Statistical Computation

Derek Li

Contents

1 Review

1.1 Complex Number

Definition 1.1. A complex number z consists of two component, real and imaginary:

$$z = x + \iota y$$

where $\iota = \sqrt{-1}$.

Property 1.1. If $z_1 = x_1 + \iota y_1, z_2 = x_2 + \iota y_2$ then

$$\begin{aligned} z_1 + z_2 &= (x_1 + x_2) + \iota(y_1 + y_2) \\ z_1 z_2 &= (x_1 x_2 - y_1 y_2) + \iota(x_1 y_2 + x_2 y_1) \end{aligned}$$

Property 1.2. $\exp(\iota\theta) = \cos(\theta) + \iota \sin(\theta)$.

Property 1.3. $z = x + \iota y = r \exp(\iota\theta)$ where $r = |z| = \sqrt{x^2 + y^2}, x = r\cos(\theta), y = r\sin(\theta)$.

Property 1.4. $\exp(\iota(\theta_1 + \theta_2)) = \cos(\theta_1 + \theta_2) + \iota \sin(\theta_1 + \theta_2)$.

1.2 Markov Chain

Definition 1.2 (Transition Density)**.** Transition density $q(\mathbf{x}, \mathbf{y})$ is the conditional density of $\mathbf{X}_i$ given $\mathbf{X}_{i-1} = \mathbf{x}$, i.e.,

$$q(\mathbf{x}, \mathbf{y}) = q(\mathbf{x} \to \mathbf{y}) = q(\mathbf{y}|\mathbf{x})$$

Definition 1.3 (Stationary)**.** The Markov chain is stationary with stationary (invariant) density $f(\mathbf{x})$ if

$$f(\mathbf{y}) = \int \cdots \int q(\mathbf{x}, \mathbf{y}) f(\mathbf{x}) \mathrm{d}\mathbf{x}$$

Definition 1.4 (Reversibility)**.** A transition density $q(\mathbf{x}, \mathbf{y})$ will have $f(\mathbf{x})$ as its stationary density if we have

$$f(\mathbf{x})q(\mathbf{x}, \mathbf{y}) = f(\mathbf{y})q(\mathbf{y}, \mathbf{x}) \text{ (Reversibility condition)}$$

1.3 Linear Algebra

Definition 1.5 (Lower Triangular Matrix)**.** We define lower triangular matrix to be

$$L = \begin{pmatrix} a_{11} & 0 & 0 & \cdots & 0 \\ a_{21} & a_{22} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{pmatrix}$$

where $a_{ij} = 0$ for $i < j$.

Definition 1.6 (Upper Triangular Matrix)**.** We define upper triangular matrix to be

$$U = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ 0 & a_{22} & a_{23} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & a_{nn} \end{pmatrix}$$

where $a_{ij} = 0$ for $i > j$.

Note. For upper and lower triangular matrices, A^{-1} exists iff $a_{11}, \cdots, a_{nn}$ are all non-zero.

Example 1.1. Suppose we solve $A\mathbf{x} = \mathbf{b}$ for lower triangular matrix A, where $\mathbf{b} = (b_1 \quad \cdots \quad b_n)^T$ and $\mathbf{x} = (x_1 \quad \cdots \quad x_n)^T$ Then

$$\begin{aligned} x_1 &= \frac{b_1}{a_{11}} \\ x_2 &= \frac{b_2 - a_{21}x_1}{a_{22}} \\ x_3 &= \frac{b_3 - a_{31}x_1 - a_{32}x_2}{a_{33}} \\ &\vdots \\ x_n &= \frac{b_n - a_{n1}x_1 - a_{n2}x_2 - \cdots - a_{n,n-1}x_{n-1}}{a_{nn}} \end{aligned}$$

Note 1. The algorithm for upper triangular matrix is similar.
Note 2. In R, we use `backsolve` and `forwardsolve`.

Definition 1.7 (Norm)**.** Suppose $\mathbf{x} = (x_1 \quad \cdots \quad x_n)^T$ is a vector. A norm $\|\mathbf{x}\|$ satisfies the following conditions:

1. $\|\mathbf{x} + \mathbf{y}\| \leqslant \|\mathbf{x}\| + \|\mathbf{y}\|$.
2. $\|a\mathbf{x}\| = |a|\|\mathbf{x}\|$.
3. $\|\mathbf{x}\| = -$ implies $\mathbf{x} = \mathbf{0}$.

Note. $\|\mathbf{x}\|$ gives a measure of the size or length of $\mathbf{x}$.

Example 1.2 (General L_p Norm)**.**

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}, p \geqslant 1$$

Example 1.3 (Euclidean Norm)**.**

$$\|\mathbf{x}\|_2 = \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2}$$

Example 1.4 (Manhattan Distance)**.**

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$$

Example 1.5 (L_∞ Norm)**.**

$$\|\mathbf{x}\|_\infty = \max_{1 \leqslant i \leqslant n} |x_i|$$

Definition 1.8 (Frobenius Norm)**.** The Frobenius norm of an $m \times n$ matrix is

$$\|A\|_F = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2 \right)^{1/2}$$

Definition 1.9 (L_p Norm for Matrix)**.** Suppose A is an $m \times n$ matrix. We define the L_p norm of A as

$$\|A\|_p = \sup_{\mathbf{x}} \frac{\|A\mathbf{x}\|_p}{\|\mathbf{x}\|_p} = \sup_{\mathbf{x}:\|\mathbf{x}\|_p = 1} \|A\mathbf{x}\|_p$$

Property 1.5.

1. $\|AB\|_p \leqslant \|A\|_p \|B\|_p$.
2. $\|A^k\|_p \leqslant \|A\|_p^k$.
3. $\|I\|_p = \|AA^{-1}\|_p = 1 \Rightarrow \|A^{-1}\|_p \geqslant \dfrac{1}{\|A\|_p}$.
4. For any vector $\mathbf{x}$, $\|A\mathbf{x}\|_p \leqslant \|A\|_p \|\mathbf{x}\|_p$.
5. When $p = 2$,
$$\|A\|_2 = \sqrt{\text{Maximum eigenvalue of } A^T A}$$

Note. If A is symmetric then $\|A\|_2$ is the maximum absolute eigenvalue of A.

Example 1.6 (L_∞ Norm for Matrix)**.** Consider vectors $\mathbf{x}$ whose elements are all ± 1, i.e., $\|\mathbf{x}\|_\infty = 1$. We maximize $\|A\mathbf{x}\|_\infty$ by taking $\mathbf{x}$ so that $\sum_{j=1}^{n} a_{ij}x_j$ is maximized and

$$\|A\|_\infty = \max_{1 \leqslant i \leqslant m} \sum_{j=1}^{n} |a_{ij}|$$

Example 1.7 (L_1 Norm for Matrix)**.** Consider vectors $\mathbf{x}$ whose elements are one 1 and $(n-1)$ 0s, i.e., $\|\mathbf{x}\|_1 = 1$. $A\mathbf{x}$ picks out one column of A and

$$\|A\|_1 = \max_{1 \leqslant j \leqslant n} \sum_{i=1}^{m} |a_{ij}|$$

Theorem 1.1. Suppose A is an $n \times n$ matrix with real-valued eigenvalues $\lambda_1, \cdots, \lambda_n$. Then

$$\max_{1 \leqslant k \leqslant n} |\lambda_k| \leqslant \|A\|_p, p \geqslant 1$$

Proof. Suppose that $A\mathbf{v} = \lambda_k \mathbf{v}$ where $\|\mathbf{v}\|_p = 1$. Then for all $k = 1, \cdots, n$,

$$|\lambda_k| = \|A\mathbf{v}\|_p \leqslant \|A\|_p$$

□

Note. The result holds if there are complex-valued eigenvalues where $|\lambda_k|$ is the modulus if λ_k is complex-valued.

Definition 1.10 (Diagonally Dominant Matrix)**.** An $n \times n$ matrix A is (row) diagonally dominant if

$$|a_{ii}| \geqslant \sum_{j \neq i} |a_{ij}|, i = 1, \cdots, n$$

If $\geqslant$ is replaced by $>$ then A is strictly diagonally dominant.

Property 1.6. A strictly diagonally dominant matrix is invertible.

Theorem 1.2 (Gershgorin Circle Theorem)**.** Suppose A is an $n \times n$ matrix with elements $\{a_{ij}\}$. Define $r_i = \sum_{j \neq i} |a_{ij}|$ and

$$C_i = \{z \in \mathbb{C} : |z - a_{ii} \leqslant r_i|\}$$

which is a circle on the complex plane centered at a_{ii} with radius r_i. If λ is an eigenvalue of A, then $\lambda \in C_i$ for some i.

1.3.1 Application: Condition Numbers

Theorem 1.3. If $\|B\|_p < 1$ for some p then

$$(I - B)^{-1} = I + B + B^2 + \cdots = \sum_{k=0}^{\infty} B^k$$

We use $(A + E)^{-1} - A^{-1}$ to check the sensitivity of A^{-1} to round-off error, where E is small relative to A.

We have

$$\begin{aligned}(A + E)^{-1} &= A^{-1}(I + EA^{-1})^{-1} \\ &= A^{-1}[I - EA^{-1} + (EA^{-1})^2 + \cdots] \\ &\approx A^{-1} - A^{-1}EA^{-1}\end{aligned}$$

and thus

$$(A + E)^{-1} - A^{-1} \approx -A^{-1}EA^{-1}$$

Then

$$\|(A + E)^{-1} - A^{-1}\|_p \approx \|A^{-1}EA^{-1}\|_p \leqslant \|A^{-1}\|_p^2 \|E\|_p$$

and

$$\frac{\|(A + E)^{-1} - A^{-1}\|_p}{\|A^{-1}\|_p} \leqslant \|A^{-1}\|_p \|E\|_p = \kappa_p(A) \frac{\|E\|_p}{\|A\|_p}$$

where $\kappa_p(A) = \|A\|_p \|A^{-1}\|_p \geqslant 1$ is called the condition number of A. If the condition number of A is large then the numerical solution of $A\mathbf{x} = \mathbf{b}$ may be unstable.

Example 1.8. Let

$$A = \begin{pmatrix} 1 & 1 - \varepsilon \\ 1 + \varepsilon & 1 \end{pmatrix}$$

then

$$A^{-1} = \begin{pmatrix} \varepsilon^{-2} & \varepsilon^{-1} - \varepsilon^{-2} \\ -\varepsilon^{-1} - \varepsilon^{-2} & \varepsilon^{-2} \end{pmatrix}$$

for $\varepsilon > 0$ and small. We have $\|A\|_1 = \|A\|_\infty = 2 + \varepsilon$ and $\|A^{-1}\|_1 = \|A^{-1}\|_\infty = 2\varepsilon^{-2} + \varepsilon^{-1}$. Thus

$$\kappa(A) = 4\varepsilon^{-2} + 4\varepsilon^{-1} + 1 \approx 4\varepsilon^{-2}$$

for small ε.

2 Basics

2.1 Floating Point

2.1.1 Floating Point Representation

Definition 2.1. A ***floating point number*** is represented by three components: (S, F, E) where S is the sign of the number (± 1), F is a fraction (lying between 0 and 1), E is an exponent. S, F, E are all represented as binary digits (bits). The ***floating point representation*** of x, $\text{fl}(x)$ is

$$\text{fl}(x) = S \times F \times 2^E$$

Note. x and $\text{fl}(x)$ need not be the same, since $\text{fl}(x)$ is a binary approximation to x, and there are only a finite number of floating point numbers.

2.1.2 Round-Off Error

Mathematical operations introduce further approximation errors

$$f(\text{fl}(x)) = f(x + \varepsilon) \approx f(x) + \varepsilon f'(x)$$

and the goal is to make the round-off error $|f(x) - \text{fl}(f(\text{fl}(x)))|$ as small as possible.

2.1.3 Machine Epsilon and Other Constants

For a given real number x, we have

$$|\text{fl}(x) - x| \leqslant U|x| \text{ or } \text{fl}(x) = x(1 + u), |u| \leqslant U$$

where U is ***machine epsilon*** or ***machine unit***. U is machine dependent but very small. In R, $U = 2^{-52} = 2.220 \times 10^{-16}$.

Other machine dependent constants include:

1. The minimum and maximum positive floating point numbers: $x_{\min} = 2^{-1022} = 2.225 \times 10^{-308}$ and $x_{\max} = 2^{1024} - 1 = 1.798 \times 10^{308}$.
2. The maximum integer: $2147383647 = 2^{31} - 1$.

2.1.4 Overflow and Underflow Error

Definition 2.2. If the result of a floating point operation exceeds $x_{\max}$, then the value returned is `Inf`.

Note. `Inf` indicates an ***overflow error***.

Definition 2.3. If the result of a floating point operation is undefined then `NaN` is returned.

Definition 2.4. An ***underflow error*** occurs when the result of a floating point calculation is smaller (in absolute value) than $x_{\min}$.

Note. There are two possible outcomes: an error is reported or an exact 0 is returned. The latter outcome may cause problems in subsequent computations (e.g., division by 0).

Note. There are some ways to avoid overflow and underflow errors:

1. Use logarithmic scale: Changes multiplication/division into addition/subtraction, e.g., `lgamma`, `lfactorial`, `lchoose`.
2. Use series expansions (e.g., Taylor series).

Example 2.1. For x close to 0, $\dfrac{\exp(x)-1}{x} \approx 1$. Naive computation of $\dfrac{\exp(x)-1}{x}$ is problematic for x close to 0 due to possible round-off and underflow errors:

$$\frac{\mathrm{fl}(\exp(x)-1)}{\mathrm{fl}(x)} \neq \frac{\exp(x)-1}{x}$$

We solve the problem by using a series approximation, for $|x| \leqslant \varepsilon$,

$$\frac{\exp(x)-1}{x} = \frac{x + x^2/2 + x^3/6 + \cdots}{x} = 1 + \frac{x}{2} + \frac{x^2}{6} + \cdots$$

2.1.5 Catastrophic Cancellation

Suppose $z_1 = g_1(x_1, \cdots, x_n)$ and $z_2 = g_2(x_1, \cdots, x_n)$. We want to compute $y = z_1 - z_2$. What we actually compute is

$$y^* = \mathrm{fl}(\mathrm{fl}(z_1) - \mathrm{fl}(z_2))$$

where $\mathrm{fl}(z_1) = z_1(1+u_1)$ and $\mathrm{fl}(z_2) = z_2(1+u_2)$. We have

$$\mathrm{fl}(z_1) - \mathrm{fl}(z_2) = \underbrace{z_1 - z_2}_{y} + \underbrace{z_1 u_1 - z_2 u_2}_{\text{error}}$$

If z_1 and z_2 are large but $y = z_1 - z_2$ is small then the magnitude of the error may be larger than the magnitude of y - ***catastrophic cancellation***.

2.2 Sparse Matrices

Definition 2.5. We say an $n \times n$ matrix is ***sparse*** if it has $k \times n$ non-zero elements where $k \ll n$.

Note 1. An $n \times n$ matrix needs at least n non-zero elements to be invertible.

Note 2. Sparse matrices are useful because we need only store non-zero elements and their row and column indices; multiplication by and addition to 0 are free operations.

2.3 Application: Computation of Probability Distributions

Question: Suppose X_i are independent discrete r.v.s. taking values $0, \cdots, l$ with

$$P(X_i = x) = p(x), x = 0, \cdots, l$$

Define $S = X_1 + \cdots + X_n$ and find the probability distribution of S.

2.3.1 Brute Force Approach

Start with $n = 2$ and proceed inductively:

$$p_2(x) := P(X_1 + X_2 = x) = \sum_{y=0}^{x} P(X_1 = y, X_2 = x - y)$$

$$p_3(x) := P(X_1 + X_2 + X_3 = x) = \sum_{y=0}^{x} P(X_1 + X_2 = y, X_3 = x - y)$$

$$\vdots$$

$p_k(x)$ requires $x+1$ multiplications and to evaluate $p_k(x)$ for $x = 0, \cdots, kl$, we need

$$N(k) = \sum_{x=0}^{kl}(x+1) \approx \frac{(kl)^2}{2} \text{ multiplications}$$

Thus the total number of multiplications is

$$\sum_{k=2}^{n} N(k) \approx \frac{n^3 l^2}{6} = O(n^3 l^2)$$

2.3.2 Probability Generating Function

Definition 2.6. If X is a discrete r.v. taking values $0, 1, \cdots$, then its ***probability generating function*** is

$$\phi(t) = \mathbb{E}[t^X] = \sum_{x=0}^{\infty} P(X = x)t^x$$

Note. If X takes values $0, \cdots, l$, then $P(X = x)$ can be recovered from evaluating $\phi(t)$ at $l+1$ distinct (non-zero) points $t_0, \cdots, t_l$.

If $\phi(t) = \mathbb{E}[t^{X_i}]$, then the probability generating function of S is

$$\mathbb{E}[t^S] = \mathbb{E}[t^{X_1+\cdots+X_n}] = [\phi(t)]^n$$

Thus we can recover $P(S = x)$ for $x = 0, \cdots, nl$ by evaluating $[\phi(t)]^n$ at $t_0, \cdots, t_{nl}$. We have $nl+1$ linear equations in $nl+1$ unknowns, and solving typically requires $O(n^3 l^3)$ operations, which is slower than the brute force approach.

2.3.3 Discrete Fourier Transform (DFT)

A choice for $t_0, \cdots, t_{nl}$ are complex exponentials

$$t_j = \exp\left(-2\pi\iota \frac{j}{nl+1}\right), j = 0, \cdots, nl$$

where $\iota = \sqrt{-1}$. Since $p(x) = 0$ for $x = l+1, \cdots, nl$, we have

$$\phi(t_j) = \sum_{x=0}^{l} p(x) \exp\left(-2\pi\iota \frac{jx}{nl+1}\right) = \sum_{x=0}^{nl} p(x) \exp\left(-2\pi\iota \frac{jx}{nl+1}\right)$$

$\phi(t_0), \cdots, \phi(t_{nl})$ is the ***discrete Fourier transform*** (DFT) of $p(0), \cdots, p(nl)$, and thus, the DFT of $P(S = 0), \cdots, P(S = nl)$ is $[\phi(t_0)]^n, \cdots, [\phi(t_{nl})]^n$. Hence, given $\phi(t_0), \cdots, \phi(t_{nl})$, we can compute the probability distribution of S using the inverse DFT:

$$P(S = x) = \frac{1}{nl+1} \sum_{j=0}^{nl} [\phi(t_j)]^n \exp\left(2\pi\iota \frac{jx}{nl+1}\right), x = 0, \cdots, nl$$

Naive computation of $P(S = x)$ using DFT requires $O(n^3 l^2)$ multiplications; but with divide-and-conquer algorithm, we can reduce the number of multiplications by a factor of n.

In R, if `x` is a vector of length n we can compute its DFT with `fft(x)` and the inverse DFT with `fft(tx, inv=T) / length(x)`:

```
probs = # The vector for P(X=x)
dft = fft(probs)
dft.s = dtf^n # S=X1+...+Xn
idft.s = fft(dft.s, inv=T) / length(probs)
Re(idft.s) # Real component of idft.s, or P(S=x)
```

Note. `fft` is the fast Fourier transform, which is an efficient algorithm for computing the DFT when the length of the sequence is a product of small primes.

2.4 Application: Image Processing

Question: We observe an image denoted by $x(i,j).i = 1, \cdots, m, j = 1, \cdots, n$, where (i,j) denotes a pixel location. We want:

1. Denoising: Think of $\{x(i,j)\}$ as a image corrupted by noise

$$x(i,j) = \underbrace{s(i,j)}_{\text{True}} + \underbrace{\varepsilon(i,j)}_{\text{Noise}}$$

2. Compression: Approximate $x(i,j)$ by $x^*(i,j)$ where

$$x^*(i,j) = \sum_{k=1}^{p} \beta_k \phi_k(i,j)$$

where $p \ll m \times n$ and $\phi_1, \cdots, \phi_p$ are known functions.

2.4.1 Transformation

Define X to be the $m \times n$ matrix whose elements are $x(i,j)$. Define orthogonal matrices H_1 $(m \times m)$ and H_2 $(n \times n)$ and define $\widehat{X} = H_1 X H_2$, which has the same dimensions as X. Since for orthogonal matrix H, $H^{-1} = H^T$ and so $X = H_1^T \widehat{X} H_2^T$. Assume the noisy image model $X = S + E$, if H_1 and H_2 are chosen appropriately,

$$\widehat{X} = \underbrace{H_1 S H_2}_{\text{Sparse}} + \underbrace{H_1 E H_2}_{\approx 0}$$

Therefore,

1. Denoising: Given $\widehat{X}$, find a transformation $\widehat{X} \mapsto T(\widehat{X})$ and define the denoised image

$$X_{\text{dn}} = H_1^T T(\widehat{X}) H_2^T$$

where we assume the smallest elements of $\widehat{X}$ are due to noise and set these equal to 0

$$T(\widehat{X})(i,j) = 0, |\widehat{X}(i,j)| \leqslant \text{Threshold}$$

2. Compression: The same idea is used for compression: for some T,

$$X_{\text{c}} = H_1^T T(\widehat{X}) H_2^T$$

Note. T is usually defined more deterministically. The form of T depends on the amount of compression and the type of image.

2.4.2 Hadamard Matrices and Walsh-Hadamard Transform

Definition 2.7. A ***Hadamard matrix*** is an $n \times n$ matrix whose elements are all ± 1 with orthogonal rows s.t. $HH^T = nI$.

Note 1. $H^{-1} = \dfrac{H^T}{n}$.

Note 2. Hadamard matrices only exist if $n = 1, n = 2$, or n is a multiple of 4.

Note 3. We focus on the case where $n = 2^k$ since it is simple to construct and we can write the Hadamard matrix as a product of sparse matrices. We start with the trivial 1×1 Hadamard matrix $H_1 = 1$, and then define $H_2, H_4, H_8, \cdots$ recursively:

$$H_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$
$$H_{2^k} = \begin{pmatrix} H_{2^{k-1}} & H_{2^{k-1}} \\ H_{2^{k-1}} & -H_{2^{k-1}} \end{pmatrix}$$

for $k = 2, 3, \cdots$.

Note 4. H_2 is symmetric and so H_{2^k} is symmetric and thus $H_{2^k}^{-1} = \dfrac{H_{2^k}}{2^k}$.

Definition 2.8. Given arbitrary matrices A and B, the ***Kronecker product*** $A \otimes B$ is

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{pmatrix}$$

for an $m \times n$ matrix A.

Property 2.1. Assume below that any matrix sums, products or inverses are well-defined.

1. $A \otimes (B + C) = (A \otimes B) + (A \otimes C)$.
2. $(B + C) \otimes A = (B \otimes A) + (C \otimes A)$.
3. $A \otimes (B \otimes C) = (A \otimes B) \otimes C$.
4. $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$.
5. $(A \otimes B)^T = A^T \otimes B^T$.
6. $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$.

Note. For Hadamard matrices, $H_{2^k} = H_2 \otimes H_{2^{k-1}}$. We rewrite it as $H_{2^k} = (H_2 I_2) \otimes (I_{2^{k-1}} H_{2^{k-1}})$ and using the property, we have

$$H_{2^k} = (H_2 \otimes I_{2^{k-1}})(I_2 \otimes H_{2^{k-1}})$$

Repeating the process with $H_{2^{k-1}}, H_{2^{k-2}}, \cdots$, we get

$$H_{2^k} = \underbrace{(H_2 \otimes I_{2^{k-1}})(I_2 \otimes H_2 \otimes I_{2^{k-2}})(I_4 \otimes H_2 \otimes I_{2^{k-3}}) \cdots (I_{2^{k-1}} \otimes H_2)}_{k=\log_2(n) \text{ terms}}$$

Definition 2.9. Given an $n \times n$ Hadamard matrix H and a vector $\mathbf{x}$ of length n, we define its ***Walsh-Hadamard transform*** by $\hat{\mathbf{x}} = H\mathbf{x}$.

Note 1. Given the W-H transform, we can recover $\mathbf{x}$

$$\mathbf{x} = \frac{1}{n} H^T \hat{\mathbf{x}}$$

Note 2. If $n = 2^k$, since $H = H^T$, then

$$\mathbf{x} = \frac{1}{n} H \hat{\mathbf{x}}$$

2.5 Application: Denoising

Question: Suppose we observe $\mathbf{x} = (x_1, \cdots, x_n)^T$ where we assume that

$$\mathbf{x} = \mathbf{s} + \mathbf{e} = \text{Signal} + \text{Noise}$$

We want to recover or estimate the signal $\mathbf{s}$.

2.5.1 Assumption

Assume $\mathbf{s}$ is structured so that its W-H transform $\widehat{\mathbf{s}} = H\mathbf{s}$ contains mostly 0s

$$\widehat{\mathbf{x}} = H\mathbf{x} = \underbrace{H\mathbf{s}}_{\text{Sparse}} + \underbrace{H\mathbf{e}}_{\text{Relatively small}}$$

2.5.2 Thresholding

We shrink smaller components of $\widehat{\mathbf{x}}$ towards 0, and then estimate $\mathbf{s}$ by the inverse W-H transform of the thresholded $\widehat{\mathbf{x}}$. Thresholded W-H transform $\widehat{\mathbf{x}}_s$ is an estimate of the W-H transform of $\mathbf{s}$, and thus we can estimate $\mathbf{s}$ by the inverse W-H transform

$$\widetilde{\mathbf{s}} = \frac{1}{n} H^T \widehat{\mathbf{x}}_s$$

Define thresholds $\lambda_1, \cdots, \lambda_n \geqslant 0$. The ***hard thresholding*** is to modify $\widehat{\mathbf{x}}$ as follows:

$$\widehat{\mathbf{x}}_s = \begin{pmatrix} \widehat{x}_1 I(|\widehat{x}_1| \geqslant \lambda_1) \\ \vdots \\ \widehat{x}_n I(|\widehat{x}_n| \geqslant \lambda_n) \end{pmatrix}$$

The ***soft thresholding*** is to modify $\widehat{\mathbf{x}}$ as follows:

$$\widehat{\mathbf{x}}_s = \begin{pmatrix} \text{sgn}(\widehat{x}_1)(|\widehat{x}_1| - \lambda_1)_+ \\ \vdots \\ \text{sgn}(\widehat{x}_n)(|\widehat{x}_n| - \lambda_n)_+ \end{pmatrix}$$

where $\text{sgn}(y)$ is the sign of y, and y_+ equals y if $y > 0$ and 0 if $y \leqslant 0$.

Typically we set $\lambda_1 = 0$, and use knowledge of the problem to decide $\lambda_2, \cdots, \lambda_n$; or take $\lambda_2 = \cdots = \lambda_n$ and choose the common value based on tools such as half normal plots.

2.5.3 The Fast W-H Transform

A Hadamard matrix H consists of ± 1 so computation of $H\mathbf{x}$ involves only additions and subtractions, but naive computation involves $n(n-1) = O(n^2)$ additions and subtractions, which is less than ideal if n is very large. We can write H as a product of sparse matrices to reduce complexity.

Example 2.2 ($n = 2^3 = 8$)**.** The 8×8 Hadamard matrix is

$$H_8 = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{pmatrix}$$

Naive computation of $H_8\mathbf{x}$ needs 56 additions and subtractions. But if $H_8 = A^3$ where

$$A = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix}$$

Computation of $AAA\mathbf{x}$ needs $3 \times 8 = 24$ additions and subtractions.

2.5.4 R code for FWHT

The function `fwht` below computes the W-H transform of data in a vector `x`.

```
fwht = function(x) {
  h=1
  len = length(x)
  while (h < len) {
    for (i in seq(1, len, by=h*2)) {
      for (j in seq(i, i+h-1)) {
        a = x[j]
        b = x[j+h]
        x[j] = a + b
        x[j+h] = a - b
      }
    }
  h = 2 * h
  }
  x
}
```

We can compute the inverse W-H transform using `fwht` by dividing the output by the length of the vector.

2.6 Fast Fourier Transform (FFT)

Definition 2.10 (Discrete Fourier Transform)**.** Suppose we have data $x_0, \cdots, x_{n-1}$, and define $\widehat{x}_0, \cdots, \widehat{x}_{n-1}$ by

$$\widehat{x}_j = \sum_{t=0}^{n-1} \exp\left(-2\pi\iota \frac{j}{n} t\right) x_t$$

where $\iota = \sqrt{-1}$.

Property 2.2 (Inverse DFT)**.** Given DFT, recover the original sequence by

$$x_t = \frac{1}{n} \sum_{j=0}^{n-1} \exp\left(2\pi\iota \frac{j}{n} t\right) \widehat{x}_j$$

Proof. For complex numbers z,

$$\sum_{j=0}^{n-1} z^j = \begin{cases} n, & z = 1 \\ \dfrac{1-z^n}{1-z}, & \text{otherwise} \end{cases}$$

Thus if $z = \exp\left(\dfrac{2\pi\iota t}{n}\right)$ for an integer t. we have

$$\sum_{j=0}^{n-1} z^j = \sum_{j=0}^{n-1} \exp\left(2\pi\iota \frac{t}{n} j\right) = \frac{1-\exp(2\pi\iota t)}{1-\exp(2\pi\iota t/n)} = 0$$

since $\exp(2\pi\iota t) = \cos(2\pi t) + \iota \sin(2\pi t) = 1$. Hence,

$$\begin{aligned} \frac{1}{n}\sum_{j=0}^{n-1} \exp\left(2\pi\iota \frac{j}{n} t\right) \widehat{x}_j &= \frac{1}{n}\sum_{j=0}^{n-1}\sum_{s=0}^{n-1} \exp\left(2\pi\iota \frac{t-s}{n} j\right) x_s \\ &= \frac{1}{n}\sum_{s=0}^{n-1} x_s \sum_{j=0}^{n-1} \exp\left(2\pi\iota \frac{t-s}{n} j\right) \\ &= x_t \end{aligned}$$

since

$$\sum_{j=0}^{n-1} \exp\left(2\pi\iota \frac{t-s}{n} j\right) = \begin{cases} n, & s = t \\ 0, & s \neq t \end{cases}$$

□

Definition 2.11 (Matrix Formulation of DFT)**.** Define $\mathbf{x} = (x_0, \cdots, x_{n-1})^T$ and $\widehat{\mathbf{x}} = (\widehat{x}_0, \cdots, \widehat{x}_{n-1})^T$. Then

$$\widehat{\mathbf{x}} = F\mathbf{x}$$

where F is an $n \times n$ matrix whose jth row and kth column is

$$f_{jk} = \exp\left(-2\pi\iota \frac{(j-1)(k-1)}{n}\right)$$

The elements of F^{-1} are

$$\overline{f}_{jk} = \frac{1}{n}\exp\left(2\pi\iota \frac{(j-1)(k-1)}{n}\right)$$

Note 1. Using the matrix form directly, we need $O(n^2)$ additions and multiplications to compute the DFT (and its inverse).

Note 2. We can write F as a product of sparse matrices, but unlike the W-H transform, factorization of the DFT matrix is more complicated.

2.6.1 FFT Derivation

Assume n is a product of prime numbers $n_1, \cdots, n_k : n = n_1 \times \cdots \times n_k$.

2.6.1.1 Case I: Even Number and Product of Small Prime Numbers

Assume n is even, then

$$\begin{aligned}\widehat{x}_j &= \sum_{t=0}^{n/2-1} \exp\left(-2\pi\iota\frac{j}{n}2t\right) x_{2t} + \sum_{t=0}^{n/2-1} \exp\left(-2\pi\iota\frac{j}{n}(2t+1)\right) x_{2t+1} \\ &= \underbrace{\sum_{t=0}^{n/2-1} \exp\left(-2\pi\iota\frac{j}{n/2}t\right) x_{2t}}_{\text{DFT of } x_0,x_2,\cdots} + \exp\left(-2\pi\iota\frac{j}{n}\right) \underbrace{\sum_{t=0}^{n/2-1} \exp\left(-2\pi\iota\frac{j}{n/2}t\right) x_{2t+1}}_{\text{DFT of } x_1,x_3,\cdots}\end{aligned}$$

Hence, the DFT of $x_0, \cdots, x_{n-1}$ is a linear combination of the DFT of the even and odd indices. Our rearrangement into DFT of odd and even indices can be written in matrix form as

$$\widehat{\mathbf{x}} = \underbrace{\begin{pmatrix} I & \Omega \\ I & -\Omega \end{pmatrix}}_{\text{Sparse}} \underbrace{\begin{pmatrix} F_{n/2} & 0 \\ 0 & F_{n/2} \end{pmatrix}}_{\text{Sparser}} P\mathbf{x}$$

where Ω is a diagonal matrix (sparse) and P is a permutation matrix (sparse), i.e., if n is even, we can write F as a product of two sparse matrices and a matrix that is sparser than F ($n^2/2$ 0s).

If $n/2$ is divisible by a prime number n', we can perform a similar decomposition of $F_{n/2}$ and F is now the product of sparser matrices. When $n_1, \cdots, n_k$ are small then we need $O(n\ln(n))$ additions and multiplications.

2.6.1.2 Case II: Prime Number with Zero-Padding

Definition 2.12 (Zero Padding)**.** Add 0s to the end of the sequence so that the length of the ***zero padded*** sequence is a product of small prime numbers:

$$x_0, \cdots, x_{n-1}, \underbrace{0, \cdots, 0}_{m}$$

with $n + m = n_1 \times \cdots \times n_k$ where $n_1, \cdots, n_k$ are small primes.

Note 1. The function `nextn` is useful for zero-padding.

Note 2. Adding 0s to a sequence changes the nature of the sequence - creating a large discontinuity, which is reflected in the DFT.

2.6.2 Analysis of DFT Approach

For the application in computation of probability distributions with DFT approach, we take $m \geqslant nl = 1$ where m is a product of small prime numbers, and follow the steps:

1. Define $\widehat{p}_i(0), \cdots, \widehat{p}_i(m-1)$ to be the DFT of $p_i(0), \cdots, p_i(m-1)$ for $i = 1, \cdots, n$.
2. Define
$$\widehat{p}_s(k) = \prod_{i=1}^{n} \widehat{p}_i(k), k = 0, \cdots, m-1$$
3. Inverse DFT: $P(S=0), \cdots, P(S=m-1)$ is the inverse DFT of $\widehat{p}_s(0), \cdots, \widehat{p}_s(m-1)$.

The number of multiplications at each step is:

1. DFT: $n \times O(m\ln(m)) = O(nm\ln(m))$.
2. Product of DFTs: $O(nm)$.

3. Inverse DFT: $O(m \ln(m))$.

The total number of multiplications is $O(nm \ln(m))$ and thus if $m \approx nl$, the number of multiplications is $O(n^2 l \ln(nl))$ versus $O(n^3 l^2)$ for the brute force algorithm.

3 Generation of Random Variates

3.1 Generation of Random Numbers

Example 3.1 (Importance Sampling)**.** Suppose we want to estimate

$$I = \int \cdots \int g(\mathbf{x}) \mathrm{d}\mathbf{x}$$

for some integrand $g : \mathbb{R}^p \to \mathbb{R}$. If f is a probability density function on $\mathbb{R}^p$, then

$$I = \int \cdots \int g(\mathbf{x}) \mathrm{d}\mathbf{x} = \int \cdots \int \frac{g(\mathbf{x})}{f(\mathbf{x})} f(\mathbf{x}) \mathrm{d}\mathbf{x} = \mathbb{E}_f \left[\frac{g(\mathbf{X})}{f(\mathbf{X})} \right]$$

where $\mathbf{X}$ has a density f. We can use the law of large numbers to estimate the expected value provided $\mathrm{Var}_f \left[\dfrac{g(\mathbf{X})}{f(\mathbf{X})} \right] < \infty$. Take $\mathbf{X}_1, \cdots, \mathbf{X}_n$ independent from f, LLN gives

$$\widehat{I} = \frac{1}{n} \sum_{i=1}^{n} \frac{g(\mathbf{X}_i)}{f(\mathbf{X}_i)} \approx \int \cdots \int g(\mathbf{x}) \mathrm{d}\mathbf{x}$$

Note. We choose f satisfying precision and expediency:

1. Precision: Minimize the variance of $\widehat{I}$.
2. Expediency: Be able to sample from f.

Example 3.2 (Monte Carlo Estimation of π)**.** If X and Y are independent Unif$(-1, 1)$ r.v.s., then

$$P(X^2 + Y^2 \leqslant 1) = \frac{\pi}{4}$$

We generate independent pairs and have

$$\widehat{\pi} = \frac{4}{n} \sum_{i=1}^{n} I(X_i^2 + Y_i^2 \leqslant 1)$$

3.2 Generation of Unif$(0, 1)$

To generate pseudo-random $U_1, U_2, \cdots$, we generate integers $V_1, V_2, \cdots$ from a uniform distribution on $\{1, \cdots, N\}$ and define $U_i = \dfrac{V_i}{N+1}$ for $i = 1, 2, \cdots$. Note that $U_1, U_2, \cdots$ are uniform on the set $\{1/(N+1), \cdots, N/(N+1)\}$. If N is large enough, $U_1, U_2, \cdots$ are independent Unif$(0, 1)$ r.v.s.:

$$\sup_{0 \leqslant x \leqslant 1} |P(U_i \leqslant x) - x| \leqslant \frac{1}{N}$$

3.2.1 Linear Congruential RNG

Define $V_1, V_2, \cdots$ via the recursion:

$$V_{k+1} = (aV_k + b) \mod m$$

for some integers a, b, and m.

Note 1. The initial value V_0 is the ***seed*** of the RNG.

Note 2. $V_1, V_2, \cdots$ take values in the set $\{0, \cdots, m-1\}$.

Note 3. If $b = 0$ then we have a ***multiplicative congruential*** RNG.

Note 4. We have $V_{k+p} = V_k$ for some $p \leqslant m$, and p is the ***period*** of the RNG.

Property 3.1. If $b = 0$, then the maximum possible period is $m - 1$. Furthermore, if m is prime, and

$$a^{(m-1)/q} \mod m \neq 1$$

for every prime factor q of $m - 1$ then the RNG has period $m - 1$.

Example 3.3. Take $m = 5$ and $m - 1 = 4$ has a single prime factor 2. We need $a^2 \mod 5 \neq 1$ so we can take $a = 3$ (for example).

Example 3.4. Let m to be the largest possible prime number $m = 2^{31} - 1$. We can take $a = 16807$ or 48271, or 397204094.

3.2.2 Combining Unif$(0, 1)$ RNGs

Combination increases the period of the RNG.

Example 3.5 (Wichmann-Hill RNG)**.** Combine three multiplicative congruential RNGs:

$$\begin{aligned} V_{k+1}^{(1)} &= 171 V_k^{(1)} \mod 30269 \\ V_{k+1}^{(2)} &= 172 V_k^{(2)} \mod 30307 \\ V_{k+1}^{(3)} &= 170 V_k^{(3)} \mod 30323 \end{aligned}$$

where the periods are short ($\approx 3 \times 10^4$). Then

$$U_k = \left(\frac{V_k^{(1)}}{30269} + \frac{V_k^{(2)}}{30307} + \frac{V_k^{(3)}}{30323} \right) \mod 1$$

where the period is

$$p = \frac{30268 \times 30306 \times 30322}{4} = 6.9536 \times 10^{12}$$

3.2.3 Shift Register Method

We use the binary representation of Unif$(0, 1)$. Suppose $Z_1, Z_2, \cdots$ are independent binary r.v.s. with

$$P(Z_k = 0) = P(Z_k = 1) = \frac{1}{2}$$

then

$$U = \sum_{k=1}^{\infty} \frac{Z_k}{2^k} \sim \text{Unif}(0, 1)$$

In practice, we define U as a finite sum

$$U = \sum_{k=1}^{r} \frac{Z_k}{2^k}$$

where r is the number of bits.

We generate $\{Z_k\}$ via ***exclusive-or*** operations for binary variables x and y. We construct $\{Z_k\}$ as follows:

$$Z_k = Z_{k-p} \oplus Z_{k-p+q}, 1 < q < p$$

and

$$U_n = \sum_{k=1}^{r} \frac{Z_{n-s(k)}}{2^k}$$

for some shifts $\{s(k)\}$.

Recall. If Z_1 and Z_2 are independent, and $Z_3 = Z_1 \oplus Z_2$, then Z_3 is independent of Z_1 and Z_2.

Note 1. For the shifts, we need $s(k) - s(k-1) \gg p$.

Note 2. Initialization of shift register RNGs is much complicated since Z_k is a function of Z_{k-p} and Z_{k-p+q} and U_n depends on r values of $\{Z_k\}$.

Note 3. We need a $p \times r$ matrix of binary seeds.

Example 3.6 (Lewis-Payne RNG)**.** $p = 98, q = 27$, and $s(k) = 100p(k-1)$ s.t. $s(k)-s(k-1) = 100p$. The period is $2^{98} - 1$.

Example 3.7 (Mersenne Twister)**.** The period is $2^{19937} - 1$.

3.3 Testing Unif$(0, 1)$ RNGs

We need to check:

1. Uniformity on $[0, 1]$: For $0 \leqslant a < b \leqslant 1$,

$$\frac{1}{n} \sum_{i=1}^{n} I(a \leqslant U_i \leqslant b) \approx b - a$$

2. Uniformity of k-tuples on $[0, 1]^k$: For $A \subset [0, 1]^k$,

$$\binom{n}{k}^{-1} \sum_{(i_1, \cdots, i_k)} I[(U_{i_1}, \cdots, U_{i_k}) \in A] \approx \text{Volume}(A)$$

3. Independence: U_i independent of $U_{i+1}, U_{i+2}, \cdots$.

3.4 RNGs in R

The function `RNGkind` that allows a user to specify the RNG used to generate Unif$(0, 1)$ r.v.s. and the method used to generate normal r.v.s..

3.5 Methods for Continuous Distribution

3.5.1 Inverse Method

Suppose F is a univariate distribution and we want to generate $X \sim F$.

Definition 3.1. For a general univariate distribution function F, we define

$$F^{-1}(t) = \inf\{x : F(x) \geqslant t\}, 0 < t < 1$$

Property 3.2. If F is a univariate distribution function with inverse F^{-1} and $U \sim$ Unif$(0, 1)$, then

$$X = F^{-1}(U) \sim F$$

Proof. We need to show $P(F^{-1}(U) \leqslant x) = F(x)$ or equivalently $[F^{-1}(U) \leqslant x] = [U \leqslant F(x)]$. By definition of F^{-1}, $[U \leqslant F(x)]$ implies $[F^{-1}(U) \leqslant x]$. If $F^{-1}(U) \leqslant x$ then $F(x + \varepsilon) \geqslant U, \forall \varepsilon > 0$. F is right continuous so $[F^{-1}(U) \leqslant x]$ implies $[U \leqslant f(x)]$. □

Example 3.8 (Exponential Distribution)**.** $F(x) = 1 - \exp(-\lambda x)$ for $x \geqslant 0, \lambda > 0$. Solving $F(F^{-1}(t)) = t$ for $F^{-1}(t)$, we have

$$F^{-1}(t) = -\frac{\ln(1-t)}{\lambda}$$

Thus $X = -\dfrac{\ln(1-U)}{\lambda}$ has an exponential distribution. Since $1 - U \sim \text{Unif}(0,1)$ so we define $X = -\dfrac{-\ln(U)}{\lambda}$.

Example 3.9 (Logistic Distribution)**.** $F(x) = \dfrac{\exp(x)}{1+\exp(x)}$. Solving $F(F^{-1}(t)) = t$, we have

$$F^{-1}(t) = \ln\left(\frac{t}{1-t}\right)$$

which is called logit function. Thus $X = \ln\left(\dfrac{U}{1-U}\right)$ has a Logistic distribution.

Example 3.10 (Approximation of Euler's Constant)**.** The Euler's constant is

$$\begin{aligned}\gamma &= \lim_{m\to\infty}\left[\sum_{k=1}^{m}\frac{1}{k} - \ln(m)\right] \\ &= \int_1^\infty \left(\frac{1}{\lfloor x\rfloor} - \frac{1}{x}\right)\mathrm{d}x \\ &= \int_1^\infty x^2\left(\frac{1}{\lfloor x\rfloor} - \frac{1}{x}\right)x^{-2}\mathrm{d}x\end{aligned}$$

where $f(x) = x^{-2}$ is a density function on $[1,\infty)$. If we can sample $X_1, \cdots, X_n$ from $f(x)$, we can estimate γ by

$$\widehat{\gamma} = \frac{1}{n}\sum_{i=1}^{n} X_i^2\left(\frac{1}{\lfloor X_i\rfloor} - \frac{1}{X_i}\right)$$

The distribution function is $F(x) = 1 - x^{-1}$ whose inverse is $F^{-1}(t) = (1-t)^{-1}$. We can use inverse method to sample from $f(x)$.

```
n = 1000000
u = runif(n)
x = 1 / (1 - u)
gammahat = mean(x^2 * (1 / floor(x) - 1 / x))
```

3.5.2 Rejection Sampling

Assume F is continuous with density function f and $F^{-1}(t)$ is not easily computable. Suppose we want to sample X from a density f. We define a proposal density g s.t. $f(x) \leqslant Mg(x)$ for all x and some $M < \infty$. We sample Y from g and $U \sim \text{Unif}(0,1)$ where Y and U are independent, and define $T = \dfrac{f(Y)}{Mg(Y)}$. If $U \leqslant T$, then set $X = Y$; if $U > T$, then reject and repeat until acceptance. The algorithm works: Given independent $Y \sim g$ and $U \sim \text{Unif}(0,1)$,

$$\begin{aligned}P(X \leqslant x) &= P\left(Y \leqslant x \middle| U \leqslant \frac{f(Y)}{Mg(Y)}\right) \\ &= \frac{P(Y \leqslant x, U \leqslant f(Y)M^{-1}g^{-1}(Y))}{P(U \leqslant f(Y)M^{-1}g^{-1}(Y))}\end{aligned}$$

Since $Y \perp U$, the joint density of (Y, U) is

$$h(y,u) = \begin{cases} g(y), & 0 \leqslant u \leqslant 1 \\ 0, & \text{otherwise} \end{cases}$$

Therefore,

$$\begin{aligned} P(\text{Accept})P\left(U \leqslant \frac{f(Y)}{Mg(Y)}\right) &= \int_{-\infty}^{\infty}\int_{0}^{f(y)M^{-1}g^{-1}(y)} g(y)\mathrm{d}u\mathrm{d}y \\ &= \frac{1}{M}\int_{-\infty}^{\infty} f(y)\mathrm{d}y \\ &= \frac{1}{M} \end{aligned}$$

and

$$\begin{aligned} P\left(Y \leqslant x, U \leqslant \frac{f(Y)}{Mg(Y)}\right) &= \int_{-\infty}^{x}\int_{0}^{f(y)M^{-1}g^{-1}(y)} g(y)\mathrm{d}u\mathrm{d}y \\ &= \frac{1}{M}\int_{-\infty}^{x} f(y)\mathrm{d}y \\ &= \frac{P(X \leqslant x)}{M} \end{aligned}$$

Note 1. The probability of acceptance of a given proposal is $\frac{1}{M}$.
Note 2. If f and g are close then M will be close to 1.
Note 3. We can evaluate M by maximizing $\frac{f(x)}{g(x)}$ but we do not need to find the smallest possible M with $f(x) \leqslant Mg(x)$ since rejection sampling will work with a sub-optimal M with a lower probability of acceptance.
Note 4. f and g can be joint density functions or probability mass functions.

Example 3.11 (Half-Normal Distribution with Exponential Proposal)**.** Suppose we want to sample X from a half-normal distribution whose density is

$$f(x) = \frac{2}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right), x \geqslant 0$$

i.e., if $X \sim \mathcal{N}(0,1)$, then $|X| \sim f$. Since X takes values on $[0,\infty)$, a natural proposal distribution is exponential

$$g(y) = \exp(-y), y \geqslant 0$$

since the tails of the exponential are heavier than those of the normal distribution so M should be finite.

To fine M, we need to maximize $\frac{f(x)}{g(x)}$ over $x \geqslant 0$, i.e.,

$$\max \ln[f(x)] - \ln[g(x)]$$

After calculation, we find $\frac{f(x)}{g(x)}$ is maximized at $x = 1$ and

$$M = \frac{f(1)}{g(1)} = 1.315489$$

and the probability of acceptance of a given proposal is

$$\frac{1}{M} = 0.76$$

The code to generate half-normal r.v.s. is:

```
x = NULL
count = 0
total = 0 # Number of proposals generated
while (count < 100) {
   reject = T
   while (reject) {
      y = rexp(1)
      u = runif(1)
      total = total + 1
      if (u <= 2*dnorm(y)/(1.315489*dexp(y))) {
         x = c(x, y)
         count = count + 1
         reject = F
      }
   }
}
```

Example 3.12 (Cauchy Distribution)**.** Suppose we want to sample X from a Cauchy distribution whose density is

$$f(x) = \frac{1}{\pi(1+x^2)}$$

The distribution function is

$$F(x) = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}(x)$$

and

$$F^{-1}(t) = \tan\left[\pi\left(t - \frac{1}{2}\right)\right], 0 < t < 1$$

We can generate r.v.s. from a Cauchy distribution using the inverse method bug floating point evaluation of $\tan(x)$ is not always straightforward.

We can write $f(x)$ as a mixture of two densities

$$f(x) = \frac{1}{2}f_1(x) + \frac{1}{2}f_2(x)$$

where

$$f_1(x) = \frac{2}{\pi(1+x^2)}, |x| \leqslant 1$$
$$f_2(x) = \frac{2}{\pi(1+x^2)}, |x| > 1$$

We know that if $X \sim f_1$, then $X^{-1} \sim f_2$. Therefore, we generate Z from f_1 and $U \sim \text{Unif}(0,1)$ with $Z \perp U$. If $U > \frac{1}{2}, X = Z$; if $U < \frac{1}{2}, X = \frac{1}{Z}$. Hence, we can use rejection sampling to sample from

f_1. Taking g to be a uniform distribution on $[-1.1]$ is a reasonable choice, and $\frac{f_1(x)}{g(x)}$ is maximized at $x = 0$. Thus

$$M = \frac{f_1(0)}{g(0)} = \frac{4}{\pi} = 1.273$$

and

$$P(\text{Accept}) = \frac{\pi}{4} = 0.785$$

3.6 Sampling from Mixture Densities

Suppose we want to sample from a density $f(x)$ which can be written as a mixture of k components:

$$f(x) = \lambda_1 f_1(x) + \cdots + \lambda_k f_k(x)$$

where $f_1(x), \cdots, f_k(x)$ are densities, $\lambda_1 + \cdots + \lambda_k = 1$. We sample a discrete r.v. J from a discrete distribution with $P(J = j) = \lambda_j$ (and we can do with a single Unif$(0, 1)$ r.v.). Given $J = j$, sample X from $f_j(x)$. The algorithm works best if k is small or if $\lambda_1 = \cdots = \lambda_k = \frac{1}{k} : J = \lceil kU \rceil$.

3.6.1 Application: Walker's Alias Method

Suppose we want to sample X from a discrete distribution

$$f(x_j) = P(X = x_j) = p_j, j = 1, \cdots, k$$

where $p_1 + \cdots + p_k = 1$. We can write f as a mixture of k components each with weight k^{-1} :

$$f(x) = P(X = x) = \frac{1}{k} f_1(x) + \cdots + \frac{1}{k} f_k(x), x = x_1, \cdots, x_k$$

$f_1, \cdots, f_k$ are discrete distribution putting mass at two points:

$$f_j(x) = \begin{cases} \tau_j, & x = x_j \\ 1 - \tau_j, & x = a_j \end{cases}$$

where $a_j \in \{x_1, \cdots, x_k\}$ is called an alias.

Given $\tau_1, \cdots, \tau_k$ and $a_1, \cdots, a_k$, we sample X from f as follows: Generate $U_1 \sim$ Unif$(0, 1)$ and set $J = \lceil kU_1 \rceil$; generate $U_2 \sim$ Unif$(0, 1)$ and define $X = x_J$ if $U_2 \leqslant \tau_J$ and $X = a_J$ if $U_2 > t_J$.

Note 1. We require a separate algorithm to construct $\tau_1, \cdots, \tau_k$ and $a_1, \cdots, a_k$.

Note 2. Walker's alias method is used by the R function `sample` when the option `replace=T` is given.

Example 3.13 (Binomial Distribution)**.** Take $X \sim$ Binom$(3, 0.4)$:

$$f(x) = P(X = x) = \binom{3}{x} 0.4^x 0.6^{3-x}, x = 0, 1, 2, 3$$

where $f(0) = 0.216, f(1) = 0.432, f(2) = 0.288, f(3) = 0.064$. We need to write

$$f(x) = \frac{1}{4} \sum_{i=0}^{3} f_i(x)$$

with $f_i(x) = \tau_i$ is $x = i$ and $f_i(x) = 1 - \tau_i$ if $x = a_i$ where $a_i \in \{0, 1, 2, 3\}$. We let

$$\begin{aligned} \tau_0 &= 0.272 \quad a_0 = 1 \\ \tau_1 &= 1 \quad \text{No alias} \\ \tau_2 &= 0.408 \quad a_2 = 0 \\ \tau_3 &= 0.256 \quad a_3 = 2 \end{aligned}$$

3.7 Generation of Normal Random Variables

3.7.1 Inverse Method

Define the $\mathcal{N}(0, 1)$ distribution function

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t$$

$\Phi(x)$ is strictly increasing so we can define its inverse by $\Phi(\Phi^{-1}(t)) = t$ for $0 \leqslant t \leqslant 1$. Thus given $U \sim \text{Unif}(0, 1), X = \Phi^{-1}(U) \sim \mathcal{N}(0, 1)$.

Note 1. It is the default method in R.

Note 2. Though $\Phi^{-1}(t)$ is not a nice function, it is very well approximated.

3.7.2 Box-Muller Method

If X_1 and X_2 are independent $\mathcal{N}(0, 1)$, then their joint density is

$$f(x_1, x_2) = \frac{1}{2\pi} \exp\left(-\frac{x_1^2 + x_2^2}{2}\right)$$

Convert to polar coordinates: $X_1 = R\cos(\Theta)$ and $X_2 = R\sin(\Theta)$, and (R, Θ) has joint density

$$g(r, \theta) = \frac{1}{2\pi} \exp\left(-\frac{r^2}{2}\right) r, r > 0, 0 \leqslant \theta < 2\pi$$

where $g(r, \theta) = g_1(r)g_2(\theta)$ and so $R \perp \Theta$. $\Theta \sim \text{Unif}(0, 2\pi)$ and $R = \sqrt{V}$ where V is exponential with mean 2. We generate R from $g_1(r)$ and Θ from $g_2(\theta)$, and

$$X_1 = R\cos(\Theta), X_2 = R\sin(\Theta)$$

If U_1 and U_2 are independent Unif$(0, 1)$, then we can define

$$\Theta = 2\pi U_1, R = \sqrt{-2\ln(U_2)}$$

3.7.3 Kinderman-Ramage Method

Consider half-normal distribution

$$f(x) = \frac{2}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

and we write $f(x)$ as a mixture of two distributions

$$\begin{aligned} f(x) &= \lambda_1 f_1(x) + \lambda_2 f_2(x) \\ &= 0.884 \times \underbrace{\text{Triangular density}}_{0.90 - 0.41x} + 0.116 f_2(x) \end{aligned}$$

Note 1. It is easy to generate from the triangular density $f_1(x)$: U_1 and U_2 are independent Unif$(-1, 1)$, then $V = \dfrac{2.216|U_1 + U_2|}{2}$ has density $f_1(x)$.

Note 2. It is not easy to generate from $f_2(x)$.

3.7.4 Monty Python Method

We generate independent r.v.s. U_1 and U_2 s.t. (U_1, U_2) have a uniform distribution on

$$\mathcal{B} = [0, \sqrt{2\pi}] \times \left[0, \frac{1}{\sqrt{2\pi}}\right]$$

and divide $\mathcal{B}$ into 4 regions - depending on which region (U_1, U_2), we can define a r.v. X with a half-normal distribution.

Define

$$f(x) = \frac{2}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

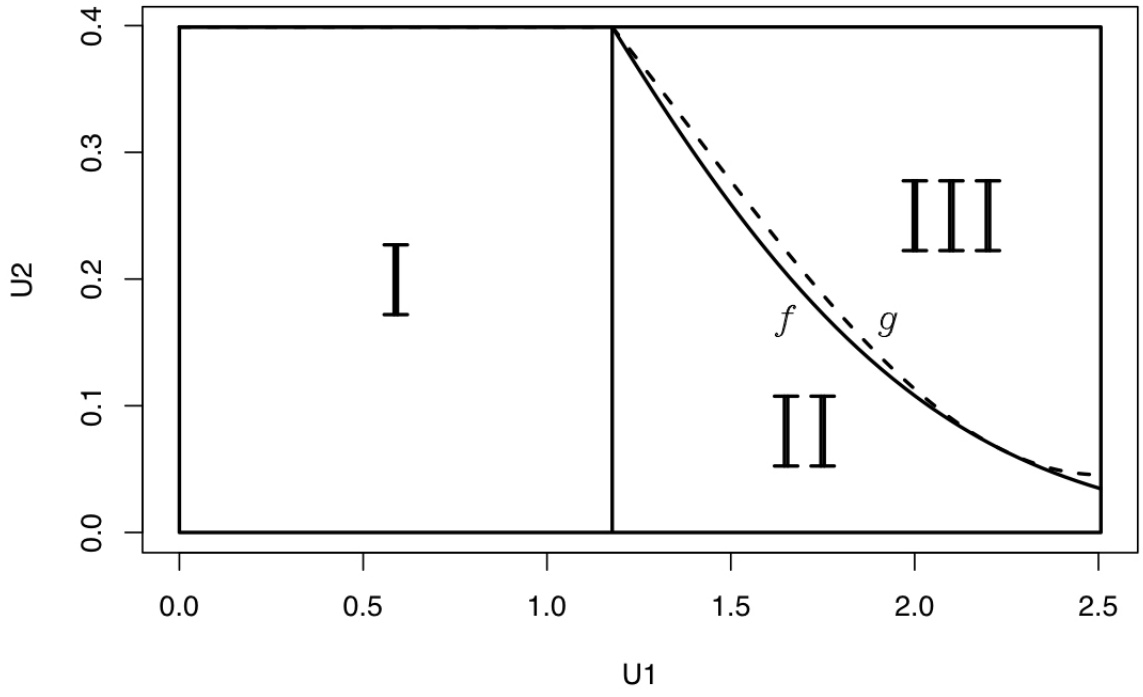
Define $g(x)$ to be $f(x)$ rotated and rescaled into $\mathcal{B}$ for x s.t.

$$f(x) > \frac{1}{\sqrt{2\pi}} = \frac{1}{b} \text{ or } x < \sqrt{\ln(4)} = a$$

i.e.,

$$g(x) = \frac{1}{b} - \frac{a}{b-a}\left[f\left(\frac{a(b-x)}{b-a}\right) - \frac{1}{b}\right]$$

for $\sqrt{\ln(4)} = a \leqslant x \leqslant b = \sqrt{2\pi}$. We can refine regions I, II, and III in terms of $f(x)$ and $g(x)$.



We generate (U_1, U_2) on $\mathcal{B}$: $U_1 \sim \text{Unif}(0, \sqrt{2\pi})$ and $U_2 \sim \text{Unif}(0, 1/\sqrt{2\pi})$. If $(U_1, U_2) \in$ I, then $X = U_1$; if $(U_1, U_2) \in$ II, then $X = U_1$; if $(U_1, U_2) \in$ III, then $X = \dfrac{a(b-U_1)}{b-a}$; otherwise, we need to generate X from the tail $(x > \sqrt{2\pi})$ of the half-normal distribution (by rejection sampling with a shifted exponential proposal).

3.7.5 Sum of Uniforms

We sum k independent Unif$(0,1)$ r.v.s. $U_1, \cdots, U_k$ and define

$$X = \frac{U_1 + \cdots + U_k - k/2}{\sqrt{k/12}}$$

where the normalization guarantees $\mathbb{E}[X] = 0$ and $\text{Var}[X] = 1$.

Note. $k = 12$ works well.

3.8 Markov Chain Monte Carlo

3.8.1 Construction of Reversible Markov Chain

We first assume that $f(\mathbf{x})q(\mathbf{x},\mathbf{y}) > f(\mathbf{y})q(\mathbf{y},\mathbf{x})$. Define $q^*(\mathbf{x},\mathbf{y}) = \alpha(\mathbf{x},\mathbf{y})q(\mathbf{x},\mathbf{y})$ s.t.

$$f(\mathbf{x})q^*(\mathbf{x},\mathbf{y}) = f(\mathbf{y})q^*(\mathbf{y},\mathbf{x})$$

The solution is

$$\alpha(\mathbf{x},\mathbf{y}) = \frac{f(\mathbf{y})q(\mathbf{y},\mathbf{x})}{f(\mathbf{x})q(\mathbf{x},\mathbf{y})}$$
$$\alpha(\mathbf{y},\mathbf{x}) = 1$$

We can do the similar thing if $f(\mathbf{x})q(\mathbf{x},\mathbf{y}) < f(\mathbf{y})q(\mathbf{y},\mathbf{x})$. In general,

$$\alpha(\mathbf{x},\mathbf{y}) = \min\left\{\frac{f(\mathbf{y})q(\mathbf{y},\mathbf{x})}{f(\mathbf{x})q(\mathbf{x},\mathbf{y})}, 1\right\}$$

Note. $q^*(\mathbf{x},\mathbf{y})$ may not be a transition density (unless $\alpha(\mathbf{x},\mathbf{y}) = 1$ for all $\mathbf{x},\mathbf{y}$). Given $\mathbf{X}_{i-1} = \mathbf{x}$, we can fix by allowing $\mathbf{X}_i = \mathbf{x}$ w.p.

$$\alpha(\mathbf{x},\mathbf{x}) = 1 - \int\cdots\int q^*(\mathbf{x},\mathbf{y})\mathrm{d}\mathbf{y}$$

which ideally should be small.

3.8.2 Metropolis-Hastings Algorithm

Suppose we want to generate $\mathbf{X}_i$ from $f(\mathbf{x})$ and we have a proposal transition density $q(\mathbf{x},\mathbf{y})$. Given $\mathbf{X}_{i-1} = \mathbf{x}$, we generate $\mathbf{Y}$ from $q(\mathbf{x},\mathbf{y})$ (density in $\mathbf{y}$ for each $\mathbf{x}$) and $U \sim$ Unif$(0,1)$ independent of $\mathbf{Y}$. If $U \leqslant \alpha(\mathbf{X}_{i-1},\mathbf{Y})$, then $\mathbf{X}_i = \mathbf{Y}$; if $U > \alpha(\mathbf{X}_{i-1},\mathbf{Y})$, then $\mathbf{X}_i = \mathbf{X}_{i-1}$.

We want to sample $\mathbf{X}_i$ s.t.

$$\frac{1}{n}\sum_{i=1}^{n} h(\mathbf{X}_i) \text{ converges to } \int\cdots\int h(\mathbf{x})f(\mathbf{x})\mathrm{d}\mathbf{x}$$

as fast as possible for any function h, and the convergence speed is determined largely by

$$\text{Var}\left[\frac{1}{n}\sum_{i=1}^{n} h(\mathbf{X}_i)\right] \approx \frac{1}{n}\left\{\text{Var}[h(\mathbf{X}_i)] + 2\sum_{s=1}^{\infty}\text{Cov}[h(\mathbf{X}_i), h(\mathbf{X}_{i+s})]\right\}$$

The choice of $q(\mathbf{x},\mathbf{y})$ is important − it determines $\alpha(\mathbf{x},\mathbf{y})$, i.e., how often $\mathbf{X}_{i-1} = \mathbf{X}_i$, and how quickly $\mathbf{X}_i$ move around the space − we want to make the autocovariance terms small.

Example 3.14. Suppose we want to generate X_i from

$$P(X_i = x) = \binom{2}{x} 0.3^2 0.7^{2-x}, x = 0, 1, 2$$

Using simple transition matrix

$$Q = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{pmatrix}$$

s.t. $q(x, y) = \frac{1}{3}$ for $x, y = 0, 1, 2$. Our acceptance probability is

$$\alpha(x, y) = \min\left\{\frac{f(y)q(y, x)}{f(x)q(x, y)}, 1\right\} = \min\left\{\frac{\binom{2}{y}}{\binom{2}{x}} 0.3^{y-x} 0.7^{x-y}, 1\right\}$$

3.8.2.1 Application to Bayesian Inference

$\alpha(\mathbf{x}, \mathbf{y})$ depends on f only via the ratio $\frac{f(\mathbf{y})}{f(\mathbf{x})}$ and we only need to know $f(\mathbf{x})$ up to a multiplicative constant. Hence, we do not need to know the constant to sample from the posterior density.

3.8.2.2 Random Walk (Metropolis) Sampler

$q(\mathbf{x}, \mathbf{y}) = g(\mathbf{y} - \mathbf{x})$ for some density g :

$$\alpha(\mathbf{x}, \mathbf{y}) = \min\left\{\frac{f(\mathbf{y})g(\mathbf{x} - \mathbf{y})}{f(\mathbf{x})g(\mathbf{y} - \mathbf{x})}, 1\right\}$$

Given $\mathbf{X}_{i-1} = \mathbf{x}$, we would generate the proposal $\mathbf{Y}$ by

$$\mathbf{Y} = \mathbf{x} + \mathbf{Z}$$

where the density of $\mathbf{Z}$ is g.

Note. If g is symmetric around 0, then

$$\alpha(\mathbf{x}, \mathbf{y}) = \min\left\{\frac{f(\mathbf{y})}{f(\mathbf{x})}, 1\right\}$$

3.8.2.3 Independence (Hastings) Sampler

$q(\mathbf{x}, \mathbf{y}) = g(\mathbf{y})$: For each i, the distribution of the proposal $\mathbf{Y}$ is independent of $\mathbf{X}_{i-1}$

$$\alpha(\mathbf{x}, \mathbf{y}) = \min\left\{\frac{f(\mathbf{y})g(\mathbf{x})}{f(\mathbf{x})g(\mathbf{y})}, 1\right\}$$

Note. The independence sampler somewhat resembles rejection sampling (but without the rejection).

Example 3.15 (Poisson Distribution)**.** Suppose we want to sample X_i from a Poisson distribution with mean $\lambda > 0$. We use independence sampler with a geometric proposal

$$g(x) = (1 - \theta)\theta^x, x = 0, 1, \cdots$$

whose expected value is $\frac{\theta}{1-\theta}$. We choose θ s.t. $\frac{\theta}{1-\theta} = \lambda$, i.e., $\theta = \frac{\lambda}{1+\lambda}$. We can sample from a geometric distribution by $\lfloor V \rfloor$ where V as an exponential distribution with mean $-\frac{1}{\ln(\theta)}$.

The code for $\lambda = 5$ is:

```
lambda = 5
x = 5
samp = NULL
theta = lambda / (1 + lambda)
for (i in 1:10000) {
   v = -rexp(1) / log(theta)
   y = floor(v)
   u = runif(1)
   if (u <= dpois(y, lambda) * theta^(x-y) / dpois(x, lambda)) x = y
   samp = c(samp, x)
}
eprob = NULL
for (i in 0: 10) eprob = c(eprob, sum(samp==i) / 10000)
```

3.8.3 Practical Issues of MCMC

MCMC is often very sensitive to initial conditions. We can discard the first m iterations of the MCMC algorithm.

Note. This is important when sampling high dimensional random vectors.

It is useful to treat the output of an MCMC algorithm as a time series, and we can look at time series plots to see when the output has achieved stationarity, and autocorrelation $\widehat{\rho}(1), \cdots$.

The effective sample size is

$$n_{\text{eff}} = \left[1 + 2\sum_{s=1}^{\infty} \rho(s)\right]^{-1} n$$

4 Numerical Linear Algebra

4.1 Solving Linear Equations

Theorem 4.1 (Sherman-Morrison-Woodbury Formula/Woodbury Matrix Identity)**.**

$$(A+UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1}+VA^{-1}U)^{-1}VA^{-1}$$

Note. If A and C are diagonal matrices, then computation of $(A+UCV)^{-1}$ is easy.

Example 4.1. Let $\mathbf{u}$ and $\mathbf{v}$ be vectors of length n and so $\mathbf{u}\mathbf{v}^T$ has rank 1. We can now use the Woodbury matrix identity setting $A = I, C = 1, U = \mathbf{u}, V = \mathbf{v}^T$

$$\begin{aligned}(I+\mathbf{u}\mathbf{v}^T)^{-1} &= I - I\mathbf{u}(1+\mathbf{v}^T I\mathbf{u})^{-1}\mathbf{v}^T I \\ &= I - \frac{1}{1+\mathbf{v}^T\mathbf{u}}\mathbf{u}\mathbf{v}^T\end{aligned}$$

and thus

$$A^{-1}\mathbf{b} = \mathbf{b} - \frac{\mathbf{v}^T\mathbf{b}}{1+\mathbf{v}^T\mathbf{u}}\mathbf{u}$$

Example 4.2. Define

$$A = \begin{pmatrix} 1 & 1-\varepsilon \\ 1+\varepsilon & 1 \end{pmatrix} = \underbrace{\begin{pmatrix} 0 & -\varepsilon \\ \varepsilon & 0 \end{pmatrix}}_{B_\varepsilon} + \underbrace{\begin{pmatrix} 1 \\ 1 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix}^T}_{\mathbf{v}\mathbf{v}^T}$$

where $B_\varepsilon^{-1} = B_{1\varepsilon}$ (both off-diagonal matrices). We can apply the Woodbury identity to evaluate $A^{-1}\mathbf{b}$:

$$\begin{aligned}A^{-1}\mathbf{b} &= B_\varepsilon^{-1}\mathbf{b} - \frac{1}{1+\mathbf{v}^T B_\varepsilon^{-1}\mathbf{v}} B_\varepsilon^{-1}\mathbf{v}\mathbf{v}^T B_\varepsilon^{-1}\mathbf{b} \\ &= B_\varepsilon^{-1}\mathbf{b} - \frac{\mathbf{v}^T B_\varepsilon^{-1}\mathbf{b}}{1+\mathbf{v}^T B_\varepsilon^{-1}\mathbf{v}} B_\varepsilon^{-1}\mathbf{v}\end{aligned}$$

The R code is:

```
eps = 1.e-7
B = matrix(c(0, -eps, eps, 0), ncol=2, byrow=T)
b = c(1, 1)
v = c(1, 1)
solve(B, b) - sum(v*solve(B, b)) * solve(B, v) / (1 + sum(v*solve(B, v)))
```

Definition 4.1 (Pivoting)**.** Pivoting means exchanging rows. Let P be a permutation matrix where each row and column has exactly one 1 and $n-1$ 0s, then $P\mathbf{b}$ rearranges elements of $\mathbf{b}$ while PA rearranges rows of A.

Gaussian elimination with partial pivoting is essentially the approach used by R function `solve`. For some permutation matrix P, we find lower and upper triangular matrices L and U s.t.

$$PA = LU$$

Then

$$\begin{aligned}PA\mathbf{x} &= P\mathbf{b} \\ L\underbrace{U\mathbf{x}}_{\mathbf{y}} &= P\mathbf{b}\end{aligned}$$

We solve $L\mathbf{y} = P\mathbf{b}$ for $\mathbf{y}$ and then solve $U\mathbf{x} = \mathbf{y}$ for $\mathbf{x}$.

4.2 Matrix Factorizations

4.2.1 Cholesky Factorization

If A is a symmetric ($A = A^T$) and positive definite ($\mathbf{x}^T A\mathbf{x} > 0$ for $\mathbf{x} \neq \mathbf{0}$) $n \times n$ matrix, then we can write

$$A = LL^T$$

where L is lower triangular.

4.2.1.1 Computation of L

Define A_k to be the upper left $k \times k$ sub-matrix of A where $A_1 = a_{1,1}$ and $A_n = A$. A_k is symmetric positive definite so $A_k = L_k L_k^T$ where L_k is lower triangular and L_k is a sub-matrix of L_{k+1}.

Define $\mathbf{v}_{k-1} = (l_{k,1}, \cdots, l_{k,k-1})^T$ and $\mathbf{a}_{k-1} = (a_{1,k}, \cdots, a_{k-1,k})^T$, then

$$\begin{aligned} A_k &= \begin{pmatrix} A_{k-1} & \mathbf{a}_{k-1} \\ \mathbf{a}_{k-1}^T & a_{k,k} \end{pmatrix} \\ &= \begin{pmatrix} A_{k-1} & L_{k-1}\mathbf{v}_{k-1} \\ \mathbf{v}_{k-1}^T L_{k-1}^T & l_{k,1}^2 + \cdots + l_{k,k}^2 \end{pmatrix} \end{aligned}$$

Thus we have

$$L_{k-1}\mathbf{v}_{k-1} = \mathbf{a}_{k-1} \text{ (Lower triangular system)}$$

and

$$l_{k,k} = \sqrt{a_{k,k} - (l_{k,1}^2 + \cdots + l_{k,k-1}^2)}$$

Then we can successively compute $L_1 = \sqrt{a_{1,1}}, L_2, \cdots, L_n = L$:

$$L_k = \begin{pmatrix} L_{k-1} & & 0 \\ l_{k,1} & \cdots & l_{k,k-1} \quad l_{k,k} \end{pmatrix}$$

Example 4.3. Let

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

Then

$$L_1 = l_{1,1} = \sqrt{a_{1,1}} = \sqrt{2}$$

and

$$L_2 = L = \begin{pmatrix} \sqrt{2} & 0 \\ l_{2,1} & l_{2,2} \end{pmatrix}$$

We have

$$\sqrt{2} l_{2,1} = a_{1,2} = 1 \Rightarrow l_{2,1} = \frac{1}{\sqrt{2}}$$

and

$$l_{2,2} = \sqrt{a_{2,2} - l_{2,1}^2} = \sqrt{\frac{3}{2}}$$

Thus

$$L = \begin{pmatrix} \sqrt{2} & 0 \\ \sqrt{1/2} & \sqrt{3/2} \end{pmatrix}$$

4.2.1.2 Application: Generating Multivariate Normal Random Vectors

Suppose we want to generate a random vector $\mathbf{X}$ from a p-variate normal distribution with mean vector $\mathbf{0}$ and $p \times p$ covariance matrix C where we assume C is positive definite.

We generate $Y_1, \cdots Y_p$ independent $\mathcal{N}(0,1)$ r.v.s. and define $\mathbf{Y} = (Y_1, \cdots, Y_p)^T$, and compute L in the Cholesky factorization of C. We define $\mathbf{X} = L\mathbf{Y}$ and $\mathbf{X} \sim \mathcal{N}_p(\mathbf{0}, LL^T = C)$.

Note. In R, we use function `chol` that returns L^T.

4.3 Iterative Matrix Method

Suppose we solve $A\mathbf{x} = \mathbf{b}$ for $\mathbf{x}$ where A is an $n \times n$ matrix with n very large. If A is not too complicated then we can solve iteratively, i.e., find a sequence $\{\mathbf{x}_k\}$ s.t. $\mathbf{x}_k$ converges to the solution. We write $A = A_1 + A_2$ where A_1 is nice (e.g., diagonal, lower or upper triangular) and define $\mathbf{x}_1, \mathbf{x}_2, \cdots$ s.t.

$$A_1\mathbf{x}_{k+1} = \mathbf{b} - A_2\mathbf{x}_k \text{ or } \mathbf{x}_{k+1} = A_1^{-1}\mathbf{b} - A_1^{-1}A_2\mathbf{x}_k$$

4.3.1 Jacobi and Gauss-Seidel Algorithm

Suppose we want to solve

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$$

For each $i = 1, \cdots, n$, we have

$$b_i = \sum_{j=1}^{n} a_{ij}x_j = a_{ii}x_i + \sum_{j \neq i} a_{ij}x_j$$

so that

$$x_i = \frac{1}{a_{ii}} \left(b_i - \sum_{j \neq i} a_{ij}x_j \right)$$

Then,

1. Jacobi algorithm:

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j \neq i} a_{ij}x_j^{(k)} \right)$$

2. Gauss-Seidel algorithm:

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j<i} a_{ij}x_j^{(k+1)} - \sum_{j>i} a_{ij}x_j^{(k)} \right)$$

Note. The Gauss-Seidel algorithm computes $\mathbf{x}_{k+1}$ in place, which is more efficient for memory; while for the Jacobi algorithm, we need to store both $\mathbf{x}_{k+1}$ and $\mathbf{x}_k$.

Let $A = L + D + U$:

$$A = \begin{pmatrix} 0 & 0 & \cdots & 0 & 0 \\ a_{21} & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{n,n-1} & 0 \end{pmatrix} + \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{pmatrix} + \begin{pmatrix} 0 & a_{12} & \cdots & a_{1,n-1} & a_{nn} \\ \vdots & \cdots & \ddots & \cdots & \vdots \\ 0 & 0 & \cdots & a_{n-1,n-1} & a_{n-1,n} \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix}$$

We can write both the Gauss-Seidel and Jacobi iterations in terms of L, D and U :

1. Jacobi: $\mathbf{x}_{k+1} = D^{-1}[\mathbf{b} - (L+U)\mathbf{x}_k]$.
2. Gauss-Seidel: $(L+D)\mathbf{x}_{k+1} = \mathbf{b} - U\mathbf{x}_k$ or $\mathbf{x}_{k+1} = (L+D)^{-1}(\mathbf{b} - U\mathbf{x}_k)$.

4.3.1.1 Convergence of Gauss-Seidel and Jacobi Algorithm

Suppose $\mathbf{x}_k \to \mathbf{x}^*$ as $k \to \infty$, then

1. Jacobi: $\mathbf{x}^* = D^{-1}[\mathbf{b} - (L+U)\mathbf{x}^*]$ or $(L+D+U)\mathbf{x}^* = \mathbf{b}$.
2. Gauss-Seidel: $\mathbf{x}^* = (L+D)^{-1}(\mathbf{b} - U\mathbf{x}^*)$ or $(L+D+U)\mathbf{x}^* = \mathbf{b}$.

4.3.1.2 Comment

1. When n is large and A is relatively sparse, then iterative method can be much more efficient than direct method.
2. Iterative method is easy to program.
3. We can extend the basic idea behind the Gauss-Seidel algorithm to other problems, such as back-fitting, coordinate descent, etc.
4. We can improve the algorithm with successive over-relaxation (S0OR) method.

4.3.1.3 Application: Quadratic Minimization

Suppose we want to minimize the quadratic function

$$g(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A\mathbf{x} - \mathbf{x}^T\mathbf{b} + c$$

where A is symmetric positive definite, i.e., $A^T = A$ and $\mathbf{x}^T A\mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$.

We can write out $g(\mathbf{x})$ explicitly:

$$g(\mathbf{x}) = g(x_1, \cdots, x_n) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j - \sum_{i=1}^{n} x_i b_i + c$$

Now fix $\{x_i : i \neq k\}$ and minimize $g(\mathbf{x})$ w.r.t. x_k :

$$\frac{\partial}{\partial x_k} g(\mathbf{x}) = \sum_{i=1}^{n} a_{ik}x_i - b_k = a_{kk}x_k + \sum_{i \neq k} a_{ik}x_i - b_k$$

Setting the partial derivative to 0, we have

$$x_k = \frac{1}{a_{kk}}\left(b_k - \sum_{i \neq k} a_{ik}x_i\right)$$

which is simply a Gauss-Seidel iteration.