Problem Set 2: Projections and Low Rank Representations

February 28, 2018

Problem set is due March 8th.

1 Linear Algebra Warm Up

(1) Consider, $\boldsymbol{D}$, an $m \times n$ gene expression matrix, where entry D_{ij} is the number of counts for gene i in cell j. m is the number of genes, and n is the number of cells.

This data matrix can act on a column vector, $\boldsymbol{x}$, where $\boldsymbol{x}$ is an $n \times 1$ vector, $\boldsymbol{x} = (x_1, ..., x_n)$, and $D\ x = y$.

- (1.1) As defined, y is also a vector, how many entries does it have?
- (1.2) Take $x_j = \frac{1}{n}\ \ (\forall j)$, so that $\boldsymbol{x} = (\frac{1}{n}, \frac{1}{n} \cdots \frac{1}{n})$, write down an expression for y_i, an element of $\boldsymbol{y}$, and interpret this expression in terms of the rows of the data matrix, $\boldsymbol{D}$?
- (1.3) Now, lets have D^T act on a new vector, so that $\boldsymbol{z} = (z_1 \cdots z_m)$ and $\boldsymbol{z}$ is $m \times 1$. Can you design a z such that $D^T\ z = y$ and y_j is the total number of transcripts counted in cell j?

2 Analysis of synthetic gene module data

The goal of this problem is to have you think in greater depth about the meaning of PCA and non-negative matrix factorization on transcriptional profiling data. There is a file of synthetic data on the github called ‘pset2syndata.csv’ under week 2. The file has data generated for 11 ’genes’ across 500 cells.

(2.1) Down-load and plot data as a heat-map. You can plot using imagesc or a matrix plot comman in python, matlab, R.

(2.2) Perform principle component analysis on this data to return a set of eigenvectors and eigenvalues of the covariance matrix. Please proceed by writing a short procedure for doing PCA by (a) mean centering the data (b) calculating the gene by gene covariance matrix and then (c) finding the eigenvalues and eigenvectors of this matrix.

(2.3) Plot a 3D projection of the data using the first three principle components, also plot a curve showing the decay of the eigenvalues.

(2.4) Perform NNMF on the data for two inner dimension values. Select the inner dimensions based upon your eigen-value decay plots in the previous problem. Briefly describe how you choose the dimensions. Plot the W and H matrices of the data set. Comment in words on the difference between the PCA eigenvectors and the NNMF vectors relative to the underlying gene modules.

Note: Packages for performing NNMF on the data are available in matlab, R, and python. Please see the github for direction on this

3 Analysis of healthy immune data

The goal of this problem is to have you gain direct experience with projecting the human immune data with PCA and NNMF. I think you will also see the geometric beauty in this data and be inspired to think more about what it means. This problem set is done on the healthy1.csv data you previously worked with.

(3.1) Write a short procedure to perform perform PCA on the healthy 1 sample and plot a 2 and 3D projection of the data. You will want to mean center the data.

(3.2) Study the genes in the first and second eigenvectors, and consider both the positive and negative entries. How would you select genes in these principle components to focus on? What is the mean and std of the coefficients in this vector? Pick the top positive and negative entries in components 1-4 and comment on their names–just take the top 2 genes in the + and - for these components. Can you say anything about the point clouds in the data based upon these genes?

(3.3) Perform NNMF on the data using functions in R, matlab, or scikit learn.

Note prior to the NNMF, normalize genes by their standard deviation. You can try the procedure both ways, but make sure to factor the normalized data at the minimum.

Select a dimension for the decomposition based upon inspection of the PCA eigenspectrum. It is interesting to go above and below this dimension and study transitions in the factorization. For this problem, you are required to pick a single inner dimension, k, and plot the W and H matrix for this k. I used $k = 10$, and had very interesting results.

(3.4) Use a hierarchical clustering to cluster W and H using a procedure of your choosing and plot the clustered version of W and H. What do you notice? Select 4 genes from a W component of interest and comment on their function.

Optional but very interesting: (3.4). It is interesting to make a 3D scatter plot of the genes projected into the first PCs and then to color this plot by the weightings of the cells in the NNMF H matrix.

(3.5) If you have time and would like to go further, you can imagine extending this analysis to incorporate the AML sample.