STA 6366, Homework.3

Carson Slater *Baylor University*

(1) The transition matrix of a Markov chain is shown below. Find the stationary distribution for this Markov Chain. You may use R or solve this by hand.

$$\mathbf{P} = \begin{bmatrix} 0.6 & 0.3 & 0.1 \\ 0.4 & 0.4 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}$$

We have that the stationary distribution for this Markov Chain is:

$$\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$$

So the stationary distribution is the solution to the system

$$\begin{aligned} 0.6\pi_1 + 0.4\pi_2 + 0.1\pi_3 &= \pi_1 \\ 0.3\pi_1 + 0.4\pi_2 + 0.1\pi_3 &= \pi_2 \\ 0.1\pi_1 + 0.2\pi_2 + 0.8\pi_3 &= \pi_3, \end{aligned}$$

subject to $\pi_1 + \pi_2 + \pi_3 = 1$. Moving all variables to the left-hand side, we then wind up with the system

$$\begin{aligned} -0.4\pi_1 + 0.4\pi_2 + 0.1\pi_3 &= 0 \\ 0.3\pi_1 - 0.6\pi_2 + 0.1\pi_3 &= 0 \\ 0.1\pi_1 + 0.2\pi_2 - 0.2\pi_3 &= 0 \\ \pi_1 + \pi_2 + \pi_3 &= 1, \end{aligned}$$

which can be expressed as

$$\underbrace{\begin{bmatrix} -0.4 & 0.4 & 0.1 \\ 0.3 & -0.6 & 0.1 \\ 0.1 & 0.2 & -0.2 \\ 1 & 1 & 1 \end{bmatrix}}_{\mathbf{A}} \begin{bmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \end{bmatrix} = \underbrace{\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}}_{\mathbf{b}}$$

So we have $\mathbf{A}\boldsymbol{\pi}' = \mathbf{b}$.

```r
# ------------------- Problem 1 -------------------
# Helper function using QR decomposition
solve_steady_state <- function(A_mat, b_vec) {
  A_mat |>
    qr.solve(b_vec)
}
# fmt:skip
A <- matrix(
  c(-0.4,  0.4,  0.1,
     0.3, -0.6,  0.1,
     0.1,  0.2, -0.2,
     1.0,  1.0,  1.0),
  nrow = 4,
  byrow = TRUE
)

# Define the vector b (zeros + sum-to-one constraint)
b <- c(0, 0, 0, 1)

# Calculate the probabilities
pi_probs <- solve_steady_state(A, b)

# Assign names for clarity and print
names(pi_probs) <- c("pi_1", "pi_2", "pi_3")
print(pi_probs)
```

```
##      pi_1      pi_2      pi_3
## 0.3448276 0.2413793 0.4137931
```

(2) The general form of a 2x2 Markov chain is shown below, expressed in terms of a and b. Find the stationary distribution for this Markov chain. Interpret your result.

$$\mathbf{P} = \begin{bmatrix} 1-a & a \\ b & 1-b \end{bmatrix}$$

We know the stationary distribution satisfies $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$:

$$\begin{bmatrix} \pi_1 & \pi_2 \end{bmatrix} \begin{bmatrix} 1-a & a \\ b & 1-b \end{bmatrix} = \begin{bmatrix} \pi_1 & \pi_2 \end{bmatrix}$$

This gives the following system of equations:

$$\begin{aligned} (1-a)\pi_1 + b\pi_2 &= \pi_1 \\ a\pi_1 + (1-b)\pi_2 &= \pi_2 \end{aligned}$$

Subtracting π_1 from both sides of the first equation yields:

$$-a\pi_1 + b\pi_2 = 0 \implies a\pi_1 = b\pi_2$$

We apply the sum-to-one constraint $\pi_1 + \pi_2 = 1$ by substituting $\pi_2 = 1 - \pi_1$:

$$\begin{aligned} a\pi_1 &= b(1 - \pi_1) \\ a\pi_1 + b\pi_1 &= b \\ \pi_1(a + b) = b &\implies \pi_1 = \frac{b}{a + b} \end{aligned}$$

Substituting π_1 back into the constraint gives $\pi_2 = 1 - \frac{b}{a+b} = \frac{a}{a+b}$. Thus, the stationary distribution is:

$$\boldsymbol{\pi} = \begin{bmatrix} \frac{b}{a+b} & \frac{a}{a+b} \end{bmatrix}$$

Hence, the long-run probability of being in a specific state is proportional to the probability of transitioning *into* that state from the other. For instance, if b (the probability of moving from state 2 to 1) is much larger than a (the probability of moving from 1 to 2), the chain will spend a proportionally larger fraction of time in state 1.

(3) There are four models described below for a signal of length five (i.i.d., weight matrix, first-order Markov and MDD). For each sequence CCGAT and CATAT, find the likelihood of the sequence occurring under each model (so 2 probabilities per model, 8 in total). Note that this is an $N = k = 5$ situation.

a. IID: The nucleotides occur with the following probabilities:

$$p_A = 0.2, \quad p_C = 0.1, \quad p_G = 0.1, \quad p_T = 0.6$$

If all of them are i.i.d. then we have it that the likelihood for the signal CCGAT would be

$$\begin{aligned} P(\text{CCGAT}) &= P(C)P(C)P(G)P(A)P(T) \\ &= (0.1)(0.1)(0.1)(0.2)(0.6) \\ &= 0.00012. \end{aligned}$$

Then, the likelihood for the signal CATAT would be

$$\begin{aligned} P(\text{CATAT}) &= P(C)P(A)P(T)P(A)P(T) \\ &= (0.1)(0.2)(0.6)(0.2)(0.6) \\ &= 0.00144. \end{aligned}$$

b. Weight matrix: The weight matrix (in order A, C, G, T) is

$$\begin{bmatrix} .2 & .3 & .2 & .1 & .1 \\ .1 & .2 & .15 & .6 & .6 \\ .3 & .4 & .6 & .1 & .15 \\ .4 & .1 & .05 & .2 & .15 \end{bmatrix}$$

Using the weight matrix, we would have that the likelihood for the signal CCGAT would be

$$\begin{aligned} P(\text{CCGAT}) &= P(C|i=1)P(C|i=2)P(G|i=3)P(A|i=4)P(T|i=5) \\ &= (0.3)(0.4)(0.6)(0.1)(0.15) \\ &= 0.00108. \end{aligned}$$

Then the likelihood for the signal CATAT would be

$$\begin{aligned} P(\text{CATAT}) &= P(C|i=1)P(A|i=2)P(T|i=3)P(A|i=4)P(T|i=5) \\ &= (0.3)(0.3)(0.05)(0.1)(0.15) \\ &= 0.0000675. \end{aligned}$$

c. First-order Markov: The probabilities at position 1 are the same from (a), and the transition matrix is

$$\begin{bmatrix} .1 & .8 & .05 & .05 \\ .35 & .1 & .1 & .45 \\ .3 & .2 & .2 & .3 \\ .6 & .1 & .25 & .05 \end{bmatrix}$$

Using the first-order Markov model (assuming the matrix order A, C, G, T), the likelihood for the signal CCGAT is:

$$\begin{aligned} P(\text{CCGAT}) &= P(C)P(C|C)P(G|C)P(A|G)P(T|A) \\ &= (0.1)(0.1)(0.1)(0.3)(0.05) \\ &= 0.000015. \end{aligned}$$

Then, the likelihood for the signal CATAT is:

$$\begin{aligned} P(\text{CATAT}) &= P(C)P(A|C)P(T|A)P(A|T)P(T|A) \\ &= (0.1)(0.35)(0.05)(0.6)(0.05) \\ &= 0.0000525. \end{aligned}$$

d. Maximal dependence decomposition. Exploratory analysis indicates that position 2 has the strongest influence over the other positions. Marginal probabilities for position 2 are found to be $p_a = 0.2$, $p_c = 0.3$, $p_g = 0.1$, $p_t = 0.4$. Given the nucleotide at position 2, the weight matrix for the other 4 spots are shown below (Note that I am omitting $\mathbf{W}_G, \mathbf{W}_T$ as they are not needed for this problem):

$$\mathbf{W}_A = \begin{bmatrix} .4 & .1 & .2 & .2 \\ .3 & .4 & .1 & .3 \\ .2 & .1 & .3 & .2 \\ .1 & .4 & .4 & .3 \end{bmatrix}, \quad \mathbf{W}_C = \begin{bmatrix} .2 & .1 & .2 & .8 \\ .5 & .1 & .2 & .1 \\ .2 & .1 & .3 & .05 \\ .1 & .7 & .3 & .05 \end{bmatrix}, \quad \mathbf{W}_G = \cdots, \quad \mathbf{W}_T = \cdots$$

Under the Maximal Dependence Decomposition (MDD) model, the likelihood of the sequence depends on the marginal probability of the nucleotide at position 2, and the conditional probabilities of the remaining positions from the corresponding weight matrix.

For the signal CCGAT, position 2 is C. Using the marginal probability p_c and the weight matrix $\mathbf{W}_C$ (where the columns correspond to positions 1, 3, 4, and 5 respectively), the likelihood is:

$$\begin{aligned} P(\text{CCGAT}) &= P(C \text{ at pos } 2)P(C \text{ at pos } 1|\mathbf{W}_C)P(G \text{ at pos } 3|\mathbf{W}_C)P(A \text{ at pos } 4|\mathbf{W}_C)P(T \text{ at pos } 5|\mathbf{W}_C) \\ &= (0.3)(0.5)(0.1)(0.2)(0.05) \\ &= 0.00015. \end{aligned}$$

Then, for the signal CATAT, position 2 is A. Using the marginal probability p_a and the weight matrix $\mathbf{W}_A$, the likelihood is:

$$\begin{aligned} P(\text{CATAT}) &= P(A \text{ at pos } 2)P(C \text{ at pos } 1|\mathbf{W}_A)P(T \text{ at pos } 3|\mathbf{W}_A)P(A \text{ at pos } 4|\mathbf{W}_A)P(T \text{ at pos } 5|\mathbf{W}_A) \\ &= (0.2)(0.3)(0.4)(0.2)(0.3) \\ &= 0.00144. \end{aligned}$$

(4) Suppose we have a stick of length 1 and break it randomly (and independently) in 2 places. Find the probability that the 3 pieces (U_1, U_2, U_3) can form a triangle (i.e., that each piece is smaller than the sum of the other 2 pieces).

Considering a stick broken into three pieces, with two breakpoints X_1 and X_2. Let $U_1 = X_{(1)}$, $U_2 = X_{(2)} - X_{(1)}$, and $U_3 = 1 - X_{(2)}$. Let $U_{(3)}$ be the max order statistic for all U_i. We also have $\sum_{i=1}^{3} U_i = 1$. We want to know the probability of $P(U_1 \leq U_2 + U_3 \cap U_2 \leq U_1 + U_3 \cap U_3 \leq U_1 + U_2)$. We then have it that $U_{(3)} \leq 0.5$ to satisfy the constraints. This implies that at most, any of the three U_i values can be 0.5. So we write this as

$$\begin{aligned} &P(U_1 \leq U_2 + U_3 \cap U_2 \leq U_1 + U_3 \cap U_3 \leq U_1 + U_2) \\ &\quad = P(U_1 \leq 0.5 \cap U_2 \leq 0.5 \cap U_3 \leq 0.5). \end{aligned}$$

We can compute this probability leveraging the complement,

$$\begin{aligned}
&1 - P(U_1 > 0.5 \cup U_2 > 0.5 \cup U_3 > 0.5) \\
&\quad = 1 - P(U_1 > 0.5) - P(U_2 > 0.5) - P(U_3 > 0.5) \\
&\qquad - \underbrace{P(U_1 > 0.5 \cap U_2 > 0.5) - P(U_2 > 0.5 \cap U_3 > 0.5)}_{=0} \\
&\qquad - \underbrace{P(U_1 > 0.5 \cap U_3 > 0.5) + P(U_1 > 0.5 \cap U_2 > 0.5 \cap U_3 > 0.5)}_{=0}.
\end{aligned}$$

By symmetry, we have that $P(U_1 > 0.5) = P(U_2 > 0.5) = P(U_3 > 0.5) \implies P(U_1 > 0.5 \cup U_2 > 0.5 \cup U_3 > 0.5) = 3P(U_1 > 0.5)$, and so then

$$\begin{aligned}
P(U_1 > 0.5) &= P(X_1 > 0.5 \cap X_2 > 0.5) \\
&= \frac{1}{2} \cdot \frac{1}{2} \\
&= \frac{1}{4}.
\end{aligned}$$

By substitution, this leaves us with

$$1 - P(U_1 > 0.5 \cup U_2 > 0.5 \cup U_3 > 0.5) = 1 - 3 \cdot \frac{1}{4} = \frac{1}{4}.$$

(5) Consider the question posed at the end of the slides. If we were to instead test the hypothesis that the genes occur at a somewhat regular interval, what approach might you recommend? You do not need to derive the null distribution or anything like this, but you should discuss the pros and cons of your approach and any challenges that may arise. (Note: this is very open-ended and I don't have a single correct answer in mind)

Thinking about the question in the slides, I personally would intuit that although $U_{\min}$ or $U_{\max}$ might not be the greatest for determining irregular or regular spacing on their own, perhaps you'd expect their normalized difference to be significantly similar to the expected normalized distance under uniform spacing.

To formalize this, let U_i represent the normalized distance between consecutive genes, such that $\sum U_i = 1$. Under the null hypothesis of completely random gene placement, the joint distribution of these spacings follows a standard Dirichlet distribution:

$$(U_1, U_2, \ldots, U_{n+1}) \sim \text{Dirichlet}(1, 1, \ldots, 1)$$

Conversely, under an alternative hypothesis of strict spatial regularity, the spacings converge to a constant, meaning $U_i \approx \frac{1}{n+1}$ for all i. Therefore, we can evaluate the normalized difference using the range statistic:

$$R = U_{\max} - U_{\min}$$

Under regular spacing, we expect $R \approx 0$. By comparing our observed R to the expected range derived from the null Dirichlet distribution, we establish a highly intuitive test for periodic gene placement.

(6) Derive the expected value of the Greenwood statistic using the mean and variance of a Dirichlet random variable.

Let $U_1, \ldots, U_{n-1} \sim \text{Dir}(1, 1, \ldots, 1)$ under H_0. The Greenwood statistic is defined as

$$G = \sum_{i=1}^{n+1} U_i^2.$$

For $\mathbf{U} \sim \text{Dir}(\alpha_1, \ldots, \alpha_{n+1})$ with $\alpha_0 = \sum_i \alpha_i$, the mean and variance of each component are:

$$\mathbb{E}[U_i] = \frac{\alpha_i}{\alpha_0}, \qquad \text{Var}(U_i) = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)}.$$

Derivation of $\mathbb{E}[G]$. Using $\mathbb{E}[U_i^2] = \text{Var}(U_i) + (\mathbb{E}[U_i])^2$ for each i:

With $\alpha_i = 1$ for all i and $\alpha_0 = n + 1$:

$$\mathbb{E}[U_i] = \frac{1}{n+1}, \qquad \text{Var}(U_i) = \frac{1 \cdot n}{(n+1)^2(n+2)} = \frac{n}{(n+1)^2(n+2)}.$$

Therefore,

$$\begin{aligned}
\mathbb{E}[U_i^2] &= \frac{n}{(n+1)^2(n+2)} + \frac{1}{(n+1)^2} \\
&= \frac{n}{(n+1)^2(n+2)} + \frac{n+2}{(n+1)^2(n+2)} \\
&= \frac{2(n+1)}{(n+1)^2(n+2)} \\
&= \frac{2}{(n+1)(n+2)}.
\end{aligned}$$

Summing over all $n + 1$ components:

$$\mathbb{E}[G] = \sum_{i=1}^{n+1} \mathbb{E}[U_i^2] = (n+1) \cdot \frac{2}{(n+1)(n+2)} = \frac{2}{n+2}.$$

(7) Use either the CV, Greenwood or LRT methods to test the hypothesis that the genes occur in a regular interval. There are 9 genes, creating the following (unnormalized) gap sizes:

$$(62,\ 71,\ 88,\ 94,\ 103,\ 112,\ 119,\ 131,\ 148,\ 72)$$

Whatever method you use, give an interpretation of the test statistic and interpret the results.

```
## The normal theory approximate p-value for testing the null hypothesis
## that there is no significant clustering is 0.04339584
```

We conduct an upper tail hypothesis test for H_0, that the genes occur at regularly-spaced intervals, relying on the assumption that our standardized Greenwood test statistic

$$Z = \frac{G - \mathbb{E}[G]}{\sqrt{\mathrm{Var}(G)}} \dot{\sim} \mathcal{N}(0, 1).$$

We compute our p-value, $P(Z > z) = 0.0434$. Since this p-value falls below the significance level $\alpha = 0.05$, we reject H_0. Thus, there is sufficient evidence to conclude that the gene locations exhibit significant clustering, and the observed spacing is not consistent with random placement along the gene sequence.

Appendix

```
knitr::opts_chunk$set(
  dev = "cairo_pdf",
  fig.width = 5,
  fig.height = 5,
  fig.align = 'center',
  echo = FALSE,
  message = FALSE,
  warning = FALSE,
  error = FALSE,
  results = 'markup'
)

# Load required libraries
library("tidyverse")
library("patchwork")
library("glue")
library("scales", warn.conflicts = FALSE)
library("extrafont")
library("tinytex")
library("knitr")
library("tidyr")
library("latex2exp")
library("MASS")
library("kableExtra")

theme_set(theme_minimal(base_family = "Roboto Condensed"))

conflicted::conflicts_prefer(
  readr::col_factor(),
  purrr::discard(),
  dplyr::lag(),
  readr::parse_date(),
  kableExtra::group_rows(),
  dplyr::select
)
# ------------------- Problem 1 -------------------
# Helper function using QR decomposition
solve_steady_state <- function(A_mat, b_vec) {
  A_mat |>
    qr.solve(b_vec)
}
# fmt:skip
A <- matrix(
  c(-0.4,  0.4,  0.1,
     0.3, -0.6,  0.1,
```

```
    0.1,  0.2, -0.2,
    1.0,  1.0,  1.0),
  nrow = 4,
  byrow = TRUE
)

# Define the vector b (zeros + sum-to-one constraint)
b <- c(0, 0, 0, 1)

# Calculate the probabilities
pi_probs <- solve_steady_state(A, b)

# Assign names for clarity and print
names(pi_probs) <- c("pi_1", "pi_2", "pi_3")
print(pi_probs)
# ------------------- Problem 7 -------------------
U <- c(62, 71, 88, 94, 103, 112, 119, 131, 148, 72)

L <- sum(U)
n <- length(U) - 1 # number of points, not gaps

# Greenwood statistic (normalized gaps)
G <- sum((U / L)2)

# Mean and variance
E_G <- 2 / (n + 2)
Var_G <- (4 * n) / ((n + 2)2 * (n + 3) * (n + 4))

# Standardized statistic
Z <- (G - E_G) / sqrt(Var_G)

# Two-sided test
cat(
  "The normal theory approximate p-value for testing the null hypothesis\nthat
  pnorm(Z)
)
```