Stat501_Homework6

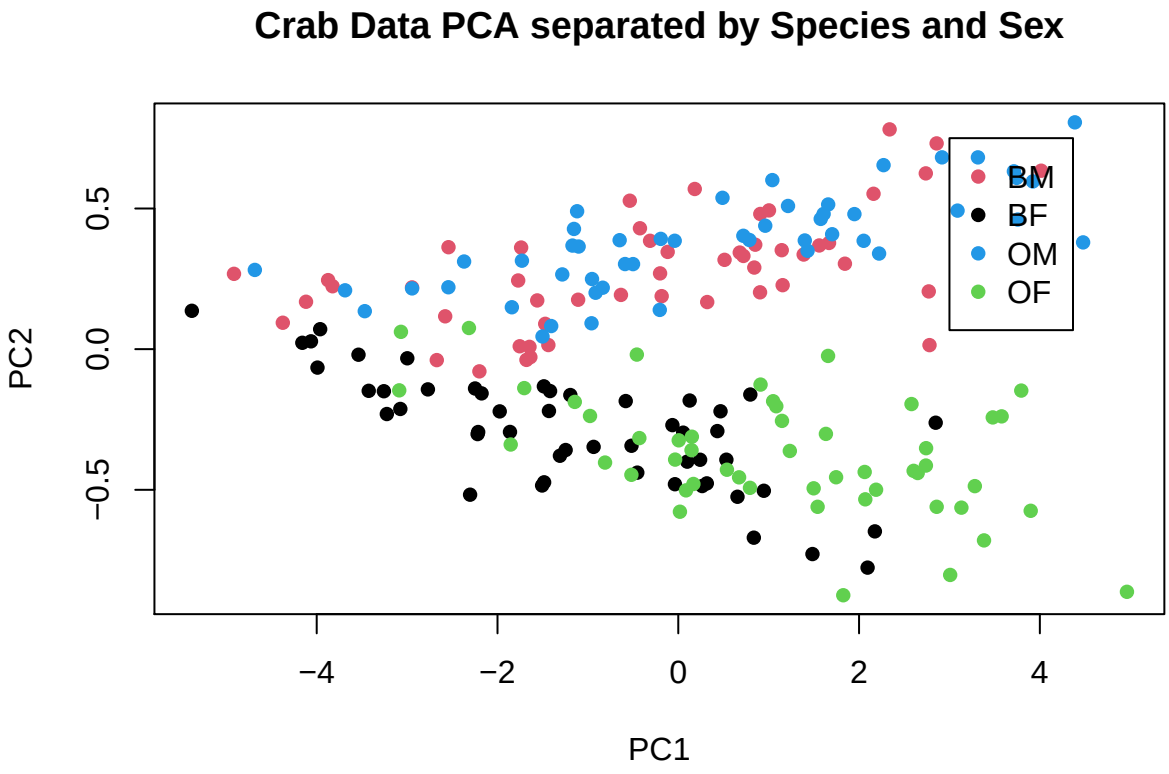
Kelby Kies

4/7/2021

Question 1: Consider the crabs dataset in R used in Exam1.

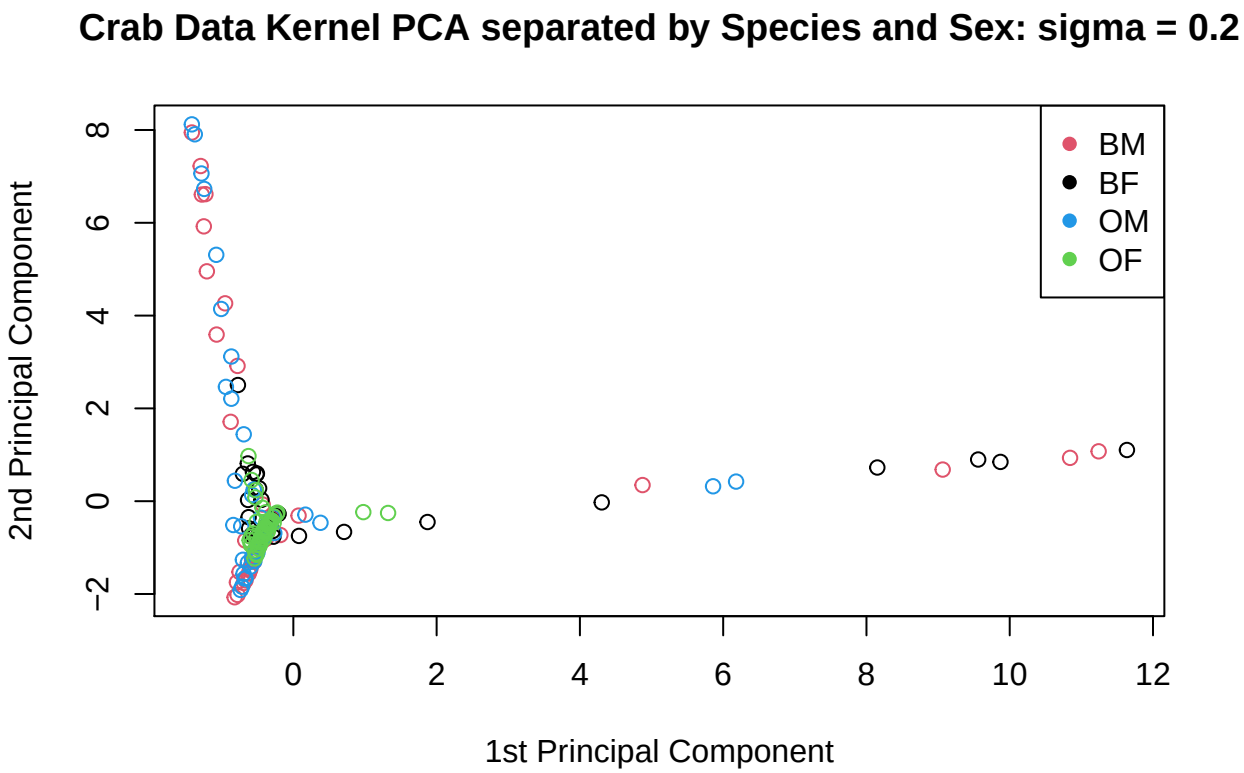
part A.) Use principal components analysis to reduce the dimensionality of the crabs dataset into two dimensions.Display the results. Is there any distinctiveness in the four species/sex combinations? [10 points]

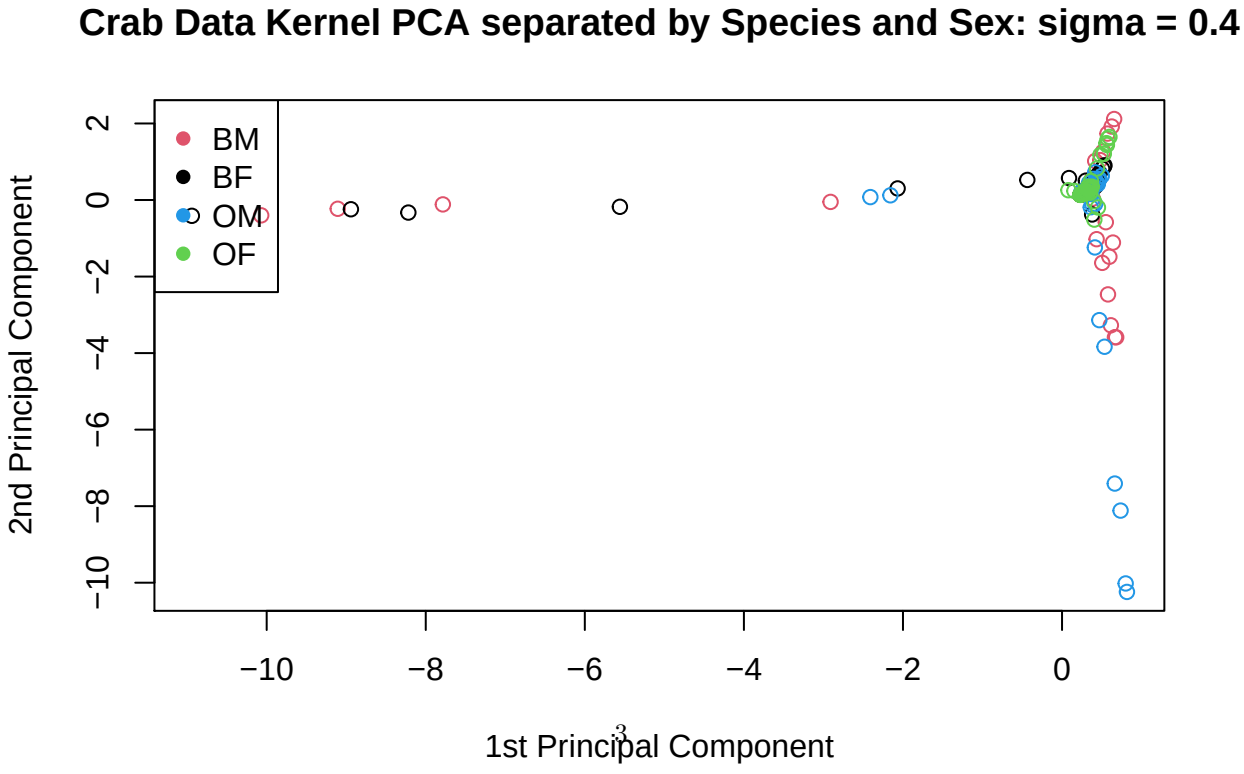
```
## [1] "PCA SUMMARY:"

## Importance of components:
##                             PC1     PC2     PC3     PC4     PC5
## Standard deviation     2.1883 0.38947 0.21595 0.10552 0.04137
## Proportion of Variance 0.9578 0.03034 0.00933 0.00223 0.00034
## Cumulative Proportion  0.9578 0.98810 0.99743 0.99966 1.00000
```

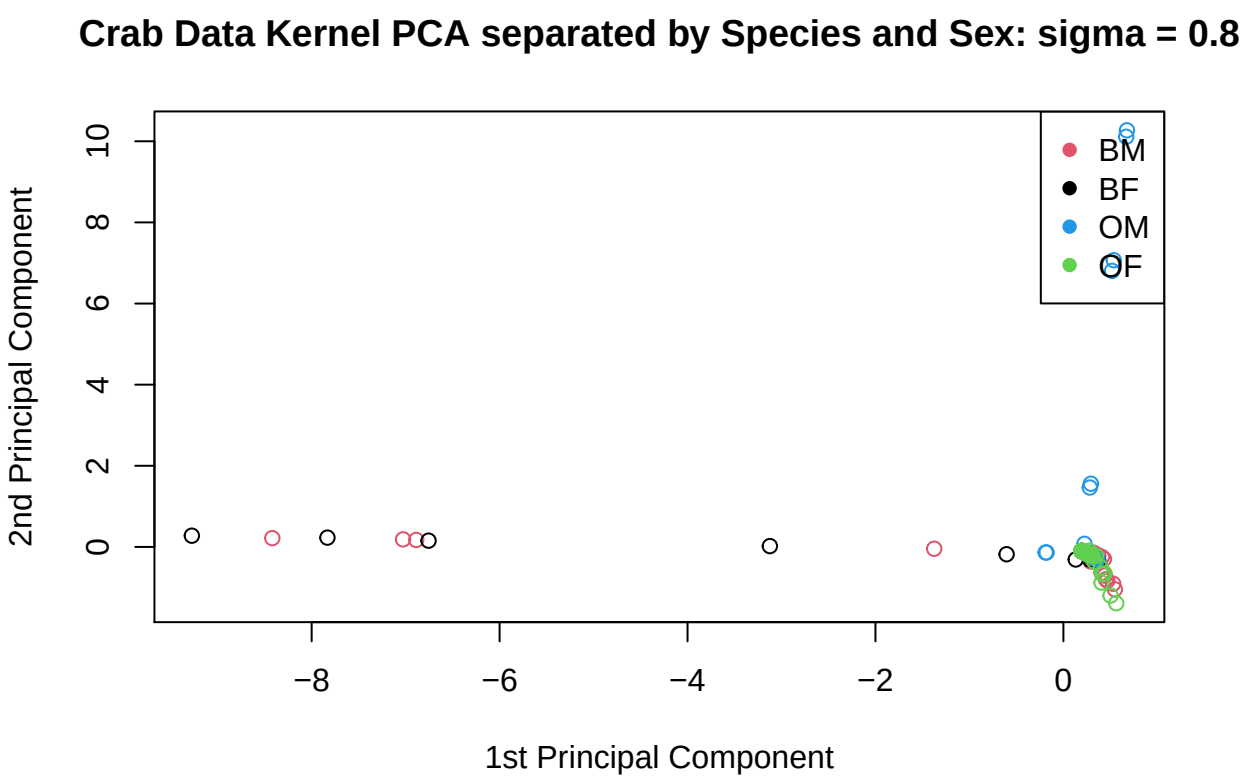


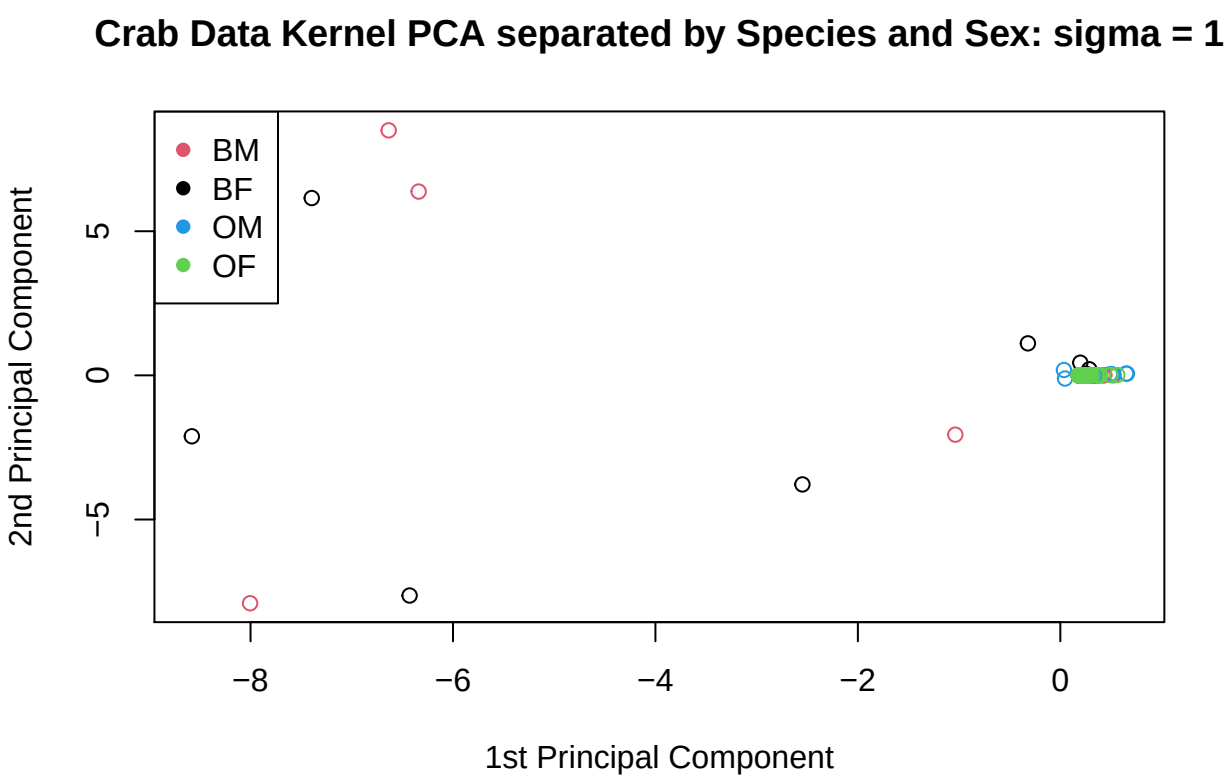
Is there any distinctiveness in the four species/sex combinations? [10 points] PC1 accounts for ~98% of the variance in the data. When we account for PC2 we are looking at 99% of the total variation in the data. The biggest distinction is between Blue/Males, Orange/Males and Orange/Females crabs, thus shows a difference between sex of crab rather than the color of the crab. We can see that these 4 groups are slightly clustered together, but eventually form their own clusters correlated with sex(M/F).

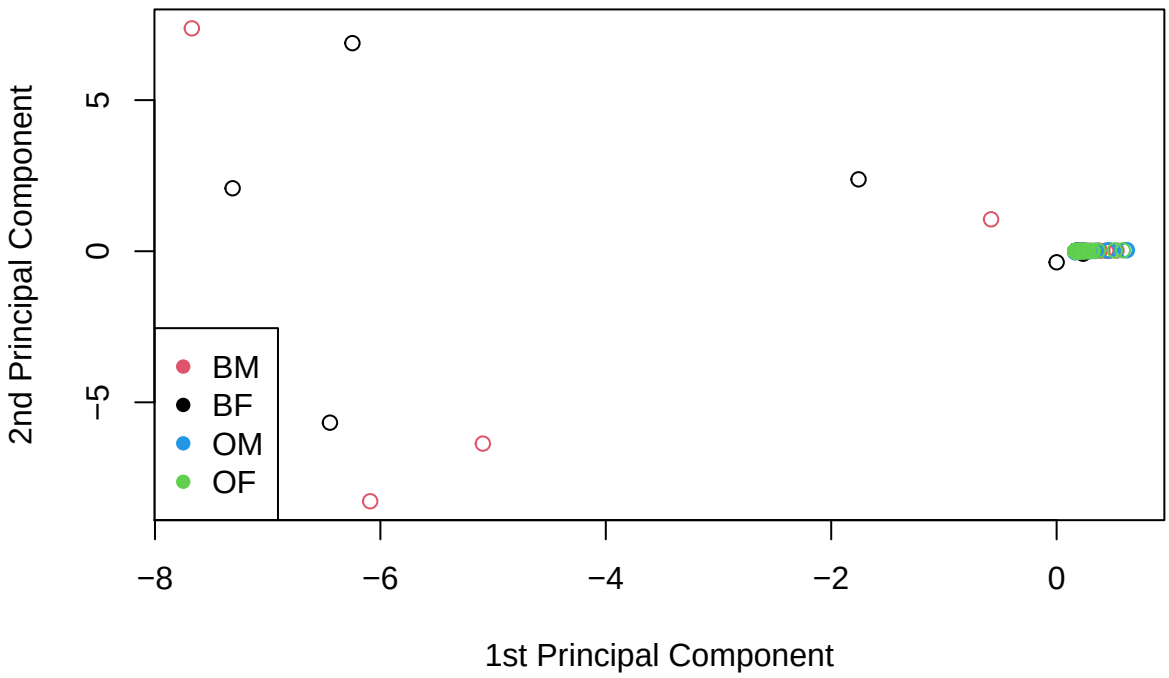
part b.) Perform a kernel principal components analysis with two features and display the results. You may use the Gaussian radial basis funcion. But display the results for different values of $\sigma = 0.2, 0.4, 0.8, 1.0, 1.5, 3$. [20 points]

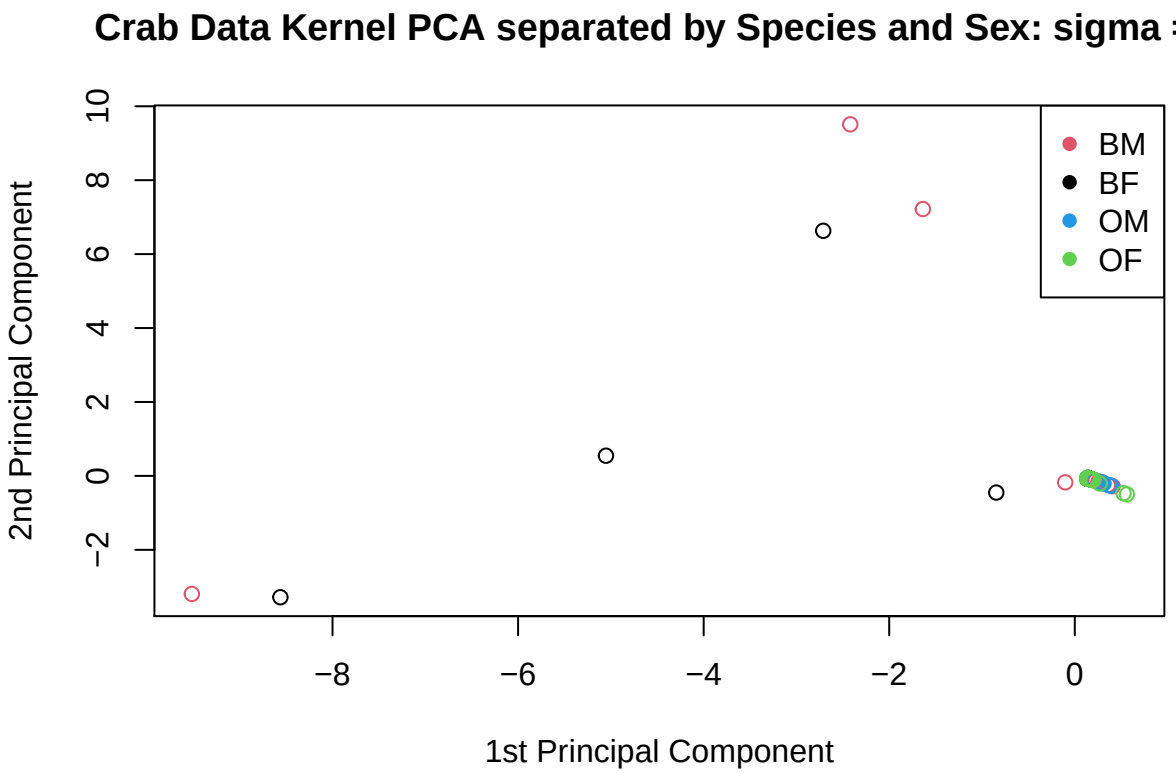





Crab Data Kernel PCA separated by Species and Sex: sigma = 0.8
2nd Principal Component
1st Principal Component
BM
BF
OM
OF
0
2
4
6
8
10
−8
−6
−4
−2
0


Crab Data Kernel PCA separated by Species and Sex: sigma = 1
BM
BF
OM
OF
2nd Principal Component
5
0
−5
−8
−6
−4
−2
0
1st Principal Component


Crab Data Kernel PCA separated by Species and Sex: sigma = 1.5
2nd Principal Component
1st Principal Component
BM
BF
OM
OF
5
0
−5
−8
−6
−4
−2
0



Question 2:

Part A.) We will see whether the digits are distinguishable. To do so, we will first prepare the dataset by rooting out those pixels (coordinates) which do not contribute to categorization. Do so, using appropriate univariate but simultaneous methods, at the 5% level. For the remainder of this question, we will be focused on the 100 most significant coordinates (in terms of the p-value for the above test). [10 points]

```
## [1] "Here are the final pixel(coordinate) variables I will use\n      (Top 100 based on significane
```

```
##   [1] "V170" "V154" "V138" "V11"  "V57"  "V123" "V28"  "V10"  "V122" "V56"
##  [11] "V139" "V41"  "V107" "V12"  "V186" "V73"  "V72"  "V27"  "V163" "V179"
##  [21] "V155" "V29"  "V40"  "V147" "V158" "V142" "V174" "V126" "V189" "V127"
##  [31] "V205" "V196" "V146" "V143" "V162" "V180" "V190" "V195" "V131" "V159"
##  [41] "V130" "V173" "V212" "V164" "V178" "V175" "V111" "V106" "V204" "V213"
##  [51] "V13"  "V110" "V45"  "V114" "V206" "V148" "V115" "V220" "V144" "V30"
##  [61] "V145" "V221" "V128" "V58"  "V157" "V91"  "V191" "V129" "V160" "V171"
##  [71] "V161" "V44"  "V141" "V188" "V197" "V169" "V229" "V46"  "V26"  "V194"
##  [81] "V132" "V211" "V98"  "V88"  "V113" "V185" "V176" "V119" "V125" "V112"
##  [91] "V118" "V95"  "V153" "V14"  "V181" "V222" "V99"  "V207" "V89"  "V102"
```

part Ai.) Evaluate whether the variance-covariance matrices are all equal across all digits.[10 points]

```
source('~/Desktop/stat_501/BoxMTest.R')

BoxMTest(new_zip[,-1], cl = as.factor(new_zip$digits), alpha=0.05)
```

```
## [1] 10
## ------------------------------------------------
##  MBox Chi-sqr. df P
## ------------------------------------------------
##        Inf        Inf       45450       0.0000
## ------------------------------------------------
## Covariance matrices are significantly different.

## $MBox
##   0
## Inf
##
## $ChiSq
##   0
## Inf
##
## $df
## [1] 45450
##
## $pValue
## 0
## 0
```

Using the BoxMTest() function we see that the covariance matrices are significantly different across the digits when comparing at the 5% significance level.

part Aii.) If these are not equal, we will assume that they are the true values of the individual group dispersion matrices. Derive a likelihood ratio test statistic for testing differences in mean effects across different digits. [10 points]

See attached pdf.

part B.) We will now use principal components to reduce dimensionality of the original dataset. Note that the images for the different digits have different means and chanracteristics, therefore, it would be preferred to remove the effect of the digit-specific means before performing the principal components analysis. (Transformed versions of these means need to be added back before proceeding much further.) Use the principal components and determine the number of components needed to explain at least 80% of the total variation in the data, at the 5% level of significance. [20 points]

```
## [1] "Full PCA Summary:"
```

```
## Importance of components:
##                            PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     4.9382 3.45452 3.25458 3.13346 2.91044 2.60558 2.52921
## Proportion of Variance 0.1425 0.06975 0.06191 0.05739 0.04951 0.03968 0.03739
## Cumulative Proportion  0.1425 0.21228 0.27419 0.33157 0.38108 0.42076 0.45815
##                            PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     2.3470 2.28672 2.15368 2.00573 1.92153 1.84086 1.70888
## Proportion of Variance 0.0322 0.03056 0.02711 0.02351 0.02158 0.01981 0.01707
## Cumulative Proportion  0.4904 0.52091 0.54802 0.57153 0.59311 0.61292 0.62999
##                           PC15    PC16    PC17    PC18    PC19    PC20    PC21
## Standard deviation     1.64613 1.59833 1.52917 1.46943 1.37483 1.34733 1.30032
## Proportion of Variance 0.01584 0.01493 0.01367 0.01262 0.01105 0.01061 0.00988
## Cumulative Proportion  0.64583 0.66076 0.67442 0.68704 0.69809 0.70870 0.71858
##                           PC22    PC23    PC24    PC25    PC26    PC27    PC28
## Standard deviation     1.27583 1.25744 1.21775 1.20171 1.17219 1.15231 1.11526
## Proportion of Variance 0.00951 0.00924 0.00867 0.00844 0.00803 0.00776 0.00727
## Cumulative Proportion  0.72810 0.73734 0.74601 0.75445 0.76248 0.77024 0.77751
##                           PC29    PC30    PC31    PC32    PC33   PC34    PC35
## Standard deviation     1.09894 1.07353 1.05248 1.04347 1.01693 0.9699 0.94143
## Proportion of Variance 0.00706 0.00674 0.00647 0.00636 0.00604 0.0055 0.00518
## Cumulative Proportion  0.78457 0.79130 0.79778 0.80414 0.81019 0.8157 0.82086
##                           PC36    PC37   PC38    PC39    PC40    PC41    PC42
## Standard deviation     0.92824 0.91031 0.8965 0.86597 0.84969 0.84447 0.83217
## Proportion of Variance 0.00504 0.00484 0.0047 0.00438 0.00422 0.00417 0.00405
## Cumulative Proportion  0.82590 0.83074 0.8354 0.83982 0.84404 0.84821 0.85226
##                           PC43    PC44    PC45    PC46    PC47    PC48    PC49
## Standard deviation     0.81957 0.80392 0.78615 0.76930 0.75309 0.73921 0.71916
## Proportion of Variance 0.00393 0.00378 0.00361 0.00346 0.00331 0.00319 0.00302
## Cumulative Proportion  0.85618 0.85996 0.86357 0.86703 0.87035 0.87354 0.87656
##                           PC50    PC51    PC52    PC53    PC54    PC55    PC56
## Standard deviation     0.71022 0.70111 0.69620 0.68610 0.67799 0.65946 0.65176
## Proportion of Variance 0.00295 0.00287 0.00283 0.00275 0.00269 0.00254 0.00248
## Cumulative Proportion  0.87951 0.88239 0.88522 0.88797 0.89066 0.89320 0.89568
##                           PC57    PC58    PC59    PC60    PC61    PC62    PC63
## Standard deviation     0.63976 0.63294 0.62577 0.62110 0.60243 0.58350 0.57534
## Proportion of Variance 0.00239 0.00234 0.00229 0.00225 0.00212 0.00199 0.00193
## Cumulative Proportion  0.89807 0.90041 0.90270 0.90496 0.90708 0.90907 0.91100
##                           PC64    PC65   PC66    PC67    PC68    PC69    PC70
## Standard deviation     0.56860 0.56026 0.5554 0.54834 0.53422 0.52947 0.52507
## Proportion of Variance 0.00189 0.00183 0.0018 0.00176 0.00167 0.00164 0.00161
## Cumulative Proportion  0.91289 0.91473 0.9165 0.91829 0.91996 0.92159 0.92321
##                           PC71    PC72    PC73    PC74    PC75    PC76    PC77
## Standard deviation     0.52453 0.51081 0.50751 0.50226 0.49707 0.49068 0.48591
## Proportion of Variance 0.00161 0.00153 0.00151 0.00147 0.00144 0.00141 0.00138
## Cumulative Proportion  0.92481 0.92634 0.92784 0.92932 0.93076 0.93217 0.93355
##                           PC78    PC79    PC80    PC81    PC82    PC83    PC84
## Standard deviation     0.47855 0.47560 0.46732 0.45638 0.45133 0.44189 0.43813
## Proportion of Variance 0.00134 0.00132 0.00128 0.00122 0.00119 0.00114 0.00112
## Cumulative Proportion  0.93489 0.93621 0.93749 0.93870 0.93990 0.94104 0.94216
##                           PC85    PC86    PC87    PC88    PC89    PC90    PC91
## Standard deviation     0.4341 0.42665 0.42616 0.42499 0.41231 0.40875 0.40415
## Proportion of Variance 0.0011 0.00106 0.00106 0.00106 0.00099 0.00098 0.00095
## Cumulative Proportion  0.9433 0.94432 0.94539 0.94644 0.94743 0.94841 0.94937
##                           PC92    PC93    PC94    PC95    PC96    PC97    PC98
```

```
## Standard deviation     0.40307 0.40127 0.39504 0.39109 0.38672 0.38157 0.37837
## Proportion of Variance 0.00095 0.00094 0.00091 0.00089 0.00087 0.00085 0.00084
## Cumulative Proportion  0.95032 0.95126 0.95217 0.95306 0.95394 0.95479 0.95562
##                           PC99  PC100   PC101   PC102   PC103   PC104   PC105
## Standard deviation     0.37577 0.3690 0.36822 0.36674 0.36108 0.36077 0.35987
## Proportion of Variance 0.00083 0.0008 0.00079 0.00079 0.00076 0.00076 0.00076
## Cumulative Proportion  0.95645 0.9573 0.95804 0.95882 0.95959 0.96035 0.96110
##                          PC106   PC107  PC108  PC109   PC110   PC111   PC112
## Standard deviation     0.35226 0.35120 0.3464 0.3457 0.34108 0.33371 0.32934
## Proportion of Variance 0.00073 0.00072 0.0007 0.0007 0.00068 0.00065 0.00063
## Cumulative Proportion  0.96183 0.96255 0.9633 0.9639 0.96463 0.96528 0.96591
##                          PC113   PC114   PC115  PC116   PC117   PC118   PC119
## Standard deviation     0.32744 0.32436 0.32255 0.3196 0.31696 0.31295 0.31015
## Proportion of Variance 0.00063 0.00061 0.00061 0.0006 0.00059 0.00057 0.00056
## Cumulative Proportion  0.96654 0.96716 0.96776 0.9684 0.96895 0.96952 0.97008
##                          PC120   PC121   PC122   PC123   PC124   PC125  PC126
## Standard deviation     0.30977 0.30532 0.30290 0.29909 0.29774 0.29442 0.2937
## Proportion of Variance 0.00056 0.00054 0.00054 0.00052 0.00052 0.00051 0.0005
## Cumulative Proportion  0.97064 0.97119 0.97172 0.97225 0.97277 0.97327 0.9738
##                          PC127   PC128   PC129   PC130   PC131   PC132   PC133
## Standard deviation     0.28936 0.28452 0.28302 0.27986 0.27741 0.27451 0.27268
## Proportion of Variance 0.00049 0.00047 0.00047 0.00046 0.00045 0.00044 0.00043
## Cumulative Proportion  0.97427 0.97474 0.97521 0.97566 0.97611 0.97655 0.97699
##                          PC134   PC135   PC136   PC137  PC138  PC139   PC140
## Standard deviation     0.27120 0.26913 0.26644 0.26354 0.2626 0.2607 0.25807
## Proportion of Variance 0.00043 0.00042 0.00041 0.00041 0.0004 0.0004 0.00039
## Cumulative Proportion  0.97742 0.97784 0.97826 0.97866 0.9791 0.9795 0.97985
##                          PC141   PC142   PC143   PC144   PC145   PC146   PC147
## Standard deviation     0.25704 0.25403 0.25314 0.25118 0.24999 0.24644 0.24539
## Proportion of Variance 0.00039 0.00038 0.00037 0.00037 0.00037 0.00035 0.00035
## Cumulative Proportion  0.98024 0.98062 0.98099 0.98136 0.98173 0.98208 0.98243
##                          PC148   PC149   PC150   PC151   PC152   PC153   PC154
## Standard deviation     0.24261 0.24149 0.24036 0.23886 0.23397 0.23265 0.23147
## Proportion of Variance 0.00034 0.00034 0.00034 0.00033 0.00032 0.00032 0.00031
## Cumulative Proportion  0.98278 0.98312 0.98345 0.98379 0.98411 0.98442 0.98474
##                          PC155  PC156  PC157   PC158   PC159   PC160   PC161
## Standard deviation     0.22908 0.2274 0.2259 0.22360 0.22295 0.22189 0.21940
## Proportion of Variance 0.00031 0.0003 0.0003 0.00029 0.00029 0.00029 0.00028
## Cumulative Proportion  0.98504 0.9853 0.9856 0.98594 0.98623 0.98652 0.98680
##                          PC162   PC163   PC164   PC165   PC166   PC167   PC168
## Standard deviation     0.21833 0.21436 0.21396 0.21338 0.21196 0.21068 0.20943
## Proportion of Variance 0.00028 0.00027 0.00027 0.00027 0.00026 0.00026 0.00026
## Cumulative Proportion  0.98708 0.98734 0.98761 0.98788 0.98814 0.98840 0.98866
##                          PC169   PC170   PC171   PC172   PC173   PC174   PC175
## Standard deviation     0.20668 0.20630 0.20499 0.20232 0.20086 0.19929 0.19864
## Proportion of Variance 0.00025 0.00025 0.00025 0.00024 0.00024 0.00023 0.00023
## Cumulative Proportion  0.98891 0.98915 0.98940 0.98964 0.98987 0.99011 0.99034
##                          PC176   PC177   PC178   PC179   PC180  PC181  PC182
## Standard deviation     0.19451 0.19291 0.19076 0.18906 0.18816 0.1873 0.1853
## Proportion of Variance 0.00022 0.00022 0.00021 0.00021 0.00021 0.0002 0.0002
## Cumulative Proportion  0.99056 0.99078 0.99099 0.99120 0.99140 0.9916 0.9918
##                         PC183   PC184   PC185   PC186   PC187   PC188   PC189
## Standard deviation     0.1837 0.18236 0.18153 0.18080 0.17932 0.17853 0.17668
## Proportion of Variance 0.0002 0.00019 0.00019 0.00019 0.00019 0.00019 0.00018
```

```
## Cumulative Proportion  0.9920 0.99220 0.99239 0.99259 0.99277 0.99296 0.99314
##                           PC190   PC191   PC192   PC193   PC194   PC195   PC196
## Standard deviation     0.17443 0.17308 0.17266 0.17069 0.16845 0.16719 0.16675
## Proportion of Variance 0.00018 0.00018 0.00017 0.00017 0.00017 0.00016 0.00016
## Cumulative Proportion  0.99332 0.99350 0.99367 0.99384 0.99401 0.99417 0.99433
##                           PC197   PC198   PC199   PC200   PC201   PC202   PC203
## Standard deviation     0.16520 0.16410 0.16201 0.16162 0.16010 0.15886 0.15855
## Proportion of Variance 0.00016 0.00016 0.00015 0.00015 0.00015 0.00015 0.00015
## Cumulative Proportion  0.99449 0.99465 0.99480 0.99495 0.99510 0.99525 0.99540
##                           PC204   PC205   PC206   PC207   PC208   PC209   PC210
## Standard deviation     0.15646 0.15586 0.15425 0.15371 0.15165 0.15060 0.14913
## Proportion of Variance 0.00014 0.00014 0.00014 0.00014 0.00013 0.00013 0.00013
## Cumulative Proportion  0.99554 0.99568 0.99582 0.99596 0.99610 0.99623 0.99636
##                           PC211   PC212   PC213   PC214   PC215   PC216   PC217
## Standard deviation     0.14797 0.14650 0.14478 0.14377 0.14164 0.14123 0.14031
## Proportion of Variance 0.00013 0.00013 0.00012 0.00012 0.00012 0.00012 0.00012
## Cumulative Proportion  0.99649 0.99661 0.99673 0.99685 0.99697 0.99709 0.99720
##                           PC218   PC219   PC220   PC221  PC222  PC223  PC224
## Standard deviation     0.13886 0.13684 0.13502 0.13474 0.1335 0.1327 0.1307
## Proportion of Variance 0.00011 0.00011 0.00011 0.00011 0.0001 0.0001 0.0001
## Cumulative Proportion  0.99732 0.99743 0.99753 0.99764 0.9977 0.9979 0.9980
##                          PC225  PC226   PC227   PC228   PC229   PC230   PC231
## Standard deviation     0.1278 0.1276 0.12637 0.12414 0.12327 0.12223 0.12116
## Proportion of Variance 0.0001 0.0001 0.00009 0.00009 0.00009 0.00009 0.00009
## Cumulative Proportion  0.9980 0.9981 0.99823 0.99832 0.99841 0.99850 0.99858
##                           PC232   PC233   PC234   PC235   PC236   PC237   PC238
## Standard deviation     0.12013 0.11845 0.11647 0.11539 0.11244 0.11172 0.11087
## Proportion of Variance 0.00008 0.00008 0.00008 0.00008 0.00007 0.00007 0.00007
## Cumulative Proportion  0.99867 0.99875 0.99883 0.99890 0.99898 0.99905 0.99912
##                           PC239   PC240   PC241   PC242   PC243   PC244   PC245
## Standard deviation     0.10912 0.10798 0.10721 0.10553 0.10501 0.10167 0.09973
## Proportion of Variance 0.00007 0.00007 0.00007 0.00007 0.00006 0.00006 0.00006
## Cumulative Proportion  0.99919 0.99926 0.99933 0.99939 0.99946 0.99952 0.99958
##                           PC246   PC247   PC248   PC249   PC250   PC251   PC252
## Standard deviation     0.09836 0.09311 0.09264 0.09111 0.08637 0.08171 0.08117
## Proportion of Variance 0.00006 0.00005 0.00005 0.00005 0.00004 0.00004 0.00004
## Cumulative Proportion  0.99963 0.99968 0.99973 0.99978 0.99983 0.99986 0.99990
##                           PC253   PC254   PC255   PC256
## Standard deviation     0.07876 0.07367 0.06514 0.02578
## Proportion of Variance 0.00004 0.00003 0.00002 0.00000
## Cumulative Proportion  0.99994 0.99997 1.00000 1.00000

## [1] "PC Variance Proportion at the 5% significance level"

## [1] 1 0
## [1] 2 0
## [1] 3 0
## [1] 4 0
## [1] 5 0
## [1] 6 0
## [1] 7 0
## [1] 8 0
## [1] 9 0
## [1] 10  0
```

```
## [1] 11  0
## [1] 12  0
## [1] 13  0
## [1] 14  0
## [1] 15  0
## [1] 16  0
## [1] 17  0
## [1] 18  0
## [1] 19  0
## [1] 20  0
## [1] 21  0
## [1] 22  0
## [1] 23  0
## [1] 24  0
## [1] 25  0
## [1] 26  0
## [1] 27  0
## [1] 28  0
## [1] 29  0
## [1] 30  0
## [1] 31  0
## [1] 32  1
## [1] 33  1
## [1] 34  1
## [1] 35  1
## [1] 36  1
## [1] 37  1
## [1] 38  1
## [1] 39  1
## [1] 40  1
## [1] 41  1
## [1] 42  1
## [1] 43  1
## [1] 44  1
## [1] 45  1
## [1] 46  1
## [1] 47  1
## [1] 48  1
## [1] 49  1
## [1] 50  1
## [1] 51  1
## [1] 52  1
## [1] 53  1
## [1] 54  1
## [1] 55  1
## [1] 56  1
## [1] 57  1
## [1] 58  1
## [1] 59  1
## [1] 60  1
## [1] 61  1
## [1] 62  1
## [1] 63  1
## [1] 64  1
```

```
## [1] 65  1
## [1] 66  1
## [1] 67  1
## [1] 68  1
## [1] 69  1
## [1] 70  1
## [1] 71  1
## [1] 72  1
## [1] 73  1
## [1] 74  1
## [1] 75  1
## [1] 76  1
## [1] 77  1
## [1] 78  1
## [1] 79  1
## [1] 80  1
## [1] 81  1
## [1] 82  1
## [1] 83  1
## [1] 84  1
## [1] 85  1
## [1] 86  1
## [1] 87  1
## [1] 88  1
## [1] 89  1
## [1] 90  1
## [1] 91  1
## [1] 92  1
## [1] 93  1
## [1] 94  1
## [1] 95  1
## [1] 96  1
## [1] 97  1
## [1] 98  1
## [1] 99  1
## [1] 100   1
## [1] 101   1
## [1] 102   1
## [1] 103   1
## [1] 104   1
## [1] 105   1
## [1] 106   1
## [1] 107   1
## [1] 108   1
## [1] 109   1
## [1] 110   1
## [1] 111   1
## [1] 112   1
## [1] 113   1
## [1] 114   1
## [1] 115   1
## [1] 116   1
## [1] 117   1
## [1] 118   1
```

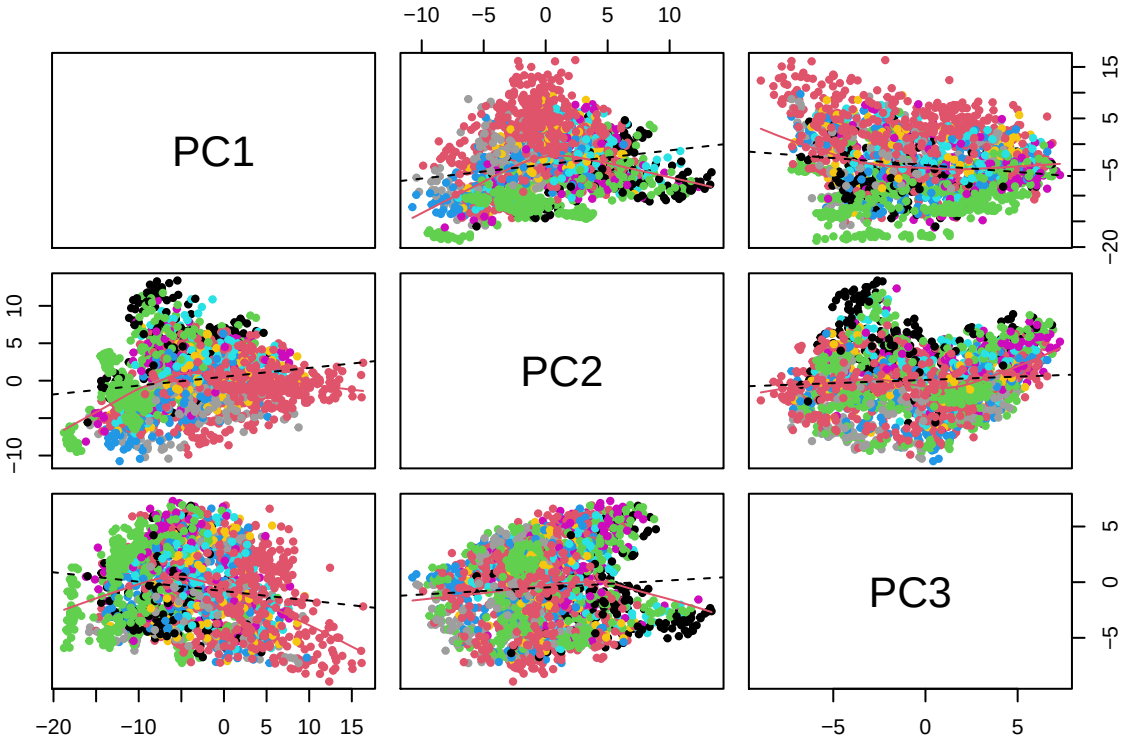
```
## [1] 119   1
## [1] 120   1
## [1] 121   1
## [1] 122   1
## [1] 123   1
## [1] 124   1
## [1] 125   1
## [1] 126   1
## [1] 127   1
## [1] 128   1
## [1] 129   1
## [1] 130   1
## [1] 131   1
## [1] 132   1
## [1] 133   1
## [1] 134   1
## [1] 135   1
## [1] 136   1
## [1] 137   1
## [1] 138   1
## [1] 139   1
## [1] 140   1
## [1] 141   1
## [1] 142   1
## [1] 143   1
## [1] 144   1
## [1] 145   1
## [1] 146   1
## [1] 147   1
## [1] 148   1
## [1] 149   1
## [1] 150   1
## [1] 151   1
## [1] 152   1
## [1] 153   1
## [1] 154   1
## [1] 155   1
## [1] 156   1
## [1] 157   1
## [1] 158   1
## [1] 159   1
## [1] 160   1
## [1] 161   1
## [1] 162   1
## [1] 163   1
## [1] 164   1
## [1] 165   1
## [1] 166   1
## [1] 167   1
## [1] 168   1
## [1] 169   1
## [1] 170   1
## [1] 171   1
## [1] 172   1
```

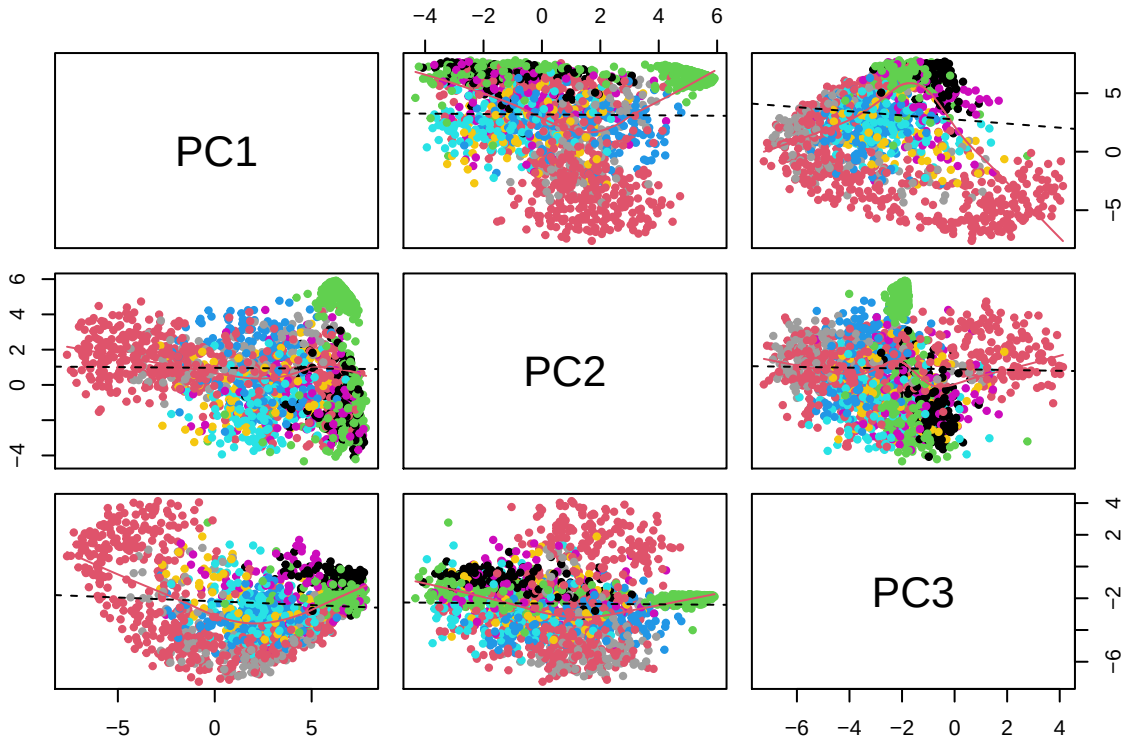
```
## [1] 173   1
## [1] 174   1
## [1] 175   1
## [1] 176   1
## [1] 177   1
## [1] 178   1
## [1] 179   1
## [1] 180   1
## [1] 181   1
## [1] 182   1
## [1] 183   1
## [1] 184   1
## [1] 185   1
## [1] 186   1
## [1] 187   1
## [1] 188   1
## [1] 189   1
## [1] 190   1
## [1] 191   1
## [1] 192   1
## [1] 193   1
## [1] 194   1
## [1] 195   1
## [1] 196   1
## [1] 197   1
## [1] 198   1
## [1] 199   1
## [1] 200   1
## [1] 201   1
## [1] 202   1
## [1] 203   1
## [1] 204   1
## [1] 205   1
## [1] 206   1
## [1] 207   1
## [1] 208   1
## [1] 209   1
## [1] 210   1
## [1] 211   1
## [1] 212   1
## [1] 213   1
## [1] 214   1
## [1] 215   1
## [1] 216   1
## [1] 217   1
## [1] 218   1
## [1] 219   1
## [1] 220   1
## [1] 221   1
## [1] 222   1
## [1] 223   1
## [1] 224   1
## [1] 225   1
## [1] 226   1
```

```
## [1] 227   1
## [1] 228   1
## [1] 229   1
## [1] 230   1
## [1] 231   1
## [1] 232   1
## [1] 233   1
## [1] 234   1
## [1] 235   1
## [1] 236   1
## [1] 237   1
## [1] 238   1
## [1] 239   1
## [1] 240   1
## [1] 241   1
## [1] 242   1
## [1] 243   1
## [1] 244   1
## [1] 245   1
## [1] 246   1
## [1] 247   1
## [1] 248   1
## [1] 249   1
## [1] 250   1
## [1] 251   1
## [1] 252   1
## [1] 253   1
## [1] 254   1
## [1] 255   1
## [1] 256   1
```

I first summarized the PCA I ran on the digit mean centered data. Then I used PCs.proportion.variation.enuff() to test at the 80% significance level. Both tests confirm that 80% of the total variation of the data is found within the first 32 principal components at the 5% signifcance level.

part B i.) Display the components (using color or characters for each digit) using appropriate methods. Compare with the displays obtained using the reduced dataset. [10 points]

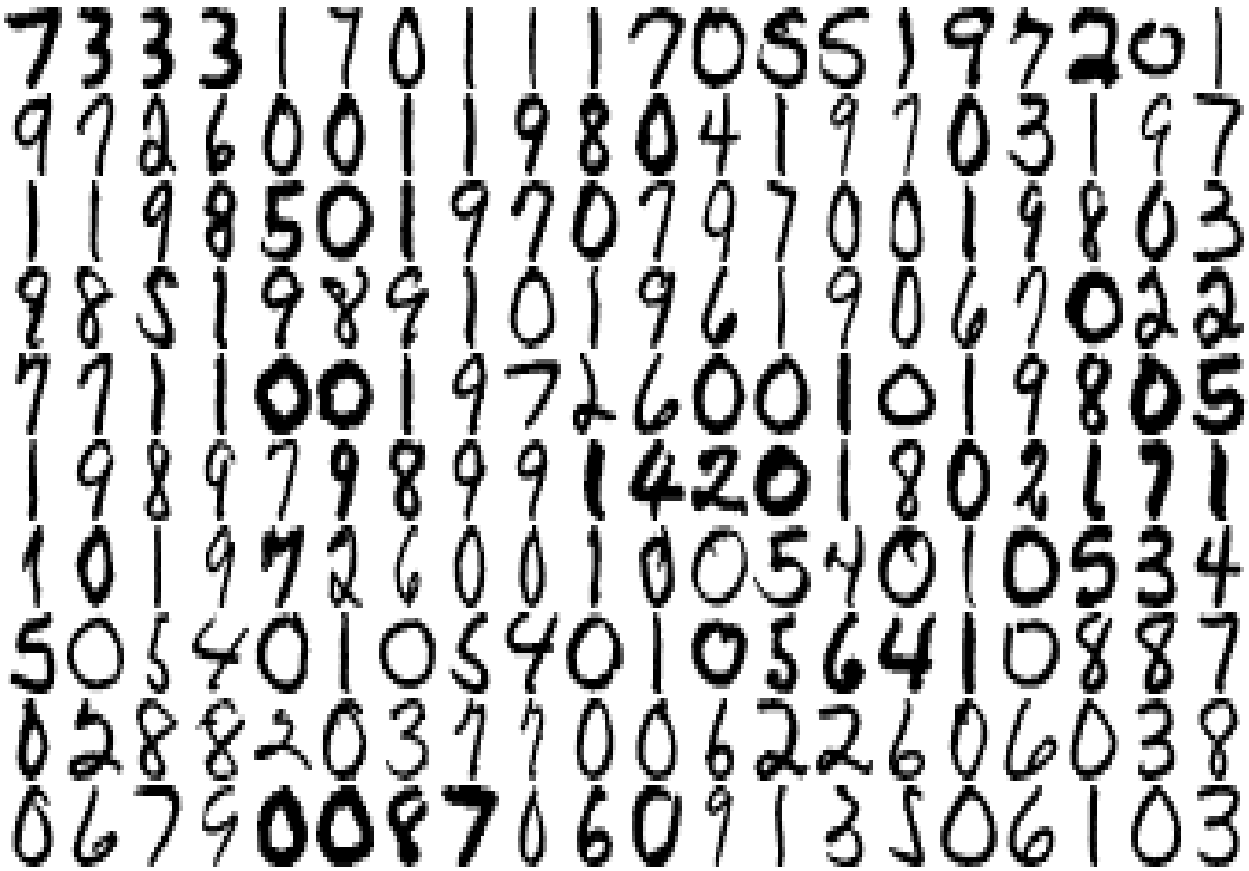


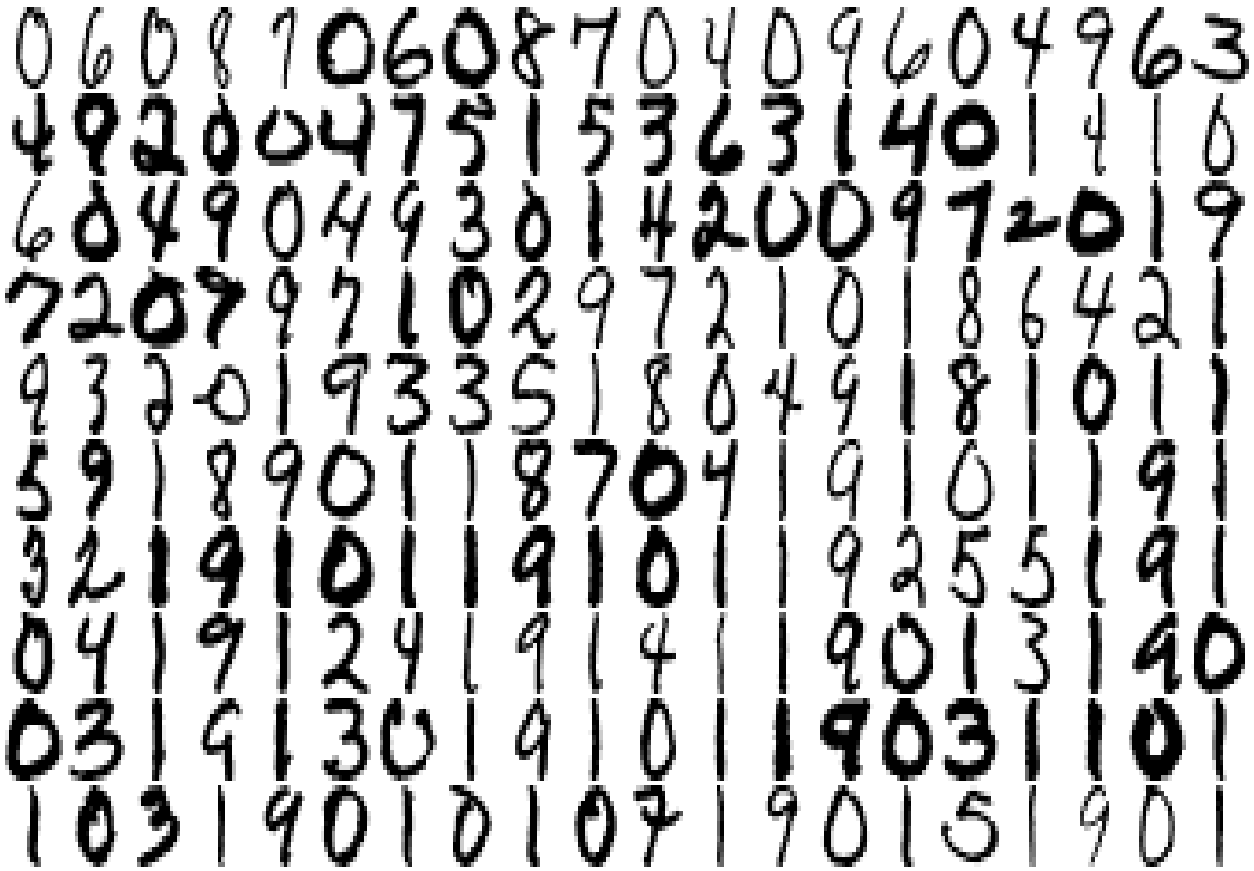


I wasn't sure how to show the first 32 principal components for the full data set or what the best way to compare the full to reduced dataset was. Here I am showing a comparison of the first 3 principal components between the full and reduced data sets. Both plots were created using digit mean centered data, the only difference is the first used all of the data and the second uses the reduced dataset from part.A) I think that the reduced shows better separation between the digits, namely 1, 2 and maybe 0 or 8. This is most likely because we reduced the data down to only include the 100 most significant pixels that contribute to the categorization.

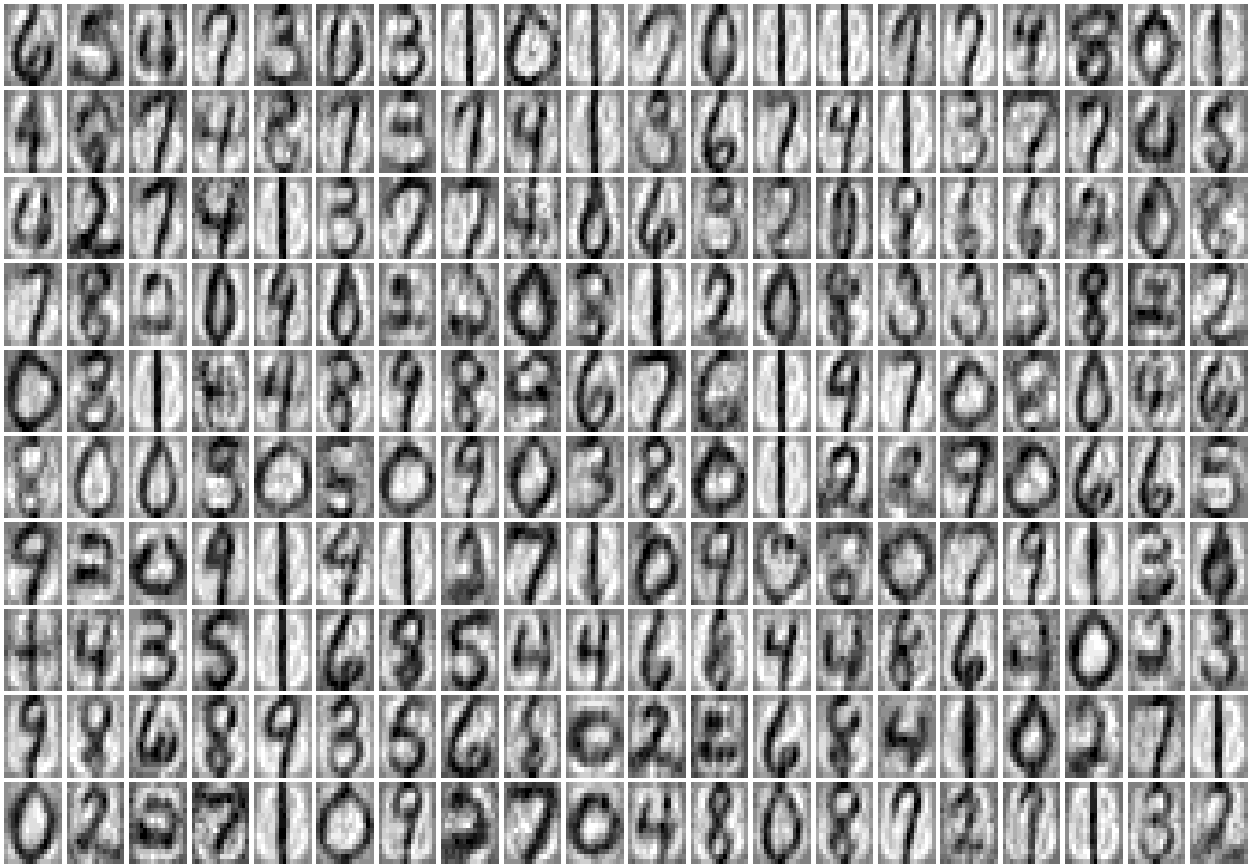
part Bii.) One aspect of using PCA is to obtain a lower-rank representation of the images. Suppose that we have $\Sigma = \Gamma\ \Lambda\ \Gamma$' in terms of the spectral decomposition. Assume that Γ_q contains the first q eigenvectors (columns) of Σ . Then the variance of $\Gamma_q\Gamma_q'X$ has a lower-rank (q) is than Σ and indeed does not contain the smaller eigenvalues (largely random noise, and not a source of variability owing to signal). Using this transformation, display the digits in the raw form (use a 10 × 20 grid) and in terms of the lower-rank form. Comment on the results. [10 points]

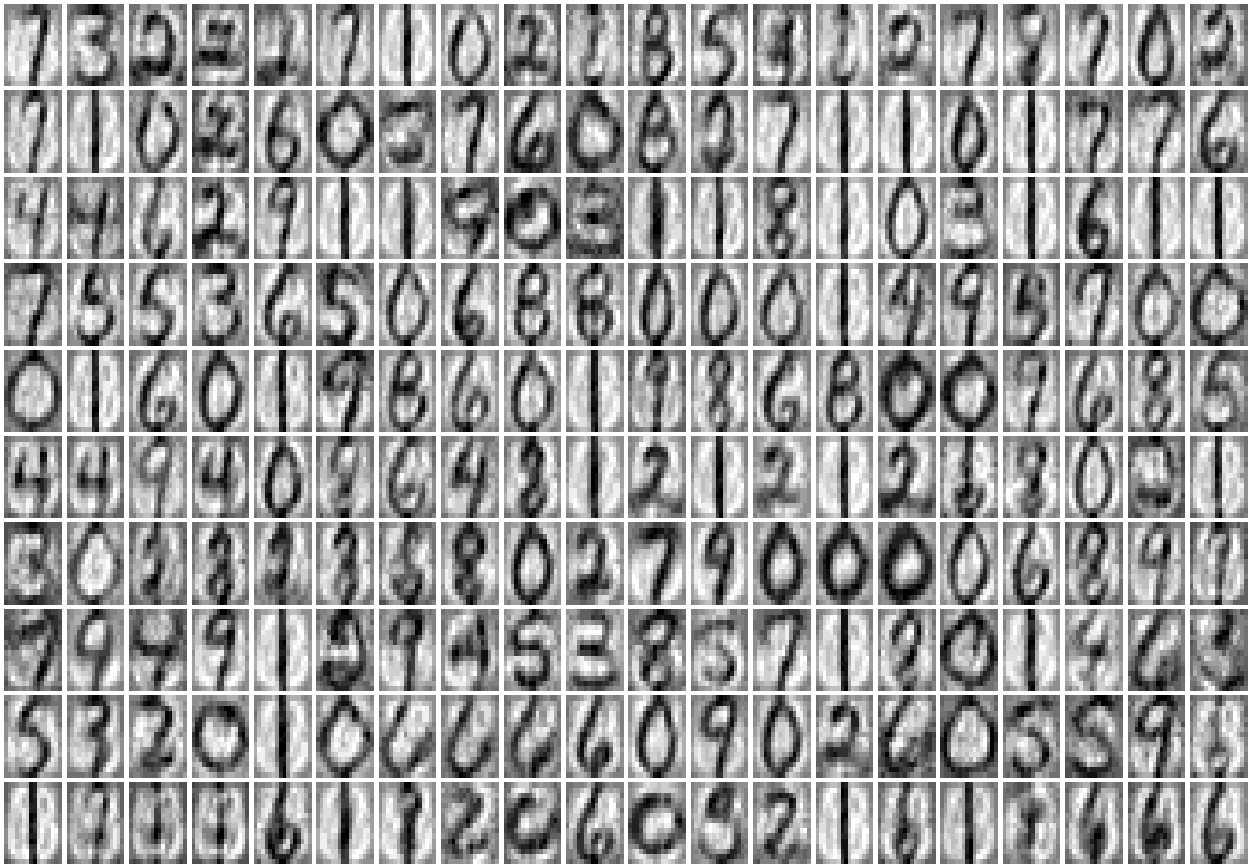
```
zip <- as.matrix(zip)
# Raw Data
par(mfrow = c(10,20), mar = rep(0.1, 4))
for (i in 1:2000)
  image(z = matrix(zip[i,], ncol = 16, by = F)[,16:1], axes = F, col = gray(15:0/15))
```

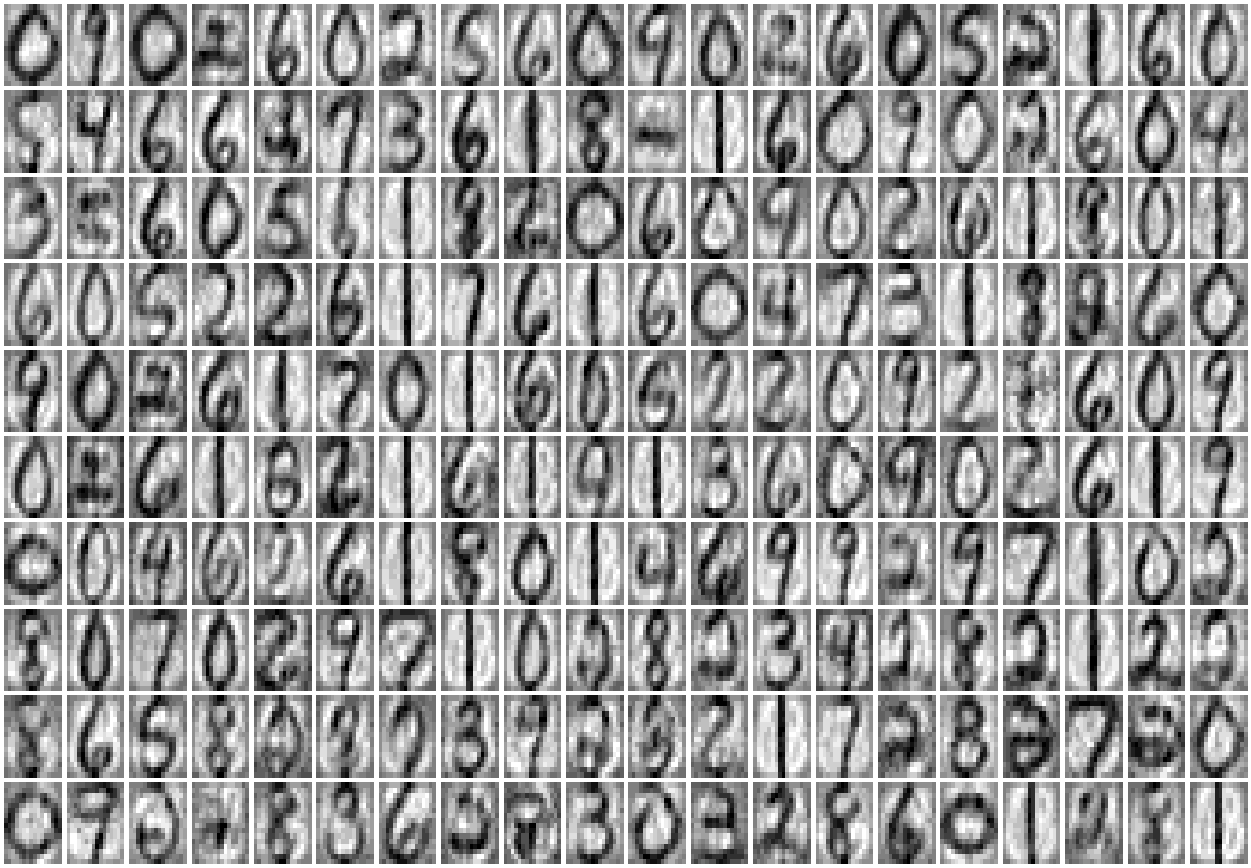


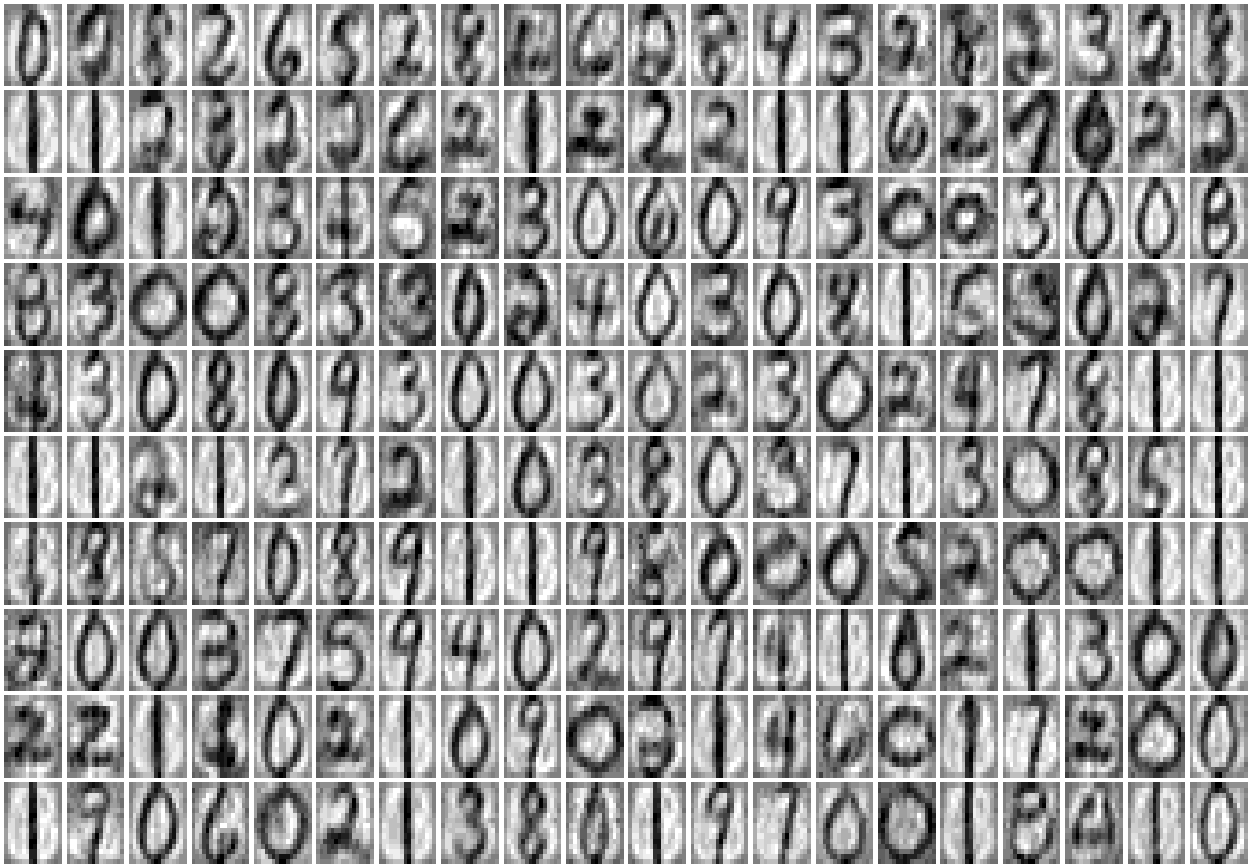


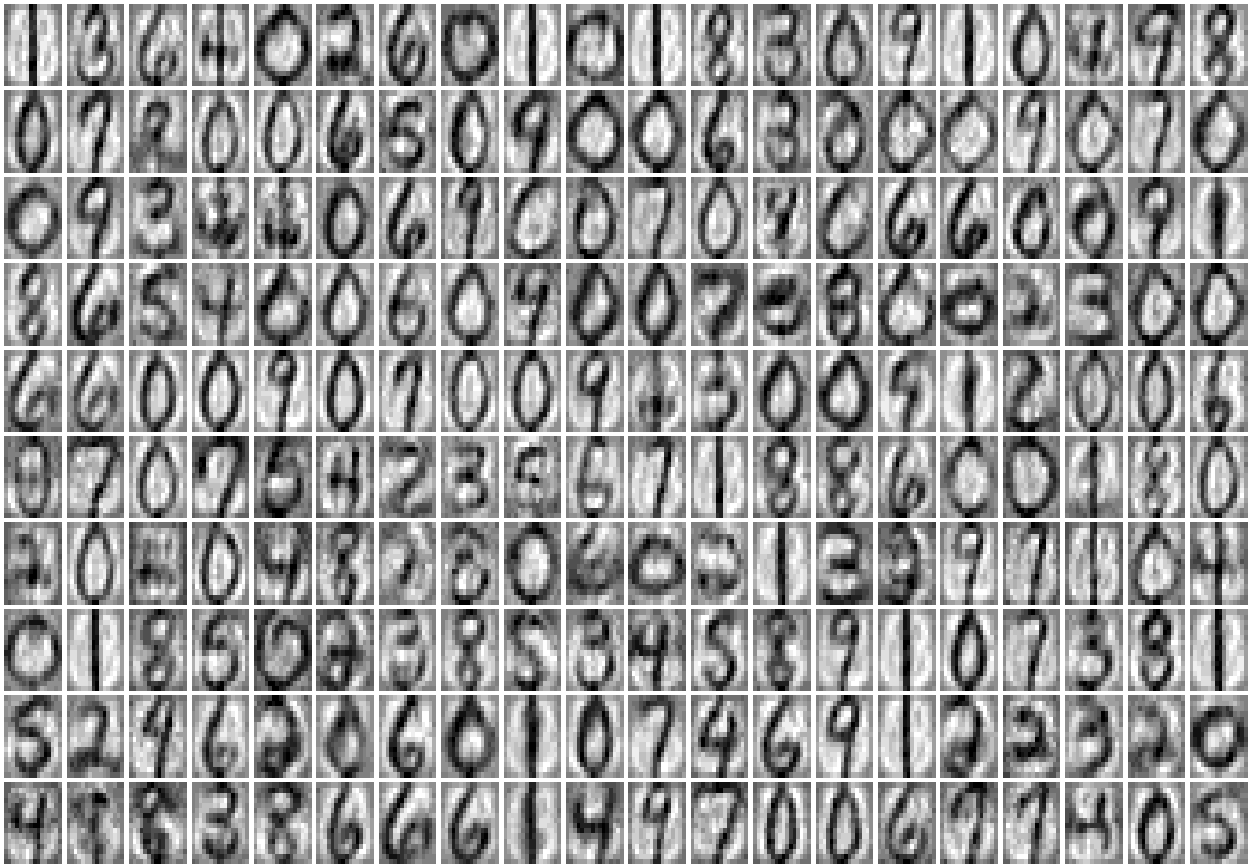
```
# Reduced Data
gamma_q <- pca_zip$rotation[,1:32]
gamma_q_t <- t(gamma_q)
test <- gamma_q %*% gamma_q_t %*% t(zip)
test <- t(test)

par(mfrow = c(10,20), mar = rep(0.1, 4))
for (i in 1:2000)
  image(z = matrix(test[i,], ncol = 16, by = F)[,16:1], axes = F, col = gray(16:0/16))
```

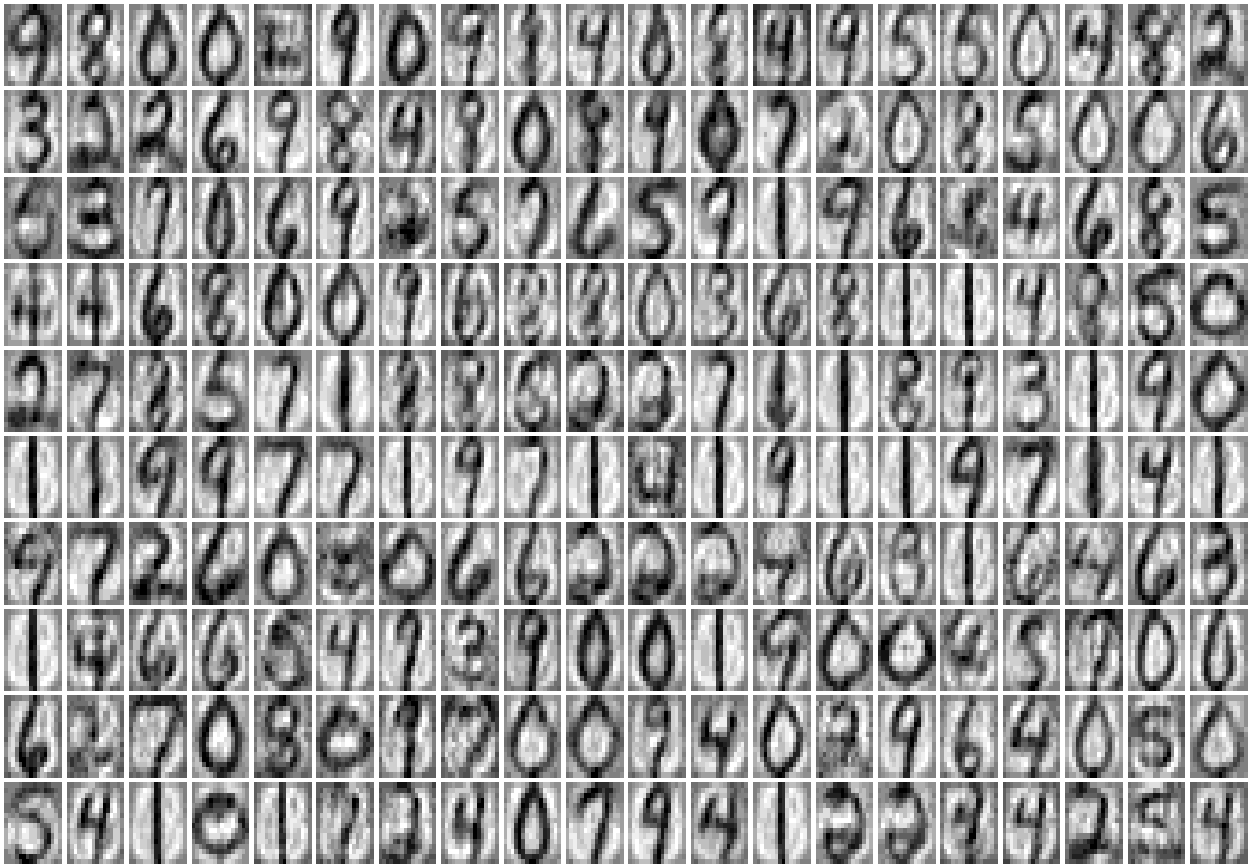


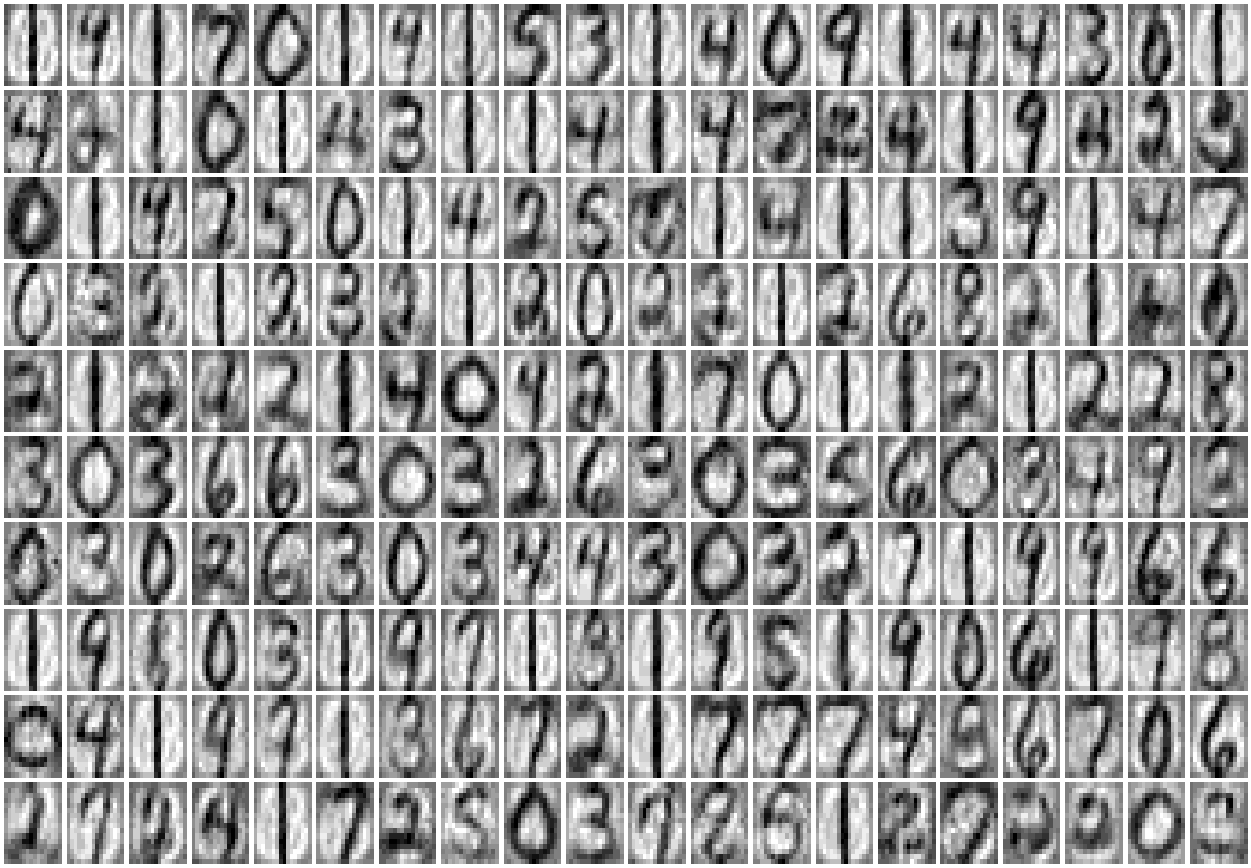


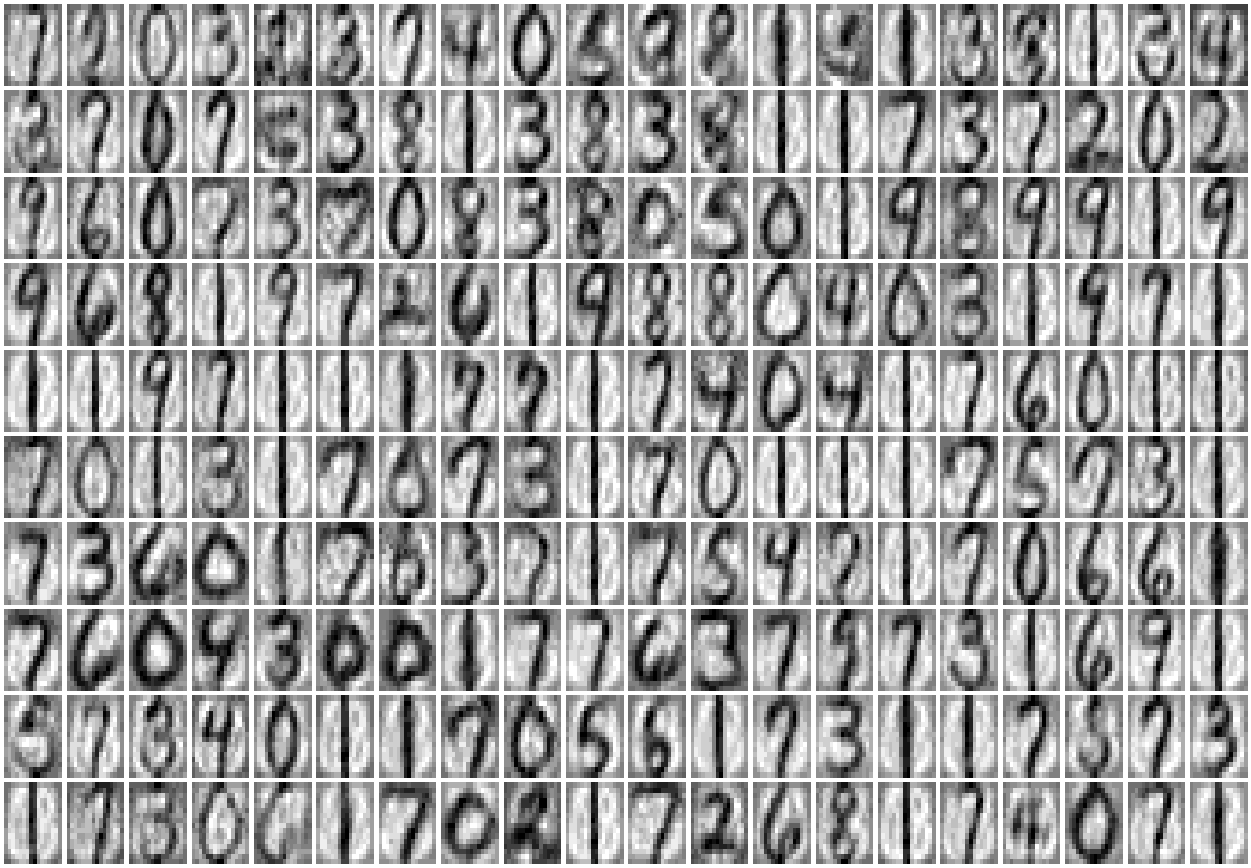


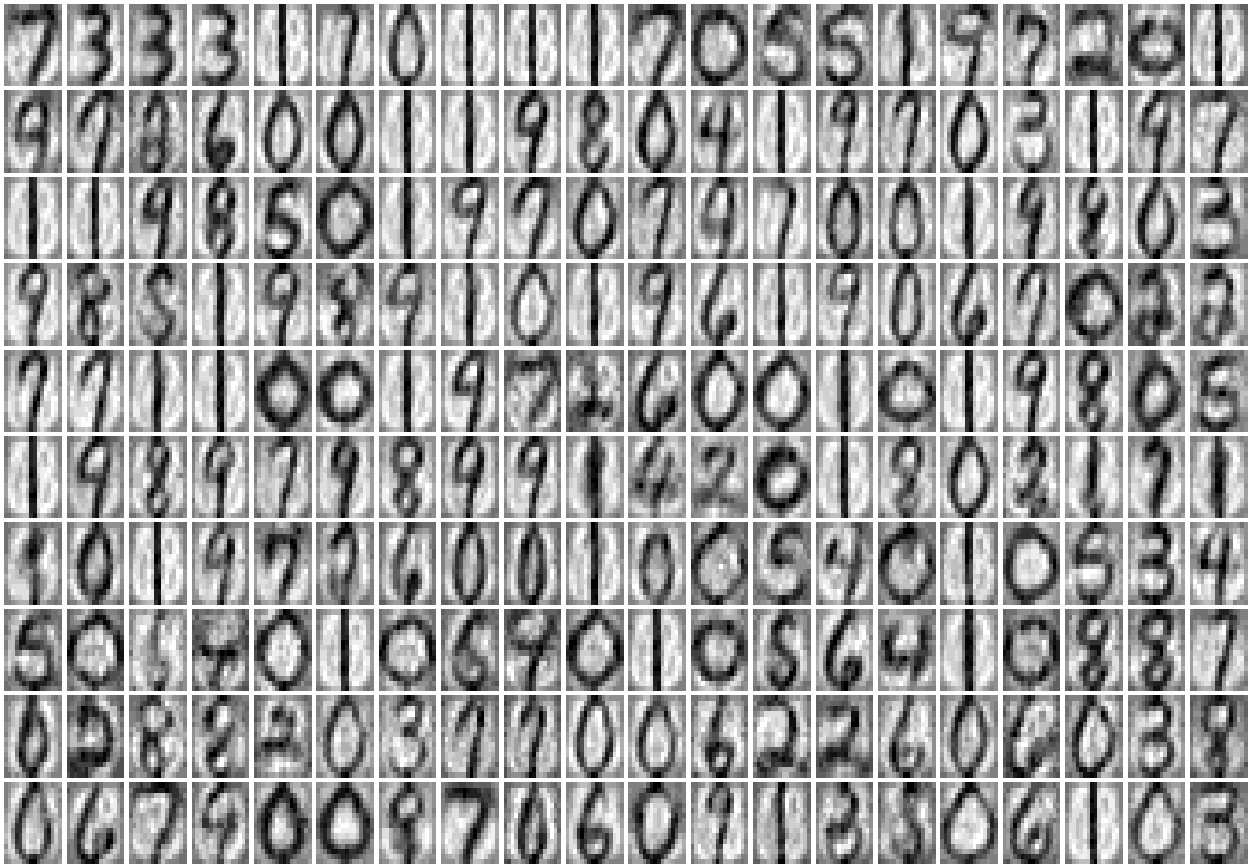


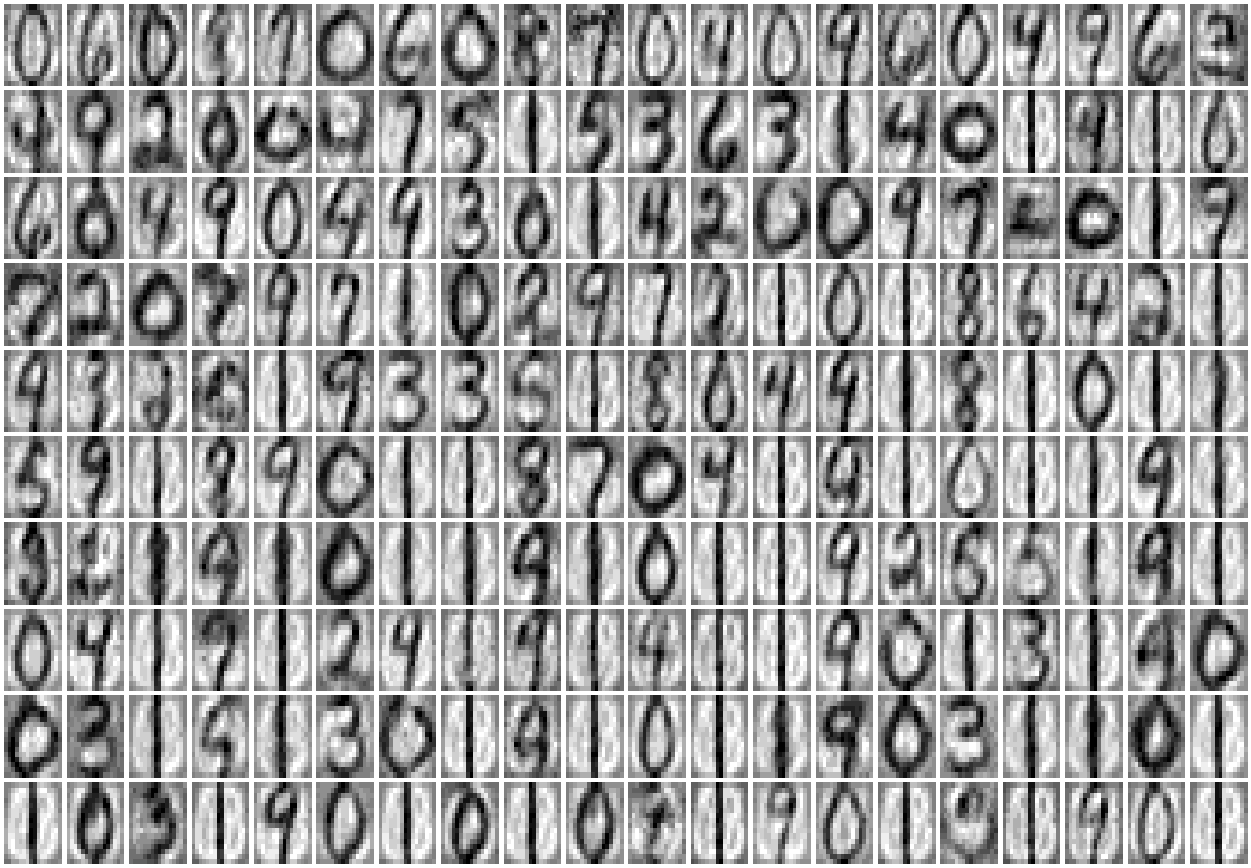












I understand that the reduced data is supposed to present a clearer image because we are using principal components and spectral decomposition to reduce the noise of the data. My reduced data does not look clearer to me. THe backgrounds are slightly different, but I used the same gray scale. Although the digits are still readable.

HW 6: Question 2 A.ii) Derive a LRT statistic for testing differences b/w mean effects across digits.

Assume the var-covariance matrices from part A.i) are the true values because they are not equal.

Definition of LRT Statistic:

$$\Lambda = \frac{\max_{\{\Sigma\}} L(\mu_0, \Sigma)}{\max_{\{\mu_1, \Sigma\}} L(\mu_1, \Sigma)} = \frac{\max_{\{\Sigma\}} \left(L(\mu, \Sigma_0) \cdot L(\mu, \Sigma_1) \cdots L(\mu, \Sigma_9)\right)}{\max_{\{\mu_i, \Sigma\}} L(\mu_0, \Sigma_0) \cdot L(\mu_1, \Sigma_1) \cdots L(\mu_9, \Sigma_9)}$$

where $\underset{1\times 10}{\mu_0} = (\mu, \mu, \mu \cdots \mu)$ ← means are the same

&

$\underset{1\times 10}{\mu_1} = (\mu_0, \mu_1, \mu_2 \cdots \mu_9)$ ← means are different

&

$\Sigma = (\Sigma_0, \Sigma_1, \cdots \Sigma_i, \cdots \Sigma_9)$ where Σ_i is the true var-cov matrix for digit i.

To make this easier I am going to define the numerator & denominator separately, & then combine at the end.

Defining the denominator

$$L(\mu_1, \Sigma) = \prod_{i=1}^{9} L(\mu_i, \Sigma_i)$$

$$\rightarrow = f(X; \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{p/2} \cdot |\Sigma_i|^{1/2}} \exp\left\{\frac{1}{2}(x_i - \mu_i)' \Sigma_i^{-1} (x_i - \mu_i)\right\}$$

$$= \prod_{j=1}^{n_i} (2\pi)^{-np/2} \cdot |\Sigma_i|^{-n/2} \cdot \exp\left\{-\frac{1}{2} \sum_{j=1}^{n_i} (x_{ij} - \mu_i)' \Sigma_i^{-1} (x_{ij} - \mu_i)\right\}$$

$$= (2\pi)^{-np/2} \cdot |\Sigma_i|^{-n/2} \cdot \exp\left\{-\frac{1}{2} \sum_{j=1}^{n_i} \operatorname{tr} (x_{ij} - \mu_i)' \Sigma_i^{-1} (x_{ij} - \mu_i)\right\}$$

$$= 2\pi^{-np/2} \cdot |\Sigma_i|^{-n/2} \cdot \exp\left\{-\frac{1}{2} \operatorname{tr} \Sigma_i^{-1} \sum_{j=1}^{n_i} (x_{ij} - \mu_i)(x_{ij} - \mu_i)'\right\}$$

$$= 2\pi^{-n_i p} \cdot |\Sigma_i|^{-n_i/2} \exp\left\{-\frac{1}{2} \operatorname{tr} \Sigma_i^{-1} \cdot \Sigma_i n_i\right\}$$

$$= 2\pi^{-\frac{n_i p}{2}} \cdot |\Sigma_i|^{n_i/2} \cdot \exp\left\{-\frac{n_i p}{2}\right\}$$ Plug back in

$$= (2\pi)^{\frac{n_0 p}{2}} \cdot |\Sigma_0|^{-n_0/2} \cdot \exp\left\{-\frac{n_0 p}{2}\right\} \cdot (2\pi)^{\frac{n_1 p}{2}} \cdot |\Sigma_1|^{n_1/2} \cdot \exp\left\{-\frac{n_1 p}{2}\right\} \cdots 2\pi^{\left(\frac{n_9 p}{2}\right)} |\Sigma_9|^{-\frac{n_9}{2}}$$

$$= 2\pi^{-\frac{(n_0 + n_1 \cdots n_9) \cdot p}{2}} \cdot \left(|\Sigma_0|^{-n_0/2} \cdots |\Sigma_1|^{(-n_1/2)}\right) \cdot \exp\left\{-\frac{(n_0 + n_1 + \cdots n_9) \cdot p}{2}\right\} \exp\left\{-\frac{n_9 \cdot p}{2}\right.$$

Defining the numerator

$$L(\mu_0, \Sigma) = \prod_{i=1}^{q} L(\mu, \Sigma_i) = \prod_{i=1}^{q} f(X, \mu, \Sigma_i)$$

$$= \frac{1}{(2\pi)^{p/2} \cdot |\Sigma_i|^{1/2}} \cdot \exp\left\{-\frac{1}{2}(x_i - \mu)' \cdot \Sigma_i^{-1} (x_i - \mu)\right\}$$

$$= \prod_{j=1}^{n_i}$$

$$= 2\pi^{-\frac{n_i \cdot p}{2}} \cdot |\Sigma_i|^{-n_i/2} \cdot \exp\left\{-\frac{1}{2} \sum_{j=1}^{n_i} (x_{ij} - \mu) \cdot \Sigma_i^{-1} (x_{ij} - \mu)\right\}$$

this can't reduce this time.

← Plug back in.

$$= (2\pi^{-\frac{n_0 p}{2}} |\Sigma_0|^{-n_0/2} \cdot \exp\left\{-\frac{1}{2} \sum_{j=1}^{n_0} (x_{0j} - \mu) \cdot \Sigma_0^{-1} (x_{0j} - \mu)\right\} \cdot$$

$$\cdot (2\pi^{-\frac{n_1 p}{2}} |\Sigma_1|^{-n_1/2} \exp\left\{-\frac{1}{2} \sum_{j=1}^{n_1} (x_{1j} - \mu) \cdot \Sigma_1^{-1} (x_{1j} - \mu)\right\} \cdots$$

$$\cdots (2\pi^{-\frac{n_q p}{2}} \cdot |\Sigma_q|^{-n_q/2} \cdot \exp\left\{-\frac{1}{2} \sum_{j=1}^{n_q} (x_{qj} - \mu) \cdot \Sigma_q^{-1} (x_{qj} - \mu)\right\})$$

The final LRT statistic is the ratio of the two maximized likelihood functions. Here, I am denoting the numerator MLE as $\hat{\ }$ & denom as $\tilde{\ }$

$$\Lambda = \frac{2\pi^{-\frac{(n_0 + n_1 + \cdots n_q)p}{2}} (|\hat{\Sigma}_0|^{-n_0/2} \cdot |\hat{\Sigma}_1|^{-n_1/2} \cdots |\hat{\Sigma}_q|^{-n_q/2}) \cdot \exp\left\{-\frac{1}{2} \sum_{j=1}^{n_0} x_{0j} - \hat{\mu}) \cdot \hat{\Sigma}_0^{-1} (x_{0j} - \hat{\mu}\right.}{\cdot \exp\left\{-\frac{1}{2} \sum_{j=1}^{n_1} (x_{1j} - \hat{\mu}) ' \hat{\Sigma}_1^{-1} (x_1 - \hat{\mu})\right\} \cdots \exp\left\{-\frac{1}{2} \sum_{j=1}^{n_q} (x_{qj} - \hat{\mu})' \cdot \hat{\Sigma}_q^{-1} (x_q - \hat{\mu}_2)\right\}}$$

$$2\pi^{-\frac{(n_0 + n_1 + \cdots n_q)p}{2}} \cdot (|\tilde{\Sigma}_0|^{-n_0/2} \cdot |\tilde{\Sigma}_1|^{-n_1/2} \cdots |\tilde{\Sigma}_q|^{-n_q/2}) \cdot \exp\left\{-\frac{(n_0 + n_1 + \cdots n_q)p}{2}\right\}$$