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

Problem Set 6

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 $X = \mu = E[X]$.

A:

$$f(x) \approx \sum_{n=0}^{1} \frac{f^{(n)}(\mu)}{n!}(x-\mu)^n = f(\mu) + f'(\mu)(x-\mu)$$

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

A:

$$\begin{aligned}
\mathrm{Var}[f(X)] &\approx \mathrm{Var}[f(\mu) + f'(\mu)(x-\mu)] \\
&= \mathrm{Var}[f'(\mu)(x-\mu)] \\
&= (f'(\mu))^2 \mathrm{Var}[X]
\end{aligned}$$

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:

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

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 $\text{Var}[Q] = k\theta^2$, show that $E[K] = k\theta$, and $\text{Var}[K] = k\theta + k\theta^2$.

A: For the mean of K, we have

$$E[K] = E_Q[E[K|Q]] = E[Q] = k\theta$$

.

For the variance of K, because we know that the mean and variance of Poisson distribution are the same, according to the law of total variance, we have

$$\begin{aligned} Var[K] &= E[Var(K|Q)] + Var[E(K|Q)] \\ &= E[Q] + Var[Q] \\ &= k\theta + k\theta^2 \end{aligned}$$

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

$$\begin{aligned} K' \mid Q, s &\sim \text{Poisson}(sQ) \\ Q &\sim \text{Gamma}(\frac{1}{\phi}, \mu\phi). \end{aligned} \tag{2}$$

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

A: For the mean of K, we have

$$E[K'] = E_Q[E[K'|Q]] = E[sQ] = s\mu = \mu'$$

.

For the variance of K,, similar to part a, we have

$$\begin{aligned} Var[K'] &= E[Var(K'|Q)] + Var[E(K'|Q)] \\ &= E[sQ] + Var[sQ] \\ &= sE[Q] + s^2 Var[Q] \\ &= s\mu + s^2\phi\mu^2 \\ &= \mu' + \phi\mu'^2 \end{aligned}$$

(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

A:

$$E[Y] = E[K'/s] = \frac{1}{s}E[K'] = \mu$$

and

$$Var[Y] = Var[K'/s] = \frac{1}{s^2}Var[K'] = \frac{\mu}{s} + \phi\mu^2$$

(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? Recall that the mean-variance relationship of a Gamma-Poisson random variable is $\sigma^2 = \mu + \phi\mu^2$.

A: The true mean-variance relationship for the normalized counts Y is $\sigma^2 = \frac{\mu}{s} + \phi\mu^2$, which does not match the mean-variance relationship for a Gamma-Poisson random variable ($\sigma^2 = \mu + \phi\mu^2$).

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].

The solution notebook is here.

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.