Bi/BE/CS 183 2021-2022
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 **10am Thursday January 20th**.

- 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.

Problem 1 (24 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 this problem you will derive a key property of this estimator, namely that it is the **B**est **L**inear **U**nbiased **E**stimator (BLUE).

(a) (3 points) 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) (3 points) A linear estimator $\hat{\beta}$ is unbiased if $\hat{\beta} = \beta$ whenever $\epsilon = 0$. Show that the least square estimator is unbiased

(c) (6 points) Consider a matrix D such that $DX = 0$, show $\hat{\beta} = (X^{\dagger} + D)\mathbf{y}$ is also an unbiased linear estimator.

(d) (12 points) i. (2 points) Consider the estimation error of a linear estimator of the form $\hat{\beta} = By$, i.e.,

$$|\beta - \hat{\beta}|.$$

Show that the estimation error is equal to $|B\epsilon|$ when $\hat{\beta}$ is unbiased.

ii. (10 points) Clearly, we would like for B in our linear estimator to be "small", so that the error $|B\epsilon|$ can be small. Show that for any unbiased linear estimator $\hat{\beta} = B\mathbf{y}$, we have

$$\sum_{i,j} B_{ij}^2 \geq \sum_{i,j} X_{ij}^{\dagger 2}, \tag{3}$$

thus concluding that the least square estimator is the BLUE. This is known as the Gauss-Markov theorem. (Hint: consider the result of part (c), first show that all unbiased linear estimator can be written as $\hat{\beta} = (X^\dagger + D)\mathbf{y}$, where $DX = 0$)

Problem 2 (24 points)

In the context of scRNA-seq analysis, gene expression matrix X is often rank-deficient, where one of the column 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 $\hat{\beta} = X^{\dagger} y = (X^T X)^{-1} X^T y$.

(a) (12 points) Show that if $X^T X$ is not invertible, then 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 $X^T X$ may arise

(b) (12 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 $X^{\dagger}$,

$$X^+ = \lim_{\delta \to 0} \left(X^T X + \delta^2 I \right)^{-1} X^T. \tag{4}$$

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).

Problem 3 (24 points)

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

(a) (12 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 $\rho_{X_1 X_2 \cdot Y} \neq 0$ 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.

(b) (12 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 4 (28 points)

Here you will explore how to use (1) linear and (2) logistic 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 (3) 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.