Matrix Completion Using PCA and SoftImpute

Shrey Patel

12/8/2021

1 Getting Data Ready

```
# loading data set
data(chicago)
X <- chicago
head(X)
```

```
##   death pm10median pm25median  o3median  so2median    time tmpd
## 1   130 -7.4335443         NA -19.59234  1.9280426 -2556.5 31.5
## 2   150         NA         NA -19.03861 -0.9855631 -2555.5 33.0
## 3   101 -0.8265306         NA -20.21734 -1.8914161 -2554.5 33.0
## 4   135  5.5664557         NA -19.67567  6.1393413 -2553.5 29.0
## 5   126         NA         NA -19.21734  2.2784649 -2552.5 32.0
## 6   130  6.5664557         NA -17.63400  9.8585839 -2551.5 40.0
```

```
# creating matrix with missing value
X1 <- na.omit(X)
X1 <- scale(X1)
dim(X1)
```

```
## [1] 719   7
```

```
head(X1)
```

```
##           death  pm10median pm25median   o3median  so2median      time
## 4023 0.73157218 -0.66403580 -0.8166056 -1.3800817 -0.8405208 -1.812097
## 4029 0.16453227 -0.04739007 -0.9659269 -1.1723713  1.3303686 -1.794118
## 4035 1.22773211  0.03787407  2.2387377 -1.3874351  0.5396463 -1.776138
## 4041 1.08597213  0.32556617  1.8041692 -1.5235392  1.7860686 -1.758158
## 4047 0.02277229 -0.08958263  0.3626445 -1.3230635  0.1706265 -1.740178
## 4053 0.73157218 -1.00882838 -0.5409355  0.9603717 -1.5175527 -1.722198
##            tmpd
## 4023 -0.2788074
## 4029 -1.8924043
## 4035 -1.6234714
## 4041 -1.3545386
## 4047 -0.8704596
## 4053 -1.1393924
```

```
chi.omit <- 200
set.seed (1234)
in.row <- sample ( 719, chi.omit)
in.col <- sample (1:7, chi.omit , replace = TRUE)
X1.omit <- X1
```

```r
index.omit <- cbind (in.row , in.col)
X1.omit[index.omit] <- NA
ismiss <- is.na(X1.omit)
# X1.omit is the one with missing data
```

2 Matrix Completion Using PCA

```r
fit.pca <- function(X , M) {
  pcob <- prcomp(X)
  with(pcob,
       x[, 1:M, drop = FALSE] %*%
         (t(rotation[, 1:M, drop = FALSE]))
  )
}

Mat.Complete <- function(X,thresh,maxiter){
  # Step 1 #
  # calculating Xhat
  Xhat <- X1.omit
  xbar <- colMeans(X1.omit , na.rm = TRUE)
  Xhat[index.omit] <- xbar[in.col]

  # Step 2 #
  # initializing progress variables
  rel_err <- 1                              # relative error
  iter <- 0                                 # iterator
  mssold <- mean((scale(X1.omit, xbar, FALSE)[!ismiss])^2) # mse of non-missing elements (old version)
  mss0 <- mean(X1.omit[!ismiss]^2)          # mse of non-missing elements

  while(rel_err > thresh) {
    iter <- iter + 1
    # Step 2(a)
    Xapp <- fit.pca(Xhat , M = 1)
    # Step 2(b)
    Xhat[ismiss] <- Xapp[ismiss]
    # Step 2(c)
    mss <- mean(((X1.omit - Xapp)[!ismiss])^2)
    rel_err <- (mssold - mss)/mss0
    mssold <- mss
    cat("Iter:", iter, "MSS:", mss, "Rel. Err:", rel_err, "\n")
    if (iter >= maxiter){
      cat("WARNING: Maximum iterations reached. Exiting the loop.")
      break
    }
  }
  list <- list("x"=Xapp, "iter"=iter, "relerr"=rel_err)
  return(list)
}

tic("PCA")
output = Mat.Complete(X1.omit, 1e-7, 30)
```

```
## Iter: 1 MSS: 0.70007 Rel. Err: 0.2921999
## Iter: 2 MSS: 0.6977939 Rel. Err: 0.002300821
## Iter: 3 MSS: 0.697601 Rel. Err: 0.0001949948
## Iter: 4 MSS: 0.697576 Rel. Err: 2.527417e-05
## Iter: 5 MSS: 0.6975714 Rel. Err: 4.680467e-06
## Iter: 6 MSS: 0.6975703 Rel. Err: 1.152808e-06
## Iter: 7 MSS: 0.6975699 Rel. Err: 3.439953e-07
## Iter: 8 MSS: 0.6975698 Rel. Err: 1.142639e-07
## Iter: 9 MSS: 0.6975698 Rel. Err: 3.999657e-08
```

```
toc()
```

```
## PCA: 0.08 sec elapsed
```

```
cat("Correlation between imputed and actual values: ",
      cor(output$x[ismiss], X1[ismiss]))
```

```
## Correlation between imputed and actual values:  0.4386683
```

3 Matrix Completion Using SoftImpute

```
fit.sftimp <- function(X , M) {
  sftimp <- softImpute(X)
  with(sftimp,
       u[, 1:M, drop = FALSE] %*%
         (d[1:M] * t(v[, 1:M, drop = FALSE ]))
  )
}

Mat.Complete2 <- function(X,thresh,maxiter){
  # Step 1 #
  # calculating Xhat
  Xhat <- X1.omit
  xbar <- colMeans(X1.omit , na.rm = TRUE)
  Xhat[index.omit] <- xbar[in.col]

  # Step 2 #
  # initializing progress variables
  rel_err <- 1                              # relative error
  iter <- 0                                 # iterator
  mssold <- mean((scale(X1.omit, xbar, FALSE)[!ismiss])^2) # mse of non-missing elements (old version)
  mss0 <- mean(X1.omit[!ismiss]^2)          # mse of non-missing elements

  while(rel_err > thresh) {
    iter <- iter + 1
    # Step 2(a)
    Xapp <- fit.sftimp(Xhat , M = 1)
    # Step 2(b)
    Xhat[ismiss] <- Xapp[ismiss]
    # Step 2(c)
    mss <- mean(((X1.omit - Xapp)[!ismiss])^2)
    rel_err <- (mssold - mss)/mss0
    mssold <- mss
    cat("Iter:", iter, "MSS:", mss, "Rel. Err:", rel_err, "\n")
```

```
    if (iter >= maxiter){
      cat("WARNING: Maximum iterations reached. Exiting the loop.")
      break
    }
  }
  list <- list("x"=Xapp, "iter"=iter, "relerr"=rel_err)
  return(list)
}

tic("softImpute")
output2 = Mat.Complete2(X1.omit, 1e-7, 30)
```

```
## Iter: 1 MSS: 0.7000672 Rel. Err: 0.2922028
## Iter: 2 MSS: 0.6977914 Rel. Err: 0.002300505
## Iter: 3 MSS: 0.6975982 Rel. Err: 0.0001952451
## Iter: 4 MSS: 0.6975731 Rel. Err: 2.542873e-05
## Iter: 5 MSS: 0.6975684 Rel. Err: 4.685838e-06
## Iter: 6 MSS: 0.6975676 Rel. Err: 8.108962e-07
## Iter: 7 MSS: 0.6975669 Rel. Err: 7.175049e-07
## Iter: 8 MSS: 0.6975671 Rel. Err: -1.560591e-07
```

```
toc()
```

```
## softImpute: 0.14 sec elapsed
```

```
cat("Correlation between imputed and actual values: ",
      cor(output2$x[ismiss], X1[ismiss]))
```

```
## Correlation between imputed and actual values:  0.4385084
```