Data Science and AI for Neuroscience Summer Course

Instructor: Tara Chari
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. Each of the i rows is a cell, and each of the j columns is a gene. Thus 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 $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ are linearly independent if the equation

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

can only be satisfied by $a_i = 0$ for $i = 1, \ldots, n$. This means no vector is a linear combination of the other vectors.

i. Find the rank of G

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

(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.

(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 (see diagram below):

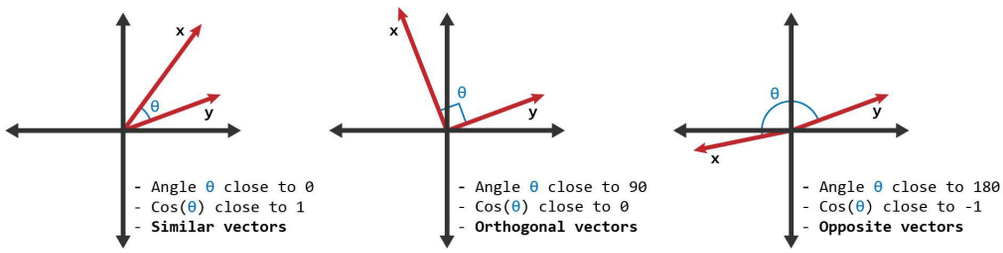

Figure 1: Examples of cosine distance between vectors from https://www.learndatasci.com/glossary/cosine-similarity/

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.

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.

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 conditions). 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 constants $a \geq 0$ and $b \geq 0$,

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

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

Exercise 2

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

$$\mathbf{y} = A\mathbf{x} + \epsilon \tag{1}$$

where ϵ 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 $\mathbf{y}$ (e.g. phenotype, protein activity etc) and expression levels of different genes, (1) corresponds to the following,

$$\mathbf{y} = X\beta + \epsilon, \tag{2}$$

where X is the observed gene expression matrix and β is a vector of regression coefficients that are to be estimated. From the lectures, you learned about one such estimator called the least square estimator, $\hat{\beta} = X^{\dagger}y = (X^TX)^{-1}X^Ty$. 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 β)

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

$$X^{+} = \lim_{\delta \to 0} \left(X^T X + \delta^2 I\right)^{-1} X^T. \tag{3}$$

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 ρ where

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

(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 ρ_{XY} being the regular Pearson correlation.