Stat 3005 – Nonparametric Statistics

Assignment 4

Due: 15 November (Friday) at 1800

Fall 2024

Exercise 4.1 (★★☆ — Cancer image data analysis (80%)). In RGB format, a color image $C = (C_{ijk}) \in \mathbb{R}^{n_1 \times n_2 \times 3}$ can be represented by a tensor, i.e., a multidimensional array, of $n_1 \times n_2$ pixel values of the three primary colors (red, green and blue). Each value $C_{i,j,k}$ is rescaled to $[0, 1]$ so that it represents the intensity of the kth primary color at the (i, j) pixel. For example, the (i, j)th pixel is the triple $(C_{i,j,1}, C_{i,j,2}, C_{i,j,3})$.

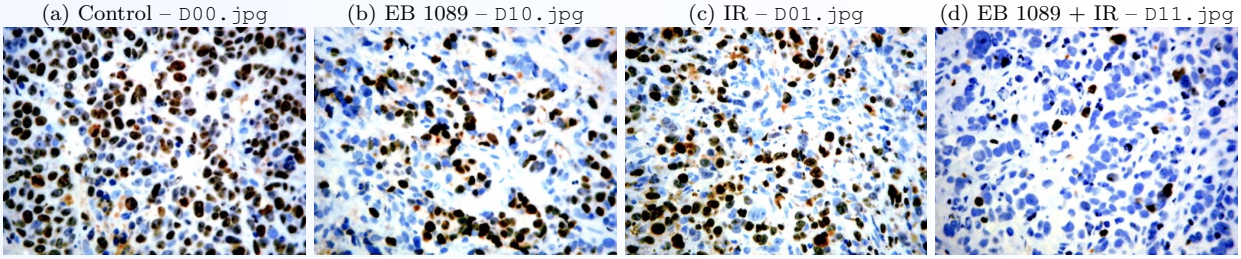
- It is black if $C_{i,j,1:3} = (0, 0, 0)$.
- It is white if $C_{i,j,1:3} = (1, 1, 1)$.
- It is blue if $C_{i,j,1:3} = (0, 0, 1)$.
- It is yellow if $C_{i,j,1:3} = (1, 1, 0)$.
- It is purple if $C_{i,j,1:3} = (0.5, 0, 0.5)$.

A greyscale (monochrome) image can be formed by a linear combination of the intensities of the three primary colors. Typically, the greyscale version of C is a matrix $G = (G_{i,j}) \in \mathbb{R}^{n_1 \times n_2}$ defined as

$$G_{i,j} = 0.22C_{i,j,1} + 0.71C_{i,j,2} + 0.07C_{i,j,3}, \tag{4.1}$$

for $i = 1, \ldots, n_1$ and $j = 1, \ldots, n_2$. The closer $G_{i,j}$ is to zero, the darker is the (i, j)th pixel. In this exercise, we are interested in performing histological analysis of cancer treatment. In particular, we compare images of tumor tissue section published in Clinical Cancer Research by Sundaram et al. (2003); see this article (click here). The following images show the proliferative activity of tumors exposed to (a) control, (b) EB 1089, (c) radiation, and (d) EB 1089 + radiation by Ki67 staining of paraffin-embedded tumor tissue sections. The dark regions indicate cancer cells. The images can be downloaded here. Our goal is to test if there is any cancer treatment effect.

(a) Control – `D00.jpg` (b) EB 1089 – `D10.jpg` (c) IR – `D01.jpg` (d) EB 1089 + IR – `D11.jpg`



1. (20%) Complete the following `R` function `get.gray` for producing grayscale images according to (4.1). Use this function to produce grayscale version of the above 4 images. Plot them.

```
#install.packages("jpeg")
img = rep(list(NA),4)
img[[1]] = jpeg::readJPEG("D00.jpg")
img[[2]] = jpeg::readJPEG("D10.jpg")
img[[3]] = ...                          # complete this line
img[[4]] = ...                          # complete this line
# plots of color images
par(mfrow=c(1,4),mar=c(0,0,1,0))
for(i in 1:4){
   plot(0:1,0:1,cex=0,axes=FALSE,main=c("D00","D10","D01","D11")[i])
   rasterImage(img[[i]],0,0,1,1)
}
# produce grayscale images
get.gray = function(img, w=c(0.22,0.71,0.07)){
   img[,,1]*w[1] + ...                  # complete this line
}
img0 = rep(list(NA),4)
for(i in 1:4){
   img0[[i]] = get.gray(img[[i]])
}
# plot of grayscale images
par(mfrow=c(1,4),mar=c(0,0,1,0))
for(i in 1:4){
   plot(0:1,0:1,cex=0,axes=FALSE,main=c("D00","D10","D01","D11")[i])
   rasterImage(...)                     # complete this line
}
```

2. (20%) Complete the following R code to produce the empirical CDFs of the grayscale pixel values of the four cases on the same plot.

```
col = c("royalblue","firebrick3","forestgreen","goldenrod2")
par(mfrow=c(1,1),mar=c(4.5,5,3,2))
for(j in 1:4){
   data = c(img0[[j]])
   plot(ecdf(...),                          # complete this line
      col=col[j],add=j>1,ylab=expression(widehat(F)(x)),xlab=expression(x))
}
treatment = c("(a) Control","(b) EB 1089","(c) IR","(d) EB 1089 + IR")
legend("bottomright",treatment,col=col,lwd=3,bg="white")
```

3. (20%) Consider the cancer treatments (b) EB 1089 and (c) IR. We suspect that the cancer treatments (b) and (c) have different effects on killing the cancer cells. Denote the CDFs of the grayscale pixel values of the treatments (b) and (c) by

$$F_B(\cdot) \qquad \text{and} \qquad F_C(\cdot).$$

   By Sundaram et al. (2003), a lower value of the CDF indicates a lighter image because of the presence of a higher fraction of apoptotic cells. So, we may statistically verify this claim by testing whether the distributions of the grayscale pixel values are the same. Write down the null and alternative hypotheses.

4. (20%) Use your preferred nonparametric method to test the hypotheses stated in part 4.

Hints: See Remark 4.1. Don't read the hints unless you have no ideas and have tried for more than 15 mins.

Exercise 4.2 (★★★ — One-sided distribution test (20%))**.** Revisit Exercise 4.1. Suppose we suspect the cancer treatment (d) is better than treatment (a). This exercise walks you through the procedures and logic for performing one-sided distribution test. Assume that the CDFs of the grayscale pixel values of the treatments (a) and (d), denoted as $F_A(\cdot)$ and $F_D(\cdot)$, respectively, are continuous and strictly increasing. Let

$$A \sim F_A(\cdot) \qquad \text{and} \qquad D \sim F_D(\cdot)$$

be independent RVs. Suppose that

$$F_A(t) > F_D(t), \qquad t \in \mathbb{R}.$$

1. (✂) Prove that we can represent A and D as $A = F_A^{-1}(U)$ and $D = F_D^{-1}(V)$, where $U, V \overset{\text{IID}}{\sim} \text{Unif}(0,1)$.
2. (✂) Prove that $F_A^{-1}(u) < F_D^{-1}(u)$ for all $u \in (0,1)$.
3. (✂) Prove that $\mathsf{E}(A) < \mathsf{E}(D)$.

So, the takeaway of the above properties is that

$$\boxed{F_A(t) > F_D(t), \quad t \in \mathbb{R}} \qquad \Rightarrow \qquad \boxed{\mathsf{E}(A) < \mathsf{E}(D)}.$$

Note that the inequality sign is reverse! So, we should test

$$H_0 : F_A(t) = F_D(t) \ \text{ for all } t \in \mathbb{R} \qquad \text{against} \qquad H_1 : \sup_{t\in\mathbb{R}}\{F_A(t) - F_D(t)\} > 0,$$

where H_1 essentially means that the image `D00.jpg` corresponding to treatment (a) is darker than the image `D11.jpg` corresponding to treatment (d), implying that treatment (d) is better than treatment (a). So,

$$\boxed{\text{Worse treatment}} \ \Leftrightarrow \ \boxed{\text{Darker image}} \ \Leftrightarrow \ \boxed{\text{Smaller pixel values}} \ \Leftrightarrow \ \boxed{\text{Larger CDF values}}.$$

Let $\widehat{F}_A(\cdot)$ be the empirical CDF based on the pixel values $A_1, \ldots, A_m$, and $\widehat{F}_D(\cdot)$ be the empirical CDF based on the pixel values $D_1, \ldots, D_n$. The one-sided two-sample Kolmogorov–Smirnov (KS) test statistic is

$$K^+_{m,n} = \sqrt{\frac{mn}{m+n}} \sup_{t\in\mathbb{R}} \left\{\widehat{F}_A(t) - \widehat{F}_D(t)\right\} = \sqrt{\frac{mn}{m+n}} \max_{i=1,\ldots,m} \left\{\frac{i}{m} - \widehat{F}_D(A_{(i)})\right\}, \tag{4.2}$$

where the second equality is true when there is no tied observations. Under H_0, as $m, n \to \infty$,

$$K^+_{m,n} \xrightarrow{\text{d}} \sup_{t\in[0,1]} (\mathbb{B}_t - t\mathbb{B}_1) =: K^+,$$

where $\mathbb{B}_t$ is a standard Brownian motion on $[0,1]$. In R, (4.2) can be computed by the following R-code:

```
A = ...        # the pixel value of image corresponding to treatment A
D = ...        # the pixel value of image corresponding to treatment D
ks.test(A, D, alternative="greater")
```

4. (20%) Find the p-value of the above test. What is your conclusion?

Hints: See Remark 4.2. Don't read the hints unless you have no ideas and have tried for more than 15 mins.

Remark 4.1 (Hints for Exercise 4.1)**.**

3. Check `help(ecdf)` for more information about the function `ecdf`. Example 5.14 is related.
4. The alternative hypothesis is the hypothesis that you suspect true.
5. You may use two-sample two-sided KS test. See Examples 5.10–5.11.

Remark 4.2 (Hints for Exercise 4.2)**.**

1. Try probability inverse transform; see Examples 5.4 and B.25.
2. Apply some inverse function on both side, and then change of variable.
3. If the integrand is smaller, then the integral is ...

It is not hard to write the code for computing the one-sided two-sample KS test statistic. You may try the following comparison:

```
> ks.test.statistic.2sided = function(x,y){
+     m = length(x)
+     n = length(y)
+     FY_at_X = ecdf(y)(sort(x))       # ecdf(y) is a function. We can compute, e.g., ecdf(y)(0.4)
+     K_plus  = sqrt(n/(1+n/m))*max( (1:m)/m - FY_at_X     )
+     K_minus = sqrt(n/(1+n/m))*max( FY_at_X - (0:(m-1))/m )
+     K = max(K_plus, K_minus)
+     c(two.sided=K, less=K_minus, greater=K_plus)
+ }
>
> # simulated dataset
> n = 5
> m = 8
> x = rnorm(n)
> y = rnorm(m)
>
> # self-written version
> ks.test.statistic.2sided(x,y)
two.sided      less   greater
0.5262348 0.3069703 0.5262348
>
> # built-in version
> c(ks.test(x,y, alternative="two.sided")$statistic,
+   ks.test(x,y, alternative="less")$statistic,
+   ks.test(x,y, alternative="greater")$statistic) * sqrt(m*n/(m+n))
        D       D^-       D^+
0.5262348 0.3069703 0.5262348
```

Instructions: Please follow the instructions stated on the last page of assignment 1.