Homework 8

This homework will be due next Thursday (Dec 6th).

1. Suppose you perform single cell RNA sequencing on 2 cells each from two different stages of a cancer. You want to compare gene expression between these two stages of cancer in order to identify genes that contribute to varying disease severity. You determine the expression read counts of 2000 genes per cell (See table below for example). Some of these genes have extreme expression differences between cells regardless of whether the cells are from the same or different stages of the cancer. For example, they are nearly absent or highly expressed. Before making inferences about the biological differences between the two stages of the cancer based on the varying gene expression of the cells, you would like to first assess some pairwise relationships to build a framework for general patterns. You decide to compute pairwise measures of relationships between (1) each cell; (2) between cell types; and (3) between each gene.

A) Suggest a measure of relatedness (or more than one)  
B) Explain your goals—that is, what kind of relatedness that you are trying to characterize with your measure(s)  
C) Explain your choice of measure(s) in terms of your goals.  
D) Suggest possible problems with your chosen measure (e.g., when some value is zero, the measure goes to infinity).

Example read counts:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Gene | Cell Type 1 | | Cell Type 2 | |
| Stage 1A | Stage1B | Stage 4A | Stage4B |
| Upp2 | 1 | 20 | 0 | 0 |
| Dusp15 | 14 | 33 | 30 | 71 |
| Dpy19l1 | 111 | 655 | 343 | 108 |
| Mtx2 | 1770 | 995 | 1496 | 475 |
| Alkbh1 | 9 | 581 | 1 | 29 |
| Man1c1 | 3 | 699 | 848 | 112 |
| AU023762 | 2 | 0 | 0 | 56 |
| Plscr3 | 0 | 6 | 0 | 0 |
| Tas1r3 | 3 | 0 | 0 | 0 |

2. Suppose we now have the normal coordinate system and there is a list of vectors on this coordinate, the following is their coordinate values:

(1, 0.5) (1.5, 0.5) (1.3, 1) (2, 1) (2.5, 1) (2.4, 1.5) (2, 2) (3, 2.5) (4, 2.5) (4, 3) (4.5, 3) (5, 3)

Please find the PC1 and PC2 for these vectors based on the PCA algorithm. You should write down the formula of the original coordinates based on the vector PC1 and vector PC2, and also write down the formula of the vector PC1 and vector PC2 by using the original coordinates. The coefficients in the 1st formula is called scores (or PCA coordinates) and those in the second are called loadings.

Also, scatter plot the vectors both in the original coordinates and coordinates of PC1 and PC2. Please paste your results in the word file and also submit your code, your results should include one PC1/2 values table and two scatter figures.

The PC1/2 values table should like this:

|  |  |  |
| --- | --- | --- |
|  | Principal component 1 | Principal component 2 |
| 0 | … | … |
| 1 | … | … |
| … | … | … |

Hints: You may use any programming languages and any available packages. If you use Python, you can use pandas package to manipulate the data and matplotlib.pyplot to plot the figure, you could also use PCA imported from sklearn.decomposition package. The csv file of the data is provided for your convenience, you can use “pd.read\_csv()” to import the data.

The following links should be a very good resource for these packages:

<https://pandas.pydata.org/pandas-docs/stable/10min.html>

<https://matplotlib.org/index.html>

<https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html>