2023 SISBID Dimension Reduction Lab

Genevera I. Allen & Yufeng Liu

PCA LAB Using Digits Data

Data set - Digits Data. Either use all digits or choose 2-3 digits if computational speed is a problem. Looking at 3's, 8's and 5's are interesting.

Problem 1 - PCA

Problem 1a - Apply PCA to this data.

Problem 1b - Do the first several PCs well separate different digits? Why or why not?

Problem 1c - Use the first several PCs and PC loadings to evaluate the major patterns in the digits data. Can you come up with a description of the pattern found by each of the first five PCs?

Problem 1d - How many PCs are needed to explain 95% of the variance? You must decide how many PCs to retain. Which do you pick and why?

Problem 2 - MDS

Problem 2a - Apply MDS (classical or non-metric) to this data. Try out several distance metrics and different numbers of MDS components.

Problem 2b - Which distance metric is best for this data? Which one reveals the most separation of the digits?

Problem 2c - Compare and contrast the MDS component maps to the dimension reduction of PCA. Which is preferable?

Problem 3 - ICA.

Problem 3a - Apply ICA to this data set.

Problem 3b - Which value of K did you use? Why? What happens when you slightly change your chosen K?

Problem 3c - Interpret the independent image signals found. Do any other them accurately reflect the different digits? Which ones?

Problem 4 - NMF

Problem 4a - Apply NMF to this data set.

Problem 4b - Which value of K did you use? Why? What happens when you slightly change your chosen K?

Problem 4c - Interpret the NMF basis factors.

Problem 5 - UMAP

Problem 5a - Apply UMAP on this data set.

Problem 5b - Try using different distances and numbers of neighbors. How does this change your results?

Problem 5c - Interpret the UMAP projections. Does this identify meaningful groups of digits?

Problem 6 - tSNE

Problem 6a - Apply tSNE on this data set

Problem 6b - Try changing the perplexity. How does this change your results?

Problem 6c - Interpret the tSNE projections. Does this identify meaningful groups of digits?

Problem 7 - Comparisons.

Problem 7a - Compare and contrast PCA, MDS, NMF ICA, TSNE, and UMAP on this data set. Which one best separates the different digits? Which one reveals the most interesting patterns?

Problem 7b - Overall, which method do you recommend for this data set and why?

Additional Data set - NCI Microarray data

(If you have time - take a further look at this data set using various methods for dimension reduction. Also you may be interested in trying MDS to visualize this data.)

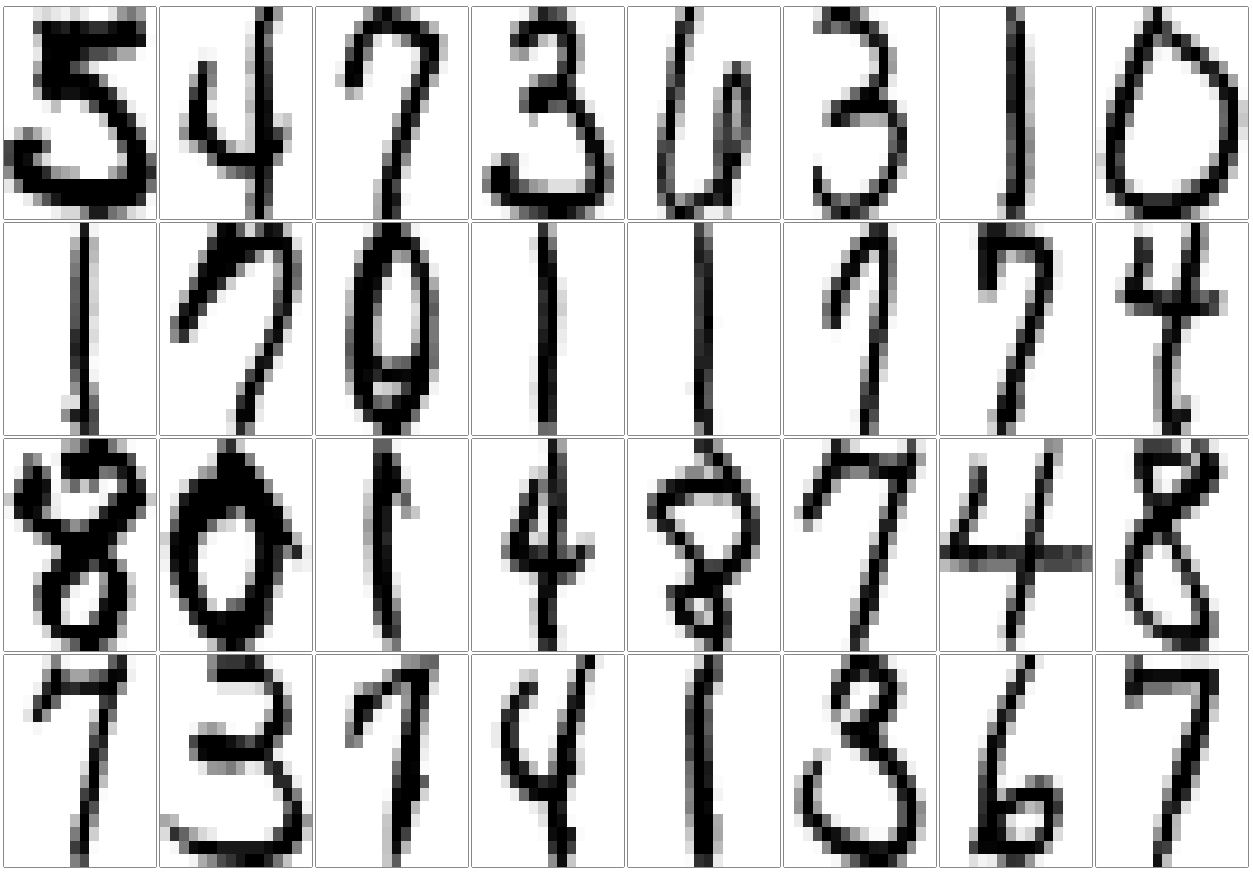
R scripts to help out with the Dimension Reduction Lab

Don't peek at this if you want to practice coding on your own!!

```
library(ISLR)
library(ggplot2)
library(tidyr)
```

Load in data and visualize

```
#code for digits - ALL
rm(list=ls())
load("UnsupL_SISBID_2023.Rdata")

#visulaize
#pdf("temp.pdf")
par(mfrow=c(4,8), mar=c(.1,.1,.1,.1))
for(i in 1:32){
  imagedigit(digits[i,])
}
```



```
#dev.off()
```

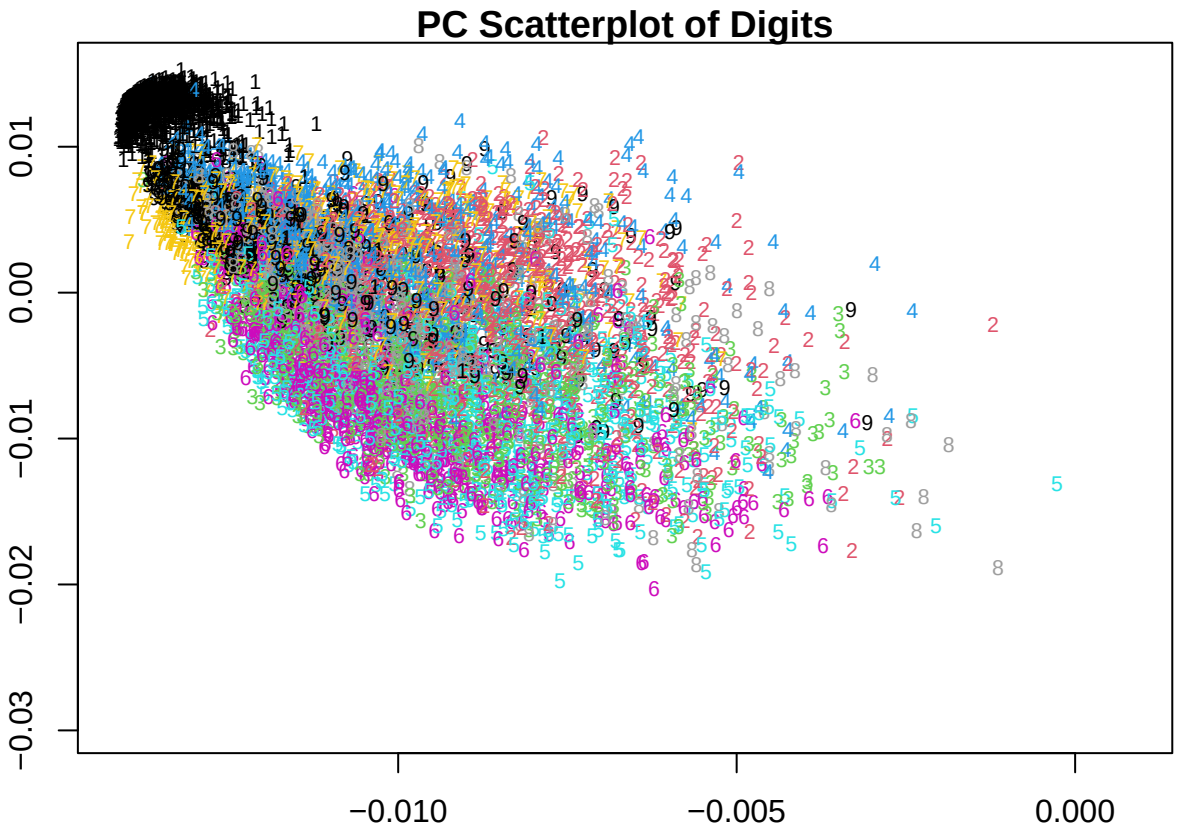
##Problem 1 - PCA

PCA - take SVD to get solution don't center and scale to retain interpretation as images

```
########Problem 1 - PCA
#PCA - take SVD to get solution
#don't center and scale to retain interpretation as images
svdd = svd(digits)
U = svdd$u
V = svdd$v #PC loadings
D = svdd$d
Z = digits%*%V #PCs
```

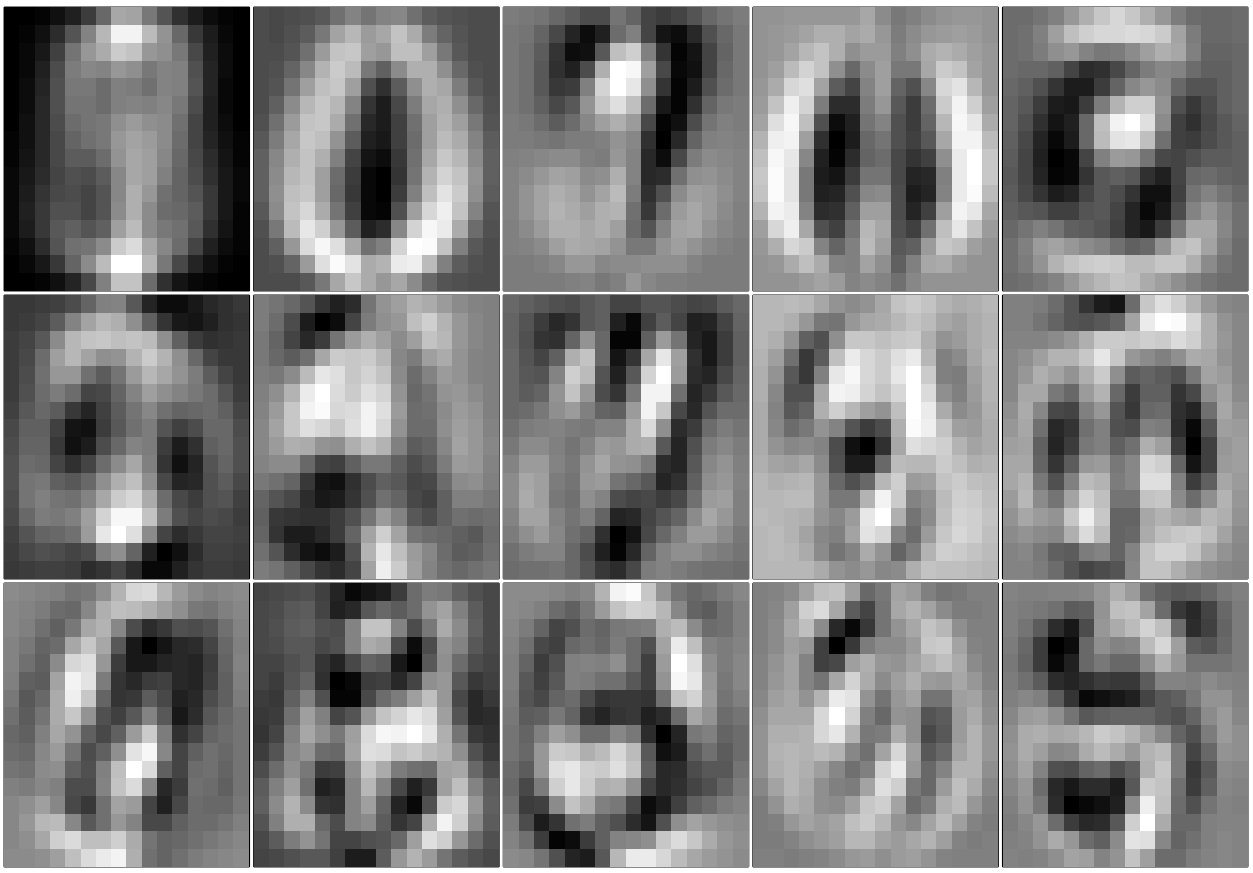
PC scatterplot

```
i = 1; j = 2;
par(mfrow=c(1,1), mar=c(3,3,1,1))
plot(U[,i],U[,j],type="n", xlab = "PC1", ylab = "PC2", main = "PC Scatterplot of Digits")
text(U[,i],U[,j],rownames(digits),col=rownames(digits),cex=.7)
```



PC loadings

```
#PC loadings
par(mfrow=c(3,5),mar=c(.1,.1,.1,.1))
for(i in 1:15){
  imagedigit(V[,i])
}
```

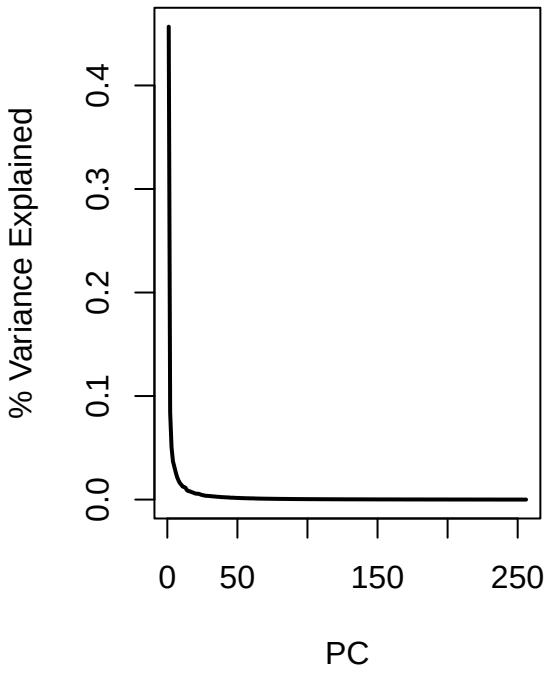


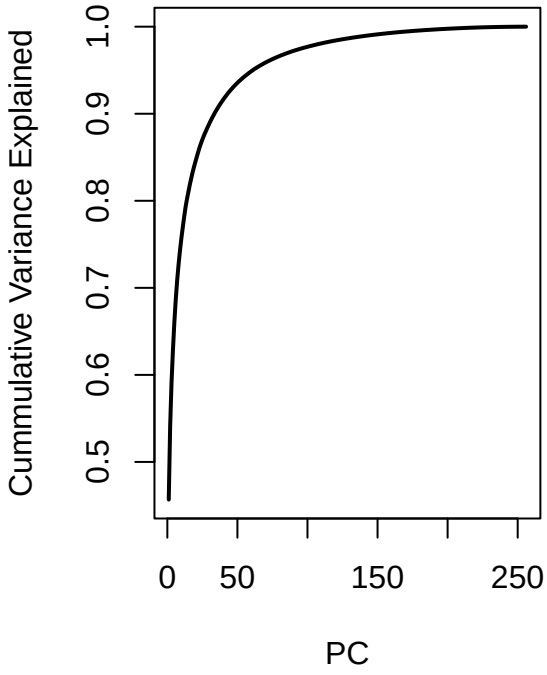
Variance Explained

```
#Variance Explained
varex = 0
cumvar = 0
denom = sum(D^2)
for(i in 1:256){
  varex[i] = D[i]^2/denom
  cumvar[i] = sum(D[1:i]^2)/denom
}
```

Screeplot

```
par(mfrow=c(1,2))
plot(1:256,varex,type="l",lwd=2,xlab="PC",ylab="% Variance Explained")
plot(1:256,cumvar,type="l",lwd=2,xlab="PC",ylab="Cummulative Variance Explained")
```



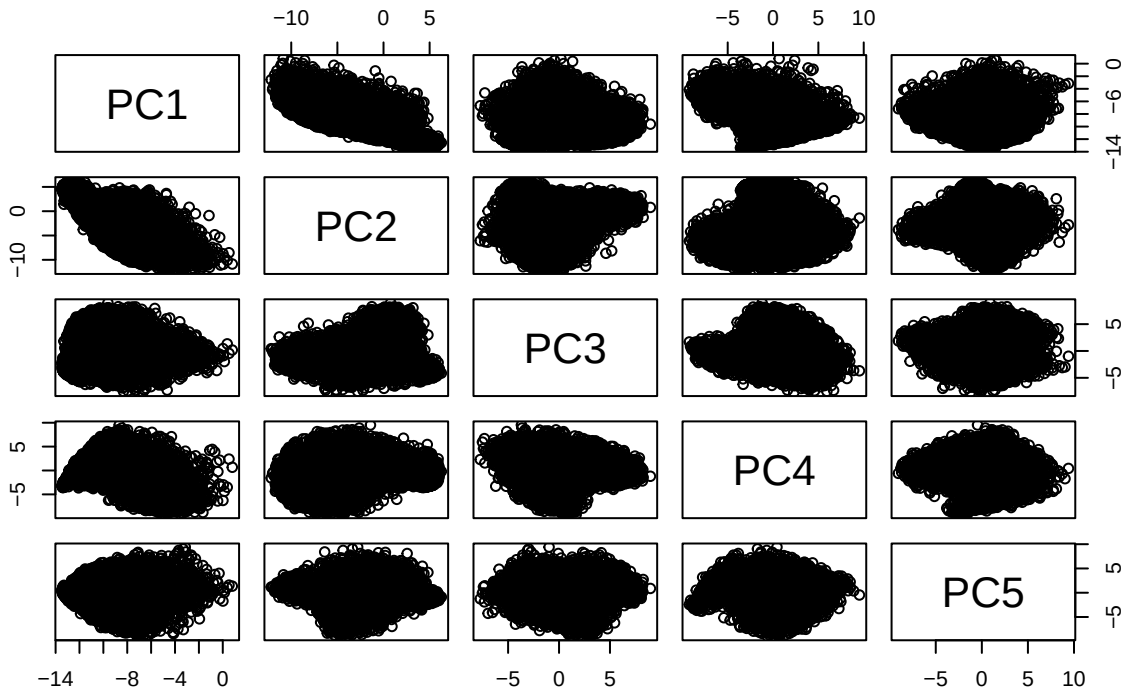


Pairs Plot

```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
Z_sub = Z[,1:5]
comp_labels<-c("PC1","PC2","PC3","PC4", "PC5")
pairs(Z_sub, labels = comp_labels, main = "Pairs of PC's for Digits Data")
```



Problem 2 - MDS

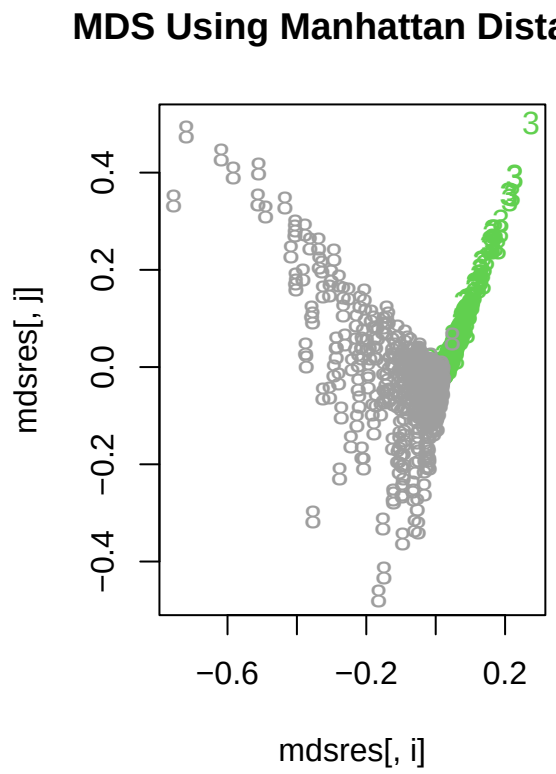
classical MDS (Note, this may take some time - try only on 3's and 8's)

```
dat38 = rbind(digits[which(rownames(digits)==3),],digits[which(rownames(digits)==8),])
dim(dat38)
```

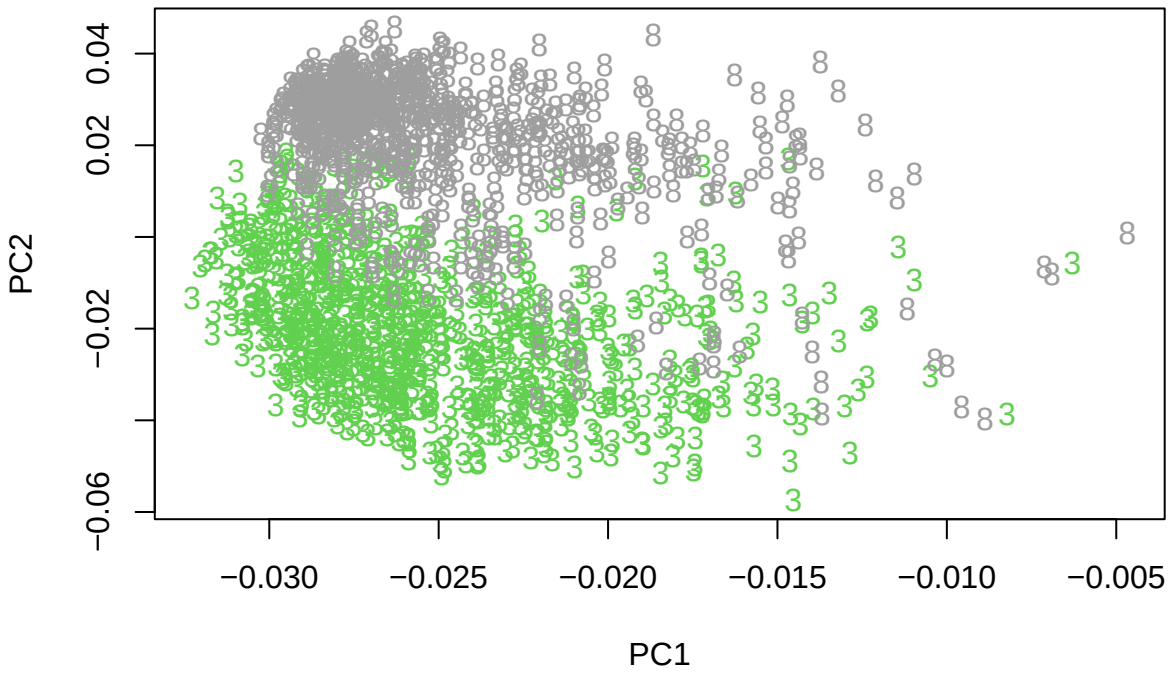
```
## [1] 1532  256
```

```
#PCA for comparison
svdd = svd(dat38)
U = svdd$u
V = svdd$v #PC loadings
D = svdd$d
Z = dat38%*%V #PCs
```

```
#MDS
Dmat = dist(dat38,method="maximum") #Manhattan (L1) Distance
mdsres = cmdscale(Dmat,k=10)

i = 1; j = 2;
par(mfrow=c(1,2))
plot(mdsres[,i],mdsres[,j],type="n", main = "MDS Using Manhattan Distance")
text(mdsres[,i],mdsres[,j],rownames(dat38),col=rownames(dat38))
```



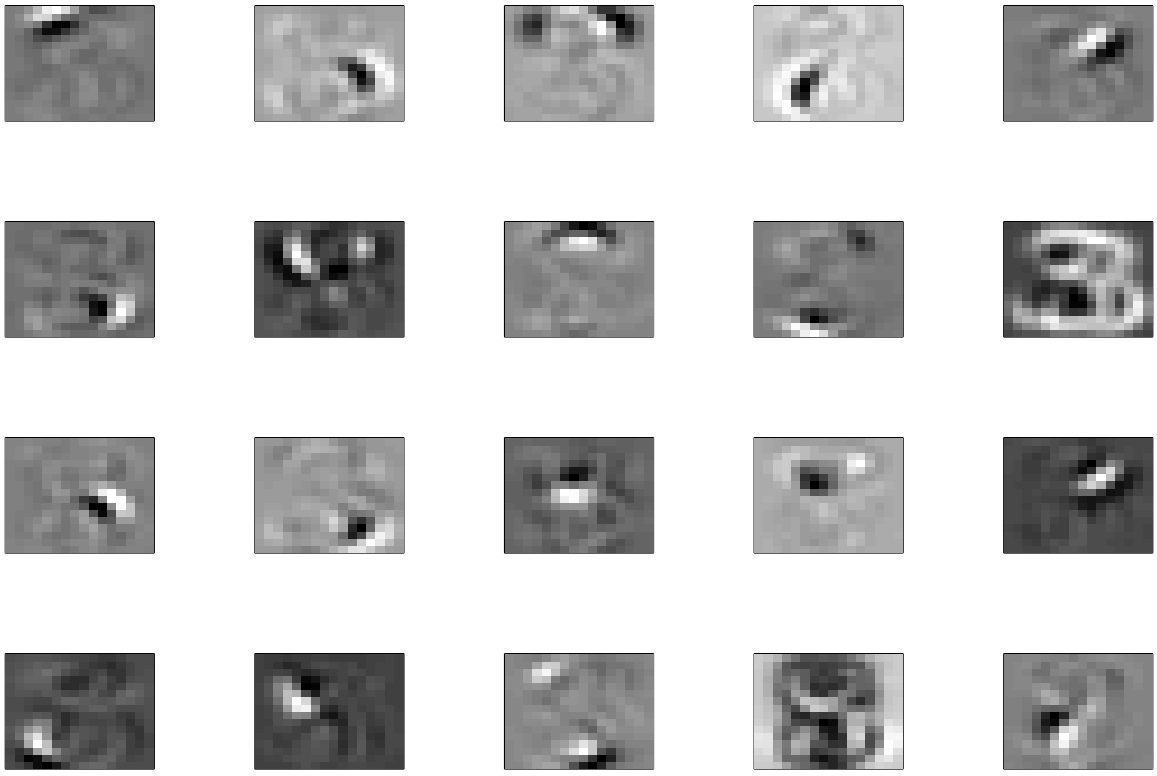

```
plot(U[,i],U[,j],type="n",xlab="PC1",ylab="PC2")
text(U[,i],U[,j],rownames(dat38),col=rownames(dat38))
```



Problem 3 - ICA

```
library(fastICA)
require("fastICA")
```

```
K = 20
icafit = fastICA(t(dat38),n.comp=K)

#plot independent source signals
options(width = 60)
par(mfrow=c(4,5),mar = c(2, 2, 2, 2))
for(i in 1:K){
  imagedigit(icafit$S[,i])
}
```



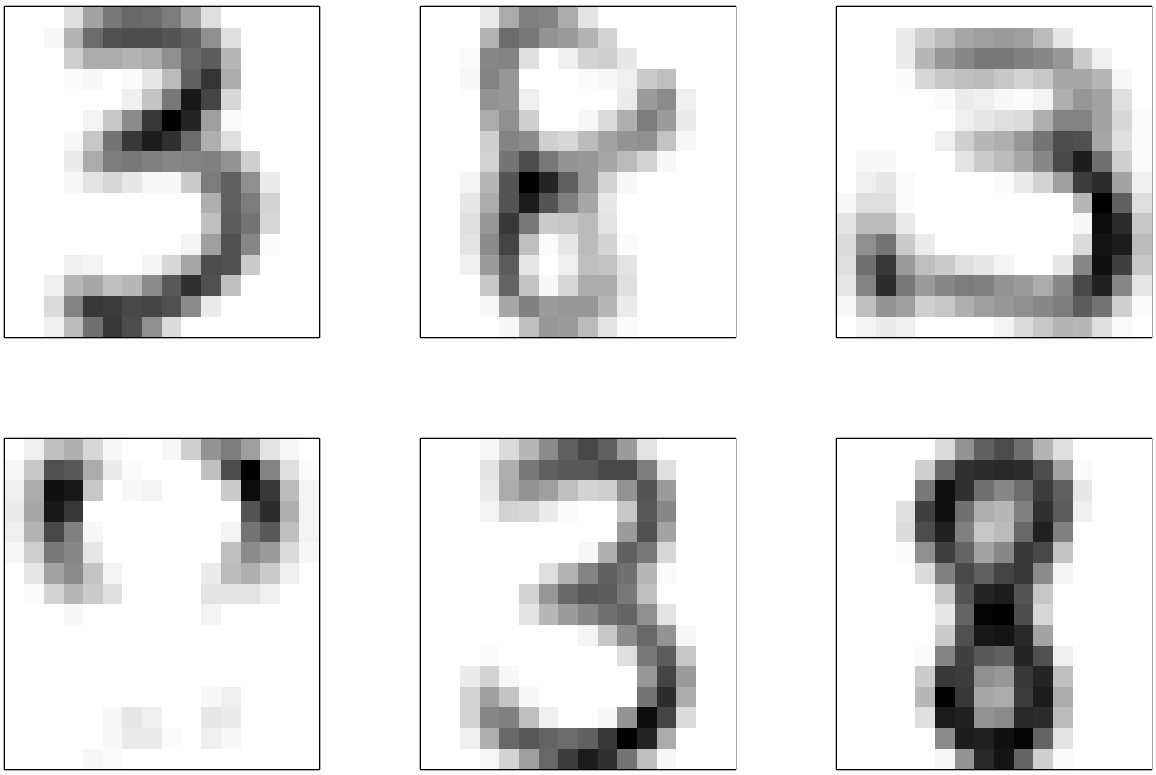
Problem 4 - NMF

```
#Note that this may take a while to run - try using smaller values of K
library(NMF)
```

```
## Loading required package: registry
## Loading required package: rngtools
## Loading required package: cluster
## NMF - BioConductor layer [OK] | Shared memory capabilities [NO: bigmemory] | Cores 2/2
##   To enable shared memory capabilities, try: install.extras('
## NMF
## ')
```

```
K = 6
nmffit = nmf(dat38+1,rank=K)
W = basis(nmffit)
```

```
H = coef(nmffit)

#plot NMF basis factors
options(width = 60)
par(mfrow=c(2,3),mar = c(2, 2, 2, 2))
for(i in 1:K){
  imagedigit(H[i,])
}
```



Problem 5 - UMAP

Load Packages

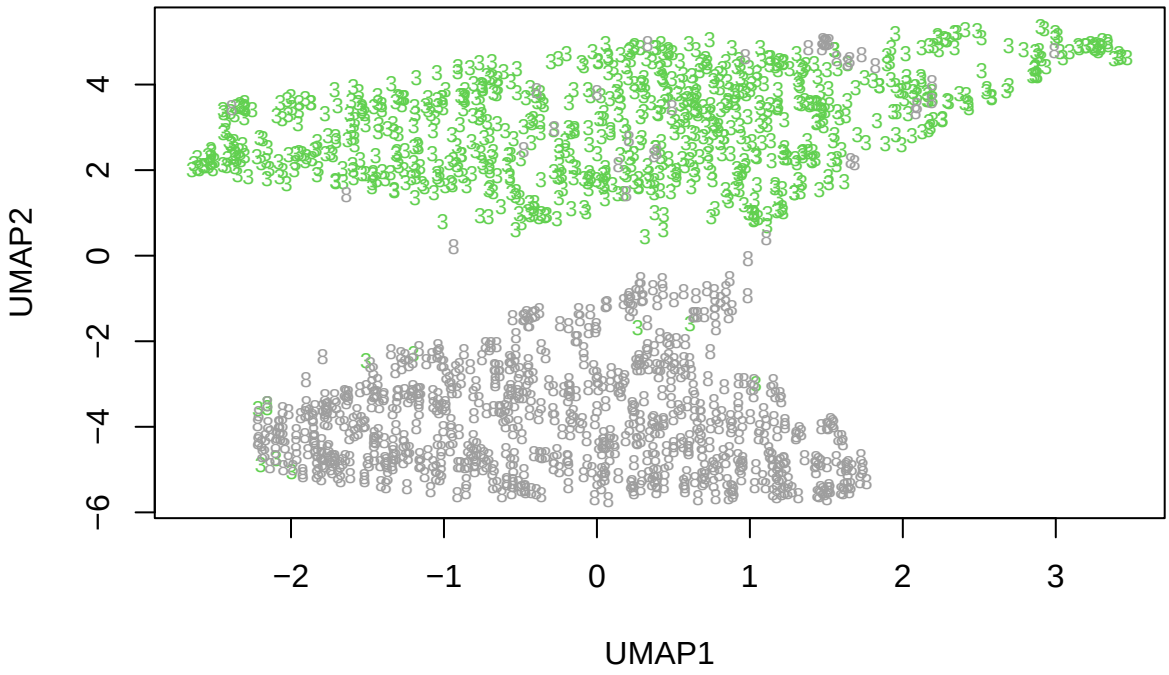
```
library(umap)
library(Rtsne)
```

Run UMAP

```
digits.umap = umap(dat38)
```

Plot UMAP

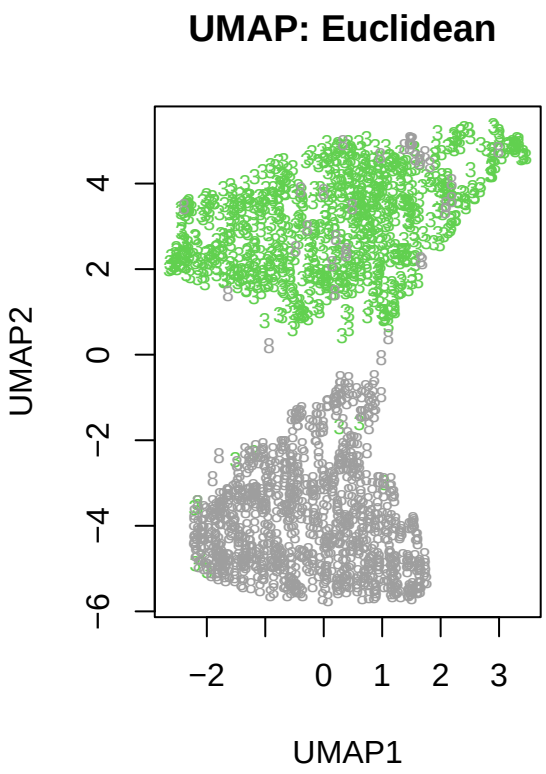
```
plot(digits.umap$layout[,1],y=digits.umap$layout[,2], type ='n', main = "UMAP on Digits 3,8 ", xlab = "
text(digits.umap$layout[,1],y=digits.umap$layout[,2],rownames(dat38),col=rownames(dat38),cex=.7)
```

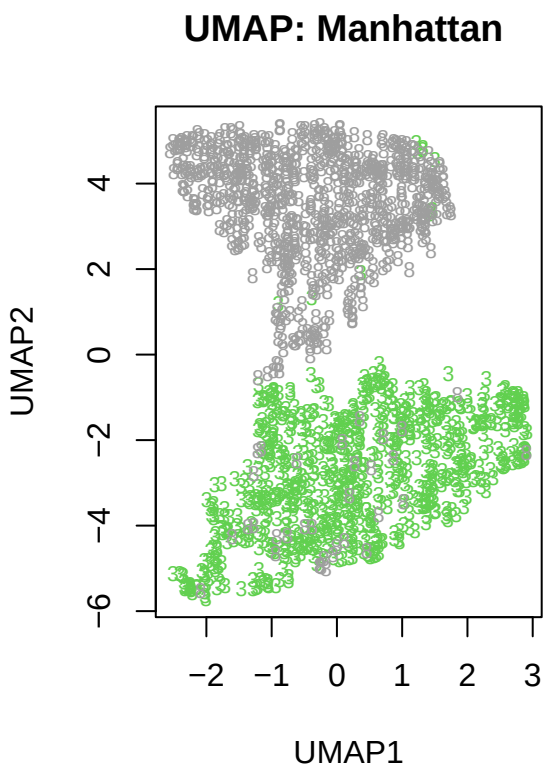


Try

Changing Hyperparameters

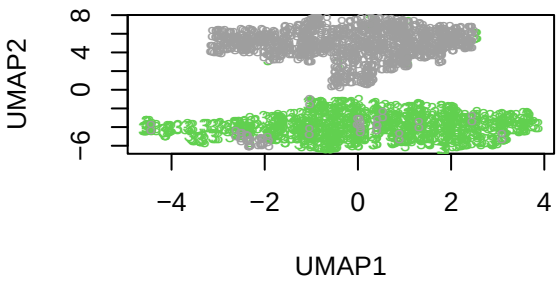
```
digits.umap1 = umap(dat38,metric="euclidean")
digits.umap2 = umap(dat38,metric="manhattan")


par(mfrow=c(1,2))
plot(digits.umap1$layout[,1],y=digits.umap1$layout[,2], type ='n', main = "UMAP: Euclidean ", xlab = "U
text(digits.umap1$layout[,1],y=digits.umap1$layout[,2],rownames(dat38),col=rownames(dat38),cex=.7)

plot(digits.umap2$layout[,1],y=digits.umap2$layout[,2], type ='n', main = "UMAP: Manhattan ", xlab = "U
text(digits.umap2$layout[,1],y=digits.umap2$layout[,2],rownames(dat38),col=rownames(dat38),cex=.7)
```

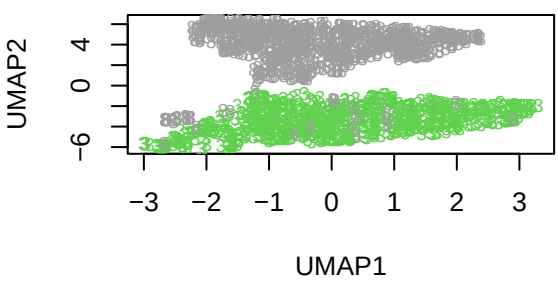


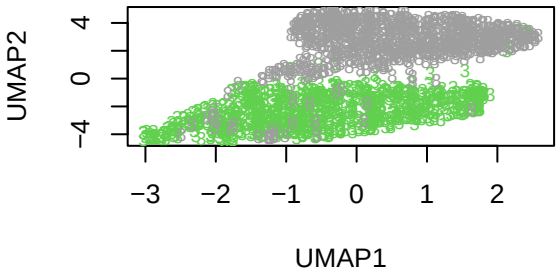


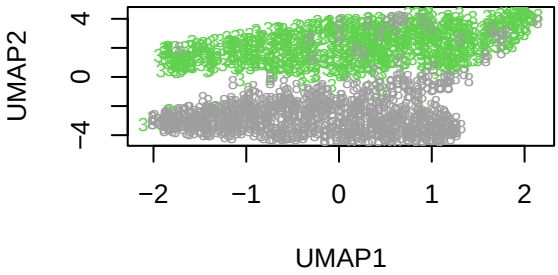

```
nneighs = c(5,10,50,100)
digits.umap = list()
for(i in nneighs){
  digits.umap[[i]] = umap(dat38,n_neighbors=i)
}
```

```
par(mfrow=c(2,2))
for(i in nneighs){
  plot(digits.umap[[i]]$layout[,1],y=digits.umap[[i]]$layout[,2], type ='n', xlab = "UMAP1", ylab = "UM
  text(digits.umap[[i]]$layout[,1],y=digits.umap[[i]]$layout[,2],rownames(dat38),col=rownames(dat38),ce
}
```







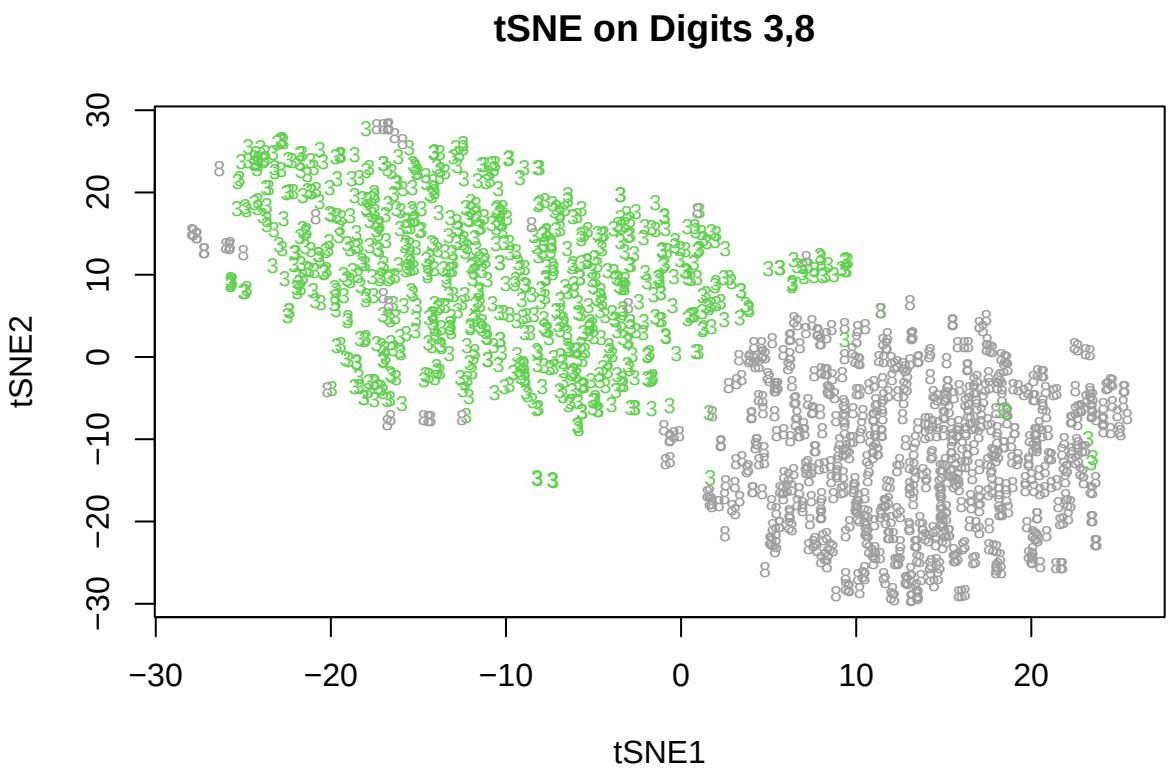


Problem 6 - tSNE

Run tSNE

```
tsne_digit <- Rtsne(as.matrix(dat38))
```

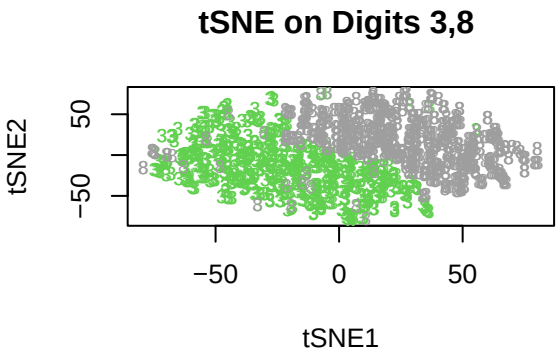
```
plot(tsne_digit$Y[,1],y=tsne_digit$Y[,2], type ='n', main = "tSNE on Digits 3,8 ", xlab = "tSNE1", ylab
text(tsne_digit$Y[,1],y=tsne_digit$Y[,2],rownames(dat38),col=rownames(dat38),cex=.7)
```

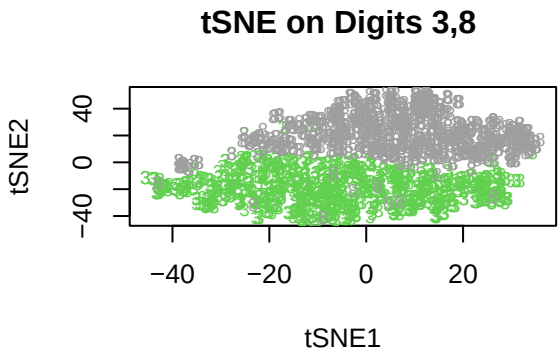


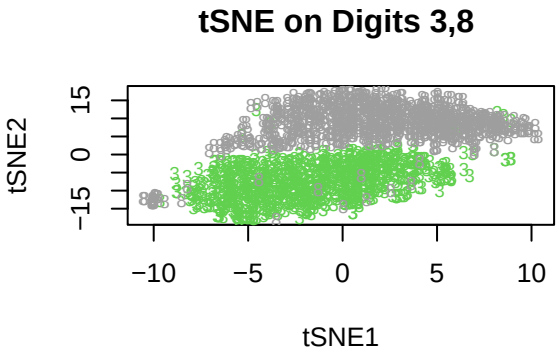
Try Changing Hyperparameters (Perplexity)

```
perps = c(2,10,100)
tsne_digit = list()
for(i in perps){
  tsne_digit[[i]] <- Rtsne(as.matrix(dat38),perplexity=i)
}
```

```
par(mfrow=c(2,2))
for(i in perps){
  plot(tsne_digit[[i]]$Y[,1],y=tsne_digit[[i]]$Y[,2], type ='n', main = "tSNE on Digits 3,8 ", xlab = "
  text(tsne_digit[[i]]$Y[,1],y=tsne_digit[[i]]$Y[,2],rownames(dat38),col=rownames(dat38),cex=.7)
}
```


tSNE on Digits 3,8
tSNE2
50
−50
−50
0
50
tSNE1


tSNE on Digits 3,8
tSNE2
40
0
−40
−40
−20
0
20
tSNE1


tSNE on Digits 3,8
tSNE2
15
0
−15
−10
−5
0
5
10
tSNE1