Homework 5: Gene expression distributions

Bi/BE/CS183 - Winter 2020
Profs. Matt Thomson and Lior Pachter
Due on Moodle by Thursday February 20 at 12:00pm

Problem 1: Mixture model for a two-component cell population (25 points)

Consider a cell population composed of two different cell-types. In one cell-type, a gene, **Bimdl**, is '**on**' and present at an average copy number of λ_{on} and in the second cell-type Bimdl is present at an average of λ_{off}, molecules. In both cell-types, Bimdl counts per cell, b, are a Poisson distributed random variable, so that $b \sim \text{Poi}(\lambda_{on})$ in Bimdl **on** cells.

(A) Construct a mathematical expression for the gene expression distribution, $P(b)$ that describes a mixed cell population where w_on is the fraction of **on** cells and $(1 - w_{on})$ is the fraction of **off** cells.

(B) Now imagine, that I have a mixed cell population containing Bimdl-on and Bimdl-off cells. I know that $\lambda_{\text{on}} = 20$ in Bimdl-on cells and $\lambda_{off} = 2$ in Bimdl-off. I aim to infer the composition of the cell population(the relative proportion of **on** (w_{on}) and **off** $(1 - w_{on})$ cells) from single cell gene expression measurements.

Construct the log-likelihood function $L(w_{on})$ for the cell population using your expression in (A).

(C) Write a simple program that **generates** synthetic data from a cell population with $w_{on} = .3$ and $w_{off} = .7$ cells (and $\lambda_{on} = 30$, $\lambda_{off} = 2$).

Now, perform MLE using this synthetic cell population as a generative model (where we can measure b without sampling noise).

Draw 100 samples from the synthetic data and construct and plot the log likelihood function $L(w_1)$. Can you find w_1 through maximization of $L(w_1|\text{observations})$.

(D) Now consider the problem of performing this inference in the presence of a Poisson measurement process. Extend the program you wrote in (C) to include a Poisson measurement step where $P(\hat{b}|b) = Poi(\hat{b}, p \times b)$ where p is the transcript capture probability. For simplicity, plot the histograms of your synthetic data for $p = .1$,$p = .5$ and $p = .8$. What do you observe? How does measurement noise impact the qualitative shape of the distribution?

Problem 2: Cell sampling and parameter estimation error (25 points)

In single cell data analysis, a common problem is understanding how many cell measurements are needed to estimate a quantity up to a desired accuracy. In this problem, you will numerically and analytically solve such a problem using the framework of maximum likelihood estimation.

Consider the Poisson model of gene expression, and consider a gene present at $\lambda_0 = 10$ copies per cell on average. How many cells must be sampled to estimate λ_0 with an average L2 error, $||\hat{\lambda} - \lambda||_2$ of 10%?

(A) Using your computer, generate synthetic data by sampling from a Poisson distribution and perform MLE of λ on the synthetic data. Plot $(\lambda - \lambda_0)^2$, as a function of n on a log-log plot (explore at least $n = 1$ to $n = 50$).

(B) Now, we want to understand the slope and intercept of your error plot. Recall from lecture that:

$$\langle(\lambda - \lambda_0)^2\rangle \; > \frac{1}{n\;I},$$

where I is the Fisher information, λ_0 is the true value $\lambda_0 = 10$, λ is the estimate, and n is the number of measurements. For a Poisson distribution, $I = \frac{1}{\lambda}$.

Plot the the theoretical error bound on the same plot as your simulated data. Can you explain the slope and intercept?

(C) Consider the error bound provided by the Fisher information for the Poisson distribution:

$$\epsilon = |\hat{\lambda} - \lambda| > \sqrt{\frac{\lambda}{n}}$$

Now, derive a result for the relative error, $\frac{\epsilon}{\lambda}$. How many samples are required to estimate λ with a relative error bound of ϵ_0 (for example $\epsilon_0 = .1$)?

Problem 3: Quantifying PBMC mRNA variation (25 points)

In this problem, you will analyze the coefficient of variation in single cell mRNA-seq data collected from human immune cells.

Down-load the filtered gene count matrix `Feature / cell matrix (filtered)` from 10x genomics dataset **1k PBMCs from a Healthy Donor (v3 chemistry)** available on the following link:

`https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.0/pbmc_1k_v3`

The link for the filtered gene expression matrix is

`http://cf.10xgenomics.com/samples/cell-exp/3.0.0/pbmc_1k_v3/pbmc_1k_v3_filtered_feature_bc_matrix.tar.gz`

The filename should be `pbmc_1k_v3_filtered_feature_bc_matrix.tar.gz`

(A) **First, plot a histogram of the log of total number of counts per cell across the data set**. Calculate the min, max, mean and variance of this distribution. Comment on each number.

(B) Now, normalize the data by dividing each column by the total number of gene counts in that column and then multiply by 10^5.

For each gene in the data set, calculate the mean, standard deviation, and coefficient of variation using the column normalized data. Generate a scatter plot where each point is $(\log(\mu_{\text{gene}}), \log(\mathrm{C_v}))$, for a single gene. You will need to exclude genes (rows) with zero total counts.

(C) Perform a least squares linear fit to the points, $(\log(\mu_{\text{gene}}), \log(\mathrm{C_v}))$ in the scatter plot.

Specifically, consider the linear equation:

$$a \log(\mu_{\text{gene}}) + b = \log(\mathrm{C_v})$$

for all data points. Solve this linear equation for a and b and report the slope and intercept of the best fit line and comment relative to expectations for a Poisson model.

Problem 4: Analyzing the impact of immuno-suppressing drugs on the human immune cells

The single cell profiling center at Caltech has collected a large data set studying the impact of various immunomodulatory drugs on the human immune system. The data set contains primary human immune cells under resting, inflammatory, and drug treated conditions. In this problem, you will use a non-parametric approach to discover genes impacted by a common immunomodulatory, cyclosporine, on human immune cells.

We will specifically write a simple differential expression program that uses the L1 metric to quantify differences between gene expression distributions.

We will do so by comparing a control data set to a data set where the immune cells have been treated with a drug called cyclosporine.

Your program should first download, column normalize, and log transform each data set using the following links:

`https://www.dropbox.com/s/lub5uxih985ytb6/Cyclosporin%20A_CD3.mtx?dl=0`

`https://www.dropbox.com/s/493vbdbs9ggbhd5/CONTROL_CD3.1.mtx?dl=0`

`https://www.dropbox.com/s/hp0a5ikd1y0mv1v/gene_drug_screen.csv?dl=0`

Second, the program should construct a gene expression histogram for each gene in the data set. Specifically, find the minimum and maximum value for the gene across both data sets. Then, bin this range into 20 evenly space bins. Now construct a normalized histogram that contains the fraction gene counts for the control and drug treated sample, and calculate:

$$Z(g) = \sum_{i=1}^{20} |P(g_i; \text{control}) - P(g_i; \text{drug})|$$

where $|*|$ indicates absolute value.

Find $Z(g)$ for each gene in the data set, and plot a histogram of these values.

Select the genes that have the top five largest values of Z and plot their gene expression histogram in the two conditions. Use the internet to look up the function of each gene and comment on what you find.

(C) Discuss how you would extend your program to calculate the statistical significance of these scores. What additional data would you like to collect in order to asses the likelihood that the gene expression distribution for a given gene is the same in the drug treated and control sample?

Useful functions and resources:

- numpy.random.poisson

- scipy.io.mmread

Extra non-required problems: Sampling noise and measures of variation

In this problem, you will analyze the impact of Poisson measurement noise on the Fano factor and coefficient of variation for Poisson gene expression models. Consider a gene whose mRNA is Poisson distributed in a cell population with mean expression N_0. You measure the number of mRNA molecules in each cell using a Poisson measurement process with sampling 'sensitivity' f. Calculate the coefficient of variation $(\frac{\sigma}{\mu})$ as a function of the measurement sensitivity, f.

To do so, model the measurement process as a Poisson process:

$$P(\hat{m}) = \sum_m P(\hat{m}|m) \; \frac{\exp(-N_0)N_0^m}{m!}$$

$$P(\hat{m}|m) = \frac{\exp(-fm) \; (fm)^{\hat{m}}}{\hat{m}!}$$

(A) Show that $E[\hat{m}] = fN_0$.

(B) Calculate $E[(\hat{m} - E[\hat{m}])^2]$, the variance of $\hat{m}$.

(C) Using A and B, calculate the coefficient of variation for $\hat{m}$.

For (B), you can use the result that for a Poisson distribution with $E[x] = \lambda$, $E[x^2] = \lambda + \lambda^2$.