Bi/BE/CS 183 2022-2023
Instructor: Lior Pachter
TAs: Tara Chari, Meichen Fang, Zitong (Jerry) Wang

Problem Set 2

Submit your solutions as a single PDF file via Canvas by **8am Tuesday January 24th**.

- If writing up problems by hand, please use a pen and not a pencil, as it is difficult to read scanned submission of pencil work. Typed solutions are preferred.
- For problems that require coding, Colab notebooks will be provided. Please copy and save the shared notebook and edit your own copy, which you should then submit by including a clickable link in your submitted homework. Prior to submission make sure that you code runs from beginning to end without any error reports.

  See the class Intro to Colab on how to produce a shareable link for your notebook.

Problem 1 (16 points)
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 $\mathbf{y}$ 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. In class, you learned about one such estimator called the least square estimator, $\hat{\beta} = X^{\dagger}y = (X^TX)^{-1}X^Ty$.

In the context of scRNA-seq analysis, gene expression matrix X is often rank-deficient, where at least one of the column of X is a linear combination of the other columns. In this case, X^TX has no inverse, and so we must use a different estimator instead of $\hat{\beta}$.

(a) (8 points) Show that if X^TX is not invertible, then at least one column of X is a linear combination of the others. This implies that there is only one way in which problem with invertibility of X^TX may arise

(b) (8 points) Show that when X is rank-deficient, the least-square problem does not admit a unique solution.

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

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

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

Problem 2 (16 points)

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

(a) (8 points) 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 the partial correlation between X and Y given Z, denoted $\rho_{X_1 X_2 \cdot Y}$, is non-zero 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.

(b) (8 points) Partial correlation is meant to measure the relationship between two random variables after removing any linear dependency with a third random variable. Consider the set of random variables X, Y and Z with strong non-linear dependence,

$$Y = X^2 + Z.$$

Show that the partial correlation $\rho_{XY \cdot Z} = 0$ when X and Z are independent, standard Gaussians, even though X and Y are not independent given Z. Therefore, zero partial correlation does not imply conditional independence. Note that zero partial correlation does imply conditional independence under the special condition that all variables are jointly normal.

Problem 3 (12 points)

A RNAseq analysis pipeline typically starts by removing cells whose total RNA level is below a threshold value, as these are likely to be artifacts (e.g. empty droplets with spurious RNA detected). We will explore a potential issue with this pre-processing step when the threshold value is not chosen with care. Suppose the expression level of two genes, X_1 and X_2, can be modeled as follows,

$$X_1 = Z + \epsilon_1,$$
$$X_2 = Z + \epsilon_2,$$

where Z is a negative binomial random variable with $\mathbb{E}(Z) = n(1-p)/p$ and $\text{Var}(Z) = n(1-p)/p^2$, ϵ_1 and ϵ_2 are independent, identically distributed Poisson random variable with $\mathbb{E}(\epsilon_1) = \mathbb{E}(\epsilon_1) = \lambda$ and are also both independent of Z. For this question, set $\lambda = 40$, $n = 1$, $p = 0.3$.

We provide a blank Problem 3 notebook here.

(a) (4 points) Write a script that generates 100,000 pairs of $(X_1\ X_2)$ corresponding to 100,000 cells, and compute the correlation between the set of sample values.

(b) (4 points) Using the same set of sample values, create a filtered set by removing all paired instances where $X_1 + X_2 < 60$. Compute the correlation between the remaining pair of values in the filtered set, and make a scatter plot showing the original and the filtered set of points (use different colors).

(c) (4 points) Repeat part (b) using a different condition of $X_1 + X_2 < 80$, report the correlation of the resulting data points and a corresponding scatter plot, how does it compare with your result in part (a) and (b)?

Problem 4 (28 points)

Here you will explore how to use (1) linear regression to model gene count relationships, and investigate the assumptions these models will make. Utilizing the metadata from single-cell datasets, you will also apply (2) partial correlations to remove the influence of possibly confounding variables from your calculations of correlation between genes and their expression profiles. See the Problem 4 notebook here.

Your edited version of the notebook *must be submitted* for this problem. Reminder to check that your notebook runs all the way through with the the `Runtime` → `Restart` and `Runtime` → `Run All` commands.

Problem 5 (28 points)

Here you will explore how to use a *spatial* RNA-seq dataset to perform (1) logistic regression to extract genes which are cell type markers and (2) spatial (auto)correlation analysis to recover spatially-variant gene relationships, which may or may not map to cell type markers. This will combine using gene-count matrices and gene-coordinate matrices, where 2D coordinates are given for the genes in the tissue. See the Problem 5 notebook here.

Your edited version of the notebook *must be submitted* for this problem. Reminder to check that your notebook runs all the way through with the the `Runtime` → `Restart` and `Runtime` → `Run All` commands.