**Bi/BE/CS 183 2021-2022**
**Instructor: Lior Pachter**
**TAs: Tara Chari, Zitong (Jerry) Wang, Meichen Fang**

**Problem Set 2 Solution**

**Problem 1**

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

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

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,

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

where $X$ is the observed gene expression matrix and $\beta$ 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 one of the column of is a linear combination of the others. In this case, $X^TX$ has no inverse, and so we must use a different estimator instead of $\hat{\beta} = X^{\dagger}y = (X^TX)^{-1}X^Ty$.

(a) (8 points) Show that if $X^TX$ 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 $X^TX$ may arise

**A:** If $X^TX$ is not invertible, then there exists a non-zero vector a such that $X^TXa = 0$, and thus $a^TX^TXa = 0$.

$$\begin{aligned} 0 &= a^TX^TXa \\ &= (Xa)^TXa \\ &= \sum_i (Xa)_i^2 \end{aligned}$$

Therefore, there exists a non-zero vector a such that $Xa = 0$, which means one column of $X$ is a linear combination of the others.

(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).

**A:** Since $X$ is rank-deficient, there exists a non-zero vector a such that $Xa = 0$.
If there exists an optimal solution $\hat{\beta}$ to the least-square problem, $\hat{\beta} + a$ is also an optimal solution because $||y - X(\hat{\beta} + a)|| = ||y - X\hat{\beta}||$.
Therefore, the least-square problem does not admit a unique solution.

**Problem 2**

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 $\rho_{X_1X_2 \cdot Y} \neq 0$ 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, $\rho_{X_1X_2} = cov(X_1, X_2) = 0$.

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 $cov(X_1, Y) \neq 0$.

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

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

(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.

**A:** To show that $\rho_{XY \cdot Z} = 0$, we just need to show $\rho_{XY} = \rho_{XZ}\rho_{YZ}$, with $\rho_{XY}$ being the regular Pearson correlation. Since X and Z are independent, standard Gaussians, $cov(X, X^2) = E[X^3] - E[X]E[X^2] = 0$ and $cov(X, Z) = 0$.

Then

$$cov(X, Y) = cov(X, X^2 + Z) = cov(X, Z) + cov(X, X^2) = 0$$

.
Therefore,

$$\rho_{XY} = \frac{cov(X,Y)}{\sigma(X)\sigma(Y)} = 0$$
$$\rho_{XZ}\rho_{YZ} = \frac{cov(X,Z)cov(Y,Z)}{\sigma(X)\sigma(Y)\sigma^2(Z)} = 0$$

and

$$\rho_{XY} = \rho_{XZ}\rho_{YZ}$$

**Problem 3**

Here is an example solution notebook.

**Problem 4**

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 solution notebook here.

**Problem 5**

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 solution notebook here.