Problem Set 1: Inference and Single cell analysis

February 20, 2018

1 Data Orientation: option 1 of 2

(1) Please see Figure S2 in the following paper from Elowitz and Cai on gene counting using single molecule approaches.

Dynamic Heterogeneity and DNA Methylation in Embryonic Stem Cells Singer et al. Mol Cell 2014. (http://www.cell.com/cms/attachment2016385158/2037094618/mmc1.pdf)

Looking at Figure S2, estimate the position of the peak (mode) in the gene count histogram for **(1.1)** Oct4 **(1.2)** Rest **(1.3)** Stat3 and **(1.4)** Fgfr2 (estimate half-max location). This exercise is meant to give you an intuition for the magnitude of typical mRNA count numbers in cells.

(1.5) How many counts would you expect to detect for each gene with the $p = .1$ sampling rate of Drop-seq? To say this in another way, given a gene with mode g, you would expect this mode to shift to $g\ p$ with Poisson sampling in single cell mRNA-seq. For each gene above, what is $p\ g$ for the value you stated in 1.1 thrugh 1.4?

2 Inference problems with Poisson distributed gene expression data

In class we showed that a very simple model of stochastic transcription generates Poisson distributed gene expression values ($g \sim \text{Poi}(\lambda)$) with a rate λ.

(2.1) Using a function such as poissrand in Matlab learn how to generate samples from a Poisson distribution with a fixed lambda. Generate 500 samples for $\lambda = 1, 5, 50$ and plot a histogram of each resulting simulated data set.

(2.2) Take the data set you generated for $\lambda = 1, 5, 50$, and perform parameter estimation on this data. Plot your estimate as a function of sample number.

(2.3) Plot your estimation error as $\sqrt{(\lambda - \widehat{\lambda})^2}$ for samples ranging from 1 to 500. How many samples do you need to achieve error of less than 10% of the true value of λ in each case?

3 Inference problems Poisson mixtures: option 2 of 2

(3.1) First, imagine I give you a mixture of cells, and I tell you that the population contains 70% brain cells and 30% muscle cells. I also tell you that brain cells and muscle cells both have Poisson distributed expression of a gene called actin, but that gene has different mean count abundance in each cell-type, so that $\lambda_m \neq \lambda_b$.

Go to the class Github site and down-load the file, brainmuscle1.csv. This file contains gene count data samples from a synthetic cell population. Design a procedure to infer λ_m, λ_b from the data. What are your resulting values for λ_b, λ_m?

(3.2) Now, I give you a cell population, and I tell you that the cell population is mixture of two different types of cells. The actin gene is present at an average copy number of $\lambda_m = 100$ in muscle cells and an average copy number of $\lambda = 5$ in neurons.

Apply ML estimation to determine the fraction of neurons and muscle cells from the (data brain-muscle2.csv). You might need to enforce constraints on the weighting parameter, w.

Note you can use a procedure like fminsearch, FindMinimum (mathematica) or your own method (gradient descent or even direct griding are easy to implement in this space) to perform.

(3.3) Thought problem: How would you structure the problem to infer $w, \lambda_m,$ and λ_b. What constraints do you need to apply to w during the minimization procedure?

4 Data analysis of public cancer data: Required

We are going to construct a naive cancer detector by using the tools we developed this week in class. Please go to the class GitHub and down-load the four data sets.

(4.1) Using the healthy1 data, first plot μ_i (the mean) versus σ_i^2 (variance) for all genes in the data set on a log-log plot. Fit the data to a linear model and report the slope and y-intercept.

(4.2) Calculate and plot the coefficient of variation for each gene in the data set.

The coefficient of variation is:

$$\mathrm{C_v} = \frac{\sigma}{\mu}$$

The coefficient of variation and the fano factor are two useful measures of dispersion in the data. For a Poisson distribution $\mu = \sigma^2$ (the mean and variance are equal), so the fano factor is 1. For a Poisson process, the $\mathrm{C_v} = \frac{1}{\sqrt{\lambda}}$. For this reason, it is common in the literature to study noise in gene expression via C_v^2.

In the case of the healthy 1 sample, please plot μ vs C_v^2 and fit the resulting points with a linear model or estimate the slope of the line through visual inspection (plot and fit the log).

Calculate and plot the fano factor. Comment on the result. Think about how data normalization impacts mean and variance.

(4.3) For each gene, plot the number of zero count cells in a histogram.

Building the model

You are free, in the following work, to use any model of gene expression that you like. You will be graded on your discussion and producing the required plots. One possible model for the gene expression distributions is a gamma distribution with a modification to account for the zeros. I will develop that model in the following questions.
If you choose a modified model, then the following directions will not be useful for you. But please produce the indicated plots.

(4.4) Write a function that performs the following operations for a gene. (i) find the zero count measurements (ii) take the non-zero values and use ML estimation to find gamma distribution parameters α and β that are ML estimates. You can perform this operation using direct maximization of gamma OR you can use the following functions that are included in matlab and R. Matlab: gamfit. R: fitdistr.

Once you have written function, you can apply it to all the genes in the data set.

Plot a scatter plot of your α and β values for each gene in the data set.

(4.6) Develop a procedure to plot the underlying data on the same plot as your MLE from the gamma distribution. To do this, you will need to histogram the gene count values (focus on the non-zero values), and then you will need to plot the corresponding value of the gamma distribution. Hint: the histogram pools values across an entire bin, so when you plot gamma, you, in fact, plot $\text{Gamma}(g, \alpha, \beta) \times \delta$.

(4.7) Using your gamma ML estimation function, now perform ML estimation across all the genes in the aml 1 sample. Plot the resulting values of alpha and beta on a scatter plot with the healthy 1 α and β values. Make sure to set your axis to something like [0,10] and [0,1000], so that you can see structure.

(4.8) Calculate confidence intervals for the MLE estimates for healthy1 and aml1 and discuss how this procedure works in words. You can use values returned by the R, python or matlab package. Isolate and 3 distributions in the data set where the estimated values of α and β in aml1 lie significantly outside the confidence intervals determined in healthy1. Plot these distributions and find the name of the associated genes. Look up the function of these genes, and write them down.

(4.9) For each gene you now have a probabilistic model:

$$P(g_i|\text{Healthy 1}) = w_i\ \mathbb{1}_0(g_i) + (1 - w_i)\ \text{Gamma}(g_i, \alpha_i, \beta_i)$$
$$P(g_i|\text{AML 1}) = w_i\ \mathbb{1}_0(g_i) + (1 - w_i)\ \text{Gamma}(g_i, \alpha_i, \beta_i)$$

where w_i is the number of zero-count cells for gene i and $\mathbb{1}_0$ is an indicator function that is 1 when $g = 0$ and 0 otherwise. In words, this function gives the probability of observing a value g_i for gene i given the values of α and β you calculated above.

Given a set of measurements from a new sample, say, healthy 2, we have an expression for the gene-wise log-likelihood ratio the hypothesis health vs hypothesis aml.

$$A(g_i) = \log(\frac{P(g_i|\text{Healthy 1})}{P(g_i|\text{AML 1})})$$

Construct a function that can calculate $A(g_i)$ on single cell samples from the healthy 2 and aml 2 data sets.

For a single cell, you can calculate:

$$\sum_{i=1}^{n} A(g_i)$$

where i runs over genes to sum evidence across genes.

(4.10) Apply your function A_i to single cells from the healthy 2 sample and the aml 2 sample. Please plot the cumulative evidence for the cancer or aml hypothesis. what do you observe? **Note:** I randomized the ordering of the cells in the data set before running the calculation.

How many cells and genes are required to achieve an evidence value of 10? Comment on the scaling of evidence with cell number.
Optional: Make a plot of the cumulative evidence as a function of genes and cells. Can you determine a lower bound for classification as a function of gene and cell number.

Optional: Calculate false positive and false negative rates from the model and determine optimal values for the decision threshold T.