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

Problem Set 6

Submit your solutions as a single PDF file via Canvas by **10am Friday February 18th**.

- 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 (20 points)

The delta method for finding variance-stabilizing transformations makes use of the following relation,

$$\mathrm{Var}[f(X)] \approx (f'(\mu))^2 \mathrm{Var}[X],$$

where $\mu = E[X]$. We will derive this approximate relation using Taylor expansion.

(a) (10 points) Under certain regularity conditions, a function $f(x)$ can be expressed as an infinite sum consisting of powers of x known as a Taylor series,

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n,$$

where $f^{(n)}(a)$ denotes the n-th derivative of f evaluated at the point a. Write down the first-order Taylor expansion of $f(X)$ around the point $X = \mu = E[X]$.

(b) (10 points) Using your Taylor expansion, show that $\mathrm{Var}[f(X)] \approx (f'(\mu))^2 \mathrm{Var}[X]$, where $\mu = E[X]$.

Problem 2 (40 points)

When performing single-cell RNAseq, there is usually variability in read-depth (total number of reads) between cells, which could be due to both variable gene expression and variable efficiency with which molecules are sampled from the mRNA pool of a cell during the measurement process. One way to adjust for these variabilities (assuming they are not of interest to you) is to divide each gene count by the total count for each cell, which we will refer to as the size factor s, so we will have a different size factor for each cell. In this problem. you will explore an issue associated with this simple adjustment when combined with variance-stabilizing by the delta method, and **practice applying the law of total variance** as discussed in lecture 4.

(a) (10 points) Consider a simple model of gene count represented by a Gamma-Poisson random variable K, equivalently known as a negative-binomial random variable, with parameters k and θ:

$$\begin{aligned} K \mid Q &\sim \mathrm{Poisson}(Q) \\ Q &\sim \mathrm{Gamma}(k, \theta). \end{aligned} \tag{1}$$

The Poisson level of this hierarchical model corresponds to sampling noise and the Gamma level models additional variation between genes. Using the fact that $E[Q] = k\theta$ and $\mathrm{Var}[Q] = k\theta^2$, show that $E[K] = k\theta$, and $\mathrm{Var}[K] = k\theta + k\theta^2$.

(b) (10 points) Now, consider a model with size factors s,

$$\begin{aligned} K' \mid Q, s &\sim \mathrm{Poisson}(sQ) \\ Q &\sim \mathrm{Gamma}(\mu, \phi). \end{aligned} \tag{2}$$

Find an expression for the mean and variance of K' in terms of the parameters $\mu' = s\mu$ and ϕ, show your work.

(c) (10 points) Finally, suppose we were to normalize our gene count by the size factor as mentioned earlier,

$$Y = K'/s$$

Find an expression for the mean and variance of Y in terms of the parameters s, μ, and ϕ, show your work

(d) (10 points) What is the problem with performing variance stabilization on Y using a transformation for Gamma-Poisson (negative-binomial) random variable derived by the Delta method? especially if the size factor s vary a lot between cells. Recall that the mean-variance relationship of a Gamma-Poisson random variable is $\sigma^2 = \mu + \phi\mu^2$.

Problem 3 (40 points)

For this problem you will be exploring various methods for variance stabilization commonly used to transform single-cell datasets. The Problem notebook is here.

Often single-cell count matrices are variance stabilized e.g. using the log1p transformation. This theoretically decouples the variance from the mean expression (per cell) to allow for differential expression testing, particularly as common regression models assume homoscedascity, and comparing cells to determine cell types/populations within the larger group, using clustering algorithms [1].

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.

References

1. Svensson, V. *Variance stabilizing scRNA-seq counts — What do you mean "heterogeneity"?* en. `https://www.nxn.se/valent/2017/10/15/variance-stabilizing-scrna-seq-counts`. Accessed: 2022-2-3. Oct. 2017.