

# Tarea 1

Pablo, Sofía, Román

26/1/2020

## Ejercicio 3

### Las flores de Fisher y Anderson

3a

```
X <- iris3[, , 1] #1 is for Setosa
s_mn <- apply(X = X, MARGIN = 2, FUN = mean)

#corrected mean square
dim_set <- dim(X)[1]
s_mn_matrix <- matrix(rep(s_mn, each = dim_set), nrow = dim_set)
A <- X - s_mn_matrix
# for(i in 1:4){
#   A[,i] <- A[,i] * A[,i]
# }
A <- t(A) %*% A
#unbiased sample covarianse
Sx <- (1/(dim_set - 1) * A)
```

3b

```
#eigenpar
eigen_list <- eigen(Sx)
```

3c

```
U <- eigen_list$vectors
L <- diag(eigen_list$values)

#compute Sx = ULU'
Sx_prim = U %*% L %*% t(U)

#compute UU' & U'U
UUt <- U %*% t(U)
UtU <- t(U) %*% U
```

#3d

```

matplot_setosa <- iris %>%
  filter(Species == "setosa") %>%
  select(Sepal.Length, Sepal.Width, Petal.Length) %>%
  ggpairs() +
  theme_bw() +
  labs(title = "SETOSA", x = "", y = "")

matplot_veris <- iris %>%
  filter(Species == "versicolor") %>%
  select(Sepal.Length, Sepal.Width, Petal.Length) %>%
  ggpairs() +
  theme_bw() +
  labs(title = "VERSICOLOR", x = "", y = "")

matplot_virg <- iris %>%
  filter(Species == "virginica") %>%
  select(Sepal.Length, Sepal.Width, Petal.Length) %>%
  ggpairs() +
  theme_bw() +
  labs(title = "VIRGINICA", x = "", y = "")

matplot_all <- iris %>%
  ggscatmat(color = 'Species') +
  theme_bw() +
  labs(title = "FLORES", x = "", y = "", color = "Especie de\n flor")

```

```

## Warning in ggscatmat(., color = "Species"): Factor variables are omitted in
## plot

```

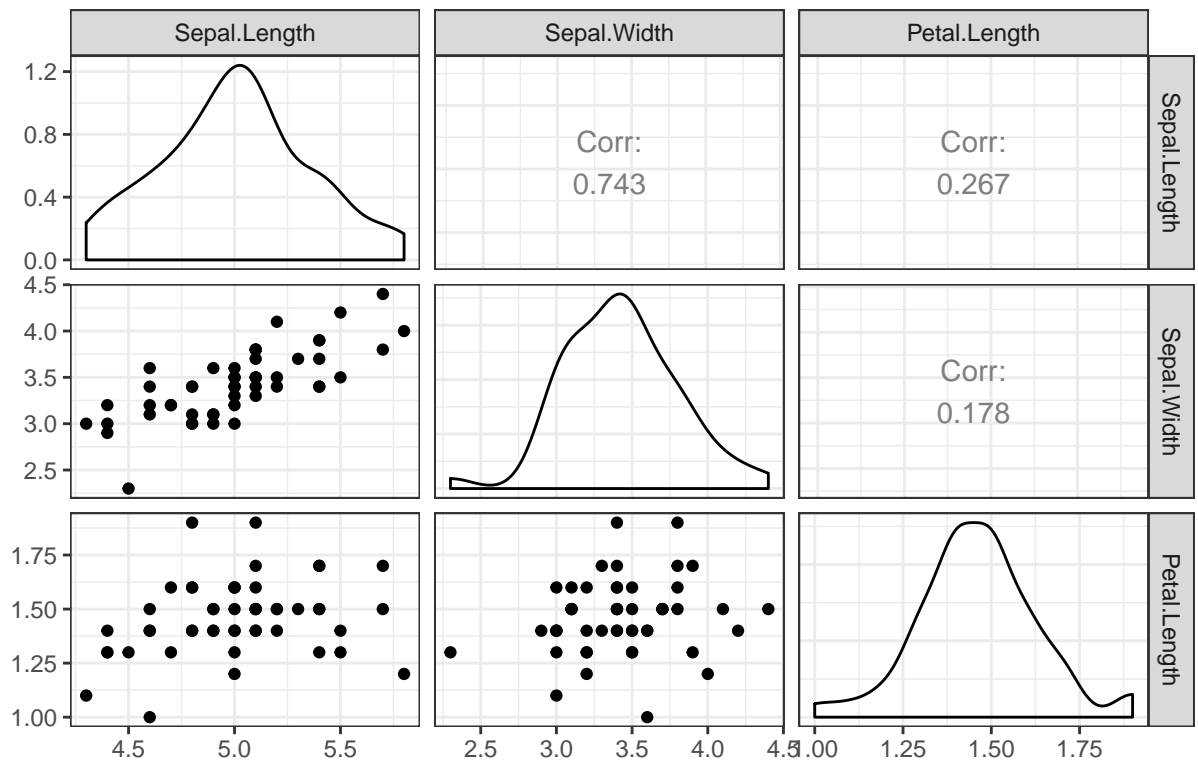
```

#we omitted this graph because it looked very heaped
# matplot_all2 <- iris %>%
#       ggpairs(mapping = aes(color = 'Species')) +
#       theme_light() +
#       labs(title = "Figura 1", x = "", y = "", color = "Especie de\n flor")

```

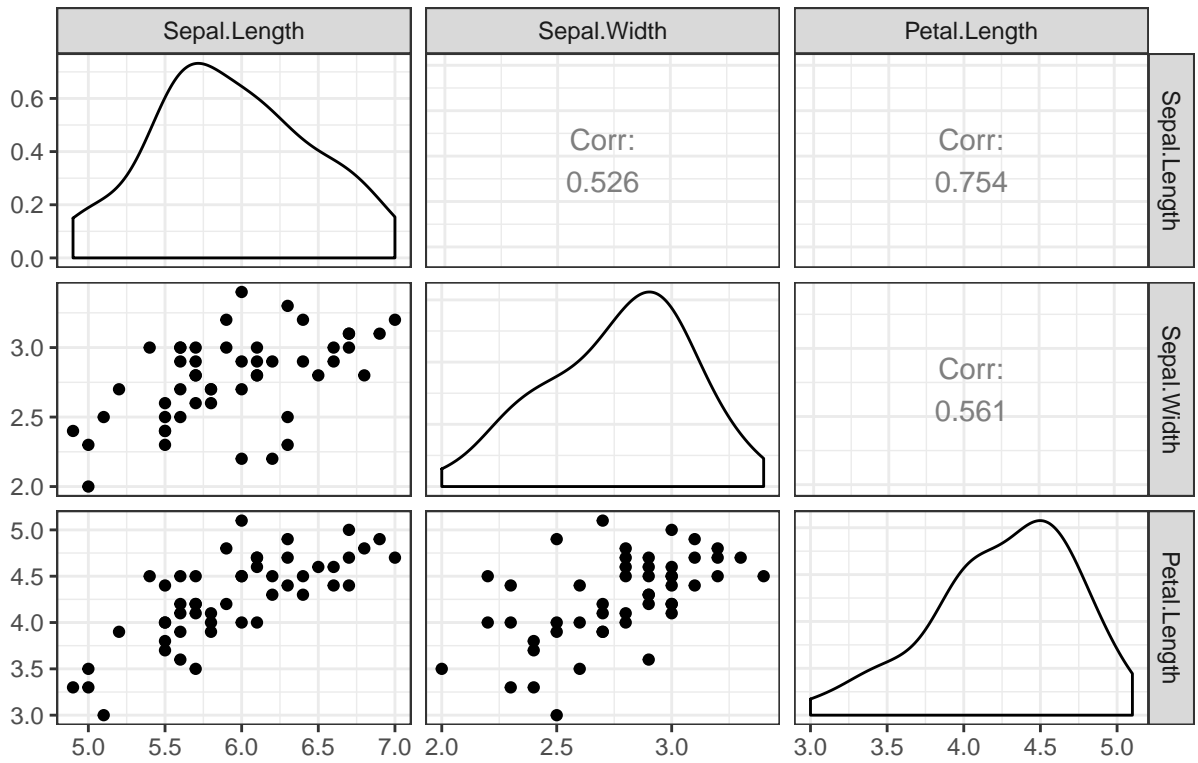
```
matplot_setosa
```

## SETOSA



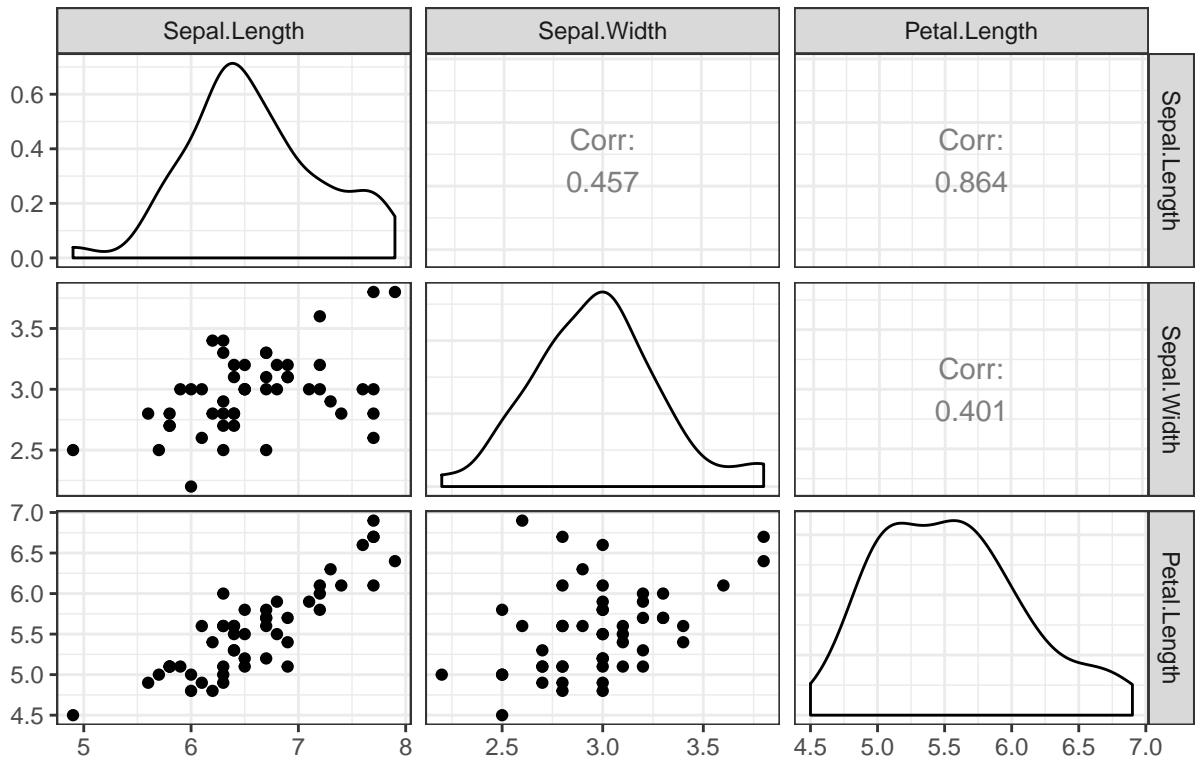
matplot\_veris

## VERSICOLOR



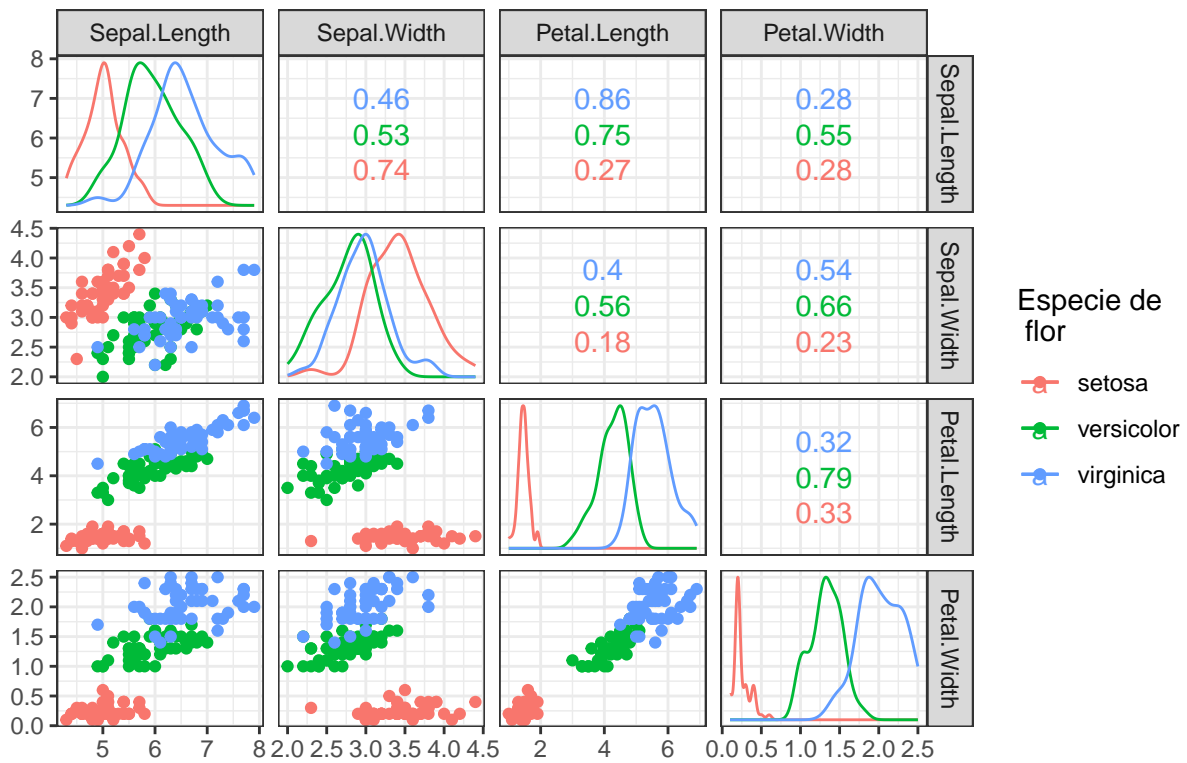
matplot\_virg

## VIRGINICA



matplot\_all

## FLORES



## Ejercicio 4

### Flores de Fisher y Anderson parte II.

4a

Sabemos que  $Y^5 = X^3 + X^4$ . Entonces, para encontrar  $C$  tal que

$$Y = XC,$$

se puede notar que  $C$  debe ser de la forma

```
# \[
# C=
# \left[
# \begin{array}{c|c}
# I_{\{4,4\}} & \\
# \begin{array}{c}
# 0 \ \ 0 \ \ 1 \ \ 1 \end{array}
# \end{array}
# \right]
```

```

Y <- cbind(X,(X[, 'Petal L.'] + X[, 'Petal W.']))
colnames(Y) <- c('Sepal L.', 'Sepal W.', 'Petal L.', 'Petal W.', 'PL + PW')

#C
C <- diag(x = 1, nrow = 4)
C <- cbind(C,c(0,0,1,1))

#check if Y = XC
testY <- Y == X %*% C

```

4b

```

#covariance matrix
dim_Y <- dim(Y)[1]
Sy <- (1/(dim_Y -1)) * t(Y) %*% Y

#eigenpair
eigen_listY <- eigen(Sy)

```

4c

```

testSy <- t(C) %*% Sx %*% C

```

## Ejercicio 5

### EDA de los Indicadores de la CNBV

La CNBV publica indicadores financieros de manera mensual con el objetivo de proporcionar estadísticas descriptivas para reflejar la evolución de la condición financiera de la Banca Múltiple.

```

doc1 <- "DatosCNBVModificados1.csv"
doc2 <- "DatosCNBVModificados2.csv"

data_cnbv1 <- read.csv(file = doc1)
data_cnbv2 <- read.csv(file = doc2)

```

### Tipo de variables

### Preguntas de investigación

- Observar si existe una diferencia en el promedio del *monto* y *flujo por mes* en los distintos periodos.
- Observar si existe una diferencia en el promedio del *monto* y *flujo por mes* en la Banca Múltiple.
- ¿Existirá una tendencia en la distribución del *monto* y *flujo por mes* en la Banca Múltiple?
- ¿Existe alguna relación entre *monto* y *flujo por mes*?
- ¿Hubo cambios en el orden de *montos* y *flujos por mes* en la Banca Múltiple?

## Patrones relevantes

```
# union of the two data to see duplicated data
colnames(data_cnbv2) <- colnames(data_cnbv1)
data_cnbv <- rbind(data_cnbv1,data_cnbv2)
data_cnbv$Monto[data_cnbv$Monto == ""] <- NaN
data_cnbv$Flujo_Mes[data_cnbv$Flujo_Mes == ""] <- NaN

#duplicated data
dupl_data <- duplicated(data_cnbv,incomparables = FALSE)
dup_count <- sum(as.numeric(dupl_data))

data_cnbv <- data_cnbv[-which(dupl_data),]

#Number of NA's
na_data <- is.na(data_cnbv$Monto) | is.na(data_cnbv$Flujo_Mes)
na_count <- sum(as.numeric(na_data))
data_cnbv <- data_cnbv[-which(na_data),]
```

## Reportar aberraciones en los datos

Primero que nada, se enfatiza que los datos en el archivo .csv publicado por la CNBV contiene datos reptidos, 1255, de igual manera contiene 296 datos faltantes, por lo que se optó por eliminar dichos datos. De igual manera (distribuciones)

Es de observar que