

## Data Science and AI for Neuroscience Summer Course

Instructors: Tara Chari

[tchari@caltech.edu](mailto:tchari@caltech.edu)

### Group Exercises: Regression Basics

#### Exercise 1

Consider the following gene expression matrix  $G$ , representing the expression level of four genes across three cells. The  $i$ -th row of  $G$  is the number of transcribed molecule counts of genes in the  $i$ -th cell, and the  $j$ -th column is the number of transcribed molecule counts of the  $j$ -th gene across all cells:

$$G = \begin{bmatrix} 1 & 2 & 3 & 3 \\ 3 & 1 & 9 & 4 \\ 1 & 4 & 3 & 5 \end{bmatrix}.$$

- (a) One of the most fundamental characteristics of a matrix  $A$  is its rank, which corresponds to the maximal number of linearly independent columns/rows of  $A$ . A set of vectors  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$  are linearly independent if the equation

$$a_1 \mathbf{v}_1 + a_2 \mathbf{v}_2 + \dots + a_n \mathbf{v}_n = \mathbf{0}$$

can only be satisfied by  $a_i = 0$  for  $i = 1, \dots, n$ .

- i. Find the rank of  $G$

**A:** One possible way to compute the rank is using Gaussian elimination:

$$G = \begin{bmatrix} 1 & 2 & 3 & 3 \\ 3 & 1 & 9 & 4 \\ 1 & 4 & 3 & 5 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 2 & 3 & 3 \\ 0 & -5 & 0 & -5 \\ 0 & 2 & 0 & 2 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 2 & 3 & 3 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

There are only two non-zero rows. Therefore, the rank of  $G$  is 2.

- ii. What does the rank of  $G$  suggest in terms of the relationship in expression level between the four genes?

**A:** It suggests that the expression level of four genes are not independent and two genes are redundant in terms of expression information.

- (b) Suppose you are interested in selecting only highly expressed genes for further analysis.

- i. Find a vector  $v$  such that the product  $v^T G$  contains the mean expression level for every gene in  $G$ .

**A:** Since we have three cells,  $v = [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]^T$ .

(c) In many applications, we are interested in finding cells that are “similar” to one another in terms of their gene expression profiles, where the gene expression profile of a cell corresponds to a row vector of the gene expression matrix. Finding how similar two cells are amounts to computing how “close” their corresponding expression vectors are under a particular distance function.

- i. Consider three valid distance functions between a pair of vectors  $\mathbf{x}$  and  $\mathbf{y}$ , the  $L_1$  distance,  $L_2$  distance, and cosine similarity ( $c$ ),

$$L_1(\mathbf{x}, \mathbf{y}) := \|\mathbf{x} - \mathbf{y}\|_1 = \sum_{i=1}^n |x_i - y_i|,$$

$$L_2(\mathbf{x}, \mathbf{y}) := \|\mathbf{x} - \mathbf{y}\|_2 = \sqrt{\sum_{i=1}^n (x_i - y_i)^2},$$

$$c(\mathbf{x}, \mathbf{y}) := \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2} = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2} \sqrt{\sum_{i=1}^n y_i^2}}.$$

Note that the cosine similarity is simply a measure of the cosine of the angle between two vectors. For each distance function above, construct a distance matrix  $D$ , where  $D_{ij}$  denotes the distance between the  $i$ -th and  $j$ -th row/cell of  $G$ .

**A:** For the  $L_1$  distance, using above definition, we have:

$$D_{12}^{L_1} = 2 + 1 + 6 + 1 = 10$$

$$D_{13}^{L_1} = 0 + 2 + 0 + 2 = 4$$

$$D_{23}^{L_1} = 2 + 3 + 6 + 1 = 12$$

$$\therefore D^{L_1} = \begin{bmatrix} 0 & 10 & 4 \\ 10 & 0 & 12 \\ 4 & 12 & 0 \end{bmatrix}$$

For the  $L_2$  distance, we have:

$$D_{12}^{L_2} = \sqrt{2^2 + 1^2 + 6^2 + 1^2} = \sqrt{42}$$

$$D_{13}^{L_2} = \sqrt{0^2 + 2^2 + 0^2 + 2^2} = \sqrt{8}$$

$$D_{23}^{L_2} = \sqrt{2^2 + 3^2 + 6^2 + 1^2} = \sqrt{50}$$

$$\therefore D^{L_2} = \begin{bmatrix} 0 & \sqrt{42} & \sqrt{8} \\ \sqrt{42} & 0 & \sqrt{50} \\ \sqrt{8} & \sqrt{50} & 0 \end{bmatrix}$$

For the cosine similarity, we have:

$$D_{12}^c = \frac{3 + 2 + 27 + 12}{\sqrt{1^2 + 2^2 + 3^2 + 3^2} \sqrt{3^2 + 1^2 + 9^2 + 4^2}} = \frac{44}{\sqrt{23} \sqrt{107}} = 0.887$$

$$D_{13}^c = \frac{1 + 8 + 9 + 15}{\sqrt{1^2 + 2^2 + 3^2 + 3^2} \sqrt{1^2 + 4^2 + 3^2 + 5^2}} = \frac{33}{\sqrt{23} \sqrt{51}} = 0.964$$

$$D_{23}^c = \frac{3 + 4 + 27 + 20}{\sqrt{1^2 + 4^2 + 3^2 + 5^2} \sqrt{3^2 + 1^2 + 9^2 + 4^2}} = \frac{44}{\sqrt{51} \sqrt{107}} = 0.731$$

$$\therefore D^c = \begin{bmatrix} 1 & 0.887 & 0.964 \\ 0.887 & 1 & 0.731 \\ 0.964 & 0.731 & 1 \end{bmatrix}$$

- ii. For each distance function, which two cells in  $G$  are most similar to one another? For  $L_1$  and  $L_2$  distance, this corresponds to a pair of row vectors with the smallest distance, whereas for the cosine similarity, this corresponds to a pair of row vectors with the smallest angle.

**A:** For all three measures, cell 1 and cell 3 are most similar to one another.

- iii. When comparing the gene expression profile of different cells, it may be useful to compare the relative gene expression level instead of absolute expression level (consider comparing cells pooled from different growth condition). In such a case, we want to use a distance function  $f$  that is invariant to scaling. Specifically, we would like  $f$  to satisfy the property that for any scaling constant  $a \geq 0$ ,

$$f(\mathbf{x}, \mathbf{y}) = f(a\mathbf{x}, a\mathbf{y}).$$

Show that the cosine similarity satisfies this property, while the  $L_1$  and  $L_2$  distances do not.

**A:** For the  $L_1$  distance,

$$L_1(a\mathbf{x}, a\mathbf{y}) := \|a\mathbf{x} - a\mathbf{y}\|_1 = \sum_{i=1}^n a|x_i - y_i| = a\|\mathbf{x} - \mathbf{y}\|_1,$$

For the  $L_2$  distance,

$$L_2(a\mathbf{x}, a\mathbf{y}) := \|a\mathbf{x} - a\mathbf{y}\|_2 = \sqrt{\sum_{i=1}^n a^2(x_i - y_i)^2} = a\|\mathbf{x} - \mathbf{y}\|_2$$

For the cosine similarity,

$$c(a\mathbf{x}, a\mathbf{y}) := \frac{a\mathbf{x} \cdot a\mathbf{y}}{\|a\mathbf{x}\|_2 \|a\mathbf{y}\|_2} = a^2 \frac{\mathbf{x} \cdot \mathbf{y}}{a^2 \|\mathbf{x}\|_2 \|\mathbf{y}\|_2} = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2}$$

## Exercise 2

Consider a linear model involving variables  $\mathbf{x}$  and  $\mathbf{y}$ , i.e.

$$\mathbf{y} = A\mathbf{x} + \epsilon \quad (1)$$

where  $\epsilon$  represents random “noise”. We are often interested in estimating  $\mathbf{x}$  given  $\mathbf{y}$  and  $A$ . For example, if we are trying to fit a linear model between some variable of interest  $\mathbf{y}$  and expression levels of different genes, (1) corresponds to the following,

$$\mathbf{y} = X\beta + \epsilon, \quad (2)$$

where  $X$  is the observed gene expression matrix and  $\beta$  is a vector of regression coefficients that are to be estimated. From the previous Lectures, you learned about one such estimator called the least square estimator,  $\hat{\beta} = X^\dagger \mathbf{y} = (X^T X)^{-1} X^T \mathbf{y}$ . Here we will investigate why  $\hat{\beta}$  is the least squares estimator.

- (a) Derive the least square estimator  $\hat{\beta} = X^\dagger \mathbf{y}$  by showing that  $\hat{\beta}$  minimizes the squared error, i.e.

$$\hat{\beta} = \arg \min_{\beta} \|X\beta - \mathbf{y}\|_2^2$$

(Hint: express the squared error as a product of vectors and take its derivative with respect to  $\beta$ )

**A:**

$$\begin{aligned} \frac{d\|X\beta - \mathbf{y}\|_2^2}{d\beta} &= 2(X\beta - \mathbf{y})^\top X \\ 2(X\hat{\beta} - \mathbf{y})^\top X &= 0 \\ \hat{\beta}^\top X^\top X - \mathbf{y}^\top X &= 0 \\ X^\top X\hat{\beta} &= X^\top \mathbf{y} \\ \hat{\beta} &= (X^\top X)^{-1} X^\top \mathbf{y} \end{aligned}$$

- (b) A linear estimator  $\hat{\beta}$  is unbiased if  $\hat{\beta} = \beta$  whenever  $\epsilon = 0$ . Show that the least square estimator is unbiased.

**A:** Assume  $\epsilon = 0$  and write the least square estimator as follows:

$$\begin{aligned}\hat{\beta} &= (X^T X)^{-1} X^T y \\ &= (X^T X)^{-1} X^T (X\beta + \epsilon) \\ &= (X^T X)^{-1} X^T X\beta \\ &= \beta\end{aligned}$$

Therefore  $\hat{\beta} = \beta$  and  $\hat{\beta}$  is unbiased.

*Comment:* In the context of scRNA-seq analysis, gene expression matrix  $X$  is often rank-deficient, where one of the columns (genes) of is a linear combination of the others. In this case,  $X^T X$  has no inverse, and so we must use a different estimator instead of  $\hat{\beta} = X^\dagger y = (X^T X)^{-1} X^T y$ .

*Comment:* To deal with rank-deficient data matrix  $X$ , we use the Moore-penrose inverse  $X^+$  instead  $X^\dagger$ ,

$$X^+ = \lim_{\delta \rightarrow 0} (X^T X + \delta^2 I)^{-1} X^T. \quad (3)$$

The Moore-penrose inverse is well-defined even when  $X^T X$  is not invertible. Furthermore, it generates a solution  $\hat{\beta} = X^+ y$  to the least-square problem (In fact,  $\hat{\beta}$  has the smallest  $l_2$  norm among all least-square solutions).

### Exercise 3

In this question, we will explore the relationship between dependence of random variables and their partial correlation.

Recall that Pearson correlation is denoted as  $\rho$  where

$$\rho_{XY} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\mathbb{E}[X Y] - \mathbb{E}[X] \mathbb{E}[Y]}{\sqrt{\mathbb{E}[X^2] - (\mathbb{E}[X])^2} \sqrt{\mathbb{E}[Y^2] - (\mathbb{E}[Y])^2}} \quad (4)$$

- (a) Consider the activity of two independently expressed genes as binary random variables  $X_1$  and  $X_2$ , where  $P(X_1 = 0) = P(X_2 = 0) = \frac{1}{2}$  and  $P(X_1 = 1) = P(X_2 = 1) = \frac{1}{2}$ . Furthermore, consider the sum of these two random variables  $Y = X_1 + X_2$  as the total number of active genes. Show that  $\rho_{X_1 X_2 \cdot Y} \neq 0$  even though  $X_1 \perp\!\!\!\perp X_2$ , thus independence does not imply zero partial correlation, where the partial correlation is defined as follows,

$$\rho_{XY \cdot Z} = \frac{\rho_{XY} - \rho_{XZ}\rho_{ZY}}{\sqrt{1 - \rho_{XZ}^2}\sqrt{1 - \rho_{ZY}^2}},$$

with  $\rho_{XY}$  being the regular Pearson correlation.

**A:** To show that  $\rho_{X_1X_2 \cdot Y} \neq 0$ , we just need to show  $\rho_{X_1X_2} \neq \rho_{X_1Y}\rho_{X_2Y}$ , with  $\rho_{XY}$  being the regular Pearson correlation.

Since  $X_1$  and  $X_2$  are independent,  $\rho_{X_1X_2} = \text{cov}(X_1, X_2) = 0$ .

Thus, we need to show that  $\rho_{X_1Y}\rho_{X_2Y} \neq 0$ . By symmetry,  $\rho_{X_1Y} = \rho_{X_2Y}$ . So we only need to show that  $\rho_{X_1Y} \neq 0$  which means that  $\text{cov}(X_1, Y) \neq 0$ .

$$\begin{aligned} E[X_1] &= \frac{1}{2} \\ E[Y] &= 1 \\ E[X_1Y] &= 1 \times P(X_1 = 1, X_2 = 0, Y = 1) + 2 \times P(X_1 = 1, X_2 = 1, Y = 2) \\ &= 1 \times \frac{1}{4} + 2 \times \frac{1}{4} = \frac{3}{4} \\ \text{cov}(X_1, Y) &= E[X_1Y] - E[X_1]E[Y] = \frac{3}{4} - \frac{1}{2} \neq 0 \end{aligned}$$

Therefore,  $\rho_{X_1X_2 \cdot Y} \neq 0$ .