Lab 5

Elizabeth McHugh

11:59PM March 18, 2021

Create a 2x2 matrix with the first column 1's and the next column iid normals. Find the absolute value of the angle (in degrees, not radians) between the two columns in absolute difference from 90 degrees.

```
norm_vec = function(v){       #Computes norm of a vector
  sqrt(sum(v^2))
}

angle_btwn_deg = function(v1, v2){                      #Computes angle btwn two columns in absolute d
  theta_rad = (acos((t(v1) %*% v2) / (norm_vec(v1) * norm_vec(v2))))
  abs(90 - (theta_rad * 180 / pi))
}

X = cbind(1, rnorm(2))            #2x2 matrix with first column 1's, second column iid normals

angle_btwn_deg(X[,1], X[,2])
```

```
##          [,1]
## [1,] 48.23669
```

Repeat this exercise `Nsim = 1e5` times and report the average absolute angle.

```
Nsim = 1e5

angles = array(NA, Nsim)      #Stores each iteration of computed angles

for (i in 1:Nsim){              #Repeats angle between computation 10,000 times

  X = cbind(1, rnorm(2))

  angles[i] = angle_btwn_deg(X[,1], X[,2])
}

mean(angles)        #Average of 10,000 computations
```

```
## [1] 44.96579
```

Create a n x 2 matrix with the first column 1's and the next column iid normals. Find the absolute value of the angle (in degrees, not radians) between the two columns. For n = 10, 50, 100, 200, 500, 1000, report the average absolute angle over `Nsim = 1e5` simulations.

```
n_array = c(10, 50, 100, 200, 500, 1000)

Nsim = 1e5

angles = matrix(NA, nrow = Nsim, ncol = length(n_array))
angles_calculated = data.frame("n", "average angle between")

for (i in 1:length(n_array)){
  n = n_array[i]

  for (j in 1:Nsim){
    X = cbind(1, rnorm(n))
    angles[j,i] = angle_btwn_deg(X[,1], X[,2])
  }

  angles_calculated[i,1] = n
  angles_calculated[i,2] = mean(angles[,i])

}

angles_calculated
```

```
##   X.n. X.average.angle.between.
## 1   10          15.326353958008
## 2   50         6.54626497125094
## 3  100         4.59682238128833
## 4  200         3.23585027029219
## 5  500         2.04961602447556
## 6 1000         1.44622292426442
```

What is this absolute angle converging to? Why does this make sense?

The absolute angle difference is converging to 0. This makes sense, since random vectors in a high dimensional space will be nearly orthogonal.

Create a vector y by simulating n = 100 standard iid normals. Create a matrix of size 100 x 2 and populate the first column by all ones (for the intercept) and the second column by 100 standard iid normals. Find the R^2 of an OLS regression of `y ~ X`. Use matrix algebra.

```
Nsim = 100

y = c(rnorm(100))      #vector of 100 standard iid normals

X = cbind(1, y)        #matrix with column one as 1's and column two as vector y

H = X %*% solve(t(X) %*% X) %*% t(X)     #Hat matrix H
y_hat = H %*% y                        #y hat = Hy
y_bar = mean(y)
SSR = sum((y_hat - y_bar)^2)
SST = sum((y - y_bar)^2)
R_squared = SSR / SST

R_squared
```

```
## [1] 1
```

Write a for loop to each time bind a new column of 100 standard iid normals to the matrix X and find the R^2 each time until the number of columns is 100. Create a vector to save all R^2's. What happened??

```
N = 100
X = cbind(1, rnorm(Nsim))
R_squared_vec = array(NA, N - 2)

for (i in 1:(N - 2)){
  X = cbind(X, rnorm(100))

  H = X %*% solve(t(X) %*% X) %*% t(X)
  y_hat = H %*% y
  y_bar = mean(y)
  SSR = sum((y_hat - y_bar)^2)
  SST = sum((y - y_bar)^2)

  R_squared_vec[i] = SSR / SST
}

R_squared_vec
```

```
##  [1] 0.1048795 0.1225507 0.1248143 0.1516812 0.1723363 0.1848039 0.1951556
##  [8] 0.2042017 0.2120199 0.2122463 0.2140001 0.2336006 0.2336420 0.2515251
## [15] 0.2517246 0.2519791 0.2793848 0.2799883 0.2811352 0.3079637 0.3311607
## [22] 0.3417327 0.3474498 0.3476829 0.3547627 0.3615389 0.3649946 0.3747268
## [29] 0.3754795 0.3768441 0.3797153 0.4177693 0.4179567 0.4469131 0.4473574
## [36] 0.4651909 0.4817101 0.4934779 0.4957270 0.5270792 0.5348129 0.5423722
## [43] 0.5536383 0.5552376 0.5654152 0.5656582 0.5677724 0.5678326 0.5919747
## [50] 0.5958491 0.5962397 0.6181192 0.6221603 0.6273792 0.6294834 0.6319679
## [57] 0.6322562 0.6399157 0.6416169 0.6557870 0.6813430 0.6868741 0.6894530
## [64] 0.7196555 0.7248053 0.7250304 0.7261330 0.7491132 0.7494414 0.7494706
## [71] 0.7626759 0.7632934 0.7644834 0.7994048 0.8211416 0.8262456 0.8283324
## [78] 0.8379695 0.8391413 0.8569796 0.8794294 0.8812956 0.8894989 0.8984680
## [85] 0.8995434 0.8997546 0.9525217 0.9546981 0.9546983 0.9546985 0.9557039
## [92] 0.9584850 0.9602151 0.9602706 0.9772392 0.9998332 0.9998913 1.0000000
```

Test that the projection matrix onto this X is the same as I_n. You may have to vectorize the matrices in the `expect_equal` function for the test to work.

```
pacman::p_load(testthat)

I = diag(N)
expect_equal(I, H, test_that_tolerance = 1e-13)
```

Add one final column to X to bring the number of columns to 101. Then try to compute R^2. What happens?

```
X = cbind(X, rnorm(100))

 H = X %*% solve(t(X) %*% X) %*% t(X)
```

```
  y_hat = H %*% y
  y_bar = mean(y)
  SSR = sum((y_hat - y_bar)^2)
  SST = sum((y - y_bar)^2)

  R_squared = SSR / SST

  R_squared
```

Why does this make sense?

Any error occurs. This makes sense, since the matrix has more columns than rows, at least one row must be linearly dependent, thus the inverse of X transpose X is non-existant since X^t X is rank deficient.

Write a function spec'd as follows:

```
#' Orthogonal Projection
#'
#' Projects vector a onto v.
#'
#' @param a   the vector to project
#' @param v   the vector projected onto
#'
#' @returns   a list of two vectors, the orthogonal projection parallel to v named a_parallel,
#'            and the orthogonal error orthogonal to v called a_perpendicular
orthogonal_projection = function(a, v){

  H = v %*% t(v) / norm_vec(v)^2
  a_parallel = H %*% a
  a_perpendicular = a - a_parallel

  list(a_parallel = a_parallel, a_perpendicular = a_perpendicular)
}
```

Provide predictions for each of these computations and then run them to make sure you're correct.

```
orthogonal_projection(c(1,2,3,4), c(1,2,3,4))
```

```
## $a_parallel
##      [,1]
## [1,]    1
## [2,]    2
## [3,]    3
## [4,]    4
##
## $a_perpendicular
##      [,1]
## [1,]    0
## [2,]    0
## [3,]    0
## [4,]    0
```

```
#prediction: a_parallel is (1,2,3,4), a_perpendicular = 0
orthogonal_projection(c(1, 2, 3, 4), c(0, 2, 0, -1))
```

```
## $a_parallel
##      [,1]
## [1,]    0
## [2,]    0
## [3,]    0
## [4,]    0
##
## $a_perpendicular
##      [,1]
## [1,]    1
## [2,]    2
## [3,]    3
## [4,]    4
```

```
#prediction: a_parallel = 0 a_perpendicular = (1,2,3,4)
result = orthogonal_projection(c(2, 6, 7, 3), c(1, 3, 5, 7))
t(result$a_parallel) %*% result$a_perpendicular
```

```
##               [,1]
## [1,] -3.552714e-15
```

```
#prediction: 0 (parallel and perpendicular components should be orthogonal)
result$a_parallel + result$a_perpendicular
```

```
##      [,1]
## [1,]    2
## [2,]    6
## [3,]    7
## [4,]    3
```

```
#prediction: get back original vector (2,6,7,3)
result$a_parallel / c(1, 3, 5 ,7)
```

```
##           [,1]
## [1,] 0.9047619
## [2,] 0.9047619
## [3,] 0.9047619
## [4,] 0.9047619
```

```
#prediction: proportion of original vector length
```

Let's use the Boston Housing Data for the following exercises

```
y = MASS::Boston$medv
X = model.matrix(medv ~ ., MASS::Boston)
p_plus_one = ncol(X)
n = nrow(X)

head(X)
```

```
##   (Intercept)    crim zn indus chas   nox    rm  age    dis rad tax ptratio
## 1           1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3
## 2           1 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8
## 3           1 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8
## 4           1 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7
## 5           1 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7
## 6           1 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7
##    black lstat
## 1 396.90  4.98
## 2 396.90  9.14
## 3 392.83  4.03
## 4 394.63  2.94
## 5 396.90  5.33
## 6 394.12  5.21
```

Using your function `orthogonal_projection` orthogonally project onto the column space of X by projecting y on each vector of X individually and adding up the projections and call the sum `yhat_naive`.

```
yhat_naive = rep(0, n)

for (j in 1:p_plus_one){
  yhat_naive = yhat_naive + orthogonal_projection(y, X[,j])$a_parallel

}
```

How much double counting occurred? Measure the magnitude relative to the true LS orthogonal projection.

```
yhat = X %*% solve(t(X) %*% X) %*% t(X) %*% y
sqrt(sum(yhat_naive^2)) / sqrt(sum(yhat^2))
```

```
## [1] 8.997118
```

Is this ratio expected? Why or why not?

Such a ratio is expected to vary from 1, because yhat_naive is not yhat (thus double counting occurs).

Convert X into V where V has the same column space as X but has orthogonal columns. You can use the function `orthogonal_projection`. This is the Gram-Schmidt orthogonalization algorithm.

```
V = matrix(NA, nrow = n, ncol = p_plus_one)
V[ , 1] = X[ , 1]

for (j in 2:p_plus_one){
  V[ , j] = X[ , j]
  for (k in 1:(j-1)){
    V[ , j] = V[ , j] - orthogonal_projection(X[ , j], V[ , k])$a_parallel
  }

}

V[ , 7] %*% V[ , 9]
```

```
##               [,1]
## [1,] -2.140346e-11
```

Convert V into Q whose columns are the same except normalized

```
Q = matrix(NA, nrow = n, ncol = p_plus_one)

for (j in 1:p_plus_one){
  Q[ , j] = V[ , j] / norm_vec(V[ , j])

}
```

Verify Q^T Q is I_{p+1} i.e. Q is an orthonormal matrix.

```
I = diag(1, p_plus_one)

expect_equal(t(Q) %*% Q, I, tolerance = 1e-12)
```

Is your Q the same as what results from R's built-in QR-decomposition function?

```
Q_from_Rs_builtin = qr.Q(qr(X))

expect_equal(Q, Q_from_Rs_builtin)
```

Is this expected? Why did this happen?

Yes. In general, orthogonal and orthonormal bases are not unique, so there is no reason to expect our Q and the Q computed from the built in QR-decomposition function to be the same.

Project y onto colsp[Q] and verify it is the same as the OLS fit. You may have to use the function `unname` to compare the vectors since the entries will likely have different names.

```
yhat = X %*% solve(t(X) %*% X) %*% t(X) %*% y
expect_equal(unname(yhat), unname(Q %*% t(Q) %*% y))
```

Project y onto colsp[Q] one by one and verify it sums to be the projection onto the whole space.

```
yhat_naive = rep(0, n)

for (j in 1:p_plus_one){
  yhat_naive = yhat_naive + orthogonal_projection(y, Q[ , j])$a_parallel

}
```

Split the Boston Housing Data into a training set and a test set where the training set is 80% of the observations. Do so at random.

```
n = nrow(X)

K = 5
n_test = round(n * 1 / K)
n_train = n - n_test

X_test = matrix(NA, nrow = n_test, ncol = ncol(X))
X_train = matrix(NA, nrow = n_train, ncol = ncol(X))
```

```
y_test = array(NA, n_test)
y_train = array(NA, n_train)


set.seed(28)
n_shuffle = sample(nrow(X))           #Randomly shuffled indexes

for (i in 1:n_test){                    #Test data set X and response y
  X_test[i, ] = X[n_shuffle[i], ]
  y_test[i] = y[n_shuffle[i]]
}

for (i in 1:n_train){                             #Training data set X and response y
   X_train[i, ] = X[n_shuffle[(i + n_test)], ]
   y_train[i] = y[n_shuffle[(i + n_test)]]
}
```

Fit an OLS model. Find the s_e in sample and out of sample. Which one is greater? Note: we are now using s_e and not RMSE since RMSE has the n-(p + 1) in the denominator not n-1 which attempts to de-bias the error estimate by inflating the estimate when overfitting in high p. Again, we're just using `sd(e)`, the sample standard deviation of the residuals.

```
mod = lm(y_train ~ X_train)
summary(mod)$sigma

s_e = sd(mod$residuals)
s_e

predict_oos = predict(mod, data.frame(X_test))

ooss_e = sd(y_hat - y)
```

Do these two exercises `Nsim = 1000` times and find the average difference between s_e and ooss_e.

```
#TODO
```

We'll now add random junk to the data so that `p_plus_one = n_train` and create a new data matrix `X_with_junk`.

```
X_with_junk = cbind(X, matrix(rnorm(n * (n_train - p_plus_one)), nrow = n))
dim(X)
```

```
## [1] 506  14
```

```
dim(X_with_junk)
```

```
## [1] 506 405
```

Repeat the exercise above measuring the average s_e and ooss_e but this time record these metrics by number of features used. That is, do it for the first column of `X_with_junk` (the intercept column), then do it for the first and second columns, then the first three columns, etc until you do it for all columns of `X_with_junk`. Save these in `s_e_by_p` and `ooss_e_by_p`.

```
#TODO
```

You can graph them here:

```
pacman::p_load(ggplot2)
ggplot(
  rbind(
    data.frame(s_e = s_e_by_p, p = 1 : n_train, series = "in-sample"),
    data.frame(s_e = ooss_e_by_p, p = 1 : n_train, series = "out-of-sample")
  )) +
  geom_line(aes(x = p, y = s_e, col = series))
```

Is this shape expected? Explain.

#TO-DO