Sample Theoretical Exercises

BIOENG-210

April 7, 2025

Question 1: Advanced Expectation Manipulation

Problem: In a pharmacokinetics study, a drug's concentration C in the blood follows the exponential decay model:

$$C(t) = C_0 e^{-\lambda t} + \epsilon(t),$$

where C_0 is the initial concentration, λ is the elimination rate, and $\epsilon(t)$ is a random error with $E[\epsilon(t)] = 0$ and variance σ^2.

1. Show that $E[C(t)] = C_0 e^{-\lambda t}$.
2. Compute $\text{Var}(C(t))$ assuming C_0 is a constant.
3. Suppose C_0 itself is a random variable with $E[C_0] = \mu_0$ and $\text{Var}(C_0) = \tau^2$. Find $E[C(t)]$ and $\text{Var}(C(t)]$.

Solution:

1. By linearity of expectation:

$$E[C(t)] = E[C_0 e^{-\lambda t} + \epsilon(t)] = C_0 e^{-\lambda t} + E[\epsilon(t)] = C_0 e^{-\lambda t}.$$

2. Since variance is additive for independent variables:

$$\text{Var}(C(t)) = \text{Var}(C_0 e^{-\lambda t}) + \text{Var}(\epsilon(t)) = 0 + \sigma^2 = \sigma^2.$$

3. Now, if C_0 is random:

$$E[C(t)] = E[C_0] e^{-\lambda t} = \mu_0 e^{-\lambda t}.$$

   Using the independence of C_0 and $\epsilon(t)$:

$$\text{Var}(C(t)) = e^{-2\lambda t}\text{Var}(C_0) + \text{Var}(\epsilon(t)) = \tau^2 e^{-2\lambda t} + \sigma^2.$$

Question 2: Marginal and Conditional Probabilities in Disease Diagnosis

Problem: A diagnostic test for a disease correctly identifies 90% of the diseased patients and has a false-positive rate of 15%. The disease prevalence in the population is 5%.

1. Compute the probability that a randomly selected individual tests positive.
2. Given that a person tested positive, what is the probability that they actually have the disease?

3. What happens to this probability if the prevalence drops to 1%?

Solution:

1. Using the law of total probability:

$$P(T^+) = P(T^+|D)P(D) + P(T^+|D^c)P(D^c),$$

where $P(T^+|D) = 0.9$, $P(D) = 0.05$, $P(T^+|D^c) = 0.15$, and $P(D^c) = 0.95$.

$$P(T^+) = (0.9)(0.05) + (0.15)(0.95) = 0.045 + 0.1425 = 0.1875.$$

2. Applying Bayes' theorem:

$$P(D|T^+) = \frac{P(T^+|D)P(D)}{P(T^+)} = \frac{(0.9)(0.05)}{0.1875} = 0.24.$$

So only 24% of those who test positive actually have the disease.

3. If $P(D) = 0.01$:

$$P(T^+) = (0.9)(0.01) + (0.15)(0.99) = 0.009 + 0.1485 = 0.1575.$$

Then,

$$P(D|T^+) = \frac{(0.9)(0.01)}{0.1575} = 0.057.$$

This shows the effect of low prevalence on predictive value.

Question 3: Sampling Distributions and Hypothesis Testing

Problem: A researcher measures the blood glucose levels of 40 individuals and wants to test if the mean differs from 100 mg/dL. The sample mean is 105 mg/dL and sample variance is 16 mg^2/dL^2.

1. What is the sampling distribution of the sample mean?
2. Compute the test statistic.

Solution:

1. The sampling distribution is:

$$\bar{X} \sim N(\mu, \sigma^2/n) \approx N(100, 16/40).$$

2. The test statistic is:

$$t = \frac{105 - 100}{\sqrt{16/40}} = \frac{5}{\sqrt{0.4}} = \frac{5}{0.632} \approx 7.91.$$

From statistical tables, this yields a very small p-value, indicating strong evidence against H_0.

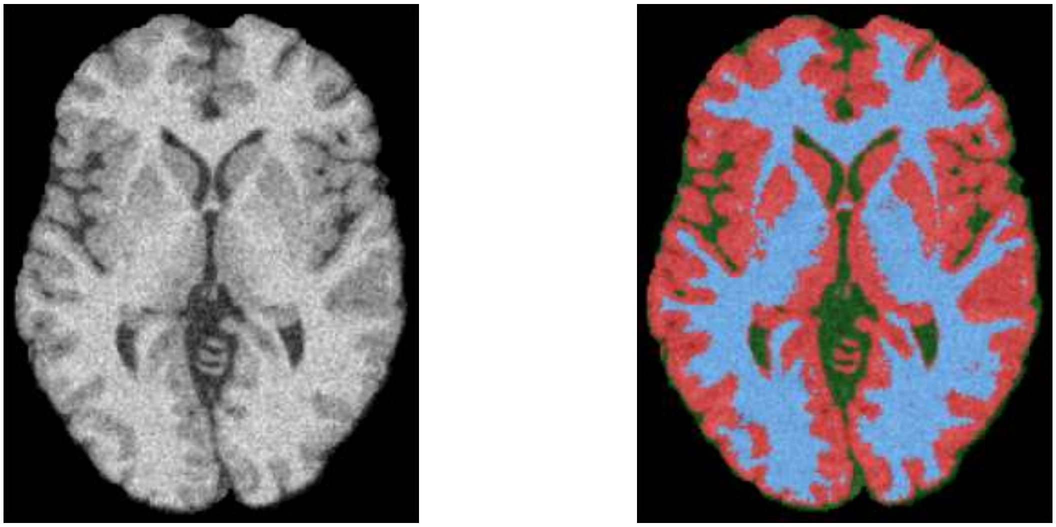

Figure 1: Left: Plane of an anatomical MRI scan. Right: Segmented anatomical image. In red, white matter; in blue, gray matter; in green, cerebrospinal fluid.

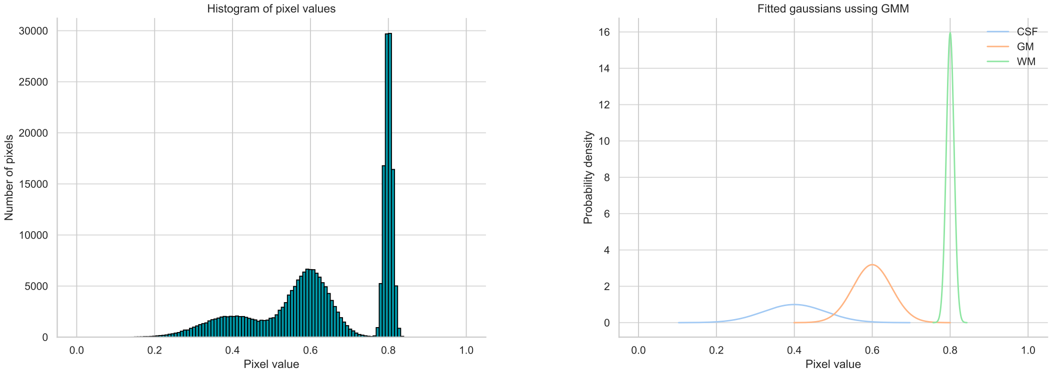

Figure 2: Left: Histogram of pixel values of the image (background excluded). Right: Gaussian distribution corresponding to the pixel values of the 3 tissue types.

Question 4: MRI case study

You are working as a data analyst in a neuroscience lab where they work with anatomical MRI images (magnetic resonance intensity), as the one shown in Figure 1, left. After pre-processing the images, you need to segment the image to infer what pixels correspond to white matter (WM), gray matter (GM) and cerebrospinal fluid (CSF) Figure 1, right.

One of the clinicians in the lab tells you that these 3 types of tissues are easy to distinguish in the images, that pixel intensity is mostly sufficient to distinguish them. Therefore, you decide to plot the histogram of pixel values and observe Figure 2.

It seems that indeed the distribution of pixel value (X) has 3 distinct modes and you decide to model it as such. One model that seems accurate to you is a mixture of gaussians (TK:Fig) where you model the whole pdf of pixel values as a combination of 3 gaussians with different means and variances:

$$p(X) = \pi_w \mathcal{N}(\mu_w, \sigma_w) + \pi_g \mathcal{N}(\mu_g, \sigma_g) + \pi_c \mathcal{N}(\mu_c, \sigma_c)$$

Where $\pi_w + \pi_g + \pi_c = 1$, $\pi_w, \pi_g, \pi_c \geq 0$. Here $\mathcal{N}(\mu, \sigma)$ denotes a gaussian distribution with mean μ and standard deviation σ.

a) At a first glance, this expression looks a bit strange to you and want to check whether $p(x)$ is a valid pdf. Check that $p(x)$ fullfils the properties of a pdf.

   Solution: To check that $p(x)$ is a valid pdf, we need to check two properties:

   (a) $p(x) \geq 0$ for all x.
   (b) $\int_{-\infty}^{+\infty} p(x)dx = 1$.

   Since the gaussian pdfs are non-negative for all x:

$$\pi_w \mathcal{N}(\mu_w, \sigma_w) + \pi_g \mathcal{N}(\mu_g, \sigma_g) + \pi_c \mathcal{N}(\mu_c, \sigma_c) \geq \pi_w + \pi_c + \pi_g \geq 0, \forall x$$

   For the second property:

$$\int_{-\infty}^{+\infty} p(x)dx = \pi_w \int_{-\infty}^{+\infty} \mathcal{N}(\mu_w, \sigma_w)dx + \pi_g \int_{-\infty}^{+\infty} \mathcal{N}(\mu_g, \sigma_g)dx + \pi_c \int_{-\infty}^{+\infty} \mathcal{N}(\mu_c, \sigma_c)dx$$

   Since $\int_{-\infty}^{+\infty} \mathcal{N}(\mu, \sigma)dx = 1$, we have:

$$\int_{-\infty}^{+\infty} p(x)dx = \pi_w + \pi_g + \pi_c = 1$$

   Therefore, $p(x)$ is a valid pdf.

b) To further understand this distribution, you want to know what the total mean of this distribution is. Compute $\mathbb{E}[X]$ as a function of π_w, π_g, π_c and the means and variances of the normal distributions.

   Solution: By linearity of the expectation operator, the total mean of the distribution is given by:

$$\mathbb{E}[X] = \pi_g \mathbb{E}[\mathcal{N}(\mu_g, \sigma_g)] + \pi_w \mathbb{E}[\mathcal{N}(\mu_w, \sigma_w)] + \pi_c \mathbb{E}[\mathcal{N}(\mu_c, \sigma_c)]$$

   We substitute the expectations of the normal distributions:

$$\mathbb{E}[X] = \pi_g \mu_g + \pi_w \mu_w + \pi_c \mu_c$$

   Therefore the mean of the mixture of gaussians is a weighted sum of the means of the individual gaussians, where the weights are given by the mixing coefficients π_w, π_g, π_c.

From this point onwards, to simplify the problem we will work only with two gaussians, so that the CSF is not segmented. Therefore, the pdf is now:

$$p(X) = \pi_w \mathcal{N}(\mu_w, \sigma_w) + (1 - \pi_w)\mathcal{N}(\mu_g, \sigma_g)$$

The clinician from your lab has been extremely nice and manually segmented the image for you (this is pain, but it is done a lot to train large neural networks!). You want to use this information to estimate the parameters $\pi_w, \mu_w, \mu_g, \sigma_w, \sigma_g$, so you introduce a new random variable Z which corresponds to the type of tissue. Therefore, for each pixel i in the image you have its pixel value x_i and its tissue type z_i which so that $z_i = 0$ for gray matter and $z_i = 1$ for white matter. The joint probability density of these variables is given by:

$$p(X, Z) = \pi_w^z \mathcal{N}(\mu_w, \sigma_w)^z (1 - \pi_w)^{1-z} \mathcal{N}(\mu_w, \sigma_w)^{1-z}$$

Note that substitution of $z = 0$ or $z = 1$ simply yields a gaussian distribution with the mean and variance for that particular tissue.+ You now want to ensure that the marginal distributions make sense in terms of known distributions-

c) First, show that the marginal distriburion corresponds to $p(X)$. Do it by marginalizing so that $p(X) = \sum_{z=0}^{1} p(X, Z)$

Solution: To compute the marginal distribution $p(X)$, we need to sum over the possible values of Z. If we evaluate first for $z = 0$:

$$p(X, Z = 0) = (1 - \pi_w)\mathcal{N}(\mu_g, \sigma_g)$$

And for $z = 1$:

$$p(X, Z = 1) = \pi_w \mathcal{N}(\mu_w, \sigma_w)$$

Therefore, the marginal distribution is given by:

$$p(X) = p(X, Z = 0) + p(X, Z = 1) = (1 - \pi_w)\mathcal{N}(\mu_g, \sigma_g) + \pi_w \mathcal{N}(\mu_w, \sigma_w)$$

d) Now compute the marginal $p(Z)$ by integrating over all possible values of x, that is $p(Z) = \int_{-\infty}^{+\infty} p(X, Z)dX$. Do it separately for $z = 0$ and $z = 1$.

Solution: As suggested, we compute $p(Z = 0)$ first:

$$p(Z = 0) = \int_{-\infty}^{+\infty} p(X, Z = 0)dx = (1 - \pi_w) \int_{-\infty}^{+\infty} \mathcal{N}(\mu_g, \sigma_g) = 1 - \pi_w$$

Similarly, for $z = 1$:

$$p(Z = 1) = \int_{-\infty}^{+\infty} p(X, Z = 1)dx = \pi_w \int_{-\infty}^{+\infty} \mathcal{N}(\mu_w, \sigma_w) = \pi_w$$

We can compact this into the pdf of a bernoulli distribution with $p = \pi_w$:

$$p(Z) = \pi_w^z (1 - \pi_w)^{1-z}$$

At this point you feel more prepared to estimate the parameters from the image. Assume you have n pixels, each of them has a pixel value x_i and z_i.

e) First, show that the log-likelihood as a function of the 5 parameters is given by:

$$l(\pi_w, \mu_w, \sigma_w, \mu_g, \sigma_g) = \sum_{i=1}^{n} z_i \big(\log(\pi_w) - \frac{1}{2} \log\big(2\pi\sigma_w^2\big) - \frac{(x_i - \mu_w)^2}{2\sigma_w^2} \big) + (1 - z_i) \big(\log(\pi_g) - \frac{1}{2} \log\big(2\pi\sigma_g^2\big) - \frac{(x_i - \mu_g^2)}{2\sigma_g^2} \big)$$

Solution: The likelihood of n independent samples is given by: $\mathcal{L}(\pi_w, \mu_w, \sigma_w, \mu_g, \sigma_g) = \prod_{i=1}^{n} p(x_i, z_i)$

Upon substituting the expression for $p(x_i, z_i)$ we have and applying the logarithm:

$$l(\pi_w, \mu_w, \sigma_w, \mu_g, \sigma_g) = \sum_{i=1}^{n} z_i (\log(\pi_w) + \log(\mathcal{N}(\mu_w, \sigma_w)) + (1 - z_i)(\log(1 - \pi_w) + \log(\mathcal{N}(\mu_g, \sigma_g)))$$

Substituting the expressions for the normal distributions yields the mentioned expression.

f) At this point you wonder if your job is worth dealing with this expression... Nonetheless, you persevere and and decide to start by find the maximum likelihood estimate of π_w. By $\frac{\partial l}{\partial \pi_w} = 0$, show that $\hat{\pi}_w = \frac{1}{n} \sum_{i=1}^{n} z_i$.

Solution: We compute the derivative with respect the parameter

$$\frac{\partial l}{\partial \pi_w} = \sum_{i=1}^{n}(\frac{z_i}{\hat{\pi}_w} - \frac{1-z_i}{1-\hat{\pi}_w}) = 0$$

We solve to find $\hat{\pi}_w$:

$$\frac{1}{\hat{\pi}_w}\sum_{i=1}^{n} z_i - \frac{1}{1-\hat{\pi}_w}(n - \sum_{i=1}^{n} z_i) = \frac{(1-\hat{\pi}_w)\sum_{i=1}^{n} z_i - \hat{\pi}_w n + \pi \sum_{i=1}^{n} z_i}{\hat{\pi}_w(1-\hat{\pi}_w)}$$

Discarding the trivial solutions $\hat{\pi}_w = 0$ and $\hat{\pi}_w = 1$, we cancel out the denominator and we have:

$$(1-\hat{\pi}_w)\sum_{i=1}^{n} z_i - \hat{\pi}_w n + \hat{\pi}_w \sum_{i=1}^{n} z_i = \sum_{i=1}^{n} z_i - \hat{\pi}_w n = 0$$

Therefore:

$$\hat{\pi}_w = \frac{1}{n}\sum_{i=1}^{n} z_i$$

Which is simply the fraction of white matter pixels in the image.

g) That was not so bad... (hopefully). Similarly, show now that the expressions for μ_w and σ_w are given by:

$$\hat{\mu}_w = \frac{\sum_{i=1}^{n} z_i x_i}{\sum_{i=1}^{n} z_i}$$

$$\hat{\sigma}_w^2 = \frac{\sum_{i=1}^{n} z_i (x_i - \mu_w)^2}{\sum_{i=1}^{n} z_i}$$

Solution: We substitute the pdf of the normal and compute the derivative with respect the parameter μ_w:

$$\frac{\partial l}{\partial \mu_w} = \sum_{i=1}^{n} z_i(\frac{x_i - \hat{\mu}_w}{\sigma_w^2}) = 0$$

Since σ_w^2 is positive, we can multiply by it and we have:

$$\sum_{i=1}^{n}(z_i x_i) - \hat{\mu}_w \sum_{i=1}^{n} z_i = 0$$

Therefore:

$$\hat{\mu}_w = \frac{\sum_{i=1}^{n} z_i x_i}{\sum_{i=1}^{n} z_i}$$

For the variance, we compute the derivative with respect to σ_w^2:

$$\frac{\partial l}{\partial \sigma_w^2} = \sum_{i=1}^{n} z_i(\frac{(x_i - \hat{\mu}_w)^2}{2\sigma_w^4} - \frac{1}{2\sigma_w^2}) = 0$$

Again, since σ_w^2 is positive, we can multiply by it and we have:

$$\sum_{i=1}^{n} z_i(x_i - \hat{\mu}_w)^2 - \hat{\sigma}_w^2 \sum_{i=1}^{n} z_i = 0$$

Therefore:

$$\hat{\sigma}_w^2 = \frac{\sum_{i=1}^{n} z_i(x_i - \hat{\mu}_w)^2}{\sum_{i=1}^{n} z_i}$$

With this information, you are ready to take the segmented image and fit your gaussian mixture model! As a final step, you want to compare the average pixel value of the white matter μ_w to some reference value that a colleague has given to you. Assume that you have an image of $n = 512^2$ pixels and that 30 % of them have been identified as white matter. The value you have computed for the mean is $\hat{\mu_w} = 0.802$ and the variance is $\hat{\sigma_w} = 0.01$; the mean given by your colleague is $\mu_w = 0.8$. You decide to do a t-test to test whether the to values for the mean are different or not.

h) Write down the test by stating the null hypothesis and the alternative.

Solution:

- $H_0 : \mu_w = 0.8$
- $H_1 : \mu_w \neq 0.8$

Note that since we test for difference the test is two-sided.

i) Compute the t-statistic for the test, think carefully what is the number of samples used to compute the mean.

Solution: Since we have $n = 512^2$ pixels and 30% of them are white matter, we have $n_{WM} = 0.3 \cdot n = 0.3 \cdot 512^2$. Therefore, the t-statistic is given by:

$$t = \frac{\hat{\mu}_w - \mu_w}{\hat{\sigma}_w / \sqrt{n_{WM}}} = \frac{0.802 - 0.8}{0.01/\sqrt{0.3 \cdot 512^2}} = 56.09$$

We already see that this is very large.

j) To what reference distribution should you compare in this case? Do you think it is appropiate to approximate it with a normal?

Solution: The reference distribution is a t-distribution with $n_{WM} - 1$ degrees of freedom. As the number of samples is very large, we could also use a normal distribution since the t-distribution converges to a normal distribution as the degrees of freedom increase.

k) What is the p-value of the test? How do you interpret this result?

Solution: Seeing the t-statistic, the p-value is essentially 0. This is due to performing the test with an extremely large number of samples. This is called the fallacy of hypothesis testing where, as the number of samples used for test increases, the test becomes more sensitive to small differences and everything becomes significant. Remember that significance only relates to the type I error (false positives) and not to the type II error (false negatives). The test might be significance, but we can still be having an extremely large type II error.

Question 5: Sampling Distribution and Hypothesis Testing in Gene Expression Data

Problem: In a study of gene expression for *Gene X*, the expression levels in healthy individuals are assumed to be normally distributed. Let

$$X_1, X_2, \ldots, X_{16} \sim N(\mu, 100)$$

denote the expression levels measured from a random sample of 16 individuals. We are interested in:

a) Deriving the sampling distribution of the sample mean $\bar{X}$.

b) Testing the hypothesis given that the known benchmark expression level in healthy individuals is $\mu_0 = 50$ and an observed sample mean is $\bar{x} = 55$. In particular, test

$$H_0 : \mu = 50 \quad \text{vs.} \quad H_1 : \mu > 50,$$

by computing the corresponding z-score and p-value, and interpret the result.

Solution:

a) **Sampling Distribution of $\bar{X}$:**
Since the individual observations are independent and normally distributed, the sum

$$S = \sum_{i=1}^{16} X_i$$

follows

$$S \sim N(16\mu, 16 \cdot 100) = N(16\mu, 1600).$$

The sample mean is given by

$$\bar{X} = \frac{S}{16}.$$

By the properties of normal distributions under linear transformations, it follows that

$$\bar{X} \sim N\left(\mu, \frac{1600}{16^2}\right) = N\left(\mu, \frac{100}{16}\right) = N(\mu, 6.25).$$

b) **Hypothesis Testing:**
Under $H_0 : \mu = 50$, the sampling distribution of $\bar{X}$ is

$$\bar{X} \sim N\left(50, \frac{100}{16}\right) = N(50, 6.25).$$

The z-score is computed as

$$z = \frac{\bar{x} - \mu_0}{\sqrt{100/16}} = \frac{55 - 50}{\sqrt{6.25}} = \frac{5}{2.5} = 2.$$

For the one-sided alternative $H_1 : \mu > 50$, the p-value is

$$p = P(Z \geq 2) \approx 0.0228.$$

Interpretation: A p-value of approximately 0.0228 indicates that there is a 2.28% probability of observing a sample mean of 55 or higher if the true mean were 50. This low p-value provides evidence against H_0, suggesting that the gene expression level in healthy individuals is significantly higher than 50.

Question 6: Affine Transformations of Random Variables

Let $\mathbf{X}$ be a d-dimensional random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Let $\mathbf{Y} = A\mathbf{X} + \mathbf{b}$, where A is a $n \times d$ matrix and $\mathbf{b}$ is a n-dimensional vector.

(a) Show that the mean of $\mathbf{Y}$ is $A\boldsymbol{\mu} + \mathbf{b}$.

(b) Show that the covariance matrix of $\mathbf{Y}$ is $A\boldsymbol{\Sigma}A^T$.

Solution:

(a)

$$\begin{aligned}\mathbb{E}(\mathbf{Y}) &= \mathbb{E}(A\mathbf{X} + \mathbf{b}) = \mathbb{E}(A\mathbf{X}) + \mathbb{E}(\mathbf{b}) \\ &= A\mathbb{E}(\mathbf{X}) + \mathbf{b} = A\boldsymbol{\mu} + \mathbf{b}\end{aligned}$$

(b)

$$\begin{aligned}\mathrm{Var}(\mathbf{Y}) &= \mathbb{E}\left[(\mathbf{Y} - \mathbb{E}\mathbf{Y})(\mathbf{Y} - \mathbb{E}\mathbf{Y})^T\right] \\ &= \mathbb{E}\left[(A\mathbf{X} + \mathbf{b} - A\boldsymbol{\mu} - \mathbf{b})(A\mathbf{X} + \mathbf{b} - A\boldsymbol{\mu} - \mathbf{b})^T\right] \\ &= \mathbb{E}\left[(A\mathbf{X} - A\boldsymbol{\mu})(A\mathbf{X} - A\boldsymbol{\mu})^T\right] \\ &= \mathbb{E}\left[A(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T A^T\right] \\ &= A\mathbb{E}\left[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T\right] A^T \\ &= A\boldsymbol{\Sigma}A^T\end{aligned}$$

Question 7: Maximum Likelihood Estimation

Let $x_1, x_2, \ldots, x_n$ be independent samples from the following distribution:

$$P(x \mid \theta) = \theta x^{-\theta-1} \quad \text{where } \theta > 1,\ x \geq 1$$

Find the maximum likelihood estimator of θ.

Solution:

$$\begin{aligned}L(\theta \mid x_1, x_2, \ldots, x_n) &= \prod_{i=1}^{n} \theta x_i^{-\theta-1} = \theta^n \prod_{i=1}^{n} x_i^{-\theta-1} \\ \ln L(\theta \mid x_1, x_2, \ldots, x_n) &= n \ln \theta - (\theta + 1) \sum_{i=1}^{n} \ln x_i \\ \frac{\delta \ln L}{\delta \theta} &= \frac{n}{\theta} - \sum_{i=1}^{n} \ln x_i = 0 \\ \theta_{mle} &= \frac{n}{\sum_{i=1}^{n} \ln x_i}\end{aligned}$$

Since $\theta > 1$, any $\theta_{mle} \leq 1$ has a zero probability of generating any data, so our best estimate of θ when $\theta_{mle} \leq 1$ is $\theta_{mle} = 1$. Therefore, the final answer is $\theta_{mle} = \max(1, \frac{n}{\sum_{i=1}^{n} \ln x_i})$.

However, we will still accept $\theta_{mle} = \frac{n}{\sum_{i=1}^{n} \ln x_i}$.

Question 8: Linear Regression

Recall that if we model our input data as linear plus Gaussian noise in the y-values, $Y \mid \mathbf{x} \sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}, \sigma^2)$, then the maximum likelihood estimator is the $\mathbf{w}$ that minimizes the residual sum of squares

$$\sum_{i=1}^{n} (X_i^\top \mathbf{w} - y_i)^2,$$

where the samples are $X_1, X_2, \ldots, X_n$ and their targets are $y_1, y_2, \ldots, y_n$.

Let's model noise with a Laplace distribution instead of a normal distribution. The probability density function (PDF) of Laplace(μ, b) is

$$P(y) = \frac{1}{2b} \exp\left(-\frac{|y-\mu|}{b}\right).$$

Show that if we model our input data as a line plus Laplacian noise in the y-values, i.e.

$$Y \mid \mathbf{x} \sim \text{Laplace}(\mathbf{w}^\top \mathbf{x}, b),$$

then the maximum likelihood estimator is the $\mathbf{w}$ that minimizes the *sum of absolute residuals*

$$\sum_{i=1}^{n} \left| X_i^\top \mathbf{w} - y_i \right|.$$

Solution: We wish to maximize the log-likelihood:

$$\begin{aligned}
\ln \prod_{i=1}^{n} P(y_i \mid X_i) &= \sum_{i=1}^{n} \ln\left(\frac{1}{2b} e^{-|y_i - X_i^\top \mathbf{w}|/b}\right) \\
&= -\frac{1}{b} \sum_{i=1}^{n} |y_i - X_i^\top \mathbf{w}| - n \ln(2b),
\end{aligned}$$

which is equivalent to minimizing

$$\sum_{i=1}^{n} |X_i^\top \mathbf{w} - y_i|$$

.

Question 9: Design matrix

You have the data shown in Table 1 and you want to build a design matrix for a linear regression model. The 3 variables are categorical and take the following values:

- R: 1, 2
- C: 1, 2, 3
- T(treatment): A, B, C

You want to build a design matrix with R, C and T to predict a variable of interest y, such that:

$$y_i = \mu_i + \varepsilon_i$$

Where $\mu_i = f(r, c, t)$ and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

R	C	T
1	1	A
1	3	B
1	3	C
2	1	B
2	2	A
2	3	B

Table 1: Recorded samples

1. Write down the design matrix for this data. Hint: The model should have 6 parameters (why?) and the design matrix should have 6 columns.
2. Write down the regression model in matricial form.
3. Interpret the parameters of the model in terms of the original variables.

(a) The design matrix is constructed as follows:

$$\left[\begin{array}{c|cc|ccc|ccc}
 & \multicolumn{2}{c|}{\overbrace{\hspace{3em}}^{R}} & \multicolumn{3}{c|}{\overbrace{\hspace{5em}}^{C}} & \multicolumn{3}{c}{\overbrace{\hspace{5em}}^{T}} \\
\text{Intercept} & 1 & \color{red}{2} & 1 & 2 & \color{red}{3} & A & \color{red}{B} & C \\
\hline
1 & 1 & \color{red}{0} & 1 & 0 & \color{red}{0} & 1 & \color{red}{0} & 0 \\
1 & 1 & \color{red}{0} & 0 & 0 & \color{red}{1} & 0 & \color{red}{1} & 0 \\
1 & 1 & \color{red}{0} & 0 & 0 & \color{red}{1} & 0 & \color{red}{0} & 1 \\
1 & 0 & \color{red}{1} & 1 & 0 & \color{red}{0} & 0 & \color{red}{1} & 0 \\
1 & 0 & \color{red}{1} & 0 & 1 & \color{red}{0} & 1 & \color{red}{0} & 0 \\
1 & 0 & \color{red}{1} & 0 & 0 & \color{red}{1} & 0 & \color{red}{1} & 0
\end{array}\right]$$

- We add a column for the intercept.
- The variables are categorical, so we use dummy codification.
- For each variable, we drop one column (red) as the sum of columns for a variable equals the column for the intercept, leading to linear dependence, and we require a full-rank matrix.
- The categories dropped for each variable become the **reference categories**.
- The choice of column to drop is arbitrary among categories for each variable.

(b)

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \rho_1 \\ \rho_2 \\ \delta_1 \\ \delta_2 \\ \tau_A \\ \tau_C \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{pmatrix}$$

$$= \begin{pmatrix} \beta_0 + \rho_1 + \delta_1 + \tau_A \\ \beta_0 + \rho_1 \\ \beta_0 + \rho_1 + \tau_C \\ \beta_0 + \delta_1 \\ \beta_0 + \delta_2 + \tau_A \\ \beta_0 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{pmatrix}$$

(c)

$$\beta_0 \text{ is the } \mathbb{E}[Y_i] \text{ when } X_i = (R_i, C_i, T_i) = (2, 3, B)$$

Think about unit 6, we “expect a value” of β_0 for it.

$$\delta_1 = \mathbb{E}[Y_i] - \mathbb{E}[Y_j] \text{ when } X_i = (r, 1, t), \quad X_j = (r, 3, t)$$

So it is the effect of $C = 1$ **compared to** $C = 3$, the reference.

Think about the comparison between unit 4 and unit 6.

The same of course holds for δ_2.

$$\rho_1 = \mathbb{E}[Y_i] - \mathbb{E}[Y_j] \text{ when } X_i = (1, c, t), \quad X_j = (2, c, t)$$

So it is the effect of $r = 1$ **compared to** $r = 2$, the reference.

Think about the comparison between unit 2 and unit 6.

Similarly, τ_A and τ_C are the effects of treatments A and C **compared to treatment B**, the reference (the column we dropped).