Tarea 1

Pablo,Sofía,Román

26/1/2020

Ejercicio 3

Las flores de Fisher y Anderson

3a

Se presenta a continuación la matrix de covarianza muestral insesgada $\mathbf{S}_X$, de los datos para la Iris Setosa.

```
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
A <- t(A) %*% A

#unbaised sample covarianse
Sx <- (1/(dim_set -1) * A)

kable(A,"markdown")
```

	Sepal L.	Sepal W.	Petal L.	Petal W.
Sepal L.	6.0882	4.8616	0.8014	0.5062
Sepal W.	4.8616	7.0408	0.5732	0.4556
Petal L.	0.8014	0.5732	1.4778	0.2974
Petal W.	0.5062	0.4556	0.2974	0.5442

3b

Los eigenvalores y eigenvectores de $\mathbf{S}_X$ son:

```
#eigenpar
eigen_list <- eigen(Sx)
kable(eigen_list$values, col.names = "Eigenvalores")
```

Eigenvalores
0.2364557
0.0369187
0.0267964
0.0090333

```
kable(eigen_list$vectors, "markdown", col.names = c("v1","v2","v3","v4"), label = "Eigenvectores")
```

v1	v2	v3	v4
-0.6690784	-0.5978840	0.4399628	-0.0360771
-0.7341478	0.6206734	-0.2746075	-0.0195503
-0.0965439	-0.4900556	-0.8324495	-0.2399013
-0.0635636	-0.1309379	-0.1950675	0.9699297

3c

Mostraremos que:

- $ULU^T = S_X$
- $U^TU = UU^T = I_{4\times4}$

Checamos entrada por entrada si el error relativo $\varepsilon = \frac{\|A-\tilde{A}\|_\infty}{\|\tilde{A}\|_\infty}$es menor a cierta tolerancia. Fijemos la tolerancia numérica de tol $= e^{-8}$ para la norma del supremo de la matriz a comparar.

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

#compute Sx = ULU'
Sx_prim = U %*% L %*% t(U)
testSx <- norm(Sx_prim - Sx,"I") / norm(Sx,"I")
```

Observamos que el error relativo $\varepsilon_1 = 3.2866807 \times 10^{-16}$ es menor que la tolerancia, por tanto se cumple la igualdad numéricamente.

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

test_uut <- norm(UUt - diag(1,nrow = 4),"I") / norm(UUt,"I")
test_utu <- norm(UtU - diag(1,nrow = 4),"I") / norm(UtU,"I")
```

También podemos ver que el error relativo $\varepsilon_2 = 1.7746221 \times 10^{-15}$ y el error relativo $\varepsilon_2 = 1.6115581 \times 10^{-15}$ es menor que la tolerancia, por tanto se cumple la igualdad numéricamente para ambos casos.

3d

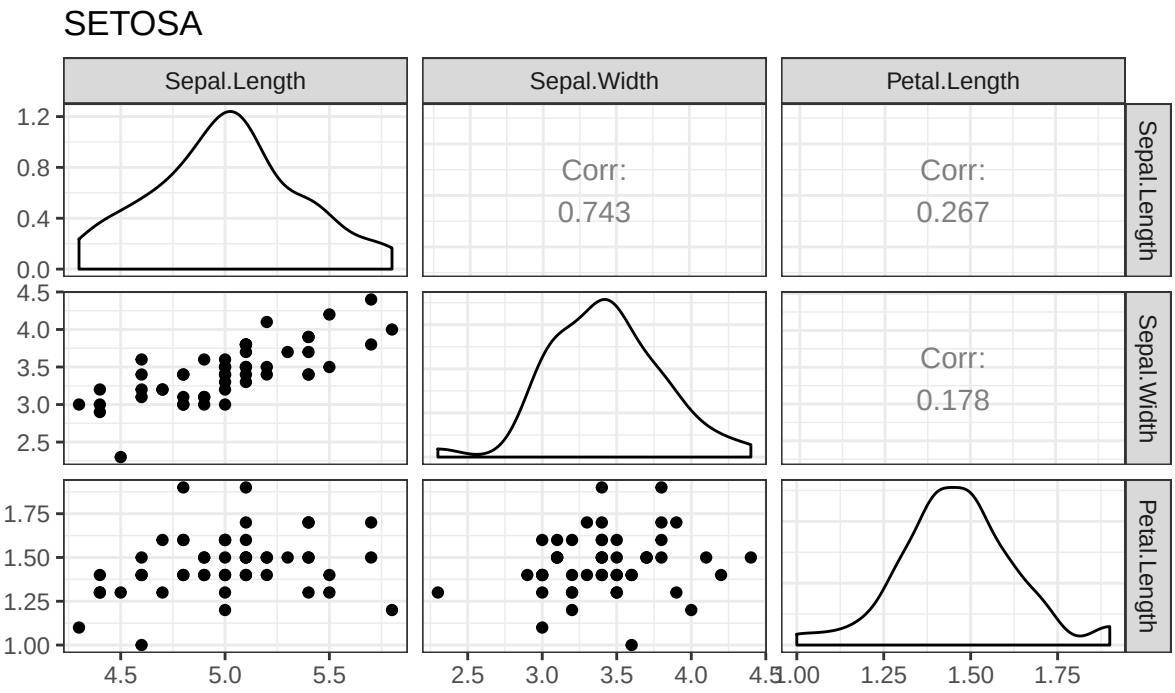
Se mostrará un cuatro gráficas de dispersión. Una para cada categoría de Iris y la última es una gráfica conjunta. Esto con el mootivo de estudiar por separado y con mayor legibilidad las gráficas por cada tipo y la última con el fin de comparar la distribución para cada especie de flor.

```
matplot_setosa <- iris %>%
                  filter(Species == "setosa") %>%
                  select(Sepal.Length, Sepal.Width, Petal.Length) %>%
```

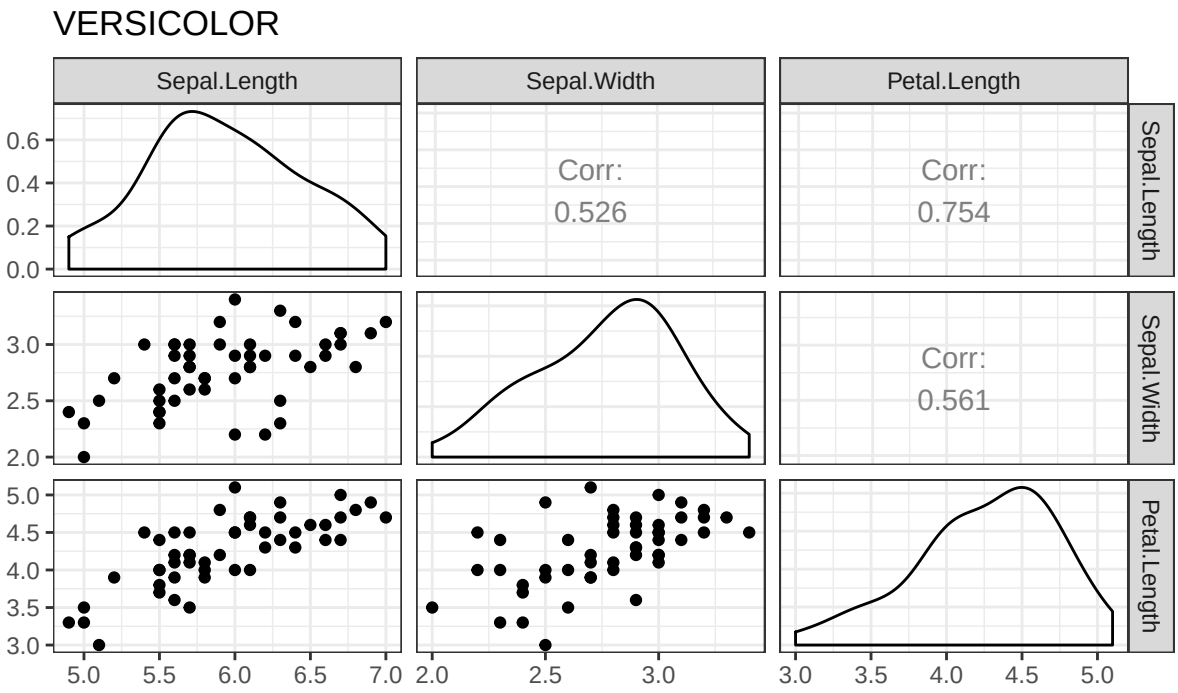
```
                  ggpairs() +
                  theme_bw() +
                  labs(title = "SETOSA", x = "", y = "") +
                  theme(plot.margin = margin(1,.3,1,.3,"cm"))

matplot_veris <- iris %>%
                  filter(Species == "versicolor") %>%
                  select(Sepal.Length, Sepal.Width, Petal.Length) %>%
                  ggpairs() +
                  theme_bw() +
                  labs(title = "VERSICOLOR", x = "", y = "") +
                  theme(plot.margin = margin(1,.3,1,.3,"cm"))

matplot_virg <- iris %>%
                  filter(Species == "virginica") %>%
                  select(Sepal.Length, Sepal.Width, Petal.Length) %>%
                  ggpairs() +
                  theme_bw() +
                  labs(title = "VIRGINICA", x = "", y = "") +
                  theme(plot.margin = margin(1,.3,1,.3,"cm"))



matplot_all <- iris %>%
                  ggscatmat(color = 'Species') +
                  theme_bw() +
                  labs(title = "FLORES", x = "", y = "", color = "Especie de\n flor") +
                  theme(plot.margin = margin(1.1,.35,1.1,.35,"cm"))
```

```
## 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
```

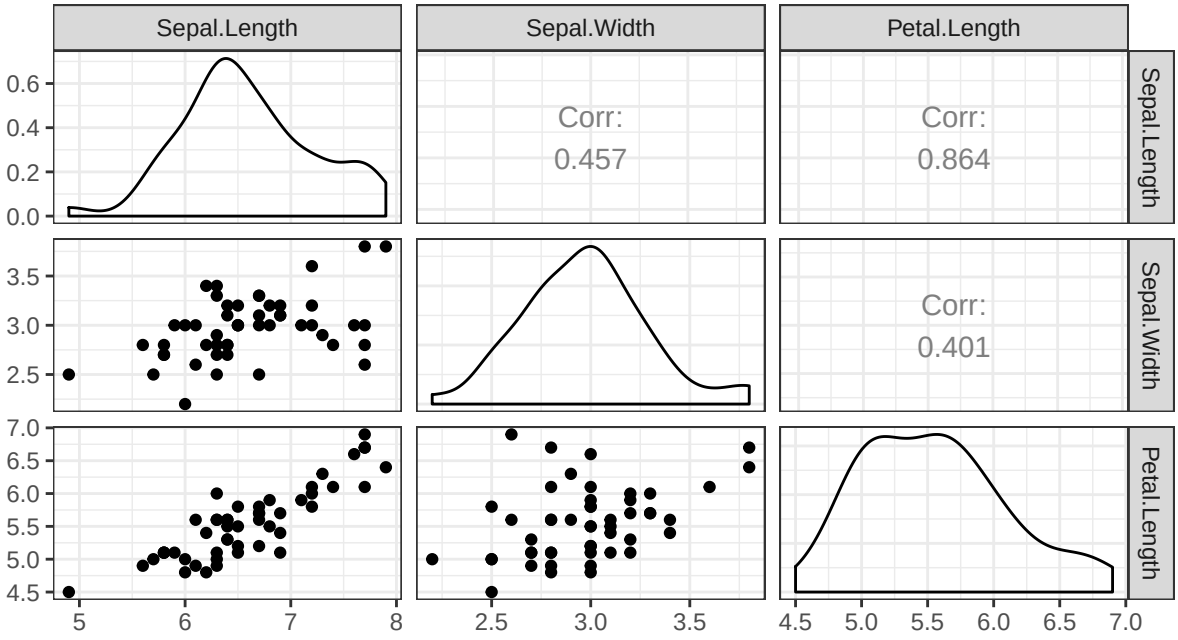



```
matplot_veris
```

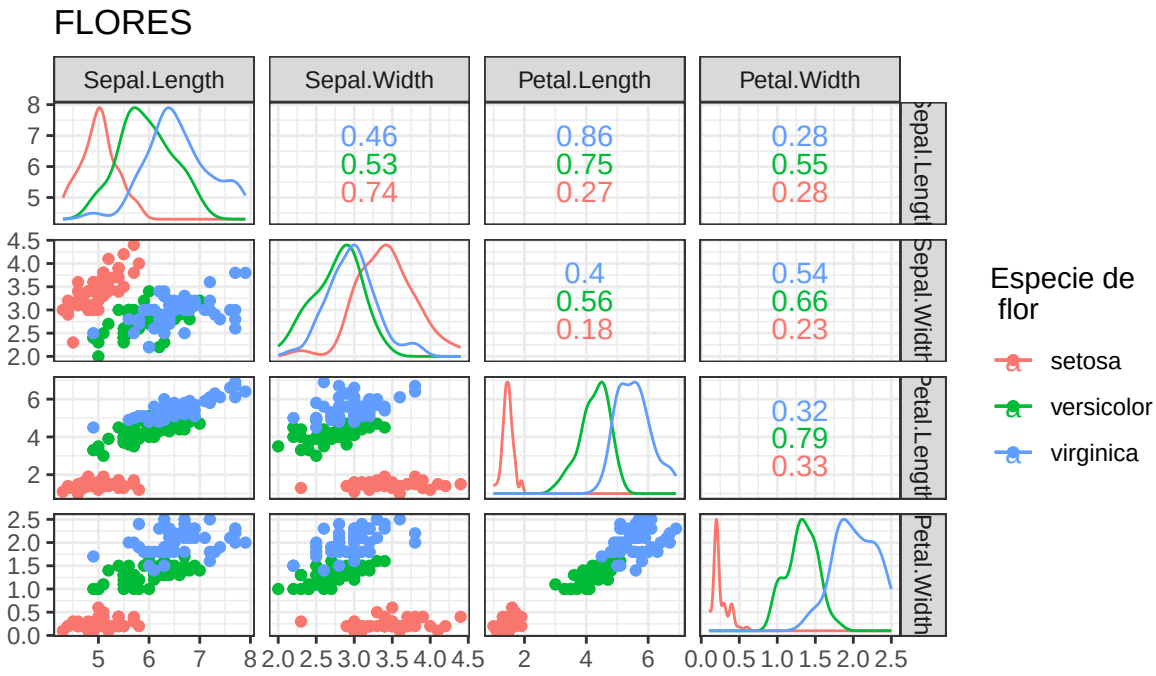



```
matplot_virg
```

VIRGINICA




```
matplot_all
```



Se observa que hay una relación lineal positiva entre el largo y ancho del petalo, como el largo y el ancho del sépalo. Hay correlaciones positivas para estos casos, donde las corelaciones muestrales mayores se presentan en la especie virgínica.

Ejercicio 4

Flores de Fisher y Anderson parte II.

4a

Sabemos que $Y^5 = X^3 + X^4$. Entonces, para encontar 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 <- norm(Y - X %*% C,"I")/norm(X %*% C,"I")
```

Es de notar que el error relativo $\varepsilon_4 = 0$ es menor que la tolerancia, por tanto $Y = XC$, numéricamente.

4b

La matriz de covarianzas corregidas esta dada por

```
# 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
# #unbaised sample covarianse
# Sx <- (1/(dim_set -1) * A)
#
# kable(A,"markdown")

#covariance matrix of Y

#mean of columns and dim
s_mnY <- apply(X = Y, MARGIN = 2, FUN = mean)
dim_Y <- dim(Y)[1]
s_mn_matrixY <- matrix(rep(s_mnY,each = dim_Y), nrow = dim_Y)


#corrected mean squear
B <- Y - s_mn_matrixY
Sy <- (1/(dim_Y -1)) * t(B) %*% B

kable(Sy)
```

	Sepal L.	Sepal W.	Petal L.	Petal W.	PL + PW
Sepal L.	0.1242490	0.0992163	0.0163551	0.0103306	0.0266857
Sepal W.	0.0992163	0.1436898	0.0116980	0.0092980	0.0209959
Petal L.	0.0163551	0.0116980	0.0301592	0.0060694	0.0362286
Petal W.	0.0103306	0.0092980	0.0060694	0.0111061	0.0171755
PL + PW	0.0266857	0.0209959	0.0362286	0.0171755	0.0534041

.

Los eigenpares de la matriz S_Y están dados por

```
#eigenpair
eigen_listY <- eigen(Sy)

kable(eigen_listY$values, col.names = "Eigenvalores")
```

Eigenvalores
0.2442194
0.0748382
0.0330587
0.0104918
0.0000000

```
kable(eigen_listY$vectors, "markdown", col.names = c("v1","v2","v3","v4","v5"), label = "Eigenvectores")
```

v1	v2	v3	v4	v5
-0.6569405	0.0384643	0.7528919	0.0101712	0.0000000
-0.7118498	0.2954275	-0.6365906	0.0272931	0.0000000
-0.1252167	-0.5479581	-0.0891735	0.5854720	-0.5773503
-0.0755588	-0.2060501	-0.0447951	-0.7851715	-0.5773503
-0.2007755	-0.7540082	-0.1339687	-0.1996995	0.5773503

4c

Igualmente mostraremos numéricamente que $S_Y = C^T S_X C$ con la norma del supremo.

```
#compare relative error
tol <- exp(-8)
testSy <- (norm(t(C) %*% Sx %*% C - Sy, "I")) / (norm(t(C) %*% Sx %*% C,"I"))
```

Es de notar que el error relativo $\varepsilon_5 = 1.5831216 \times 10^{-16}$ es menor que la tolerancia, entonces se cumple la igualdad.

Ejercicio 5

EDA de los Indicadores de la CNBV

La CNBV publica indicadores financieros de manera mensual con el objetivo de 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