369 FINAL PROJECT

Project description: In this project, you will analyze a data set involving mice infected with the bacterium F. Tularensis, using basic tools from linear algebra, including projections, eigenvectors, and the singular value decomposition. The project requires the use of Matlab[1]. Note that Matlab is freely available to you by our university license. It is also accessible for free on the computers in our Weber 205 lab. See the instructions on our course Canvas page.

Project rules: This project contains 5 problems, worth 150 points total. You can discuss these problems with your classmates, but you should write up solutions on your own. Please submit your completed project as a PDF file to Canvas.

Your completed project should contain:

- **All the code you used to answer the questions below;**
- **All the output of the codes you used to answer the questions below, including all figures;**
- **Answers to ALL of the prompts and questions below;**
- **Labels for EACH PAGE of your submitted project, indicating which problem the page is addressing.**

Feel free to make use of the Canvas discussion page if you get stuck.

[1] You could also try using free Matlab clones such as Octave and SciLab

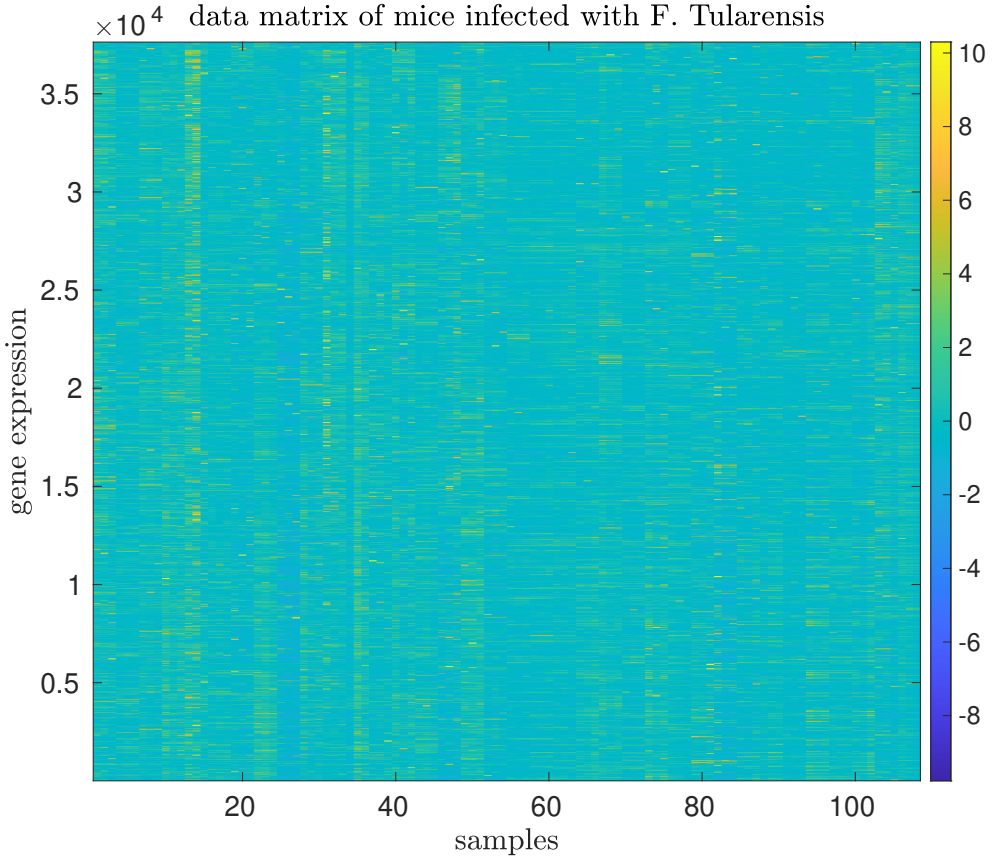

FIGURE 1. Data matrix X. Each column of X represents gene expressions coming from a mouse sample. Samples are taken from a mouse's lung or spleen. Mice are from three groups: an uninfected control group, a group infected with the virulent strain Schu4, and another group infected with the less virulent strain LVS.

1. (20 points) Download the file "Kringynormalized.mat" from Canvas, and place it in your current Matlab directory. Then, load the data and store it as a matrix, X, using the following commands:

```
load Kingrynormalized.mat    %load the data matrix
X = Kingrynorm;              %store the data matrix as X
```

The matrix X has $n = 37,632$ rows and $m = 108$ columns. Check this using the following command:

```
[n,m] = size(X)
```

Each column of X comes from a sample taken from a mouse. The $37,632$ numerical values in each column represent gene expressions in the associated sample. The samples may be from an (uninfected) control group, or from a group of mice infected with either of two F. Tularensis strains, called Schu4 and LVS. As the strains primarily affect the lung and spleen, the gene expressions are taken from either of these organs.

Columns 1-6 of X are gene expressions sampled from the lungs of 6 uninfected mice, while columns 55-60 of X are from the spleens of 6 uninfected mice. The rest of the columns are gene expressions in infected samples. Columns 7-30 of X are from the lungs of 24 mice infected with the Schu4 strain; columns 31-54 of X are from the lungs of 24 mice infected with the LVS strain; columns 61-84 of X are from the spleens of 24 mice infected with the Schu4 strain; and columns 85-108 of X are from the spleens of 24 mice infected with the LVS strain.

Abstractly, X represents a cloud of 108 data points, with each data point belonging to $\mathbb{R}^{37,632}$.

In this project, we will need to be able to compare large vectors with each other. One way to do this is via Matlab's "norm" command, which computes the length of vectors of any size. For instance, try entering the following command into Matlab:

```
norm(mean(X,2))
```

What does the output here say about the mean of the data?

2. (20 points) Compute the "thin" SVD of X, and plot the squares of the singular values:

```
[U S V] = svd(X,0);     %X = USV^T is the thin SVD of X
sigmas = diag(S);       %sigmas are the singular values of X
figure                  %open new figure
plot(sigmas.^2)         %plot the squares of the singular values
```

Check that the squares of the singular values are the eigenvalues of $X^T X$:

```
lams = eig(X'*X);              %lams are the eigenvalues of X^T*X
lams = sort(lams,'descend');   %sort the eigenvalues largest to smallest
norm(sigmas.^2-lams)           %compare the eigenvalues with the singular values
```

Check that the first column of U is the eigenvector of XX^T associated to the largest eigenvalue:

```
u1 = U(:,1);                %pull first column of U
lam1 = lams(1);             %pull first eigenvalue of XX^T
norm(X*(X'*u1)-lam1*u1)     %verify that u1 is an eigenvector with eigenvalue lam1
```

Repeat this for for a few other columns of U, to check that they are also eigenvectors of XX^T.

3. (20 points) The following code generates a "PCA plot" of the data in X.

```
U_3d = U(:,1:3);     %U_3d is formed from the first three columns of U
X_3d = U_3d'*X;      %X_3d is a 3d representation of X

figure                                                          %create new figure
plot3(X_3d(1,13),X_3d(2,13),X_3d(3,13),'sy','linewidth',10)    %highlight one sample
hold on
plot3(X_3d(1,1:6),X_3d(2,1:6),X_3d(3,1:6),'b+')                %plot control lung
plot3(X_3d(1,55:60),X_3d(2,55:60),X_3d(3,55:60),'bd')          %plot control spleen
plot3(X_3d(1,7:30),X_3d(2,7:30),X_3d(3,7:30),'ro')             %plot Schu4 lung
plot3(X_3d(1,31:54),X_3d(2,31:54),X_3d(3,31:54),'gx')          %plot LVS lung
plot3(X_3d(1,61:84),X_3d(2,61:84),X_3d(3,61:84),'m^')          %plot Schu4 pleen
plot3(X_3d(1,85:108),X_3d(2,85:108),X_3d(3,85:108),'kv')       %plot LVS spleen
legend('highlighted sample','control lung','control spleen',...
    'Schu4 lung','LVS lung','Schu4 spleen','LVS spleen')       %create legend
```

Include this plot with your completed project. Comment on any features you see in the figure.

The eigenvalues of $\frac{1}{n}X^TX$ equal the (nonzero) variances of the data cloud in the directions of the eigenvectors of XX^T. This 3-dimensional PCA plot uses directions associated to the eigenvectors of XX^T with the largest 3 eigenvalues. What fraction of the variance of the data is captured by the 3-dimensional representation of this figure?

4. (35 points) Below, we'll explore eigenvector coordinates for the data. Let $v_1, \ldots, v_n$ be orthogonal, unit length eigenvectors[2] of XX^T associated to the eigenvalues $\lambda_1 \geqslant \ldots \geqslant \lambda_n$.

In parts (i)-(ii) below, x is any vector in $\mathbb{R}^n$. In part (iii), x is a specific sample from X.

(ii) Give a formula for the projection of x onto v_i, in terms of the scalar $x \cdot v_i$ and the vector v_i.

(10 points) $\text{proj}_{v_i} x =$

(i) Give a formula for x in eigenvector coordinates, in terms of the scalars $x \cdot v_1, \ldots, x \cdot v_n$.

(10 points) If $\beta = \{v_1, \ldots, v_n\}$, then $[x]_\beta =$

(iii) One of the data points from your figure in Problem 3 is highlighted in yellow. This point is a 3-dimensional representation of the 13th column of X (which is one of the Schu4 lung samples).

What are the coordinates of the highlighed point in the figure? Express the 3 coordinates of this point, in terms of v_1, v_2, v_3 and x, where x is the 13th column of X.

(10 points) coordinates of highlighted point =

(iv) (5 points) Is it possible to reconstruct the original data matrix X, *using only the points plotted in the figure from Problem 3?* Discuss.

[2]These unit eigenvectors are not completely unique, since they could point in either direction along the axes they define. Here, assume that the signs of these eigenvectors match the signs of the corresponding eigenvectors of the matrix U from Problem 2.

5. (25 points) Store the columns of X associated to the control spleen and Schu4 spleen samples:

```
control_spleen = X(:,55:60);     %control spleen samples
Schu4_spleen = X(:,61:84);       %Schu4 spleen samples
```

Use the "orth" command to get an orthonormal basis for the span of the control spleen samples:

```
U = orth(control_spleen);        %columns of U are orthogonal unit vectors forming
                                 %a basis for the span of the control spleen samples
```

Use this basis to project the first Schu4 spleen sample onto the span of the control spleen samples. Compare the Schu4 spleen sample to its projection by taking the distance between the two:

```
    v = Schu4_spleen(:,1);       %v = first Schu4 spleen sample
    proj_Uv = U*(U'*v);          %proj_Uv = projection of v onto span of cols of U
    norm(v - proj_Uv)            %display the distance between v and its projection
```

This distance is called a *novelty*. Find novelties for all 24 of the Schu4 spleen samples with respect to the span of the control spleen samples:

```
for i=1:24                       %loop over all 24 Schu4 spleen samples
    v = Schu4_spleen(:,i);       %v = ith Schu4 spleen sample
    proj_Uv = U*(U'*v);          %proj_Uv = projection of v onto span of cols of U
    norm(v - proj_Uv)            %display novelty of v with respect to cols of U
end                              %end loop
```

(i) Find the average of these 24 novelties:

(10 points) average novelty = _____

(ii) (10 points) To interpret the novelties in (i), some frame of reference is needed. To this end, compute the novelties of each control spleen sample with respect to the span of the *other* 5 control spleen samples. Then take the average of these 6 novelties.

(10 points) average novelty = _____

(iii) (5 points) Compare the averages from part (i) and (ii). Which average is larger? Is this what you would expect? What does a larger novelty mean? Discuss.

6. (30 points) Let U be a matrix whose columns $u_1, \ldots, u_k$ are orthogonal unit vectors.

(i) Give a formula for the projection of a vector x onto the span, S, of the columns of U.
Your formula should be written in terms of the scalars $x \cdot u_1, \ldots, x \cdot u_k$ and the vectors $u_1, \ldots, u_k$.

(10 points) $\text{proj}_S x =$

(ii) Give a formula for $U^T x$ in terms of the scalars $x \cdot u_1, \ldots, x \cdot u_k$.

(10 points) $U^T x =$

(iii) (10 points) Using (i)-(ii) above, explain why the *matrix* for projection onto S is UU^T.

7. (Extra credit: 20 points) Use Fisher discriminant analysis (FDA) on the infected spleen and control spleen samples. Take the infected samples as class C^+ and the control samples as class C^-. Plot the projected samples in 1 dimension, labeling each projected sample as infected or uninfected. Comment on how well FDA separates the two classes.