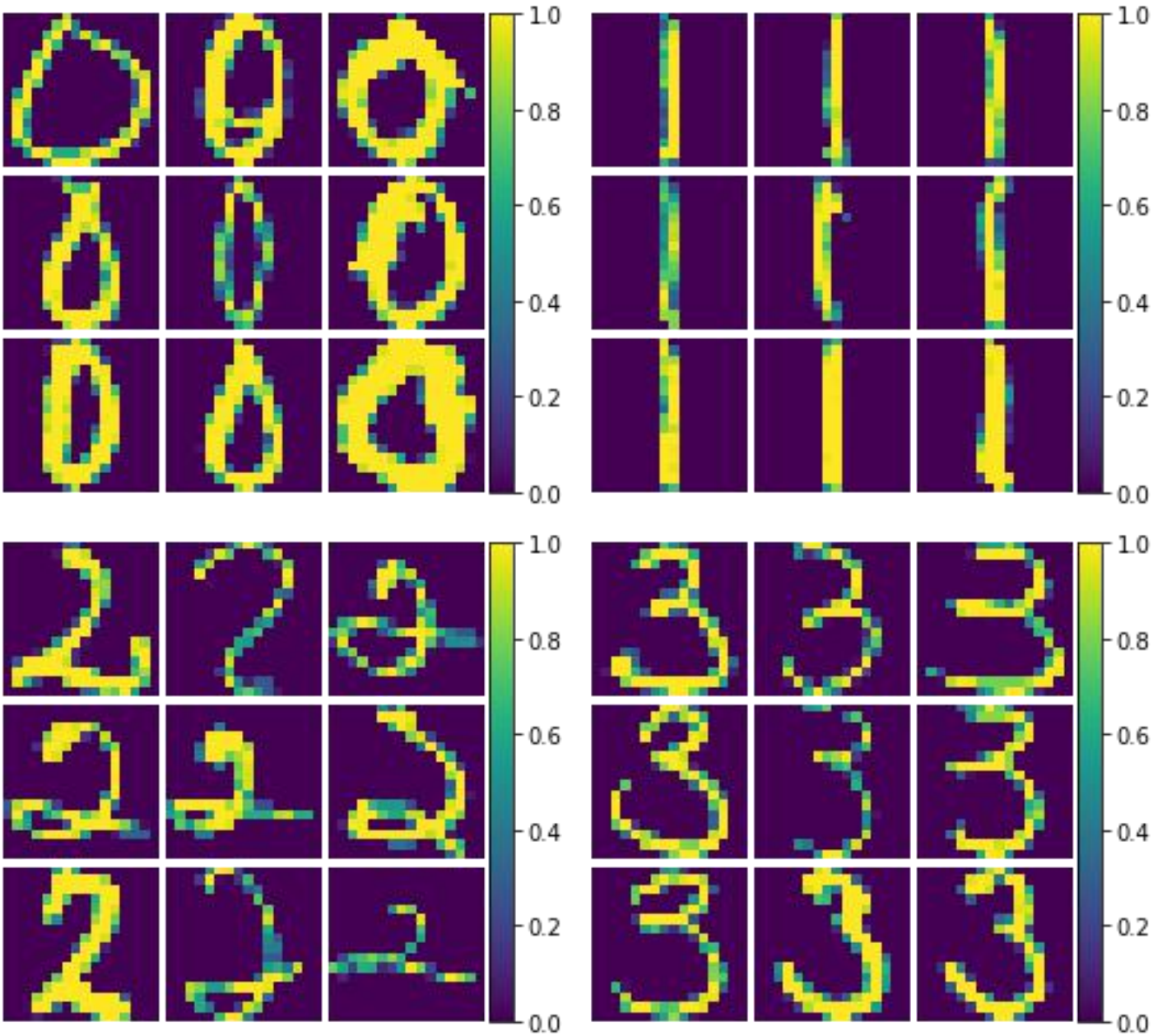
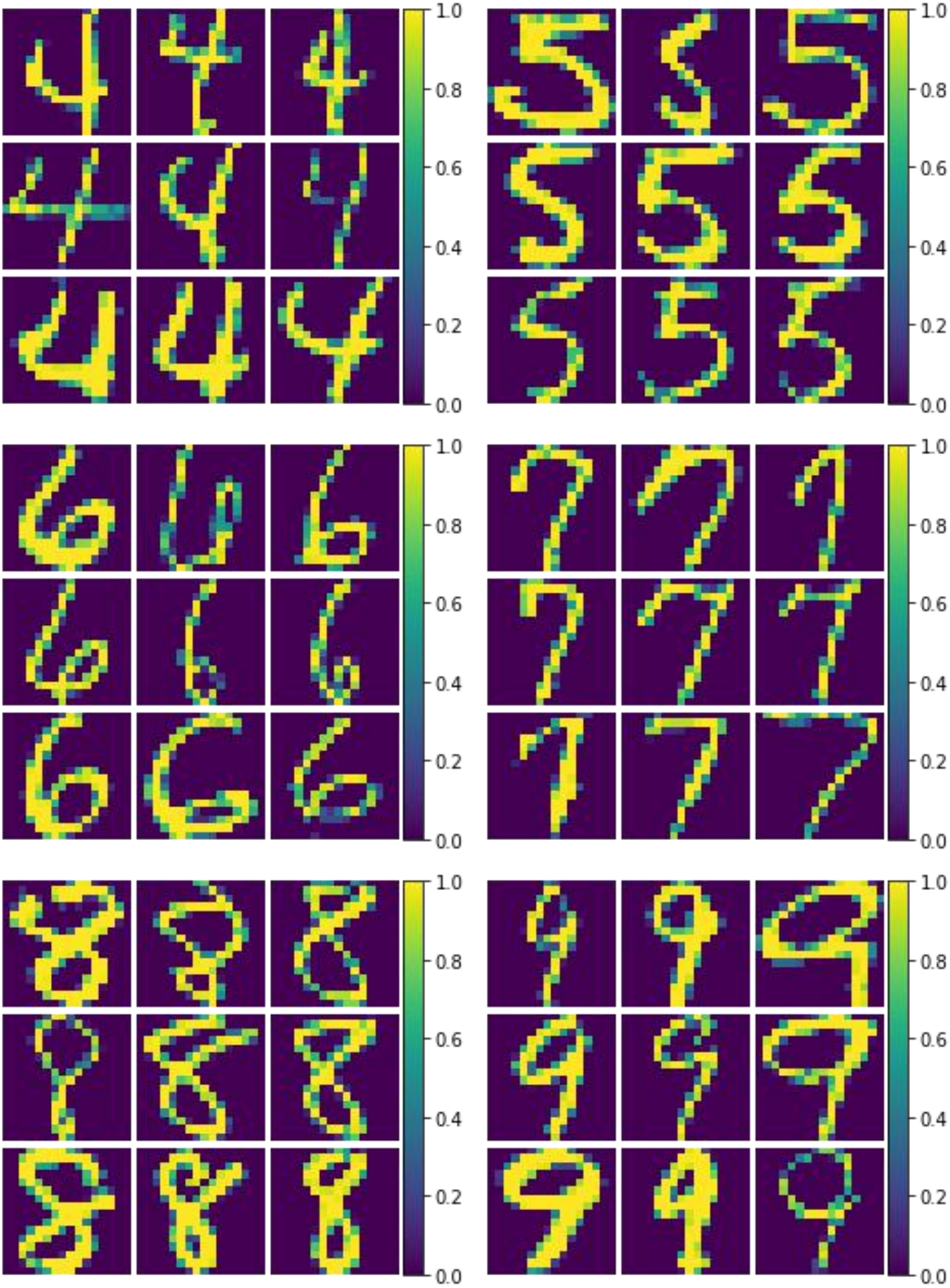
Name: Chee Kitt Win
Student Number: 45589140
STAT3006 Assignment 4

Problem 1

Part 1)

Plotted all the values to see if the plots in the later sections seem reasonable.




1.0
0.8
0.6
0.4
0.2
0.0

Part 2)

From the lecture notes, the solution to this optimization problem is

$$\hat{\mathbf{O}} = \underset{\mathbf{O}\in\mathbb{O}_{q,s}}{\arg\max}\ \text{trace}\left(\mathbf{O}^\top \left\{\frac{1}{n}\tilde{\mathbf{G}}\right\}\mathbf{O}\right)$$

where O_hat is an orthogonal matrix whose columns are the eigenvectors corresponding to the s largest eigenvalues of the covariance matrix G/n.

In the context of the question, R = O, F is the transpose of O, G is the Grammian matrix, n = 7291, q = 256 and s = 4.

To obtain the solution, I used Pythons scikit-learn library to conduct PCA. Below are the principal components. I have checked that they are of unit length, so **R_hat is simply the matrix whose columns are the principal components below, and F_hat is the transpose of that**. **The minimum value calculated using the R_hat and F_hat values is approximately 537801 as shown below (see python code for working).**

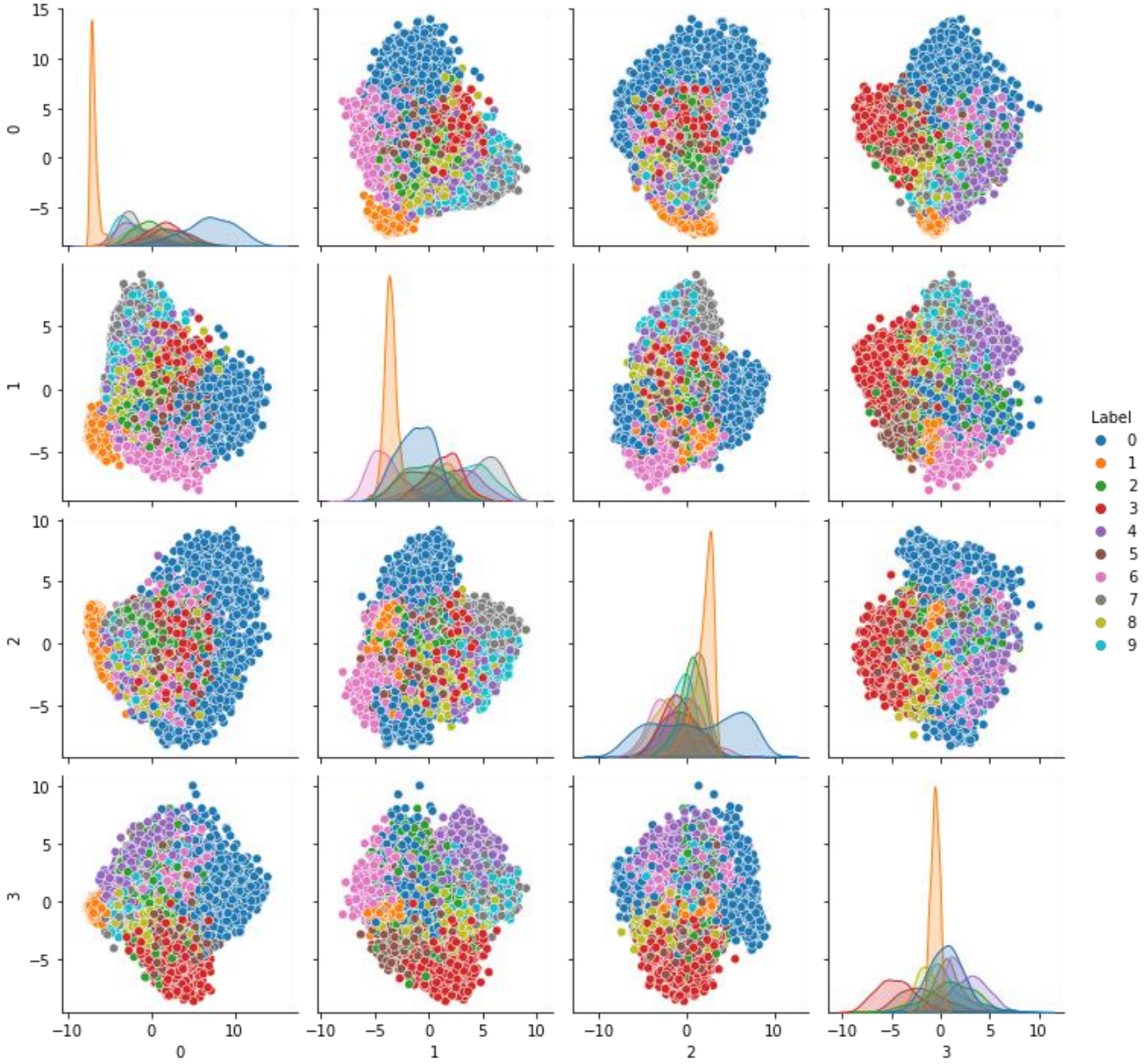
Note: Scikit-learn's documentation says that the data is centered but not scaled before carrying out the procedure. It also clarifies that the variance explained is indeed the largest eigenvalues of the covariance matrix. To verify both these points, I also did some manual checks and got the same answer.

```
PCs are:
[[-7.49436580e-05 -3.75198008e-04  5.97100977e-05 ...  3.27414017e-03
  1.40775226e-03  3.07120480e-04]
 [ 1.09434697e-03  5.63882368e-03  1.29255678e-02 ...  1.71485933e-03
  1.26625087e-03  3.59728021e-04]
 [ 4.73258827e-04  2.97612597e-03  4.58952048e-03 ...  1.68601021e-03
  1.40623810e-03  3.58303824e-04]
 [-4.33886510e-05 -3.07568708e-04 -6.39571589e-03 ... -5.96799644e-03
 -1.46911164e-06  7.70983255e-04]]

The minimum value is:
 537801.4612759223

Variance proportion explained by each PC:
 [0.17884424 0.0896704  0.06571727 0.05554546]


In [2]:
```

Part 3)

Below is a pair plot of the data using the principal components as the covariates. Although there is a lot of overlapping in the 2-d plots, we can definitely observe differences in distributions between different labels (digits) of the forward mapped observations. Similarly labeled observations tend to clump together, not straying very far away from their “color cluster”, and some digits such as 0 and 1 are well separated even in 2 dimensions as can be seen from the marginal pdf in the top left corner. This suggests that it is possible that the data is well separated in 4 dimensions, but we cannot tell just from this pairplot.

Also, we can see from the marginal pdfs that the distribution of digit 1 has very small variance and a sharp peak and is easily distinguishable from the other numbers. Digits such as 0 and 2 on the other hand, have distributions with high variance. This makes sense when looking at the plots in part 1; there are large differences in the handwritten 0’s and 2’s whereas all the 1’s are very similar.



Part 4)

From the same screenshot in Part 2), below are the variance proportions explained by the eigenvectors corresponding to the 4 largest eigenvalues. (The sum is 0.38977737289306497)

Although scikit-learn obtains the results below via the spectral decomposition of the covariance matrix G/n and not G, the variance proportions are the same since multiplying eigenvalues by a constant, then calculating the proportions will yield the same result as the constants cancel.

```
PCs are:
[[-7.49436580e-05 -3.75198008e-04  5.97100977e-05 ...  3.27414017e-03
  1.40775226e-03  3.07120480e-04]
 [ 1.09434697e-03  5.63882368e-03  1.29255678e-02 ...  1.71485933e-03
  1.26625087e-03  3.59728021e-04]
 [ 4.73258827e-04  2.97612597e-03  4.58952048e-03 ...  1.68601021e-03
  1.40623810e-03  3.58303824e-04]
 [-4.33886510e-05 -3.07568708e-04 -6.39571589e-03 ... -5.96799644e-03
 -1.46911164e-06  7.70983255e-04]]

The minimum value is:
 537801.4612759223

Variance proportion explained by each PC:
 [0.17884424 0.0896704  0.06571727 0.05554546]


In [2]:
```

Part 5)

As pointed out in Ed Discussion, I was able to get R working at this point, so from here onwards I use R instead of python.

Using ppca from Rdimtools:

I calculated **sigma2_hat** to be **0.292748**

Below are the **mu_hat** values which are simply the column means of the data, and also the first couple of **R_hat values (see the rest in R code)**.

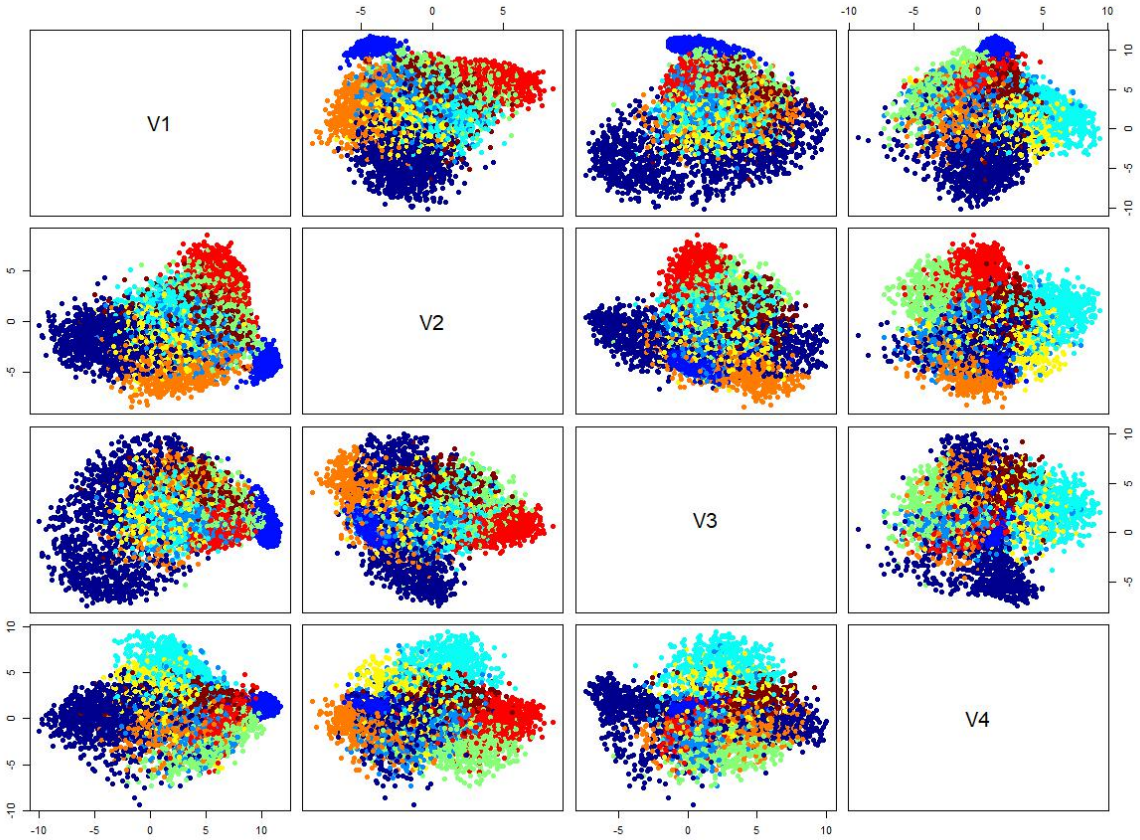
```
> mu_hat = colMeans(x)
> mu_hat
        V2          V3          V4          V5          V6          V7          V8          V9         V10         V11         V12
-0.99641736 -0.98113770 -0.95115293 -0.88773762 -0.77346839 -0.61030243 -0.36899108 -0.04576930 -0.05264052 -0.28456385 -0.50410602
       V13         V14         V15         V16         V17         V18         V19         V20         V21         V22         V23
-0.68646962 -0.81520176 -0.90615512 -0.96600315 -0.99336113 -0.98954259 -0.95133274 -0.86532314 -0.69586943 -0.43380497 -0.14343753
       V24         V25         V26         V27         V28         V29         V30         V31         V32         V33         V34
 0.13066836  0.41182362  0.41477548  0.14882924 -0.07425758 -0.34720793 -0.61392182 -0.80006213 -0.91840104 -0.97898505 -0.98276862
       V35         V36         V37         V38         V39         V40         V41         V42         V43         V44         V45
-0.92740502 -0.80057400 -0.54983665 -0.23338280 -0.02354986  0.06606666  0.21097188  0.16134536 -0.03532945 -0.08632520 -0.21367234
       V46         V47         V48         V49         V50         V51         V52         V53         V54         V55         V56
-0.49748786 -0.75960774 -0.89965149 -0.97180167 -0.97680154 -0.90895131 -0.74896777 -0.46434920 -0.18955054 -0.13014484 -0.17676505
       V57         V58         V59         V60         V61         V62         V63         V64         V65         V66         V67
-0.11284693 -0.17799328 -0.27096983 -0.17684200 -0.18531463 -0.43489247 -0.72671787 -0.90001481 -0.97690427 -0.97374325 -0.88966356
       V68         V69         V70         V71         V72         V73         V74         V75         V76         V77         V78
-0.70456837 -0.42396352 -0.21732890 -0.24074229 -0.33272500 -0.30440955 -0.34018242 -0.34819984 -0.19959813 -0.18220368 -0.41584666
       V79         V80         V81         V82         V83         V84         V85         V86         V87         V88         V89
-0.70422603 -0.89292443 -0.98112262 -0.96973351 -0.86256631 -0.67012838 -0.42274722 -0.25538486 -0.27823755 -0.36624359 -0.33399808
       V90         V91         V92         V93         V94         V95         V96         V97         V98         V99        V100
-0.31456919 -0.28907667 -0.19652544 -0.23002249 -0.45037910 -0.70119545 -0.87633315 -0.97750199 -0.96035743 -0.83136689 -0.64994294
      V101        V102        V103        V104        V105        V106        V107        V108        V109        V110        V111
-0.44317117 -0.29381895 -0.28255658 -0.31940996 -0.23995611 -0.18488095 -0.19880222 -0.18980634 -0.28970909 -0.49390344 -0.69817844
      V112        V113        V114        V115        V116        V117        V118        V119        V120        V121        V122
-0.85035455 -0.96085009 -0.94271609 -0.80844589 -0.64162735 -0.46072212 -0.34522329 -0.33011137 -0.31467659 -0.18803689 -0.12074887
      V123        V124        V125        V126        V127        V128        V129        V130        V131        V132        V133
-0.13768221 -0.18626060 -0.31927335 -0.50237759 -0.67446537 -0.81820464 -0.93757043 -0.92304581 -0.78830352 -0.63331381 -0.48603497
      V134        V135        V136        V137        V138        V139        V140        V141        V142        V143        V144
-0.41303648 -0.43077150 -0.39477616 -0.22094089 -0.11522507 -0.11746482 -0.19631354 -0.33080593 -0.47101001 -0.62227253 -0.77695844
      V145        V146        V147        V148        V149        V150        V151        V152        V153        V154        V155
-0.91585105 -0.90881511 -0.75949842 -0.60616843 -0.49838733 -0.47688644 -0.52257427 -0.47603497 -0.27293636 -0.10799314 -0.12107886
      V156        V157        V158        V159        V160        V161        V162        V163        V164        V165        V166
-0.24894377 -0.35619394 -0.43530901 -0.57268934 -0.74379139 -0.90581210 -0.89671828 -0.72395378 -0.56724894 -0.47838637 -0.49623879
      V167        V168        V169        V170        V171        V172        V173        V174        V175        V176        V177
-0.56999479 -0.51698217 -0.28318036 -0.07554437 -0.15532478 -0.32031175 -0.35619147 -0.38121273 -0.53032026 -0.73498601 -0.91431820
      V178        V179        V180        V181        V182        V183        V184        V185        V186        V187        V188
-0.90378617 -0.71132520 -0.52161295 -0.42065862 -0.45762803 -0.56248608 -0.51042902 -0.24683925 -0.05637923 -0.19805569 -0.33099369
      V189        V190        V191        V192        V193        V194        V195        V196        V197        V198        V199
-0.30505418 -0.33134920 -0.52619503 -0.75758510 -0.93110822 -0.93014785 -0.74992004 -0.52534344 -0.35618749 -0.34719284 -0.44735249
      V200        V201        V202        V203        V204        V205        V206        V207        V208        V209        V210
-0.40732369 -0.12532725  0.02026416 -0.15981580 -0.21988932 -0.20774366 -0.33708586 -0.58416527 -0.81561350 -0.95034138 -0.95776437
      V211        V212        V213        V214        V215        V216        V217        V218        V219        V220        V221
-0.83362968 -0.62045851 -0.36969593 -0.21724592 -0.19723097 -0.10587862  0.17212769  0.25563818  0.03748388 -0.07681964 -0.22228967
      V222        V223        V224        V225        V226        V227        V228        V229        V230        V231        V232
-0.47101756 -0.72274215 -0.88165204 -0.96525525 -0.98147716 -0.92649582 -0.80531861 -0.58313784 -0.29751680 -0.06522795  0.18089549
      V233        V234        V235        V236        V237        V238        V239        V240        V241        V242        V243
 0.49034357  0.48573831  0.13808929 -0.18041778 -0.47468262 -0.71973364 -0.86912248 -0.93853244 -0.98017515 -0.99665012 -0.98276972
      V244        V245        V246        V247        V248        V249        V250        V251        V252        V253        V254
-0.95105047 -0.87788040 -0.73381457 -0.50233342 -0.19854656  0.13982266  0.11643890 -0.31410835 -0.65371074 -0.83785777 -0.92219366
      V255        V256        V257
-0.95739268 -0.97928995 -0.99467782
>
```

```
> R_hat = PPCA$mle.W
> R_hat
              [,1]         [,2]         [,3]          [,4]
  [1,]  1.598437e-03  0.011543062 -0.003621360  2.786090e-04
  [2,]  8.002417e-03  0.059477738 -0.022773274  1.975086e-03
  [3,] -1.273528e-03  0.136337554 -0.035118702  4.107529e-02
  [4,] -5.306307e-02  0.277006362 -0.044632031  1.420633e-01
  [5,] -1.819299e-01  0.472276490 -0.045557957  3.020935e-01
  [6,] -4.860439e-01  0.660155739  0.003872391  4.706114e-01
  [7,] -6.161120e-01  0.610393429  0.079409277  6.670897e-01
  [8,]  1.861617e-01  0.080078493 -0.015257424  8.265926e-01
  [9,] -1.368748e-01  0.252545963  0.265318430  7.387706e-01
 [10,] -5.813322e-01  0.704830133  0.470473009  5.343710e-01
 [11,] -2.732503e-01  0.861121437  0.339262766  3.387425e-01
 [12,] -1.198680e-01  0.766689256  0.168161613  1.321166e-01
 [13,] -9.600254e-02  0.540291989  0.048973789 -3.420500e-03
 [14,] -6.109574e-02  0.306854211 -0.021738970 -4.131380e-02
 [15,] -2.330798e-02  0.111777865 -0.027722479 -2.559806e-02
 [16,] -6.367234e-03  0.018101952 -0.008338727 -7.438066e-03
 [17,]  2.652691e-03  0.034655800 -0.013305304  1.320531e-03
 [18,] -5.322693e-03  0.142461168 -0.049644671  2.187656e-02
 [19,] -9.455073e-02  0.355014754 -0.102242105  1.339091e-01
 [20,] -3.029145e-01  0.749877781 -0.165783112  3.508292e-01
 [21,] -6.848262e-01  1.224473137 -0.195737618  5.767857e-01
 [22,] -1.250479e+00  1.510382339 -0.118387497  7.213824e-01
 [23,] -1.285161e+00  1.193916149  0.070280533  8.005892e-01
 [24,] -2.924363e-01  0.134345112  0.007150534  8.473976e-01
 [25,] -4.452427e-01 -0.167097419  0.208806559  8.518397e-01
 [26,] -1.441518e+00  0.674144120  0.468247124  8.288268e-01
 [27,] -1.180818e+00  1.352446830  0.332655864  7.464788e-01
 [28,] -5.761161e-01  1.518786986  0.145239499  4.483772e-01
 [29,] -3.137015e-01  1.139353083  0.005331146  1.058880e-01
 [30,] -1.737289e-01  0.647405996 -0.088348508 -4.164180e-02
 [31,] -7.218432e-02  0.263425271 -0.070194664 -4.455516e-02
 [32,] -1.728737e-02  0.060397129 -0.024893851 -1.937422e-02
 [33,] -6.863282e-05  0.044523641 -0.022877536 -4.992944e-03
 [34,] -6.677158e-02  0.176948011 -0.089476234  1.524214e-02
 [35,] -2.740193e-01  0.480628401 -0.207033726  9.062768e-02
 [36,] -7.006810e-01  1.023685652 -0.350145204  2.193528e-01
 [37,] -1.249875e+00  1.507002224 -0.353493456  2.709327e-01
 [38,] -1.746768e+00  1.389943336 -0.068228681  1.435711e-01
```

Part 6)

Below are the first few instances of the required posterior expectations of latent variables and the plot obtained using the same package. The different colors represent different digits.

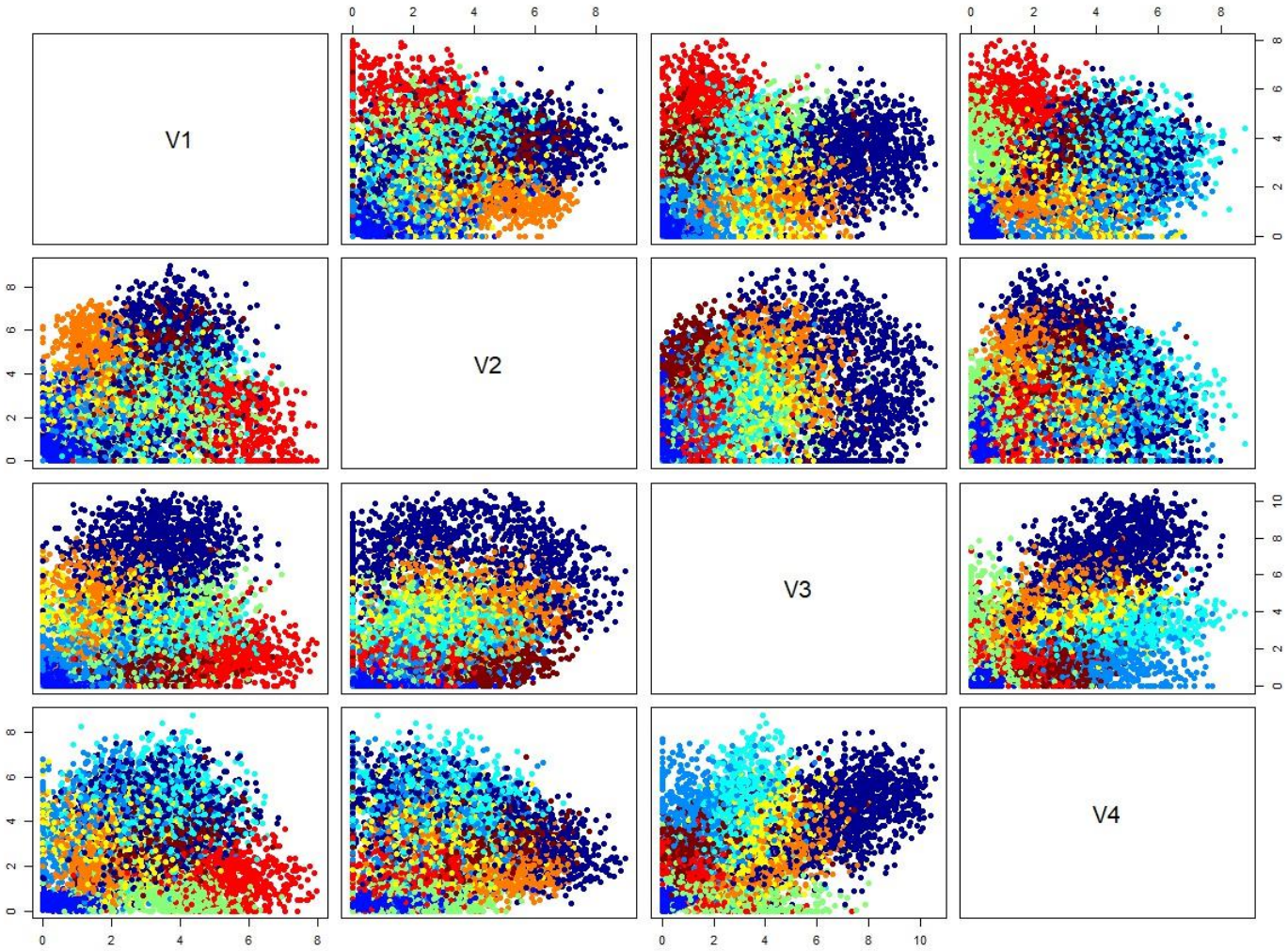
```
> PPCA$Y
                [,1]               [,2]               [,3]               [,4]
 [1,] -1.01031446101128886 -6.71282427904492618  4.09294393949569745499 -1.396035599782883274
 [2,] -3.03212723862314082 -1.64169380696406741  0.3242778371794621 25  4.660444327679612364
 [3,]  2.95288108028334939  1.44413257629585745  5.3578291529235064 77 -4.393213768311785117
 [4,]  6.44815555418329822  6.04429007619560288  0.0997055118112594 73  0.515233494437488226
 [5,]  1.24537053899668826 -0.78503305010853341  2.2707692596084876 34  8.177409732939615594
 [6,]  1.51653407398315498  0.81222139555736594  4.9663901069195253 16 -1.924951980277555652
 [7,]  4.20927445036154602 -0.03118562776357437  3.7421102470533829 10  5.182315109999372282
 [8,] 10.98653173151751083 -3.48471344586260345 -0.6960271875487982 21  1.017314340786022786
 [9,] -2.03181464425629654 -0.72776182051351035 -5.5275218192899542 30  1.870539016611964866
[10,] 10.49363776544636551 -3.49308821157536409 -0.5408568283044824 51  1.214136265857573616
[11,]  1.88749744964444899  2.14319495597742682  0.0844439249813886 68 -2.120005904671912855
[12,] -0.04858661580569559 -0.78864719487950319  8.0882823781161707 90 -2.487370287902588739
[13,] 11.25350310628525996 -4.00380755232128038 -1.2564474518078505 57  0.953493804642923659
[14,] 10.98259666098336851 -3.82079277921594374 -1.2949697842305978 31  0.896618786855193473
[15,]  6.70236739045634344  1.21678842421515987  3.5677836772566875 47 -0.688666068790277364
[16,]  6.98641654851563754  3.85216243089803712  0.8231641763482112 63  1.685386559924301864
[17,]  7.44203843670810006  0.83953614932142584  1.2195452517022395 97  1.770780942538719982
[18,] -0.27445749614282894  2.40399104559458587  5.2923138746875943 99  2.061634817816341503
[19,] -3.76239699674368966 -1.47830203321360010  2.4467397638897812 05 -2.758379965767777442
[20,]  9.33586302381087663 -4.70473252939157760  0.9043510995169637 97  0.402558423165401635
[21,]  9.17322737449218728 -3.27168861982713066  3.9204374043719854 01 -1.224574102829488309
[22,]  3.43523615509376867 -0.10415427276047801  2.7724729704038115 76 -3.040334028810641076
[23,]  5.85675763027336949  6.42343117481851511 -0.4132504147235938 62 -0.636626200077209492
[24,]  6.16169927644144799  2.88780209300540180 -0.5163245192677039 52 -3.175139303683335079
[25,]  2.70547705050096265 -0.84055458259563454  5.0515140777652112 19  2.954400441483790107
[26,]  7.27960313496876577  4.49416757637509079  0.3895337968444124 17  0.342817798974192567
[27,]  2.52407388117940057  1.62088373462267787  1.7271126038704625 04  6.082296788570584312
[28,]  6.89234294733133002  4.11784618233609478  2.6105283493978914 68 -1.447478389074432714
[29,]  6.33166703917247276  3.45311712465411302  3.4129216472292300 82 -3.061184894476797425
[30,] 10.63725010275088856 -4.52439627629858698  0.2246552701066392 71  1.192316034512445277
[31,]  0.75792976006727086 -0.98909512420038648  1.2243319835995116 75  6.035198255458447569
[32,]  3.99693121275465346 -5.41288564088768354  5.1777274838385896 06 -0.264723542933202605
[33,]  7.85228057746776376  5.01206904283278742  0.4713770792938671 29  1.302879128501228223
[34,]  7.56358357609312471  2.89312874783475005  2.0699034783384351 32 -0.995611369898624532
[35,] 11.16701288552327753 -4.03459097657542110 -0.9908104274520191 93  1.126446353822368573
[36,]  5.38096950119803985 -0.10623824620908071  2.1307391530384465 21  4.113068984946327333
[37,]  7.88788963729697823  3.11226863386552166 -0.6881021765522828 96 -0.313582504542854357
[38,]  5.79503123103431506  4.85431668424090645  0.2238899198681802 43 -1.097471511700931712
```



Part 7)

Using the autoencoder code provided by Hien, I used the rectified exponential linear unit (ReLU) activation function defined below to produce the following plot of the encoded data. Again, the different colors represent different digits.

$$f(x) = \begin{cases} 0 & \text{for} \quad x < 0 \\ x & \text{for} \quad x \geq 0 \end{cases}$$



Problem 2

Part 1)

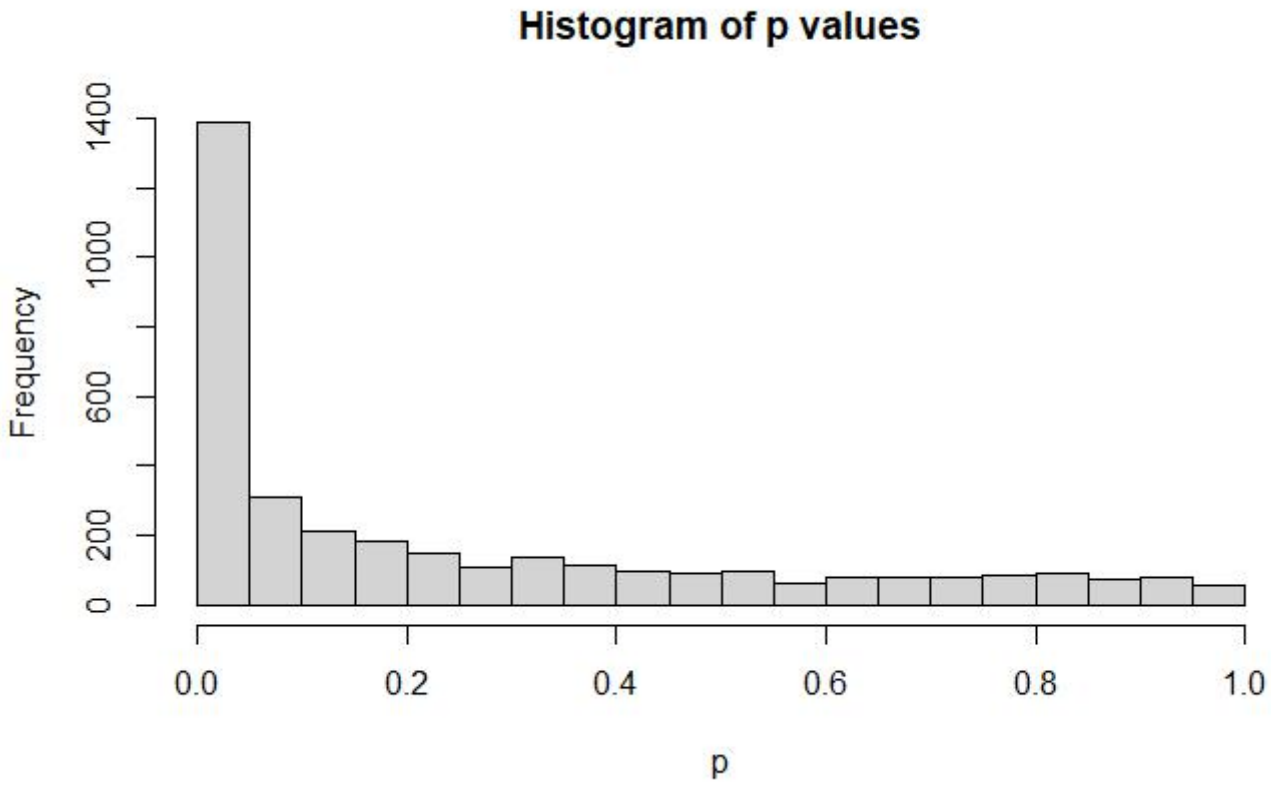
For this section I used the code that Hien provided. Inspecting the for loops in the code, it appears that sigma_sq should be equal to 1/(2*Beta) despite what the comment above the code says.

I obtained a U-statistic of 7.67428677632695e-06 and critical value of 1.23897406294995, which suggests that the null hypothesis be retained since the U statistic is less than the critical value.

Based on intuition, with 3571 dimensions and only a sample size of 47 and 25 for the ALL and AML cell types respectively, I believe that the test was not powerful enough to reject the null.
A more concrete reason is that the dimension plays a role because the U-kernel is taking as input 2 high dimensional vectors. The L2-norm square of the difference between 2 high dimensional vectors will be large, and since the U-kernel is Gaussian, this results in a smaller U-statistic.

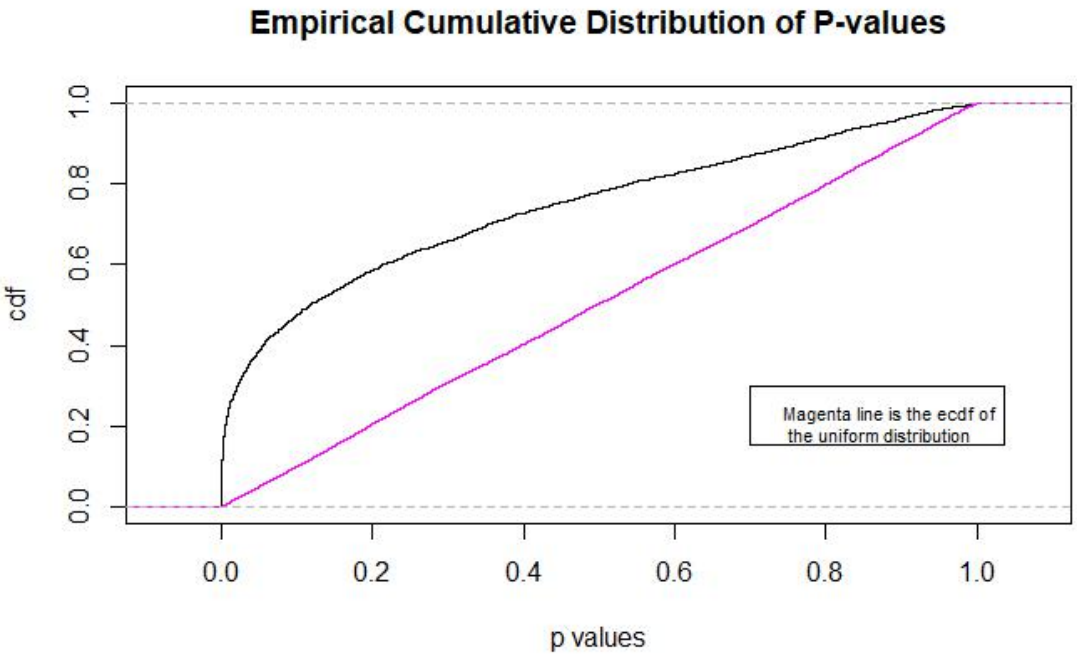
Part 2)

The p values below were obtained using a 2-sample t test, and can be obtained from the R code.



Part 3)

The distribution of p-values is not sub-uniform as this would require that the ECDF be below the magenta line. This observation suggests that there are more significant p-values than we would expect if all the means of the distribution for individual features/covariates between ALL and AML cells were indeed equal to each other. This does not conform to the conclusions made in Part 1) since features/covariates between ALL and AML cells having different means suggests that there is a difference between the distributions of the two cell types.



Part 4)

Below are integers representing indices (thought this would be more convenient rather than listing the actual names of the genes) corresponding to the genes that are significant for the Benjamini-Hochberg method. There are 949 such genes. The largest p-value under this method (corresponding to a rejected null hypothesis) is 0.0132538998571417.

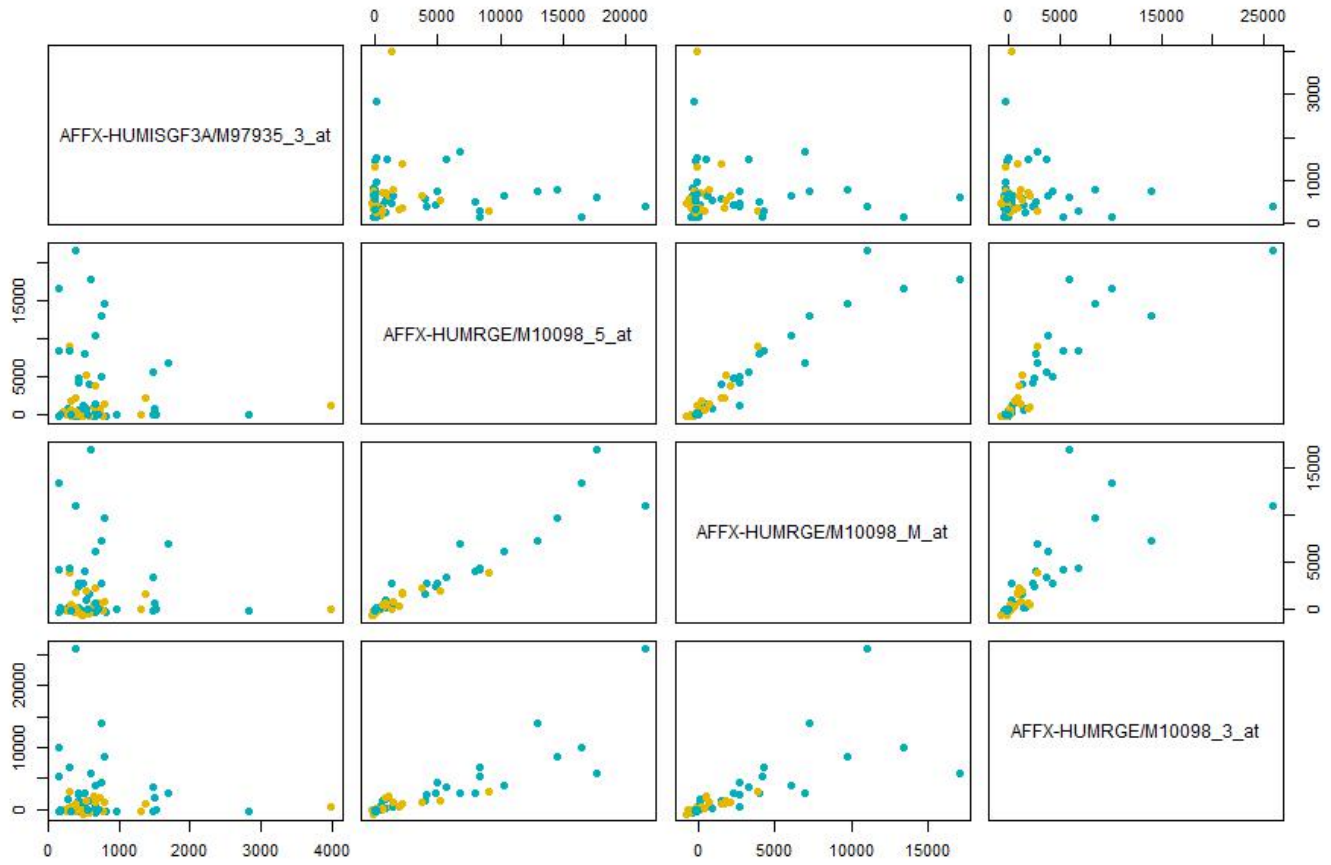
```
> BH_sig_genes
  [1]    7   16   17   26   35   38   41   46   48   52   54   56   59   65   66   69   73   76   78   84   87   88   91   93  100  103  106
 [28]  110  115  120  122  124  125  127  135  140  141  142  145  147  158  160  163  171  174  178  179  181  183  184  188  194  200  209
 [55]  210  211  214  217  219  221  230  233  236  244  245  252  258  260  264  267  269  271  272  274  276  289  290  295  296  307  313
 [82]  320  322  325  327  333  338  339  345  351  357  358  360  364  367  369  376  377  379  385  387  393  394  396  397  401  403  406
[109]  410  412  418  425  431  435  436  439  441  446  449  450  453  456  462  463  465  466  472  490  493  494  497  499  500  503  508
[136]  510  520  529  533  534  536  548  554  565  567  570  571  572  573  583  584  589  590  592  593  595  602  604  608  614  615  617
[163]  621  623  624  626  627  631  633  634  635  644  647  650  652  653  657  662  663  669  670  672  673  675  676  678  680  681  682
[190]  685  687  695  707  723  724  730  731  734  738  740  741  742  747  750  751  756  766  771  773  774  776  779  794  799  801  817
[217]  818  821  828  831  834  837  838  843  845  847  851  854  860  861  866  869  871  874  878  882  884  894  899  900  903  904  907
[244]  911  918  920  922  925  930  931  932  939  942  945  948  949  951  955  956  957  960  979  980  981  985  987  988  990  994  998
[271] 1000 1001 1004 1005 1011 1012 1014 1016 1018 1020 1028 1032 1041 1044 1045 1049 1050 1053 1057 1062 1068 1072 1081 1087 1093 1095 1098
[298] 1099 1100 1102 1104 1109 1111 1112 1115 1130 1131 1133 1134 1135 1140 1145 1146 1149 1154 1157 1161 1163 1164 1166 1167 1174 1178 1179
[325] 1182 1183 1184 1189 1192 1194 1198 1205 1214 1215 1219 1221 1222 1224 1225 1228 1232 1234 1236 1239 1240 1241 1245 1249 1251 1257 1260
[352] 1261 1263 1267 1268 1276 1283 1289 1293 1294 1301 1307 1312 1314 1318 1325 1331 1335 1336 1340 1341 1355 1356 1358 1363 1365 1371 1372
[379] 1376 1384 1392 1393 1403 1407 1409 1410 1421 1424 1429 1431 1433 1438 1446 1451 1452 1454 1463 1464 1472 1480 1483 1485 1500 1503 1504
[406] 1506 1509 1511 1515 1523 1526 1530 1531 1534 1535 1540 1543 1548 1550 1552 1556 1557 1559 1560 1563 1564 1569 1574 1578 1586 1587 1588
[433] 1592 1597 1608 1609 1612 1619 1620 1621 1622 1627 1628 1635 1639 1642 1643 1652 1656 1661 1666 1668 1675 1676 1677 1681 1684 1691 1693
[460] 1695 1698 1700 1701 1702 1703 1704 1705 1706 1709 1713 1715 1718 1719 1721 1723 1727 1729 1738 1741 1743 1747 1749 1753 1757 1762 1766
[487] 1768 1769 1775 1779 1780 1789 1804 1805 1822 1823 1825 1830 1831 1835 1845 1846 1853 1860 1863 1867 1869 1872 1876 1879 1881 1883 1890
[514] 1898 1910 1912 1914 1915 1916 1917 1919 1921 1930 1946 1949 1951 1952 1953 1960 1968 1971 1979 1982 1985 1988 1990 1995 1999 2004 2011
[541] 2016 2027 2028 2037 2048 2049 2051 2052 2060 2063 2065 2067 2068 2070 2071 2073 2079 2085 2091 2098 2101 2102 2104 2106 2109 2111 2114
[568] 2116 2118 2123 2130 2131 2132 2133 2134 2135 2136 2137 2138 2141 2145 2146 2147 2159 2161 2166 2167 2169 2171 2172 2173 2177 2180 2181
[595] 2185 2195 2196 2198 2199 2200 2204 2205 2220 2222 2225 2226 2228 2229 2230 2232 2233 2235 2238 2239 2245 2246 2249 2254 2255 2260 2265
[622] 2268 2269 2270 2276 2277 2282 2285 2286 2289 2291 2292 2293 2298 2301 2302 2304 2305 2312 2316 2317 2318 2321 2323 2327 2331 2333 2339
[649] 2353 2366 2367 2371 2374 2380 2383 2387 2392 2402 2404 2407 2412 2414 2415 2418 2419 2422 2423 2424 2431 2434 2436 2437 2439 2440 2442
[676] 2447 2449 2455 2460 2462 2468 2472 2473 2476 2481 2489 2503 2507 2517 2520 2525 2526 2529 2530 2532 2537 2546 2547 2549 2561 2562 2569
[703] 2571 2578 2582 2591 2594 2596 2597 2605 2607 2610 2612 2614 2617 2618 2622 2628 2631 2641 2648 2652 2653 2655 2656 2666 2668 2670 2671
[730] 2678 2682 2683 2686 2687 2690 2693 2695 2696 2700 2701 2713 2716 2717 2723 2727 2732 2734 2740 2744 2745 2749 2752 2753 2762 2763 2773
[757] 2774 2789 2790 2807 2810 2811 2815 2819 2820 2827 2828 2831 2833 2838 2840 2841 2850 2852 2855 2859 2862 2866 2872 2875 2881 2884 2893
[784] 2898 2911 2913 2915 2917 2921 2923 2924 2929 2931 2936 2938 2946 2953 2955 2958 2959 2964 2969 2979 2980 2984 2988 2989 2990 2991 2996
[811] 2997 3005 3009 3013 3023 3026 3029 3037 3038 3043 3055 3060 3067 3068 3072 3085 3092 3093 3095 3096 3098 3103 3104 3105 3108 3112 3116
[838] 3117 3120 3121 3122 3123 3125 3129 3130 3132 3136 3138 3139 3145 3151 3152 3153 3155 3157 3162 3163 3164 3165 3177 3178 3182 3191 3193
[865] 3197 3199 3201 3208 3214 3216 3217 3218 3221 3223 3230 3236 3259 3262 3266 3272 3276 3278 3279 3280 3281 3282 3292 3294 3296 3301 3303
[892] 3308 3309 3314 3316 3322 3323 3326 3327 3331 3336 3345 3349 3351 3358 3359 3363 3368 3370 3376 3381 3382 3388 3398 3402 3410 3414 3417
[919] 3418 3419 3422 3432 3437 3438 3441 3446 3452 3455 3463 3466 3467 3468 3469 3477 3478 3480 3484 3487 3488 3489 3492 3495 3496 3500 3501
[946] 3514 3515 3531 3566
```

And below for the Benjamini-Yekutielie method. There are 416 such genes. The largest p-value (corresponding to a rejected null hypothesis) under this method is 0.000660433510394291.

```
> BY_sig_genes
  [1]   26   35   38   65   73   76   78   91  106  110  135  140  141  142  145  160  171  181  183  211  214  219  258  264  269  276  289
 [28]  295  320  322  325  357  358  364  367  376  377  385  393  394  418  435  436  439  441  446  456  462  472  493  497  500  510  534
 [55]  548  570  583  584  589  590  615  624  626  631  635  652  657  662  672  673  675  678  685  687  695  723  734  741  742  750  751
 [82]  818  821  837  843  845  847  851  861  869  871  874  884  894  899  907  918  922  931  932  949  951  956  960  979  985  987  990
[109]  994 1001 1004 1016 1020 1044 1049 1050 1053 1072 1081 1093 1095 1098 1099 1100 1111 1112 1133 1146 1149 1157 1161 1163 1166 1179 1182
[136] 1184 1189 1198 1205 1214 1219 1222 1225 1239 1249 1251 1257 1260 1268 1283 1289 1293 1294 1307 1312 1335 1336 1356 1363 1365 1392 1407
[163] 1410 1424 1433 1438 1446 1454 1463 1483 1485 1503 1509 1511 1523 1530 1534 1535 1548 1556 1560 1563 1574 1578 1586 1587 1597 1609 1612
[190] 1620 1643 1652 1656 1661 1693 1702 1704 1706 1715 1718 1719 1723 1741 1743 1766 1768 1775 1780 1804 1805 1822 1825 1835 1846 1853 1863
[217] 1869 1879 1881 1890 1912 1916 1917 1919 1946 1953 1968 1971 1979 1985 1988 1990 2004 2011 2016 2028 2037 2051 2052 2063 2070 2071 2079
[244] 2085 2098 2101 2109 2114 2116 2118 2123 2130 2131 2134 2135 2136 2141 2145 2147 2161 2181 2195 2196 2198 2200 2220 2226 2228 2230 2235
[271] 2239 2249 2260 2268 2269 2270 2277 2282 2285 2292 2298 2301 2305 2312 2321 2323 2331 2353 2367 2387 2392 2404 2412 2419 2422 2431 2434
[298] 2440 2442 2460 2476 2481 2507 2520 2526 2530 2532 2537 2546 2547 2582 2591 2594 2596 2612 2617 2618 2641 2648 2652 2653 2655 2666 2670
[325] 2678 2682 2690 2693 2700 2713 2723 2727 2734 2749 2752 2773 2774 2789 2810 2820 2827 2828 2833 2841 2850 2859 2872 2875 2911 2917 2931
[352] 2936 2989 3009 3023 3026 3029 3038 3043 3060 3072 3092 3093 3096 3098 3103 3105 3108 3116 3117 3130 3136 3145 3162 3163 3182 3201 3214
[379] 3216 3221 3223 3230 3259 3262 3266 3272 3279 3280 3282 3292 3301 3309 3323 3336 3370 3381 3382 3398 3402 3417 3418 3419 3438 3441 3446
[406] 3455 3463 3466 3467 3469 3484 3488 3489 3492 3514 3566
```

Part 5)

From the scatterplot matrix below (of a subset of genes that I chose after a bit of exploration), we can see that there is clear dependence between some features (genes). The 2nd and 3rd genes (the two in the center) in particular seem to have strong linear correlation. The Benjamini-Hochberg method requires that the p-values are independent. In this scenario, the Benjamini_Hochberg assumptions are not met since dependence between features implies dependence between p-values. The Benjamini_Yekutieli does not require this assumption and in this situation is valid. The tradeoff is that it is more conservative and is a less powerful test as reflected in part 4.



Part 6)

6. $$\sum_{j=1}^{\infty} \frac{\delta(\min\{j,m\})}{j(j+1)} = \sum_{j=1}^{m} \frac{\delta(\min\{j,m\})}{j(j+1)} + \sum_{j=m+1}^{\infty} \frac{\delta(\min\{j,m\}}{j(j+1)}$$

$$= \sum_{j=1}^{m} \frac{\delta(j)}{j(j+1)} + \sum_{j=m+1}^{\infty} \frac{\delta(m)}{j(j+1)}$$

Substitute $\delta(k) = (2m)^{-1} k(k+1)$ as given

$$= \frac{1}{2m} \sum_{j=1}^{m} \frac{j(j+1)}{j(j+1)} + \frac{1}{2m} \sum_{j=m+1}^{\infty} \frac{m(m+1)}{j(j+1)}$$

$$= \frac{1}{2m}(m) + \frac{1}{2} \sum_{j=m+1}^{\infty} \frac{(m+1)}{j(j+1)}$$

$$= \frac{1}{2} + \frac{1}{2} \sum_{j=m+1}^{\infty} \frac{(m+1)}{j(j+1)}$$

from notes we know that $\sum_{j=1}^{\infty} \frac{1}{j(j+1)} = 1$

$$= \frac{1}{2} + \frac{m+1}{2}\left(1 - \sum_{j=1}^{m} \frac{1}{j(j+1)}\right)$$

$$= \frac{1}{2} + \frac{m+1}{2}\left(1 - \sum_{j=1}^{m}\left(\frac{1}{j} - \frac{1}{j+1}\right)\right)$$ telescoping series

$$= \frac{1}{2} + \frac{m+1}{2}\left(1 - \left(\frac{1}{1} - \frac{1}{2} + \frac{1}{2} - \frac{1}{3} + \frac{1}{3} - \dots - \frac{1}{m} + \frac{1}{m} - \frac{1}{m+1}\right)\right)$$

$$= \frac{1}{2} + \frac{m+1}{2}\left(\frac{1}{m+1}\right)$$

$$= 1 \leq 1$$ □

Let $H_{01}, \dots H_{0m}$ be tested using p-values $P_1, \dots, P_m$ and suppose that we order the p-values as follows: $P_{(1)} \leq P_{(2)} \leq \dots \leq P_{(m)}$

We reject H_{0j} via the rule $R_j = [\![P_j \leq \frac{\alpha_{FD}\,\delta(k^*)}{m}]\!]$ where

$k^* = \arg\max_{k \in [m]} \left\{ P_{(k)} \leq \frac{\alpha_{FD}\,\delta(k)}{m} \right\}$, and $\delta(k)$ is the function in the question.

Since we prove that $\sum_{j=1}^{\infty} \frac{\delta(\min\{j,m\})}{j(j+1)} \leq 1$, we know that

$FDR \leq \alpha \frac{m_0}{m} \sum_{j=1}^{\infty} \frac{\delta(\min\{j,m\})}{j(j+1)}$ can be controlled at size α_{FD}

Problem 3

Part 1)

1. From the lecture notes, we know that assuming problem 2 has a solution for some $\gamma > 0$, it is equivalent to problem 3 for some $\lambda > 0$ if $\frac{1}{2n}\sum_{i=1}^{n}(Y_i - \alpha - \beta^T X_i)^2$ and $\|\beta\|_1 + \|\beta\|_2^2$ are both convex.

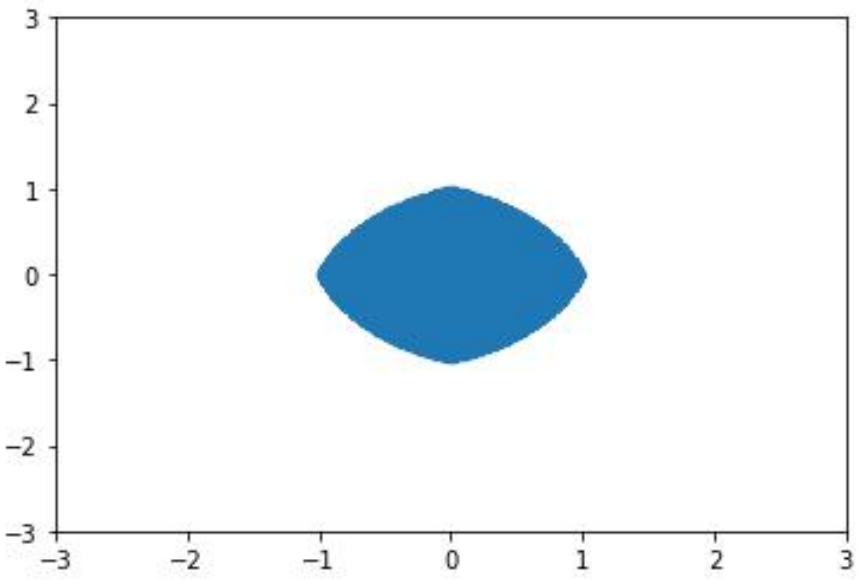
$\frac{1}{2n}\sum_{i=1}^{n}(Y_i - \alpha - \beta^T X_i)^2$ is a semi-positive definite quadratic and is hence convex.

From lecture notes, we know that $\|\beta\|_1$ and $\|\beta\|_2^2$ are individually convex. Moreover, $\|\beta\|_2^2$ is strongly convex, so $\|\beta\|_1 + \|\beta\|_2^2$ must also be convex.

Hence the conditions are met.

Part 2)

Below is a plot for gamma = 2 and q = 2 showing the admissible region for Beta in the constrained regression problem. In the lectures we saw that for lasso, the estimator can be sparse because the admissible region is diamond shaped, and the solution for the constrained problem can coincide with one of its corners (which is sparse because the sparse region is the x axis and y axis in the admissible region). Using similar reasoning, the elastic net-penalized least squares estimator **can be sparse** because the admissible region for Beta also has corners, which might coincide with the solution ("touch" the smallest contour as in lecture slides).



Part 3)

Solve Problem (2) using coordinate descent.

We want to solve the following

$$\hat{\alpha}, \hat{\beta} = \underset{\alpha, \beta \in \mathbb{R}^q}{\arg\min} \underbrace{\left(\frac{1}{2n} \sum_{i=1}^{n} (Y_i - \alpha - \beta^T X_i)^2 + \lambda (\|\beta\|_1 + \|\beta\|_2^2) \right)}_{\text{Call this } E}$$

for some $\lambda \geq 0$

Setting the partial derivative with respect to α to 0, we have

$$\frac{1}{n} \sum_{i=1}^{n} (Y_i - \alpha - \beta^T X_i)(-1) = 0$$

$$\Rightarrow \sum_{i=1}^{n} (Y_i - \beta^T X_i) = n\alpha$$

$$\Rightarrow \alpha = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \beta^T X_i)$$

For fixed $\beta = \beta^{(s-1)}$ where s denotes the sth iteration of coordinate descent,

$$\alpha^{(s)} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \beta^{T(s-1)} X_i)$$

Now fix $\alpha = \alpha^{(s-1)}$ and $\beta_{-j} = \beta_{-j}^{(s-1)}$. We can write the sub differential of the univariate function $E(\beta_j ; \alpha^{(s-1)}, \beta_{-j}^{(s-1)})$ for each $j \in [q]$ as

$$\partial E(\beta_j ; \alpha^{(s-1)}, \beta_{-j}^{(s-1)}) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \alpha^{(s-1)} - \sum_{k \neq j} \beta_k^{T(s-1)} X_{ik} - \beta_j X_{ij})(-X_{ij})$$

$$+ \lambda \begin{cases} -1 + 2\beta_j & \beta_j < 0 \\ [-1, 1] & \beta_j = 0 \\ 1 + 2\beta_j & \beta_j > 0 \end{cases}$$

$$= -\frac{1}{n} \sum_{i=1}^{n} X_{ij} (Y_i - \alpha^{(s-1)} - \sum_{k \neq j} \beta_k^{T(s-1)} X_{ik}) + \frac{1}{n} \sum_{i=1}^{n} \beta_j X_{ij}^2$$

$$+ \begin{cases} -\lambda + 2\lambda\beta_j & \beta_j < 0 \\ [-\lambda, \lambda] & \beta_j = 0 \\ \lambda + 2\lambda\beta_j & \beta_j > 0 \end{cases}$$

If $\beta_j = 0$,

$$0 \in -\frac{1}{n}\sum_{i=1}^{n} x_{ij}\left(y_i - \alpha^{(s-1)} - \sum_{k\neq j} \beta_k^{T(s-1)} x_{ik}\right) + [-\lambda, \lambda]$$

$$\Rightarrow \left|\frac{1}{n}\sum_{i=1}^{n} x_{ij}\left(y_i - \alpha^{(s-1)} - \sum_{k\neq j} \beta_k^{T(s-1)} x_{ik}\right)\right| \in [-\lambda, \lambda]$$

$$\Rightarrow \left|\sum_{i=1}^{n} x_{ij}\left(y_i - \alpha^{(s-1)} - \sum_{k\neq j} \beta_k^{T(s-1)} x_{ik}\right)\right| \leq n\lambda$$

If $\beta_j > 0$,

$$-\frac{1}{n}\sum_{i=1}^{n} x_{ij}\left(y_i - \alpha^{(s-1)} - \sum_{k\neq j} \beta_k^{T(s-1)} x_{ik}\right) + \frac{\beta_j}{n}\sum_{i=1}^{n} x_{ij}^2 + \lambda + 2\lambda\beta_j = 0$$

$$\Rightarrow \beta_j^{(s)} = \frac{\sum_{i=1}^{n} x_{ij}\left(y_i - \alpha^{(s-1)} - \sum_{k\neq j} \beta_k^{T(s-1)} x_{ik}\right) - \lambda n}{2\lambda n + \sum_{i=1}^{n} x_{ij}^2}$$

If $\beta_j < 0$, (same as above, but λ term is negative)

$$\Rightarrow \beta_j^{(s)} = \frac{\sum_{i=1}^{n} x_{ij}\left(y_i - \alpha^{(s-1)} - \sum_{k\neq j} \beta_k^{T(s-1)} x_{ik}\right) + \lambda n}{2\lambda n + \sum_{i=1}^{n} x_{ij}^2}$$

We now have update equations for the parameters. To run the algorithm, first let $(\alpha^{(0)}, \beta^{(0)}) = \theta^{(0)}$ be some initial guess of θ.

At iteration $s \in \mathbb{N}$ for each $j \in [q]$, compute

$$\theta_j^{(s)} \begin{cases} \operatorname{argmin} g(\theta_j; \theta_{-j}^{(s-1)}) & \text{if } s = j-1 \pmod p \\ \theta_j^{(s-1)} & \text{otherwise} \end{cases}$$

Run the algorithm until $s \geq \bar{s}$ for some $\bar{s}$ of our choosing.

Part 4)

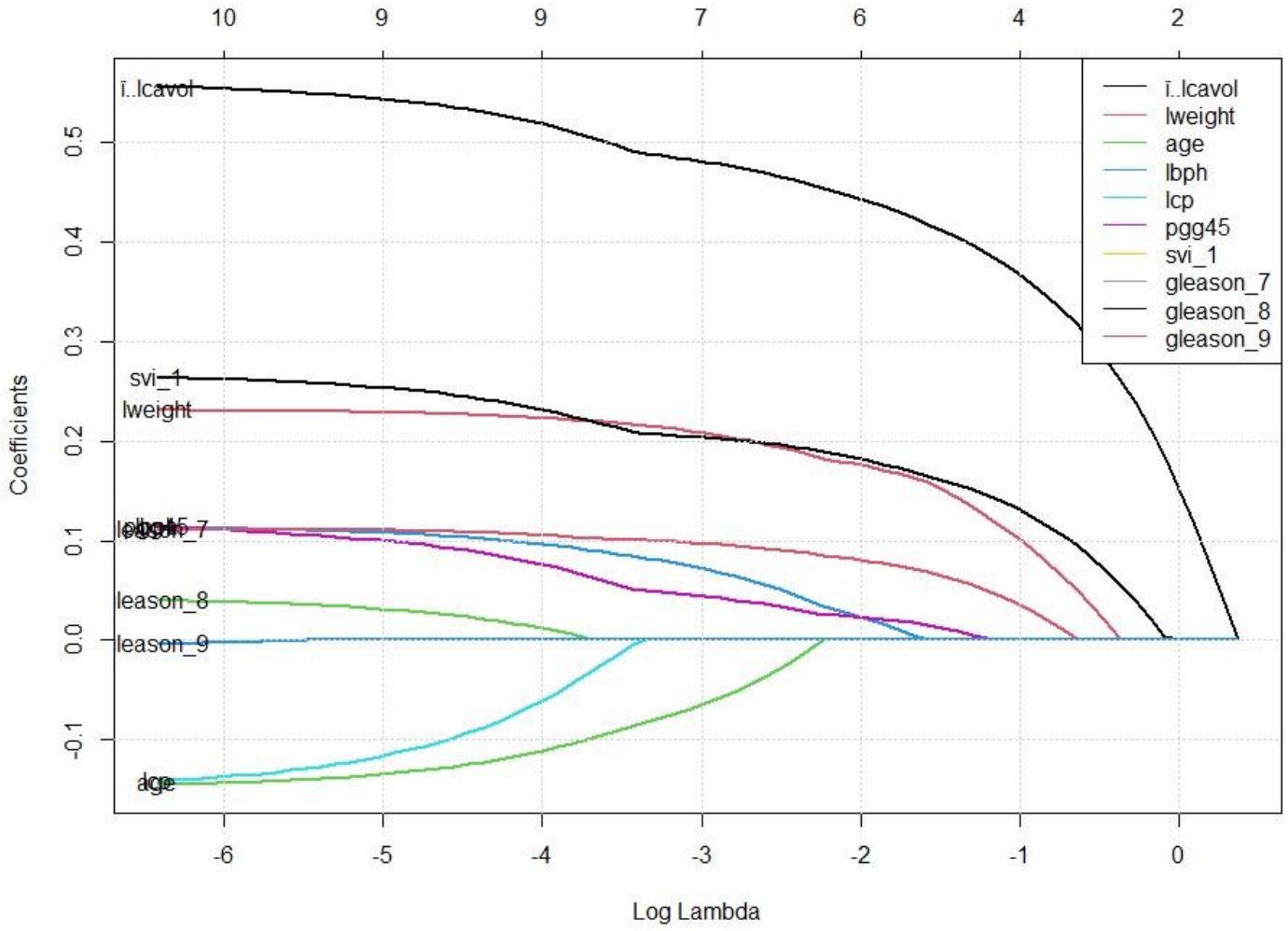
I used the the glmnet package with alpha = 0.5 to implement an elastic net regression.
Below are the values of alpha after each iteration (73 in total), followed by the last few iterations for the beta values, and finally the plot of the beta values against lambda.

Alpha values

```
> model$a0
          s0           s1           s2           s3           s4           s5           s6           s7
3.469447e-17 3.252912e-17 3.040280e-17 2.832340e-17 2.629806e-17 2.431962e-17 2.236181e-17 2.049861e-17
          s8           s9          s10          s11          s12          s13          s14          s15
1.926644e-17 2.401403e-17 2.850238e-17 3.234446e-17 3.446123e-17 3.647752e-17 3.839175e-17 4.020348e-17
         s16          s17          s18          s19          s20          s21          s22          s23
4.191325e-17 4.362913e-17 4.540489e-17 4.705733e-17 4.860486e-17 5.004672e-17 5.114013e-17 5.203343e-17
         s24          s25          s26          s27          s28          s29          s30          s31
5.285281e-17 5.360024e-17 5.428169e-17 5.490275e-17 5.550801e-17 6.116218e-17 6.636534e-17 7.116873e-17
         s32          s33          s34          s35          s36          s37          s38          s39
7.559829e-17 7.967899e-17 8.343482e-17 8.688868e-17 9.006235e-17 9.297647e-17 9.565050e-17 9.810271e-17
         s40          s41          s42          s43          s44          s45          s46          s47
1.003503e-16 1.026911e-16 1.052383e-16 1.075469e-16 1.100584e-16 1.125810e-16 1.148961e-16 1.170196e-16
         s48          s49          s50          s51          s52          s53          s54          s55
1.189663e-16 1.207502e-16 1.223585e-16 1.238549e-16 1.252255e-16 1.264794e-16 1.276261e-16 1.286744e-16
         s56          s57          s58          s59          s60          s61          s62          s63
1.296066e-16 1.304818e-16 1.312834e-16 1.320157e-16 1.326844e-16 1.332948e-16 1.338274e-16 1.343401e-16
         s64          s65          s66          s67          s68          s69          s70          s71
1.348571e-16 1.353519e-16 1.358075e-16 1.362238e-16 1.365731e-16 1.369118e-16 1.372281e-16 1.375189e-16
         s72          s73
1.377849e-16 1.379997e-16
```

Beta values

```
i..lcavol  0.50695458  0.511399895  0.515481443  0.51923500  0.52268422  0.52585153  0.52875063  0.53141633
lweight    0.22067721  0.221692310  0.222624747  0.22347429  0.22424835  0.22495360  0.22559193  0.22617744
age       -0.09969911 -0.103920376 -0.107787355 -0.11133445 -0.11458664 -0.11756687 -0.12025537 -0.12275578
lbph       0.09072291  0.092780960  0.094662567  0.09638799  0.09796941  0.09941812  0.10073579  0.10195030
lcp       -0.03402342 -0.043799581 -0.052763273 -0.06101074 -0.06859413 -0.07556156 -0.08180072 -0.08767526
pgg45      0.06456395  0.068724656  0.072567326  0.07610302  0.07935420  0.08234147  0.08501205  0.08753077
svi_1      0.22094924  0.224874956  0.228470813  0.23177762  0.23481685  0.23760812  0.24009121  0.24244363
gleason_7  0.10280214  0.103763573  0.104626628  0.10541291  0.10612930  0.10678201  0.10736384  0.10790621
gleason_8  0.00230964  0.005742573  0.008888339  0.01177677  0.01442748  0.01685856  0.01905221  0.02109495
gleason_9  .           .            .            .           .           .           .           .

i..lcavol  0.53385996  0.53609868  0.53814872  0.54002515  0.54173057  0.54330118  0.54473771  0.54605085
lweight    0.22671114  0.22719739  0.22764041  0.22804404  0.22840884  0.22874367  0.22904914  0.22932750
age       -0.12504587 -0.12714101 -0.12905707 -0.13080879 -0.13236699 -0.13382966 -0.13516929 -0.13639308
lbph       0.10306183  0.10407848  0.10500805  0.10585773  0.10662372  0.10733321  0.10798190  0.10857434
lcp       -0.09306873 -0.09801220 -0.10254071 -0.10668715 -0.11031412 -0.11378590 -0.11697403 -0.11988997
pgg45      0.08984393  0.09196419  0.09390654  0.09568506  0.09724171  0.09873007  0.10009760  0.10134846
svi_1      0.24460342  0.24658253  0.24839509  0.25005438  0.25148956  0.25287833  0.25415470  0.25532204
gleason_7  0.10840076  0.10885135  0.10926189  0.10963592  0.10996191  0.11027258  0.11055660  0.11081546
gleason_8  0.02296706  0.02468081  0.02624893  0.02768326  0.02895782  0.03015660  0.03125499  0.03225875
gleason_9  .           .           .           .           .           .           .           .

i..lcavol  0.54725082  0.54834709  0.54933210  0.5502466024  0.551152860  0.552016225  0.552813007  0.553541958
lweight    0.22958111  0.22981218  0.23002191  0.2302131068  0.230347165  0.230455571  0.230551561  0.230638390
age       -0.13751060 -0.13853085 -0.13941948 -0.1402662822 -0.141032258 -0.141735707 -0.142377433 -0.142962762
lbph       0.10911526  0.10960904  0.11004787  0.1104587094  0.110849110  0.111215389  0.111551479  0.111858444
lcp       -0.12255532 -0.12499086 -0.12704564 -0.1290662785 -0.130956977 -0.132710605 -0.134314258 -0.135778297
pgg45      0.10249184  0.10353666  0.10442523  0.1052910938  0.106381684  0.107523522  0.108593221  0.109574912
svi_1      0.25638893  0.25736374  0.25817113  0.2589785111  0.259684866  0.260332391  0.260922349  0.261460396
gleason_7  0.11105132  0.11126622  0.11144592  0.1116233688  0.111543352  0.111367567  0.111185916  0.111015674
gleason_8  0.03317562  0.03401294  0.03474013  0.0354354356  0.036045292  0.036596128  0.037096928  0.037553415
gleason_9  .           .           .          -0.0001213841 -0.000588614 -0.001118944 -0.001623727 -0.002088573

i..lcavol  0.55416975  0.554769394  0.555323686  0.555831999  0.556296621  0.556672981
lweight    0.23072673  0.230800344  0.230865734  0.230924570  0.230977866  0.231036427
age       -0.14345645 -0.143936186 -0.144381675 -0.144790007 -0.145162875 -0.145459181
lbph       0.11212218  0.112374591  0.112607896  0.112821637  0.113016853  0.113177279
lcp       -0.13694222 -0.138129232 -0.139247229 -0.140276454 -0.141217745 -0.141897140
pgg45      0.11035508  0.111138291  0.111883331  0.112574118  0.113208066  0.113694218
svi_1      0.26187855  0.262313243  0.262724482  0.263103255  0.263449606  0.263692300
gleason_7  0.11088992  0.110762016  0.110634549  0.110513015  0.110399988  0.110310801
gleason_8  0.03793855  0.038313487  0.038661323  0.038979853  0.039270592  0.039501362
gleason_9 -0.00247402 -0.002844855 -0.003195868 -0.003521757 -0.003821222 -0.004072546
> |
```


10
9
9
7
6
4
2
ï..lcavol
svi_1
lweight
leason_8
leason_9
age
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
Coefficients
-6
-5
-4
-3
-2
-1
0
Log Lambda
ï..lcavol
lweight
age
lbph
lcp
pgg45
svi_1
gleason_7
gleason_8
gleason_9