = f(X_i)/g(X_i)$
3. Approximate θ with $\hat{\theta}_{IS}$

Variance

Asymptotic variance

$$\begin{aligned}\sigma^2 &= Var(\varphi(Y) w(Y)) \\ &= \int_S \frac{\varphi^2(y) f^2(y)}{g(y)} dy - \theta^2\end{aligned}$$

Estimated asymptotic variance

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \varphi^2(Y_i) w^2(Y_i) - \hat{\theta}_{IS}^2$$

Calculation for Exact variance and confidence interval using σ^2 is same as Simple Sampling

Optimal proposal density g

Optimal choice of $g(x)$

$$g(x) \propto |\varphi(x)| f(x)$$

In general, choose $g(x)$ with heavier tail than $f(x)$

If $g(x)$ not chosen properly, $\hat{\theta}_{IS}$ may have larger variance than $\hat{\theta}_{SS}$

Condition

1. Able to sample from $g(x)$
2. Finite variance $Var(\hat{\theta}_{IS}) < \infty$

   Sufficient condition: $\int_S \varphi^2(x) f^2(x)/g(x) < \infty$

Checking finite variance

$$\int_1^{+\infty} \frac{1}{x^p} dx = \begin{cases} +\infty, & p \leq 1 \\ < +\infty, & p > 1 \end{cases}$$
$$\int_0^{1} \frac{1}{x^p} dx = \begin{cases} < +\infty, & p < 1 \\ +\infty, & p \geq 1 \end{cases}$$

If $g(x)$ is different trend from $f(x)$ then proposal is inappropriate

We can say: as $x \to 0+$, function $\frac{exp(2x)}{2x}$ behaves similarly to $\frac{1}{2x}$. However, $\int_0^{\epsilon} \frac{1}{2x} dx = +\infty$ for any small $\epsilon > 0$. Therefore, infinite variance.

Rare events

When relative s.d. is large, simple sampling is inefficient

$$\begin{aligned}\text{relative s.d.} = \frac{\text{exact s.d.}}{p_*} &= \frac{\sqrt{p_*(1-p_*)/n}}{p_*} \\ &= \frac{1}{\sqrt{np_*}} \\ \Rightarrow n &= \frac{1}{\text{relative var} \times p_*}\end{aligned}$$

where p_* is the probability of interest (e.g. $P(X > 4), X \sim N(0,1)$)

Self-Normalizing Importance Sampling

When $f(x), g(x)$ is only known up to a normalizing constants $Z_f > 0, Z_g > 0$

$$f(x) = \frac{1}{Z_f} \tilde{f}(x),\ g(x) = \frac{1}{Z_g} \tilde{g}(x),\ \tilde{w}(x) = \frac{\tilde{f}(x)}{\tilde{g}(x)}$$

With the generalized weights $\tilde{w}(x)$, we have self-normalized importance sampling estimator

$$\begin{aligned}\hat{\theta}_{SIS} &= \frac{\sum_{i=1}^n \varphi(X_i) \tilde{w}(X_i)}{\sum_{i=1}^n \tilde{w}(X_i)} \\ &= \frac{\hat{\theta}_{IS}}{\sum_{i=1}^n \tilde{w}(X_i)}\end{aligned}$$

Bias

$\hat{\theta}_{SIS}$ is bias. But bias and fluctuation decreases as sample increase

$$\text{bias}(\hat{\theta}_{SIS}) = \mathcal{O}(1/n),\ \text{fluctuation}(\hat{\theta}_{SIS}) = \mathcal{O}(1/\sqrt{n})$$

Variance

Asymptotic variance

$$\begin{aligned}\sigma_{SIS}^2 &= E_g\left[w^2(X)[\varphi(X) - \theta]^2\right] \\ w(x) &= \frac{f(x)}{g(x)} \\ &= \frac{Z_g}{Z_f} \tilde{w}(x)\end{aligned}$$

Estimated exact variance (note: do not divide by n again)

$$\frac{\hat{\sigma}_{SIS}^2}{n} = \frac{\sum_{i=1}^n \left\{\tilde{w}^2(X_i)\left[\varphi(X_i) - \hat{\theta}_{SIS}\right]^2\right\}}{\left(\sum_{i=1}^n \tilde{w}(X_i)\right)^2}$$

95% Confidence interval

$$\hat{\theta}_{SIS} \pm 1.96 \sqrt{\frac{\sum_{i=1}^n \left\{\tilde{w}^2(X_i)\left[\varphi(X_i) - \hat{\theta}_{SIS}\right]^2\right\}}{\left(\sum_{i=1}^n \tilde{w}(X_i)\right)^2}}$$

Note:

- Usually $\sigma_{SIS}^2 > \sigma_{IS}^2$
- σ_{SIS}^2 is computable if and only if Z_f, Z_g is known
- $Var(\hat{\theta}_{SIS})$ is unknown (hard to find var of ratio of 2 RV)
- Estimated $Var(\hat{\theta}_{SIS}) = n\times$ estimated exact variance

Control Variates Method

Widely used in Bayesian statistics

Want : reduce $Var(\hat{\theta})$, where $\hat{\theta}$ estimates $\theta = E_f[\varphi(X)]$

Main idea : use control variate $\hat{h}$ that is correlated with $\hat{\theta}$

Assumption: supposed we know all of following

1. an unbiased estimator $\hat{h}$ of $E_f[h(X)]$
2. $E_f[h(X)]$ and $Var(\hat{h})$
3. the value or sign of $Cov(\hat{\theta}, \hat{h})$

Construction

$$\begin{aligned}
\tilde{\theta} &= \hat{\theta} + \beta\{\hat{h} - E_f[h(X)]\} \\
E_f[\tilde{\theta}] &= E_f[\hat{\theta}] \\
Var(\tilde{\theta}) &= Var(\hat{\theta}) + \beta^2 Var(\hat{h}) + 2\beta Cov(\hat{\theta}, \hat{h}) \\
\arg_\beta \min Var(\tilde{\theta}) &= -\frac{Cov(\hat{\theta}, \hat{h})}{Var(\hat{h})} = \beta^* \\
\Rightarrow Var(\tilde{\theta} | \beta = \beta^*) &= (1-\rho^2)Var(\hat{\theta}), \rho = Cor(\hat{\theta}, \hat{h}) \\
&< Var(\hat{\theta}) \text{ if } \rho \neq 0
\end{aligned}$$

Expectation of Indicator function

$$\begin{aligned}
I_A I_B &= \begin{cases} 1, A \cap B \\ 0, \text{otherwise} \end{cases} \\
&= I_{A \cap B} \\
I_{A \cup B} &= 1 - I_{A^C} I_{B^C}
\end{aligned}$$

Therefore,

$$\begin{aligned}
&Cov[I(X_i > a), I(X_i > 0)] \\
&= E[I(X_i > a) \cdot I(X_i > 0)] - E[I(X_i > a)]E[I(X_i > 0)] \\
&\because E[I(X_i > a, X_i > 0)] = E[I(X_i > a)] \\
&\therefore Cov[I(X_i > a), I(X_i > 0)] = E\left[I(X_i > a)\right]\left[1 - P(X_i > 0)\right]
\end{aligned}$$

Estimating β^*

Hard to obtain β^* in practice. Use linear regression instead.

$$\hat{\theta} = \alpha + \beta \hat{h}$$

with sample $\hat{\theta}, \hat{h}$

Antithetic Variates Method

$$\hat{I}_{SS} = \frac{1}{2n}\sum_{i=1}^{2n} h(U_i)$$

$$\hat{I}_{An} = \frac{1}{2n}\sum_{i=1}^{n} (h(U_i) + h(1-U_i))$$

Construction

If X, X' has same distribution (but not independent), then

$$\begin{aligned}
&2Cov(X, X') \\
&= E\{[g(U_1) - g(U_2)][g(1-U_1) - g(1-U_2)]\} \leq 0
\end{aligned}$$

Where X, X' is generated with $g(U)$ and $g(1-U)$ respectively.

$$Var(\frac{X+X'}{2}) \leq \frac{1}{2}Var(X)$$

Supporting facts

Constructs the Antithetic Variates Method

1. $X = F^{-1}(U) = h(U)$, $X' = F^{-1}(1-U) = h(1-U)$ has same distribution F

   $U \sim Unif(0,1)$, $F^{-1}(U)$ is quantile function, X generated from inversion method
2. If $g(\cdot)$ is monotone function (either increasing/decreasing), then

   $$[g(u_1) - g(u_2)][g(1-u_1) - g(1-u_2)] \leq 0$$

   for any $u_1, u_2 \in [0,1]$

Calculate An Var

$$\begin{aligned}
Var(\hat{I}_{SS}) &= \frac{1}{2n}Var[h(U)] \\
Var(\hat{I}_{An}) &= \frac{1}{4n}Var[h(U) + h(1-U)] \\
&= \frac{1}{2n}\left\{Var\left[h(U)\right] + Cov\left[h(U), h(1-U)\right]\right\}
\end{aligned}$$

Note: $Cov\left[h(U), h(1-U)\right]$

$$\begin{aligned}
&= E\left[h(U) \cdot h(1-U)\right] - E(h(U))E(h(1-U)) \\
&= E\left[h(U) \cdot h(1-U)\right] - E(h(U))^2 \\
&= E\left[h(U) \cdot h(1-U)\right] - \hat{I}^2
\end{aligned}$$

Since $E(h(U)) = E(h(1-U))$

Also, to calculate the empirical var, we need $n \times M$ samples.

$n :=$ num of sample, $M :=$ num of trials

An Var Example

Estimate $\int_0^1 x^2 dx, X \sim Unif(0,1)$

$$\hat{I}_{SS} = \frac{1}{2n}\sum_{i=1}^{2} n(U_i^2)$$

$$\hat{I}_{AN} = \frac{1}{2n}\sum_{i=1}^{n}(U_i^2 + (1-U_i)^2)$$

Expectation-Maximization (EM)

EM algorithm is used to find the MLE for a particular class of models, with unobserved latent variables

Key points

- iterative method
- finds maximum likelihood estimate of parameters in statistical models
- models contain either missing data or unobserved latent variables

Required knowledge

- Convex function

  $$H_f \geq 0$$

  Positive semi-definite hessian matrix, or non-negative second derivatives
- Jesen Inequality

  Let f be a convex function and X be a random variable

  $$f(E[X]) \leq E[f(X)]$$

  $$\Rightarrow \log\left(\int \varphi(x)q(x)dx\right) \geq \int \log[\varphi(x)]q(x)dx$$
- Maximum Likelihood Estimation

  with data D and parameter θ

  $$\max_\theta L(D;\theta) = \prod_{i=1}^{n} f(D|\theta)$$

  $$\hat{\theta}_{ML} = \arg\max_{\theta \in \Theta} L(D;\theta)$$

  Issue: MLE exist but no closed-form expression

Latent Variable Model

Goal: Compute MLE $\hat{\theta}_{ML}$ (parameter) from Y (data)
Trick: use Latent (unobserved variables/hidden state) Z
Original problem: $\theta \to Y$
Hierarchical model: $\theta \to Z \to Y$

EM algorithm

Steps:

1. Initialize θ_0
2. E-Step: in the k th iteration
   given $\theta^{(k)}$, calculate $\alpha_i^{(k,j)}, i \in [1,n], j \in Z$
3. M-Step: update
   $\theta^{(k+1)}$ with $\alpha_i^{(k,j)}$
4. Iterate between E-step and M-step until convergence
   $|\theta^{(k+1)} - \theta^{(k)}| < \epsilon$

Expectation (E-step)

Given $\theta^{(k)}$, calculate

$$\begin{aligned} Q(\theta|\theta^{(k)}) &= \sum_{i=1}^{n} \int_{z_i \in Z} \{\log p(y_i, z_i|\theta\} p(z_i|y_i, \theta^{(k)}) dz_i \\ &= E_Z[\ell^c(Y, Z; \theta)|Y, \theta^{(k)}] \\ &= E_Z[\log P(Y, Z|\theta)|Y, \theta^{(k)}] \end{aligned}$$

$\ell^c :=$ complete log-likelihood

$$\ell^c(Y, Z; \theta) = \log p(Y, Z|\theta) = \sum_{i=1}^{n} \log p(y_i, z_i|\theta)$$

Maximization (M-step)

Calculate $\theta^{(k+1)}$

$$\theta^{(k+1)} = \arg\max_{\theta \in \Theta} Q(\theta|\theta^{(k)})$$

Example: Mixture of Normals

Problem setup

Model:

$$\begin{aligned} p(y|\theta) = p \cdot \frac{1}{\sqrt{2\pi\sigma_1^2}} exp\left(-\frac{(y-\mu_1)^2}{2\sigma_1^2}\right) \\ + (1-p) \cdot \frac{1}{\sqrt{2\pi\sigma_2^2}} exp\left(-\frac{(y-\mu_2)^2}{2\sigma_2^2}\right) \end{aligned}$$

Parameter:

$$\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, p)$$

Goal: Find the MLE of θ

Finding complete log-likelihood function

Define $Z = \{1, 2\}$, if data belong to Normal 1 or 2
$\ell^c(Y, Z; \theta) = \log p(Y, Z|\theta)$

$$\begin{aligned} &= \sum_{i: z_i = 1} \log(p) - \log(\sigma_1) - \frac{(y-\mu_1)^2}{2\sigma_1^2} \\ &\quad + \sum_{i: z_i = 2} \log(1-p) - \log(\sigma_2) - \frac{(y-\mu_2)^2}{2\sigma_2^2} \\ &= \sum_{i=1}^{n} \{I(z_i = 1) \cdot \left[\log(p) - \log(\sigma_1) - \frac{(y-\mu_1)^2}{2\sigma_1^2}\right] \\ &\quad + I(z_i = 2) \cdot \left[\log(1-p) - \log(\sigma_2) - \frac{(y-\mu_2)^2}{2\sigma_2^2}\right]\} \end{aligned}$$

Finding Q-function

$Q(\theta|\theta^{(k)}) = E_Z[\log p(Y, Z|\theta)|Y, \theta^{(k)}]$

$$\begin{aligned} Q(\theta|\theta^{(k)}) =& E_Z(I(z_i = 1)|Y, \theta^{(k)}) \cdot f_1(p, \sigma_1, \mu_1, y) \\ &+ E_Z(I(z_i = 2)|Y, \theta^{(k)}) \cdot f_2(p, \sigma_2, \mu_2, y) \\ =& p(z_i = 1|Y, \theta^{(k)}) \cdot f_1(p, \sigma_1, \mu_1, y) \\ &+ p(z_i = 2|Y, \theta^{(k)}) \cdot f_2(p, \sigma_2, \mu_2, y) \\ =& \alpha_i^{(k,1)} f_1(\cdot) + \alpha_i^{(k,2)} f_2(\cdot) \end{aligned}$$

Using Bayes rule $p(z_i = 1|Y, \theta^{(k)})$

$$\begin{aligned} &= \frac{p(y_i|z_i = 1, \theta^{(k)}) p(z_i = 1|\theta^{(k)})}{\sum_{j=1}^{2} p(y_i|z_i = j, \theta^{(k)}) p(z_i = j|\theta^{(k)})} \\ &= \alpha_i^{(k,1)} \end{aligned}$$

where $p(y_i|z_i = 1, \theta^{(k)}) p(z_i = 1|\theta^{(k)})$

$$p^{(k)} \cdot \frac{1}{\sqrt{2\pi\sigma_1^{2(k)}}} exp\left(-\frac{(y_i - \mu_1^{(k)})^2}{2\sigma_1^{2(k)}}\right)$$

and $\alpha_i^{(k,2)} = 1 - \alpha_i^{(k,1)}$
Finally, we have

$$\begin{aligned} Q(\theta|\theta^{(k)}) = \sum_{i=1}^{n} \{\alpha_i^{(k,1)} \cdot \left[\log(p) - \log(\sigma_1) - \frac{(y-\mu_1)^2}{2\sigma_1^2}\right] \\ + \alpha_i^{(k,2)} \cdot \left[\log(1-p) - \log(\sigma_2) - \frac{(y-\mu_2)^2}{2\sigma_2^2}\right]\} \end{aligned}$$

Iterate to find MLE estimators

Solving $\theta^{(k+1)} = \arg\max_{\theta \in \Theta} Q(\theta|\theta^{(k)})$ by FOC

$$\begin{aligned} \mu_1^{(k+1)} &= \frac{\sum_{i=1}^{n} \alpha_i^{(k,1)} y_i}{\sum_{i=1}^{n} \alpha_i^{(k,1)}} \\ \mu_2^{(k+1)} &= \frac{\sum_{i=1}^{n} \alpha_i^{(k,2)} y_i}{\sum_{i=1}^{n} \alpha_i^{(k,2)}} \\ p^{(k+1)} &= \frac{\sum_{i=1}^{n} \alpha_i^{(k,1)}}{n} \\ \sigma_1^{2(k+1)} &= \frac{\sum_{i=1}^{n} \alpha_i^{(k,1)} (y_i - \mu_1^{(k+1)})^2}{\sum_{i=1}^{n} \alpha_i^{(k,1)}} \\ \sigma_2^{2(k+1)} &= \frac{\sum_{i=1}^{n} \alpha_i^{(k,2)} (y_i - \mu_2^{(k+1)})^2}{\sum_{i=1}^{n} \alpha_i^{(k,2)}} \end{aligned}$$

Example: Zero-Truncated Poisson

Problem setup

Model:

$$p(Y_i = k|\lambda) = \frac{1}{1 - e^{-\lambda}} \cdot \frac{\lambda^k e^{-\lambda}}{k!}, k \geq 1$$

Parameter: λ
Goal: Find MLE of λ

Finding complete log-likelihood function

Define $Z =$ number of zeros
$\log p(Y, Z|\lambda) =$

$$(\sum y_i) \log \lambda - (n + z)\lambda + Const$$

Finding Q-function

$Q(\lambda|\lambda^{(k)}) = E_Z[\log p(Y, Z|\lambda)|Y, \lambda^{(k)}]$

$$\begin{aligned} &= (\sum y_i) \log \lambda - [n + E_Z(z|Y, \lambda^{(k)})]\lambda + Const \\ &= (\sum y_i) \log \lambda - n\left(1 + \frac{exp(-\lambda^{(k)})}{1 - exp(-\lambda^{(k)})}\right)\lambda + Const \\ &= (\sum y_i) \log \lambda - \frac{n}{1 - exp(-\lambda^{(k)})}\lambda + Const \end{aligned}$$

Note:

$$
\begin{aligned}
P(Z=z_i) &= P(Y=0)^{z_i}[1-P(Y=0)] \\
&= exp(-\lambda z_i)[1-exp(-\lambda)] \\
E(z_i) &= \frac{exp(-\lambda)}{1-exp(-\lambda)} \\
E_Z(z) &= \sum_{i=1}^{n} E(z_i) \\
&= n\frac{exp(-\lambda)}{1-exp(-\lambda)}
\end{aligned}
$$

Iterate to find MLE estimators

Solving $\lambda^{(k+1)} - \arg\max \lambda Q(\lambda|\lambda^{(k)})$

$$
\lambda^{(k+1)} = \frac{(1-exp(-\lambda^{(k)}))\sum y_i}{n}
$$

Markov Chain

Stochastic processes

Sequence of random variable indexed by a time index $t \geq 0$

$$
X = \{X_t\}_{t\geq 0}
$$

Discrete stochastic processes: discrete $t, t = 0, 1, 2, \cdots$
Continuous stochastic processes: continuous $t, t \in [0, +\infty)$

Markov Property

Distribution of X_t only depends upon X_{t-1}

$$
P(X_t \in A|X_0, \cdots, X_{t-1}) = P(X_t \in A|X_{t-1})
$$

for any set A

Transition (one-step)

Transition of a Markov chain determines its property.

$$
\begin{aligned}
p_{ij} &= P(X_{t+1}=j|X_t=i) \\
p_{ij} &\geq 0,\ \forall(i,j) \\
\sum_j p_{ij} &= 1,\ \forall i
\end{aligned}
$$

transition probability from state i to state j at time $t+1$
Note:

- We assume Markov chain X is homogeneous in time: $\Rightarrow P(X_{t+1}=j|X_t=i)$ does not change with time t

Transition matrix (one-step)

If X has finite K states (possible positions), then transition probabilities constitute a $P_{K\times K}$ matrix.

$$
P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1K} \\ p_{21} & p_{22} & \cdots & p_{2K} \\ \vdots & \vdots & & \vdots \\ p_{K1} & p_{K2} & \cdots & p_{KK} \end{pmatrix}
$$

Multi-Step Transition (m-step)

Transition from one state to another over some fixed number of steps m

$$
p_{ij}(m) = P(X_{t+m}=j|X_t=i), m = 1, 2, \cdots
$$

m-step transition probability from state i to state j

$$
p_{ij}(m+1) = \sum_k p_{ik}(m)p_{kj}, m = 1, 2, \cdots
$$

(recursion formula) view $p_{ij}(m)$ as sum over all possible 'paths' with length m that connects i to j

Transition matrix (m-step)

Note: $P^{(1)} = P$

$$
P^{(m)} = \begin{pmatrix} p_{11}(m) & p_{12}(m) & \cdots & p_{1K}(m) \\ p_{21}(m) & p_{22}(m) & \cdots & p_{2K}(m) \\ \vdots & \vdots & & \vdots \\ p_{K1}(m) & p_{K2}(m) & \cdots & p_{KK}(m) \end{pmatrix}
$$

From recursion formula

$$
P^{(m)} = P^m = P \cdot P \cdots P
$$

State Distribution

$\pi^{(0)}$ is the initial distribution ($t=0$) over all possible states (row vector)

$$
\begin{aligned}
\pi^{(0)} &= (p_1, p_2, \cdots, p_k) \\
\pi^{(t)} &= \pi^{(t-1)}P \\
&= \pi^{(0)}P^t
\end{aligned}
$$

Stationary Distribution

Invariant/Stationary distribution: $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_K)$
transition tends towards steady-state probability

$$
\begin{aligned}
\lim_{t\to\infty} p_{ij}^{(t)} &= \pi_j \\
\lim_{t\to\infty} P^t &= \begin{pmatrix} \pi_1 & \pi_2 & \cdots & \pi_K \\ \pi_1 & \pi_2 & \cdots & \pi_K \\ \vdots & \vdots & & \vdots \\ \pi_1 & \pi_2 & \cdots & \pi_K \end{pmatrix} \\
&= \begin{pmatrix} \pi \\ \pi \\ \vdots \\ \pi \end{pmatrix} = \mathbf{1}\boldsymbol{\pi} \\
\boldsymbol{\pi} &= (\pi_1, \cdots, \pi_K) \\
\mathbf{1} &= (1, 1, \cdots, 1)^T
\end{aligned}
$$

If such $\boldsymbol{\pi}$ exist

$$
\lim_{t\to\infty} \pi^{(0)}P^t = \pi^{(0)}\mathbf{1}\boldsymbol{\pi} = \boldsymbol{\pi}
$$

Regardless of initial state, Markov chain converge to $\boldsymbol{\pi}$

Solving stationary distribution

- Stationary distribution $\boldsymbol{\pi}$ exist and unique if Markov chain is irreducible and positive recurrent
- If $\boldsymbol{\pi}$ exists and is unique, $\lim_{t\to\infty} P^t = \mathbf{1}\boldsymbol{\pi}$ holds true if Markov chain is irreducible, positive current and aperiodic
- To find stationary distribution $\boldsymbol{\pi}$ given transition matrix P, solve $\boldsymbol{\pi}P = \boldsymbol{\pi}$, or use detailed balance condition (form $x_0(x_j)$ then sub to $\sum x_i = 1$)

Note:

- Instead of solving $\pi P = \pi$, let the last equation be $\sum_k \pi_k = 1$ (else result will be $\boldsymbol{\pi} = 0$)
- Draw the diagram and for transient state, $\pi_i = 0$
- Solving $\pi P = \pi \Leftrightarrow \pi(P-1)^T = 0$

Irreducible

A Markov Chain X is called irreducible if for all pairs of states i, j, there exists a $t > 0$ s.t. $p_{ij}(t) > 0$

accessible	:= state i can transit to state j with positive probability
communicate	:= state i and j are accessible to each other
class	:= states can communicate with each other
irreducible	:= all states belonging to same class
leaking probability	:= probability escape from the current class

Recurrent and Transient state

Recurrent state i: Markov chain returns to state i with probability 1. State reoccurs for infinite number of times

$$f_i = P(\text{ever returning to state } i) = 1$$

Transient state i: $f_i < 1$. State reoccurs for finite number of times

Recurrent and Transient chain

Let τ_{ii} be the time of first return to state i

$$\tau_{ii} = \min\{t > 0 : X_t = i | X_0 = i\}$$

Irreducible Markov chain X is recurrent if $P(\tau_{ii} < \infty) = 1$ for some (and hence for all) state i. Else, X is transient
Alternatively, for recurrent chain:

$$\sum_t p_{ii}(t) = \infty, \ \forall i$$

Note:
It's easier to draw the states to determine if chain is recurrent

Positive Recurrent

Irreducible Markov chain X is called positive recurrent if

$$E[\tau_{ii} < \infty] \ \forall i$$

Otherwise, it is called null recurrent
Note:

- All states in a communication class C are all together either positive recurrent, null recurrent, or transient
- In an irreducible Markov chain, all states must together be positive recurrent, null recurrent or transient.
- If a Markov chain only has a finite number of states, and if it is irreducible, then it must be positive recurrent.

Positive Recurrent (alternative condition)

Positive recurrence has a stationary pmf $\pi(\cdot)$ on the state space of X s.t.

$$\sum_i \pi_i p_{ij}(t) = \pi_j, \ \forall j, t \geq 0$$

If at time t, $\pi^{(t)} = \pi \Rightarrow \pi^{(t)} = \pi = \pi^{(t+1)}$
Note:
Every irreducible Markov chain with finite number of states has a unique stationary distribution

Aperiodic

An irreducible chain X is called aperiodic if for some (and hence for all) i

$$\text{Greatest common divisor of } \{t : p_{ii}(t) > 0\} = 1$$

Simply, if chain has both 2, 3 periods then it's aperiodic

Convergence Theorem (Ergodic Theorem)

X is ergodic:
If $X = \{X_1, X_2, \cdots\}$ is a positive recurrent and aperiodic Markov Chain, then its stationary distribution $\pi(\cdot)$ is the unique probability
The following holds

1. $p_{ij}(t) \to \pi_j$ as $t \to \infty \ \forall i, j$
2. (Ergodic Theorem) For a function $h(x)$, if $E_\pi[|h(X)|] < \infty$, then

$$\frac{1}{N} \sum_{k=1}^{N} h(X_k) \to E_\pi[(h(X))]$$

as $N \to \infty$, with probability 1

where $E_\pi[h(X)] = \sum_i h(i)\pi_i$, the expectation of $h(x)$ with respect to $\pi(\cdot)$

Finding Stationary probability

In general, solve system of equations or detailed balance condition (explained here)
However, suppose we have positive number $x_j, j = 1, 2, \cdots, K$ (finite state space), such that

$$x_i p_{ij} = x_j p_{ji}, i \neq j, \sum_{j=1}^{K} x_j = 1$$

$\Rightarrow \pi$ satisfies $\pi_j \propto x_j, j = 1, 2, \cdots, K$ because $\{\pi_j, j = 1, 2, \cdots, K\}$ are the unique solution to $\pi P = \pi$

Simulation of Discrete Markov Chains

```r
# transition matrix
P <- rbind(c(0.2, 0.8), c(0.6, 0.4))
# total number of steps
N <- 5000
# path taken
path <- rep(0, N)
path[1] <- 1 # starting point
# simulation
for (i in 2:N) {
    path[i] <- sample(
        # next state space
        c(1, 2),
        size = 1,
        # transition matrix
        P[path[i-1], ]
    )
}
```

Bayesian Inference

Data: $Y = \{y_1, \cdots, y_n\}$
Parameters: $\theta = (\theta_1, \cdots, \theta_p)$ which lies in a set Θ
Model (Likelihood): $p(Y|\theta) \Leftrightarrow L(Y; \theta)$
Prior distribution: $\pi(\theta)$
Posterior distribution: $\pi(\theta|Y)$ (inference based on)

$$\pi(\theta|Y) = \frac{p(Y|\theta)\pi(\theta)}{\int_\Theta p(Y|\theta)\pi(\theta)d\theta} \propto p(Y|\theta)\pi(\theta)$$

Difficulty in Posterior Calculation

- normalizing constant does not have closed form (closed form only available when prior is conjugate $\Rightarrow$ prior and posterior fall into the same parametric family)

- When θ is multi-dimensional, difficult to find conjugate priors for the entire vector of θ

Metropolis Algorithm

General, always require transition kernel $Q(\theta^{(t)}, \theta)$

1. Initial state $\theta^{(t)}$
2. Generate θ^* from density $q(\theta|\theta^{(t)}) = Q(\theta^{(t)}, \theta)$
3. Compute acceptance probability

$$\begin{aligned}\alpha(\theta^{(t)}, \theta^*) &= \min\left(1, \frac{\pi(\theta^*|Y)Q(\theta^*, \theta^{(t)})}{\pi(\theta^{(t)}|Y)Q(\theta^{(t)}, \theta^*)}\right) \\ &= \min\left(1, \frac{p(Y|\theta^*)\pi(\theta^*)Q(\theta^*, \theta^{(t)})}{p(Y|\theta^{(t)})\pi(\theta^{(t)})Q(\theta^{(t)}, \theta^*)}\right)\end{aligned}$$

4. Set next state

$$\theta^{(t+1)} = \begin{cases} \theta^*, & p = \alpha(\theta^{(t)}, \theta^*) \\ \theta^{(t)}, & p = 1 - \alpha(\theta^{(t)}, \theta^*) \end{cases}$$

Transition Kernels

Properties

- Probability of transit to all states = 1

$$\int_\Theta Q(\theta_a, \theta)d\theta = 1$$

- If transition kernel is symmetric

$$Q(\theta^*, \theta^{(t)}) = Q(\theta^{(t)}, \theta^*)$$

Common symmetric transition kernel

- Uniform kernel

$$Q(\theta_a, \theta_b) = \frac{1}{2\delta} \sim Unif(\theta_a - \delta, \theta_a + \delta)$$

- Normal kernel

$$Q(\theta_a, \theta_b) = \frac{1}{\sqrt{2\pi\delta^2}} exp\left\{-\frac{(\theta_b - \theta_a)^2}{2\delta^2}\right\}$$

Metropolis-Hasting Algorithm

1. set $\theta^{(0)}$
2. for $t \in [0, T-1]$ do

   [1] Propose θ^* from density $q(\theta|\theta^{(t)}) = Q(\theta^{(t)}, \theta)$

   [2] Compute acceptance probability

$$\alpha(\theta^{(t)}, \theta^*) = \min\left(1, \frac{p(Y|\theta^*)\pi(\theta^*)Q(\theta^*, \theta^{(t)})}{p(Y|\theta^{(t)})\pi(\theta^{(t)})Q(\theta^{(t)}, \theta^*)}\right)$$

   [3] Generate $U \sim Unif(0, 1)$

   [4] If $U < \alpha(\theta^{(t)}, \theta^*)$, set $\theta^{(t+1)} = \theta^*$

   [5] Else set $\theta^{(t+1)} = \theta^{(t)}$
3. end for

If symmetric kernel (Random Walk), then $Q(\theta^{(t)}\theta)$ cancels

MH conditions

Want stationary distribution

$$\int_\Theta \pi(\theta_a)K(\theta_a, \theta)d\theta_a = \pi(\theta)$$

Sufficient condition: detailed balanced conditions holds

$$\pi(\theta_a)K(\theta_a, \theta_b) = \pi(\theta_b)K(\theta_b, \theta_a)$$

MH algorithm converge to stationary distribution $\pi(\theta|Y)$ if

$$\begin{aligned}&\pi(\theta^{(t)}|Y)Q(\theta^{(t)}, \theta^*)\alpha(\theta^{(t)}, \theta^*) \\ &= \pi(\theta^*|Y)Q(\theta^*, \theta^{(t)})\alpha(\theta^*, \theta^{(t)})\end{aligned}$$

since transition kernel in MH is $K(\theta^{(t)}, \theta^*) \approx Q(\theta^{(t)}, \theta^*)\alpha(\theta^{(t)}, \theta^*)$

MH tricks

- Parameter space

  Transform parameters to unbounded real line

  e.g. $\theta = Var(x) \geq 0 \Rightarrow \log(\theta) \in \boldsymbol{R}$
- Initial value $\theta^{(0)}$

  Select maximised parameter for log posterior $\log \pi(\theta|Y)$

  Can retrive Hessian matrix (usually negative definite matrix)
- Normal proposal kernel

  optimal acceptance rate: 0.234

  optimal variance (d :=dimension of θ, best ≥ 3)

$$\sigma^2 = c^2\Sigma, \ \Sigma = (-H)^{-1}, \ c = \frac{2.4}{\sqrt{d}}$$

- Burn-in

  drop initial t_0 draws as burn-in
- Thinning

  thin the chain by taking 1 from every 10 draws etc
- Diagnostics

  check trace plots have stabilized visually

  check autocorrelations $\{\theta^{(t)}\}_{t=1}^T$ are decreasing fast

Gibbs Sampler

Idea: sample conditional distribution $\pi(\theta_i|Y, \theta_j), j \neq i$ to retrieve posterior distribution $\pi(\theta|Y), \theta = (\theta_1, \cdots, \theta_d)$

1. Initialize $\theta^{(0)}$
2. At step $t \in [0, T-1]$

   Sample $\theta_1^{(t+1)} \sim \pi(\theta_1|\theta_2^{(t)}, \theta_3^{(t)}, \cdots, \theta_d^{(t)}, Y)$

   Sample $\theta_2^{(t+1)} \sim \pi(\theta_2|\theta_1^{(t+1)}, \theta_3^{(t)}, \cdots, \theta_d^{(t)}, Y)$

   ...

   Sample $\theta_d^{(t+1)} \sim \pi(\theta_d|\theta_1^{(t+1)}, \theta_2^{(t+1)}, \cdots, \theta_{d-1}^{(t+1)}, Y)$
3. Set $\theta^{(t+1)} = (\theta_1^{(t+1)}, \cdots, \theta_d^{(t+1)})$
4. Repeat the steps until time T. Output $(\theta^{(1)}, \cdots, \theta^{(T)})$

Ex: Multivariate Dirichlet Density

Dirichlet Density

$$\begin{aligned}&f(x_1, x_2, \cdots, x_d) \propto x_1^{\alpha_1 - 1}x_2^{\alpha_2 - 1}\cdots x_d^{\alpha_d - 1} \\ &\sum_{i=1}^d x_i = 1, \ x_i > 0\end{aligned}$$

Conditional distribution

$$x_i|x_j \sim Beta(\alpha_i - 1, 2), j \neq i$$

Marginal distribution

$$x_i \sim Beta\left(\alpha_i, \sum_{j\neq i} \alpha_j\right)$$

Note: for change of variable remember

$$x = g(y)$$
$$\Rightarrow f_x(x) = f_y(g^{-1}(y))\left|\frac{dy}{dx}\right|$$

Ex: Posterior of a Normal Model

Let $\pi(\mu, \sigma^2) \approx \frac{1}{\sigma^2}$ (improper prior)
$Y = \{y_1, \cdots, y_n\} \sim N(\mu, \sigma^2)$ iid

$$\begin{aligned}\pi(\mu, \sigma^2|Y) &\propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{(y_i-\mu)^2}{2\sigma^2}\right\} \cdot \frac{1}{\sigma^2} \\ &\propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+1} exp\left\{-\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2\sigma^2}\right\}\end{aligned}$$

key trick: Let $\tau = 1/\sigma^2$

$$\pi|\sigma^2, Y \sim N\left(\bar{y}, \frac{\sigma^2}{n}\right)$$
$$\tau|\mu, Y \sim Gamma\left(\frac{n}{2}, \frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2}\right)$$