CS 498 AML
Homework 3
Chao Xu (Dustin)
chaox2@illinois.edu

Number of PCs->	**0N**	**1N**	**2N**	**3N**	**4N**	**0c**	**1c**	**2c**	**3c**	**4c**
Dataset I	4.54247067	0.38345031	0.175563	0.14178365	0.16083836	4.54311903	0.38461353	0.17781528	0.14444051	0.16083836
Dataset II	4.54247067	0.64109318	0.71562849	0.90839291	1.11565786	4.54953899	0.64864211	0.75062113	0.94197282	1.11565786
Dataset III	4.54247067	1.29037245	1.96724039	2.65084114	3.65327973	4.55747296	1.32346215	2.11974805	3.02737992	3.65327973
Dataset IV	4.54247067	0.79994274	0.82808255	0.98494977	1.194	4.56619867	0.84061416	1.2070898	1.27119197	1.194
Dataset V	4.54247067	1.91776775	3.3317221	4.5482572	5.13926667	4.919928	2.83567943	4.6514345	4.97124727	5.13926667

Observation: We could see that, as the noise gets larger, using fewer principal components gives a more accurate estimate of the original dataset (i.e. the one without noise).

jupyter Untitled Last Checkpoint: 几秒前 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted

Run Code

In [1]:
```r
# import necessary libraries
library(fscaret) # for mse
library(doParallel)
# accelerate using parallel computing
registerDoParallel(makeCluster(detectCores()))
```

In [2]:
```r
# INPUT: 150rows, 4cols
dataIN1 <- as.matrix(read.csv('hw3-data/dataI.csv'))
dataIN2 <- as.matrix(read.csv('hw3-data/dataII.csv'))
dataIN3 <- as.matrix(read.csv('hw3-data/dataIII.csv'))
dataIN4 <- as.matrix(read.csv('hw3-data/dataIV.csv'))
dataIN5 <- as.matrix(read.csv('hw3-data/dataV.csv'))
dataIN6 <- as.matrix(read.csv('hw3-data/iris.csv'))
dataIN <- list(dataIN1, dataIN2, dataIN3, dataIN4, dataIN5, dataIN6)
mean <- dataIN
dataIN_noiseless <- dataIN
restructed <- list(dataIN, dataIN, dataIN, dataIN, dataIN, dataIN, dataIN, dataIN)
```

In [3]:
```r
# calculate the mean and center the data
for (i in 1:length(dataIN)){
    for (j in 1:4){
        mean[[i]][,j] <- mean(dataIN[[i]][,j])
        dataIN[[i]][,j] <- dataIN[[i]][,j] - mean[[i]][,j]
    }
}
# center the noisy dataset using the noiseless mean
for (i in 1:5){
    for (j in 1:4){
        dataIN_noiseless[[i]][,j] <- dataIN_noiseless[[i]][,j] - mean[[6]][,j]
    }
}
```

In [4]:
```r
# calculate the covariance matrix
covtmp <- matrix(0,4,4)
cov <- list(covtmp, covtmp, covtmp, covtmp, covtmp, covtmp)
loadings <- cov
for (i in 1:length(dataIN)){
    for (j in 1:4){
        for (k in 1:4){
            cov[[i]][j,k] <- cov(dataIN[[i]][,j],dataIN[[i]][,k])
        }
    }
}
```

In [5]:
```r
# calculate the eigenvectors of covariance matrix
for (i in 1:length(dataIN)){
    loadings[[i]] <- eigen(cov[[i]])$vectors
}
```

In [6]:
```r
# calculate MSE when for different number of principal components & different datasets
error <- matrix(0.0,8,5)
# loop the number of principal components (pc = 0 will be in the next code cell)
for (pc in 1:4){
    # loop the noisy datasets
    for (i in 1:5){
        # restructed[[1:4]] stores restruction from noiseless mean & pcs, [[5:8]] from respective noisy
        restructed[[pc]][[i]] <- dataIN_noiseless[[i]] %*% loadings[[6]][,1:pc] %*% t(loadings[[6]][,1:pc])
        restructed[[pc+4]][[i]] <- dataIN[[i]] %*% loadings[[i]][,1:pc] %*% t(loadings[[i]][,1:pc])
        for (k in 1:4){
            restructed[[pc]][[i]][,k] <- restructed[[pc]][[i]][,k] + mean[[6]][,k]
            restructed[[pc+4]][[i]][,k] <- restructed[[pc+4]][[i]][,k] + mean[[i]][,k]
        }
        error[pc,i] <- MSE(restructed[[pc]][[i]], dataIN6, 600)*4
        error[pc+4,i] <- MSE(restructed[[pc+4]][[i]], dataIN6, 600)*4
    }
}
error <- t(error)
```

In [7]:
```r
# calculate MSE for 0N 0c situations (PCs = 0)
error_pc0 <- matrix(0.0,2,5)
for (j in 1:5){
    error_pc0[1,j] <- MSE(mean[[6]], dataIN6, 600)*4
    error_pc0[2,j] <- MSE(mean[[j]], dataIN6, 600)*4
}
```

In [8]:
```r
# produce the final 5*10 table
finalTable <- cbind(as.matrix(error_pc0[1,]),error[,1:4],as.matrix(error_pc0[2,]),error[,5:8])
colnames(finalTable) = c("0N", "1N", "2N", "3N", "4N", "0c", "1c", "2c", "3c", "4c")
write.csv(finalTable, file = "chaox2-numbers.csv", row.names = F)
```

In [9]:
```r
# reconstruction of dataset II, expanded onto 2 pcs, where mean and pcs are computed from the dataset of version II
requiredRestruction <- restructed[[2+4]][[2]]
colnames(requiredRestruction) = c("Sepal.Length","Sepal.Width","Petal.Length","Petal.Width")
write.csv(requiredRestruction, file = "chaox2-recon.csv", row.names = F)
```

CS 498 AML
Homework 3

Chao Xu (Dustin)
chaox2@illinois.edu