MTH 511: Problem Set

Instructor: Dootika Vats

November 9, 2021

Sampling

1. Show that is $U \sim U(0,1)$, then for any a, b,

$$(b-a) * U + a \sim U(a,b)$$

2. Use the inverse transform method to sample from a geometric distribution, where for $0 < p < 1$ and $q = 1 - p$,

$$\Pr(X = i) = pq^{i-1}, \quad i \geq 1, \quad \text{where } q = 1 - p\,.$$

3. List as many appropriate proposal distributions as you can think of for the following target distributions:
   - Binomial
   - Bernoulli
   - Geometric
   - Negative Binomial
   - Poisson
4. (Using R) Draw 10,000 draws from a Binomial$(20, .75)$ distribution using an accept-reject sampler.

5. In an accept-reject algorithm, we need to find c such that

$$\frac{p_j}{q_j} \leq c \quad \text{forall } j \text{ for which } p_i > 0\,.$$

And, the probability of accepting in any iteration is $1/c$. Why is c guaranteed to be more than 1?

6. Simulate from a Negative Binomial(n, p) using the inverse transform and accept-reject methods. Implement in R with $n = 10$ successes and $p = .30$.

7. Simulate from the following "truncated Poisson distribution" with pmf:

$$\Pr(X = i) = \frac{e^{-\lambda}\lambda^i/i!}{\sum_{j=0}^{m} e^{-\lambda}\lambda^j/j!} \quad i = 0, 1, 2, \ldots, m\,.$$

Implement in R with $m = 30$ and $\lambda = 20$.

8. Implement a an algorithm to sample from a Zero Inflated Binomial distribution. Can you think of an application of such a distribution?

9. Sample from a Bern$(p^3(1-p)^2)$ distribution using only draws from a Bern(p) distribution. Implement in R for $p = .40$.

10. Using the inverse transform method, simulate from Exp(λ) for any $\lambda > 0$. Implement this for $\lambda = 5$.

11. Use the inverse transform method to obtain samples from the Weibull(α, λ)

$$f(x) = \alpha\lambda x^{\alpha-1} e^{-\lambda x^\alpha}\,.$$

12. Sample following the following distribution using two different methods:

$$f(x) = \begin{cases} \frac{3}{4}(1-x^2) & \text{if } x \in (-1, 1) \\ 0 & \text{otherwise} \end{cases}$$

13. Ross: exercise 5.1 - 5.9

Exercises

1. Give a method for generating a random variable having density function

$$f(x) = e^x/(e-1), \quad 0 \leqslant x \leqslant 1$$

2. Give a method to generate a random variable having density function

$$f(x) = \begin{cases} \frac{x-2}{2} & \text{if } 2 \leqslant x \leqslant 3 \\ \frac{2-x/3}{2} & \text{if } 3 \leqslant x \leqslant 6 \end{cases}$$

3. Use the inverse transform method to generate a random variable having distribution function

$$F(x) = \frac{x^2 + x}{2}, \quad 0 \leqslant x \leqslant 1$$

4. Give a method for generating a random variable having distribution function

$$F(x) = 1 - \exp(-\alpha x^{\beta}), \quad 0 < x < \infty$$

A random variable having such a distribution is said to be a Weibull random variable.

5. Give a method for generating a random variable having density function

$$f(x) = \begin{cases} e^{2x}, & -\infty < x < 0 \\ e^{-2x}, & 0 < x < \infty \end{cases}$$

6. Let X be an exponential random variable with mean 1. Give an efficient algorithm for simulating a random variable whose distribution is the conditional distribution of X given that $X < 0.05$. That is, its density function is

$$f(x) = \frac{e^{-x}}{1 - e^{-0.05}}, \quad 0 < x < 0.05$$

Generate 1000 such variables and use them to estimate $E[X|X < 0.05]$. Then determine the exact value of $E[X|X < 0.05]$.

7. (The Composition Method) Suppose it is relatively easy to generate random variables from any of the distributions $F_i, i = 1, \ldots, n$. How could we generate a random variable having the distribution function

$$F(x) = \sum_{i=1}^{n} p_i F_i(x)$$

where $p_i, i = 1, \ldots, n$, are nonnegative numbers whose sum is 1?

8. Using the result of Exercise 7, give algorithms for generating random variables from the following distributions.

(a) $F(x) = \frac{x+x^3+x^5}{3}, 0 \leqslant x \leqslant 1$

(b) $F(x) = \begin{cases} \frac{1-e^{-2x}+2x}{3} & \text{if} 0 < x < 1 \\ \frac{3-e^{-2x}}{3} & \text{if} 1 < x < \infty \end{cases}$

(c) $F(x) = \sum_{i=1}^{n} \alpha_i x^i, 0 \leqslant x \leqslant 1, \quad$ where $\alpha_i \geqslant 0, \quad \sum_{i=1}^{n} \alpha_i = 1$

9. Give a method to generate a random variable having distribution function

$$F(x) = \int_0^{\infty} x^y e^{-y}\, dy, \quad 0 \leqslant x \leqslant 1$$

[Hint: Think in terms of the composition method of Exercise 7. In particular, let F denote the distribution function of X, and suppose that the conditional distribution of X given that $Y = y$ is

$$P\{X \leqslant x|Y = y\} = x^y, \quad 0 \leqslant x \leqslant 1$$

14. Ross: exercise 5.15 - 5.23

what information?

(b) Show that the rejection method reduces in this case to generating a random variable X having distribution G and then accepting it if it lies between a and b.

15. Give two methods for generating a random variable having density function

$$f(x) = xe^{-x}, \quad 0 \leqslant x < \infty$$

and compare their efficiency.

16. Give two algorithms for generating a random variable having distribution function

$$F(x) = 1 - e^{-x} - e^{-2x} + e^{-3x}, \quad x > 0$$

17. Give two algorithms for generating a random variable having density function

$$f(x) = \frac{1}{4} + 2x^3 + \frac{5}{4}x^4, \quad 0 < x < 1$$

18. Give an algorithm for generating a random variable having density function

$$f(x) = 2xe^{-x^2}, \quad x > 0$$

19. Show how to generate a random variable whose distribution function is

$$F(x) = \frac{1}{2}(x + x^2), \quad 0 \le x \le 1$$

using

(a) the inverse transform method;
(b) the rejection method;
(c) the composition method.

Which method do you think is best for this example? Briefly explain your answer.

20. Use the rejection method to find an efficient way to generate a random variable having density function

$$f(x) = \frac{1}{2}(1 + x)e^{-x}, \quad 0 < x < \infty$$

21. When generating a gamma random variable with parameters $(\alpha, 1), \alpha < 1$, that is conditioned to exceed c by using the rejection technique with an exponential conditioned to exceed c, what is the best exponential to use? Is it necessarily the one with mean α, the mean of the gamma $(\alpha, 1)$ random variable?

22. Give an algorithm that generates a random variable having density

$$f(x) = 30(x^2 - 2x^3 + x^4), \quad 0 \leqslant x \leqslant 1$$

Discuss the efficiency of this approach.

23. Give an efficient method to generate a random variable X having density

$$f(x) = \frac{1}{.000336}x(1 - x)^3, \quad .8 < x < 1$$

24. In Example 5f we simulated a normal random variable by using the rejection technique with an exponential

15. (Using R)

(a) Implement an accept-reject sampler to sample uniformly from the circle $\{x^2 + y^2 \le 1\}$ and obtain 10000 samples and estimate the probability of acceptance. Does it approximately equal $\pi/4$?

(b) Now consider sampling uniformly from a p-dimensional sphere (a circle is $p = 2$). Consider a p-vector $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ and let $\|\cdot\|$ denote the

Euclidean norm. The pdf of this distribution is

$$f(\mathbf{x}) = \frac{\Gamma\left(\frac{p}{2}+1\right)}{\pi^{p/2}} I\{\|\mathbf{x}\| \leq 1\}\,.$$

Use a uniform p-dimensional hypercube to sample uniformly from this sphere. Implement this for $p = 3, 4, 5$, and 6. What happens as p increases?

16. (Using R)

    (a) Using accept-reject and a standard normal proposal, obtain samples from a truncated standard normal distribution with pdf:

    $$f(x) = \frac{1}{\Phi(a) - \Phi(-a)} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} I(-a < x < a)\,,$$

    where $\Phi(\cdot)$ is the CDF of a standard normal distribution. Run for $a = 4$ and $a = 1$. What are the differences between the two settings.

    (b) Now consider a multivariate truncated normal distribution, where for $\mathbf{x} = (x_1, x_2, \ldots, x_p)$, the pdf is

    $$f(\mathbf{x}) = \left(\frac{1}{\Phi(a) - \Phi(-a)}\right)^p \left(\frac{1}{\sqrt{2\pi}}\right)^p e^{-\mathbf{x}^T\mathbf{x}/2} I(-a < \mathbf{x} < a)\,.$$

    Implement an accept-reject sampler with proposal distribution $N_p(0, I)$ with $a = 4$ and $p = 3, 10$ and with $a = 1$ and $p = 3, 10$. Describe the differences between these settings.

17. Implement an accept-reject sampler to draw from a Gamma$(\alpha, 1)$ for $\alpha > 1$. *Hint:* See the discussion following Example 5e in Ross.

    Using the above method, can you draw samples from Gamma(α, β), for any β?

18. In accept-reject sampling, why is $c \geq 1$?

19. Use ratio-of-uniforms method to sample from a truncated exponential distribution with density

    $$f(x) = \frac{e^{-x}}{1 - e^{-a}} \quad 0 < x < a\,.$$

    How efficient is this algorithm?

20. Use ratio-of-uniforms method to sample from the distribution with density

$$f(x) = \frac{1}{x^2} \quad x \geq 1\,.$$

21. Use ratio-of-uniforms method to draw samples from a t_ν distribution for $\nu \geq 1$.

22. Draw 10^4 samples from $N_3(\mu, \Sigma)$ where

$$\mu = \begin{pmatrix} 5 \\ 0 \\ -2 \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} 1 & .9 & -.3 \\ .9 & 1 & .1 \\ -.3 & .1 & 1 \end{bmatrix}$$

23. In R, for any specified $p \times p$ matrix Σ, how will you check whether the matrix is positive-definite?

Importance Sampling

1. Estimate $\int_0^1 e^x dx$ using importance sampling.

2. Consider the problem of estimating the k-th moment of a Beta(α, β) distribution. For which values of α and β are we sure to obtain importance estimators of the kth moment with a finite variance for a uniform proposal distribution.

3. In the previous problem, give other examples of importance proposal distributions that will give finite variance of the importance estimator? For what values of α and β is the finite variance of the estimator not guaranteed?

4. Consider estimating the mean of a standard Cauchy distribution using importance sampling with a normal proposal distribution. Does the estimator have finite variance?

5. Considering a density $f(x)$ and an importance proposal $g(x)$. Suppose

$$\sup_x \frac{f(x)}{g(x)} < \infty \,.$$

   In order to estimate the mean of the target density, is there any benefit to using importance sampling over accept-reject sampling?

6. For a target distribution π and a proposal g, if

$$\sup_x \frac{\pi(x)}{g(x)} < \infty$$

   then we know that the simple importance estimator has finite variance. Does the weighted importance estimator also have finite variance?

7. For some known $y_i \in \mathbb{R}$, $i = 1, \ldots, n$ and some $\nu > 2$, suppose the target density is

$$\pi(x) \propto e^{-x^2/2} \prod_{i=1}^{n} \left(1 + \frac{(y_i - x)^2}{\nu}\right)^{-(\nu+1)/2} .$$

   Generate ys using the following code for $\nu = 5$

```
set.seed(1)
n <- 50
nu <- 5
y <- rt(n, df = nu)
```

Implement an importance sampling estimator with a $N(0,1)$ proposal to estimate the first moment of this distribution. Does the weighted importance sampling estimator have finite variance? What happens if $\nu = 1$ and $\nu = 2$?

8. Suppose interest is in estimating

$$\theta = \int_0^{10} \exp\{-2|x-5|\}\, dx\,.$$

- What is the optimal simple importance proposal distribution? (*Hint:* look up Laplace distribution) and what is the corresponding simple importance sampling estimator?
- Implement a weighted importance sampling procedure with the same proposal distribution from above. How do the final estimators compare?

The quantity θ can be written as

$$\theta = \int_0^{10} 10\exp\{-2|x-5|\}\, \pi(x) dx\,,$$

where $\pi(x)$ is the density of a $U[0,10]$. We know we can do IID sampling that is, sample $X_1, X_2, \ldots, X_N$ from Unif$[0,10]$ and estimate θ. But using importance sampling, we can reduce the variance. First, note that the optimal importance distribution here is

$$g^*(z) \propto 10\exp\{-2|z-5|\}\, I(-10 < z < 10)\,.$$

The function h id identical to the density of a Laplace (Double exponential) distribution. For a Laplace random variable with parameters μ and b

$$l(x) = \frac{1}{2b}\exp\left\{-\frac{|z-\mu|}{b}\right\} \quad -\infty < z < \infty$$

So $h(x)$ is the density of Laplace$(5, 1/2)$, and π truncates it. So the optimum proposal is a truncated Laplace$(5, 1/2)$. However, since implementing truncation requires accept-reject, that is time consuming, so I will get

$$g(z) = \exp\{-2|z-5|\} \quad -\infty < z < \infty$$

Then,

$$\begin{aligned}\theta &= \int_0^{10} \frac{10\exp\{-2|z-5|\}\,\pi(x)}{g(x)} g(z)dz \\ &= \int_0^{10} g(z) \\ &= \int_{-\infty}^{\infty} I(0<z<10)g(z) \\ &= \mathrm{E}_g\left(I(0<z<10)\right)\,.\end{aligned}$$

Thus, the simple importance sampling estimator is

$$\hat{\theta}_g = \frac{1}{N}\sum_{t=1}^{N} I(0<Z_t<10)\,.$$

Now suppose, in this example, we did not know the normalized density $\pi(x)$. We know that $\pi(x) \propto 1$ that's it. So,

$$\theta = \int_0^{10} \frac{\exp\{-2|z-5|\}}{a}\, a\tilde{\pi}(x)dx = \int \frac{\exp\{-2|z-5|\}}{a}\; aI(0<x<10)dx\,.$$

Obtain $Z_1,\ldots,Z_N \sim \text{Laplace}(5,1/2)$ and the weights are

$$w(Z_t) = \frac{I(0<Z_t<10)}{\exp\{-2|z-5|\}}\,.$$

So the weighted importance sampling estimator is,

$$\hat{\theta}_w = \frac{\sum_{t=1}^{N} 10\exp\{-2|z-5|\}\,w(Z_t)}{\sum_{t=1}^{N} w(Z_t)}\,.$$

Data Analysis

1. *Simple linear regression*: Load the `cars` dataset in R:

   ```
   data(cars)
   ```

   Fit a linear regression model using maximum likelihood with response y being the distance and x being speed. Remember to include an intercept term in X by making the first column as a column of 1s. *Do not use inbuilt functions in R to fit the model.*

2. *Multiple linear regression*: Load the `fuel2001` dataset in R:

   ```
   fuel2001 <- read.csv("https://dvats.github.io/assets/fuel2001.csv",
   row.names = 1)
   ```

   Fit the linear regression model using maximum likelihood with response `FuelC`. Remember to include an intercept in X.

3. *Simulating data in R:*

   Let $X \in \mathbb{R}^{n \times p}$ be the design matrix, where all entries in its first column equal one (to form an intercept). Let $x_{i,j}$ be the (i,j)th element of X. For the ith case, $x_{i1} = 1$ and $x_{i2}, \ldots, x_{ip}$ are the values of the $p-1$ predictors. Let y_i be the response for the ith case and define $y = (y_1, \ldots, y_n)^T$. The model assumes that y is a realization of the random vector

   $$Y \sim N_n(X\beta_*, \sigma_*^2 I_n)\,,$$

   where $\beta_* \in \mathbb{R}^p$ are unknown regression coefficients and $\sigma_*^2 > 0$ is the unknown variance.

   For our simulation, let's pick $n = 50, p = 5, \sigma^2 = 1/2$ and generate the entries of β_* as p independent draws from $N(0,1)$:

   ```
   set.seed(1)
   n <- 50
   p <- 5
   sigma2.star <- 1/2
   beta.star <- rnorm(p)
   beta.star  # to output
   [1] -0.6264538  0.1836433 -0.8356286  1.5952808  0.3295078
   ```

We will create the design matrix $X \in \mathbb{R}^{n \times p}$, so that $x_{i1} = 1$ and the other entries are from $N(0, 1)$.

```
X <- cbind(1, matrix(rnorm(n*(p-1)), nrow = n, ncol = (p-1)))
```

Now we will generate a realization of $Y \sim N_n(X\beta_*, \sigma_*^2 I_n)$:

```
y <-  X %*% beta.star + rnorm(n, mean = 0, sd = sqrt(sigma2.star))
```

In this way we have generate *simulated* data to be used in regression.

4. Find the MLE estimator of β and σ^2 from the previous dataset. Is it close to β_* and σ_*^2? Find the ridge regression solution with $\lambda = 0.01, 1, 10, 100$.

5. *Regression: an equivalent optimization*

   In our original setup $X \in \mathbb{R}^{n \times p}$, all entries in its first column equal to one to form an intercept. The MLE estimate (when it exists) is

   $$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} (y - X\beta)^T (y - X\beta) \,.$$

   Define X_{-1} be the matrix X with its first column removed. Let $\bar{y} = n^{-1} \sum_{i=1}^n y_i$ and $\bar{x}^T = n^{-1} 1_n^T X_{-1} = (n^{-1} \sum_{i=1}^n x_{i2}, \ldots, n^{-1} \sum_{i=1}^n x_{ip})$. Let $\tilde{y} = (y_1 - \bar{y}, \ldots, y_n - \bar{y})^T$ and $\tilde{X} = X_{-1} - 1_n \bar{x}^T$. Then $\tilde{y}$ is the centered response and $\tilde{X}$ is the centered design matrix.

   Suppose that

   $$\begin{aligned} \hat{\beta}_{-1} &= \arg\min_{\tilde{\beta} \in \mathbb{R}^{p-1}} (\tilde{y} - \tilde{X}\tilde{\beta})^T (\tilde{y} - \tilde{X}\tilde{\beta}) \\ \hat{\beta}_1 &= \bar{y} - \bar{x}^T \hat{\beta}_{-1} \,. \end{aligned}$$

   Then $(\hat{\beta}_1, \hat{\beta}_{-1}^T)^T$ is equivalent to $\hat{\beta}$ above. Verify this for the dataset generated in Exercise 3.

6. *Logistic Regression:* Often in regression, the response may be a 0 or 1. That is, the response is a Bernoulli random variable. Let the covariate vector for the ith observation be $x_i = (1, x_{i2}, \ldots, x_{ip})^T$. Suppose y_i is a realization of Y_i where

   $$Y_i \sim \text{ Bern } \left(\frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)} \right) .$$

   Find the MLE of β_*.

7. Generate data according to the logistic regression model above with $n = 50$, and use Newton-Raphson's and gradient ascent algorithm to find the MLE.

8. (Using R) Find the MLE of a $\Gamma(\alpha, 1)$ distribution using Newton-Raphson's method. Set $\alpha = 4$ and $n = 10$. Rerun the Newton-Raphson's algorithm with different starting values.

9. (Using R) Find the MLE of a location Cauchy distribution with density

$$f(x) = \frac{1}{\pi} \frac{1}{1 + (x - \mu)^2} \quad \mu \in \mathbb{R}, x \in \mathbb{R}\,.$$

   Set $\mu = -2$ and $n = 4,100$, and run the Newton-Raphson's algorithm with different starting values. Are starting values more impactful here than compared to the Gamma problem? Why or why not? Now repeat the same for the gradient ascent algorithm.

10. (Modified Newton-Raphson): It is possible to "overshoot" when using Newton-Raphson's algorithm, when the objective function is not concave. In these scenarios, we use a modified Newton-Raphson's approach, with step factors.

    Suppose the kth iteration is such that $f'(\theta_{(k)}) > 0$ (that is the function is increasing at $\theta_{(k)}$), and NR takes us to $\theta_{(k+1)}$ where $f'(\theta_{(k+1)}) < 0$, so that now the function is decreasing. This means we may have overshot! As a compromise, we may want to implement the following algorithm:

$$\theta_{(k+1)} = \theta_{(k)} - \lambda_{(k)} \frac{f'(\theta_{(k)})}{f''(\theta_{(k)})}$$

    where $\lambda_{(k)}$ is a step-factor sequence chosen at every iteration so that

$$f(\theta_{(k+1)}) > f(\theta_{(k)})\,.$$

    That is, the next value must be such that we achieve an increase in the objective function. You can choose $\lambda_{(k)}$ so that

    (a) If $f(\theta_{(k+1)}) > f(\theta_k)$, then $\lambda_{(k)} = 1$. Else $\lambda_{(k)} = 1/2$, and recalculate $f(\theta_{(k+1)}$

    (b) If $f(\theta_{(k+1)}) > f(\theta_k)$, then continue, else set $\lambda_{(k)} = 1/2^2$ and so on...

    Implement this modified algorithm for the Gamma and Cauchy examples.

11. (Using R) Find the MLE of (μ, σ^2) for the $N(\mu, \sigma^2)$ distribution using Newton-

Raphson's method. Compare with the closed form estimates.

12. Find the MLE of (α, μ) for a Pareto distribution with density

$$f(x) = \frac{\alpha \mu^\alpha}{x^{\alpha+1}} \quad x \geq \mu, \quad \mu, \alpha > 0\,.$$

13. (Using R) Using both Newton-Raphson and gradient ascent algorithm, maximize objective function

$$f(x) = \cos(x) \quad x \in [-\pi, 3\pi]\,.$$

14. (Using R) Find the MLE of (α, β) for the Beta distribution:

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}\,.$$

Run algorithm for $n = 5$ and $n = 100$ samples.

15. Following the steps from class, write the EM algorithm for a mixture of K Gaussians, for any general K. That is, the distribution is

$$f(x|\theta) = \sum_{k=1}^{K} \pi_k f_k(x|\mu_k, \sigma_k^2)\,.$$

16. (Using R) Consider the `faithful` dataset in R, which contains waiting time between eruptions and the duration of each eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA. First, run the following code

```
data(faithful)
plot(density(faithful$eruptions))
```

You will see that the length of the eruptions looks like a bimodal distribution. For any given eruption, let X_i be the eruption time. Let

$$Z_i = \begin{cases} 1 & X_i \text{ has short eruptions} \\ 2 & X_i \text{ has long eruptions} \end{cases}$$

Thus Z_i is a *latent* variable which is not observed. Let π_1 and π_2 be the probability of short and long eruptions, respectively. Assume that the joint distribution of

(X, Z) is

$$f(x, z|\theta) = \pi_1 f_1(x|\mu_1, \sigma_1^2) I(Z = 1) + \pi_2 f_2(x|\mu_2, \sigma_2^2) I(Z = 2)\,.$$

Implement the EM algorithm for this example.

17. (Using R) For the same dataset `faithful`, we will fit a *multivariate* mixture of Gaussians for both the eruption time and waiting times. Let X_i be the eruption time and Y_i be the waiting time for the ith eruption. Let

$$Z_i = \begin{cases} 1 & X_i \text{ and } Y_i \text{ has short eruptions and short wait times} \\ 2 & X_i \text{ and } Y_i \text{ has long eruptions and long wait times} \end{cases}.$$

First, we want to find the EM steps for this. The joint distribution of the observed $t = (x, y)$ and the latent variable z is

$$f(t, z|\theta) = \pi_1 f_1(t|\mu_1, \Sigma_1) I(Z = 1) + \pi_2 f_2(t|\mu_2, \Sigma_2^2) I(Z = 2)\,,$$

where $\mu_c \in \mathbb{R}^2$, $\Sigma_c \in \mathbb{R}^{2\times 2}$ and

$$f_c(t \mid \mu_c, \sigma_c^2) = \left(\frac{1}{2\pi}\right) \frac{1}{|\Sigma_c|^{1/2}} \exp\left\{-\frac{(t-\mu_c)^T \Sigma_c^{-1} (t-\mu_c)}{2}\right\}.$$

Similar to the one dimensional case, set up the EM algorithm for this two-dimensional case, and then implement this on the Old Faithful dataset.

18. Repeat the previous exercises for four latent class defined as

$$Z_i = \begin{cases} 1 & X_i \text{ and } Y_i \text{ has short eruptions and short wait times} \\ 2 & X_i \text{ and } Y_i \text{ has long eruptions and long wait times} \\ 3 & X_i \text{ has short eruptions and } Y_i \text{ has long wait times} \\ 4 & X_i \text{ has long eruptions and } Y_i \text{ has short wait times} \end{cases}.$$

19. (EM algorithm for multinomial) Suppose $y = (y_1, y_2, y_3, y_4)$ has a multinomial distribution with probabilities

$$\left(\frac{1}{2} + \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right).$$

The joint distribution of y is

$$g(y \mid \theta) = \frac{(\sum y_i)!}{\prod_{i=1}^{4} y_i!} \left(\frac{1}{2} + \frac{\theta}{4}\right)^{y_1} \left(\frac{1-\theta}{4}\right)^{y_2} \left(\frac{1-\theta}{4}\right)^{y_3} \left(\frac{\theta}{4}\right)^{y_4} .$$

Suppose you observe $y = (125, 18, 20, 34)$. Also, suppose that the complete data is $(z_1, z_2, y_2, y_3, y_4)$ where $z_1 + z_2 = x_1$. That is, the first variable y_1 is broken into two groups, with the new probabilities are

$$\left(\frac{1}{2}, \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right) .$$

The complete data distribution is

$$f(z, y \mid \theta) = \frac{(z_1 + z_2 + y_2 + y_3 + y_4)!}{z_1! z_2! y_2! y_3! y_4!} \left(\frac{1}{2}\right)^{x_1} \left(\frac{\theta}{4}\right)^{x_2} \left(\frac{1-\theta}{4}\right)^{y_2} \left(\frac{1-\theta}{4}\right)^{y_3} \left(\frac{\theta}{4}\right)^{y_4} .$$

The E-Step is

$$q(\theta|\theta_{(k)}) = \mathrm{E}_{\theta_{(k)}} \left[\log f(z, y \mid \theta) \mid y_1, y_2, y_3, y_4\right] .$$

Write the above expectation explicitly, and then write the M-step. Implement this in R and return the estimate of θ.

20. Repeat Exercise 19 using Monte Carlo EM.

Resampling

1. *Comparing different cross-validation techniques*: Generate a dataset using the following code:

   ```
   set.seed(10)
   n <- 100
   p <- 50
   sigma2.star <- 4
   beta.star <- rnorm(p, mean = 2)
   beta.star
   ```

   Generate a dataset (y, X) to fit the linear regression model

   $$y = X\beta + \epsilon\,.$$

   Implement the holdout method, leave-one-out, 10-fold, and 5-fold cross-validation over 500 replications. Keep a track of the CV error from each method and compare the performance of all cross-validation methods.

2. *cars dataset*: Estimate the prediction error for the cars dataset using 10-fold, 5-fold, and LOOCV.

3. *mtcars dataset*: Consider the `mtcars` dataset from 1974 Motor Trend US magazine, that comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles. Load the dataset using

   ```
   data(mtcars)
   ```

   There are 10 covariates in the dataset, and `mpg` (miles per gallon) is the response variable. Fit a ridge regression model for this dataset and find an optimal λ using 1-fold, 5-fold, and LOOCV cross-validation. Choose the best $\lambda \in \{10^{-8}, 10^{-7.5}, \ldots, 10^{7.5}, 10^{8}\}$. Make sure you make the X matrix such that the first column is a column of 1s.

4. *Old faithful:* Fit a mixture of Gaussians for the old faithful `waiting` times. Use cross-validation to choose the appropriate number of classes $C = 2, 3, 4$. Use cross-validation, AIC, and BIC for comparisons.

5. *Old faithful:* Repeat the previous exercise for the bivariate data fitting for $C = 2, 3, 4$.

6. *Seeds dataset:* Download the seeds dataset from

   `https://archive.ics.uci.edu/ml/datasets/seeds`

   This dataset contains information about *three* varieties of wheat (last column of the dataset). There are 7 covariate information. Fit a 7-dimensional Gaussian mixture model algorithm with $C = 3$ and estimate the mis-classification rate using cross-validation.

7. *Seeds dataset:* For the same dataset, with $C = 3$, use cross-validation to find out which of the 7 covariates best helps identify between the three kinds of wheat.

8. Generate n observations from a Normal distribution with mean μ and variance σ^2. Use code below:

   ```
   set.seed(1)
   mu <- 5
   sig2 <- 3
   n <- 100
   my.samp <- rnorm(n, mean = mu, sd = sqrt(sig2))
   ```

   Construct bootstrap confidence intervals for estimating the mean of a normal distribution using both parameteric and nonparametric bootstrap methods and compare the confidence intervals with the usual normal distribution confidence intervals for μ.

9. Repeat the previous exercise for estimating the mean of t_{10} distribution from sample of size 50. Are the bootstrap confidence intervals similar to the intervals using CLT? Why or why not?

10. Repeat again to estimate the mean of a Gamma$(.05, 1)$ distribution from a sample of size $n = 50$. Are the bootstrap confidence intervals similar to the intervals using CLT? Why or why not?

11. For Exercise 1, fit a bridge regression model with $\lambda = 5$, $\alpha = 1.5$, and construct 95% parametric and nonparametric bootstrap confidence intervals for of the 50 βs. In repeated simulations, what is the coverage probability of each the confidence intervals. What percentage of the confidence intervals contain the true vector of β, `beta.star`?

12. Obtain 95% bootstrap confidence intervals for each ridge regression coefficient β for the chosen λ value in the *mtcars* Exercise 3.

Bayesian Methods

1. Suppose $\pi(x) = c\tilde{\pi}(x)$ and $g(x) = c'\tilde{g}(x)$ where *both* c and c' are unknown. Suppose there exists a finite M such that

$$\sup_x \frac{\tilde{\pi}(x)}{\tilde{g}(x)} \leq M\,.$$

   Prove that the accept-reject algorithm that accepts a proposal with probability $\dfrac{\tilde{\pi}}{M\tilde{g}}$ provides a draw from π.

2. Suppose $Y|\mu \sim N(\mu, \sigma^2)$ where σ^2 is known, and consider the prior on μ to be $\mu \sim N(m_0, s_0^2)$. What is the posterior distribution of μ and what is the posterior mean? Write the posterior mean as a linear combination of the data mean and the prior mean.

3. Suppose $Y_1, \ldots, Y_n|\lambda \overset{iid}{\sim} \text{Poisson}(\lambda)$ and prior $\lambda \sim \text{Gamma}(\alpha, \beta)$. What is the posterior distribution of λ?

4. Suppose $Y_1, \ldots, Y_n \mid \mu, \nu \overset{ind}{\sim} t_\nu(\mu)$, which is a t distribution with ν degrees of freedom with mean μ. In addition, assume the priors

$$\mu \sim N(m_0, s_0^2) \text{ and } \nu \sim \text{Truncated Gamma}(a_0, b_0, (2, \infty))\,,$$

   What is the posterior distribution of μ?

   Implement a MH algorithm for sampling from this posterior distribution. Generate your own data with $n = 50$ and $\nu_0 = 3$. How different is the posterior distribution with $\nu_0 = 500$?

5. Suppose $Y_1, \ldots, Y_n \mid \mu \overset{ind}{\sim} t_\nu(\mu)$, which is a t distribution with ν degrees of freedom with mean μ. In addition, assume the prior

$$\mu \sim N(0, 1)\,.$$

   Write the posterior distribution for this problem, and implement the A-R algorithm for different values of μ and n.

6. Consider a Bayesian reliability model, where the observed failure times of a lamp is distributed as

$$T_1, \ldots, T_n \mid \lambda, \beta \sim \text{Weibull}(\lambda, \beta)\,.$$

We assume priors

$$\beta \sim \text{Gamma}(a_0, b_0) \quad \text{and } \lambda \sim \text{Gamma}(a_1, b_1)\,.$$

Now consider the following LCD projector data. To test the manufacturers claim of expected lamp life in an LCD projector being 1500 hours, identical lamps were placed in 31 projectors for various models and their time to failure was recorded. The following were the times to failure:

```
proj.hour <- c(387,182,244,600,627,332,418,300,798,584,
660,39,274,174,50, 34,1895,158,974,345,1755,1752,473,81,954,
1407,230,464,380,131,1205)
```

(a) For $a_0 = 1$, $b_0 = 1$, $a_1 = 2.5$ and $b_1 = 2350$, implement a Metropolis-Hastings algorithm to draw from the joint posterior distribution.

7. Suppose $Y_1, \ldots, Y_n \mid \mu \sim N(\mu, 1)$ and assume the prior

$$\mu \sim t_{\nu_0}$$

where ν_0 is the degrees of freedom where the truncated Gamma is has support $(2, \infty)$.

(a) What is the posterior distribution μ, ν?

(b) Write a MH algorithm to sample from the above posterior distribution.

(c) Generate $n = 100$ data points. Set $m_0 = 0$, $s_0^2 = 1$, $a_0 = 2$, $b_0 = .1$, and run the MH algorithm described above.

8. Implement a MH algorithm to sample from a p-dimensional sphere in Problem 15 (b) in the Sampling section (above). For $p = 5$, is A-R better than MH in terms quality of estimation per unit second? For $p = 20$, which is better?

9. Implement a MH algorithm to sample from a multivariate truncated normal distribution in Problem 16 (b) in the Sampling section (above). For $p = 5$, is A-R better than MH in terms quality of estimation per unit second? For $p = 20$, which is better?

10. (Bayesian linear regression): Consider the Bayesian linear regression model. The likelihood is

$$y_1, \ldots, y_n \mid \beta, \sigma^2 \overset{iid}{\sim} N(X_i\beta, \sigma^2)\,.$$

The parameters of interest are β and σ^2, just like regular MLE. We assume priors:

$$\beta \sim N_p(\mu, \sigma^2) \quad \text{and} \quad \sigma^2 \sim \text{IG}(a, b)\,,$$

where a, b, and μ are hyperparameters that need to be chosen according to the dataset.

(a) What is the posterior distribution of (β, σ^2)?

(b) Implement a MH algorithm to sample from the posterior distribution from the `fuel2001` dataset in the `alr3` package.

11. Suppose $X_1, \ldots, X_N \overset{iid}{\sim} F$ and $Y_1, \ldots, Y_N$ is a Markov chain with F as the stationary distribution. Consider two estimates of the mean of F:

$$\bar{X}_N := \frac{1}{N}\sum_{t=1}^{N} X_t \quad \text{and} \quad \bar{Y}_N := \frac{1}{N}\sum_{t=1}^{N} Y_t\,.$$

Which estimator is better? In other words, which estimator has smaller variance?